密级：　　保密期限：

北京郵電大學

博士学位论文

题目：城市视频监控网络中车辆搜索关键技术研究

学　　号：2013010114

姓　　名：刘鑫辰

专　　业：计算机科学与技术

导　　师：马华东

学　　院：计算机学院

二〇一八年六月十五日

独创性（或创新性）声明

本人声明所呈交的论文是本人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢中所罗列的内容以外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得北京邮电大学或其他教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

申请学位论文与资料若有不实之处，本人承担一切相关责任。

本人签名：_____ 日期：2018.6.21

关于论文使用授权的说明

本人完全了解并同意北京邮电大学有关保留、使用学位论文的规定，即：北京邮电大学拥有以下关于学位论文的无偿使用权，具体包括：学校有权保留并向国家有关部门或机构送交论文，有权允许学位论文被查阅和借阅；学校可以公布学位论文的全部或部分内容，有权允许采用影印、缩印或其它复制手段保存、汇编学位论文，将学位论文的全部或部分内容编入有关数据库进行检索。(保密的学位论文在解密后遵守此规定)

本学位论文不属于保密范围，适用本授权书。

本人签名：_____ 日期：2018.6.21

导师签名：_____ 日期：2018.6.21

城市视频监控网络中车辆搜索关键技术研究

摘 要

车辆是城市视频监控系统中重要的目标。近年来，监控视频中车辆相关的研究逐渐成为热点，如车辆检测、车辆跟踪、车辆分类、车牌识别等。车辆搜索，即给定一个查询车辆，在城市视频监控网络中搜索与其身份相同的车辆，可以帮助管理人员快速准确地在城市中寻找、定位、跟踪目标车辆。然而，车辆搜索面临“搜不准”和“搜得慢”两大挑战。“搜不准”，一方面是由于车辆自身外观的相似性和多样性，另一方面是由于无约束城市监控中多变的环境因素。“搜得慢”，不仅是由于城市监控视频巨大的数据规模，而且由于车辆多模态特征的复杂性。

因此，本文针对城市视频监控网络提出了一种融合多模态数据的渐进式车辆搜索框架，并从车辆外观特征的表示与学习、车辆唯一标识即车牌的有效利用、监控网络中时空关系的挖掘三个方面提出了一系列方法与模型。通过在城市监控数据上的大量实验，我们验证了所提出框架与方法的准确性与高效性。本文的主要贡献具体如下：

（1）融合多模态数据的渐进式车辆搜索框架。该框架综合特征域和时空域进行逐步求精地搜索，具体来说：一是特征域内由粗到精地搜索，即先采用外观特征快速查找相似车辆，再使用车牌信息实现精确搜索；二是在时空域内，利用监控网络中的时空信息由近及远地搜索。实验分析表明，这种渐进式搜索框架不仅能够显著降低车辆搜索的时间消耗，同时保证了车辆搜索的准确性。

（2）基于车辆外观特征的相似车辆搜索方法。针对抓拍图像和视频两种查询数据，我们分别提出了两种基于深度卷积神经网络（Convolutional Neural Network，CNN）的车辆外观表示方法：NuFACT 和 CAN。NuFACT 方法能够从抓拍车辆图像中提取车辆的纹理、颜色、类别等多级特征，并通过零空间度量学习将上述特征融合为一种具有区分力的、鲁棒的特

征。CAN 方法能够提取视频中多张图像的共有信息和互补信息，自动学习视角不变性特征，增强了车辆外观特征的区分力和鲁棒性。

（3）车牌图像超分辨率与验证结合的精确车辆搜索方法。针对无约束监控环境中低质的车牌图像，我们提出了一种基于域先验生成对抗网络的图像超分辨率方法进行车牌图像增强。针对监控数据中车辆数量很大而每个车辆样本较少的问题，本文采用一种基于对偶神经网络（Siamese Neural Network，SNN）的车牌验证方法，实现了车牌图像的快速准确匹配。通过将车牌增强与验证结合，进一步提高了车辆搜索的准确性。

（4）基于邻接图与时空相似度模型的搜索结果重排序。通过挖掘城市监控网络中的时空信息，如车辆被拍摄的时间、摄像头的位置、摄像头邻接关系等，我们设计了一种摄像头邻接图模型表示视频监控网络的空间拓扑，提出了一种基于多层感知机的时空相似度模型（Spatio-Temporal Similarity Model，STSM），通过 STSM 估计车辆间的时空相似性对搜索结果进行重排序，得到优化的车辆搜索结果。

最后，我们构建了一个融合多模态数据的渐进式车辆搜索原型系统，并在真实视频监控数据上验证了上述框架与方法的有效性。

关键词：车辆搜索，车辆重识别，渐进式搜索，视频监控网络，多模态数据

RESEARCH ON KEY TECHNIQUES OF VEHICLE SEARCH IN URBAN VIDEO SURVEILLANCE NETWORKS

ABSTRACT

Vehicles have become an indispensable part of human life as well as a significant class of objects in urban surveillance systems. Many researchers in multimedia and computer vision fields have focused on vehicle-related research, such as detection, tracking, fine-grained categorization, and license plate recognition. Nevertheless, vehicle search or re-identification, which can find the same vehicle in large-scale surveillance videos with a given query, is an important but frontier area. Through the ubiquitous surveillance network, vehicle search can quickly tell users where and when the vehicle was in the city. However, the main challenge is how to guarantee both of accuracy and efficiency in vehicle search. On one hand, the variety and similarity of vehicles in uncertain environments make it difficult to match the same vehicle accurately. On the other hand, efficient vehicle search is a challenging task due to the huge volume of surveillance data and complexity of vehicle features.

This thesis presents a multi-modal and progressive vehicle search framework for large-scale urban surveillance scenes. In particular, we propose a series of models and approaches from three aspects of vehicle search: vehicle appearance feature learning and representation, license plate verification based vehicle matching, and vehicle search re-ranking based on spatiotemporal information in video surveillance networks. Furthermore, through extensive experiments on a large-scale dataset collected from real-world surveillance system, we validate the accuracy and efficiency of the proposed framework and methods for vehicle search. The main contributions of this thesis are as follows:

(1) Multi-modal and progressive vehicle search framework. The proposed framework utilizes the multi-modal data in large-scale video surveillance networks to achieve progressive search in both feature domain and spatiotemporal

domain. Specifically, the progressive search process includes two aspects. One is the coarse-to-fine search in the feature space, i.e., first obtaining similar vehicles with appearance features, then matching the target vehicles with license plates. The other is the near-to-distant search with spatiotemporal information in surveillance networks. Extensive experiments demonstrate that our framework can achieve accurate vehicle search and reduce the time cost through the progressive search manner.

(2) Appearance feature based similar vehicle search. We propose two novel deep CNN-based approaches, i.e., NuFACT and CAN, for single-shot and video-based vehicle search, respectively. For single-shot vehicle search, the NuFACT can extract multi-level appearance features from vehicle images. Then a null space-based metric learning method is adopted to fuse these features as a discriminative and robust representation. For video-based vehicle search, the CAN can learn common features and complementary features from multiple frames in vehicle videos. Then the significant features from images of different viewpoints and resolutions can be enhanced by the attention mechanism to make the fused features more separable in the feature space.

(3) Accurate vehicle search by combining license plate super-resolution and verification. To overcome the low quality of license plates captured in unconstrained surveillance scenes, a domain prior generative adversarial network for license plate super-resolution (SR) is designed to generate a high-resolution plate from a low-resolution one. Moreover, since there are large numbers of vehicles but small numbers of samples for each vehicle, a Siamese Neural Network (SNN) based plate verification method is proposed for accurate plate verification instead of recognition. By integrating license plate SR and verification, the accuracy of vehicle search is significantly improved.

(4) Search result re-ranking based on camera neighboring graph and spatiotemporal similarity. To achieve near-to-far search in physical space, contextual information, such as the timestamps, locations of cameras, distances between neighboring cameras, is exploited to build a neighboring graph for rep-

resenting the topology of the surveillance network. Furthermore, we propose a multi-layer perceptron based the spatiotemporal similarity model (STSM) to estimate the spatiotemporal similarity between two vehicles. Finally, the results of appearance and license plate-based vehicle search are re-ranked to obtain the optimized results.

To validate the proposed framework and methods, we build a prototype system of multi-modal data based progressive vehicle search. Extensive experiments on data from real video surveillance system demonstrate that the proposed progressive framework can find the target vehicle accurately and efficiently.

KEY WORDS: Vehicle Search, Vehicle Re-identification, Progressive Search, Video Surveillance Networks, Multi-modal Data

目 录

第一章 绪论 1
1.1 引言 1
1.2 车辆搜索 2
1.2.1 车辆搜索系统结构 2
1.2.2 车辆搜索基本特点 4
1.2.3 车辆搜索研究内容 5
1.3 面向视频监控网络的车辆搜索相关工作 7
1.3.1 车辆特征表示 7
1.3.2 视频监控网络中对象搜索 9
1.4 研究内容与主要贡献 11
1.5 论文结构 13
第二章 基于多级外观特征融合的相似车辆搜索 15
2.1 引言 15
2.2 问题描述 16
2.3 车辆多级外观特征表示 17
2.3.1 纹理特征 17
2.3.2 颜色特征 18
2.3.3 语义属性特征 18
2.4 基于零空间度量学习的多级特征融合 19
2.4.1 零空间度量学习 19
2.4.2 多级特征融合 20
2.5 实验结果与分析 21
2.5.1 数据集 21
2.5.2 实验设置 25
2.5.3 方法对比 26
2.6 本章小结 31

第三章　基于跨视角注意力神经网络的相似车辆搜索 33
3.1　引言 33
3.2　问题描述 36
3.3　面向视频车辆搜索的跨视角注意力网络框架 37
3.3.1　特征学习网络 37
3.3.2　注意力聚合网络 39
3.3.3　网络训练 40
3.4　实验结果与分析 42
3.4.1　数据集与实验设置 42
3.4.2　基于单张图像的方法与基于视频的方法对比 43
3.4.3　基于视频的车辆搜索方法对比 46
3.4.4　讨论 48
3.5　本章小结 49
第四章　车牌图像增强与验证结合的精确车辆搜索 51
4.1　引言 51
4.2　相关工作 55
4.2.1　车牌识别 55
4.2.2　图像超分辨率 56
4.3　基于域先验生成对抗网络的车牌图像增强 57
4.3.1　域先验生成对抗网络框架 58
4.3.2　生成器网络 59
4.3.3　判别器网路 59
4.3.4　对抗损失函数 60
4.4　基于车牌验证的精确车辆搜索 61
4.4.1　对偶神经网络结构 61
4.4.2　网络训练 62
4.4.3　精确车辆搜索 63
4.4.4　车辆搜索与车牌增强的结合 63
4.5　实验结果与分析 64
4.5.1　数据集 64
4.5.2　实验设置 65

4.5.3 方法对比 65
4.6 本章小结 71
第五章 多模数据融合的渐进式车辆搜索系统 73
5.1 应用背景 73
5.2 多模数据融合的渐进式车辆搜索框架 75
5.3 多模数据融合的渐进式车辆搜索原型系统 77
5.3.1 车辆数据收集子系统 78
5.3.2 车辆搜索子系统 80
5.3.3 系统运行环境 82
5.4 系统测试 82
5.4.1 车辆检测模块 83
5.4.2 车辆搜索系统 83
5.4.3 搜索效率分析 86
5.5 本章小结 86
第六章 总结与展望 89
6.1 论文工作总结 89
6.2 未来工作展望 90
参考文献 93
致 谢 101
攻读学位期间发表的学术论文目录 103

表格索引

1-1 互联网搜索与物联网搜索对比 4
2-1 不同方法在 VeRi 数据集上的结果比较 28
2-2 不同方法在 VehicleID 数据集上的结果比较..... 29
3-1 基于单张图像的车辆搜索方法与基于视频的车辆搜索方法结果对比..... 45
3-2 不同外观相似车辆搜索方法在 VIVID 数据集上的准确率对比 47
4-1 不同方法在 VeRi 数据集上的车辆搜索结果比较 71
4-2 不同方法在 VeRi 数据集上的车辆搜索结果比较 72
5-1 不同方法在 VeRi 数据集上的结果对比 84

插图索引

1-1 车辆搜索示例：利用多模态数据在城市监控视频中搜索目标车辆 2
1-2 面向城市视频监控网络的车辆搜索系统基本结构.....3
1-3 车辆搜索的典型应用 5
1-4 车辆特征表示相关研究 8
1-5 视频监控中的人员搜索研究..... 10
1-6 车辆搜索面临的挑战 12
1-7 论文章节结构与关系图 14
2-1 城市监控摄像头抓拍的车辆图像示例..... 16
2-2 基于多级特征融合与零空间度量学习的车辆搜索框架 16
2-3 基于零空间度量学习的多级特征融合..... 20
2-4 VeRi 数据集车辆图像样本示例 22
2-5 VeRi 数据集轨迹统计分布 23
2-6 VeRi 数据集中车辆颜色与类别比例..... 23
2-7 VeRi 数据集中摄像机空间距离标注..... 24
2-8 VehicleID 数据集中的车辆图像样本..... 25
2-9 不同方法在 VeRi 数据集上的 CMC 曲线对比..... 28
2-10 不同方法在 VehicleID 数据集上的 CMC 曲线对比 29
2-11 NuFACT 在 VeRi 数据集上的搜索结果示例..... 30
3-1 跨视角注意力网络（CAN）的动机 34
3-2 跨视角注意力网络框架 38
3-3 跨视角注意力网络的训练过程 41
3-4 VIVID 数据集中的车辆视频数据样本 42
3-5 基于单张图像与基于视频方法在 VIVID 数据集上的 CMC 曲线对比..... 45
3-6 VIVID 数据集中的测试样本及对应的注意力权重..... 48
4-1 利用车牌信息的精确车辆搜索 52
4-2 基于车辆搜索与域先验生成对抗网络的车牌超分辨率基本思想 52
4-3 基于车辆搜索与域先验生成对抗网络的车牌图像增强方法结果展示..... 53
4-4 基于对偶神经网络的车牌验证与精确车辆搜索框架 55

4-5 基于车辆搜索和域先验生成对抗网络的车牌图像增强框架 58
4-6 判别网络中空间分割层的结构 60
4-7 用于车牌匹配的对偶神经网络结构与训练 62
4-8 通过 EasyPR 软件对不同超分辨率方法的评估 67
4-9 通过 EasyPR 软件对基于车辆搜索的多张车牌超分辨率方法的评估 69
4-10 不同车牌超分辨率方法的生成结果对比 70
5-1 渐进式车辆搜索系统应用：嫌疑车辆搜索 74
5-2 多模数据融合的渐进式车辆搜索框架 75
5-3 视频监控网络及对应的摄像头邻接图示例 76
5-4 VeRi 数据集的时空信息统计 77
5-5 渐进式车辆搜索原型系统总体设计 78
5-6 监控摄像头间的空间距离矩阵 80
5-7 车辆搜索系统输入处理与结果显示界面 81
5-8 车辆检测结果实例图 83
5-9 渐进式车辆搜索框架在 VeRi 数据集的搜索结果实例 85
5-10 渐进式搜索时取不同比例候选车辆的时间消耗与 mAP 86

第一章 绪论

1.1 引言

随着我国经济的快速发展，机动车已经成为人民生产生活中不可缺少的组成部分。据统计，截至 2016 年末，全国民用汽车保有量已达 1.86 亿辆，公路交通完成旅客运输量 154.3 亿人次、货物运输量 334.1 亿吨，汽车是我国承担旅客和货物运输的最主要方式[1]。但是，机动车快速增长的同时带来了交通事故、交通拥堵等社会问题，这些问题不仅阻碍经济的发展，更威胁着人民生命财产安全。据统计，2016 年我国发生机动车交通事故 21.3 万起、死亡 5.88 万人，直接财产损失 11.5 亿元[1]。为确保公路交通安全、高效运行，视频监控系统已在我国各级城市道路广泛部署，在交通管理、路况监控、违章取证、犯罪侦查等方面起到至关重要的作用。

为实现城市交通监控系统的信息化、智能化，视频监控系统中车辆的相关研究引起了工业界和学术界的广泛关注。近些年，在多媒体、机器视觉、图像处理、智能交通领域的重要国际会议 (如 ACMMM，CVPR，ICCV，ECCV，ITSC) 和主流国际期刊 (如 IEEE Transactions on Multimedia，IEEE Transactions on Image Processing，IEEE Transactions on Intelligent Transportation Systems) 上陆续刊登了一些视频监控中车辆相关的学术论文。智能交通领域旗舰会议 IEEE ITSC 一直将面向车辆的图像分析作为专门的环节。CVPR 2016 组织了自动交通监控研讨会（Automatic Traffic Surveillance Workshop）供专家学者交流。CVPR 2017 组织了交通监控研讨与挑战赛（Traffic Surveillance Workshop and Challenge），吸引了众多研究人员参与车辆定位、分类等挑战项目。目前学术界关注的热点问题包括：监控图像中的车辆检测[2]、车辆分类[3]、视角估计[4]、车牌识别[5]，监控视频中的车辆跟踪[6]、车速估计[7]、行为分析[8]，监控网络中的轨迹分析[9]、摄像头校正[10]，等等。此外，各大安防设备厂商也推出了多款智能交通摄像头，如浙江大华的“易”系列卡口一体化抓拍单元[11]、海康威视的 iDS-2CD9371-K(S) 智能交通摄像机[12] 等，将较为成熟的卡口车牌识别、车型识别、颜色识别、违章抓拍等技术集成到摄像头终端中。与上述研究课题相比，大规模城市监控网络中的车辆搜索同样是一个十分重要和值得研究的课题。

图 1-1　车辆搜索示例：利用多模态数据在城市监控视频中搜索目标车辆

1.2　车辆搜索

车辆搜索（Vehicle Search），是指输入一个查询车辆描述（包括属性、车牌、图像或视频等），并指定搜索时间段与空间范围，在大规模城市视频监控网络中搜索与查询车辆身份相同的车辆。如图1-1所示，车辆搜索系统能够告知用户目标车辆在何时何地出现。通过车辆搜索，交通管理部门可以快速、准确、便捷地在海量交通监控数据中发现、定位、跟踪目标车辆。因此，本文以城市视频监控网络为背景，研究以车辆为目标的物理实体搜索问题中关键技术。

下面，本节将从车辆搜索的系统结构、基本特点、主要研究内容三个方面进行简要介绍。

1.2.1　车辆搜索系统结构

如图1-2所示，一个典型的车辆搜索系统通常由视频监控网络、车辆搜索引擎、用户三部分构成。其中，视频监控网络是感知物理世界的基础设施，在车辆搜索系统中视频监控网络通过车辆检测、跟踪等技术，将出现在城市中的车辆提取并发送到搜索引擎进行处理。搜索引擎由位于数据中心的多个服务器组成，其基本功能包括数据存储、图像处理、对象检测、特征提取、车牌识别、车辆索引等。用户包括一般社会用户、城市管理者、系统管理员，用户可以输入不同的查询信息（如车辆号牌、抓拍图像、特征描述等），系统根据不同的特征粒度与权限范围实现不同层次的

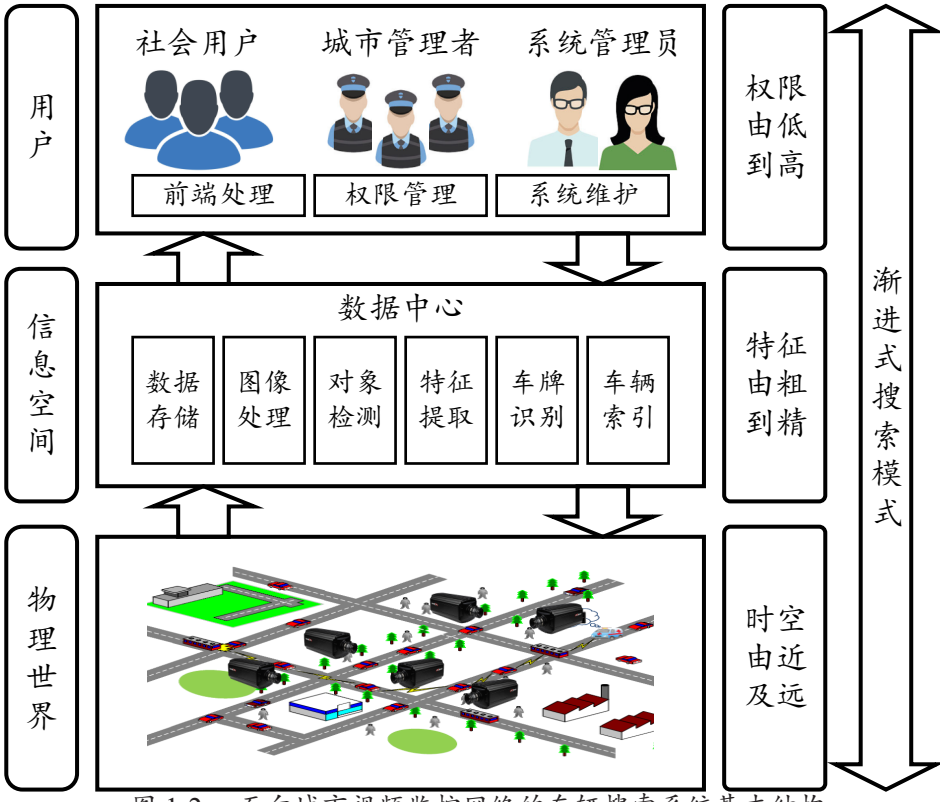


图 1-2 面向城市视频监控网络的车辆搜索系统基本结构

搜索服务。

系统的工作流程主要包括三个步骤：

- 视频监控网络通过部署在城市中的摄像头检测、跟踪拍摄到的车辆，并将车辆图像、时间戳、空间位置等信息实时发送到数据中心进行存储和处理；
- 部署于数据中心的搜索引擎接收到图像和时空信息等多模态数据后，提取车辆的视觉特征、车牌信息、时空情景信息，并根据上述信息建立多级索引，实现高效检索；
- 用户通过客户端登录并访问搜索系统，输入抓拍图像、语义描述、车牌号码、时空信息等查询信息，系统根据用户权限不同为用户提供相似车辆搜索、精确车辆搜索等不同粒度的搜索服务。

为保证城市视频监控网络中车辆搜索的高效性、准确性、安全性，车辆搜索系统遵循渐进式的搜索模式，如图1-2所示。具体来说，在用户层面，系统根据权限的由低到高，提供不同粒度、不同安全级别的搜索方式和搜索结果。在特征空间中，车辆搜索采用由粗到精地方式，即在车辆匹配时先采用车辆外形、颜色、类别等外观

表 1-1　互联网搜索与物联网搜索对比

特点	互联网搜索	物联网搜索
搜索目标	信息实体	物理对象
交互方式	人-机交互	人-机-物交互
数据来源	信息网络空间	信息网络空间 + 物理世界
结果呈现	信息排序	对象描述 + 时空状态
实时性	非实时	实时

特征快速查找外观相似的车辆，再使用车牌信息进行精确的车辆搜索。在物理世界中，系统根据查询车辆的时空信息由近及远地搜索，利用时空相似度模型对搜索结果进行重排序。

1.2.2　车辆搜索基本特点

对于传统互联网搜索引擎，用户能够实现信息空间或网络空间中的搜索。车辆搜索作为物联网搜索的一个具体实例，其本质是实现对物理世界中对象的搜索[13]。物联网搜索与互联网搜索的对比，如表1-1所示。因此，车辆搜索的主要特点主要包括:

- **搜索目标：**车辆搜索的搜索需求更加全面，不仅包括信息描述（如车辆的类型、颜色、车牌等），而且包括车辆在物理世界的时间、地点、状态等;
- **交互方式：**车辆搜索需要实现用户、信息网络空间、物理世界的交互，不仅包括从物理世界到信息网络空间再到用户的被动信息提供，也包括从用户到信息网络空间再到物理世界的主动式参与;
- **数据来源：**除了传统信息空间的数据来源，车辆搜索的主要数据来源是通过部署在城市中的视频监控网络感知的多模态数据，相比互联网搜索具有更大的规模、更高的复杂度及动态特性;
- **结果呈现：**车辆搜索对用户呈现的结果除了包括与传统互联网搜索类似的车辆图像、类型、颜色、品牌等信息，还要包括车辆存在的位置、时间及状态等。
- **实时性：**车辆物理世界中的时空位置、状态等是实时变化的，车辆搜索系统通过视频监控系统对物理世界的感知也是实时进行的。因此，车辆搜索系统需要查询的信息是实时的。

(a) 犯罪侦查　　(b) 停车管理

(c) 智能交通　　(d) 城市计算

图 1-3　车辆搜索的典型应用

车辆搜索可以广泛应用于智能交通、停车管理、犯罪侦查、城市计算等方面[14–16]，如图1-3所示。

1.2.3　车辆搜索研究内容

作为新兴的研究领域，物联网搜索在物理对象感知、对象特征提取、多模态数据挖掘、高效索引构建、实际应用等层面都面临着传统互联网搜索中不曾遇到的挑战。目前，车辆搜索作为物联网搜索的重要实例，主要围绕以下几方面展开研究：

（1）多媒体传感网感知理论

多媒体传感网，如城市视频监控网络，是车辆搜索系统感知物理世界的基础。目前，学术界和工业界针对多媒体传感网的感知理论已进行了深入的研究，包括多媒体传感器的覆盖分析、节点协作，网络的传输策略、路由算法等[17]。多媒体传感器的覆盖与协作通常针对场景中目标的定位、跟踪、分类等问题，在资源有限条件下实现对感知对象的准确定位、跟踪等任务。多媒体传感网的传输、组网策略及路由算法，则根据多媒体感知数据的大规模、实时性特点，在网络资源有限的条件下实现高效、快速的数据传输。

（2）基于视觉信息的车辆外观特征表示

视觉信息是视频监控网络能够获取的最主要信息，因此车辆的视觉特征表示是基于外观车辆搜索的重要研究内容。计算机视觉和多媒体领域针对监控图像中的车辆检测、分类、视角估计、跟踪等问题提出了许多视觉表示模型。早期的视觉特征主要为人工设计的特征描述子，如颜色直方图、尺度不变特征、梯度直方图等，特征的表达能力和对监控环境的鲁棒性较差。最近几年，深度卷积神经网络（CNN）通过有监督学习的方式从大规模数据中学习有效的特征表示，在计算机视觉领域取得重要进展[18]。通过引入 CNN 模型，车辆外观表示也取得了巨大的进步，并在车辆检测、分类、视角估计等问题上取得了优秀的结果[3,19].

（3）面向监控视频的车牌信息提取

车辆牌照是识别车辆身份的唯一性标识。基于图像的车牌识别已经发展为一项较为成熟的技术，被广泛应用于道路卡口车牌识别、停车场管理系统等有约束场景[5]。现有车牌识别技术通常包括车牌定位、车牌校正、字符分割、字符识别等主要步骤，通常对车牌图像质量具有较高要求。因此，车牌识别系统中的摄像头通常安装于有约束场景，如高速路收费站、停车场出入口、十字路口等，并且需要停车杆、闪光灯、龙门架、传感器等辅助设备。但是在无约束的城市监控环境下，车辆的行驶行为、方向、速度无法受到约束，监控摄像头受到拍摄距离、光照、遮挡等不确定因素的影响。在这种情况下，车牌图像可能产生模糊、形变、遮挡等问题，如何提取具有区分力和鲁棒性的车牌特征并应用于车辆搜索，仍需进一步研究。

（4）面向监控网络的时空关系建模

监控网络中的时空情景信息，如对象出现的地点、时间、速度、方向、摄像头邻接关系、城市道路拓扑等，对车辆搜索具有重要的作用。时空信息建模已在许多多摄像头监控系统中得到广泛研究[20,21]。现有研究主要通过贝叶斯估计、概率图模型等方法，对人或车辆等物体在监控网络中出现的时空关系进行建模。时空关系作为一种先验知识，可以实现时空上由近及远的搜索方式，从而提高车辆搜索的效率。然而，面对城市中复杂的交通环境，车辆的速度、方向、行为等状态具有巨大的不确定性，如何有效对时空关系进行建模并用于车辆搜索，仍存在很大挑战。

（5）车辆搜索系统信息安全

车辆搜索系统通过视频监控网络感知城市中的人、车、物等物理对象，收集了大量多模态的感知数据（如人脸、车牌等）。由于车辆一般与人员关系密切，因此带来诸多安全与隐私保护问题。如果忽视这些问题，将对为用户、城市管理者、服务提供者带来巨大的经济和名誉损失。因此，信息安全与个人隐私保护是设计物联网

搜索系统及车辆搜索系统的重要问题[13]。搜索系统必须对不同用户设定不同的访问权限，使得他们能够获取不同权限粒度的搜索结果，从而保证车辆的隐私信息不被恶意使用。

本文主要围绕基于视觉信息的车辆外观特征表示、监控视频中车牌信息的有效提取、监控网络中时空关系建模等相关问题展开研究。

1.3 面向视频监控网络的车辆搜索相关工作

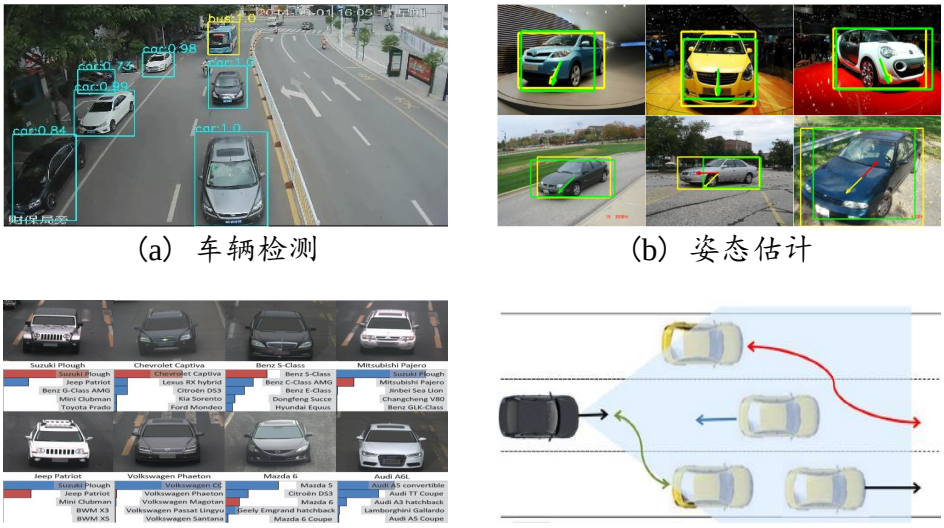
如前文所述，面向监控视频的车辆相关研究当前已成为多媒体、计算机视觉等领域的热点。因此，本文将介绍车辆特征表示、视频监控中的对象搜索两个方面的相关工作。

1.3.1 车辆特征表示

车辆在多媒体和计算机视觉领域一直是重要的研究对象，如图1-4所示。针对不同任务，如车辆检测、车辆分类、视角估计、车辆跟踪等，研究人员通过手工设计的描述子或基于 CNN 的特征学习方法对车辆的视觉特征表示进行建模[2–5,19,22–28]。

其中，监控图像中的车辆检测是指确定车辆在监控图像中的位置，通常用包围盒的坐标表示，车辆检测是车辆搜索的前提。Kembhavi 等人[23] 提出使用多种手工设计的图像特征如颜色概率图、方向梯度直方图提取车辆及其周围区域的特征，然后采用偏最小二乘法对特征进行降维，实现快速的车辆检测。李波[2] 提出一种基于核密度估计和边缘模型的运动车辆检测方法, 主要用于检测视频中的运动车辆。Zhang 等人[24] 提出采用三维变形模型表示车辆在图像中的位置、朝向、形状及视觉特征，通过拟合图像中的车辆与预定义的三维车辆模型，能够准确定位车辆的位置。

车辆作为一种类型丰富的对象，可以按照不同粒度进行分类，由粗粒度到细粒度包括按功能性分类（如小轿车、卡车、运动型多功能车、旅行车等）、按厂商分类（如一汽、北京、大众等）、按车款分类（如一汽红旗 H7、丰田凯美瑞、宝马三系等）、按年代分类（如 2007 年款、2012 年款、2017 年款宝马三系等）。例如，Krause 等人[26] 提出将图像中的二维车辆映射为三维表示，并提取多种局部特征用于车辆分类。Yang 等人[28] 提出一个大规模车辆细粒度分类数据集 CompCars，标注了车辆的品牌、车款、年代信息，并探究了 CNN 在车辆细粒度分类、车型验证、属性预测等方面的性能。Sochor 等人[29] 提出使用三维包围盒检测图像中的车辆，然后采用空间展开的方法表示车辆，最后采用 CNN 学习并提取车辆特征实现了优异的分类结果。



(a) 车辆检测　　(b) 姿态估计

(c) 细粒度分类　　(d) 行为分析

图 1-4　车辆特征表示相关研究

Fu 等人[3] 将视觉注意力机制引入 CNN 中，挖掘车辆图像中最具区分力的区域并提取鲁棒的特征，在多个数据集上取得了优异的分类结果。车辆搜索作为身份级的对象搜索问题，也是一种最细粒度的分类问题，可以借鉴细分类问题中提取车辆全局和局部特征的思想，挖掘能够区分不同身份车辆的的视觉特征。

车辆视角估计又称姿态估计，是指估计车辆在图像中的拍摄视角或朝向，可用于判断车辆的行驶方向。He 等人[30] 同时考虑车辆检测与姿态估计两种任务，提出一种结构化核方法对车辆外观进行参数建模，通过一种级联离散-连续算法实现快速的姿态估计。Yang 等人[4] 同样提出检测与视角估计联合学习，采用一种自动掩码神经网络自动从大规模训练图像中学习物体的各部分特征。Su 等人[19] 提出使用三维模型渲染生成大规模带有视角标注的物体图像，并训练一个类别相关的视角估计 CNN，实现对车辆等物体的视角估计。车辆的姿态估计有助于车辆在不同摄像头、不同视角下的特征提取，因此在车辆搜索中需要充分考虑车辆的视角信息。

车辆的牌照是车辆身份的唯一性标识，因此车牌识别是车辆识别的最有效手段，不仅是学术界的研究热点，而且已应用于各种实际场景中[5,25]。传统车牌识别技术通常包括车牌定位、车牌校正、字符分割、字符识别等主要步骤。首先，通过车牌定位可以将一整张监控图像帧中车辆的车牌区域定位。然后，通过车牌校正将车牌变换为角度端正的车牌。对于校正后的车牌，通常基于车牌制造标准和图像处理技术将整张车牌图像分割为仅包含单一字符的图像块。最后，提取每个图像块的视觉特征，使用分类器或神经网络对每个字符进行识别。因此，车牌识别系统中的摄像头

通常安装于有约束场景，如高速路收费站、停车场出入口、十字路口等，并且需要停车杆、闪光灯等辅助设备，以便获得字符清晰、角度端正、对比度高的车牌图像。

基于监控视频，可以实现视频中的车辆跟踪[2,6,31]，进而估计车辆行驶速度[7,10]或分析车辆的行为[8,27]。例如，李波[2] 提出使用多种特征表示车辆的视觉特征，然后采用均值漂移及粒子滤波算法实现多车辆跟踪。Matei 等人[6] 提出将运动特征和视觉特征结合，通过匹配车辆在不同摄像头中的连续图像，实现车辆的跨摄像头跟踪。Xiang 等人[31] 为解决车辆跟踪中角度变化很大导致的外观变化，提出了一种基于基于部件的粒子滤波框架，通过建模车辆三维部件与视角的关系，提取视角鲁棒的视觉特征，从而实现稳定的车辆跟踪。

1.3.2 视频监控网络中对象搜索

1.3.2.1 车辆搜索

车辆搜索目前仍处于初步探索阶段，相关方法主要包括两大类：第一类是基于图像的搜索方法，第二类是基于监控网络中多模态信息的方法。

其中，基于图像的车辆搜索方法仅考虑监控视频中车辆的视觉特征实现相同车辆的匹配与搜索[32–36]。例如，Feris 等人[32,33] 设计了一种基于属性的车辆搜索系统，首先通过预训练的分类器给车辆图像加上类型、颜色、形状等属性标签，然后通过属性标签在图像库中搜索包含某种属性的车辆。Liu 等人[34] 提出了一种基于耦合聚类损失的混合 CNN 模型，使得不同车辆在特征空间中更加容易区分。Zhang 等人[35] 采用三元组损失函数引导 CNN 训练，为提高网络的学习能力提出了一种改进的三元组采样方法生成训练样本，提高了车辆搜索的准确率。Yan 等人[36] 关注于车辆的多粒度属性如颜色、车型、身份，并提出了一种基于多粒度列表排序的多任务 CNN 模型，实现视觉相似车辆的搜索。

基于多模态信息的车辆搜索方法不仅考虑图像中的视觉信息，而且利用监控网络中的情景信息，如时空信息[37–39]。例如，Shen 等人[38] 提出使用链式马尔科夫随机场从大规模时空信息中给出候选车辆，然后采用视觉特征和时空信息匹配候选车辆中与查询车辆最相似的车辆。Wang 等人[39] 首先使用 CNN 提取车辆的全局和局部特征进行初步匹配，然后采用时空正则化模型优化搜索结果。

此外，研究人员还关注于大规模监控网络中的车辆轨迹分析[9,40]。例如，王龙飞[9] 提出使用城市中部署的卡口摄像头识别车牌信息，从而实现车辆在整个城市中的轨迹跟踪，进而对城市中车辆的出行规律进行分析和预测。Liu 等人[40] 提出了一

图 1-5　视频监控中的人员搜索研究

种路径排序算法，根据车牌信息获得已知的车辆轨迹，并根据已有轨迹和城市路网数据预测车辆在无摄像头区域的行驶轨迹。

1.3.2.2　人员搜索

监控视频中人员搜索一直是计算机视觉领域的研究热点，其目标是给定查询人员的描述（如图像、衣着、年龄等），在大规模视频监控网络中搜索与目标人员身份相同的人[41]。这一领域的研究主要集中在两个方面：第一个是特征表示[42–47]，第二个是度量学习[44,48–53]。

特征表示主要关注于设计和学习具有强区分力的图像特征，同时对多摄像机环境下的光照、视角、遮挡变化有较强的鲁棒性，通过图像特征匹配达到监控视频中人员搜索的目的，如图1-5所示。随着视觉计算研究的发展，采用的特征从低层特征，如纹理、颜色直方图等，发展到中层特征，如纹理分布、显著性特征、对称先验等，最近基于深度学习的高层特征取得了最好的性能。例如，Farenzena 等人[43] 提出使用颜色和局部纹理表示人员外观。Zhao 等人[45] 提出了一种显著性区域匹配方法，能够有效捕捉图像中人员的显著性区域，以克服人体形变带来的噪声。Liao 等人[44] 提出了一种对光照和视角鲁棒的局部最大响应特征，取得了手工设计特征中的最优性能。Li 等人[42] 将 CNN 引入人员搜索问题，通过在大量训练数据上自动学习具有区分力的特征，取得了比手工设计特征更好的准确性。Zheng 等人[47] 提出了一个大规模的人员搜索数据集，并在该数据集上综合比较了多种算法，成为人员搜索研究的测试基准。

度量学习又称子空间学习，其基本思想是通过从训练数据学习到的映射矩阵 W，将人员图像特征向量映射到度量空间中，在度量空间中满足：同一个人的样本之间距离较小，而不同人的样本之间的距离较大，使得不同人的样本更容易区分。度量学习从特征向量的度量角度入手，在度量空间中对查询图像特征与数据库中图像特

征进行匹配，从而搜索出最相似的目标图像[44,48–52]。例如，Prosser 等人[48] 提出了一种支持向量排序方法，将人员重识别转化为一种排序问题，相似的人排序靠前。Zheng 等人[49] 提出了一种概率相对距离比较模型，使得正样本对的距离小于负样本对。Zhao 等人[51] 提出了一种非监督显著性学习方法，在学习阶段无需标注训练样本的身份，通过邻接约束块匹配构建图像的对应关系，对视角、姿势变化具有很好的鲁棒性。Zhang 等人[52] 提出了一种零空间度量学习方法，要求零空间中相同身份的样本距离为零，负样本对间的距离为正值，取得了度量学习方法中最好的性能。

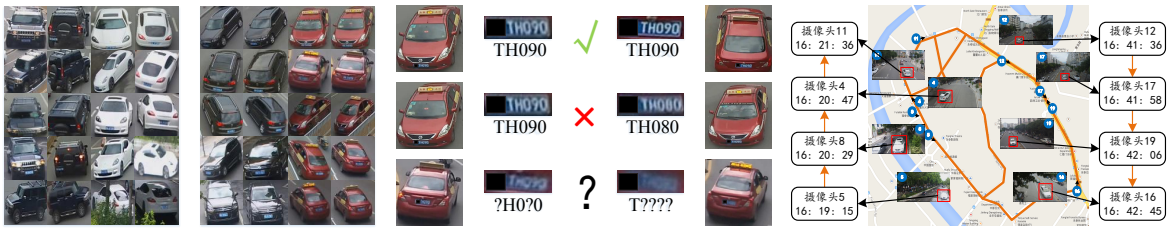
此外，人员搜索由最初基于单张抓拍图像的方法，逐渐发展到基于视频的方法，即输入查询与被查询的数据单元都为行人视频片段[54,55]。视频相比单张图像，不仅提供了更多空间视觉特征，而且提供了时间维度上的信息[56–60]。例如，Maclaughlin 等人[56] 和 Yan 等人[57] 采用了相似的思想，首先使用 CNN 对视频中每个图像帧提取空间特征，然后使用递归神经网络（RNN）或长短时记忆网络（LSTM）学习时间维度特征，取得了优异的搜索结果。Liu 等人[58] 忽略时间维度信息，将基于视频的人员匹配看作是图像集合匹配问题，通过引入注意力机制，CNN 能够挖掘多张图像中质量最好、最具区分力的图像，实现了准确的人员匹配。Zhou 等人[60] 在 CNN 与 LSTM 网络的基础上，在空间维度和时间维度都加入注意力机制，使得网络能够突出空间和时间维度的显著性区域和鲁棒性特征，取得了视频人员搜索的最优结果。

人员搜索通常通过人员图像中的外观特征进行特征匹配，人的外观不足以实现唯一性的人员搜索，仅能实现外观相似人员的搜索。我们借鉴了人员搜索中对于外观特征的提取的方法，结合车辆外观和交通监控场景下环境的特点，提取具有区分力和鲁棒性的外观特征，通过外观首先筛选出外观相似的车辆，再进行下一步的精细搜索，以提高车辆搜索的效率。

1.4 研究内容与主要贡献

对于搜索系统，尤其是面向大规模视频监控的车辆搜索，主要面临“搜不准”和“搜得慢”两大挑战，即如何在保证车辆匹配准确性的同时提高车辆搜索的效率。

“搜不准”，一方面是由于车辆自身外观的相似性和多样性，另一方面是由于无约束城市监控中极端多变的环境因素。如图1-6（a）所示，车辆的类型、颜色十分丰富，因此要求视觉特征能够具有较强的表达能力。此外，相同品牌、款式的车辆通常外观非常相似，所以外观特征必须具有细节区分能力。如图1-6（a）所示，无约束城市交通场景下，由于环境光照的多变性、拍摄视角的任意性、拍摄背景的复杂性、



(a) (b) (c)

图 1-6 车辆搜索面临的挑战

前景物体的遮挡等因素，使得不同摄像头或相同摄像头在不同时间拍摄的同一车辆外形差异（类内差异）巨大，但相似角度下相似车辆的外观差异（类间差异）却可能很小。因此，如何设计或学习具有外观表达力、细节区分力、环境鲁棒性的视觉特征，是车辆搜索的第一项挑战。

由于仅通过外观很难准确判断同一车辆，因此车牌作为车辆的唯一性标识必须参与车牌搜索过程。但同样由于无约束监控场景的不确定性，使得车牌识别方法无法正确识别低质图像中的车牌字符。即便能够识别车牌字符，但只要有一个或几个字符识别错误，就会使车辆无法正确匹配，如图1-6（b）所示。因此，如何有效利用监控视频中的车牌信息，是车辆搜索的第二项挑战。

“搜得慢”，不仅是因为城市监控视频中巨大的数据规模，而且因为车辆多模态特征的丰富性和复杂性。城市中巨大的摄像头数量和车辆数量，使得数据库中存储了巨大规模的车辆图像。由于城市摄像头的部署与道路拓扑具有相关性，因此可以提前获得摄像头的邻接关系、空间距离等信息。此外，摄像头拍摄车辆图像同时存储了拍摄的时间信息。基于上述时间、空间等情景信息，实现了一种由近及远的渐进式搜索模式，从而提高搜索效率，如图1-6（c）所示。但是，由于城市交通环境下车辆速度、方向、行为的不确定性，如何建模时空信息、提高车辆搜索的效率是本文的第三项挑战。

针对上述挑战，本文面向大规模城市交通视频监控场景，提出了一种融合多模态数据的渐进式车辆搜索框架，并从车辆外观特征的学习与表示、车辆唯一标识即车牌的有效利用、监控网络中时空关系的挖掘三个方面提出了一系列方法与模型。此外，为推进车辆搜索和相关领域的发展，本文还收集并标注了一个来自于真实城市交通监控系统的大规模车辆搜索数据集，通过在该数据集上的大量实验验证了所提出的框架与方法。本文的主要贡献具体如下：

（1）融合多模态数据的渐进式车辆搜索框架。该框架综合特征域和时空域进行

逐步求精地搜索，具体来说：一是特征域内由粗到精地搜索，即先采用外观特征快速查找相似车辆，再使用车牌信息实现精确搜索；二是在时空域内，利用监控网络中的时空信息由近及远地搜索。实验分析表明，渐进式搜索框架不仅能够显著降低车辆搜索的时间消耗，同时保证了车辆搜索的准确性。

（2）基于车辆外观特征的相似车辆搜索方法。针对抓拍图像和视频两种查询数据，我们分别提出了两种基于深度卷积神经网络（Convolutional Neural Network，CNN）的车辆外观表示方法：NuFACT 和 CAN。NuFACT 方法能够从抓拍车辆图像中提取车辆的纹理、颜色、类别等多级特征，并通过零空间度量学习将上述特征融合为一种具有区分力的、鲁棒的特征。CAN 方法能够提取视频中多张图像的共有信息和互补信息，自动学习不同距离、不同角度图像中的有效特征，增强了车辆外观特征的区分力和鲁棒性。

（3）车牌图像超分辨率与验证结合的精确车辆搜索方法。针对无约束监控环境中低质的车牌图像，我们提出了一种基于域先验生成对抗网络的图像超分辨率方法进行车牌图像增强。针对监控数据中车辆数量很大而每个车辆样本较少的问题，本文采用一种基于对偶神经网络（Siamese Neural Network，SNN）的车牌验证方法，实现了车牌图像的快速准确匹配。通过车牌增强与验证结合，进一步提高了车辆搜索的准确性。

（4）基于邻接图与时空相似度模型的搜索结果重排序。通过挖掘城市监控网络中的时空信息，如车辆被拍摄的时间、摄像头的位置、摄像头邻接关系等，我们设计了一种摄像头邻接图模型表示视频监控网络的空间拓扑，提出了一种基于多层感知机的时空相似度模型（Spatio-Temporal Similarity Model，STSM），通过 STSM 估计车辆间的时空相似性对搜索结果进行重排序，得到优化的车辆搜索结果。

最后，我们构建了一个融合多模态数据的渐进式车辆搜索原型系统，并在真实视频监控数据上验证了上述框架与方法的有效性。

1.5 论文结构

本文面向城市视频监控网络，提出了一种渐进式车辆搜索框架，针对框架中车辆外观特征表示与学习、车牌图像增强与验证、时空信息挖掘等问题开展相关研究。全文共分为六个章节，论文结构和章节关系如图1-7所示，各章内容如下：

第一章，介绍本文研究背景与应用意义，总结相关车辆搜索的研究进展，分析本文研究的挑战，概述本文的研究内容与主要贡献。

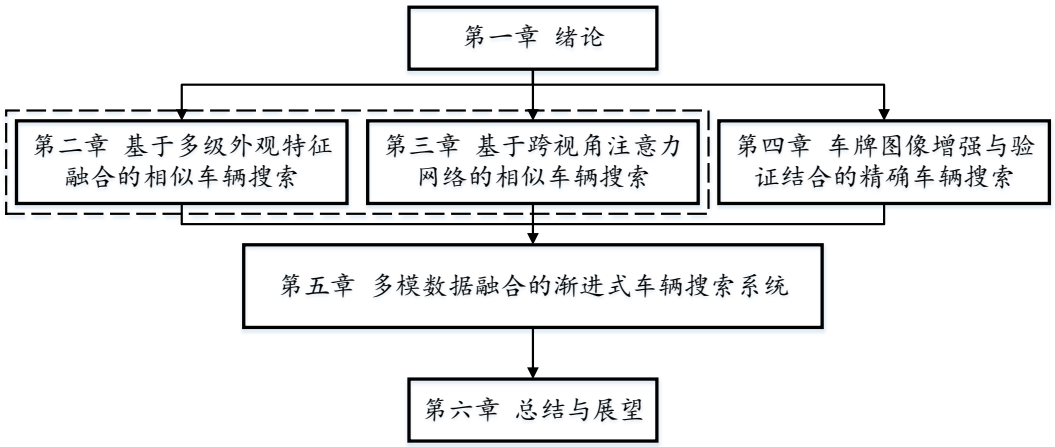


图 1-7 论文章节结构与关系图

第二章，基于多级外观特征融合的相似车辆搜索。介绍本文提出的基于零空间度量学习的车辆外观特征融合方法，并应用于外观相似车辆的搜索。

第三章，基于跨视角注意力网络的相似车辆搜索。介绍本文提出的基于跨视角注意力网络如何学习视频中车辆外观特征，并利用学习的视角不变特征搜索外观相似车辆。

第四章，车牌图像增强与验证结合的精确车辆搜索。阐述基于域先验生成对抗网络的车牌超分辨率方法和基于 SNN 的车牌验证方法，并将二者结合用于精确的车辆搜索。

第五章，多模数据融合的渐进式车辆搜索系统。介绍渐进式车辆搜索框架和原型系统的实现，并阐述如何将时空关系模型融入渐进式搜索框架。

第六章，总结本文研究工作，分析存在的不足，并给出未来研究工作的方向。

第二章 基于多级外观特征融合的相似车辆搜索

2.1 引言

本章关注于面向单幅抓拍图像的外观相似车辆搜索。假定车辆图像已由图像目标检测算法[61,62]从监控视频中抓拍得到，即每幅车辆图像仅包含一辆车，则本章中的车辆搜索任务可定义为：给定一幅查询车辆图像，在由大规模监控摄像机采集的车辆图像数据库中搜索与其身份相同的车辆，返回结果按照相似度由大到小排序。面向单幅抓拍图像的搜索是指在该任务中，匹配过程中使用的查询与被查询单元都是监控摄像头抓拍的单幅车辆图像。

与已有的车辆检测[33]、跟踪[6]、分类问题[28]不同，车辆搜索类似于近似图像检索（Near Duplicate Image Retrieval，NDIR）[63,64]、基于内容的视频搜索[65]、物理实体搜索[66]问题。对于给定的查询车辆图像，需要首先找到与其外观相似的车辆。然而，与传统近似图像检索问题不同，仅依据车辆外观很难唯一地确定身份相同的车辆，因为同一辆车在不同监控摄像头中可能存在很大的类内差异，而不同车辆在相似角度下的类间差异却可能很小，这种歧义性使车辆搜索问题具有很大的挑战。如图2-1所示，左图中第 1、2 列和第 3、4 列为相同车辆由不同摄像头拍摄的图像，右图第 1、2 列和第 3、4 列为不同车辆由相同摄像头拍摄的图像。虽然汽车牌照是车辆重识别的重要线索和依据，但在无约束的城市监控场景下，车牌可能由于光照、视角、遮挡等原因无法正确识别甚至无法拍摄到。因此，复杂城市监控场景中的车辆搜索是一项具有挑战性的问题。

为克服车辆搜索中的上述挑战，我们提出了一种基于多级特征融合的相似车辆搜索方法，如图2-2所示。该方法主要包括两个步骤：（1）基于多级特征的车辆外观特征的表示与量化，（2）基于度量学习的特征变换与匹配。在特征表示与量化中，我们使用了纹理、颜色和卷积神经网络提取的语义属性特征，将一张车辆图像表示为三种特征向量。然后，采用零空间度量学习算法将三种特征向量融合，并投影到零空间中。最后，通过计算特征在零空间中的欧式距离得到车辆的相似度。在真实城市视频监控数据上，我们综合评估了各种特征及度量学习算法的性能，并验证了所提出方法的有效性。

图 2-1　城市监控摄像头抓拍的车辆图像示例

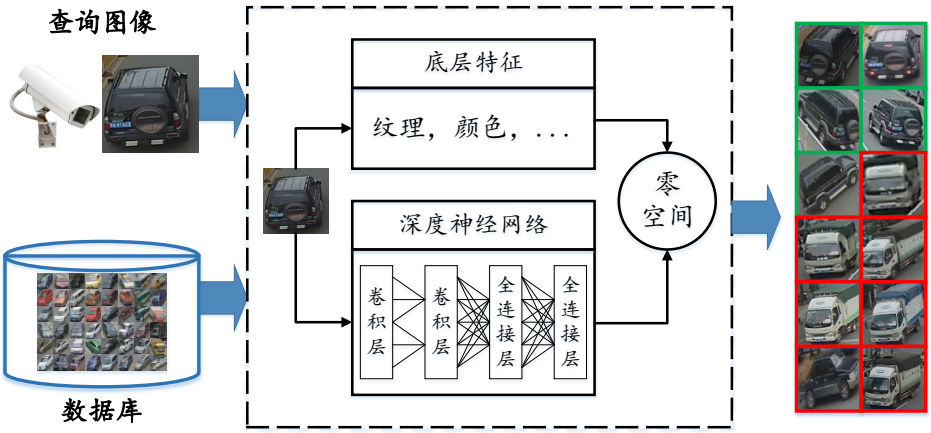


图 2-2　基于多级特征融合与零空间度量学习的车辆搜索框架

2.2　问题描述

根据前文中的定义，基于外观相似度的车辆搜索可以描述为：给定 $G=\{g_i\}_{i=1}^{N}$ 为从监控视频中采集的车辆图像数据库，包含 N 张车辆图像，它们分别属于 M 个不同身份（车牌号）的车辆 $V=\{v_j\}_{j=1}^{M}$。给定一张查询车辆图像 q，它的身份为：

$$v^{*}=argmin_{v_j\in V}dist(q,g_i) \tag{2-1}$$

其中 v^{*} 为查询图像中车辆的身份，$dist(\cdot,\cdot)$ 为某种距离度量函数。

对于本章中的车辆外观特征表示与量化，主要通过人工设计的描述子或通过卷积神经网络等手段，提取车辆图像中具有区分力的特征并量化为特征向量 $\mathbf{r}\in\mathbb{R}^{L}$，

表示如下：

$$\mathbf{r} = f(I) \tag{2-2}$$

其中 L 为特征向量长度，I 为一张查询图像或数据库图像，$f(\cdot)$ 为特征表示与量化函数。

对于基于度量学习的特征变换，首先通过有监督或无监督方法从训练图像数据学习得到投影矩阵 $W \in \mathbb{R}^{L\times d}$，然后通过投影矩阵将原始图像特征向量 $\mathbf{r}$ 变换到一个子空间，得到变换后的特征向量 $\mathbf{r}\cdot W$。此时，查询图像与数据库图像的相似度可由图像特征在子空间中的距离表示，距离越小则两张图像中的车辆外观越相似。那么，公式2-1中查询图像中车辆的身份可以表示为：

$$v^* = argmin_{v_j \in V} dist(f(q)\cdot W, f(g_i)\cdot W) \tag{2-3}$$

为了实现相似车辆准确快速的匹配，图像特征需要具有足够的表达能力以区分不同车辆的外观差异，并对城市监控环境下恶劣的环境因素具有较强的鲁棒性。度量学习方法需要能够将图像特征映射到一子空间内，使得相同车辆的样本间距离较小，不同车辆的样本间距离较大。

2.3 车辆多级外观特征表示

为实现监控视频中外观相似车辆的搜索，必须从车辆图像中提取具有有效区分力并对环境变化有较好鲁棒性的图像特征，因此我们提出一种融合纹理、颜色和语义属性的多级外观特征。为保证特征计算便利，首先对全部需要提取外观特征的车辆图像归一化为分辨率 64×128 的图像，然后再进行特征提取。

2.3.1 纹理特征

纹理是用于表示车辆局部细节的低层特征，我们采用尺度不变特征变换（Scale Invariant Feature Transform，SIFT）[67] 作为局部描述子。然后，使用 NDIR 任务中广泛采用的词袋模型（Bag-of-Visual-Word, BOW）[68] 将一幅图像中的全部 SIFT 描述子量化为同一的特征向量。

在训练阶段，首先提取所有训练图像的 SIFT 描述子，然后使用 k-means 聚类算法将所有 SIFT 描述子聚类得到 k_t 个聚类中心。这里得到的 k_t 个聚类中心的集合称为纹理码本，每个聚类中心称为一个关键字。

在搜索阶段，给定一张查询图像 q 或数据库图像 g_i，首先提取 SIFT 描述子。然后，为一张图像中提取的每一 SIFT 描述子查找码本中最接近的关键字，并对该关键字计数加一。由此，每一张图像可以得到一个长度为 k_t 的向量，每一个元素对应于码本中一个关键字的计数值。最后，对该向量进行 $L-2$ 归一化，得到图像的纹理特征向量 $\mathbf{r}_t$。

2.3.2 颜色特征

颜色是描述车辆外观的重要特征之一，但颜色受监控环境中光照变化和车辆表面镜面反射影响较大。因此，我们使用颜色名称（Color Name，CN）[69] 作为基本颜色描述子，CN 描述子是从自然场景拍摄的图像中学习得到的一种颜色表示模型，因此适合本章中的监控场景。然后，同样采用 BOW 模型对一张图像的颜色描述子量化为一致的颜色特征向量。

在训练阶段，首先将所有训练图像分割为 4×4 的图像块，计算每一图像块的 CN 描述子均值，得到所有训练数据的 CN 描述子集合。然后采用 k-means 聚类算法将 CN 描述子聚类得到包含 k_c 个关键字的颜色码本。在搜索阶段，我们采用了人员重识别中有效的 avgIDF、弱几何约束等方法[47] 提取颜色特征。首先，给定一张查询图像 q 或数据库图像 g_i，将其由上到下均分为 16 个图像带。然后，对每一个图像带，分割为 4×4 的图像块，每一图像块提取 CN 描述子，并采用训练好的颜色码本将描述子量化为长度为 k_c 的向量。最后，将 16 个图像带的特征向量按顺序连接，得到长度为 $16\times k_c$ 的颜色特征向量 $\mathbf{r}_c$。

2.3.3 语义属性特征

语义属性是指人们对车辆外观的一系列高级语义描述，如车型、品牌、车门数、车灯形状等。卷积神经网络作为一种有效的特征学习方法，能够从大量有语义属性标注的数据中学习到具有区分力的语义属性特征。因此，我们采用 GoogLeNet[70] 模型来学习语义属性特征。

在训练阶段，首先使用在 ImageNet 图像分类数据集[71] 上预训练得到的参数对 GoogLeNet 初始化。然后，将该 GoogLeNet 模型在车辆细粒度分类数据集 CompCars[28] 上进行微调训练。由于 CompCars 对车辆图像标注了细粒度的车辆品牌、型号、车门数、车灯形状等标注，因此通过微调训练，GoogLeNet 能够学习到这些语义属性特征。在搜索阶段，我们使用训练好的 GoogLeNet 模型提取查询图像 q

或数据库图像 g_i 的特征。将图像作为 GoogLeNet 网络的输入，通过前向传播计算得到第五个池化层（Pool5）的输出值，作为车辆图像的高层语义属性特征 $\mathbf{r}_s$。

2.4 基于零空间度量学习的多级特征融合

在提取了车辆图像的纹理、颜色、语义属性等多级外观特征后，需要对多级特征进行有效的融合，以实现准确的相似车辆搜索。最简单直接的特征融合方法是将上述三种特征向量直接连接，得到一个特征向量用于计算图像间的相似性。这种方法虽然能够保留全部的特征信息，但同样保留了特征中的冗余信息和干扰信息，无法挖掘特征中最有效的部分。度量学习方法，如线性判别分析（Linear Discriminant Analysis，LDA）和弗利-萨蒙变换（Foley-Sammon Transform，FST）[72]，能够得到特征中的有效成分并降低特征维度。其中，零弗利-萨蒙变换（Null Foley-Sammon Transform，NFST）最初被用于解决人脸识别中的小样本问题，能够从少量样本中学习得到有效的人脸特征[73]。Zhang 等人提出了一种面向人员重识别的核化零弗利-萨蒙变换（Kernalized Null Foley-Sammon Transform，KNFST）[52]，该方法将人员的外观特征由特征空间映射到一个零空间，取得了人员重识别的最优结果。受上述工作启发，我们提出了一种基于零空间度量学习的特征融合方法，将纹理、颜色、语义属性特征融合为有效的车辆外观表示。

2.4.1 零空间度量学习

零空间度量学习，即零弗利-萨蒙变换，是传统 LDA 算法和 FST 算法的改进。FST 算法的基本思想是从大量训练样本中学习一个投影矩阵 $W \in \mathbb{R}^{d\times m}$，并最大化其费雪判别准则如下：

$$\mathscr{J}(\mathbf{w}) = \frac{\mathbf{w}^{\top}\mathbf{S}_b\mathbf{w}}{\mathbf{w}^{\top}\mathbf{S}_w\mathbf{w}} \tag{2-4}$$

其中 $\mathbf{w}$ 为 W 中的一列，$\mathbf{S}_b$ 和 $\mathbf{S}_w$ 分别为类间散布矩阵和类内散布矩阵。通过与投影矩阵 W 相乘，原始视觉特征向量被映射到一个隐含度量空间。在该空间中，同一个对象的样本间距离要尽可能小于不同对象的样本间距离。

在此基础上，NFST 的目标是通过更严格约束学习一个零空间，约束表示如下：

$$\mathbf{w}^{\top}\mathbf{S}_w\mathbf{w} = 0 \tag{2-5}$$

$$\mathbf{w}^{\top}\mathbf{S}_b\mathbf{w} > 0 \tag{2-6}$$

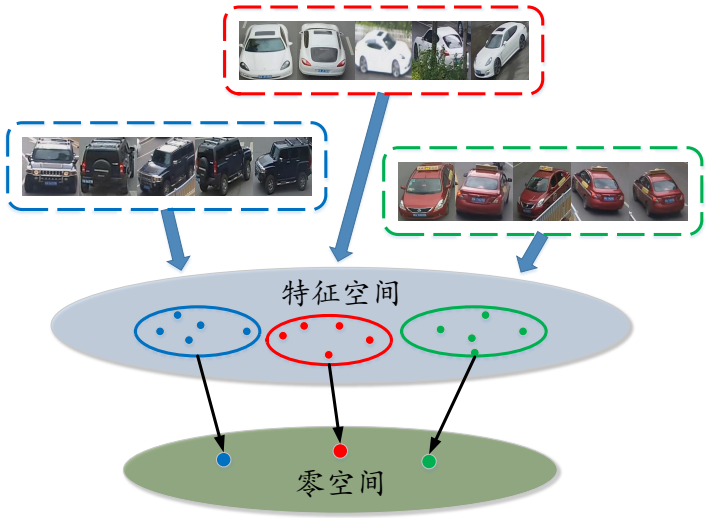


图 2-3　基于零空间度量学习的多级特征融合

在零空间中，同一对象的样本被压缩到一点，使得类内差异为零，而类间差异为正值，如图2-3所示。

更进一步，为了学习到一个更具有区分力的零空间，Zhang 等人[52] 为 NFST 引入了一个核函数 $\Phi(\cdot)$。通过核函数，原始视觉特征 $\mathbf{x}$ 首先被映射到一个高维空间。通过在大量训练数据上训练，多种视觉特征可以有效融合，为人员重识别生成一种具有强区分力的特征表示。

2.4.2　多级特征融合

车辆图像的多级外观特征，即纹理、颜色、语义属性，通过 NFST 方法融合并转换。在训练阶段，首先对所有训练图像提取前文所述的三种特征向量 $\mathbf{r}_t$、$\mathbf{r}_c$、$\mathbf{r}_s$，并将三种特征向量连接得到初始特征向量 $\mathbf{r}=(\mathbf{r}_t,\mathbf{r}_c,\mathbf{r}_s)$。然后，通过核函数 $\Phi(\cdot)$ 将特征向量 $\mathbf{r}$ 映射到高维空间，得到 $\Phi(\mathbf{r})$。最后，根据 Zhang 等人提出的 KNFST 训练算法[52]，以训练样本的车辆身份编号作为标签，训练得到投影矩阵 W。

在测试阶段，对于查询图像 q 和数据库图像 g_i，首先提取三种外观特征并得到初始外观特征 $\mathbf{r}_q$ 和 $\mathbf{r}_{g_i}$。然后，使用与训练时相同的核函数 $\Phi(\cdot)$ 将特征向量 $\mathbf{r}_q$ 和 $\mathbf{r}_{g_i}$ 映射到高维空间，得到 $\Phi(\mathbf{r}_q)$ 和 $\Phi(\mathbf{r}_{g_i})$。最后，根据训练得到的投影矩阵 W，q 和 g_i 的距离可以表示为：

$$dist(q,g_i)=||\Phi(\mathbf{r}_q)\times W-\Phi(\mathbf{r}_{g_i})\times W|| \tag{2-7}$$

其中 $||\cdot||$ 为向量的 $L-2$ 范数，即 $\Phi(\mathbf{r}_q)$ 与 $\Phi(\mathbf{r}_{g_i})$ 的欧式距离。

最终，根据上述距离度量函数，公式2-3可表示为：

$$v^* = argmin_{v_j \in V} ||\Phi(\mathbf{r}_q) \times W - \Phi(\mathbf{r}_{g_i}) \times W|| \tag{2-8}$$

通过我们提出的基于零空间度量学习与多级特征融合方法，可以根据查询图像快速准确地搜索得到与其外观最相似的车辆图像。

2.5 实验结果与分析

2.5.1 数据集

2.5.1.1 VeRi 数据集

为了评估提出的车辆搜索方法，我们收集并标注了一个综合性的车辆搜索数据集 VeRi。该数据集的原始视频来源于部署在约一平方公里城市区域内的 20 个监控摄像头，以保证收集到高质量的数据，同时能够反映真实世界交通场景的实际情况。这些摄像头拍摄了多种不同的交通场景，如十字路口、转弯路口、单向车道、双向车道、二车道道路、四车道道路等。摄像头原始视频分辨率为 1920×1080，帧率为 25 帧每秒。此外，摄像头安装角度丰富多样，包括正视角、侧视角、斜视角等（摄像机朝向角度与俯仰角度未做标注），部分摄像头之间存在重合区域。

通过上述 20 个摄像头，我们采集了连续 24 小时的原始视频数据，共计 1.43TB。然后，选取了其中下午 4：00 至 5：00 的视频，并压缩转码存储，共计 34.2GB。考虑到标注效率与质量，将原始视频以每 5 帧抽取 1 帧的方式采样为视频帧图像，共计得到 36 万个视频帧。在这些视频帧的基础上，VeRi 数据集的构建与标注包括如下步骤：

- **第一步：包围盒与轨迹标注。** 首先，给定一个原始视频帧，标注车辆在其中的位置，即包围盒（Boungding Box，BBox），并将包围盒内的车辆图像保存为独立的图像，作为车辆搜索的匹配单元。包围盒是围绕车辆的一个矩形，需记录包围盒左上角坐标、宽度、高度信息。我们仅标注了分辨率大于 64×64 且正在运动中的车辆。然后，在包围盒的基础上，将一段连续视频帧中属于同一车辆的包围盒组成一个轨迹（track）。进一步为轨迹标注摄像头编号、时间戳等信息，以供后续标注。该步骤完成后，共得到超过 60000 个车辆图像（包围盒）、约 10000 个车辆轨迹。

(a) (b)

图 2-4 VeRi 数据集车辆图像样本示例

- **第二步：颜色与车型标注。** 在进行车辆的跨摄像头关联之前，我们首先将所有的轨迹标注了 10 种颜色和 9 种类别。受“CompCars”数据集[28] 启发，所有轨迹按照“颜色-类别”层级标注。首先，每个轨迹被标注为黑、灰、白、红、绿、橙、黄、金、棕、蓝中的一种颜色。然后，每个颜色下的轨迹被标注为三厢轿车、两厢轿车、越野车、多用途车、面包车、皮卡车、大巴车、卡车和旅行车。为保证标注的质量，每个轨迹由三个人工标注员以多数投票方式进行标注。

- **第三步：跨摄像头车辆关联。** 车辆的跨摄像头关联是车辆重识别数据集最重要的标注信息，它记录了不同摄像头下的哪些车辆是同一辆车，从而在训练和测试时能够确保对相同车辆在不同摄像头下出现的正确结果。同时，该步骤也是数据集标注中最困难、最耗时的部分，因为标注人员需要对每一辆车查看所有摄像头数据，以确保相同车辆的正确标注。在标注时，标注人员采用了若干策略以提高标注效率。首先，基于上一步骤中的“颜色-类别”层级标注，给定某个摄像头中一个需要标注的轨迹，标注人员只需考虑其它摄像头中标注了相同颜色和类别的轨迹，而不用遍历其它摄像头下所有的轨迹。然后，标注人员对颜色和类别相同的轨迹进一步观察其外观细节和车牌号信息，从而确定相同车辆。此外，轨迹的时空信息，如时间戳、摄像头编号、行驶方向等，可用于辅助标注。例如，车辆在相邻摄像头出现的时间间隔应相对较小，车辆的行驶方向可以辅助判断其可能出现在哪个相邻摄像头。最终，整个数据集的标注工作共计消耗 9 名标注人员约一个月的工作量，去除无法确定的车辆，人工标注 776 个不同身份的车辆。

VeRi 数据集具有如下四个特点，使其成为一个高质量且具有挑战性的数据集：

- **第一，真实监控场景的大规模数据。** VeRi 数据集包含共计 50000 余张车辆图

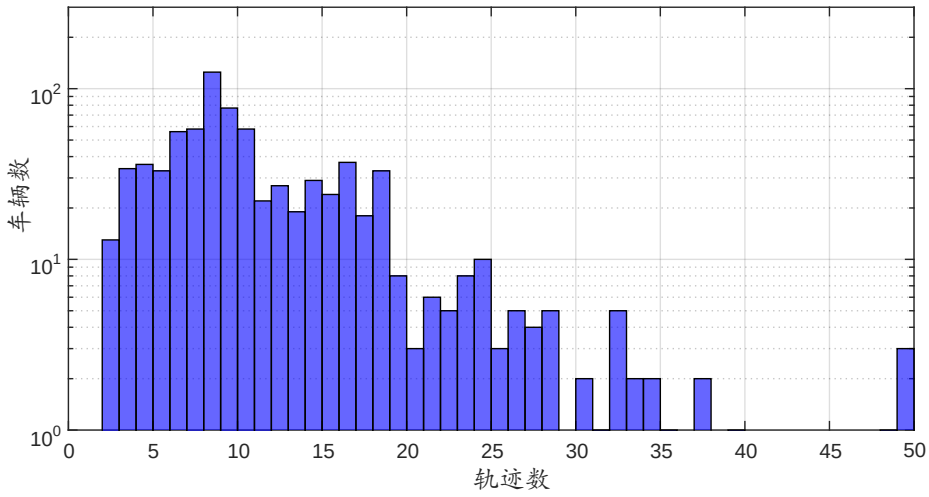


图 2-5 VeRi 数据集轨迹统计分布

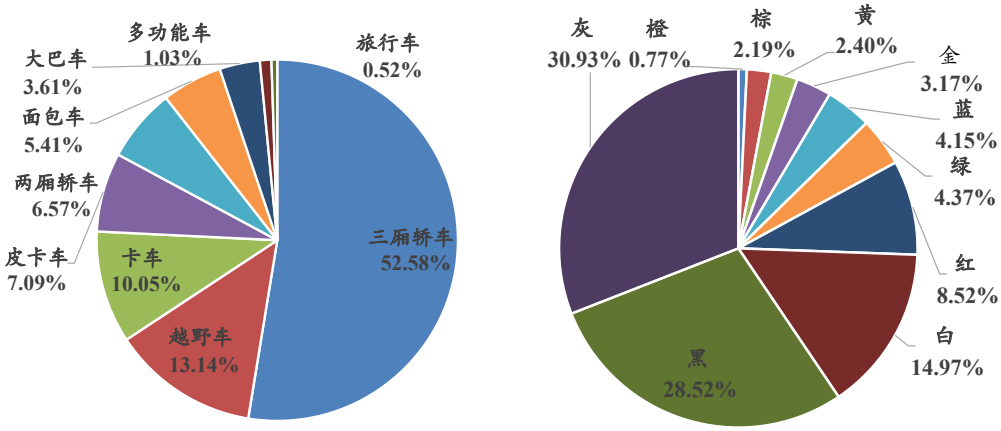


图 2-6 VeRi 数据集中车辆颜色与类别比例

像、9000 余个轨迹、776 个不同身份的车辆，从而具有足够的规模进行模型训练。车辆图像包含了丰富的拍摄视角、距离、不同的光照条件和背景，使其能够真实反映城市交通监控的情况，如图2-4所示。此外，每个车辆被至少 2 个不同摄像头拍摄，确保了车辆在不同摄像头中的高重现率，为车辆重识别研究提供了支持。车辆重现统计如图2-5所示。

- **第二，丰富的属性标注。** VeRi 数据集中的每个车辆图像都被标注了丰富的属性信息。首先，每个车辆图像记录了在原始视频帧中的包围盒坐标，因此 VeRi 数据集也可用于车辆检测等任务。此外，每个车辆不仅被标注了颜色、类别属性，部分车辆还标注了品牌信息，如宝马、奥迪、丰田、福特等 30 个品牌。图2-6展示了数据集中颜色和类别属性的分布。

图 2-7　VeRi 数据集中摄像机空间距离标注

- **第三，车辆牌照标注。** 在标注车辆跨摄像头重现的过程中，标注人员记录了每个车辆图像的车辆牌照信息（如果车牌可以识别），包括车牌区域在车辆图像中的位置、车牌字符串。最终，VeRi 数据集共计包含了 13471 张车牌图像，可供车牌检测、识别等问题研究使用。

- **第四，情景信息标注。** 通过标注车辆跨摄像头重现，标注人员发现监控视频的情景信息（contextual information）如时间戳、摄像头编号、摄像头邻接关系、摄像头距离、城市路网拓扑等，对车辆重识别具有重要的作用。因此，标注过程中每个轨迹的时间戳、摄像头编号等信息也被记录下来。此外，摄像头之间的真实距离也通过谷歌地图获取，如图2-7所示。

综上所示，VeRi 数据集是一个大规模、综合性的车辆搜索数据集，不仅可以用于车辆搜索，还可以用于车辆检测、车牌检测、车牌识别、时空信息挖掘、城市计算等相关领域的研究。

2.5.1.2　VehicleID 数据集

最近，Liu 等人[34] 也构建了一个面向车辆重识别的数据集，命名为“VehicleID”。该数据集包括大量于白天拍摄的车辆图像，数据来自于部署在一个小城市的交通监控系统。与 VeRi 数据集类似，VehicleID 中的每个车辆在不同摄像头中出现，并进行标注。VehicleID 包含了共计 26267 个不同身份车辆的 221763 张车辆图像。每个车辆被标注了车辆品牌和车型信息，如福特福克斯、丰田花冠、本田雅阁等。图2-8展示了 VehicleID 数据集中的车辆图像样本。

图 2-8 VehicleID 数据集中的车辆图像样本

VehicleID 数据集与我们提出的 VeRi 数据集存在两点主要不同：

- **第一，场景和视角丰富性。** 尽管 VehicleID 数据集包含更多的车辆图像和身份，但其车辆图像仅包含车辆正面和背面图像，且大部分在较近距离拍摄。而 VeRi 数据集中的车辆图像包含了不同的拍摄视角、距离，能够更好地反映真实应用场景中的情况，同时也使 VeRi 数据集更具挑战性。

- **第二，丰富的属性标注。** VehicleID 数据集仅标注了车辆外观信息，如车型，因此仅能用于基于外观的车辆搜索研究。而 VeRi 数据集不仅标注了外观属性，如颜色、类别，还标注了车牌区域、车牌字符、时空信息、摄像头编号等信息。因此，VeRi 数据集能够用于更丰富的研究。

2.5.2 实验设置

我们将比较多种基于外观的车辆搜索特征与方法，为验证所提出方法的有效性，实验将在 VeRi 和 VehicleID 两个数据集上进行。

在实验之前，对于 VeRi 数据集中的 776 个车辆，随机选取 200 个车辆的 11579 张图像作为测试集，即被查询数据库 G，其余 576 个车辆的 37778 张图像作为训练集。对于 200 个测试车辆，每个车辆在每个摄像头下随机选取 1 张图像作为查询图像集合 Q，共计包含 1678 张查询图像。在测试阶段，采用跨摄像头搜索的方式进行

测试，即给定一张查询图像，目标是找到该车辆在其它摄像头中的出现，而在相同摄像头中出现的该车辆图像作为中立样本，不参与准确率计算。此外，搜索采用“图像-轨迹”（“image-to-track”）的搜索方式，即被搜索的图像以轨迹为单位，这也符合真实应用中的设定和需求。一张查询图像与一个轨迹的视觉距离，用图像与轨迹中所有图像中距离最小的图像距离表示。因此，在该测试策略下，数据集共包含 1678 个查询图像和 2021 个被查询轨迹。受人员重识别研究[47] 的启发，车辆搜索的准确率采用排序第一准确率（HIT@1）、排序第五准确率（HIT@5）以及累计匹配特性曲线（CMC）评价。此外，由于 VeRi 数据集中的车辆被多个摄像头拍摄，因此给定一个查询图像有多个正确结果。为评估方法的准确率和召回率，本章也引入平均准确率均值（mean Average Precision, mAP）来综合评价相关方法。其中，对于一个查询图像，平均准确率（Average Precision，AP）表示为：

$$AP = \frac{\sum_{k=1}^{N_g} P(k) \times gt(k)}{N_{gt}} \tag{2-9}$$

其中 N_g 为数据库集合 G 的样本数，N_{gt} 为该查询图像对应的正确样本数，$P(k)$ 为排序第 k 准确率，$gt(k)$ 为一指示函数，当第 k 个结果为正确时返回 1，否则返回 0。对于全部查询图像，平均准确率均值 mAP 表示为：

$$mAP = \frac{\sum_{q=1}^{N_q} AP(q)}{N_q} \tag{2-10}$$

其中 N_q 为查询集的图像数。

对于 VehicleID 数据集，本章按照[34] 中所述，将数据集分为训练集和测试集。训练集包含 13134 个车辆、110178 张图像，测试集包含 13133 个车辆、111585 张图像。对于测试集，随机采样得到三个子集用于测试，分别包含 800、1600、2400 个车辆。与文献[34] 中相同，采用 HIT@1 和 HIT@5 作为准确率评价指标。

2.5.3 方法对比

本节在 VeRi 和 VehicleID 数据集上评估了 8 种基于外观特征的车辆搜索方法，各方法的细节如下：

- 方法一，纹理特征（**BOW-SIFT**）。该方法为我们提出的多级车辆外观特征中的纹理特征，实现方法如第2.3.1节所述。首先对查询图像和数据库图像提取 SIFT

描述子，然后使用预训练的纹理码本将图像描述子量化为特征向量。最后，每张图像由长度 10000 的纹理特征向量表示，以验证纹理特征的有效性。

- 方法二，局部最大响应表示（**LOMO**）。LOMO 描述子[44] 是人员重识别领域性能最佳的局部特征，该特征能够有效表示人员在真实监控场景下的外观特征，并对环境因素具有较好的鲁棒性。每张车辆图像由 LOMO 描述子表示为一个长度 26960 的特征向量。

- 方法三，颜色特征（**BOW-CN**）。该方法为我们提出的多级车辆外观特征中的颜色特征，实现方法如第2.3.2节所述。首先对查询图像和数据库图像提取 CN 描述子，然后使用预训练的颜色码本将图像描述子量化为特征向量。此外，图像采用了 avgIDF 和几何先验提高区分力。最后，每张图像由长度 5600 的颜色特征向量表示，以验证颜色特征对车辆搜索的有效性。

- 方法四，基于 GoogLeNet 网络的语义属性特征（**GoogLeNet**）。该方法为我们提出的多级车辆外观特征中的高级语义属性特征，实现方法如第2.3.3节所述。首先对查询图像和数据库图像使用训练好的 GoogLeNet 网络进行前向传播运算，然后提取“Pool5”层的输出向量。最后，每张图像由长度 1024 的语义属性特征向量表示，以验证语义属性特征对车辆搜索的有效性。

- 方法五，融合颜色与属性特征（**FACT**）。该方法采用后融合方法对纹理、颜色、语义属性特征的搜索结果进行融合。首先，对查询图像和数据库图像提取上述三种特征向量。然后，分别计算查询图像与数据库图像三种特征向量的余弦距离。最后，将三个余弦距离按照不同的权重求和得到最终的距离度量，本章中三种特征距离的权重分别为 0.1、0.2、0.7。

- 方法六，深度相对距离学习（**DRDL**）。DRDL 是由 Liu 等人[34] 提出的基于 CNN 与多任务学习的车辆重识别方法。该方法使用深度卷积神经网络联合学习一种有区分力的特征表达和一个度量映射，在 VehicleID 数据集上取得了最好的性能。它采用一种基于 VGG_M 网络的混合网络结构[74]，并设计了一种耦合聚类损失（Coupled Cluster Loss，CCL）学习不同车辆间的相对距离。由于 VeRi 数据集与 VehicleID 数据集的类别标签不一致，因此该方法仅在 VehicleID 数据集上测试。

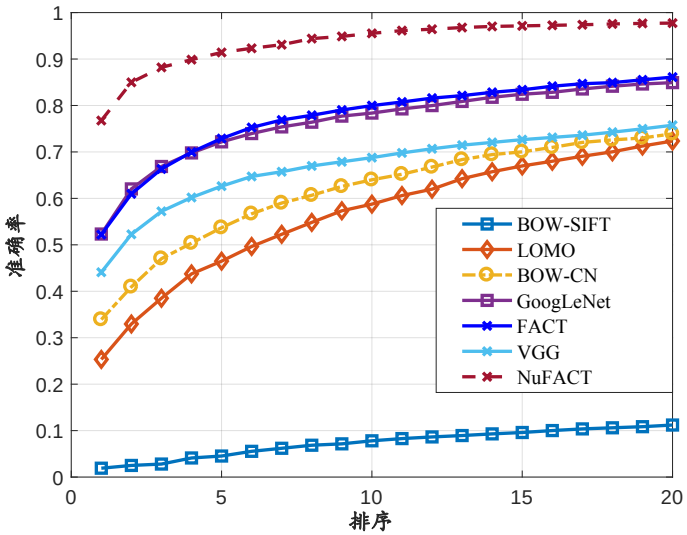


图 2-9　不同方法在 VeRi 数据集上的 CMC 曲线对比

表 2-1　不同方法在 VeRi 数据集上的结果比较

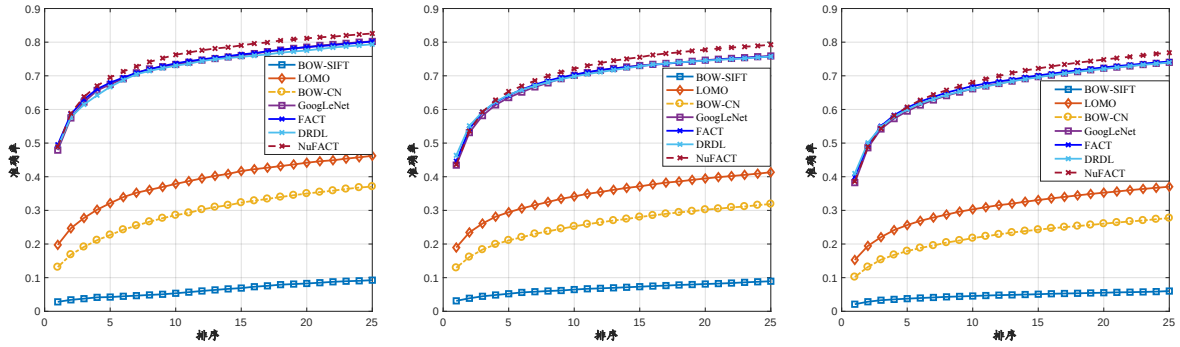
方法	mAP	HIT@1	HIT@5
BOW-SIFT	1.51	1.91	4.53
LOMO[44]	9.64	25.33	46.48
BOW-CN[47]	12.20	33.91	53.69
VGG[34]	12.76	44.10	62.63
GoogLeNet[28]	17.89	52.32	72.17
FACT	18.75	52.21	72.88
NuFACT	**48.47**	**76.76**	**91.42**

- 方法七，基于 VGG 网络的语义属性特征（**VGG**）。为评估不同卷积神经网络的效果，本节使用[34] 中采用的 VGG_M 网络结构作为特征提取器在 VeRi 数据集上进行测试。本方法使用 VGG_M 网络对查询图像和数据库图像进行前向传播运算，取“FC-7”层的输出向量作为图像的特征表示。

- 方法八，基于零空间度量学习的多级特征融合（**NuFACT**）。该方法为我们提出的车辆搜索方法，具体实现细节如第2.4节所述。

表2-1列出了上述方法在 VeRi 数据集上的 mAP、HIT@1、HIT@5 结果，图2-9展示了 CMC 曲线的对比。表2-2列出了测试方法在 VehicleID 数据集上的 HIT@1 和 HIT@5 结果，图2-10展示了 CMC 曲线的对比。从实验结果可以发现：

- 对于 VeRi 和 VehicleID 数据集，相对于基于深度卷积神经网络的方法，如



(a) 测试集大小 = 800　　(b) 测试集大小 = 1600　　(c) 测试集大小 = 2400

图 2-10　不同方法在 VehicleID 数据集上的 CMC 曲线对比

表 2-2　不同方法在 VehicleID 数据集上的结果比较

方法	测试集大小 = 800		测试集大小 = 1600		测试集大小 = 2400		平均值	
	HIT@1	HIT@5	HIT@1	HIT@5	HIT@1	HIT@5	HIT@1	HIT@5
BOW-SIFT	2.81	4.23	3.11	5.22	2.11	3.76	2.68	3.76
LOMO[44]	19.74	32.14	18.95	29.46	15.26	25.63	17.98	3.76
BOW-CN[47]	13.14	22.69	12.94	21.09	10.20	17.89	12.09	20.56
GoogLeNet[28]	47.90	67.43	43.45	63.53	38.24	59.51	43.20	60.04
FACT	**49.53**	67.96	44.63	64.19	39.91	60.49	44.69	64.21
DRDL[34]	48.91	66.71	**46.36**	64.38	**40.97**	60.02	**45.41**	63.70
NuFACT	48.90	**69.51**	43.64	**65.34**	38.63	**60.72**	43.72	**65.19**

GoogLeNet 和 VGG，手工设计的特征，即 BOW-SIFT、BOW-CN、LOMO，取得相对较低的准确率。这一现象说明深度神经网络通过在大量数据上的训练，能够学习到更具区分力和鲁棒性的特征表达。此外，直接融合多级特征的 FACT 和采用混合神经网络的 DRDL 都取得了比单一模型或特征更好的准确率。这证明了多级特征融合和多网络融合能够提供不同特征的互补特性，更有效地对车辆外观特征进行表示。最后，我们提出的 NuFACT 框架取得了两个数据集上的最佳结果。一方面表明多级特征融合能够更有效地表示车辆外观特征，另一方面表明 NFST 算法能够更有效地去除多级特征中的冗余部分，使多种特征更加互补地融合，从而达到相似车辆的准确搜索。

- 通过对比各种方法在 VehicleID 和 VeRi 数据集上的不同表现，我们发现了不同方法的若干特点。第一点，在 VehicleID 数据集上，纹理特征 LOMO 表现优于颜色特征 BOW-CN，而在 VeRi 数据集上则表现相反。通过观察两个数据集可

图 2-11　NuFACT 在 VeRi 数据集上的搜索结果示例

以发现，VehicleID 中的图像相对 VeRi 中的图像有较高的分辨率和清晰度，是的 VehicleID 中的图像具有更清晰的细节信息。此外，部分 VehicleID 图像的光照较暗，色彩饱和度较低，导致无法提取有效的颜色特征，如图2-4与2-8对比所示。因此颜色特征 BOW-CN 在 VeRi 数据集上性能优于 VehicleID 数据集上的性能。第二点，NuFACT 方法对特征融合提升的性能在 VeRi 数据集上要优于在 VehicleID 数据集。一方面原因是颜色特征的作用在 VeRi 数据集上更大，因此融合后的特征在 VeRi 数据集上效果更好。另一方面是由于 VeRi 数据集中的每个训练车辆包含平均约 64 张图像，而 VehicleID 数据集中的每个训练车辆包含平均约 8.4 张图像。因此，NFST 在训练过程能够从 VeRi 数据集中学习到更多的视觉特征，使其具有更好的鲁棒性和区分力。所以，在 VeRi 数据集上 NuFACT 对于 FACT 具有更大的提升。

图2-11展示了 NuFACT 方法在 VeRi 数据集上的搜索结果示例，每一行的蓝色框为输入查询图像，列出了排序前 16 的搜索结果。绿色框为正确搜索结果，红色框为错误搜索结果。通过（a）和（b）可以看出，NuFACT 能够准确地搜索到与查询图像相似的车辆图像。然而（c）展示了车辆外观相似性给车辆搜索带来的挑战，（d）充分表明真实监控场景中光照的变化及车辆表面的镜面反射使不同颜色的车看起来十分相似，因此搜索结果出现错误。

通过数据集的构建及实验的分析，我们得出如下结论：

- 对于准确的车辆搜索，除了需要考虑图像中的车辆外观，车辆牌照信息必须加以利用。特别地，视觉相似性仅能够提供一种粗粒度但快速的相似车辆筛选，如果要唯一地确定一辆车的身份，必须考虑车辆的唯一性标识。因此，在后续工作中，我们将进一步讨论如何利用车辆牌照信息提升车辆搜索的准确性。

- 除了车辆外观和车牌信息，视频监控网络中包含的情景信息，如时间戳、摄像头位置、相邻摄像头的空间距离等，对于搜索目标车辆是很重要的线索。因此，

如何挖掘、建模、融合上述情景信息，如何将其用于车辆搜索，也是后续工作中的重要问题。

2.6 本章小结

本章关注于面向城市监控的车辆搜索中的最基本情况，即基于抓拍车辆图像的相似车辆搜索。为提取具有区分力和鲁棒性的车辆外观特征，提出了一种多级特征表示方法。该方法采用人工设计的低层特征，包括纹理和颜色，以及由深度卷积神经网络学习得到的高层语义属性。为有效地挖掘上述多级特征的互补信息，提出采用零空间度量学习的方法融合多种特征，使得在零空间内相同车辆外观特征距离更近，不同车辆的样本距离更远，从而实现准确的相似车辆搜索。最后，通过在两个大规模公开数据集上的实验，验证了所提出方法的优异性能。

第三章 基于跨视角注意力神经网络的相似车辆搜索

3.1 引言

本章关注于基于视频的外观相似车辆搜索。假定车辆视频已由人工或自动目标检测和跟踪技术[27,62]从监控视频中截取得到，即每个车辆视频由一个连续的图像序列组成，其中每帧图像仅包含一辆车，则本章中基于视频的车辆搜索任务可定义为：给定一段截取的车辆视频，在由大规模监控摄像机采集的车辆视频数据库中搜索与其身份相同的车辆，返回结果按照相似度由大到小排序。面向车辆视频的搜索是指在该任务中，匹配过程中使用的查询输入与被查询数据都是监控视频中采集的连续图像序列。

基于抓拍图像的车辆匹配中，输入查询与数据库数据都是单幅车辆图像，目前很多工作关注于这种最基本的情况[37–39,75,76]. 从人工设计的图像描述子到基于卷积神经网络的图像特征，单幅图像车辆搜索的准确率不断提升。但是，监控视频中截取的单张图像仅能提供有限的视觉信息，特别是在无约束的城市交通监控场景中，只通过一张图像很难提取具有区分力的车辆外观特征。相比较而言，监控视频中截取的车辆视频片段包含了更多幅车辆图像，能够为车辆外观建模提供更多有效互补的信息，因此为复杂城市交通环境中准确地匹配相同车辆、区分不同车辆带来了保证。如图3-1所示，视频能够提供车辆的多视角图像，包含了互补的视觉信息。通过基于注意力权重的特征聚合，CAN 能够增强特征中具有区分力和显著性的部分，弱化冗余和歧义部分，从而使不同车辆的样本在特征空间更易区分。此外，基于视频的车辆搜索与真实应用更加接近，因为工作人员在搜索目标车辆时，仅需要截取出一段包含查询车辆的视频片段，而不用费力地在一段视频中挑选并截取最满意的一张或几张图像。因此，本章关注于面向城市监控的基于视频的车辆搜索。

但是，真实交通监控场景中基于视频的车辆搜索扔面临着两大挑战：

（1）监控环境的复杂性

由于监控环境的变化，例如光照变化、恶劣天气、复杂背景等，以及摄像头设置的不同，例如安装高度、分辨率、朝向灯，使得车辆外观在图像中的变化十分剧烈。特别是摄像机拍摄车辆时的视角变化，是其中最重要的因素之一，如图3-1所示。一

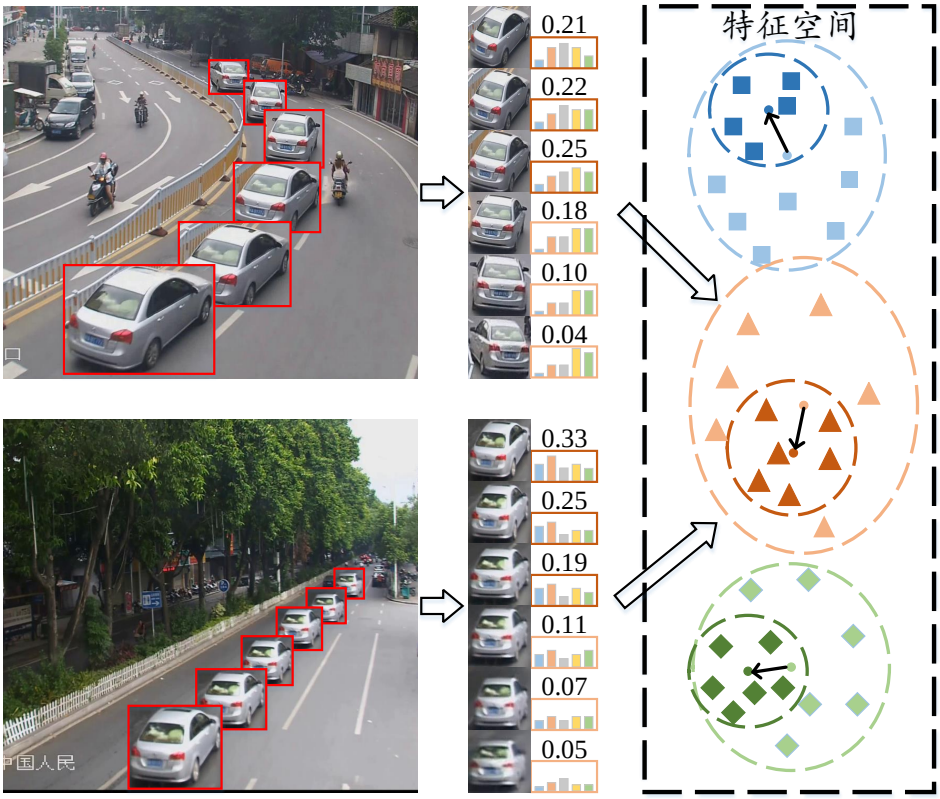


图 3-1　跨视角注意力网络（CAN）的动机

方面，不同的摄像机在不同时刻拍摄到同一个车辆的角度不同，使得同一辆车的类内差异很大。例如，我们很难仅通过同一辆车的正面和侧面图像确定其为同一辆车。另一方面，在一个摄像头拍摄到一辆车的视频片段中，该车辆可能在不同视频图像帧中呈现不同的角度。这使从不同视角图像提取车辆外观的有效表示具有很大挑战。

（2）车辆外观的多变性

此外，由于车辆是一种刚性物体，因此在监控视频中具有有限的动态特征，例如人员在视频中的动作、步态等显著特征。因此，对于车辆，很难采用已有的图像序列建模方法如递归神经网络（Recurrent Neural Network，RNN）或长短时记忆网络（Long-Short Term Memory，LSTM）提取车辆在视频中的动态模式特征。不过，相对于单张车辆图像，一段车辆视频能够提供车辆在多个视角、不同距离的图像，这些图像能够为全面的表示车辆外观提供更丰富互补的视觉信息。那么，如何有效地从视频中学习并提取视角不变的、有区分力的外观表示，是基于视频的车辆搜索的重要挑战。

为克服上述挑战，本章提出了一种面向视频车辆搜索的跨视角注意力网络（Cross-view Attentive Network，CAN）。CAN 以车辆视频片段为输入，能够学习一个

具有区分力和视角不变性的车辆外观表示。CAN 包括两个基于 CNN 的模块：

- 第一，特征学习网络（Feature Learning Network，FLN）。FLN 包含两个子网络，身份识别子网络和视角估计子网络，分别从车辆视频片段的每张图像帧学习两种中间特征表示，即身份特征和视角相关特征。其中，身份识别子网络能够识别单张图像中车辆的身份，是一个由大量不同监控环境下标注了车辆身份信息的单张车辆图像训练得到的 CNN。因此，可以通过它提取图像中具有外观区分力的特征表示。与身份识别子网络不同的是，由于车辆在不同视角下的外观变化，视角估计子网络能够学习并提取图像中与特定视角相关的车辆外观特征。通过 FLN，可以对车辆视频中的每张图像提取以上两种中间特征表示，即身份特征和视角相关特征。

- 第二，注意力聚合网络（Attentive Aggregation Network，AAN）。基于上述 FLN 提取的两种中间特征，AAN 采用了一种带有注意力机制的 CNN 模型，能够自动挖掘中间特征中最有效的部分，并弱化其中的冗余特征和噪声。对于身份特征，AAN 将学习一组权重，对于特征中对光照变化、图像分辨率等鲁棒的部分，将得到较高的权重，反之则被赋予较低的权重。对于视角相关特征，AAN 关注于多幅图像中视角最显著的部分，如车辆的前后视角图像。因此，AAN 将从视角相关特征学习一组视角显著性权重，对视角更好的图像特征给予更高的权重，视觉特征区分力较低的视角将得到较低的权重，如图3-1所示。这样，多张视频图像中互补的特征将按照不同权重相加，聚合为一个特征向量，作为该视频片段中车辆的外观表示。通过 FLN 和 AAN，CAN 不仅可以保留多个视角、分辨率下车辆的外观信息，而且能够自动挖掘最具区分力和鲁棒性的特征，为基于视频的相似车辆搜索提供了保证。

此外，目前已有的车辆搜索数据集，如上一章提出的 VeRi 和 Liu 等人提出的 VehicleID[34]，都是基于单张车辆图像的，无法有效评估基于视频的车辆搜索方法。因此，本章从真实城市监控视频中收集了一个大规模视频车辆搜索数据集——“VIVID”。该数据集不仅包含 776 个不同车辆的 8578 段视频片段，而且标注了车辆在不同摄像头中的出现，使其能够满足车辆搜索任务的需要。再者，VIVID 数据集中的视频不仅涵盖了丰富的交通场景，如直行路段、转弯路段、十字路口等，而且包含了丰富的车辆运动速度、方向、行为，如由近及远（及相反方向）直行、由左到右（及相反方向）直行、转弯、调头等。上述特点使 VIVID 数据集既能够反映真实

监控场景的情况，又具有很大的挑战性。最后，在 VIVID 数据集上，我们综合评估了不同方法的性能，验证了所提出方法的有效性。

3.2 问题描述

根据前文中的任务定义，基于视频的相似车辆搜索可以描述为：给定 $G=\{g_i\}_{i=1}^{N}$ 为从监控视频中采集的车辆视频数据库，包含 N 段车辆视频。其中，每段车辆视频 $g_i=\{I_k\}_{k=1}^{K}$，包含 K 张车辆图像，每张图像 I_k 中仅包含一辆车。数据库 G 中的视频片段分别属于 M 个不同身份（车牌号）的车辆 $V=\{v_j\}_{j=1}^{M}$。给定一段查询车辆视频 $q=\{I_k\}_{k=1}^{K}$，它的身份为：

$$v^{*}=argmin_{v_j\in V}dist(q,g_i) \tag{3-1}$$

其中 v^{*} 为查询图像中车辆的身份，$dist(\cdot,\cdot)$ 为某种距离度量函数[①]。

与基于单张图像的车辆搜索不同，需要对一段车辆视频 $S=I_1,I_2,\cdots,I_K$（S 代表 q 或 g_i）中的每张图像提取视觉特征，则视频 S 可得到特征矩阵 $R\in\mathbb{R}^{L\times K}$ 表示如下：

$$\begin{aligned} R&=f(S)\\ &=[f(I_1)^T,f(I_2)^T,\cdots,f(I_K)^T] \end{aligned} \tag{3-2}$$

其中 L 为特征向量长度，I 为一张查询图像或数据库图像，$f(\cdot)$ 为面向单张图像的特征表示函数。

为实现视频片段中车辆外观特征的距离度量，可以采用特征相似度融合的方式，将视频中的每一张图像作为独立的单元。在计算查询片段 q 与数据库片段 g_i 相似性距离时，计算所有可能的图像对之间的距离，然后将特征之间的距离进行融合得到视频片段间的距离。然而，这种方式没有考虑图像之间的相互关系，无法更有效利用互补的视觉特征。此外，当视频片段中的图像数量很大时，相似度融合方法在计算图像对的距离时需要较大的计算量。因此，本章采用特征特征融合的方法，将视频中各图像的特征通过聚合函数 $\mathscr{F}(\cdot)$ 融合为一个统一的特征表示 $\mathbf{r}\in\mathbb{R}^{L}$，再计算视频间的视觉相似性，表示如下：

$$\begin{aligned} \mathbf{r}&=\mathscr{F}(R)\\ &=\mathscr{F}([f(I_1)^T,f(I_2)^T,\cdots,f(I_K)]) \end{aligned} \tag{3-3}$$

① 为方便表示和计算，本章中假设所有查询视频与数据库视频都采样为相同长度 K。

那么，公式3-1中查询图像中车辆的身份可以表示为：

$$v^* = argmin_{v_j \in V} dist(\mathscr{F}(f(q)), \mathscr{F}(f(g_i))) \tag{3-4}$$

为了实现视频中相似车辆的准确匹配，一方面需要对单张车辆图像设计或学习具有区分力和鲁棒性的视觉特征，即得到有效的特征学习函数 $f(\cdot)$。另一方面，需要保留视频图像中有效的视觉特征，特别是车辆在不同视角下的互补视觉信息，并将这些信息融合为一个统一的特征表达，即需要设计有效的特征聚合函数 $\mathscr{F}(\cdot)$。因此，本章的主要贡献包括：提出一种跨视角注意力网络从车辆视频片段中学习并提取具有区分力和视角不变性的外观特征；其中包含了一种特征学习网络用于学习单张图像的视觉特征，及一种注意力聚合网络从连续视频图像中学习两组注意力权重，通过加权求和得到视频的统一视觉表示；最后，为评估提出的方法并推进相关领域研究，收集并标注了一个基于视频的车辆搜索数据集，在该数据集上验证了所提出方法的有效性。

3.3 面向视频车辆搜索的跨视角注意力网络框架

本章将详细介绍提出的跨视角注意力网络。CAN 以车辆视频片段为输入，对该片段中的车辆生成一个具有区分力和视角不变性的特征表示，如图3-2所示。如前文所述，本章假设车辆视频片段已由人工或现有的对象检测和跟踪[27,62] 算法截取得到，视频中的每张图像仅包含一个车辆。此外，为方便特征标注与计算，车辆视频片段被均匀采样为等长的图像序列。因此，CAN 的输入为车辆图像序列 $S = \{I_1, I_2, \cdots, I_K\}$，其中 K 为序列中的图像数。车辆图像序列 S 首先被输入到 FLN，分别经过两个子网络，提取身份特征和视角相关特征。然后，两种中间特征分别经过 AAN，通过两个子网络学习两组权重向量。最后，视频中多张图像的视觉特征通过 AAN 提取的权重向量进行加权融合，生成一个固定长度的特征表示 $\mathscr{F}(S) = \mho \in \mathbb{R}^L$。通过计算查询视频和数据库视频特征向量的距离，搜索数据库中与查询视频中车辆外观最相似的车辆视频。下面将详细介绍 CAN 的两个主要组成部分。

3.3.1 特征学习网络

特征学习网络的作用，是对输入的每张车辆图像提取具有区分力和鲁棒性的视觉特征。一方面，特征需要能够有效区分不同车辆的外观特征，并对监控场景下多变的光照、复杂的背景、不同的距离等因素具有较好的鲁棒性。另一方面，特征必

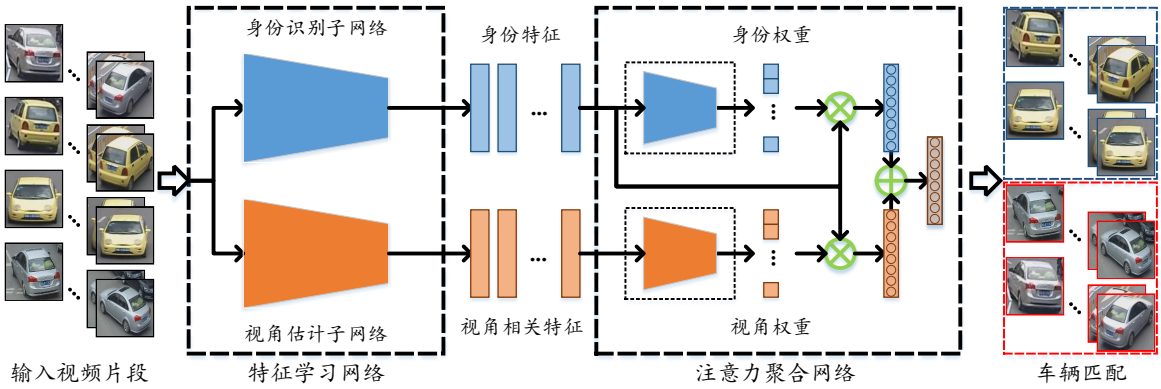


图 3-2　跨视角注意力网络框架

须能提取车辆在图像中的视角相关特征，即能够表示车辆在不同视角下的视觉特征。因此，FLN 包含两个独立的子网络：身份识别子网络和视角估计子网络。

如图3-2所示，身份识别子网络是一个深度全卷积网络（Fully Convolutional Network，FCN），网络主干中仅包含卷积层和池化层。在具体实现中，本章采用了在图像分类和对象检测中广泛采用的 50 层深度残差网络（50-layerd Residual Network，ResNet-50）[77]。ResNet-50 由一系列按层排列的残差模块（Residual Block）组成，通过跨层残差学习，可以预防极深的网络训练时出现的梯度弥散现象，从而提高网络的学习和泛化能力。本章采用 VeRi 数据集训练身份识别网络。通过将各种监控环境下的车辆图像作为训练样本，以车辆的身份编号作为监督信号，训练得到的 ResNet-50 能够有效提取不同车辆在复杂监控环境下的视觉特征。因此，给定一个输入车辆图像序列 $S=\{I_1,I_2,\cdots,I_K\}$，经过身份识别子网络的前向传播运算，以 ResNet-50 的池化-5 层（Pool5）的输出作为每张图像的身份特征，得到序列 S 的身份特征矩阵 $f_I(S)=R_I\in\mathbb{R}^{L\times K}$，其中 R_I 的第 k 列为每张车辆图像 I_k 的特征 $f_I(I_k)$，L 为 ResNet-50 输出的特征向量的长度。

近几年，已有若干工作关注于基于三维视觉模型的对象识别和细粒度分类问题。例如，Su 等人[78] 通过在大规模渲染的二维图像上训练 CNN 模型，可以有效分类三维物体图像，其结果优于传统人工设计的特征。此外，Su 等人[19] 提出一种类别相关的 CNN，通过大规模三维物体渲染图像作为训练样本，物体的视角信息作为监督信号，训练得到的 CNN 模型能够准确估计图像中对象的视角信息。在城市监控视频中，由监控摄像头拍摄的车辆视频类似于对三维车辆从不同距离和角度拍摄的一组图像，如3-1所示。受上述工作启发，本章将探索利用三维视觉模型学习车辆的视角相关特征。

为了对车辆在不同视角下互补的视觉信息建模，本章设计了一种基于 CNN 的

视角估计子网络提取车辆在特定视角下的外观特征。受 Su 等人[19] 的工作启发，本章采用一种面向物体视角估计的卷积神经网络 AlexNet[18]，该网络使用大量物体图像及物体的视角信息训练。本节中使用 AlexNet 作为视角估计子网络作为特征提取器，以其全连接-7 层（Fully Connected Layer-7，FC7）输出作为车辆图像的视角相关特征。给定一个车辆图像序列 $S=\{I_1,I_2,\cdots,I_K\}$，通过视角估计子网络可以得到其视角相关特征矩阵 $f_V(S)=R_V\in\mathbb{R}^{J\times K}$，其中 R_V 的第 k 列为每张车辆图像 I_k 的特征 $f_V(I_k)$，J 为 AlexNet 的 FC7 层输出的特征向量长度。

输入的车辆图像序列经过 FLN 后，可以得到身份特征 R_I 和视角相关特征 R_V。为了将图像序列中的多个图像特征融合为一个统一的特征向量，可以采用一些直接的方法，如平均池化或最大池化，即将多个特征向量按元素求平均值或最大值。然而，这些方法难以有效挖掘不同图像特征间的互补性，无法突出特征中最有效的部分。因此，本章提出一种注意力聚合网络进一步挖掘上述特征的表示能力。

3.3.2 注意力聚合网络

如前文所述，FLN 对一段输入车辆视频中的所有图像提取了身份特征和视角相关特征。但是，由于视频中的图像拍摄于不同的视角、距离、光照，因此不同图像中特征的显著性、区分力、鲁棒性也不相同。因此，本章提出了一种 AAN，它能够根据特征的鲁棒性、区分力从身份特征和视角相关特征学习两组注意力权重。对于身份特征，AAN 对来自于车辆清晰、光照适合、细节显著图像的特征赋予较高权重，对于细节模糊、光照较暗、距离较远图像的特征赋予较低权重。通过不同的权重，能够有效区分不同车辆的特征得到增强，而可能带来歧义性的特征将被弱化。对于视角相关特征，AAN 会通过较高权重增强来自于较显著视角的特征，如车辆正面或尾部图像，反之会弱化来自于车辆侧面的特征。通过特征的加权求和，视频中多张图像的特征中最具区分力的部分将在最终的特征表示中占有更多的比例，从而提高车辆搜索的准确率。

如图3-2所示，AAN 包含两个结构相似的子网络。一般来看，每个子网络都能够通过一个注意力函数 $g(\cdot)$，从一个输入图像序列 $S=\{I_1,I_2,\cdots,I_K\}$ 的特征 $f(S)=R\in\mathbb{R}^{L\times\ K}$ 中，学习一组注意力权重向量 $\mathbf{w}=(w_1,w_2,\cdots,w_K)$，即：

$$\mathbf{w}=g(f(S)) \tag{3-5}$$

那么，图像序列 S 加权聚合后的特征可表示为：

$$\begin{aligned}\mathbf{r} &= f(S)\cdot\mathbf{w}^T \\ &= f(s)\cdot g(f(s))^T\end{aligned} \tag{3-6}$$

为学习有效的注意力函数 $g(\cdot)$，AAN 被设计为一个三层的卷积神经网络。其中，第一层为一个包含 K 个卷积核的卷积层，生成初始的 $K\times 1\times 1$ 的注意力权重向量；第二层为一个 Sigmoid 函数层，将每一个初始权重值映射到 $[0,1]$ 区间；第三层为一个 $L-1$ 正则化层，保证权重向量所有元素的和为 1。AAN 中两个子网络区别主要在于它们的输入和输出不同。其中一个子网络的输入为车辆图像序列 S 的身份特征矩阵 $R_I\in\mathbb{R}^{L\times K}$，输出为身份识别权重 $\mathbf{w}_I\in\mathbb{R}^K$。另一个子网络为车辆图像序列 S 的视角相关特征矩阵 $R_V\in\mathbb{R}^{J\times K}$，输出为身份识别权重 $\mathbf{w}_V\in\mathbb{R}^K$。那么，通过 AAN 学习的两种注意力权重 $\mathbf{w}_I$ 和 $\mathbf{w}_V$，可以得到以身份特征为基特征的两组聚合特征：

$$\mathbf{r}_I = f_I(S)\cdot\mathbf{w}_I^T \tag{3-7}$$

$$\mathbf{r}_V = f_I(S)\cdot\mathbf{w}_V^T \tag{3-8}$$

从而得到图像序列 S 最终的特征表示：

$$\begin{aligned}\mathscr{F}(S) &= \alpha\cdot\mathbf{r}_I + (1-\alpha)\cdot\mathbf{r}_V \\ &= f_I(S)\cdot[\alpha\cdot\mathbf{w}_I^T + (1-\alpha)\cdot\mathbf{w}_V^T]\end{aligned} \tag{3-9}$$

其中 α 为平衡两种特征权重的超参数。通过上述加权聚合，来自较好视角和清晰图像的特征通过较大权重增强，来自较差视角和模糊图像的特征通过较小权重弱化。最终，查询图像序列 q 与数据库图像序列 g_i 的差别由 $\mathscr{F}(q)$ 和 $\mathscr{F}(g_i)$ 的余弦距离表示，从而实现基于视频的相似车辆搜索。

3.3.3 网络训练

CAN 学习的车辆外观特征一方面需要具有足够的表达能力以表示丰富的外观特征，另一方面需能够区分相似车辆之间极其细微的类间差别。因此，我们采用一种多任务损失函数训练 CAN，该损失函数包括两个部分。第一部分是经典的分类损失函数，即交叉熵损失，表示为 L_c，它以车辆的身份作为标签计算网络的损失值。第二部分采用一种改进的对比损失函数引导网络训练，表示为 L_C。

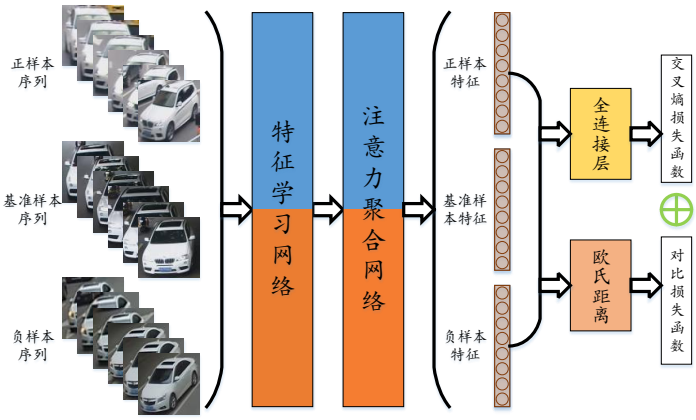


图 3-3 跨视角注意力网络的训练过程

图3-3展示了 CAN 的训练过程。一组训练样本包含三个车辆图像序列：基准序列 S_a、与基准序列车辆身份相同的正样本序列 S_p、与基准序列车辆身份不同的负样本序列 S_n。三个图像序列经过相同的 CAN 网路进行前向传播运算，输出对应的特征向量 $\mathscr{F}(S_a)$、$\mathscr{F}(S_p)$、$\mathscr{F}(S_n)$。Chopra 等人[79] 提出的原始对比损失函数 $\mathscr{L}_O(\cdot)$ 表示为：

$$\mathscr{L}_O((\mathbf{x}_1,\mathbf{x}_2,y)) = (1-y)\cdot max(m-||\mathbf{x}_1-\mathbf{x}_2||_2,0)+y\cdot||\mathbf{x}_1-\mathbf{x}_2||_2 \tag{3-10}$$

其中 x_1 和 x_2 为一对样本特征，y 为监督标签（当 x_1 与 x_2 身份相同时 $y=1$，否则 $y=0$），m 为一正数边界值。为了平衡正负样本数量，更有效训练 CAN，本章将原始对比损失函数修改为：

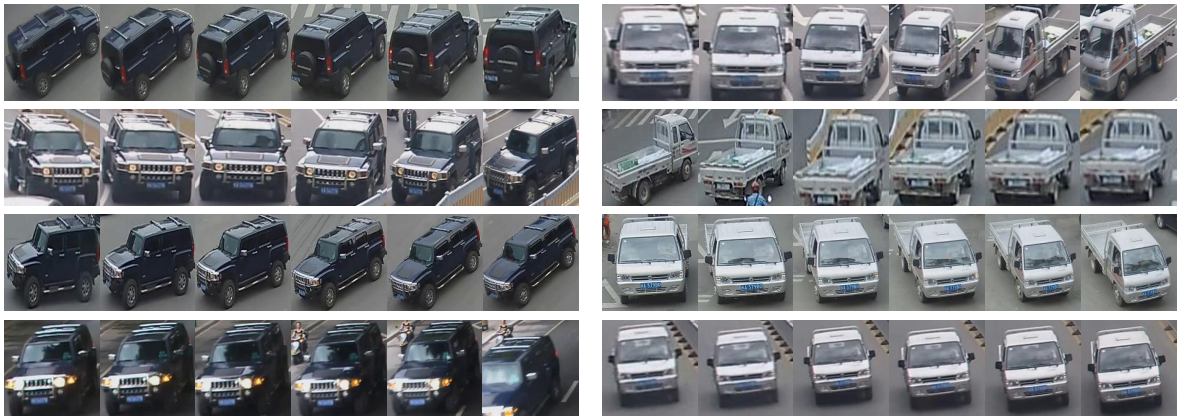
$$\begin{aligned}\mathscr{L}_C((\mathscr{F}(S_a),\mathscr{F}(S_p),\mathscr{F}(S_n))) &= L_O((\mathscr{F}(S_a),\mathscr{F}(S_p),1))+L_O((\mathscr{F}(S_a),\mathscr{F}(S_n),0)) \\ &= ||\mathscr{F}(S_a),\mathscr{F}(S_p)||_2+max(m-||\mathscr{F}(S_a),\mathscr{F}(S_n)||_2,0)\end{aligned} \tag{3-11}$$

其中 m 为一正数边界值。对比同样以三元组为输入的三元损失（Triplet Loss）[80]，改进的对比损失函数对负样本施加更严的约束，使身份不同的样本在特征空间中距离足够远，以增强特征的可区分性。通过大量样本的训练，CAN 能够学习到有效的权重函数，从而挖掘特征中最局区分性的部分，弱化特征中的冗余部分，使正负样本更易区分。

最后，整个 CAN 网络通过随机梯度下降算法进行训练，损失函数表示为：

$$L = L_c+\lambda\times L_C \tag{3-12}$$

其中 λ 为一平衡因子。具体实现时，首先采用交叉熵损失函数 L_c 训练 FLN。然后，用预训好的 FLN 提取训练样本的身份特征和视角相关特征，并组成三元组训练样本。



(a) (b)

图 3-4 VIVID 数据集中的车辆视频数据样本

下一步，使用上述训练样本和改进的对比损失函数 $\mathscr{L}_C$ 训练 AAN。最后，整个网络使用损失函数 L 进行端到端的微调训练。其中，公式3-10中的边界值设为 $m=1.0$，公式3-8中的平衡参数设为 $\alpha=0.5$，公式3-13中的平衡参数设为 $\lambda=0.1$。

3.4 实验结果与分析

3.4.1 数据集与实验设置

为评估提出的基于视频的车辆搜索方法 CAN、推动未来的相关研究工作，我们收集并标注了一个大规模基于视频车辆搜索数据集——“VIVID”。数据集原始视频来自于 20 个城市交通监控摄像头，由于其安装在任意位置、高度、朝向，能够拍摄到不同角度的车辆视频。此外，数据集包含了各种各样的车辆行为、不同的形式速度和方向，如由近及远（及其反方向）直行、从左到右（及其反方向）直行、左右转弯、反方向掉头等。更进一步，车辆视频截取自各种不同的监控环境，反映了真实城市交通场景中多变的光照、复杂的背景与遮挡等情况。

在数据集标注过程中，对于视频中出现的车辆，由人工标注员手工截取车辆图像的包围盒，并将连续车辆图像合并为一组图像序列。此外，数据集还记录了每个车辆包围盒在整个视频帧中的坐标、时间戳等信息，可用于提取车辆的运动特征或时空信息。最后，每个车辆视频标注了车辆的身份及所出现的摄像头编号，以保证车辆搜索研究所需的车辆跨摄像头重现信息。最终，VIVID 数据集包含 776 辆不同身份车辆的 8578 个视频片段（平均每辆车约 11 个片段）。类似于第2.5.1.1节提出的 VeRi 数据集，VIVID 数据集中的每个车辆视频都标注了车型、颜色、车牌、时空信息等标签。图3-4展示了部分车辆视频图像，可以看出 VIVID 数据集中包含了车辆多

种角度的视频、各种车辆行为、不同的行驶方向，是一个数据丰富又具有挑战性的综合数据集。

在本章的实验中，首先将 VIVID 数据集分为了包含 576 个车辆、6557 个片段的训练集和包含 200 个车辆、2021 个片段的测试集。对于测试集中的每个车辆，从不同摄像头拍摄的视频中选择了共计 1678 个片段作为查询集。类似于第2.5.2节采用的评价标准，我们采用排序第一准确率（HIT@1）和排序第五准确率（HIT@5）评价方法的准确率。由于一个查询可能存在多个正确结果，因此需要同时评估方法的准确率和召回率，我们采用平均准确率均值（mean Average Precision，mAP）综合评估车辆搜索方法。其中，对于一个查询图像，平均准确率（Average Precision，AP）表示为：

$$AP = \frac{\sum_{k=1}^{N_g} P(k) \times gt(k)}{N_{gt}} \tag{3-13}$$

其中 N_g 为数据库集合 G 的样本数，N_{gt} 为该查询图像对应的正确样本数，$P(k)$ 为排序第 k 准确率，$gt(k)$ 为一指示函数，当第 k 个结果为正确时返回 1，否则返回 0。对于全部查询视频，平均准确率均值 mAP 表示为：

$$mAP = \frac{\sum_{q=1}^{N_q} AP(q)}{N_q} \tag{3-14}$$

其中 N_q 为查询集的视频片段数。

3.4.2 基于单张图像的方法与基于视频的方法对比

为了探究在车辆搜索中视频能够比单张图像提供更多有效的视觉信息，本节比较了三种不同的查询样本与被查询样本匹配策略：图像-图像匹配、图像-序列匹配、序列-序列匹配。此外，本节还比较了不同的特征聚合方法，如最大池化和平均池化。在本章实验中，为了方便计算和比较，所有基于视频的搜索方法都将原始车辆视频采样得到长度为 6 帧的图像序列。为对比基于单张图像与基于视频的方法，本节使用相同的 ResNet-50 网络[77] 作为图像特征提取器。该网络首先在 ImageNet 图像分类数据集[71] 进行预训练，然后使用第2.5.1.1节提出的 VeRi 数据集中的车辆图像和身份标签微调训练。最后，提取 ResNet-50 的 Pool5 层长度 2048 的输出作为特征向量进行特征匹配。本章所有实验基于 Caffe 深度学习工具[81] 实现，各类方法细节如下：

- 方法一，图像-图像匹配（**Image-to-image matching，Img2Img**）。该方法是基

本的以图搜图方法，即对于每个查询图像序列或数据库图像随机选择一张图像 ResNet-50 的输入。提取图像特征后，计算特征间的余弦距离，返回结果以每个数据库图像与查询图像的距离有小到大排序。该方法代表了最基本的基于单张图像的相似车辆搜索方法。

- 方法二，图像-序列匹配（**Image-to-sequence matching with average pooling and maximum pooling，Img2Seq_AVG 和 Img2Seq_MAX**）。这一类匹配策略的特点是，输入查询是从查询图像序列中随机选择的一张图像，数据库数据是车辆图像序列。根据被查询序列图像特征的聚合方式不同，该策略包括两种方法。第一种是使用 ResNet-50 提取一个图像序列中所有图像的特征向量后，对这些向量计算相同位元素的平均值，对一个图像序列得到一个长度为 2048 的特征向量。第二种则是对所有图像特征的相同位元素取最大值。这两种方式是常见的视频内容分析问题（如视频分类[82]、动作识别等[83]）常用的多图像特征聚合方式。在搜索时，计算查询图像特征与聚合后图像序列特征的余弦距离。

- 方法三，序列-序列匹配（**Sequence-to-sequence matching with average pooling and maximum pooling，Seq2Seq_AVG 和 Seq2Seq_MAX**）。这一类匹配策略的特点是，输入查询和数据库数据都是车辆图像序列。根据车辆图像序列特征的聚合方式不同，该策略也包括两种方法。第一种是使用 ResNet-50 提取一个图像序列中所有图像的特征向量后，对这些向量计算相同位元素的平均值，对一个图像序列得到一个长度为 2048 的特征向量。第二种则是对所有图像特征的相同位元素取最大值。在搜索时，计算聚合后的查询图像序列特征和数据库图像序列特征的余弦距离，距离越小则相似性越大。这两种方法代表了基于视频的车辆搜索的基准方法。

- 方法四，跨视角注意力网络（**Cross-view Attentive Network，CAN**）。该方法为本章所提出的基于视频的车辆搜索方法，实现细节如第3.3.1和3.3.2节所述。

表3-1列出了上述方法在 VIVID 数据集上的准确率对比，图3-5展示了不同方法的 CMC 曲线对比。通过实验结果，可以得出如下结论：

- 首先，通过对比 Img2Img 与 Img2Seq_AVG 可以发现，平均池化方法得到的特征保存了数据库图像序列中的全部特征，因此它的准确率优于基于单张图像匹配的车辆搜索方法。但是，最大池化方法 Img2Seq_MAX 的性能差于上述两种

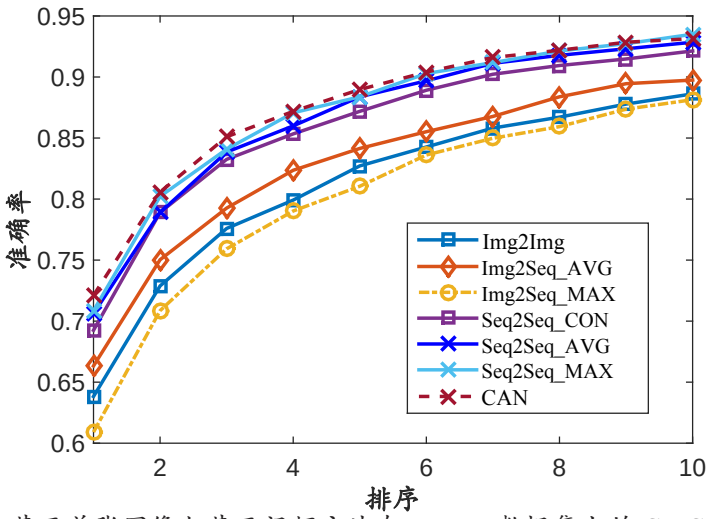


图 3-5 基于单张图像与基于视频方法在 VIVID 数据集上的 CMC 曲线对比

表 3-1 基于单张图像的车辆搜索方法与基于视频的车辆搜索方法结果对比

方法	mAP	HIT@1	HIT@5
Img2Img	30.14	63.83	82.72
Img2Seq_AVG	32.97	66.33	84.15
Img2Seq_MAX	29.22	60.91	81.05
Seq2Seq_AVG	35.98	70.62	88.38
Seq2Seq_MAX	35.87	70.86	88.38
CAN	**36.98**	**72.11**	**89.27**

方法，这是由于最大池化只保留一个位置上的最大元素而丢弃了其它元素，因此特征中的有效部分可能被丢弃。

- 两种基于图像序列的匹配方法取得了接近的准确率，并优于上述三种方法。因此，在查询数据和数据库数据都为图像序列时，平均池化和最大池化都能保存多幅图像中的有效特征。但是，这两种特征池化方法并没有挖掘多张图像特征的互补性。特别地，平均池化方法对一个图像序列中的所有图像一视同仁，将它们的特征平等地保留，无法突出其中的有效部分。

- 与上述方法对比，本章提出的 CAN 框架取得了最好的准确率。这是由于，CAN 不仅保存了图像序列中的全部成分，而且对来自较好角度、细节鲜明图像的特征赋予较高权重，从而突出特征中的显著部分，对来自模糊图像、非显著视角图像的特征赋予较低权重，从而降低冗余特征的重要性。

3.4.3 基于视频的车辆搜索方法对比

为综合评估提出的 CAN 框架，本节将 CAN 与现有基于单张图像和基于视频的车辆搜索方法进行了比较。此外，为研究动作模型对车辆搜索的作用，CAN 还与主要的基于视频的人员搜索方法进行对比。最后，本节通过实验研究了视角信息对 CAN 框架及车辆搜索的作用。各类方法实现细节与特点如下：

- 方法一，基准方法（**Seq2Seq_AVG**）。该方法是基于视频的相似车辆搜索的基准方法，实现细节同上一节所述。

- 方法二，颜色与属性融合方法（**FACT**[75]）。该方法即第2.3节提出的多级车辆外观融合方法，本节中采用该方法对车辆图像序列中的图像提取多级车辆外观特征，然后采用平均池化策略将多张图像的特征融合，从而实现基于视频的车辆搜索。

- 方法三，渐进式搜索（**Progressive**[37]）。该方法是一种基于单张车辆图像的渐进式搜索方法，采用车辆外观特征、车牌特征、时空关系等信息，渐进式地在数据库中搜索与查询车辆相同的车辆。

- 方法四，身份特征与长短时记忆网络（**ResNet + LSTM**）。该方法采用了目前基于视频的人员重识别方法中性能最好的框架[57]，该框架采用 CNN 提取单帧图像的静态视觉特征，然后使用 LSTM 网络从整个图像序列的特征中学习连续动态的特征，如人的动作、步态等。本节采用该框架对车辆视频的动态特征进行建模，以探究车辆动态特征对车辆搜索的作用。

- 方法五，身份特征加视角相关特征与长短时记忆网络（**ResNet + View + LSTM**）。该方法采用与 ResNet + LSTM 相似的网络框架。不同之处在于，输入 LSTM 的除了 ResNet 提取的身份特征，还包括视角估计网络提取的视角相关特征。通过该方法，能够探索视角信息对车辆动态特征建模的作用。

- 方法六，身份特征与身份注意力网络（**ResNet + IAN**）。该方法的网络结构仅包含 CAN 结构图3-2的上半部分，即 FLN 仅对车辆图像序列提取身份特征，然后 AAN 对身份特征学习一组身份识别权重，整个网络结构采用多分类交叉熵损失函数引导训练。本节通过对比该方法与所提出的 CAN，研究视角相关特征及其注意力权重是否有助于车辆搜索任务。

表 3-2 不同外观相似车辆搜索方法在 VIVID 数据集上的准确率对比

方法	mAP	HIT@1	HIT@5
Seq2Seq_AVG	35.98	70.62	88.38
FACT[75]	18.00	52.44	72.29
Progressive[37]	25.11	61.26	75.98
ResNet + LSTM[57]	28.11	56.20	79.14
ResNet + View + LSTM	29.42	54.53	80.75
ResNet + IAN	36.23	71.22	88.80
ResNet + IAN + CL	36.51	71.22	88.92
CAN	**36.98**	**72.11**	**89.27**

- 方法七，身份特征与采用对比损失函数训练的身份注意力网络（**ResNet + IAN + CL**）。此方法与 ResNet + IAN 方法采用相同的网络结构，即仅使用身份特征及其注意力权重对车辆外观建模。二者不同之处是本方法采用第3.3.3节提出的改进对比损失函数及三元组训练数据进行网络训练。

- 方法八，跨视角注意力网络（**CAN**）。该方法为本章所提出的基于视频的车辆搜索方法，实现细节如第3.3.1和3.3.2节所述。

表3-2列出了上述方法在 VIVID 数据集上的准确率对比。通过实验结果，可以得出如下结论：

- 首先，已有的基于单张图像的车辆搜索方法（即 FACT 和 Progressive）准确率比所有基于视频的方法都低。一方面由于 FACT 和 Progressive 中使用的车辆外观特征是人工设计的和由较浅的 CNN 学习的，无法有效表示监控环境下车辆的外观特征。另一方面由于仅适用单张车辆图像很难捕捉到足够的具有区分力的视觉特征。

- 此外，通过对比基于 LSTM 的方法与其它基于视频的车辆搜索方法，可以看出基于 LSTM 的方法 mAP 比基准方法 Seq2Seq 还要低 7.87%。这一点说明 LSTM 很难像基于视频的人员重识别任务中那样从车辆视频中学习有效的动态特征。该实验也验证了本章第3.1节的观点，即车辆作为一种刚性物体，在行驶过程中很难体现出有规律的动态模式，因此现有时序建模方法很难从车辆视频中提取对车辆搜索有用的时序动态特征。

- 更进一步，通过 ResNet + LSTM 与 ResNet + View + LSTM 的对比、ResNet +

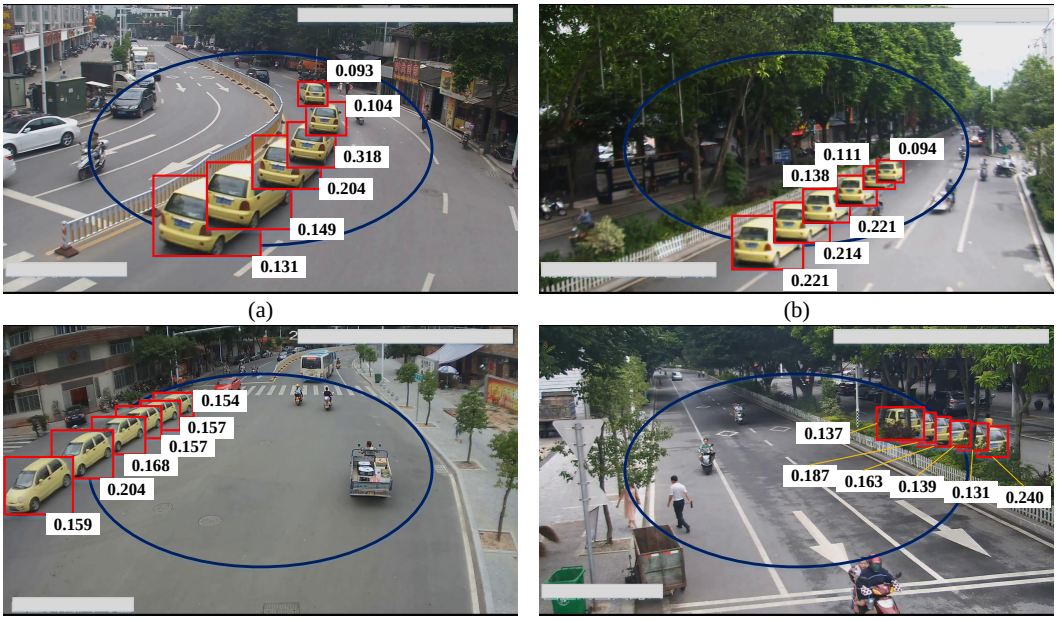


图 3-6　VIVID 数据集中的测试样本及对应的注意力权重

IAN + CL 与 CAN 的对比，可以看出加入视角相关特征的网络比仅适用身份特征的网络取得了更好的准确率。该对比表明视角相关特征对基于视频的车辆搜索任务是有效的。

- 最后，本章提出的跨视角注意力网络取得了 VIVID 数据集上的最优准确率。因此，FLN 提取的身份特征和视角相关特征都是不可或缺的车辆外观特征。AAN 从上述两种特征学习的注意力权重能够更有效地挖掘多张图像特征的互补性，使车辆特征更具区分力和鲁棒性。此外，改进的对比损失函数能够使网络更有效地训练，预防过拟合现象。

3.4.4　讨论

为直观展示 CAN 的作用，图3-6展示了 VIVID 数据集中同一个车辆的四个不同图像序列及对应的注意力权重值，其中每一行为一个序列。第一行和第二行分别展示了车辆左转和由近及远行驶的视频，可以看出车辆尾部正视角比侧视角能够获得更高的权重，而距离较远、细节较模糊的图像会获得较低的权重。因此可以证明 AAN 能够有效感知图像特征的显著性和鲁棒性，提高车辆外观特征的区分能力。第三行展示了从车辆右向左直行的视频，其中的车辆到摄像头的距离和视角基本不变，因此各张图像得到接近的权重。第四行是 CAN 的一种失败案例，这是由于图像中包

含很大的遮挡区域，使 CAN 很难提取有效的车辆外观特征和相应的注意力权重。因此，在未来的工作中可以引入区域注意力机制，使网络能够自动发现图像中的显著性区域，从而克服局部遮挡对外观特征提取的影响。

3.5 本章小结

我们提出一种全新的跨视角注意力网络 CAN，是对基于视频的相似车辆搜索的初步尝试。CAN 由两个基于 CNN 的模块组成：特征学习网络（FLN）用于从车辆图像序列中提取单张图像的车辆身份特征和视角相关特征；注意力聚合网络（AAN）用于从上述特征学习两组注意力权重。通过注意力权重，图像序列的特征通过加权聚合形成一个统一的外观特征向量。其中，来自于显著视角、细节清晰图像的特征被赋予较高权重，冗余和歧义的特征被赋予较低权重，使得车辆外观特征更具区分力。此外，为了评估提出的方法、推动相关研究发展，我们从真实监控视频中收集并标注了一个大规模视频车辆搜索数据集——VIVID。经过大量的实验，验证了所提出方法对基于视频的相似车辆搜索任务的有效性。在未来的工作中，该方法的思想也可应用于其它问题，如基于视频的人员搜索或步态识别。

第四章 车牌图像增强与验证结合的精确车辆搜索

4.1 引言

如车辆搜索任务的定义所述，给定查询车辆图像或视频，在大规模监控摄像头采集的车辆数据库中搜索与查询车辆身份相同的车辆。现有车辆搜索方法[34,38,39,75]，包括本文第2和3章所提出的 NuFACT 和 CAN 框架，主要利用车辆图像中提供的外观特征进行车辆匹配和搜索。但是，由于车辆外观具有很大的相似性，尤其是相同品牌、车系、颜色的车辆，很难仅通过车辆的外观唯一地确定相同车辆。根据外观的视觉特征仅能搜索到与查询车辆外观最相似的车辆。如图4-1所示，为了实现身份级别的精确车辆搜索，搜索系统必须考虑车辆的唯一性标识，即车辆的牌照信息。

在传统车辆身份识别中，车牌识别系统起到至关重要的作用[5,25]。通过车牌识别系统识别出车牌字符，用户可以通过输入车牌字符搜索目标车辆。这类系统通常应用于高速路收费站、停车场出入口、校园门口等有约束区域，车牌识别系统需要停车杆、闪光灯、传感器等辅助设备才能准确识别车牌字符。然而，在无约束的城市监控场景中，由于摄像头安装的位置各异、拍摄距离可能较远、环境光照随着日光或灯光变化、车辆在道路上高速行驶，监控视频中的车牌图像可能产生模糊、倾斜、遮挡等退化，这就导致现有的车牌识别系统无法准确识别车牌字符信息，如图4-1所示。因此，如何利用无约束监控视频中的车牌信息实现精确的车辆搜索，是本章的主要挑战。

为克服上述挑战，我们从两方面入手：第一，针对无约束监控场景下车牌图像的低质问题，采用一种域先验生成对抗网络（Domain Prior Generative Adversarial Network），将搜索到的外观相似车辆的低质车牌图像转换为一张清晰的车牌图像。第二，针对监控数据中车辆数量很大而每个车辆样本较少的问题，采用一种基于对偶神经网络的车牌验证方法，直接验证两个车辆的车牌图像是否相同而不识别和匹配车牌字符，实现车牌图像的快速准确匹配，从而达到车辆身份准确匹配。

具体来说，对于无约束监控视频中低质车牌图像的增强，我们提出了通过利用监控系统中不同摄像头拍摄的多张车牌图像生成一张高质量的车牌图像。如图4-2所示，给定一张带有车牌的查询车辆图像，在不同摄像头中搜索与其外观相似的车辆。

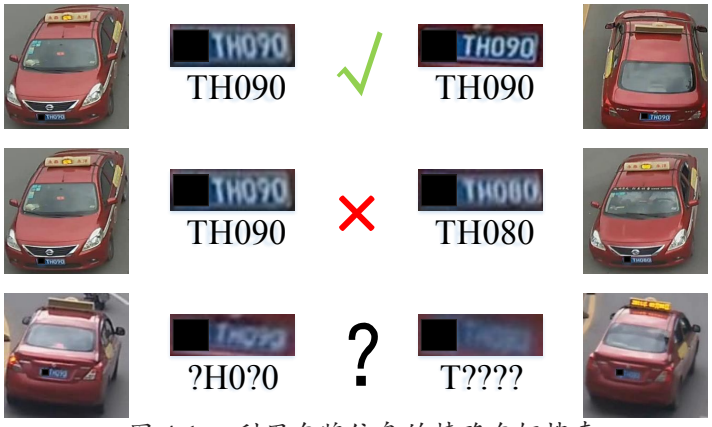


图 4-1　利用车牌信息的精确车辆搜索

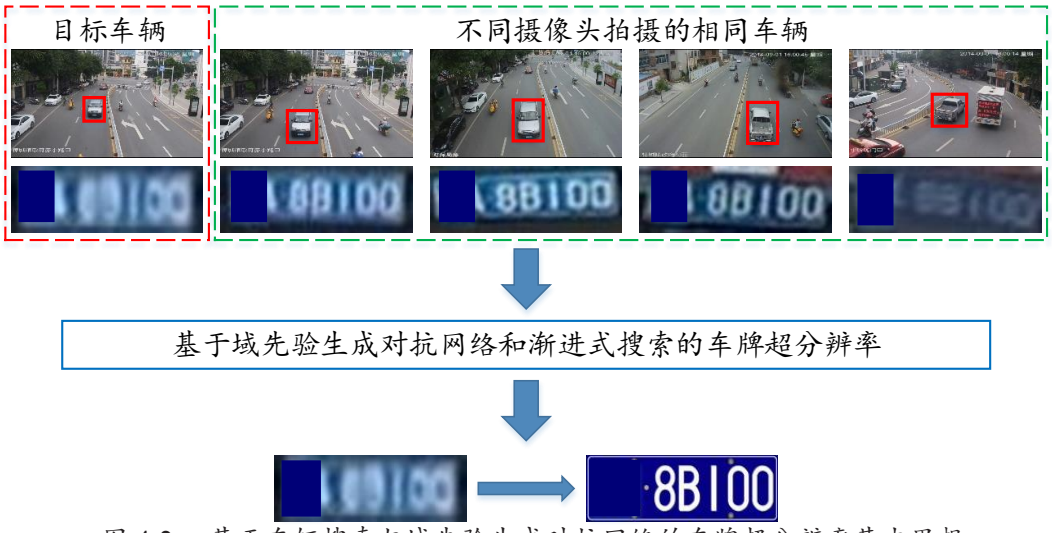


图 4-2　基于车辆搜索与域先验生成对抗网络的车牌超分辨率基本思想

下一步是如何利用这些被搜索到的车辆中的车牌图像，恢复查询车辆的真实清晰的车牌图像。毫无疑问，可以采用现有的基于多张图像的超分辨率方法（Multi-image Super-Resolution，MSR）。但是，现有的 MSR 方法主要关注于基于连续帧或视频的超分辨率，要求输入的多张图像场景基本相同、变化相对较小、前后两张图像间具有一定时序关系。而通过车辆搜索得到的车牌图像来自于不同摄像头、不同场景和环境，因此包含不同的模糊、分辨率、遮挡、仿射形变、光照、视角等状态。这给由多张车牌图像恢复为一张清晰车牌图像带来了巨大的挑战。此外，车辆搜索得到的结果可能包含于查询车辆车牌不一样的车牌，从而给车牌超分辨率带来错误的噪声信息。因此，由于以上挑战，现有的 MSR 方法无法直接用于本章的车牌超分辨率问题。

综上所述，我们提出了一种域先验生成对抗网络（DP-GAN），通过将不同摄

图 4-3　基于车辆搜索与域先验生成对抗网络的车牌图像增强方法结果展示

像头、时间、距离、视角拍摄的多张车牌图像，校正恢复为空间一致的高分辨率车牌图像。DP-GAN 包含两个对抗训练的子网络：生成器网络（Generator Network，GN）和判别器网络（Discriminator Network，DN）。GN 是一个全卷积网络（Fully Convolutional Network，FCN），用来学习一个由低分辨率（Low Resolution，LR）车牌图像向高清（High Resolution，HR）车牌图像的映射，它由一个卷积层、*N* 个残差模块、两个反卷积层组成。为了有效训练 GN，需要大量的 LR-HR 车牌图像对。其中，低分辨率图像可以从真实监控系统中收集得到，但相对应的高分辨率车牌图像很难在复杂的监控环境中得到。为解决这一问题，我们采用图像渲染的方式生成高分辨率车牌图像，即根据车牌设计与制作的标准，为每一个低分辨率车牌合成一张高分辨率车牌图像。除了对低分辨率车牌进行清晰化、提高分辨率，各种各样倾斜、形变的车牌还能通过 GN 校正为规则的车牌图像，从而提高车牌识别和车辆搜索的准确率。此外，传统 FCN 的优化目标是自动学习 LR-HR 图相对的像素级差异，而忽视了车牌整体的空间先验知识，称为域先验（Domain Priors，DP）。因此，GN 通常会生成奇异的字符，导致车牌识别和车辆搜索准确率下降。为了避免这种现象，我们训练了一个 VGG 网络[74] 作为 DP-GAN 的 DN，用于区分真实的高分辨率图像和 GN 生成的图像。更重要的是，VGG 网络中加入了一个全新的空间分割层（Spatial Split Layer，SSL），以保持车牌的空间先验知识，如车牌的整体和局部制造规则。最后，在网络训练时，对抗损失函数与内容损失函数紧密结合，增强 GN 和 DN 的学习能力。

更进一步，我们将提出的 DP-GAN 网络引入基于渐进式车辆搜索的多车牌超分辨率框架中，如图4-5所示。在车辆搜索阶段，首先采用第2章提出的外观相似车辆搜索方法 NuFACT 搜索查询车辆在不同摄像头的图像。然后，通过提出的基于 SNN 的车牌验证方法进行更精细的车辆搜索。最后，使用时空信息对搜索结果进行重排序。在车牌超分辨率（Super Resolution，SR）阶段，搜索到的多张车牌图像首先通过 DP-GAN 校正、超分辨率操作，得到多张高分辨率车牌图像。然后，将生成的多

张高分辨率车牌输入一个多图 SR-GAN（MSR-GAN）中。MSR-GAN 与 DP-GAN 结构类似，区别在于 MSR-GAN 将多张高分辨率车牌作为不同通道合并为一个数据张量作为输入。最终，通过 MSR-GAN 的前向运算，多张车牌图像被映射为一张高清车牌图像。图4-3展示了 MSR-GAN 生成的高分辨率车牌实例，一些人眼难以识别的车牌能够被恢复为清晰的车牌。此外，为评估提出的车牌增强方法，我们收集并标注了一个真实监控场景的车牌数据集。在性能评估中，传统 SR 评估标准主要考虑低分辨率图像与高分辨率图像间的像素级差异。而本章采用了车牌 SR 的下游任务，即车牌识别和车辆搜索，作为评估车牌质量的标准。使用了提出的方法后，自动车牌识别准确率由 7% 提升到 51%，甚至超过了人眼识别的能力（仅通过查询图像时 7%，采用车辆搜索时 49%）。此外，通过将提出的车牌 SR 方法引入车辆搜索任务，准确率达到了 VeRi 数据集上的最好结果。

针对无约束监控场景中车牌字符难以识别的问题，我们并不采用传统的光学字符识别（Optical Character Recognition，OCR）技术识别车牌图像中的字符和数字，而是直接提取车牌图像的视觉特征，采用车牌验证的方法计算两张车牌图像的视觉相似度，进而判断两张车牌图像是否相同。针对监控数据中车辆数量很大而每个车辆样本较少的问题，本章采用一种基于对偶神经网络 (Siamese Neural Network，SNN) 的车牌验证方法。SNN 是一种将 CNN 与度量学习方法结合的神经网络模型，最早用于手写签名验证[84]。SNN 通常由两组并行且共享参数 CNN 组成，用于学习和提取图像的多层图像特征。在训练时，SNN 采用了度量学习的思想，输入一对图像样本，如两张图像类别（身份）则标签为 1，反之标签为 0。通过对比损失函数引导网络训练，SNN 能够学习一种映射，使得相同类别（身份）的对象在特征空间中距离较近，不同类别（身份）的对象距离较远，从而实现一对样本相似度的计算。SNN 的这一特点也使其特别适合于样本类别较多但每一类的样本较少的问题，如人脸验证[79]、步态识别[85] 等。因此，本章采用大量车牌图像对，训练一个用于车牌验证的 SNN。在车辆搜索阶段，首先使用本文第2和3章提出的外观相似车辆搜索方法，为输入的查询车辆找到外观最相似的候选车辆集合。然后，使用训练好的 SNN 提取车牌图像的特征向量，计算查询车辆车牌与数据库车辆车牌的距离，实现车辆的精确搜索，如图4-4所示。

本章的主要工作与贡献如下：

- 第一，设计了一种全新的车牌超分辨率框架，首次利用不同摄像头拍摄的多张车牌图像生成高质量车牌图像。

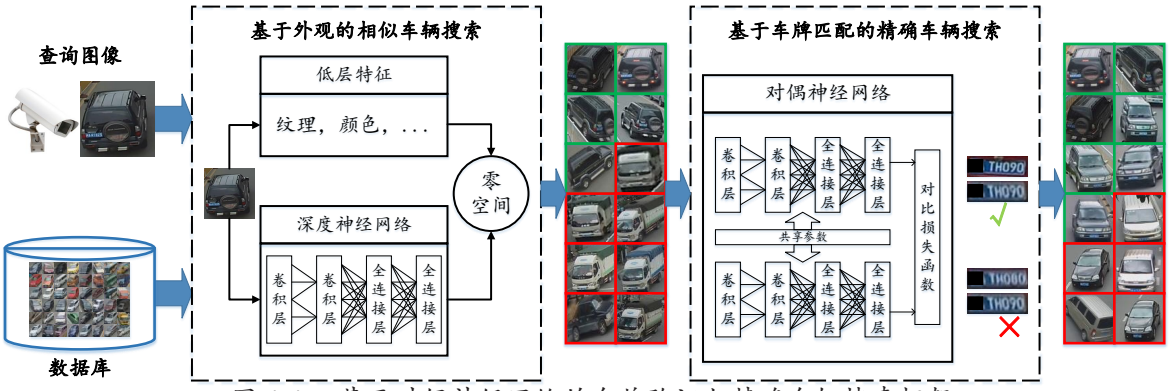


图 4-4　基于对偶神经网络的车牌验证与精确车辆搜索框架

- 第二，为生成空间结构一致、高清晰度的车牌图像，提出了一种能够同时学习训练数据分布知识和车牌制造先验知识 DP-GAN 方法，基于 DP-GAN 和多张车牌输入的 MSR-GAN 框架最终生成了高质量的车牌图像。

- 第三，在渐进式车辆搜索的框架中，为了有效利用监控视频中的车牌信息，提出了一种基于 SNN 的车牌验证方法，实现车辆的精确搜索。

- 第四，通过车牌增强与验证的有机结合，使得本章的车辆搜索方法在公开的数据集上取得了最好的准确率。

4.2　相关工作

由于本章涉及若干独立于车辆搜索的问题，如车牌识别和图像超分辨率，本节将简要介绍与以上任务相关的研究工作。

4.2.1　车牌识别

在真实世界的车辆身份识别任务中，车牌识别是最常见的技术[5,25]。传统车牌识别技术通常包括车牌检测（定位）、校正、分割、字符识别等主要步骤。首先，通过车牌定位可以将一整张监控图像帧中车辆的车牌区域定位，通常采用人工设计特征与滑动窗口匹配框架[5]。最近，基于 CNN 的对象检测方法[86,87] 通过大量标注数据训练，使车牌检测准确率进一步提高。直接从视频帧中截取的车牌图像通常包含仿射变换，车牌校正的目的是将车牌变换为角度端正的车牌，以便后续处理。对于校正后的车牌，通常基于车牌制造标准和图像处理技术将整张车牌图像分割为仅包

含单一字符的图像块。最后，提取每个图像块的视觉特征，使用分类器或神经网络对每个字符进行识别。

现有的车牌识别系统通常对车牌图像清晰度具有较高要求。因此，车牌识别系统中的摄像头通常安装于有约束场景，如高速路收费站、停车场出入口、十字路口等，并且需要停车杆、闪光灯、龙门架、传感器等辅助设备。但是，在无约束的城市监控环境下，车辆的行驶行为、方向、速度无法受到约束，监控摄像头受到拍摄距离、光照、遮挡等不确定因素的影响。在这种情况下，车牌识别系统很难正确识别监控图像中的车牌信息，更无法根据车牌号完成车辆搜索任务。为此，本章提出采用车牌验证方法代替车牌识别，即使用 CNN 直接提取车牌区域的视觉特征，计算两张车牌的视觉相似性，从而验证两张车牌是不是相同，以实现精确的车辆搜索。

4.2.2 图像超分辨率

图像超分辨率方法主要可以分为基于单张图像的超分辨率（Single-image Super Resolution，SSR）和基于多张图像的超分辨率（Multi-image Super Resolution，MSR）。

4.2.2.1 单图超分辨率

SSR 的目标是从一张低分辨率图像生成一张高分辨率图像，因其在计算机视觉和多媒体领域的重要作用受到研究者的广泛关注[88]。一般性的 SSR 方法可以分为三类：基于差值的方法、基于重建的方法、基于学习的方法。

在早些年间，基于差值的方法是基本的图像超分辨率方法，如线性插值、双线性插值、双三次差值等，通常采用预定义的计算公式将低分辨率图像转换为高分辨率图像[89,90] 尽管这类算法计算效率很高，但会在生成的结果图上造成很多瑕疵、模糊，无法达到许多真实应用中的分辨率要求[91]。基于重建的方法通常使用纹理级先验，如梯度轮廓先验[92] 和边缘先验[93] 这些先验知识通过统计方法在大量自然图像中获得，用于在生成高分辨率图像时作为约束条件。但这类算法通常能够生成比较锐利的边缘，但会忽视图像中的局部结构和复杂的细节。

近几年，基于学习的算法，又称为基于样本的算法，受到学者的广泛关注。这类方法通常使用机器学习算法，如稀疏编码[94]、字典学习[95]、回归模型[96] 等，从大量训练图像样本中学习一个从 LR 到 HR 的映射。此外，多位学者探索使用 CNN 直接从大量 LR-HR 图像对中学习一个网络，取得了超过传统方法的性能[97–99]。最近，Ledig 等人[100] 提出一种超分辨率生成对抗网络（Super Resolution Generative Adversarial

Network，SR-GAN）学习一个用于生成高分辨率图像的生成模型。通过引入一个感知损失函数和对抗式训练策略，SR-GAN 生成的图像比已有方法生成的图像有更好的视觉效果。除了一般性 SSR 方法，目前还有许多面向特定对象的 SR 方法，如生成人脸[101,102]、车牌[103]、场景[104] 等的高分辨率图像。这些方法通常将各个领域的先验知识，如人脸的结构等，融入已有的 SSR 算法，取得了特定领域中的优异效果。受以上工作启发，本章将 GAN 与车牌先验知识融合，提出一种 DP-GAN 提高车牌超分辨率效果。

4.2.2.2　多图超分辨率

本章基于车辆搜索的车牌 SR 框架实际上是一个多图超分辨率（MSR）问题。现有的 MSR 方法通常关注于视频中连续图像帧的超分辨率[105–109]。早期的工作将 SSR 算法应用于连续视频帧，但忽视了视频帧间的关系与时序信息[105]。不同的是，Liu 等人[106] 提出一种贝叶斯方法，通过估计视频中的运动、模糊、噪声等参数，实现自适应的视频超分辨率。为处理 MSR 中的运动模糊问题，Ma 等人[107] 等人提出了一种 EM 框架引导残差模糊估计和高分辨率图像重建过程。CNN 及 RNN 同样被应用于 MSR 问题，通过同时学习视频图像序列的空间映射和时序关系达到了目前最好的效果[108]。

但是，在本章的多车牌超分辨率问题中，通过车辆搜索得到的多张车牌图像由不同摄像头在不同场景下拍摄，而非连续视频帧中的图像。因此，无法采用已有方法，如贝叶斯方法或运动估计，学习图像之间的时序信息。此外，已有方法很难有效处理无约束监控场景中车牌的极端变化，如视角、光照、运动模糊等，因此无法直接应用于本章的车牌 MSR 问题。

4.3　基于域先验生成对抗网络的车牌图像增强

针对无约束监控场景下摄像头拍摄的车牌图像的退化问题，我们提出将车辆搜索得到的多个候选车辆的车牌输入一个域先验生成对抗网络（DP-GAN），将低分辨率车牌图像恢复为一张高分辨率车牌图像，如图4-5所示。本节详细阐述 DP-GAN 的原理与训练方法，并构建一个基于 DP-GAN 与多车牌输入的多图像高分辨率框架（MSR-GAN）。

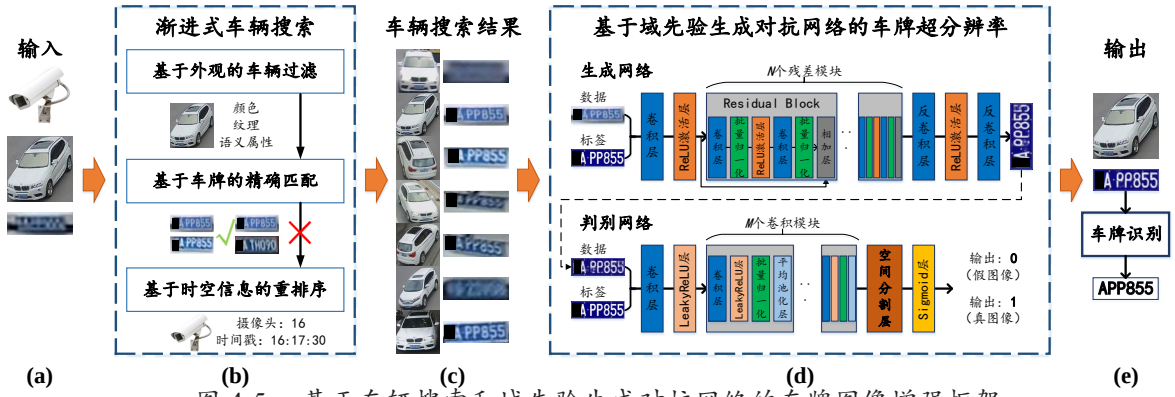


图 4-5　基于车辆搜索和域先验生成对抗网络的车牌图像增强框架

4.3.1　域先验生成对抗网络框架

一般而言，生成对抗网络（Generative Adversarial Network，GAN）[110] 包含两个并行的生成器网络（GN）和判别器网络（DN），两个网络交替、迭代训练，实现互相对抗、促进的效果，从而提高生成器网络的能力。GN 用于建模训练数据的数据分布，DN 用于估计一个输入样本来自于真实数据而非 GN 生成数据的概率。直观而言，GN 通过训练数据的分布生成足够逼真的样本，而 DN 类似于一个裁判判断 GN 生成的样本能否骗过自己。通过 GN 与 DN 的对抗式训练，GN 的生成能力与 DN 的判别能力能够显著提高。

对于车牌 SR 问题，本章提出一种 DP-GAN，如图4-5所示。与 GAN 类似，DP-GAN 中的 GN 以低分辨率车牌 P_l 为输入生成一个高分辨率车牌 P_h，DN 用于判断输入是真正的车牌 (Groundtruth)P_g 还是由 GN 生成的车牌 P_h。在训练过程中，GN 的优化目标是最小化输入 P_l 与输出 P_g 的像素差别。不同的是，DN 的目标是学习车牌制造标准的领域先验知识（Domain Prior，DP）。此外，对比现有的图像超分辨率生成对抗网络（Super Resolution Generative Adversarial Network，SR-GAN）[100]，DP-GAN 的创新点可以总结为：

- 通过大量带有域先验的合成数据训练 GAN，能够有效提高车牌 SR 性能；
- 对于 GN，提出一种合成流程，从而生成具有空间一致性的高分辨率车牌，一共后续 MSR 处理；
- 对于 DN，设计了一个空间分割层，使模型能够同时关注车牌的全局和局部先验知识。

下面各节将详细介绍上述创新点。

4.3.2　生成器网络

在低层视觉任务（如图像超分辨率和去模糊）中效果最优异的方法，通常采用全卷积网络学习一种从输入低质图像到高质量图像的映射[97]。因此，我们采用 FCN 作为 GN 的主体结构，使 GN 能够学习一个端到端的由低质车牌 P_l 到高质车牌 P_h 的映射。为了提高极深神经网络的学习能力、防止训练中的梯度弥散现象，GN 采用了层级相连的 N 个残差模块[77] 作为网络主干，如图4-5（d）所示。如 He 等人[77] 所提出的结构，每一个残差模块包含两个卷积层（64 个 3×3 的卷积核）、两个批归一化层、一个 ReLU 激活函数层。每个残差模块的输入和输出由一个旁路连接计算对应像素加和，以学习一个恒等映射，保证梯度在极深网络中的流动。在 N 个残差模块后，GN 采用一个步长为 2 的反卷积层将特征图的分辨率扩大。最后，在网络的尾部，通过一个包含 3 个 1×1 卷积核的反卷积层生成一个三通道的高分辨率车牌图像 P_h。

在一般的深度 SR 网络训练过程中，低分辨率图像通常通过高分辨率图像降采样和仿射变换得到，从而构成一组训练 LR-HR 样本对[98]。但是，对于本章中的车牌 SR 问题，大量低分辨率车牌图像可以监控视频中收集，问题在于为每一个低分辨率车牌 P_l 提供对一个的真实车牌 P_g。尽管可以采用人工收集的方式从相同车辆的车牌图像中选择最清晰、质量最好的车牌，但一方面很难准确确定清晰的标准，另一方面需要耗费大量的人力进行标注。深度 SR 网络的训练依赖大规模训练数据，低质量的训练数据将显著降低网络的性能，影响 GN 生成图像的效果。受到面向 CNN 的图像渲染[19] 启发，我们采用图像合成的方法为每一个 P_l 合成一张 P_g。因为车牌具有严格的生产标准，如固定的长宽比、颜色、字体、字符位置等，因此给定 P_l 中车牌字符的情况下，很容易根据生产标准采用图像渲染方法合成一张 P_g。在本章中，首先由人工对监控视频中收集的 P_l 标注车牌字符，然后根据车牌制作标准生成对应的 P_g，图4-10展示了一些 P_l 和 P_g 的实例。通过这些渲染的标准车牌图像，本章的车牌 SR 问题可以看作是学习一种车牌合成过程，将车牌校正与车牌 SR 合并为一个合成流程。此外，这些生成的空间一致的高分辨率车牌可用于进一步的多车牌超分辨率。

4.3.3　判别器网路

判别器网络（DN）用于区分一张车牌图像是由 GN 生成的 P_h，还是真实车牌图像 P_g，可以看做是一个二类分类问题。DN 的网络主干采用了一种类似 VGG 网络[111] 的结构，包含一个卷积层，后面连接 M 个卷积模块，每个卷积模块由一个卷积层、一个 LeakyReLU 激活函数层、一个批量归一化层和一个平均池化层组成，如

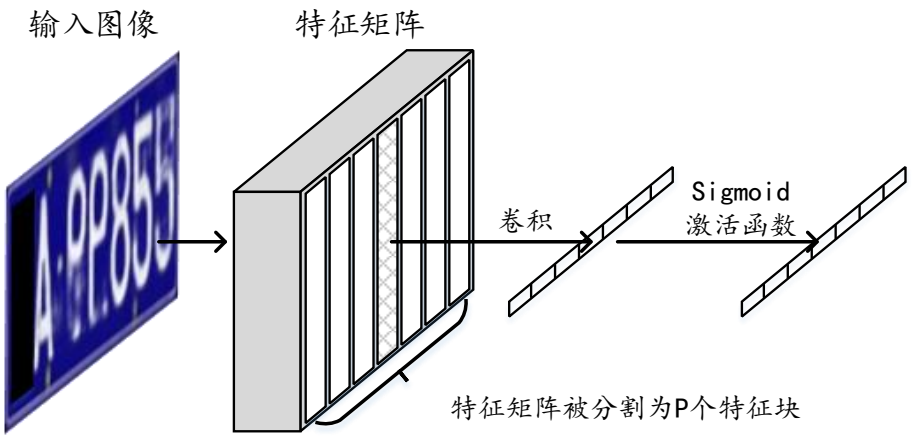


图 4-6　判别网络中空间分割层的结构

图4-5（d）所示。在本章的具体实现中采用了 3 个卷积模块，其中卷积层的的卷积核数分别为 64、128、256。LeakyReLU 层与批量归一化层的设置参考 SR-GAN[100] 中的设置。最后，为估计输入图像来自于真实图像 P_g 或 GN 生成的图像 P_h 的概率，网络尾部使用 Sigmoid 函数作为二分类问题的输出。

传统的判别器网络通常将输入图像作为一个整体，从全局图像区分其来自于 GN 还是真实数据。但是，车牌作为一类具有严格生产标准的对象，可以提供有价值的空间先验知识，如车牌长宽比、字符宽度、比例、间隔等。因此，为学习这类局部细节知识，我们在 DN 中加入了一个全新的空间分割层，如图4-5（d）所示。SSL 能够关注于每个字符位置的局部细节信息，防止相邻字符间的互相干扰。

图4-6展示了 SSL 的设计思想和细节：首先，输入车牌图像经过 DN 前向传播计算得到一个全局特征矩阵；在 SSL 层中，根据车牌中字符的位置与比例，特征矩阵被分割为 p 个兴趣区域（Region of Interest，RoI）。然后，每个 RoI 经过一个卷积层，该卷积层只包含一个大小为 $W_{RoI} \times H_{RoI}$ 卷积核，因此每个 RoI 得到一个标量值。最后，SSL 层的输出进入 Sigmoid 函数层得到每一个局部位置的概率值。通过 SSL 层的作用，DN 不仅能够考虑整张车牌图像的全局像素差异，而且能够捕捉每个字符的局部细节。在对抗训练过程中，可以将每个字符的误差反向传播到网络各层，从而进一步增强 GN 对结构较复杂字符的生成能力。

4.3.4　对抗损失函数

如 Goodfellow 等人提出的对抗训练原则[110]，DP-GAN 的目标是训练一个强大的 GN，使其能够生成与真实车牌图像 P_g 无法区分的 P_h。因此，如何为 GN 和 DN 设

计有效的对抗损失函数，是训练 DP-GAN 的关键。

对于 GN，给定一对输入训练样本 P_l 和 P_g，定义其对抗损失函数 L_G 为：

$$L_G(P_l,P_g)=L_{MSE}(P_l,P_g)+\lambda L_D \tag{4-1}$$

其中两部分分别来自于 GN 和 DN。L_{MSE} 为低分辨率车牌图像 P_l 通过 GN 生成的图像 $P_h=G(P_l)$ 与真实图像 P_g 的像素均方误差（Mean Squared Error，MSE），表示为：

$$L_{MSE}(P_l,P_g)=\frac{1}{r^2wh}\sum_{x=1}^{rw}\sum_{y=1}^{rh}(P_{g(x,y)}-G(P_l)_{(x,y)}) \tag{4-2}$$

其中 r 为超分辨率的比例系数，w 和 h 分别为 P_g 的像素宽度和高度。L_D 为 DN 中 Sigmoid 函数的输出，即 DN 估计输入图像来自于真实图像的概率。如文献[100] 中，L_D 可定义为：

$$L_D(P_l,P_g)=\sum_{n=1}^{N}-logD(G(P_l)) \tag{4-3}$$

其中 N 为一次训练迭代的批量数据大小。此外，由于 DN 中加入了 SSL 层，在具体实现中，L_D 不仅包含输入图像全局的真实性概率，还包括每个 RoI 区域的真实性概率。

在训练过程的每次迭代中，首先将一对样本 P_l 和 P_g 输入 GN 进行前向传播运算，并计算 L_{MSE}。然后，DN 以 $(P_h,0)$ 和 $(P_g,1)$ 为输入进行前向传播运算，分别计算两个样本的损失函数相加得到 L_D。在反向传播阶段，先使用 L_D 优化 DN，在使用 L_G 优化 GN。通过这种交替优化策略，GN 和 DN 以互相对抗训练方式共同提高性能。通过在 GN 优化过程中加入 DN 的损失 L_D，GN 不仅可以使 P_l 图像不断逼近真实图像 P_g，而且能够利用 DN 中学习到的车牌局部先验改善生成模型对局部字符的生成效果，减少相邻字符间的互相干扰。

4.4　基于车牌验证的精确车辆搜索

4.4.1　对偶神经网络结构

如图4-1所示，车牌字符如果能够被正确识别，则可以直接用车牌信息判断两个车辆是否是同一辆车。但是，在无约束城市交通场景中，摄像头受到光照、拍摄角度等影响，拍摄的车牌图像通常会发生变暗、模糊、形变等退化。此外，车牌识别系统通常包含车牌定位、校正、字符分割、识别等若干模块，每个模块受到上述因素

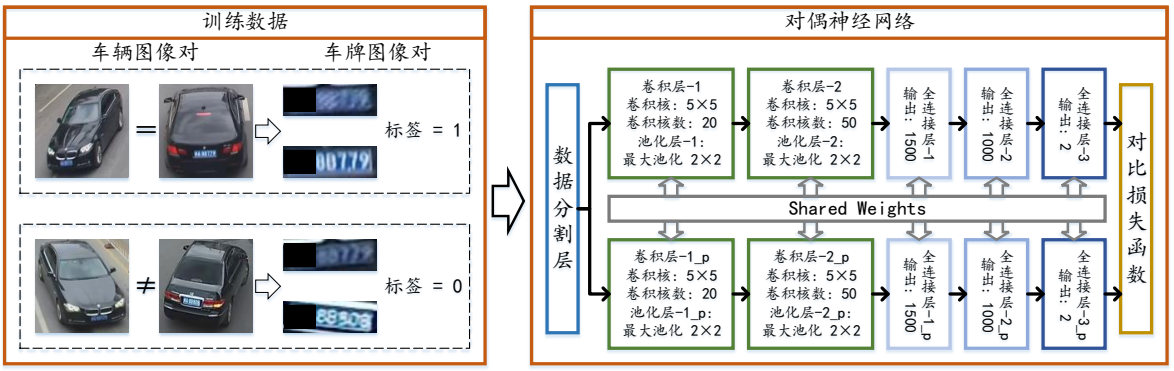


图 4-7　用于车牌匹配的对偶神经网络结构与训练

的影响都将造成车牌字符无法准确识别。因此，车牌识别技术不适合面向无约束城市监控的车辆搜索任务。为克服这一挑战，我们提出用车牌验证代替车牌识别，即通过判断两个车辆的车牌是否在视觉上相同来判断两个车辆是否相同。

SNN 最早由 Bromley 等人提出，应用于手写签名验证[84]。SNN 通常由两部分组成，一组卷积层学习图像的特征表示，一组全连接层学习一个映射函数。通过 SNN，可以从车牌图像对中提取具有区分力的视觉特征，并通过映射函数将特征投影到一个度量空间，使得相同对象的样本距离较近，不同对象的样本距离较远。因此，SNN 十分适合样本类别较多但每一类的样本较少的问题。本章中的车牌验证问题正具有这种特点，因此采用 SNN 进行车牌验证。如图4-7所示，SNN 的网络结构包含两个结构相同、共享权重的 CNN。每个 CNN 包含两个卷积层、两个最大池化层、三个全连接层。网络的具体参数，如卷积层卷积核的大小、卷积核数目、池化层窗口大小、全连接层隐单元数目等，参见图4-7中标注的参数。

4.4.2　网络训练

如图4-7所示，SNN 的训练样本由车牌图像对和对应的标签组成，如果一对车牌图像是相同车牌则标签为 1，反之标签为 0。在训练过程中，两张车牌分别通过两个 CNN 进行前向传播运算，两张图像的输出特征与标签共同输入对比损失函数层计算模型的代价。最后通过随机梯度下降算法[112] 优化 SNN 各层的网络权重。

具体来说，设 W 为 CNN 的网络权重，p_1 和 p_2 为一对输入车牌图像，则 p_1 和 p_2 通过 SNN 前向运算得到的特征可以表示为 $S_W(p_1)$ 和 $S_W(p_2)$。可以定义 p_1 和 p_2 的特征差别为一种能量函数：

$$E_W(p_1, p_2) = ||S_W(p_1) - S_W(p_2)|| \tag{4-4}$$

通过 $E_W(\cdot,\cdot)$，对比损失函数可以表示为：

$$L(W,(x_1,x_2,y)) = (1-y)\cdot max(m - E_W(x_1,x_2),0) + y\cdot E_W(x_1,x_2) \tag{4-5}$$

其中 (p_1,p_2,y) 为车牌图相对及标签组成的三元组，m 为一控制边界的正值超参数(在本章中设 $m=1$)。

4.4.3　精确车辆搜索

如图4-4所示，基于外观的相似车辆搜索方法首先搜索得到与查询车辆外观相似车辆的候选集合。对于候选车辆图像，采用通过大量车牌包围盒训练得到的车牌检测器，如 Faster R-CNN[62]，检测出车辆图像中的车牌区域，如果图像中检测到了车牌存在，则将车牌图像送入基于车牌匹配的车辆搜索模块中进行车牌验证。此时，将查询车辆的车牌图像 q_p 和候选车牌图像 g_p 输入 SNN，经过前向传播运算，以 SNN 中全连接层-2（FC2）的输入向量作为车牌的视觉特征表示 $\mathbf{r_p}$。然后，计算查询车牌与候选车牌的余弦距离，作为车牌相似度距离 $dist_p(q_p,g_p)$。设查询车辆 q 与被查询车辆 g 的外观相似度距离为 $dist_a(q,g)$，本章采用后融合（即距离融合）的方式计算两个车辆的相似度距离 $dist(q,g)$：

$$dist(q,g) = \alpha \times dist_a(q,g) + (1-\alpha)\times dist_p(q_p,g_p) \tag{4-6}$$

其中 α 为 $[0,1]$ 区间的平衡参数，本章设 $\alpha = 0.8$。最后，通过加权后的相似度距离，可以得到查询车辆与数据库车辆的精确搜索结果。

4.4.4　车辆搜索与车牌增强的结合

如本章第4.1节所述，车辆搜索方法可以根据查询车辆搜索到与其身份相同的车辆图像，根据搜索到的不同摄像头拍摄的多张车牌图像，可以采用多图像超分辨率方法为查询车辆生成一张高分辨率车牌图像，如图4-5所示。此时，生成的高质量车牌图像又可以反馈给渐进式车辆搜索框架中的车牌匹配模块，进一步提高车辆搜索的准确率。

如图4-5所示，在车辆搜索阶段，给定一张包含车牌的车辆图像，首先采用本文第2章提出的基于零空间学习与多级外观特征融合的相似车辆搜索方法（NuFACT），在数据库中匹配与查询车辆外观最相似的候选车辆。然后，对于候选车辆，采用本章第4.4节提出了基于车牌匹配的精确车辆搜索方法，进一步准确查找目标车辆。最后，

引入一种时空关系模型[76]对搜索结果进行重排序，去除时空关系上不合理的情况，如与查询车辆时间距离很近但空间距离很远的候选车辆。通过上述由粗到精的渐进式搜索方法，可以检索到与查询车辆身份相同的车辆图像集合，通过车牌定位方法即可得到与查询车辆车牌相同的车牌图像集合。但是，这些搜索到的车牌图像包括了各种各样的视角变化、仿射变换、光照变化，甚至还可能包括错误的车牌，即与查询车牌不同的车牌图像，这些极端情况将严重影响多图像超分辨率（MSR）的效果。

因此，在车牌增强阶段，首先将上述搜索得到的车牌图像分别输入训练好的DP-GAN，通过前向传播运算，将每张车牌图像恢复为一张空间结构一致的高分辨率车牌图像。然后，我们提出一种多图像超分辨率生成对抗网络（MSR-GAN），将DP-GAN生成的多张车牌图像合成为一张高质量图像。MSR-GAN采用了与DP-GAN中相同结构的GN、DN网络，主要区别在于在MSR-GAN的输入数据是由N个车牌图像按通道叠加组成的数据张量。在具体实现中，MSR-GAN采用车辆搜索结果中相似度最高的前9张车辆的车牌图像组成输入张量，训练过程采用与DP-GAN相似的策略和对抗损失函数。最后，MSR-GAN为输入的查询车辆生成一张高分辨率车牌图像，以后后续车牌识别或车辆搜索使用。

4.5 实验结果与分析

本章通过实验首先将基于域先验生成对抗网络的车牌超分辨率方法与其它图像超分辨率方法的比较，验证该方法在车牌超分辨率的优异性能。然后，评估基于车牌匹配的精确车辆搜索方法，并将该方法与基于外观的相似车辆搜索方法NuFACT结合，验证由粗到精的渐进式车辆搜索框架。最后，通过将MSR-GAN超分辨率方法与车辆搜索框架结合，验证两者结合所带来的性能提升。

4.5.1 数据集

对于车牌超分辨率方法的评估，我们以VeRi数据集为基础，收集并标注了一个车牌超分辨率数据集“VeRi_Plate”。标注过程中，人工标注员首先从VeRi数据集的776个车辆中挑选出包含车牌的图像，并手工截取其中的车牌图像。然后，为保证标注质量及生成标准车牌图像，仅保留至少有一张车牌图像能够识别出字符的车辆，共计得到746个车辆、24349张车牌图像。下一步，基于车牌生产标准和每个车辆标注的车牌字符，为每个车辆渲染一张标准车牌图像P_g，部分实例如图4-10中

“Groundtruth”一列所示。[①] 最后，按照 VeRi 数据集的划分方式，VeRi_Plate 数据集被划分为包含 594 个车辆、19524 个车牌图像的训练集，及包含 152 个车辆、4825 张车牌的测试集。对于测试集中的每个车辆，选择一张低分辨率图像作为查询图像，得到一个包含 152 张图像的查询集，部分实例如图4-10中“Query”列所示。

然后，我们采用第2章提出的大规模车辆搜索数据集 VeRi 评估车辆搜索方法。如第2.5.1.1节所述，VeRi 数据集不仅包含大量从真实交通监控视频系统手机的车辆图像数据，而且标注了车辆在不同摄像头的重现信息，为车辆搜索提供了保证。此外，VeRi 数据集还标注了每张车辆图像的车牌区域、车牌字符等信息，十分适合验证本章所提出的基于车牌验证的精确车辆搜索方法和由粗到精的渐进式车辆搜索框架。

4.5.2　实验设置

对于车牌超分辨率方法，传统的评价标准主要是像素级的图像差异，如峰值信噪比与像素均方误差。但是，如前文所述，与像素级差异评价方式不同，车牌超分辨率的目标是使生成的车牌能够被人工或计算机正确识别。因此，我们在实验中采用了两种评价标准：人工识别准确率和自动识别软件准确率。对于人工识别，12 名志愿者被邀请参与到车牌识别任务。对于自动识别，我们采用了一个开源软件“EasyPR”[②]。

对于车辆搜索方法部分，本节采用与前文相同的数据划分方式，即使用 VeRi 数据集中的 576 个车辆作为训练集，剩余 200 个车辆作为测试集，对于测试集选取了 1678 张车辆图像作为查询集。为综合评估方法的准确率，本节同样采用 mAP、HIT@1 和 HIT@5 作为评价标准，其中 mAP 定义如第2.5.2节的公式2-9和2-10所示。

4.5.3　方法对比

4.5.3.1　车牌超分辨率方法评估

本节对比了七种不同方法对 VeRi_Plate 测试集的多图超分辨率效果。对于测试集的 152 个车辆，首先使用查询图像搜索得到相似车辆集合，取前 9 个车牌图像作为以下方法的输入进行超分辨率操作。方法实现细节如下所述：

- 方法一，**Baseline**。该方法不对低分辨率车牌图像进行超分辨率，直接进行车牌识别，作为其它方法的测试基准。

① 由于隐私原因，本章中的车牌图像隐去了前两位字符。
② “EasyPR”，https://github.com/liuruoze/EasyPR

- 方法二，**VDSR**[98]。该方法采用了一种包含 20 个卷积层的单图超分辨率（Single image Super Resolution，SSR）神经网络，并在公开数据集上取得了 SSR 的最优结果。本实验中先使用 VeRi_Plate 数据集的训练数据对 VDSR 进行重新训练，然后对 9 张车牌图像分别处理，最后采用后融合的方法将 SR 后的 9 张车牌合并为 1 张车牌。

- 方法三，**IR + DDL - MSR**。该方法首先采用 Yang 等人[113] 提出的图像配准（Image Registration，IR）方法，将 9 张输入车牌图像进行校正。然后使用 Liao 等人[114] 提出的基于深度草稿集成学习（Deep Draft-ensemble Learning，DDL）的视频超分辨率方法，将配准后的 9 张车牌图像恢复为一张高分辨率车牌。

- 方法四，**IR + MSR-GAN**。该方法首先采用 Yang 等人提出的 IR 方法对 9 张低分辨率车牌进行校正。然后使用校正后的车牌训练本章提出的 MSR-GAN。最后使用训练好的 MSR-GAN 生成一张高分辨率车牌图像。

- 方法五，**SR-MSR-GAN[100]**。该方法使用 Ledig 等人提出的 SR-GAN[100] 网络，使用 EasyPR 为训练集的低分辨率车牌 P_l 选择标签图像 P_g，而不是采用本章提出的车牌渲染方法生成标准车牌图像作为标签。然后使用训练好的 SR-GAN 对 9 张输入图像做 SR 处理，最后采用后融合方法得到一张结果图像。

- 方法六，**SRS-MSR-GAN**。该方法与上述方法采用相同的 SR-GAN[100] 网络结构，唯一不同之处是采用本章提出的车牌渲染方法生成标准车牌图像作为标签 P_g 训练 SR-GAN。

- 方法七，**DP-MSR-GAN**。该方法即本章提出的面向多图超分辨率（MSR）的域先验生成对抗网络（DP-GAN）。

采用 EasyPR 对上述方法生成的图像进行车牌识别的准确率如图4-8所示。观察结果对比可以得出如下结论：

- 首先，相比基准测试，所有超分辨率方法都能提高车牌识别的准确率。

- 此外，通过 IR + DDL-MSR 和 IR + MSR-GAN 的结果可以发现，尽管采用 IR 方法将车牌校正为角度一致的图像，但后续的 MSR 方法仍然无法恢复较好的图像质量。这是因为不同摄像头拍摄的低分辨率图像具有很大的差异，已有的 MSR 方法无法有效挖掘和融合这些低分辨率图像中的互补信息。不同的是，

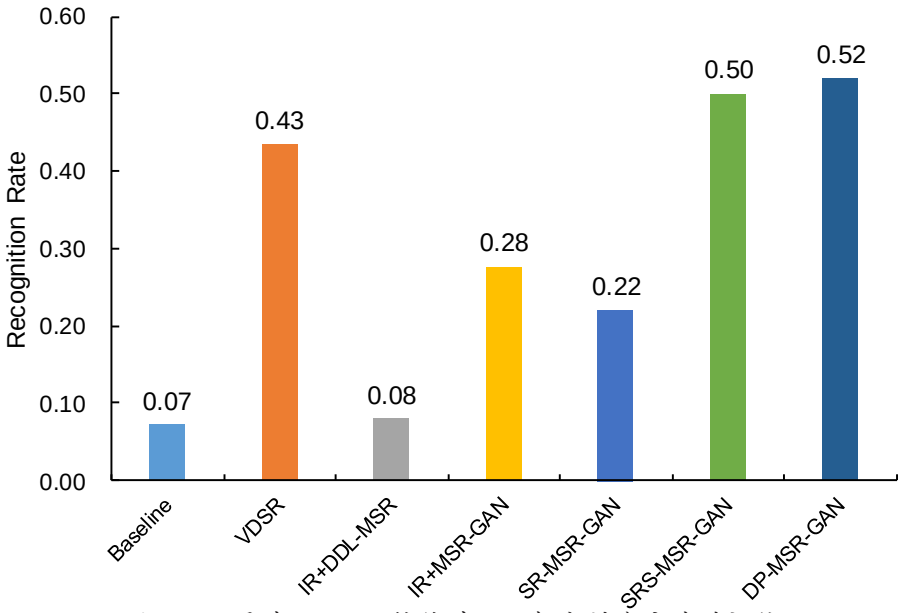


图 4-8　通过 EasyPR 软件对不同超分辨率方法的评估

通过本章提出的 DP-GAN，输入的多张低质车牌被恢复为空间一致的高分辨率车牌。因此，DP-GAN 可以有效学习多张车牌的共同信息并挖掘其中的互补信息，从而改进图像质量。

- 通过对比 SR-MSR-GAN 和 SRS-MSR-GAN，可以证明通过大量渲染数据的训练，GAN 可以学习其中有用的领域先验知识，这些知识保证了生成车牌的局部细节。VDSR 与 SRS-MSR-GAN 的对比表明 DN 与对抗损失在 GAN 训练过程中的重要作用。

- 最后，本章提出的 DP-MSR-GAN 取得了最好的结果，不仅表明基于域先验的车牌渲染为训练 GAN 提供了有效的数据支持，而且验证了加入 SSL 层的 DN 能够有效学习数据中的先验知识，保证 GN 能够捕捉车牌的局部细节，从而生成高质量的车牌图像。

4.5.3.2　车辆搜索对车牌超分辨率的作用分析

为验证渐进式车辆搜索框架对多图车牌超分辨率的作用，本节对比了使用与不使用车辆搜索的车牌超分辨率方法。对于使用了车辆搜索的方法，实验中选取前 8 张最相似车辆的车牌作为输入。以下方法都使用本章提出的 DP-GAN 作为 SR 方法：

- 方法一，**Query**。该方法直接使用 VeRi_Plate 数据集中的 152 张查询车牌图像进行车牌识别，不包含任何车辆搜索和超分辨率处理。

- 方法二，**Search-Late**。该方法先使用查询图像搜索得到最相似的 8 张车牌图像，然后使用 EasyPR 对查询图像及 8 张搜索结果进行车牌识别，最后对 9 张车牌的识别结果进行后融合。

- 方法三，**Query-SR**。该方法首先使用 DP-GAN 对查询车牌图像进行超分辨率处理，然后使用 EasyPR 识别生成的高分辨率车牌。

- 方法四，**Search-SR-AVG**。该方法首先使用 DP-GAN 对查询车牌图像及 8 张搜索结果车牌分别进行超分辨率操作，然后使用平均池化方法将 9 张高分辨率车牌融合为一张车牌，最后使用 EasyPR 识别其中的字符。

- 方法五，**Search-SR-Late**。该方法首先使用 DP-GAN 对查询车牌图像及 8 张搜索结果车牌分别进行超分辨率操作，然后使用 EasyPR 识别 9 张车牌中的字符，最后对 9 张车牌的识别结果进行后融合，得到最终的识别结果。

- 方法六，**DP-MSR-GAN**。该方法首先使用 DP-GAN 对查询车牌图像及 8 张搜索结果车牌分别进行超分辨率操作，然后将 9 张高分辨率车牌输入提出的 MSR-GAN 中生成一张高质量车牌，最后使用 EasyPR 识别其中的字符。

上述六种方法的识别准确率如图4-9所示，通过对比可以得出如下结论：

- 首先，Query 与 Search-Late 的对比表明，无论是否采用车牌超分辨率方法，渐进式车辆搜索都能够提高车牌识别的准确率。因此，多张车牌的互补信息对于车牌超分辨率具有重要作用。

- 此外，通过 Query 与 Query-SR、Search-Late 与 Search-SR-Late 的对比，可以发现本章提出的 DP-GAN 能够显著提高车牌识别的准确率。因此 DP-GAN 可以生成真正有利于车牌识别任务的高质量车牌。

- 相比较 Search-SR-AVG 及 Search-SR-Late 方法，本章提出的 DP-MSR-GAN 能够生成最好的车牌图像，取得了最高的识别准确率。这说明 DP-MSR-GAN 能够更有效地捕捉多张车牌的共同部分，挖掘其中的互补信息和先验知识，最终生成具有空间一致性的高分辨率车牌。

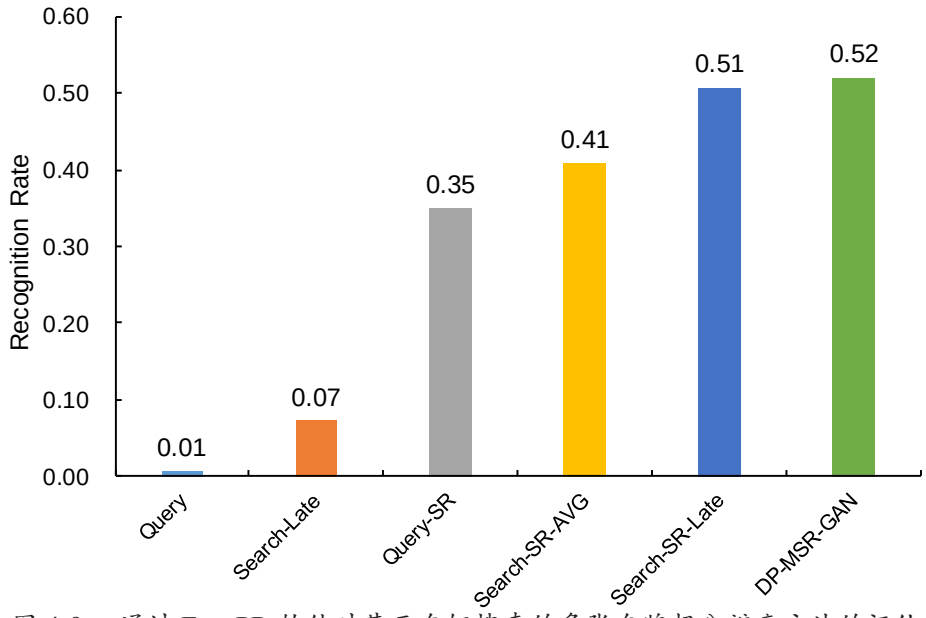


图 4-9　通过 EasyPR 软件对基于车辆搜索的多张车牌超分辨率方法的评估

4.5.3.3　车牌超分辨率的定性分析

图4-10展示了不同车牌超分辨率方法生成的车牌实例。对比两种仅使用查询车牌 Query 的方法，即 Q-VDSR 和 Q-DP-GAN，可以看出 Q-VDSR 方法生成了更多奇异形状的字符。此外，与 IR + MSR-GAN 相比，DP-GAN 能够帮助 DP-MSR-GAN 取得更好的生成效果。对比结果证明，DP-GAN 中的 GN 能够生成符合车牌全局和局部标准的车牌图像。通过大量渲染车牌参与训练，DP-GAN 能够生成具有空间一致性的高分辨率车牌。更进一步，MSR-GAN 可以充分利用多张车牌的互补信息，因此基于车辆搜索和域先验生成对抗网络的车牌增强方法 DP-MSR-GAN 取得了最好生成效果。图4-10的最后两行列出了所有方法都失败的案例，主要原因在于车牌中的“0”与“Q”、“8”与“B”十分相似，这种相似字符引起的生成误差也存在于其它字符。根据统计数据，字符“P”、“T”、“A”、“2”、“J”在车牌识别过程中准确率较高，而“Y”、“D”、“G”、“B”、“Q”的准确率最低。可以看出字符的形状相似性对于车牌超分辨率和识别具有很大影响。

4.5.3.4　基于车牌验证的车辆精确搜索方法评估

本节对比了基于 SNN 的车牌验证与基于传统图像纹理特征的方法。通过两种方法提取的车牌特征，计算查询车辆与数据库车辆的车牌相似性，然后与基于外观相

图 4-10　不同车牌超分辨率方法的生成结果对比

似性的 NuFACT 方法进行结果后融合。通过比较最终车辆搜索的准确率，比较两种方法对渐进式车辆搜索的作用。两种方法细节如下：

- 方法一，**NuFACT + Plate-SIFT**。该方法采用经典的局部描述子 SIFT[67] 作为基本图像特征，提取车牌图像中的 SIFT 描述子。然后，将车牌图像中提取的 SIFT 描述子采用词袋模型量化为长度 1000 的特征向量。词袋模型的码本在 VeRi_Plate 数据集的训练集上训练得到。最后，通过计算查询车牌 P_q 与数据库车牌 P_g 特征间的余弦距离，作为两个车辆的车牌相似性距离。

- 方法二，**NuFACT + Plate-SNN**。该方法为本章提出的基于 SNN 车牌验证方法。在训练阶段，首先从 7647 个车牌训练集中随机配成 100000 个样本对，其中正负样本对比例为 1:1。在测试阶段，采用训练好的 SNN 模型对车牌图像进行前向传播运算，取 FC2 层长度 1000 的输出作为车牌特征。最后，同样以查询车

牌 P_q 与数据库车牌 P_g 特征间的余弦距离作为两个车辆的车牌相似性距离。

表 4-1　不同方法在 VeRi 数据集上的车辆搜索结果比较

方法	mAP	HIT@1	HIT@5
NuFACT	48.47	76.76	91.42
NuFACT + Plate-SIFT	48.55	76.88	91.48
NuFACT + Plate-SNN	**50.87**	**81.11**	**92.79**

表4-1列出了不同方法在 VeRi 数据集上的搜索准确率。首先，增加了车牌信息的搜索结果优于仅使用外观特征的搜索模型。此外，通过对比两种车牌特征，SNN 提取的车牌特征能够更有效表示车牌视觉特征。因此，基于 SNN 的车牌验证方法在基于外观相似性的方法基础上提高了 2.32% 的 mAP。这也证明，基于深度学习的 SNN 可以从大量训练样本中学习到对环境变化鲁棒的视觉特征。但是，由于在无约束环境中，车牌图像由于光照、视角、运动等因素发生了严重的退化，需要对车牌进行超分辨率以进一步提高车辆搜索的准确性。

4.5.3.5　车牌增强对车辆搜索的作用分析

在基于多车图的车牌超分辨率中，车辆搜索是获得多张不同摄像头拍摄车牌的先决条件。反过来，本章提出的 DP-GAN 和 DP-MSR-GAN 也可以为查询车辆的车牌或数据库车辆的车牌提供超分辨率处理。因此，在本节的车辆搜索实验中，首先使用训练好的 DP-GAN 对 VeRi 数据集中的所有车牌进行超分辨率。对于车牌匹配中的 SNN，采用超分辨率后的训练车牌重新训练模型。在基于车牌匹配的精确车辆搜索模块中，使用超分辨率后的查询车牌及数据库车牌进行相似性匹配。表4-2对比了本文第2章中基于外观的车辆搜索方法 FACT 和 NuFACT、本章中基于车牌匹配方法与基于外观的方法结合的渐进式车辆搜索框架 PROVID、以及将 DP-GAN 加入渐进式车辆搜索的 PROVID + DP-GAN。从实验结果中可以发现，将 DP-GAN 加入基于车牌匹配的精确搜索后，车辆搜索的整体性能大幅提高（mAP 提高了 7.18%），再次证明了本章提出的车牌增强方法的有效性。

4.6　本章小结

本章关注于渐进式车辆搜索框架中基于车牌的精确车辆搜索部分。由于真实交通监控中的拍摄视角、距离、光照变化，车牌图像通常包含模糊、倾斜、变暗等退化，

表 4-2　不同方法在 VeRi 数据集上的车辆搜索结果比较

方法	mAP	HIT@1	HIT@5
FACT	18.75	52.21	72.88
NuFACT	48.47	76.76	91.42
PROVID	53.29	81.76	94.70
PROVID + DP-GAN	**60.47**	**85.52**	**95.11**

导致车牌匹配或识别等任务准确率大幅下降。因此，本章提出了一种基于域先验的生成对抗网络（DP-GAN），将监控图像中低质量车牌图像还原为清晰车牌图像，从而提高车牌匹配及车辆搜索的准确性。此外，针对无约束监控场景中车牌字符难以准确识别的挑战，本章提出了一种基于 SNN 的车牌匹配方法，该方法摒弃了车牌识别的路线，通过验证来自两张车辆图像的车牌是否是相同实现精确的车辆搜索。在我们提出的 VeRi 数据集上，本章不仅证明了车牌信息对车辆精确搜索的作用，而且验证了车辆搜索对基于多图的车牌超分辨率的重要性。最后，通过将 DP-GAN 超分辨率后的车牌反馈给车辆搜索框架，进一步提高了渐进式车辆搜索的性能。

第五章 多模数据融合的渐进式车辆搜索系统

5.1 应用背景

为验证本文研究成果的性能，在第2至4章提出的算法基础上，我们设计并实现了面向城市视频监控网络的渐进式车辆搜索原型系统。

为保证城市交通的安全运行、提高交通管理的效率，大量交通监控摄像头被广泛部署在城市道路的出入口、十字路口、高速公路收费站等关键位置。与此同时，为了实现交通监控系统的自动化、智能化水平，车辆检测、车辆跟踪、车牌识别、违章抓拍等技术已被部署到现有交通监控系统，甚至嵌入到监控摄像头终端内。相比上述技术，车辆搜索，即给定一个查询车辆的图像或视频，在大规模车辆数据库中搜索身份相同的车辆，返回目标车辆出现的时间、地点信息，仍处于初步探索阶段。目前的车辆搜索，仅限于卡口摄像头进行车牌识别后，按照输入的车牌号进行字符串匹配搜索。相对于卡口位置的高清摄像头，城市中绝大多数摄像头部署于无约束道路，摄像头拍摄的角度、距离、分辨率、环境具有很大的不确定性，基于车牌识别的搜索系统则无法满足实际需求。因此，充分利用大规模城市监控系统的图像、视频等视觉信息，挖掘时空信息、摄像头网络拓扑信息等情景信息，设计并开发面向大规模城市监控的车辆搜索系统具有十分重要的应用价值。

车辆搜索系统可以应用于很多领域，典型的应用场景如下:

(1) 嫌疑车辆搜索

作为车辆搜索的核心功能，我们提出的渐进式搜索框架可以为公安和交管部门提供嫌疑车辆搜索服务。例如在犯罪调查时，办案人员可以以案件现场摄像头抓拍的嫌疑车辆图像为输入，快速地得到该车辆在整个城市中出现的时间、地点信息，从而发现出嫌疑车辆行驶的轨迹，进而推断其可能将在哪里出现。这可以大大减少人工查看视频、排查嫌疑车辆消耗的时间，提高工作人员的办事效率。

此外，虽然我们提出的渐进式车辆搜索系统主要面向无约束监控场景，但该系统也可以与有约束环境下（如停车场、高速出入口等）的卡口摄像头系统结合。这样，通过卡口摄像机下精确的车牌信息与无约束环境下基于外观、车牌匹配的搜索结果进行联合分析，能够实现更准确的车辆搜索。例如，车辆搜索系统可以与停车

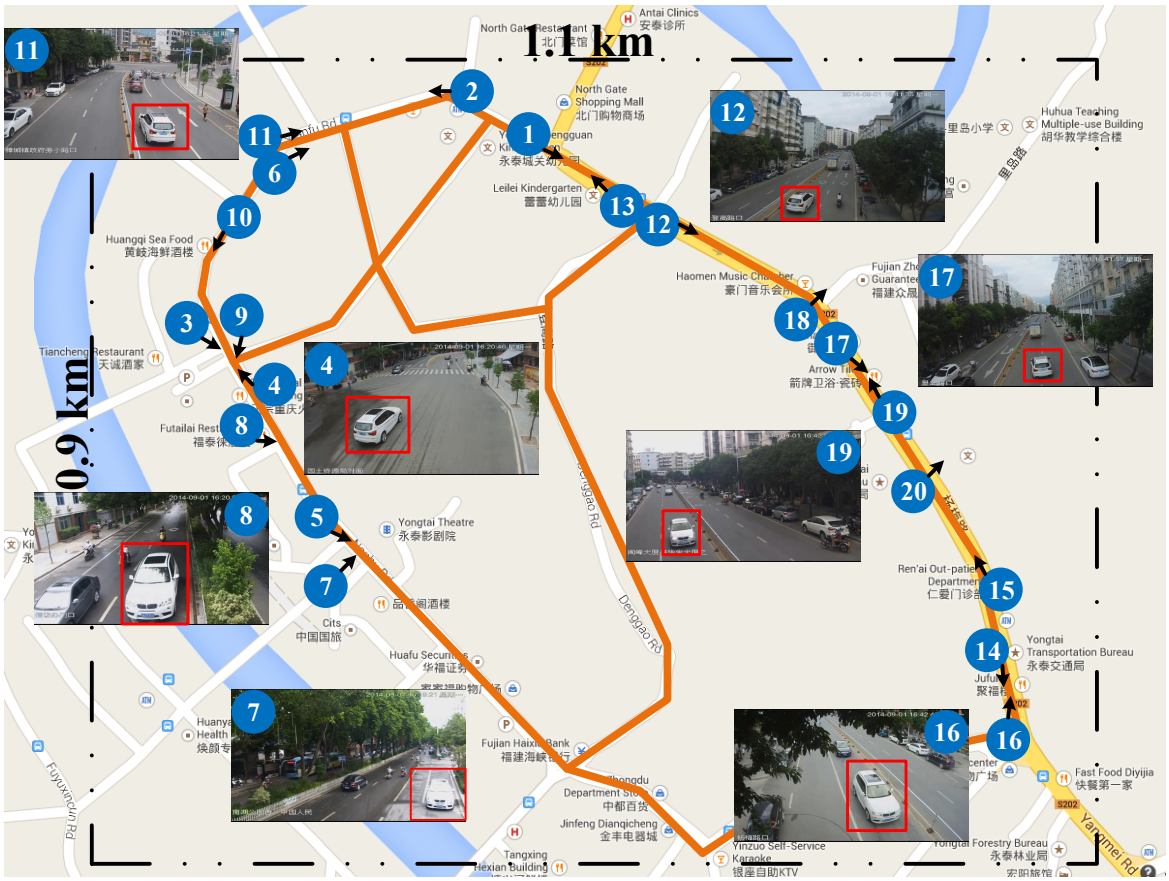


图 5-1　渐进式车辆搜索系统应用：嫌疑车辆搜索

场车牌识别系统结合，并与公安交管部门的车辆注册信息系统连接。当公安人员获取到一张嫌疑车辆的图像后，可以首先使用车辆搜索系统根据图像快速定位其出现过的停车场。然后，通过停车场车牌识别系统获取的车牌号信息，就可以查询到该车辆的所有人、注册时间、注册地等信息。借助这些信息，公安人员能够更准确、高效地进行案件侦查工作。总之，我们提出的渐进式车辆搜索系统可以作为一个面向城市监控网络的车辆搜索引擎，为交通管理、公共安全等领域提供快速、准确的车辆搜索服务，如图5-1所示。

（2）跨摄像头车辆跟踪

我们设计的车辆搜索系统也可以应用于跨多摄像头的目标车辆跟踪。例如，公安人员要追踪一辆肇事逃逸车辆在城市中的位置，首先可以在肇事发生的摄像头获取嫌疑车辆的图像。然后，可以通过车辆搜索系统在该摄像头临近的摄像头由近及远地搜索目标车辆。公安人员可以根据系统的搜索结果快速发现临近摄像头中的目标车辆，并指定该图像为新的查询进一步搜索。这样，车辆搜索系统便可以辅助公

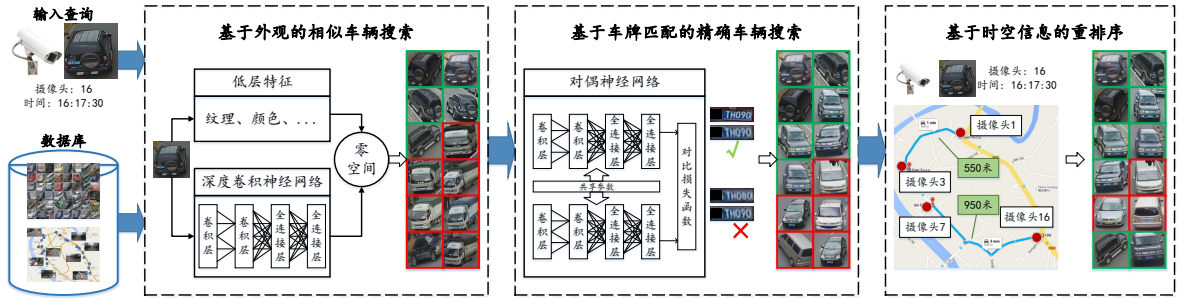


图 5-2　多模数据融合的渐进式车辆搜索框架

安人员跟踪嫌疑车辆的位置和行驶路线。

此外，车辆搜索系统也可应用于另一种跨摄像头跟踪应用，即实时汽车赛事转播。常见的汽车赛事，如达喀尔拉力赛和 F1 方程式锦标赛，通常采用多摄像车或多摄像机实时转播比赛情况。在转播某一辆赛车的画面时，通常采用人工切换的方法切换不同摄像头，从而实现跨摄像头跟踪转播。如采用我们研制的车辆搜索系统，工作人员只需指定需要转播的赛车，系统即可在邻接的摄像头根据车辆外观、号码、时间信息等追踪目标赛车。总之，车辆搜索系统可以帮助用户跨多摄像头定位并跟踪目标车辆，对于嫌疑车辆追踪或汽车赛事转播具有重要的应用价值。

5.2　多模数据融合的渐进式车辆搜索框架

在真实实践中，人们通常采用渐进式的搜索方式在大规模监控视频中寻找目标[13]。例如，公安人员想在城市监控视频中寻找一部嫌疑车辆的踪迹。通常在获得一个摄像头内的目标图像后，公安人员会根据车辆的外形、颜色、车型进行外观相似车辆的粗略筛选，从而缩小搜索范围。然后在相似车辆中进一步观察外观细节，最后通过车牌信息确定嫌疑车辆。此外，公安人员通常会按照从起始摄像头由近及远地搜索，以避免遍历所有监控视频，从而提高搜索的效率。

受到上述过程的启发，我们设计了一种融合多级特征和多模态数据的渐进式车辆搜索框架，该框架的主要思想包括两个方面：第一是特征域内由粗到精的搜索，即先使用车辆外观特征进行粗筛选，再使用车牌信息进行精确匹配；第二是时空域内由近及远的搜索，即在特征匹配时考虑车辆的时间、空间信息提高搜索的效率。渐进式搜索的概念性框架如图5-2所示，包含基于外观的相似车辆搜索、基于车牌匹配的精确车辆搜索、基于时空信息的重排序。

基于外观的相似车辆搜索以本文第2、3章的工作为基础。对于车辆图像数据，首

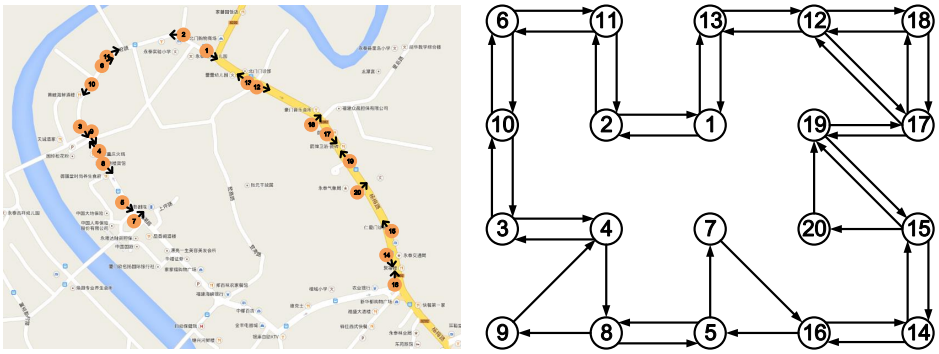


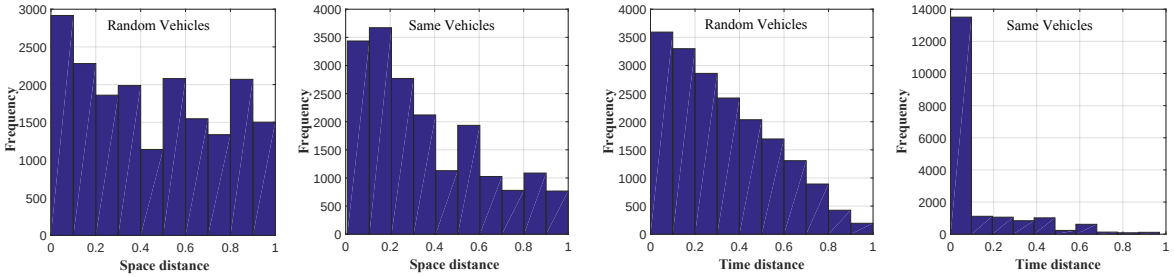
图 5-3　视频监控网络及对应的摄像头邻接图示例

先对车辆图像提取纹理、颜色、语义属性等多级外观特征。然后，使用零空间度量学习算法将多级特征融合为统一的车辆外观表示。对于车辆视频数据，采用跨视角注意力网络提取具有区分力和视角鲁棒性的车辆外观特征。通过外观特征，能够在数据库中搜索得出与查询车辆外观相似的车辆，可以相似度最大的部分图像作为候选结果进一步匹配。

基于车牌匹配的相似车辆搜索以本文第4章的工作为基础。对于外观相似车辆图像中的车牌区域，首先使用 DP-GAN 网络进行超分辨率操作，得到清晰的车牌图像。然后使用预训练的对偶神经网络提取车牌特征。根据车牌特征，可以在外观相似车辆中进一步匹配出车牌匹配度最高的候选车辆。

基于时空信息的重排序方法通过一种时空相似度模型估计查询车辆与数据库中车辆的时空相似度。视频监控网络中的时空情景信息，如图像记录的时间戳、摄像头的拓扑、距离等，已在多摄像机系统中广泛使用[20,21,115]。首先，我们根据摄像头的邻接关系构建了一个摄像头邻接图，$G=<N,E,W>$，如图5-3所示。摄像头邻接图由节点集合 $N=\{n_1,...,n_C\}$、边集合 $E=\{e_{i,j}\}$、权重集合 $W=\{w_{i,j}\}$ 构成。节点集合 N 为视频监控网络中摄像头的集合，每个节点保存了该摄像头的编号、地理坐标、配置信息等。边集合 E 记录了摄像头之间的邻接关系，如果两摄像头之间有一条直接道路连接，则这两摄像头之间存在一条边。权重集合 W 包含了边集合中每条边的长度，即两邻接摄像头间的空间距离。

为挖掘时空信息，我们在真实监控场景的数据中选择了 20000 对相同车辆的出现记录（摄像头编号与时间戳）和 20000 对随机选取的不同车辆的出现记录。然后计算每对样本出现的时间距离和距离空间，并做统计分析。图5-4中展示了将时间距离与空间距离进行归一化后的统计直方图，可以看出相同车辆是时空距离具有显著的统计规律。因此，我们采用了一种基于多层感知机（Multi-Layer Perceptron）的时



(a) 空间距离直方图.　　(b) 时间距离直方图

图 5-4　VeRi 数据集的时空信息统计

空相似度模型（Spatio-Temporal Similarity Model，STSM）。首先，给定查询车辆 q 与被查询车辆 g，我们可以计算二者的空间距离 $D_s(q,g)$ 和时间距离 $D_t(q,g)$：

$$\begin{aligned} D_s(q,g) &= |L(q) - L(g)| \\ D_t(q,g) &= |T(q) - T(g)| \end{aligned} \tag{5-1}$$

其中，$L(\cdot)$ 为车辆所在的物理位置，$T(\cdot)$ 为车辆被摄像头拍摄的时间戳。然后，我们采用一个包含两个全连接层的 MLP，$\mathscr{F}(\cdot)$，对时空相似度进行建模。两个全连接层的输入输出维度分别为 $(2,64)$ 和 $(64,1)$，激活函数分别为 ReLU 和 Sigmoid。因此，q 与 g 的时空相似度 $S_{st}(q,g)$ 可表示为：

$$S_{st}(q,g) = \mathscr{F}([D_s(q,g), D_t(q,g)]) \tag{5-2}$$

其中，$[\cdot,\cdot]$ 表示将两个元素连接为一个向量。

最后，为有效融合前文得到的车辆外观相似度 S_a、车牌相似度 S_p、时空相似度 S_{st}，我们采用了一个单层感知机 $\mathscr{G}(\cdot)$ 学习相似度融合参数。因此，车辆 q 与 g 的相似度 S 可以表示为：

$$S(q,g) = \mathscr{G}([S_a(q,g), S_p(q,g), S_{st}(q,g)]) \tag{5-3}$$

车辆搜索的结果将按照查询车辆与数据库车辆的相似度进行排序。

5.3　多模数据融合的渐进式车辆搜索原型系统

如图5-5所示，整个原型系统分为离线车辆数据收集子系统和在线车辆搜索子系统两大部分。

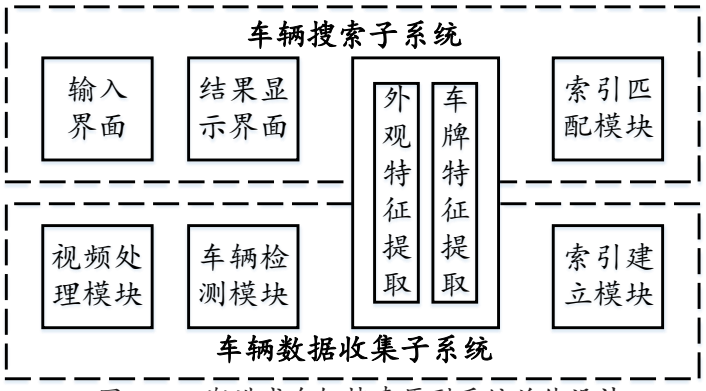


图 5-5　渐进式车辆搜索原型系统总体设计

5.3.1　车辆数据收集子系统

车辆数据收集子系统主要用于实时或离线接收监控摄像头传来的数据，从视频帧中检测车辆并存储车辆图像，然后提取车辆的外观特征、车牌特征、时间戳和摄像头编号，最后根据上述特征建立索引，存储到数据库中。该子系统主要包括如下模块：

（1）视频接收与预处理模块

数据接收与预处理主要功能包括接收摄像头视频，对原始视频进行转码、压缩。然后根据设定的采样率将视频采样为独立的视频帧图像。该模块由专门的数据预处理服务器运行，将采样后的视频帧输入车辆检测模块。

具体实现中，我们使用 FFmpeg 工具包将原始监控视频统一采用 H.264 编码压缩转码为 MP4 视频文件。然后，按照每 5 帧采样 1 帧的采样率提取视频帧图像，并使用 OpenCV 将视频帧图像重置为 1067×600 的图像。最后，记录每一帧相对于视频起始时间的偏移量、摄像头编号等元数据，将图像帧与元数据存储于数据磁盘阵列。

（2）车辆检测模块

车辆检测模块以整幅视频帧为输入，使用训练好的 Faster R-CNN 车辆检测器[62]定位车辆在视频帧中的包围盒坐标。将车辆包围盒截取并保存为车辆图像，与时间戳、摄像头编号组成三元组输入特征提取模块。车辆检测模块由配有图形处理器（Graphics Processing Unit，GPU）的车辆检测服务器运行。

具体实现时，车辆检测模块在处理前需预先训练 Faster R-CNN 车辆检测器[62]。我们预先标注了约 40000 组训练样本，包括视频图像帧、其中车辆的包围盒坐标即类别。使用 Caffe 深度学习工具包实现并训练 Faster R-CNN 网络模型，训练迭代 50000 次后保存网络模型。

在检测阶段，车辆检测模块从数据磁盘阵列读入视频帧图像，通过 Caffe 的借口调用训练好的 Faster R-CNN 模型。经过 Faster R-CNN 前行传播运算，输出视频图像帧中车辆的包围盒坐标，即左上角坐标 (x_{tl},y_{tl}) 及右下角坐标 $(x_{br}y_{br})$，并根据包围盒坐标截取车辆图像。最后，记录每一个车辆图像在当前帧中的序号、坐标、时间戳、摄像头编号等元数据，将车辆图像与元数据存储于数据磁盘阵列。

（3）多级特征提取模块

特征提取模块是系统的核心模块，以“车辆图像-时间戳-摄像头编号”三元组为输入，基于本文提出的算法提取车辆的外观、车牌特征。该模块分为外观特征提取、车牌特征提取两个子模块。其中外观特征提取子模块基于第2章提出的算法提取纹理、颜色、语义属性特征，并融合为外观特征向量。车牌特征提取子模块首先使用训练好的 Faster R-CNN 车牌检测器[62] 定位图像中的车牌区域，然后采用第4提出 DP-GAN 对车牌图像超分辨率，再使用训练好的 SNN 提取车牌特征。特征提取模块由配有 GPU 的特征提取服务器运行。

具体实现时，外观特征提取子模块基于 NuFACT 算法提取纹理、颜色、语义属性特征。其中，纹理特征采用 OpenCV 中的 SIFT 算子提取接口和词袋模型实现，纹理词袋模型训练和特征量化如第2.3.1节所述。颜色特征采用 Matlab 实现的 CN 算子提取接口[69] 和词袋模型实现，颜色词袋模型训练和特征量化如第2.3.2节所述。语义属性特征中使用的 GoogLeNet 模型[70] 使用 Caffe 实现和训练，网络训练和特征提取如第2.3.3节所述。上述三种特征使用零空间度量学习算法融合，其中映射矩阵的训练由 Matlab 代码实现，特征映射和融合过程如第2.4.2节所述。

车牌特征提取子模块需要预先训练 Faster R-CNN 车牌检测器[62]、车牌超分辨率模型 DP-GAN 及车牌匹配网络 SNN。对于车牌检测器，我们预先标注了约 10000 组训练样本，包括车辆图像及图像中车牌区域的包围盒，使用 Caffe 实现并训练。对于 DP-GAN，采用 PyTorch 实现和训练，网络训练如第4.3.4节所述，本系统标注了约 20000 对 LR-HR 车牌图像对训练网络。对于车牌匹配网络 SNN，采用 Caffe 实现和训练，网络训练如第4.4.2节所述，我们使用了约 100000 对车牌图像对用于 SNN 训练。在特征提取阶段，首先使用车牌检测器检测车牌图像中的车牌区域。如检测到车牌，则将车牌图像截取，输入 DP-GAN 进行前向传播运算，得到高分辨率车牌图像。最后使用训练好的 SNN 模型提取车牌图像的特征。

（4）多级索引建立模块

对车辆图像提取上述特征后，将车辆图像存储于图像存储服务器，索引建立模

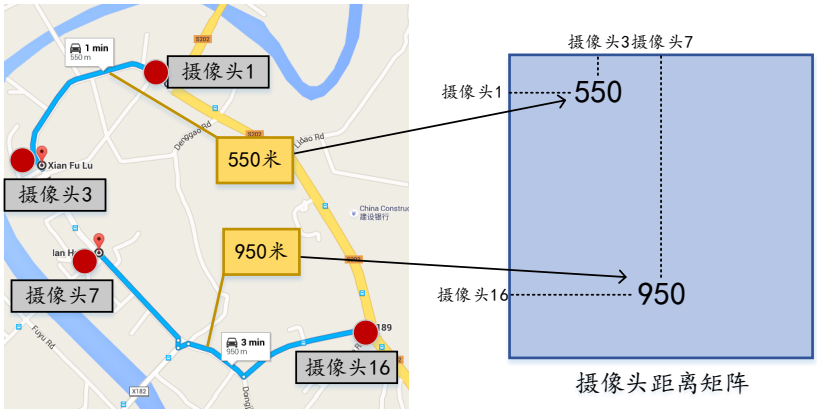


图 5-6　监控摄像头间的空间距离矩阵

块对两种特征向量分别建立外观索引、车牌索引，供搜索时匹配使用。

在实现时，我们以 1 小时为一个时间片，将所有摄像头采集车辆图像的外观特征组成外观特征矩阵、车牌特征组成车牌特征矩阵。采用 FLANN[116] 的索引建立接口对两个特征矩阵分别建立外观索引和车牌索引。在特征建立时，使用 KD-树建立索引结构，树的数量设为 4。索引建立后按照时间片顺序保存于数据服务器。

此外，为实现基于时空信息的搜索结果重排序，系统预先存储了城市监控系统中的摄像头距离矩阵，用于保存所有摄像头之间的空间距离，如图5-6所示。并且，为了实现由近及远的渐进式搜索，系统使用邻接表存储每个摄像头的邻接摄像头。

5.3.2　车辆搜索子系统

车辆搜索子系统主要负责接收用户输入的查询图像、时间戳、摄像头编号，然后提取车辆的外观、车牌特征，然后分别根据上述特征在多级索引中搜索目标车辆，返回目标车辆的图像、摄像头编号、出现时间等信息。我们假设车辆数据收集子系统离线处理来自交通监控系统的视频数据，因此不考虑数据传输、通信等问题。该子系统主要包括如下模块：

（1）输入处理与结果显示模块

该模块的功能是将用户输入的车辆图像归一化为特征提取模块所需的大小，并与时间戳、摄像头编号组成三元组传给特征提取模块。输入预处理模块部署于客户端或浏览器端供用户使用。结果显示模块将搜索结果按照相似度由大到小排序，将结果的车辆图像、时间戳、摄像头编号返回用户界面显示，可由用户选择将结果保存为本地文件。结果显示模块部署于客户端或浏览器端供用户使用。

本原型系统面向 PC 端应用，基于微软基础类库（Microsoft Fundation Classes，

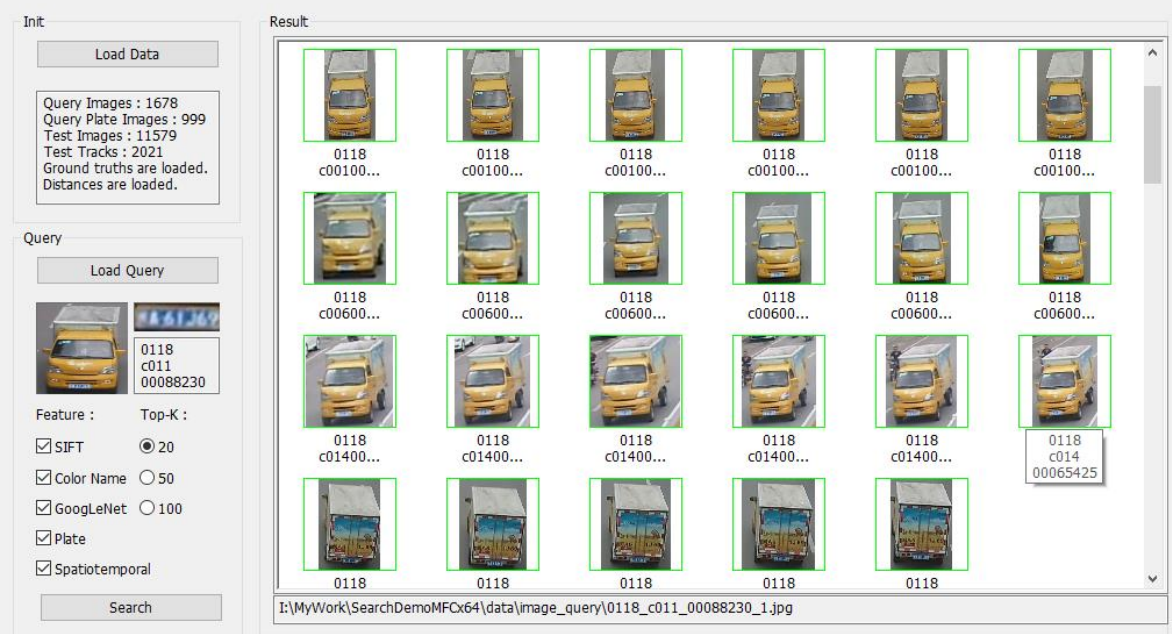


图 5-7　车辆搜索系统输入处理与结果显示界面

MFC）框架实现输入处理与结果显示模块。如图5-7所示，界面左侧为输入部分，由用户上传的查询车辆图像，接收用户输入的摄像头编号和时间戳。左侧也可选择搜索的粒度，如仅使用外观特征搜索、外观与车牌特征搜索、基于时空信息重排序。界面右侧为结果显示界面，按照相似度从大到小列出车辆的图像，图像下方显示摄像头编号、时间戳信息。

（2）多级特征提取模块

该模块的功能和实现与车辆数据收集子系统中的特征提取模块实现相同，用于提取车辆的外观、车牌特征提取实现如第5.3.1节多级特征提取模块。特征提取完成后，将车辆外观特征、车牌特征、摄像头编号、时间戳传入匹配模块。该模块同样部署于配有 GPU 的特征提取服务器。

（3）渐进式匹配模块

渐进式匹配模块将特征提取模块提取的外观特征、车牌特征依次与外观索引、车牌索引匹配。首先匹配外观特征，取外观相似度最大的前 N_a 个结果作为候选结果进行车牌特征匹配。然后取车牌相似度最大的前 N_p 个结果作为候选结果进行时空特征匹配。最后，计算查询图像与候选结果的时空相似度，将相似度最大的前 N 个结果作为最终结果传给结果显示模块。该模块部署于数据存储服务器。

具体实现时，该模块首先根据输入的时间戳确定其所在的时间片，使用 FLANN

的 load 接口将该时间片的外观索引和车牌索引加载入内存。然后，使用 FLANN 的 radiusSearch 接口，根据输入图像的外观特征在外观索引中进行基于半径的最近邻搜索。对于外观特征搜索结果，本系统取前 30% 作为候选结果。下一步，根据输入图像的车牌特征在候选结果的车牌索引中进行基于半径的最近邻搜索。此时，本系统同样取前 30% 作为候选结果。最后，通过公式5-1计算查询车辆与候选结果的时空相似性，并与外观相似性、车牌相似性加权融合得到最终的相似性排序，并将搜索结果返回结果显示界面。

5.3.3 系统运行环境

系统使用的硬件设备主要包括：

（1）图形处理服务器 1 台（Intel Xeon CPU，4 块 NVIDIA K80 GPU，14TB 固态硬盘，256G 内存）

（2）DELL 高性能视频工作站 1 台（Intel Xeon CPU，2TB 硬盘，16G 内存）

（3）20TB 存储 NVR 视频服务器 1 台

（4）20TB 存储容量的磁盘阵列 1 台

系统使用的软件环境主要包括：

（1）Ubuntu Linux 14.04.5 LTS 操作系统（图形服务器）

（2）WIndows 7 操作系统（DELL 工作站）

（3）CUDA 7.0（图形服务器）

（4）Microsoft Visual Studio 2013（DELL 工作站）

（5）OpenCV 2.4.9

（6）Matlab R2014b

（7）Python 2.7.11

（8）FFmpeg 3.4.1

（9）Caffe 深度学习工具包[81]

（10）PyTorch 深度学习工具包[117]

5.4 系统测试

由于本文各章已对各算法的性能在大规模数据集上进行了评估与分析，本节主要从数据收集子系统的车辆检测模块、车辆搜索子系统整体性能两个方面展开测试和分析。

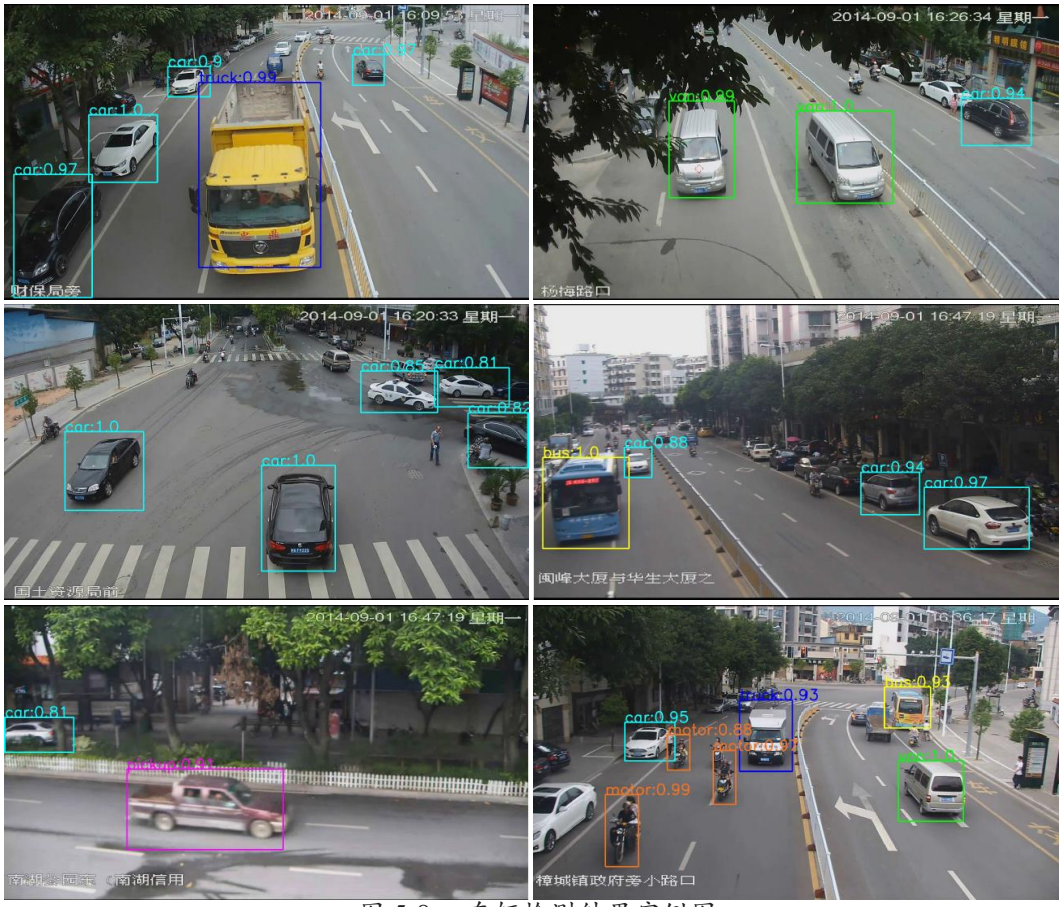


图 5-8　车辆检测结果实例图

5.4.1　车辆检测模块

由于数据收集子系统在服务器后台运行，因此各模块执行结果并不在界面显示。图5-8展示了车辆检测模块在一些典型城市交通监控场景下的车辆检测结果。

从车辆检测结果实例可以看出，本系统车辆检测模块能够有效检测多种复杂交通场景下的车辆，对拍摄视角、拍摄距离、光照变化、局部遮挡和复杂背景具有较好的鲁棒性。这是由于在训练 Faster R-CNN 车辆检测器时，训练数据规模足够大，包含了真实监控场景的不同情况，使得 Faster R-CNN 网络能够学习到具有较强鲁棒性的车辆特征。

5.4.2　车辆搜索系统

我们在本文提出的 VeRi 数据集上测试车辆搜索框架的整体准确率。这里采用逐渐增加模块的方式对比各算法，即基于外观的相似车辆搜索（NuFACT）、基于车牌

表 5-1　不同方法在 VeRi 数据集上的结果对比

方法	mAP	HIT@1	HIT@5
NuFACT	48.47	76.76	91.42
NuFACT + Plate-SNN	50.87	81.11	92.79
NuFACT + Plate-SNN + STSM	53.42	81.56	95.11
PROVID	**60.47**	**85.52**	**95.11**

匹配的精确搜索（Plate-SNN）、基于时空相似度的重排序（STSM）、基于 DP-GAN 的车牌超分辨率。同样采用 VeRi 数据集上的评价标准 mAP、HIT@1、HIT@5 评估方法准确率。

表5-1列出了不同方法在 VeRi 数据集上的结果对比。从对比中可以看出，在使用外观进行粗略搜索后，大部分外观相似车辆已经被搜索到。通过基于车牌匹配的精确搜索，准确率进一步提高，验证了车牌信息的作用。此外，基于时空信息对搜索结果重排序后，mAP 显著提升，从而表明渐进式搜索框架的有效性。最后，加入了车牌超分辨率后完整的系统框架（PROVID）取得了最好的准确率，不仅表明基于 DP-GAN 的车牌超分辨率在车牌匹配中起到重要作用，而且证明了整体系统的优异性能。

图5-9列出了部分搜索结果实例的对比，在每个查询的结果中，虚线左边的三行分别表示 FACT、FACT+Plate、FACT+Plate+STSM，虚线右边三行分别表示 NuFACT、NuFACT+Plate、NuFACT+Plate+STSM。通过对比虚线两边的结果可以看出，我们提出的外观表示模型能够更有效地融合车辆多级外观特征，实现近似车辆的搜索。通过对比每个查询中的三行结果可以看出，通过外观、车牌、时空信息的渐进式搜索，搜索准确性能够不断提高。

搜索结果中的一些例子也反映了城市监控中车辆搜索的主要难点，包括如下三个方面：

- 第一，环境因素。例如，监控环境多变的光照条件使得相同车辆显现出不同的颜色，在黑暗条件下尤其严重。此外，由于车辆表面的镜面反射，车辆在阳光照射下会反射强烈的阳光，使得图像局部曝光过度，导致车辆特征弱化。

- 第二，摄像头设置。一方面，由于城市中摄像头的型号不同，所以摄像头的分辨率、帧率、光圈、快门速度等参数不同，导致拍摄的车辆图像分辨率、对比度不同，甚至可能造成运动模糊、噪声等干扰。另一方面摄像头安装的位置、

图 5-9　渐进式车辆搜索框架在 VeRi 数据集的搜索结果实例

高度、角度各异，导致拍摄的车辆图像视角变化十分剧烈。以上原因都不仅影响车辆外观特征的提取与匹配，而且可能使得车牌区域模糊、形变，甚至无法拍摄到车牌。

- 第三，车辆外观歧义性。车辆外观的歧义性主要来自于同一品牌同一型号相同颜色的车辆。在这种情况下，仅通过外观很难判断两个车辆图像是否属于同一辆车，必须通过车牌才能确定车辆身份。此时，如果车牌无法被拍摄到、被伪造、被遮挡，那么将无法确定车辆的身份。

但是，即使是上述极端情况，渐进式车辆搜索系统可能无法给出准确搜索结果，仍能为城市监控中的车辆搜索提供一定的辅助作用。

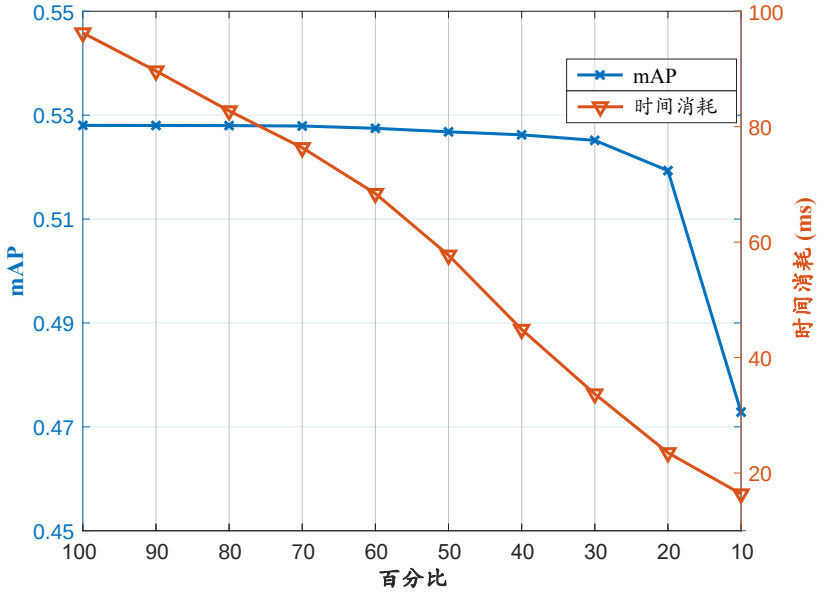


图 5-10　渐进式搜索时取不同比例候选车辆的时间消耗与 mAP

5.4.3　搜索效率分析

如第5.3.2节所述，渐进式搜索过程中，通过取外观相似结果前 N_a 个作为候选结果和车牌匹配结果前 N_p 个作为候选结果的方式逐渐减小搜索规模，提高搜索效率。本节将比较 N_a 与 N_p 取不同比例数值时系统的准确率与效率的关系。为了更有效测试搜索效率，首先在 VeRI 数据集测试集 2021 个测试轨迹基础上，增加了 99029 个无关轨迹，是规模扩大为原测试集的 50 倍。然后，在扩大后的测试集上，取 $N_a = N_p = \{10\%, 20\%, 30\%, 40\%, 50\%, 60\%, 70\%, 80\%, 90\%, 100\%\}$，分别计算 mAP 及搜索时间（以毫秒/查询计）。

测试结果如图5-10所示，从 100% 到 30% 区间，mAP 仅有极小的下降，而时间消耗从 92.4 毫秒/查询降低到 32.5 毫秒/查询。取 $N_a = N_p = 30\%$ 时，渐进式搜索框架可以降低 64.8% 的时间消耗，仅损失不到 0.01 的 mAP。因此，我们提出的渐进式车辆搜索框架不仅具有优异的准确性，同时能够大幅降低大规模监控场景下车辆搜索的时间开销。

5.5　本章小结

我们设计和实现了多模数据融合的渐进式车辆搜索原型系统，该系统主要包括车辆数据收集子系统和车辆搜索子系统两个部分。通过基于外观的相似车辆搜索、

基于车牌匹配的精确车辆搜索、基于时空信息的搜索结果重排序，该系统可以准确快速地搜索目标车辆。通过在 VeRi 数据集上的对比实验验证了车辆搜索系统优异的准确性和高效性。

第六章 总结与展望

6.1 论文工作总结

为提高城市交通监控系统的信息化、智能化水平，面向交通监控视频中车辆的相关研究引起了工业界和学术界的广泛关注。目前的研究主要关注于监控中的车辆检测[2]、车辆分类[3]、视角估计[4]、车牌识别[5]、车辆跟踪[6]、行为分析[8] 等问题。但是，与上述研究课题相比，面向城市视频监控网络的车辆搜索是一个十分重要、值得研究的课题。

车辆搜索，即给定一个车辆的查询图像或视频，在由视频监控网络收集的车辆数据库中搜索与查询车辆身份相同的车辆。车辆搜索面临“搜不准”和“搜得慢”两大挑战。一方面由于车辆自身外观的相似性和多样性，另一方面由于无约束城市监控中极端多变的环境因素，导致相同车辆外观存在很大的类内差异，不同车辆的外观却可能有较小的类间差异，给车辆搜索的准确性带来巨大挑战。此外，不仅因为城市监控视频中巨大的数据规模，而且因为车辆多模态特征的丰富性和复杂性，使得保证准确率的情况下提高搜索效率成为很大的挑战。

针对上述挑战，本文面向大规模城市交通视频监控场景，提出了一种融合多模态数据的渐进式车辆搜索框架，并从车辆外观特征的表示与学习、车牌图像的有效利用、监控网络中时空关系的挖掘三个方面提出一系列新方法与新模型。此外，为推进车辆搜索和相关领域的发展，本文还收集并标注了一个来自于真实城市交通监控系统的大规模车辆搜索数据集，通过在该数据集上的大量实验验证了所提出的框架与方法。本文的主要贡献具体如下：

（1）融合多模态数据的渐进式车辆搜索框架。该框架综合特征域和时空域进行逐步求精地搜索，具体来说：一是特征域内由粗到精地搜索，即先采用外观特征快速查找相似车辆，再使用车牌信息实现精确搜索；二是在时空域内，利用监控网络中的时空信息由近及远地搜索。实验分析表明，渐进式搜索框架不仅能够显著降低车辆搜索的时间消耗，同时保证了车辆搜索的准确性。

（2）基于车辆外观特征的相似车辆搜索方法。针对抓拍图像和视频两种查询数据，我们分别提出了两种基于深度卷积神经网络（Convolutional Neural Network，CNN）的车辆外观表示方法：NuFACT 和 CAN。NuFACT 方法能够从抓拍车辆图像

中提取车辆的纹理、颜色、类别等多级特征，并通过零空间度量学习将上述特征融合为一种具有区分力的、鲁棒的特征。CAN 方法能够提取视频中多张图像的共有信息和互补信息，自动学习不同距离、不同角度图像中的有效特征，增强了车辆外观特征的区分力和鲁棒性。

（3）车牌图像超分辨率与验证结合的精确车辆搜索方法。针对无约束监控环境中低质的车牌图像，我们提出了一种基于域先验生成对抗网络的图像超分辨率方法进行车牌图像增强。针对监控数据中车辆数量很大而每个车辆样本较少的问题，本文采用一种基于对偶神经网络（Siamese Neural Network，SNN）的车牌验证方法，实现了车牌图像的快速准确匹配。通过车牌增强与验证结合，进一步提高了车辆搜索的准确性。

（4）基于邻接图与时空相似度模型的搜索结果重排序。通过挖掘城市监控网络中的时空信息，如车辆被拍摄的时间、摄像头的位置、摄像头邻接关系等，我们设计了一种摄像头邻接图模型表示视频监控网络的空间拓扑，提出了一种基于多层感知机的时空相似度模型（Spatio-Temporal Similarity Model，STSM），通过 STSM 估计车辆间的时空相似性对搜索结果进行重排序，得到优化的车辆搜索结果。

最后，本文搭建了一个融合多模态数据的渐进式车辆搜索原型系统，并在真实视频监控数据上验证了上述框架与方法的有效性。

6.2 未来工作展望

本文对监控视频中的车辆搜索问题进行了较为深入的研究，在车辆外观特征表示、车牌信息利用、时空数据挖掘等方面取得了一定的成果。但是，车辆搜索仍是一个新兴的研究方向，在模型、方法和应用方面仍有很多问题亟待解决，距离真实应用还有较大差距。值得进一步研究的方向包括：

- 第一，基于三维视觉模型的车辆外观表示。人类视觉系统能够感知物体的三维视觉特征，从而能够有效识别和区分不同对象。车辆作为一种人工设计生产的物体，较容易获取其精细的三维模型数据。借助精细的三维模型，有助于 CNN 或其他视觉模型充分学习车辆外观的局部和全局特征。此外，车辆作为一种刚性物体，在三维视觉模型下具有很强的对称性和规律性。车辆在不同摄像头拍摄的图像，类似于同一摄像头在不同视角下对车辆拍摄的多组图像，因此具有较强三维视觉属性。因此，如何将三维视觉模型与车辆外观表示结合，为车辆搜索系统提供更具区分力和鲁棒性的视觉特征，值得我们进一步研究。

- 第二，大规模监控网络中的车辆搜索。

  本文对城市视频监控网络中的车辆搜索进行了初步探索，在相对较小规模的视频监控网络中实现了较好的搜索性能。但是，真实的城市级甚至国家级的视频监控网络规模是十分巨大的，网络中包含了数万甚至百万级的监控摄像头。在如此大规模的视频监控网络中实现准确、高效的车辆搜索服务，具有巨大的研究意义和应用价值，同时也存在巨大的挑战。一方面，视频监控数据规模巨大，使得视频处理、对象检测、特征提取、索引构建等方面必须具有高效性。如何利用云计算、边缘计算等新兴技术处理大规模监控视频数据，高效提取车辆多模态信息仍需进一步研究。另一方面，视频监控网络感知的物理空间巨大，使得网络拓扑、时空信息的规模与复杂度呈指数增长。因此，如何充分挖掘时空大数据，提高车辆搜索的高效性和准确性是重要的研究方向。总之，在未来的研究工作中，我们将研究车辆搜索在大规模视频监控网络中的关键理论与技术。

- 第三，车辆搜索与智慧城市相结合。

  智慧城市运用大数据、云计算、物联网、人工智能等技术手段，实时感知城市的状态与变化，为生产生活、城市交通、公共安全、环境保护等方面提供智能的响应与服务，从而为人民创造更高效、安全、健康的生活。车辆是城市运行的重要参与对象，车辆搜索服务以视频监控网络为基础为城市管理者提供精确车辆搜索、相似车辆搜索服务。更进一步，城市管理者可以利用车辆搜索实现智能交通、案件侦查、车辆管理，有效提高城市管理的效率。另一方面，车辆在城市中行驶受到天气状况、城市热点区域、信号灯变化、道路拥堵等因素的影响。基于城市中天气、人流、车流、热点区域等大数据，可以分析和预测城市交通的状态与变化，为车辆搜索提供帮助。将城市计算技术与交通视频监控网络结合，挖掘城市中多模态数据与车辆时空信息的关系，能够进一步提高车辆搜索的准确性和高效性。因此，车辆搜索与智慧城市的结合是未来重要的研究方向。

参考文献

[1] 中国统计年鉴—2017[EB/OL]. http://www.stats.gov.cn/tjsj/ndsj/2017/indexch.htm.

[2] 李波. 基于图像分析的车辆识别与跟踪若干关键技术研究 [D]. 湖北，武汉: 华中科技大学, 2011.

[3] Fu J, Zheng H, Mei T. Look Closer to See Better: Recurrent Attention Convolutional Neural Network for Fine-Grained Image Recognition[A]. // IEEE Conference on Computer Vision and Pattern Recognition[C]. 2017: 4476–4484.

[4] Yang L, Liu J, Tang X. Object detection and viewpoint estimation with auto-masking neural network[A]. // European Conference on Computer Vision[C]. 2014: 441–455.

[5] Wen Y, Lu Y, Yan J, et al. An algorithm for license plate recognition applied to intelligent transportation system[J]. IEEE Transactions on Intelligent Transportation Systems, 2011, 12(3): 830–845.

[6] Matei B C, Sawhney H S, Samarasekera S. Vehicle tracking across nonoverlapping cameras using joint kinematic and appearance features[A]. // IEEE Conference on Computer Vision and Pattern Recognition[C]. 2011: 3465–3472.

[7] Sochor J, Juránek R, Herout A. Traffic Surveillance Camera Calibration by 3D Model Bounding Box Alignment for Accurate Vehicle Speed Measurement[J]. Computer Vision and Image Understanding: 87–98.

[8] Dubská M, Herout A, Sochor J. Automatic Camera Calibration for Traffic Understanding[A]. // British Machine Vision Conference[C]. 2014: 8.

[9] 王龙飞. 基于车牌照的车辆出行轨迹分析方法与实践研究 [D]. 陕西，西安: 长安大学, 2011.

[10] Dubská M, Herout A, Juranek R, et al. Fully automatic roadside camera calibration for traffic surveillance[J]. IEEE Transactions on Intelligent Transportation Systems, 2015, 16(3): 1162–1171.

[11] “易”系列抓拍单元 [EB/OL]. https://www.dahuatech.com/product/lists/129.html?area=131.

[12] iDS-2CD9371-K(S) 智能交通网络摄像机 [EB/OL]. http://www.hikvision.com/cn/prgs_1065_i17012.html.

[13] Ma H D, Liu W. Progressive Search Paradigm for Internet of Things[J]. IEEE MultiMedia, 2018, 25(1): 76–86.

[14] Valera M, Velastin S A. Intelligent distributed surveillance systems: a review[J]. IEE Proceedings - Vision, Image and Signal Processing, 2005, 152(2): 192–204.

[15] Zhang J, Wang F Y, Wang K, et al. Data-driven intelligent transportation systems: A survey[J]. IEEE Transactions on Intelligent Transportation Systems, 2011, 12(4): 1624–1639.

[16] Zheng Y, Capra L, Wolfson O, et al. Urban computing: concepts, methodologies, and applications[J]. ACM Transactions on Intelligent Systems and Technology, 2014, 5(3): 38.

[17] Ma H D. Internet of Things: Objectives and Scientific Challenges[J]. Journal of Computer Science and Technology, 2011, 26(6): 919–924.

[18] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[A]. // Advances in Neural Information Processing Systems[C]. 2012: 1097–1105.

[19] Su H, Qi C R, Li Y, et al. Render for cnn: Viewpoint estimation in images using cnns trained with rendered 3d model views[A]. // IEEE International Conference on Computer Vision[C]. 2015: 2686–2694.

[20] Javed O, Shafique K, Rasheed Z, et al. Modeling inter-camera space–time and appearance relationships for tracking across non-overlapping views[J]. Computer Vision and Image Understanding, 2008, 109(2): 146–162.

[21] Xu J, Jagadeesh V, Ni Z, et al. Graph-based topic-focused retrieval in distributed camera network[J]. IEEE Transactions on Multimedia, 2013, 15(8): 2046–2057.

[22] Sun Z, Bebis G, Miller R. On-road vehicle detection: A review[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2006, 28(5): 694–711.

[23] Kembhavi A, Harwood D, Davis L S. Vehicle detection using partial least squares[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 33(6): 1250–1265.

[24] Zhang Z, Tan T, Huang K, et al. Three-dimensional deformable-model-based localization and recognition of road vehicles[J]. IEEE Transactions on Image Processing, 2012, 21(1): 1–13.

[25] Du S, Ibrahim M, Shehata M, et al. Automatic license plate recognition (ALPR): A state-of-the-art review[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2013, 23(2): 311–325.

[26] Krause J, Stark M, Deng J, et al. 3d object representations for fine-grained categorization[A]. // IEEE International Conference on Computer Vision Workshops[C]. 2013: 554–561.

[27] Sivaraman S, Trivedi M M. Looking at vehicles on the road: A survey of vision-based vehicle detection, tracking, and behavior analysis[J]. IEEE Transactions on Intelligent Transportation Systems, 2013, 14(4): 1773–1795.

[28] Yang L, Luo P, Loy C C, et al. A Large-Scale Car Dataset for Fine-Grained Categorization and Verification[A]. // IEEE Conference on Computer Vision and Pattern Recognition[C]. 2015: 3973–3981.

[29] Sochor J, Herout A, Havel J. Boxcars: 3d boxes as cnn input for improved fine-grained vehicle recognition[A]. // IEEE Conference on Computer Vision and Pattern Recognition[C]. 2016: 3006–3015.

[30] He K, Sigal L, Sclaroff S. Parameterizing object detectors in the continuous pose space[A]. // European Conference on Computer Vision[C]. 2014: 450–465.

[31] Xiang Y, Song C, Mottaghi R, et al. Monocular multiview object tracking with 3d aspect parts[A]. // European Conference on Computer Vision[C]. 2014: 220–235.

[32] Feris R, Siddiquie B, Zhai Y, et al. Attribute-based vehicle search in crowded surveillance videos[A]. // ACM International Conference on Multimedia Retrieval[C]. 2011: 18.

[33] Feris R S, Siddiquie B, Petterson J, et al. Large-scale vehicle detection, indexing, and search in urban surveillance videos[J]. IEEE Transactions on Multimedia, 2012, 14(1): 28–42.

[34] Liu H, Tian Y, Yang Y, et al. Deep Relative Distance Learning: Tell the Difference Between Similar Vehicles[A]. // IEEE Conference on Computer Vision and Pattern Recognition[C]. 2016: 2167–2175.

[35] Zhang Y, Liu D, Zha Z J. Improving triplet-wise training of convolutional neural network for vehicle re-identification[A]. // IEEE International Conference on Multimedia and Expo[C]. 2017: 1386–1391.

[36] Yan K, Tian Y, Wang Y, et al. Exploiting Multi-Grain Ranking Constraints for Precisely Searching Visually-Similar Vehicles[A]. // IEEE International Conference on Computer Vision[C]. 2017: 562–570.

[37] Liu X C, Liu W, Mei T, et al. A Deep Learning-Based Approach to Progressive Vehicle Re-identification for Urban Surveillance[A]. // European Conference on Computer Vision[C]. 2016: 869–884.

[38] Shen Y, Xiao T, Li H, et al. Learning Deep Neural Networks for Vehicle Re-ID with Visual-spatio-temporal Path Proposals[A]. // IEEE International Conference on Computer Vision[C]. 2017: 1918–1927.

[39] Wang Z, Tang L, Liu X, et al. Orientation Invariant Feature Embedding and Spatial Temporal Regularization for Vehicle Re-Identification[A]. // IEEE International Conference on Computer Vision[C]. 2017: 379–387.

[40] Liu X, Ma H D, Fu H Y, et al. Vehicle Retrieval and Trajectory Inference in Urban Traffic Surveillance Scene[A]. // International Conference on Distributed Smart Cameras[C]. 2014: 26.

[41] Gong S, Cristani M, Yan S, et al. Person re-identification[M]. Springer, 2014.

[42] Li W, Zhao R, Xiao T, et al. Deepreid: Deep filter pairing neural network for person re-identification[A]. // IEEE Conference on Computer Vision and Pattern Recognition[C]. 2014: 152–159.

[43] Farenzena M, Bazzani L, Perina A, et al. Person re-identification by symmetry-driven accumulation of local features[A]. // IEEE Conference on Computer Vision and Pattern Recognition[C]. 2010: 2360–2367.

[44] Liao S, Hu Y, Zhu X, et al. Person re-identification by local maximal occurrence representation and metric learning[A]. // IEEE Conference on Computer Vision and Pattern Recognition[C]. 2015: 2197–2206.

[45] Zhao R, Ouyang W, Wang X. Person re-identification by salience matching[A]. // IEEE International Conference on Computer Vision[C]. 2013: 2528–2535.

[46] Zhao R, Ouyang W, Wang X. Learning mid-level filters for person re-identification[A]. // IEEE Conference on Computer Vision and Pattern Recognition[C]. 2014: 144–151.

[47] Zheng L, Shen L, Tian L, et al. Scalable person re-identification: A benchmark[A]. // IEEE International Conference on Computer Vision[C]. 2015: 1116–1124.

[48] Prosser B J, Zheng W S, Gong S, et al. Person re-identification by support vector ranking.[A]. // British Machine Vision Conference[C]. 2010: 6.

[49] Zheng W S, Gong S, Xiang T. Person re-identification by probabilistic relative distance comparison[A]. // IEEE Conference on Computer Vision and Pattern Recognition[C]. 2011: 649–656.

[50] Hirzer M, Roth P M, Köstinger M, et al. Relaxed pairwise learned metric for person re-identification[A]. // European Conference on Computer Vision[C]. 2012: 780–793.

[51] Zhao R, Ouyang W, Wang X. Unsupervised salience learning for person re-identification[A]. // IEEE Conference on Computer Vision and Pattern Recognition[C]. 2013: 3586–3593.

[52] Zhang L, Xiang T, Gong S. Learning a Discriminative Null Space for Person Re-identification[A]. // IEEE International Conference on Computer Vision[C]. 2016: 1239–1248.

[53] Zhang C, Liu W, Ma H D, et al. Siamese neural network based gait recognition for human identification[A]. // IEEE International Conference on Acoustics, Speech and Signal Processing[C]. 2016: 2832–2836.

[54] Wang T, Gong S, Zhu X, et al. Person re-identification by video ranking[A]. // European Conference on Computer Vision[C]. 2014: 688–703.

[55] Zheng L, Bie Z, Sun Y, et al. Mars: A video benchmark for large-scale person re-identification[A]. // European Conference on Computer Vision[C]. 2016: 868–884.

[56] McLaughlin N, Martinez del Rincon J, Miller P. Recurrent convolutional network for video-based person re-identification[A]. // IEEE Conference on Computer Vision and Pattern Recognition[C]. 2016: 1325–1334.

[57] Yan Y, Ni B, Song Z, et al. Person re-identification via recurrent feature aggregation[A]. // European Conference on Computer Vision[C]. 2016: 701–716.

[58] Liu Y, Yan J, Ouyang W. Quality Aware Network for Set to Set Recognition[A]. // IEEE Conference on Computer Vision and Pattern Recognition[C]. 2017: 4694–4703.

[59] Zhu X, Jing X Y, Wu F, et al. Learning Heterogeneous Dictionary Pair with Feature Projection Matrix for Pedestrian Video Retrieval via Single Query Image.[A]. // AAAI Conference on Artificial Intelligence[C]. 2017: 4341–4348.

[60] Zhou Z, Huang Y, Wang W, et al. See the Forest for the Trees: Joint Spatial and Temporal Recurrent Neural Networks for Video-based Person Re-identification[A]. // IEEE Conference on Computer Vision and Pattern Recognition[C]. 2017: 6776–6785.

[61] Girshick R. Fast r-cnn[A]. // IEEE International Conference on Computer Vision[C]. 2015: 1440–1448.

[62] Ren S, He K, Girshick R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[A]. // Advances in Neural Information Processing Systems[C]. 2015: 91–99.

[63] Song J, Yang Y, Huang Z, et al. Effective multiple feature hashing for large-scale near-duplicate video retrieval[J]. IEEE Transactions on Multimedia, 2013, 15(8): 1997–2008.

[64] Xie L, Wang J, Zhang B, et al. Fine-grained image search[J]. IEEE Transactions on Multimedia, 2015, 17(5): 636–647.

[65] Mei T, Rui Y, Li S, et al. Multimedia search reranking: A literature survey[J]. ACM Computing Surveys, 2014, 46(3): 38.

[66] Meng J, Yuan J, Yang J, et al. Object instance search in videos via spatio-temporal trajectory discovery[J]. IEEE Transactions on Multimedia, 2016, 18(1): 116–127.

[67] Lowe D G. Distinctive image features from scale-invariant keypoints[J]. International Journal of Computer Vision, 2004, 60(2): 91–110.

[68] Zheng L, Wang S, Zhou W, et al. Bayes merging of multiple vocabularies for scalable image retrieval[A]. // IEEE Conference on Computer Vision and Pattern Recognition[C]. 2014: 1955–1962.

[69] Van De Weijer J, Schmid C, Verbeek J, et al. Learning color names for real-world applications[J]. IEEE Transactions on Image Processing, 2009: 1512–1523.

[70] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[A]. // IEEE Conference on Computer Vision and Pattern Recognition[C]. 2015: 1–9.

[71] Russakovsky O, Deng J, Su H, et al. Imagenet large scale visual recognition challenge[J]. International Journal of Computer Vision, 2015, 115(3): 211–252.

[72] Foley D H, Sammon J W. An optimal set of discriminant vectors[J]. IEEE Transactions on Computers, 1975, 100(3): 281–289.

[73] Guo Y F, Wu L, Lu H, et al. Null foley–sammon transform[J]. Pattern recognition, 2006, 39(11): 2248–2251.

[74] Chatfield K, Simonyan K, Vedaldi A, et al. Return of the devil in the details: Delving deep into convolutional nets[A]. // British Machine Vision Conference[C]. 2014.

[75] Liu X C, Liu W, Ma H D, et al. Large-scale vehicle re-identification in urban surveillance videos[A]. // IEEE International Conference on Multimedia and Expo[C]. 2016: 1–6.

[76] Liu X C, Liu W, Mei T, et al. PROVID: Progressive and Multi-modal Vehicle Re-identification for Large-scale Urban Surveillance[J]. IEEE Transactions on Multimedia, 2018, 20(3): 645–658.

[77] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[A]. // IEEE Conference on Computer Vision and Pattern Recognition[C]. 2016: 770–778.

[78] Su H, Maji S, Kalogerakis E, et al. Multi-view convolutional neural networks for 3d shape recognition[A]. // IEEE International Conference on Computer Vision[C]. 2015: 945–953.

[79] Chopra S, Hadsell R, LeCun Y. Learning a similarity metric discriminatively, with application to face verification[A]. // IEEE Conference on Computer Vision and Pattern Recognition[C]. 2005: 539–546.

[80] Schroff F, Kalenichenko D, Philbin J. Facenet: A unified embedding for face recognition and clustering[A]. // IEEE Conference on Computer Vision and Pattern Recognition[C]. 2015: 815–823.

[81] Jia Y, Shelhamer E, Donahue J, et al. Caffe: Convolutional architecture for fast feature embedding[A]. // ACM International Conference on Multimedia[C]. 2014: 675–678.

[82] Karpathy A, Toderici G, Shetty S, et al. Large-scale video classification with convolutional neural networks[A]. // IEEE Conference on Computer Vision and Pattern Recognition[C]. 2014: 1725–1732.

[83] Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos[A]. // Advances in Neural Information Processing Systems[C]. 2014: 568–576.

[84] Bromley J, Bentz J W, Bottou L, et al. Signature verification using a "Siamese" time delay neural network[J]. International Journal of Pattern Recognition and Artificial Intelligence, 1993, 7(04): 669–688.

[85] Zhang C, Liu W, Ma H D, et al. Siamese neural network based gait recognition for human identification[A]. // IEEE International Conference on Acoustics, Speech and Signal Processing[C]. 2016: 2832–2836.

[86] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection[A]. // IEEE Conference on Computer Vision and Pattern Recognition[C]. 2016: 779–788.

[87] Liu W, Anguelov D, Erhan D, et al. Ssd: Single shot multibox detector[A]. // European Conference on Computer Vision[C]. 2016: 21–37.

[88] Yang C Y, Ma C, Yang M H. Single-image super-resolution: A benchmark[A]. // European Conference on Computer Vision[C]. 2014: 372–386.

[89] Keys R. Cubic convolution interpolation for digital Image Processing[J]. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1981, 29(6): 1153–1160.

[90] Gonzalez R C, Woods R E, others. Addison-wesley Reading, 1992.

[91] Zhang K, Tao D, Gao X, et al. Learning multiple linear mappings for efficient single image super-resolution[J]. IEEE Transactions on Image Processing, 2015, 24(3): 846–861.

[92] Sun J, Xu Z, Shum H Y. Image super-resolution using gradient profile prior[A]. // IEEE Conference on Computer Vision and Pattern Recognition[C]. 2008: 1–8.

[93] Tai Y W, Liu S, Brown M S, et al. Super resolution using edge prior and single image detail synthesis[A]. // IEEE Conference on Computer Vision and Pattern Recognition[C]. 2010: 2400–2407.

[94] Yang J, Wright J, Huang T S, et al. Image super-resolution via sparse representation[J]. IEEE Transactions on Image Processing, 2010, 19(11): 2861–2873.

[95] Yang J, Wang Z, Lin Z, et al. Coupled dictionary training for image super-resolution[J]. IEEE Transactions on Image Processing, 2012, 21(8): 3467–3478.

[96] Kim K I, Kwon Y. Single-image super-resolution using sparse regression and natural image prior[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010, 32(6): 1127–1133.

[97] Dong C, Loy C C, He K, et al. Image super-resolution using deep convolutional networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 38(2): 295–307.

[98] Kim J, Kwon Lee J, Mu Lee K. Accurate image super-resolution using very deep convolutional networks[A]. // IEEE Conference on Computer Vision and Pattern Recognition[C]. 2016: 1646–1654.

[99] Mao H, Wu Y, Li J, et al. Super Resolution of the Partial Pixelated Images With Deep Convolutional Neural Network[A]. // ACM International Conference on Multimedia[C]. 2016: 322–326.

[100] Ledig C, Theis L, Huszar F, et al. Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network[A]. // IEEE Conference on Computer Vision and Pattern Recognition[C]. 2017: 105–114.

[101] Han Z, Jiang J, Hu R, et al. Face image super-resolution via nearest feature line[A]. // ACM International Conference on Multimedia[C]. 2012: 769–772.

[102] Jiang J, Hu R, Wang Z, et al. Face super-resolution via multilayer locality-constrained iterative neighbor embedding and intermediate dictionary learning[J]. IEEE Transactions on Image Processing, 2014, 23(10): 4220–4231.

[103] Suresh K V, Kumar G M, Rajagopalan A. Superresolution of license plates in real traffic videos[J]. IEEE Transactions on Intelligent Transportation Systems, 2007, 8(2): 321–331.

[104] Sun L, Hays J. Super-resolution from internet-scale scene matching[A]. // IEEE International Conference on Computational Photography[C].

[105] Farsiu S, Robinson M D, Elad M, et al. Fast and robust multiframe super resolution[J]. IEEE Transactions on Image Processing, 2004, 13(10): 1327–1344.

[106] Liu C, Sun D. On Bayesian adaptive video super resolution[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 36(2): 346–360.

[107] Ma Z, Liao R, Tao X, et al. Handling motion blur in multi-frame super-resolution[A]. // IEEE Conference on Computer Vision and Pattern Recognition[C]. 2015: 5224–5232.

[108] Kappeler A, Yoo S, Dai Q, et al. Video super-resolution with convolutional neural networks[J]. IEEE Transactions on Computational Imaging, 2016, 2(2): 109–122.

[109] Li Y, Li X, Fu Z, et al. Multiview Video Super-Resolution via Information Extraction and Merging[A]. // ACM International Conference on Multimedia[C]. 2016: 446–450.

[110] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets[A]. // Advances in Neural Information Processing Systems[C]. 2014: 2672–2680.

[111] Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition[J]. CoRR, 2014, abs/1409.1556.

[112] Bottou L. Large-scale machine learning with stochastic gradient descent[A]. // International Conference on Computational Statistics[C]. Springer, 2010: 177–186.

[113] Yang X, Kwitt R, Styner M, et al. Quicksilver: Fast predictive image registration - A deep learning approach[J]. NeuroImage, 2017, 158: 378–396.

[114] Liao R, Tao X, Li R, et al. Video super-resolution via deep draft-ensemble learning[A]. // IEEE International Conference on Computer Vision[C]. 2015: 531–539.

[115] Kettnaker V, Zabih R. Bayesian multi-camera surveillance[A]. // IEEE Conference on Computer Vision and Pattern Recognition[C]. 1999: 253–259.

[116] Muja M, Lowe D G. Scalable nearest neighbor algorithms for high dimensional data[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 36(11): 2227–2240.

[117] Paszke A, Gross S, Chintala S, et al. PyTorch: Tensors and dynamic neural networks in Python with strong GPU acceleration[EB/OL]. http://pytorch.org/.

致 谢

孟子曰：“天时不如地利，地利不如人和”。求学生涯行至今日如果说算是一场小小的胜利，首先要感谢计算机学科、物联网、人工智能迅猛发展之天时，其次要感谢北京邮电大学、计算机学院、物联网中心坚实平台之地利，最后但最重要的则是要感谢一路走来、不离不弃之人和。

首先，衷心感谢我的导师马华东教授。“古之学者必有师。师者，所以传道授业解惑也”。六年里，马老师向我传严谨治学、刻苦勤奋之道，授科学研究、论文著说之业，解追求真理、为人处世之惑。马老师不仅为我创造了十分优越的学习和科研环境，而且为我提供了很多到国际会议上交流的机会。从小论文逐字逐句的修改，到大论文的选题与撰写，从组会汇报的一无是处，到 ICME 的最佳学生论文，马老师对我的工作与成长都倾注了大量心血。雏鸟总要离巢，马驹终将离群，马老师的谆谆教诲将是伴我一生的宝贵财富。

然后，要感谢 BUPTMM 的组长刘武老师。刘武老师科研水平一流、学术视野开阔，是我亦师亦友的拍档。在我遇到瓶颈时总能给我启迪，在我遇到困难时总能给我帮助，他带领我在科研道路上勇往直前。感谢刘武老师与我一起熬夜修改论文、一起站在国际会议上侃侃而谈，这些并肩奋斗的日子将鼓舞我在未来不断向前。

感谢刘亮老师、傅慧源老师，是你们鼓励我走上科研道路。感谢刘亮老师在论文写作和科研方法上的帮助，感谢傅慧源老师在工程项目和系统研发方面的启发，你们一直是我学习的目标、努力的榜样。感谢物联网中心李文生老师、罗红老师、孙岩老师、段鹏瑞老师、赵东老师、张海涛老师、周安福老师，在我的成长中教会我很多。

感谢在科研路上一起经历风雨的师兄弟：吴红海、高一鸿、宋宇宁、魏汪洋、张征、卢大玮、陈建伟、王浩、张钊、郭莹莹、李双群、张明慧、周溪湉、张欢欢。感谢 BUPTMM 的成员们：刘鲲、刘培业、吕金娜、张诚、黄灏、赵晓萌、程鹏、董雄雄、刘通、齐恒、张逸凡、高文慧、李雅楠、吕浩然、孟祯、张萌、詹英等同学，很幸运与你们一同进步、一同成长。

深深感谢我的父亲刘俭先生和母亲余洁女士三十年的养育之恩。感谢我的妻子于佳对我一如既往的理解与支持。

刘鑫辰

2018 年五月于北京邮电大学

攻读学位期间发表的学术论文目录

期刊论文

[1] **Xinchen Liu** , Liu W, Mei T, Ma H D. PROVID: Progressive and Multi-modal Vehicle Re-identification for Large-scale Urban Surveillance[J]. IEEE Transactions on Multimedia, 2018, 20(3): 645–658. （SCI 收录，检索号：WOS:000425397500011）.

会议论文

[1] **Xinchen Liu** , Liu W, Mei T, Ma H D. A Deep Learning-Based Approach to Progressive Vehicle Re-identification for Urban Surveillance[A]. // European Conference on Computer Vision[C]. 2016: 869–884. （EI 收录，检索号：20164202901912）.

[2] **Xinchen Liu** , Liu W, Ma H D, Fu H Y. Large-scale Vehicle Re-identification in Urban Surveillance Videos[A]. // IEEE International Conference on Multimedia and Expo[C]. 2016: 1–6. （EI 收录，检索号：20163802815567，最佳学生论文）.

[3] **Xinchen Liu** , Ma H D, Fu H Y, Zhou M. Vehicle Retrieval and Trajectory Inference in Urban Traffic Surveillance Scene[A]. // ACM/IEEE International Conference on Distributed Smart Cameras[C]. 2014: 26. （EI 收录，检索号：20144900292976）.

[4] **Xinchen Liu** , Liu W, Ma H D. A Progressive Vehicle Search System for Video Surveillance Networks[A]. // IEEE International Conference on Multimedia Big Data[C]. 2018. （Accepted）.

[5] **Xinchen Liu** , Liu W, Ma H D. CAN: Cross-view Attentive Network for Video-based Vehicle Re-Identification[A]. // ACM Conference on Multimedia[C]. 2018. （Under review）.

[6] Gao W H, **Xinchen Liu** , Ma H D, Li Y N, Liu L. MMH: Multi-modal Hash for Instant Mobile Video Search[A]. // IEEE International Conference on Multimedia Information Processing and Retrieval[C]. 2018. （Accepted）.

[7] Liu W, **Xinchen Liu** , Ma H D, Cheng P. Beyond Human-level License Plate Super-resolution with Progressive Vehicle Search and Domain Priori GAN[A]. // ACM Conference on Multimedia[C]. 2017.

[8] Zhang Z, Liu W, Ma H D, **Xinchen Liu** . Going Clear from Misty Rain in Dark Channel Guided Network[A]. // IJCAI Workshop on AI for Internet of Things[C]. 2017.

[9] Li S Q, **Xinchen Liu** , Liu W, Ma H D, Zhang H T. A discriminative null space based deep learning approach for person re-identification[A]. // IEEE International Conference on Cloud Computing and Intelligence Systems[C]. 2016: 480–484. （EI 收录，检索号：20170603321899）.

发明专利

[1] 马华东, 傅慧源, **刘鑫辰** . 一种车辆轨迹预测分析方法 [P]. 中国: 申请号：201410551180.2, 2014–10–17.

[2] 马华东, 傅慧源, **刘鑫辰** , 张诚. 一种基于车辆轨迹向量确定车辆实际方向的方法及装置 [P]. 中国: 申请号：201510644656.1, 2015–10–08.

[3] 马华东, 刘武, **刘鑫辰**, 张海涛, 傅慧源. 一种车辆搜索方法及装置 [P]. 中国: 申请号: 201610798016.0, 2016–08–31.

[4] Ma H D, Liu W, **Xinchen Liu**, Zhang H T, Fu H Y. A Progressive Vehicle Searching Method and Device[P]. 美国: 申请号：15/350,813, 2016–11–14.