華中科技大學
HUAZHONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

基于深度学习和上下文语义的视觉内容识别与分析研究

博士论文答辩

2025年7月

华中科技大学 计算机科学与技术学院

博士生：欧新宇　　导　师：凌贺飞 教授

目录 Contents

Part
01

研究背景与现状

/ 课题来源

/ 研究背景和意义

/ 国内外研究现状

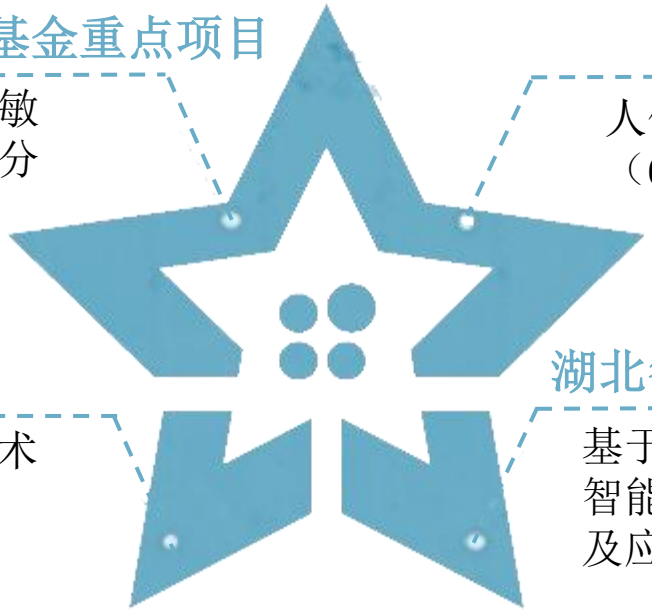
课题来源

国家自然科学基金-联合基金重点项目

网络大数据环境下的多媒体敏感内容感知、识别、检索与分析研究（U1536203）

国家自然科学基金

人像图片的语义理解方法研究（61572493）



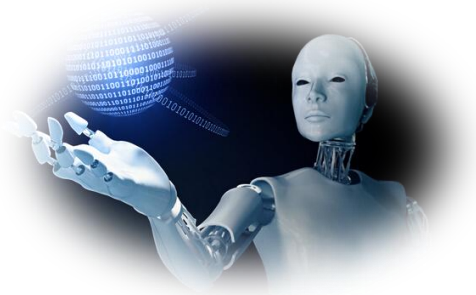
国家自然科学基金

面向社交网络的数字指纹技术研究（61272409）

湖北省自然科学基金创新项目

基于云计算的监控视频大数据智能分析与检索关键技术研发及应用（2015AAA013）

研究背景和意义



大数据时代

思科：2019年，全球每月产生视频**105EB**，占总流量的**77%**，仅YouTube每分钟将新增**300小时**（**12.5天**）视频。

艾瑞咨询：2016年，移动视频设备达到**9.1亿台**，占全部移动设备的**87.3%**。爱奇艺、优酷、腾讯视频用户超过**3亿**。

腾讯网：2017年，微信活跃用户数量达**8.46亿**，每日上传图片超过**10亿**张，视频播放超过**20亿**次，若以单人**3张/秒**的速度计算，微信每日产生的图片需要一个人花费**10.57年**才可浏览完毕。

无处不在的多媒体资源

图像、视频的直观性使其成为人类获取信息的最主要的途径，极大地丰富了人民群众的文化生活，包括：目标识别、目标检测、图像搜索、自动驾驶、卫星遥感分析、医疗辅助、三维建模等。

使用中的局限性

由于图像和视频大数据本身的特性，在处理和分析它们时依然有很多困难和挑战，主要包括：**效率**、**可用性**、**多样性**、**有用性**。

研究背景和意义

课题的提出

本论文拟在**深度学习的框架**下，结合**层次化语义关系**、**全局与局部语义关系等多种上下文关系**，针对深度学习模型在**图像视觉内容识别**、**场景解析**与**图像检索**等应用上的不足展开研究工作。

视觉内容的上下文语义

什么是视觉内容的上下文语义？

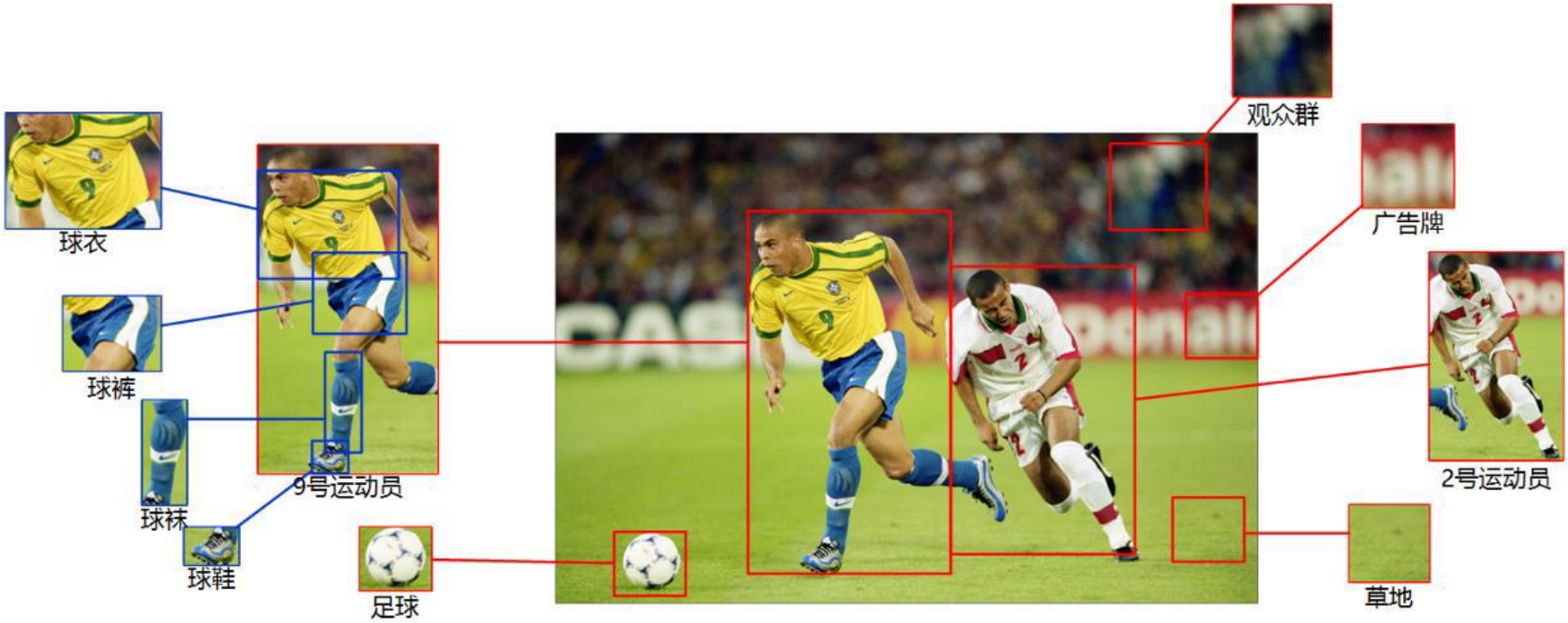


图1-1 视觉内容多种层次的语义理解

视觉内容的上下文语义

上下文语义的重要性！

单纯的对象，无法准确判断图像的所有信息

(A) 缺少上下文信息

(B) 包含上下文信息

国内外研究现状

- 基于上下文的特征表达
  - Gist全局特征(*IJCV2003*)、星座模型(CVPR2000,CVPR2003)、词袋模型(CVPR2006,CVPR2010)、空间金字塔模型(BMVC2006,IJCV2007)、部件模型(TPAMI2010)，
  - 基于CNN的全局特征(NIPS2012)、基于目标检测的局部特征(ECCV2014)
- 层次化分析方法 （空间金字塔策略、粗到细策略）
  - Laptev, Ivan. Improvements of Object Detection Using Boosted Histograms. In BMVC. 2006.
  - He, Kaiming; Zhang, Xiangyu; Ren, Shaoqing, et al. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. *TPAMI*. 2015.
  - Deng, Jia; Berg, Alexander C.; Li, Fei Fei. Hierarchical semantic indexing for large scale image retrieval. In *CVPR*. 2011.

国内外研究现状

- 多上下文建模分析方法(全局-全局、局部-局部，全局-局部)
  - ILSVRC (*VGGnet*, *overfeat*, *GoogLenet*, *ResNet*)
  - Ciresan, Dan C.; Meier, Ueli; Schmidhuber, J. U. Rgen. Multi-column deep neural networks for image classification. In *CVPR*. 2012
  - Felzenszwalb, Pedro F.; Girshick, Ross B.; Mcallester, David A., et al. Object Detection with Discriminatively Trained Part-Based Models. *TPAMI*. 2010.
  - Zhang, Ning; Donahue, Jeff; Girshick, Ross B., et al. Part-Based R-CNNs for Fine-Grained Category Detection. In *ECCV*. 2014.
  - Karpathy, Andrej; Toderici, George; Shetty, Sanketh, et al. Large-Scale Video Classification with Convolutional Neural Networks. In *CVPR*. 2014
  - Zhao, Rui; Ouyang, Wanli; Li, Hongsheng, et al. Saliency detection by multi-context deep learning. In *CVPR*. 2015.

国内外研究现状

- 传统方法（Sparse Coding\Bag of Features）
  - Yang, Jianchao; Yu, Kai; Gong, Yihong, et al. Linear spatial pyramid matching using sparse coding for image classification. In *CVPR*. 2009.
  - Lazebnik, Svetlana; Schmid, Cordelia; Ponce, Jean. Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. In *CVPR*. 2006.
- 深度学习：卷积神经网络（Convolutional Neural Network）
  - Krizhevsky, et al. Hinton, Imagenet classification with deep convolutional neural networks, In NIPS. 2012.
  - Ciresan, et al., Multi-column deep neural networks for image classification. In CVPR. 2012.
  - Simonyan, K. and A. Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv2014.
  - Szegedy, et al., Going Deeper with Convolutions, In CVPR. 2015.
  - He, K., et al., Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification, arXiv 2015.
- 关键技术
  - ReLUs/pReLU(arXiv2015), Dropout(NIPS2012), Maxout(arXiv2013), NIN(arXiv2014)

国内外研究现状

- PASCAL VOC2007/2009/2012
  - Sermanet, Pierre; Eigen, David; Zhang, Xiang, et al. OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks. *arXiv* 2013.
  - R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation, *CVPR* 2014
  - He, Kaiming; Zhang, Xiangyu; Ren, Shaoqing, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition, *PAMI* 2015.
  - Ross Girshick, Fast-RCN, In *ICCV*. 2015
  - Ren, Shaoqing; He, Kaiming; Girshick, Ross B., et al. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *NIPS*. 2015.
  - Redmon, Joseph; Divvala, Santosh Kumar; Girshick, Ross B., et al. You Only Look Once: Unified, Real-Time Object Detection. In *CVPR*. 2016.
  - Liu, Wei; Anguelov, Dragomir; Erhan, Dumitru, et al. SSD: Single Shot MultiBox Detector. In *ECCV*. 2016.
- 区域建议

  RPN(NIPS2015), Objectness (IJCV2013), BING (CVPR2014), Selective Search (IJCV2013), EdgeBoxes (ECCV2014), MCG (CVPR2014), DeepBox (arXiv2015)

国内外研究现状

- 图像分割
  - Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. *arXiv* 2014.
  - Dai, Jifeng; He, Kaiming; Sun, Jian. Convolutional feature masking for joint object and stuff segmentation. In *CVPR*. 2015
  - Chen, Liang Chieh; Papandreou, George; Kokkinos, et al. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *CoRR*. 2016.
- 场景解析
  - Zhou, Bolei; Khosla, Aditya; Lapedriza, A. Gata, et al. Places: An Image Database for Deep Scene Understanding. *CoRR*. 2016.
  - Zhou, Bolei; Zhao, Hang; Puig, Xavier, et al. Semantic Understanding of Scenes through the ADE20K Dataset. *CoRR*. 2016.
- 人脸解析
  - Guo, Dong; Sim, Terence. Digital face makeup by example. In *CVPR*. 2009.
  - Tong, Wai Shun; Tang, Chi Keung; Brown, Michael S., et al. Example-Based Cosmetic Transfer. In *CGA*. 2007.

国内外研究现状

- 传统方法
  - SIFT/BOW/VLAD/LSH/Sparse-coding/…
- 深度学习
  - Artem Babenko, et al., Neural Codes for Image Retrieval, In *ECCV*. 2014
  - Ali Sharif Razavian, et al,. CNN Features Off-the-Shelf: An Astounding Baseline for Recognition, In *CVPR* 2014
  - Wan, Ji; Wang, Dayong; Hoi, Steven Chu Hong, et al. Deep Learning for Content-Based Image Retrieval: A Comprehensive Study. In *ACMMM*. 2014.
  - Gong, Y., et al., Multi-scale Orderless Pooling of Deep Convolutional Activation Features. In *ECCV*. 2014.
  - Hanjiang Lai, Yan Pan, Ye Liu, Shuicheng Yan. Simultaneous Feature Learning and Hash Coding with Deep Neural Networks, In *CVPR*. 2015
  - Exploiting Local Features from Deep Networks for Image Retrieval, In *CVPRW*. 2015
  - Deep Semantic Ranking Based Hashing for Multi-Label Image Retrieval, In *CVPR*. 2015

Part
02

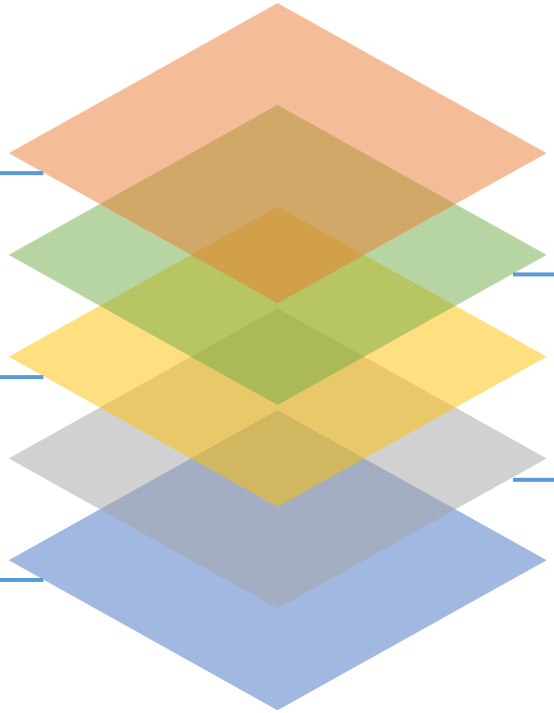
研究内容与难点

/ 挑战问题

/ 研究内容

挑战问题

计算机视觉任务中的挑战问题



大数据环境中搜索空间太大引起的效率问题

场景解析中额外背景类造成的误判问题

大数据环境中样本多样性问题

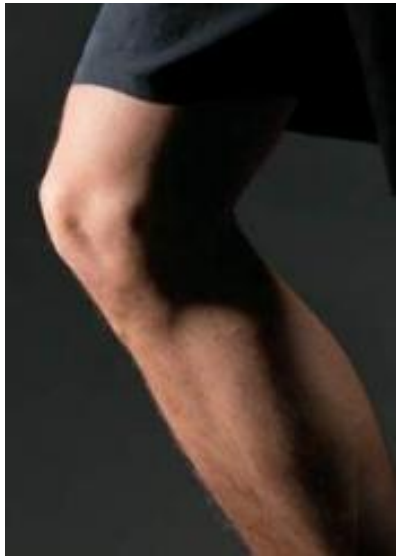
上下文融合时语义保持困难的问题

场景中难目标识别问题

挑战问题

1 大数据环境中样本多样性问题





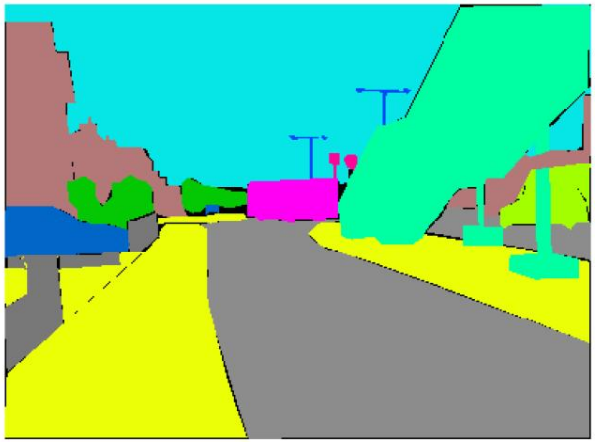



挑战问题

2 场景中难目标识别问题

挑战问题

3 场景解析中额外背景类造成的误判问题



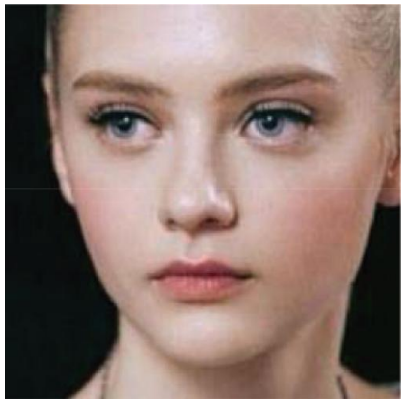
语义/实例分割

场景解析

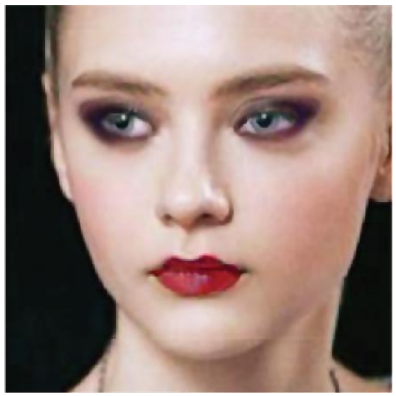
黑洞现象

挑战问题

4 上下文融合时语义保持困难的问题



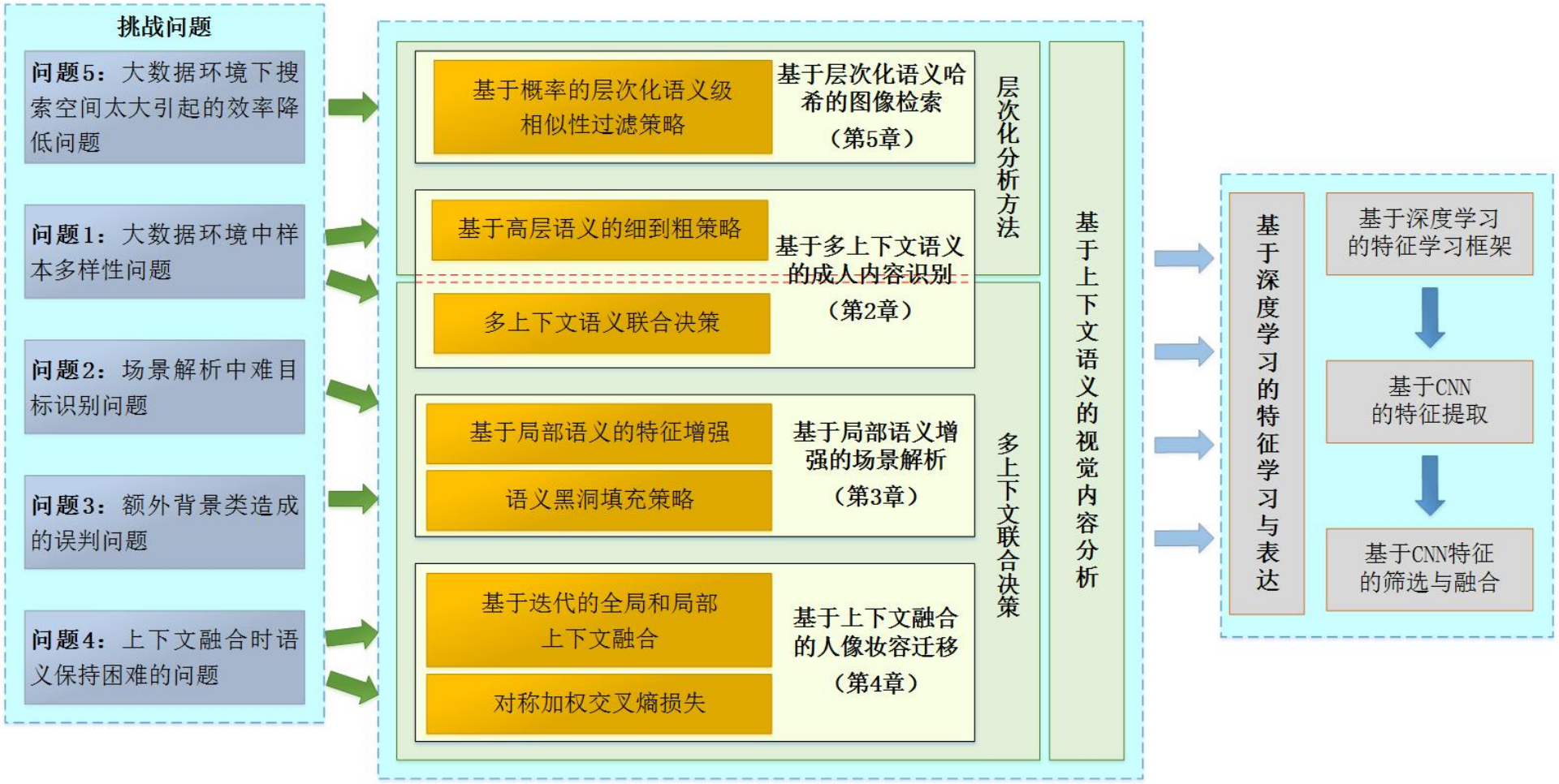
如何保持视觉上的自然？



挑战问题

5 大数据环境中搜索空间太大引起的效率问题

研究内容


挑战问题
问题5：大数据环境下搜索空间太大引起的效率降低问题
问题1：大数据环境中样本多样性问题
问题2：场景解析中难目标识别问题
问题3：额外背景类造成的误判问题
问题4：上下文融合时语义保持困难的问题
基于概率的层次化语义级相似性过滤策略
基于层次化语义哈希的图像检索（第5章）
层次化分析方法
基于高层语义的细到粗策略
基于多上下文语义的成人内容识别（第2章）
多上下文语义联合决策
基于局部语义的特征增强
语义黑洞填充策略
基于局部语义增强的场景解析（第3章）
多上下文联合决策
基于迭代的全局和局部上下文融合
对称加权交叉熵损失
基于上下文融合的人像妆容迁移（第4章）
基于上下文语义的视觉内容分析
基于深度学习的特征学习与表达
基于深度学习的特征学习框架
基于CNN的特征提取
基于CNN特征的筛选与融合

Part
03

关键技术与进展

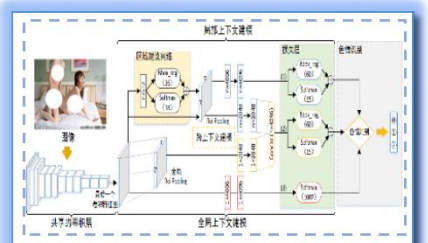
/ 基于多上下文语义的成人内容识别

/ 基于局部语义增强的场景解析

/ 基于上下文融合的人像妆容迁移

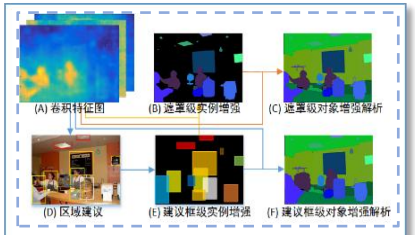
/ 基于层次化语义哈希的图像检索

关键技术与进展



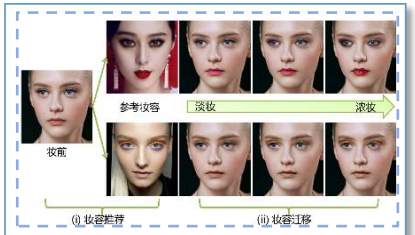
基于多上下文语义的成人内容识别

成人内容识别是图像内容识别的一个具体应用，它是互联网应用健康发展的基础工作。



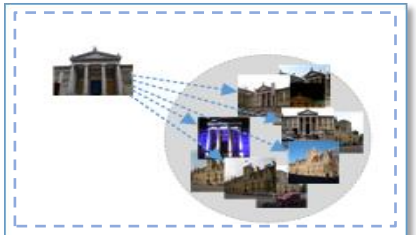
基于局部语义增强的场景解析

场景解析是计算机视觉的一个重要任务，它在自动驾驶、卫星遥感图像分析、图像搜索、机器人导航、室内三维建模等多个领域都有较广泛的潜在应用。



基于上下文融合的人像妆容迁移

化妆已成为目前大众日常生活的一部分，这给基于人脸的身份验证系统带来了巨大挑战。研究一种自动化妆系统对理解化妆后人脸身份验证具有重要意义。



基于层次化语义哈希的图像检索

基于内容的图像检索可以帮助我们从海量的数据中找到符合我们需求的样本。它被广泛应用到搜索引擎、电子商务以及各种Web 2.0的教学系统中。

基于多上下文语义的成人内容识别

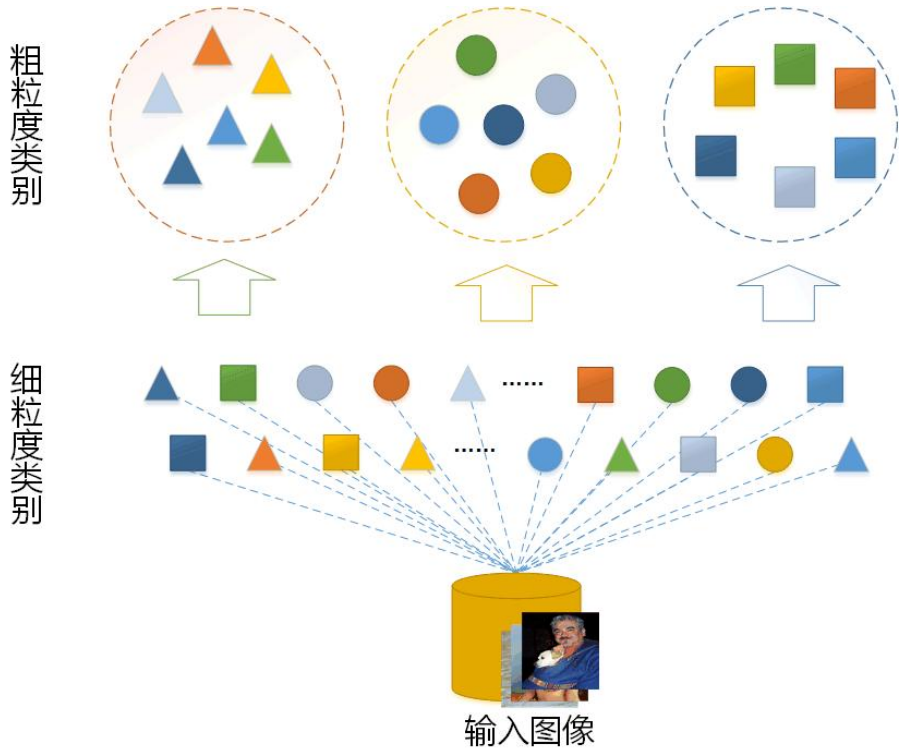
问题提出

成人内容识别是图像识别的一个具体应用，由于任务特殊性它涉及到**图像分类**和**对象检测**两个方面的关键技术。在处理成人内容识别时，面临两个关键挑战：

- **数据规模大**：数百万，甚至上千万
- **多样性**：图像内容、图像尺度、分辨率、类型等

基于多上下文语义的成人内容识别

解决方案一：细到粗策略



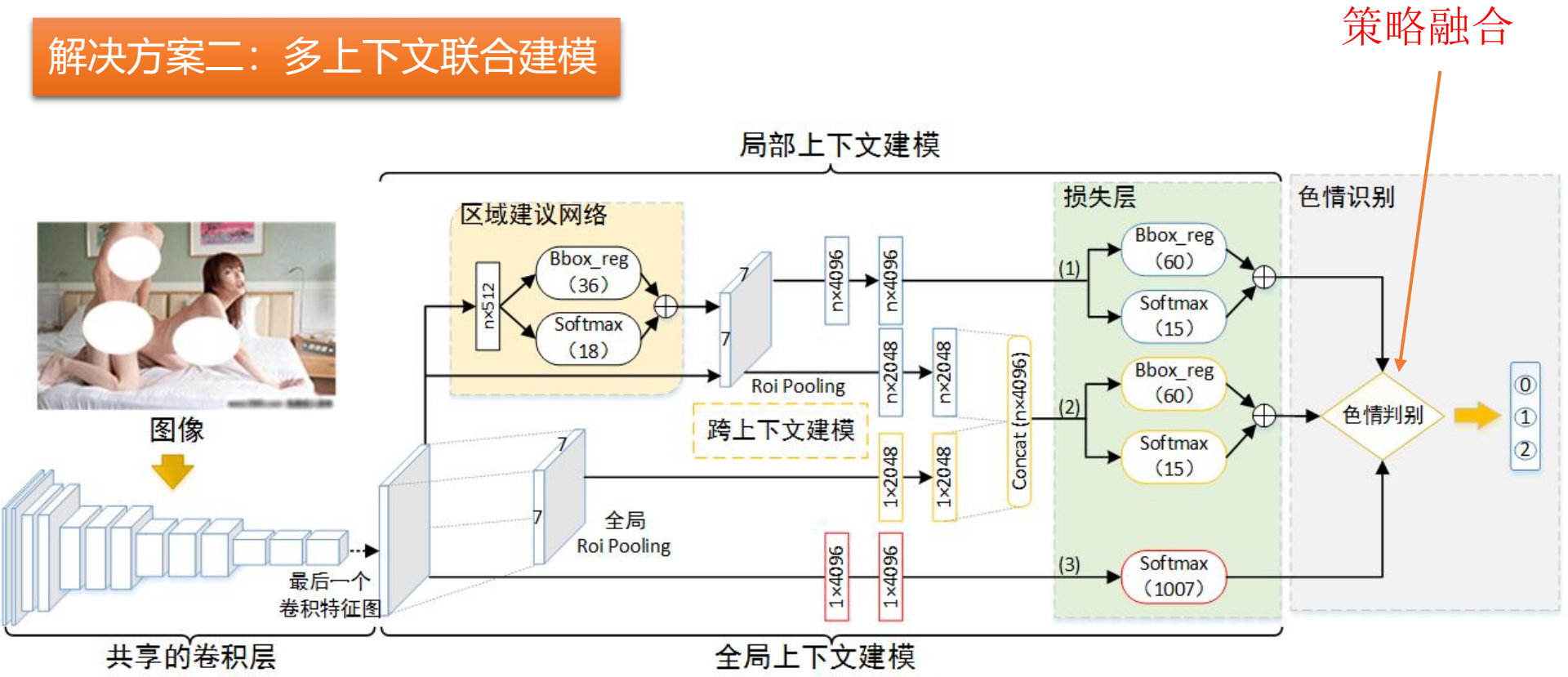
$$y_f = g(I)$$

$$y_c = T(y_f)$$

其中，$T: y_f \mapsto y_c$ 是一个映射函数，符号“↦”表示直接将细粒度类别y_f映射成为粗粒度类别y_c。函数 $g(I)$ 是深度神经网络前向推理的结果。

基于多上下文语义的成人内容识别

解决方案二：多上下文联合建模



$$\mathcal{F}_{DMCN} = max\left((1 - w_1 - w_2) \cdot \widetilde{\mathcal{F}}_{global},\ w_1 \cdot \widetilde{\mathcal{F}}_{local},\ w_2 \widetilde{\mathcal{F}}_{cross}\right)$$

$$\widetilde{\mathcal{F}}_{global} = \phi(\mathcal{F}_{local}),\ \widetilde{\mathcal{F}}_{local} = \psi\left(\phi(\mathcal{F}_{local}, t_1)\right),\ \widetilde{\mathcal{F}}_{cross} = \psi\left(\phi(\mathcal{F}_{cross}, t_2)\right)$$

基于多上下文语义的成人内容识别

实验 – 训练过程

算法 2-1：深度多上下文网络（DMCNet）训练过程

步骤 1：在 *Sensitive* 数据集上使用 *Imagenet* 预训练模型[3]训练一个新的深度模型，该模型作为基准 Baseline 模型和全局上下文模型，同时使用该模型去初始化*步骤 2* 和*步骤 3* 中的卷积层。我们称卷积层为共享卷积层。

步骤 2：使用*步骤 1* 中的网络作为预训练模型，训练区域建议网络，其中共享卷积层的参数固定不变。

步骤 3：使用*步骤 2* 中训练好的区域建议网络作为建议区域生成器，训练一个可表征局部上下文信息的对象检测网络。该检测网络仍然使用*步骤 1* 中训练好的基准模型作为初始权重。共享卷积层和区域建议网络的参数固定不变。

步骤 4：保持共享卷积层，区域建议网络的参数固定不变，组合局部上下文特征和全局上下文特征，并微调新的混合层用于生成跨上下文信息。建议区域仍然由*步骤 2* 中生成的网络产生。

步骤 5：将*步骤 2*、3 和 4 中训练好的模型组合起来形成统一的多上下文框架，作为最终的 DMCN 模型。

基于多上下文语义的成人内容识别

实验结果：细到粗策略的多模型评价

表 2-2 细到粗策略在全局上下文建模中的性能评估。实验基于三种流行的深度网络结构 Alexnet[23]，VGG16[3]和 GoogLeNet[4]，并在 *Sensitive* 数据集上完成。

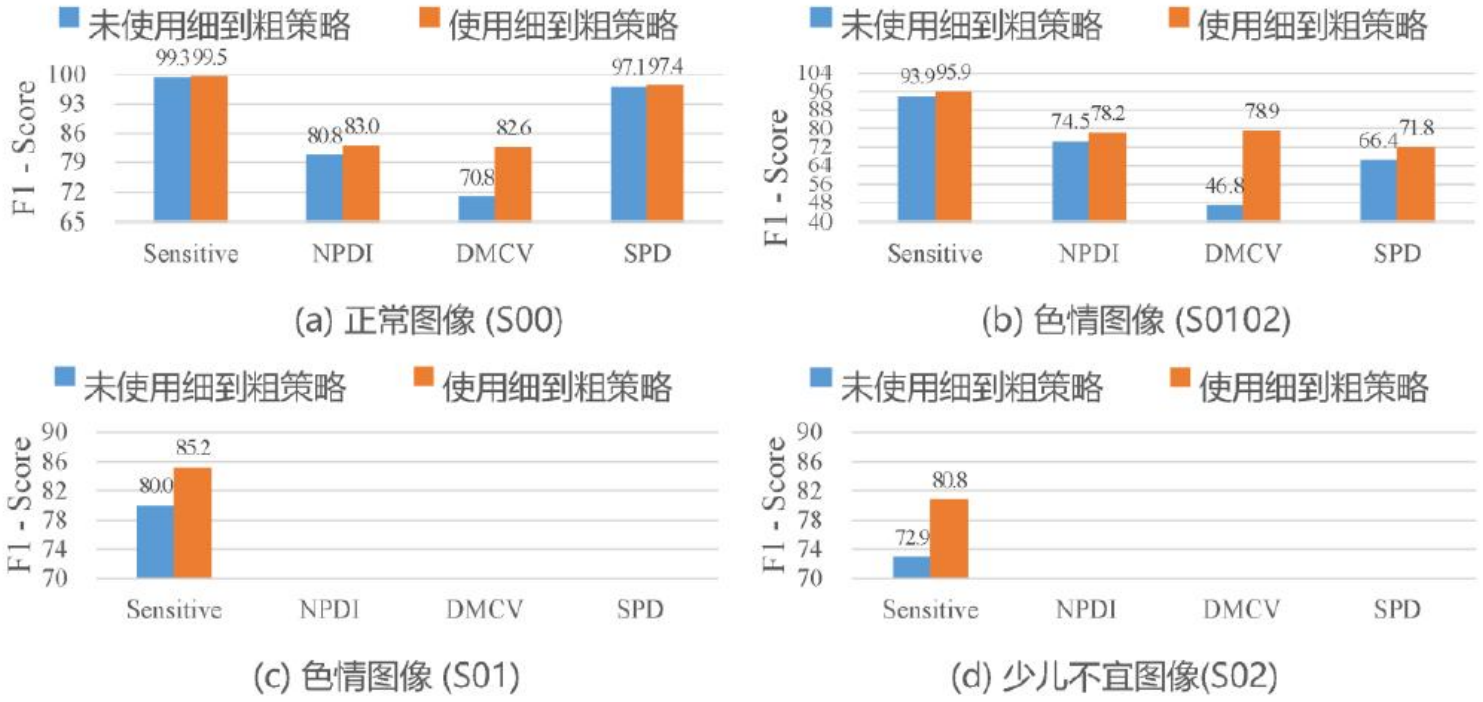
		S00	S01	S02	S0102	时间消耗（毫秒）
Baseline	Alexnet	99.0	73.23	57.9	91.8	32
	VGG16	99.3	80.0	72.9	93.9	145.4
	GoogLeNet	98.6	67.0	72.8	92.3	101.2
Global-Context	Alexnet	99.3	79.0	74.5	94.1	47
	VGG16	**99.5**	**85.2**	**80.8**	**95.9**	158.5
	GoogLeNet	99.4	81.7	77.5	94.8	117.6

注：S00：正常图像，S01：成人图像，S02：少儿不宜图像，S0102：成人图像（S01+S02）

[3] Simonyan, Karen; Zisserman, Andrew. Very Deep Convolutional Networks for Large-Scale Image Recognition. *CoRR*. 2014.
[4] Szegedy, Christian; Liu, Wei; Jia, Yangqing, et al. Going deeper with convolutions. In *CVPR*. 2015.
[23] Krizhevsky, Alex; Sutskever, Ilya; Hinton, Geoffrey E. ImageNet classification with deep convolutional neural networks. In *NIPS*. 2012.

基于多上下文语义的成人内容识别

实验结果：细到粗策略的多数据集评价



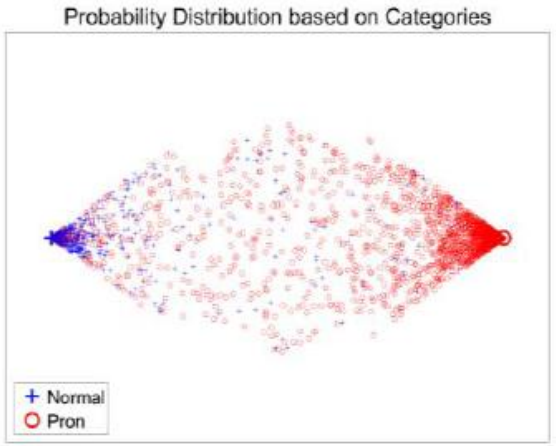
(a) 正常图像 (S00)　(b) 色情图像 (S0102)

(c) 色情图像 (S01)　(d) 少儿不宜图像(S02)

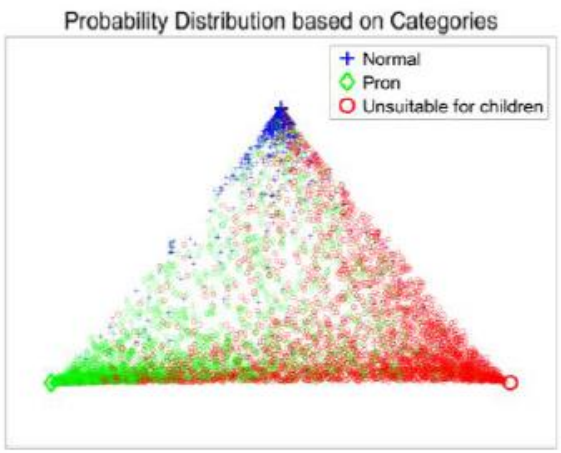
图 2-5 细到粗策略在四个成人数据集上 F1-Score 评估结果。(a-b) 分别显示了正常图像（S00）和成人图像（S0102）在 4 个数据集上的性能。其中，*Sensitive* 数据集中的类别 S0102 是类别 S01 和 S02 的组合。(c-d)反映的是 *Sensitive* 数据集上类别成人图像 S01 和少儿不宜图像（S02）的性能。

基于多上下文语义的成人内容识别

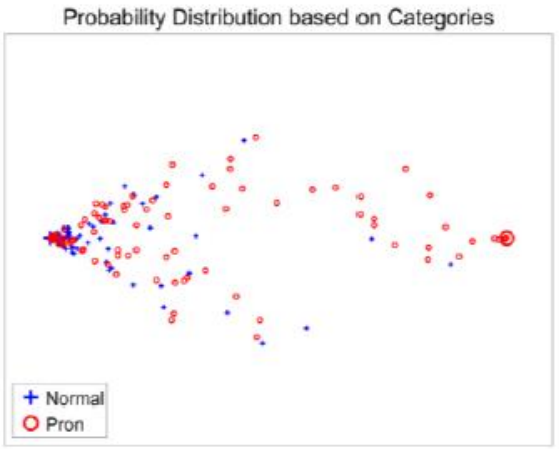
实验结果：细到粗策略的多数据集评价



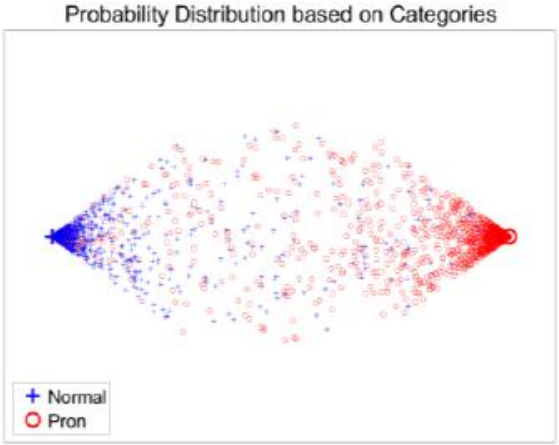
（a）*Sensitive*, *L2*，未使用细到粗策略



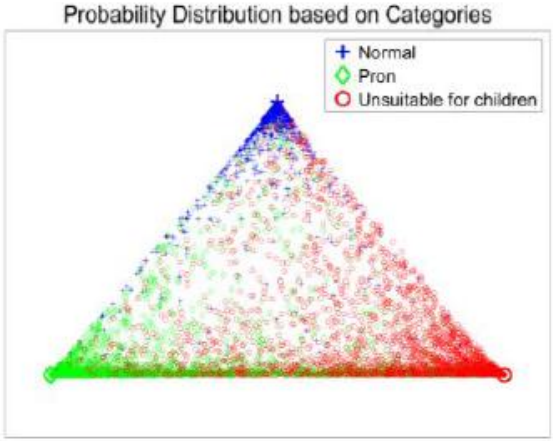
（b）*Sensitive*, *L3*，未使用细到粗策略



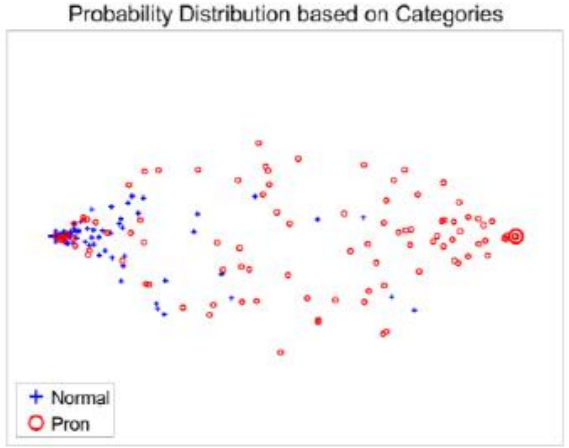
（c）*DMCV*, *L2*，未使用细到粗策略



（d）*Sensitive*, *L2*，使用细到粗策略



（e）*Sensitive*, *L3*,使用细到粗策略



（f）*DMCV*, *L2*，使用细到粗策略

基于多上下文语义的成人内容识别

实验结果：多上下文联合建模

表 2-3 多上下文建模在四个数据集上的性能评估

		F1-Score			
		S00	S01	S02	S01+S02
Sensitive	Global-Content	**99.5**	85.2	80.8	**95.9**
	Local-Content (t=0.99)	94.6	**19.8**	9.8	15.2
	Cross-Content (t=0.99)	95.5	20.0	**9.8**	15.5
	Multi-Content (w=0.47)	**99.5**	85.3	80.9	**95.9**
NPDI	Global-Content	83.0			78.2
	Local-Content (t=0.3)	77.7			79.2
	Cross-Content (t=0.3)	78.6			80.1
	Multi-Content (w=0.32)	**85.2**			**85.3**
DMCV	Global-Content	**82.6**			78.9
	Local-Content (t=0.2)	66.0			70.4
	Cross-Content (t=0.2)	66.6			71.1
	Multi-Content (w=0.27)	82.3			**80.4**
SPD	Global-Content	97.4			71.8
	Local-Content (t=0.6)	96.6			63.1
	Cross-Content (t=0.6)	**97.5**			63.7
	Multi-Content (w=0.48)	**97.5**			**74.7**

注：S00：正常图像，S01：成人图像，S02 少儿不宜图像，S0102：成人图像（S01+S02）

基于多上下文语义的成人内容识别

实验结果：多上下文联合建模

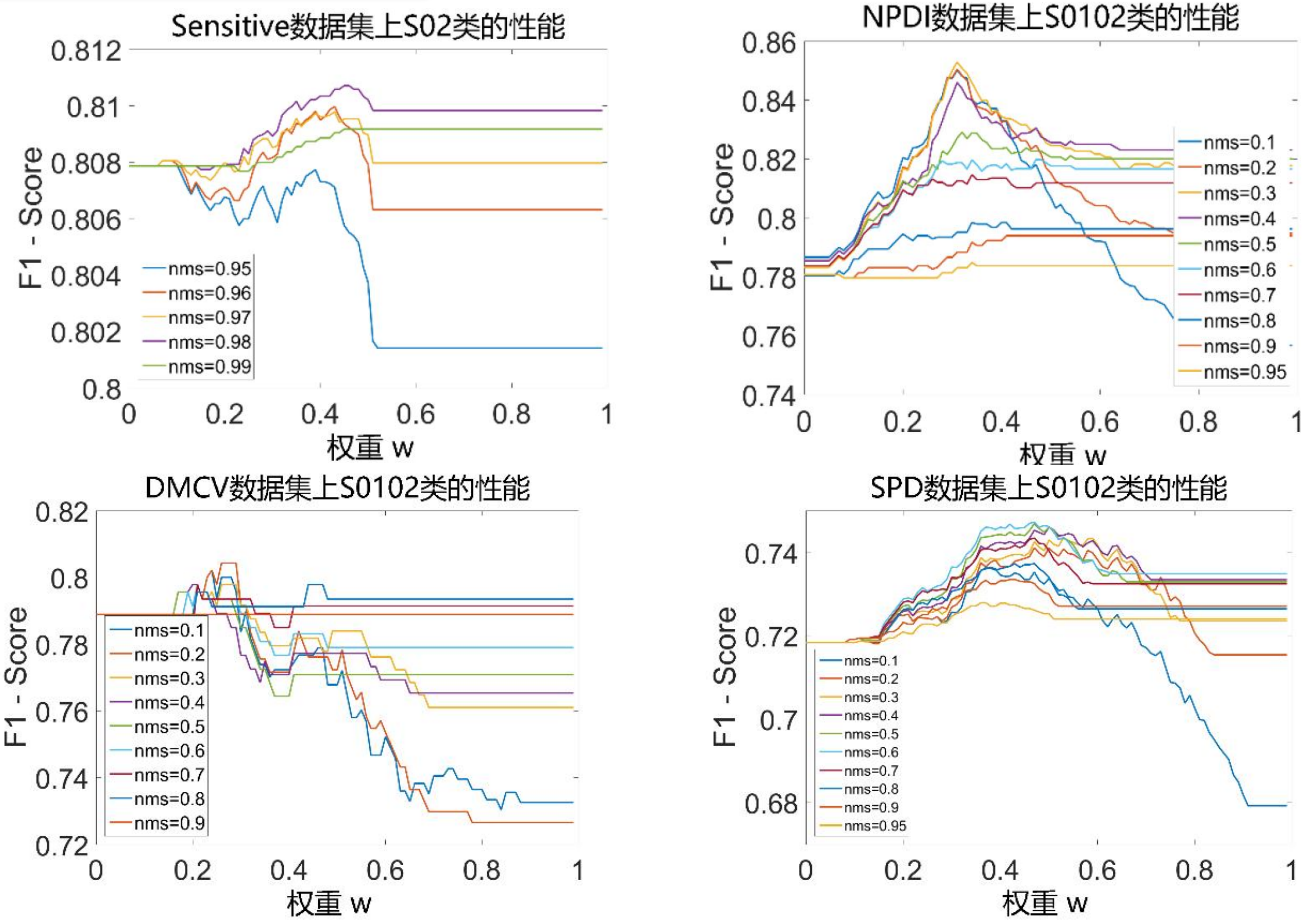


图 2-7 多上下文建模在四个数据集上的评估结果。不同的图使用不同的融合权重w和不同的建议框选择阈值t。

基于多上下文语义的成人内容识别

实验结果：多上下文联合建模

表 2-4 DMCNet 模型与三个对比模型在 *Sensitive* 数据集上的性能评估结果

方法	召回率			准确率			F1-Score			精确度	
	S01	S02	S0102	S01	S02	S0102	S01	S02	S0102	L2	L3
AGNet	80.2	56.4	86.0	55.9	**92.9**	**98.8**	65.9	70.2	92.0	98.3	96.3
Incremental Learning	58.7	26.5	68.5	14.3	30.0	35.8	23.0	28.1	47.0	75.1	72.3
Baseline-VGG16	83.4	62.6	89.2	76.9	87.4	99.1	80.0	72.9	93.9	98.7	96.9
DMCNet	**88.0**	**73.9**	**93.4**	**82.6**	89.3	98.7	**85.3**	**80.9**	**95.9**	**99.1**	**97.8**

表 2-5 DMCNet 模型与三个对比模型在三个数据集上的泛化性能评估结果

数据集	方法	召回率		准确率		F1-Score		精确度
		普通	成人	普通	成人	普通	成人	
NPDI	AGNet	88.3	69.8	74.7	**85.6**	80.8	76.9	79.0
	Incremental Learning	76.2	51.6	71.8	57.2	73.9	54.3	63.9
	Baseline-VGG16	**92.2**	64.0	71.9	89.2	80.8	74.5	78.1
	DMCNet	85.0	**85.5**	**85.4**	85.1	**85.2**	**85.3**	**85.3**
DMCV	AGNet	87.0	65.7	71.9	83.3	78.7	73.5	76.4
	Incremental Learning	23.0	86.9	63.9	52.8	33.8	65.7	54.8
	Baseline-VGG16	**91.0**	33.3	58.0	78.6	70.8	46.8	62.3
	DMCNet	86.0	**76.8**	**78.9**	**84.4**	**82.3**	**80.4**	**81.4**
SPD	AGNet	99.8	49.2	94.4	96.9	97.0	65.8	94.4
	Incremental Learning	78.9	40.3	91.8	18.4	84.9	25.2	74.8
	Baseline-VGG16	**99.9**	49.9	94.4	**99.3**	97.1	66.4	94.6
	DMCNet	99.3	**63.1**	**95.8**	91.7	**97.5**	**74.7**	**95.4**

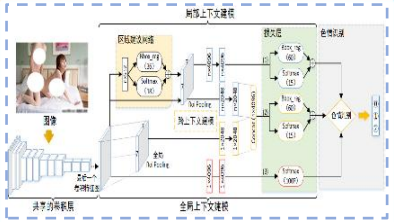
基于多上下文语义的成人内容识别

本章贡献

在本章中，我们提出一种基于深度学习的多上下文框架（Deep Multi-Context Network，DMCNet）用于成人图像和视频的识别。本章工作主要有以下几点贡献：

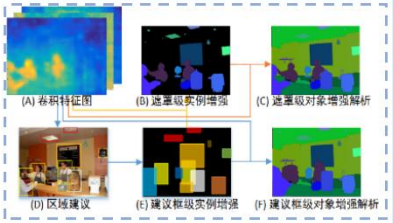
1. 设计了一种基于深度卷积神经网络的多上下文体系结构
2. 提出了一种精心设计的多上下文融合策略
3. 设计了一个基于高级语义的细到粗策略
4. 模块化的设计允许通过更新组件来提升整体系统性能
5. 收集了三个专门用于成人内容识别的具有不同特色的数据集

关键技术与进展



基于多上下文语义的成人内容识别

成人内容识别是图像内容识别的一个具体应用，它是互联网应用健康发展的基础工作。

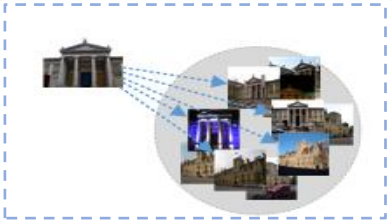


基于局部语义增强的场景解析

场景解析是计算机视觉的一个重要任务，它在自动驾驶、卫星遥感图像分析、图像搜索、机器人导航、室内三维建模等多个领域都有较广泛的潜在应用。

基于上下文融合的人像妆容迁移

化妆已成为目前大众日常生活的一部分，这给基于人脸的身份验证系统带来了巨大挑战。研究一种自动化妆系统对理解化妆后人脸身份验证具有重要意义。



基于层次化语义哈希的图像检索

基于内容的图像检索可以帮助我们从海量的数据中找到符合我们需求的样本。它被广泛应用到搜索引擎、电子商务以及各种Web 2.0的教学系统中。

基于局部语义增强的场景解析

问题提出

场景解析是计算机视觉的重要任务，它与语义分割、实例分割有很大的相似性，但又不完全相同。它**不仅关注对象**，同时也**关注背景区域**。换句话说，场景解析需要处理样本的每一个像素。

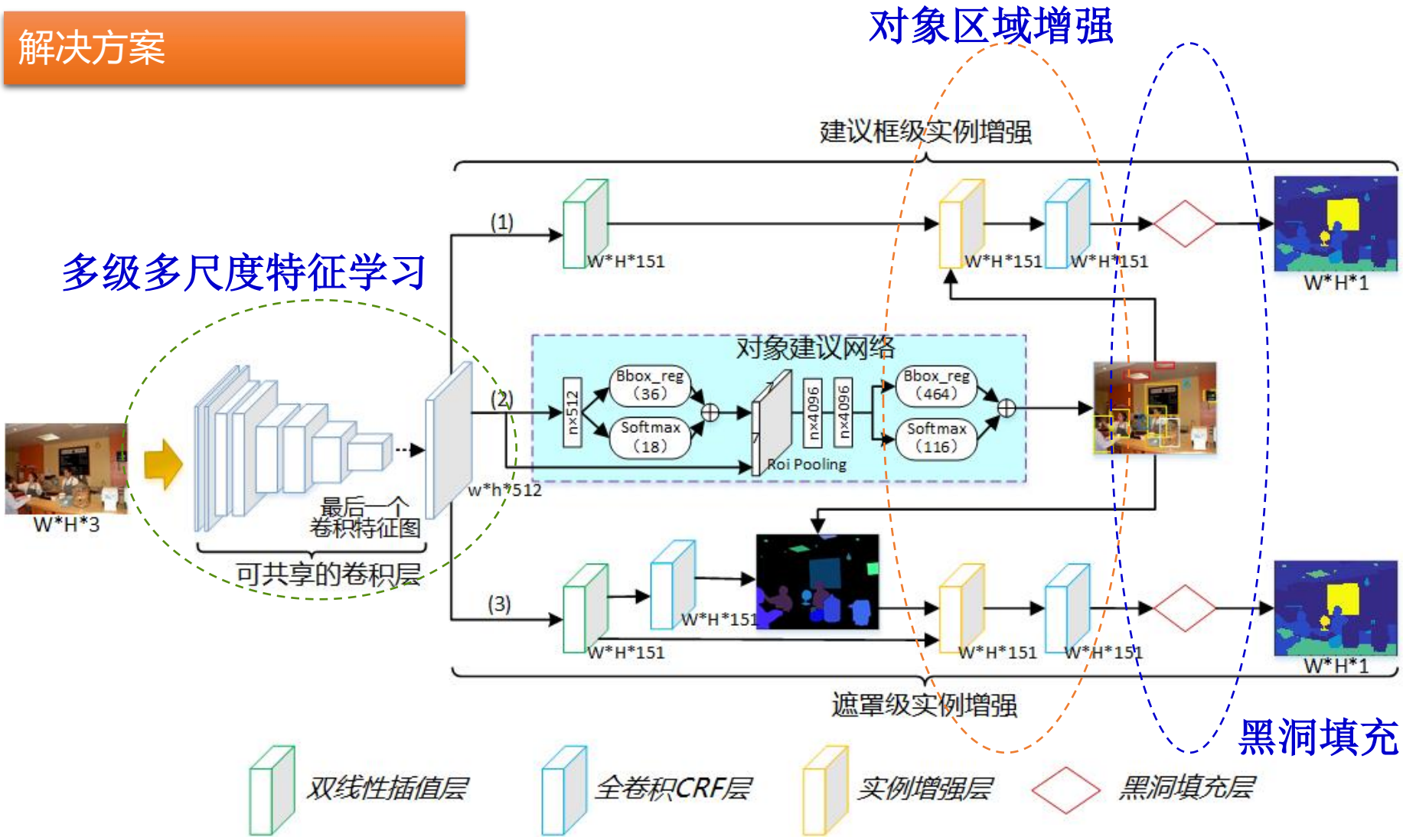
场景解析在自动驾驶、互联网视频搜索、卫星遥感图像分析、图像搜索、机器人导航、室内三维建模等多个领域都有较广泛的潜在应用。

难目标识别问题 → 对象区域增强

额外背景类造成的误判问题 → 黑洞填充

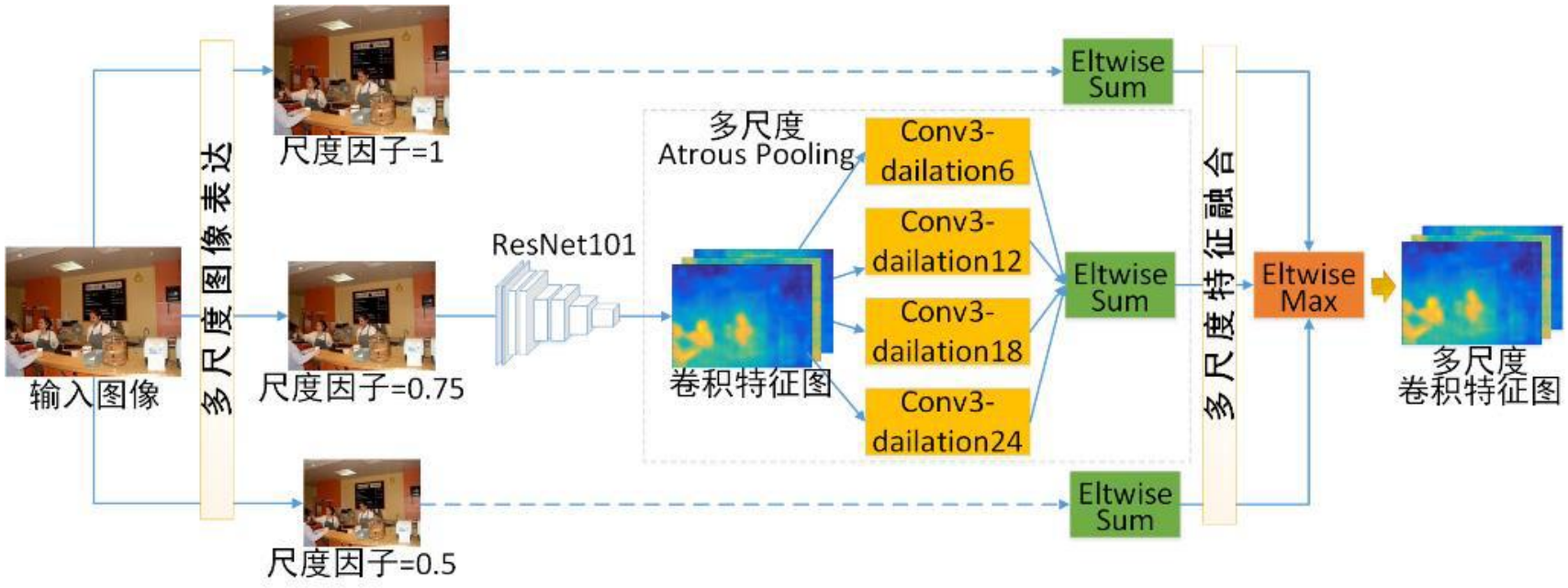
基于局部语义增强的场景解析

解决方案



基于局部语义增强的场景解析

解决方案 – 多级多尺度特征学习



$$\mathcal{F}_M = \max_{m\in[1,\ldots,M]} \sum_{n=1}^{N} \mathcal{F}_{(m,n)}$$

基于局部语义增强的场景解析

解决方案 – 对象区域增强

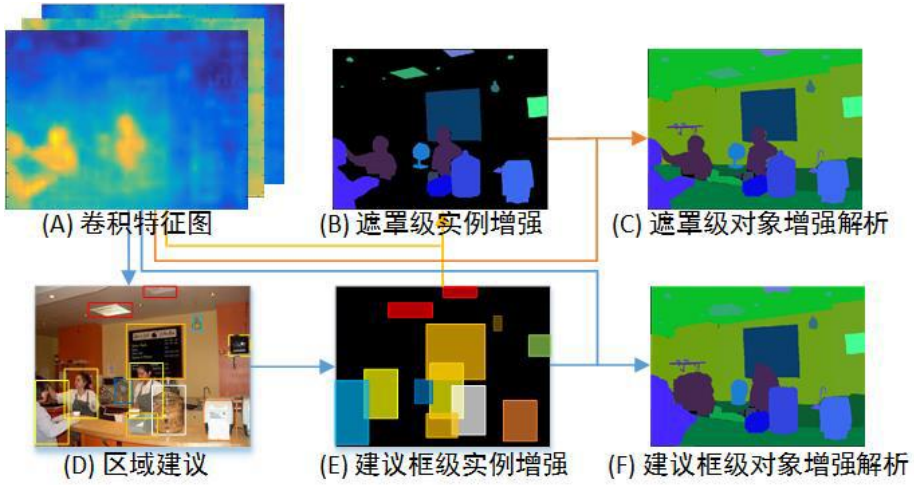


图3-6 对象增强的流程图

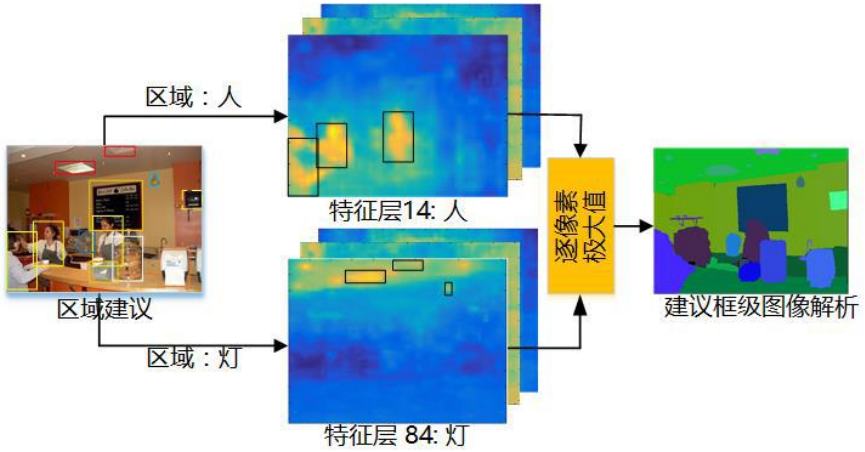
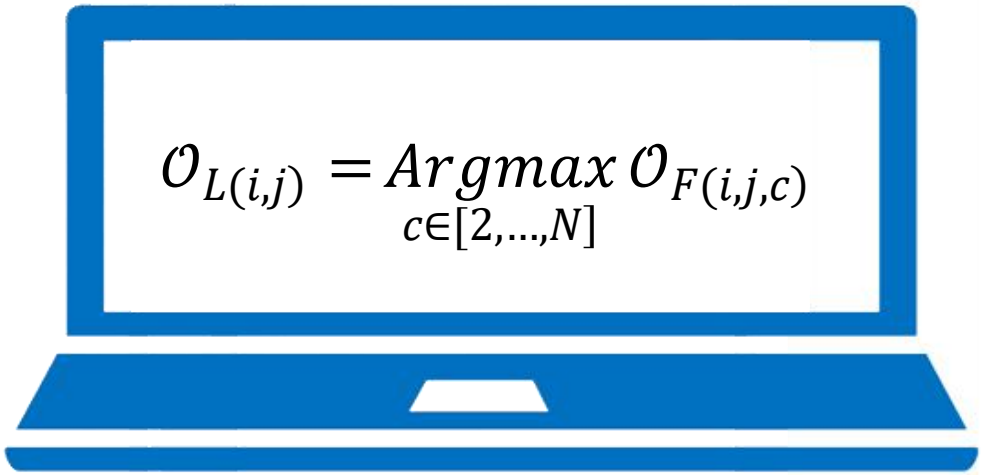


图3-1 对象区域增强原理示意图

- 建议框级实例 $B_i(W,H,c_i)=\begin{cases}1 & p\in R_i(t_i)\\ 0 & \text{otherwise}\end{cases}$
- 遮罩级实例 $M_i(W,H,c_i)=\begin{cases}1 & p\in F(W,H,c_i)\cdot R_i(t_i,c_i)>t\\ 0 & \text{otherwise}\end{cases}$
- 基于实例的增强 $F_{ME}(W,H,C)=F(W,H,c_i)\cdot M\left(W,H,c_j\right)\times w\times p_i\times c^*$

基于局部语义增强的场景解析

解决方案 – 黑洞填充策略



基于局部语义增强的场景解析

实验 – 训练过程

算法 4-1：对象区域增强网络（OENet）训练过程

步骤 1：在 *MSCOCO* 数据集[90]上预训练一个 ResNet101[21, 22]深度卷积神经网络模型。

步骤 2：使用*步骤 1* 中预训练的模型作为初始化网络，并用交叉熵损失训练多级多尺度特征提取网络（FEN）。这个网络用于初始化其他网络，同时作为基准模型进行性能评估。

步骤 3：依据文献[118]的建议，采用交叉验证方式搜索全连接 CRF 参数。

步骤 4：使用*步骤 2* 中预训练模型进行初始化，训练对象建议网络 OPN。在该步骤中，我们首先训练 RPN 子网络，然后使用 RPN 生成的建议框训练检测网络。整个过程中，共享卷积层始终保持固定不定。

步骤 5：按照图 3-3 的结构输出统一的 OENet 模型，该模型通过集成第 2 步、第 4 步训练生成的模型，CRF 模块、区域增强模块和黑洞填充策略模块得到。

[118] Chen, Liang Chieh; Papandreou, George; Kokkinos, Iasonas, et al. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *CoRR*. 2016.

基于局部语义增强的场景解析

实验结果

表 3-1 OENet 各种策略在 *SceneParsing150* 上的性能评估

多级多尺度	建议框级增强	遮罩级增强	黑洞填充	平均 IoU	像素准确率
				30.9	74.0%
√				35.0	75.5%
√	√			35.6	75.1%
√	√	√		36.5	75.7%
√	√	√	√	38.4	77.9%

表 3-3 区域建议在 *SceneParsing150* 上的性能评估

方法	检测精度	平均 IoU	像素准确率
Groundtruth	100%	47.8	80.3%
Faster RCNN[101]	84.4%	38.4	77.9%
Region FCN[175]	82.3%	38.0	77.7%

基于局部语义增强的场景解析

实验结果

表3-4 各种算法在*SceneParsing150*上的性能对比

方法	平均 IoU	像素准确率
SegNet[180]	21.6	71.0%
Cascade-SegNet[131]	27.5	71.8%
FCN8s[174]	29.4	71.3%
DilatedNet [179]	32.3	73.6%
DeepLabv2[118]	34.3	75.3%
Cascade-DilatedNet[131]	34.9	74.5%
Baseline	30.9	74.0%
OENet	38.4	77.9%

表3-5 各种算法在*Cityscapes*验证集上的性能对比

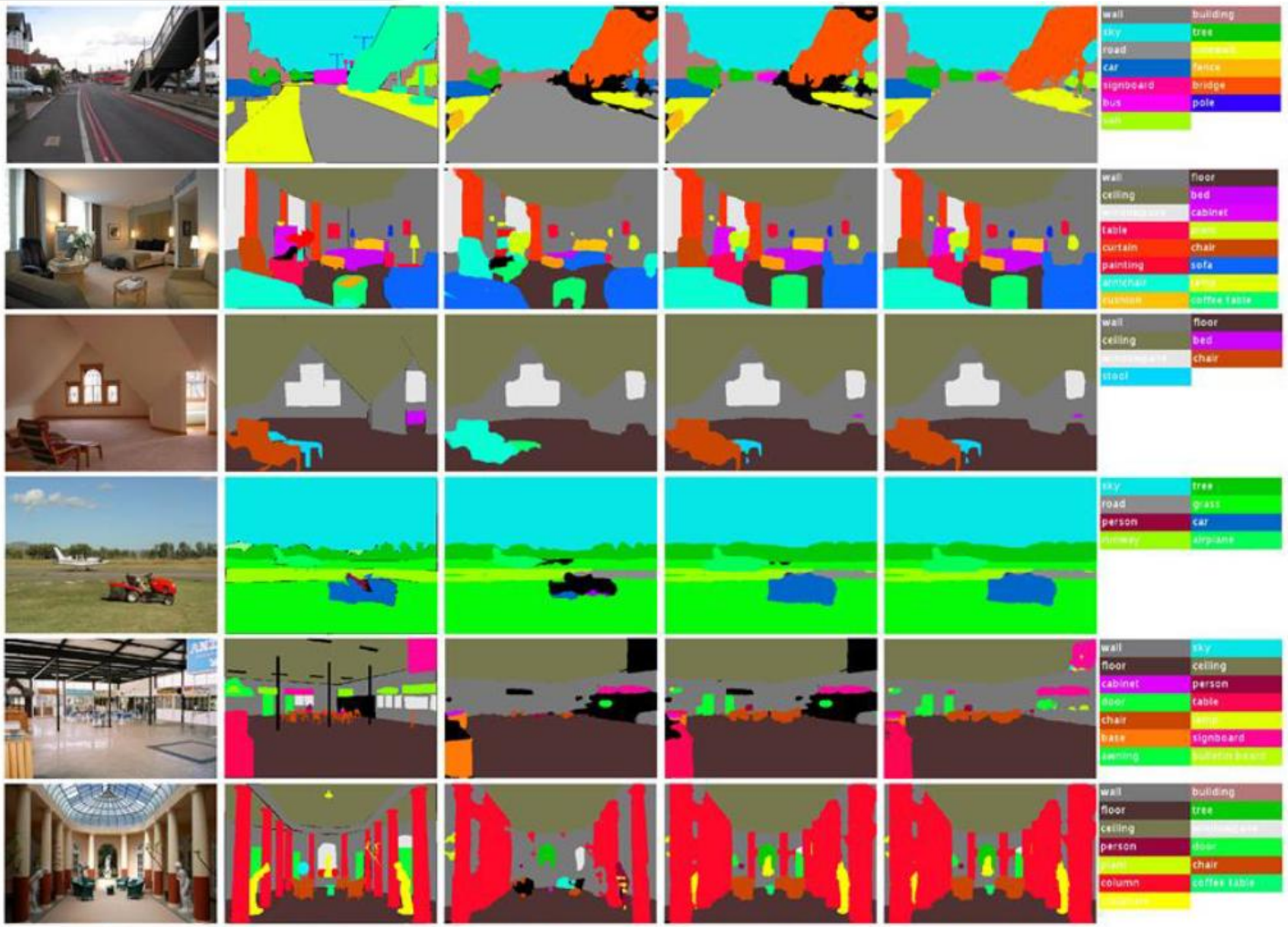
方法	平均 IoU
VGG16	
DeepLabv2-VGG16[118]	62.9
FCN[174]	63.4
Pixel-level Encoding	64.3
DPN[125]	66.8
DiatedNet[179]	67.1
Adelaide[128]	68.6
ResNet-101	
DeepLabv2-Resnet101[118]	71.4
Baseline	69.3
Baseline-高分辨率	71.3
OENet	70.0
OENet-高分辨率	73.1

注：高分辨率：模型训练测试都基于 705×705 的高分辨率图像片，未进行尺度缩放

基于局部语义增强的场景解析

实验结果

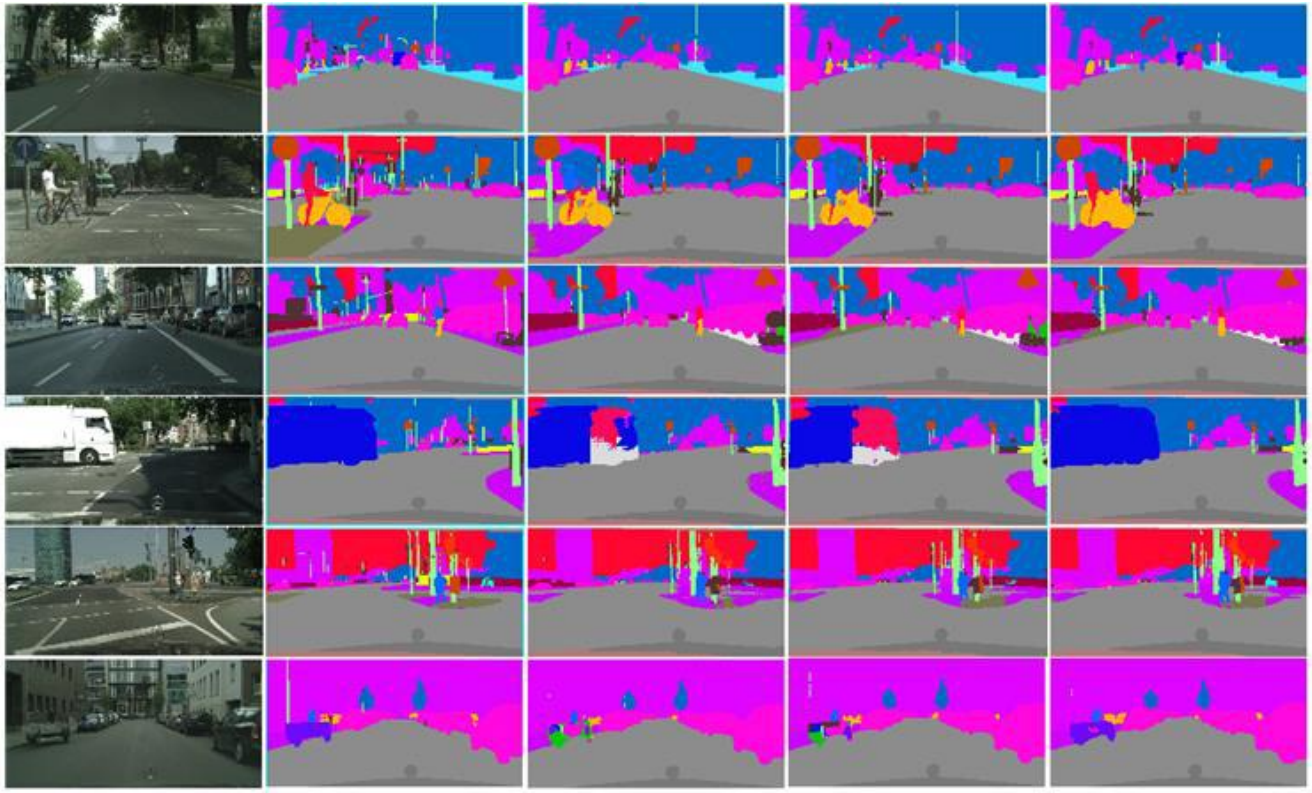
图3-7 *SceneParsing150*数据集上场景解析结果示意图



基于局部语义增强的场景解析

实验结果

图 3-9 *Cityscape*数据集上场景解析结果示意图



基于局部语义增强的场景解析

实验结果 – 失败案例分析

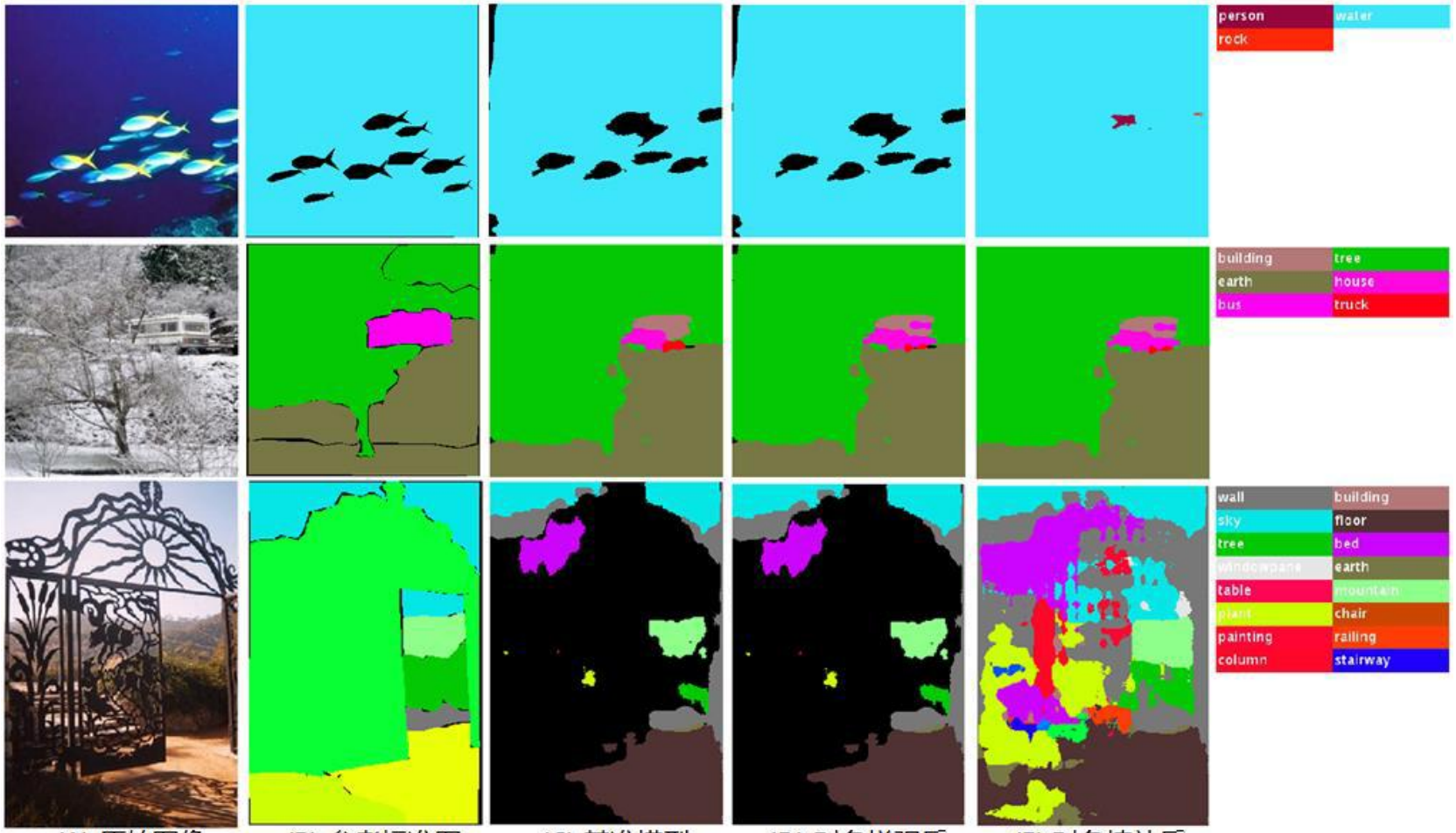


图3-8 *SceneParsing150* 数据集上错误案例示意图

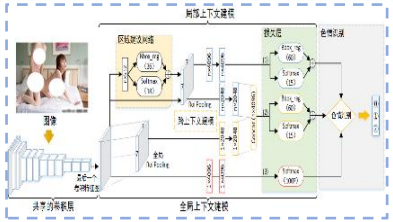
基于局部语义增强的场景解析

本章贡献

在本章中，我们提出了一种利用上下文语义互补性的对象区域增强网络（Objectness Region Enhancement Network，OENet），利用传统的图像分类网络来实现场景分割任务。本章主要有以下三个贡献：

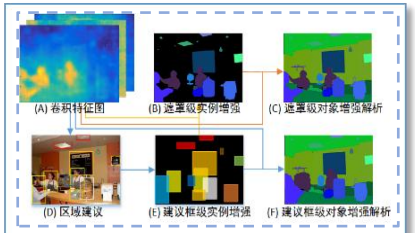
1. 提出了一个**统一的框架**用于处理场景解析任务。受益于**模块化设计**方案，我们提出的算法不仅仅可以通过更换卷积或检测模块来提高整体性能，也可以将对象增强和黑洞填充应用到其他系统以提高系统对对象的解析能力。
2. 提出了一种**基于对象区域增强**的方法用于召回那些在标准的分割网络中无法识别的对象。
3. 提出了一种**黑洞填充的技术**，用于解决那些被错误地分类到不存在的额外背景类的像素。

关键技术与进展



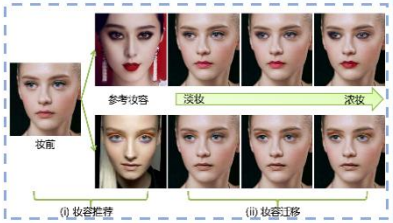
基于多上下文语义的成人内容识别

成人内容识别是图像内容识别的一个具体应用，它是互联网应用健康发展的基础工作。



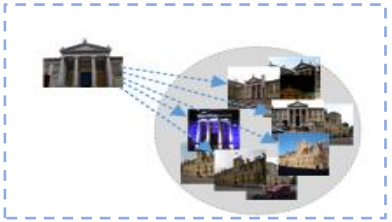
基于局部语义增强的场景解析

场景解析是计算机视觉的一个重要任务，它在自动驾驶、卫星遥感图像分析、图像搜索、机器人导航、室内三维建模等多个领域都有较广泛的潜在应用。



基于上下文融合的人像妆容迁移

化妆已成为目前大众日常生活的一部分，这给基于人脸的身份验证系统带来了巨大挑战。研究一种自动化妆系统对理解化妆后人脸身份验证具有重要意义。



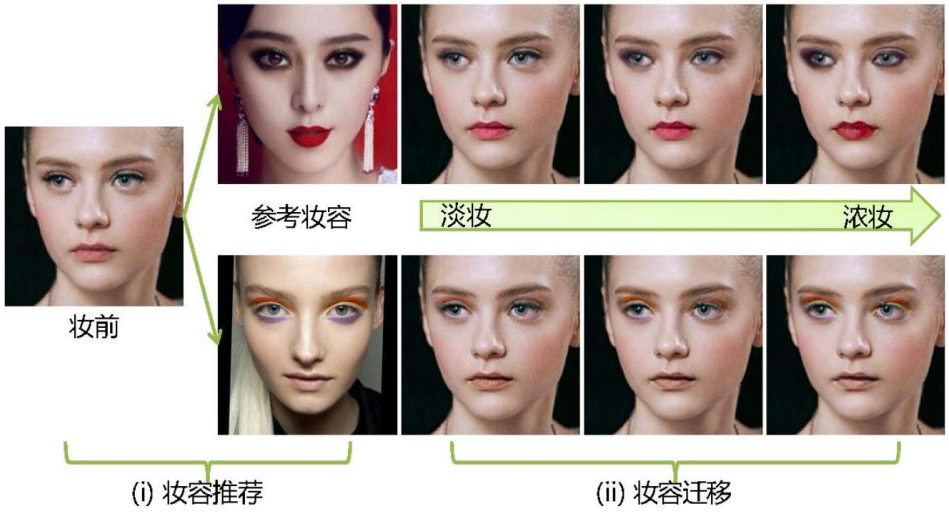
基于层次化语义哈希的图像检索

基于内容的图像检索可以帮助我们从海量的数据中找到符合我们需求的样本。它被广泛应用到搜索引擎、电子商务以及各种Web 2.0的教学系统中。

基于上下文融合的人像妆容迁移

问题提出

化妆已成为目前大众日常生活的一部分，这给基于人脸的身份验证系统带来了巨大挑战。研究一种自动化妆系统对理解化妆后人脸身份验证具有重要意义。本章从一个有趣的任务出发，通过研究人像妆容迁移实现对人像面部信息的理解和分析。

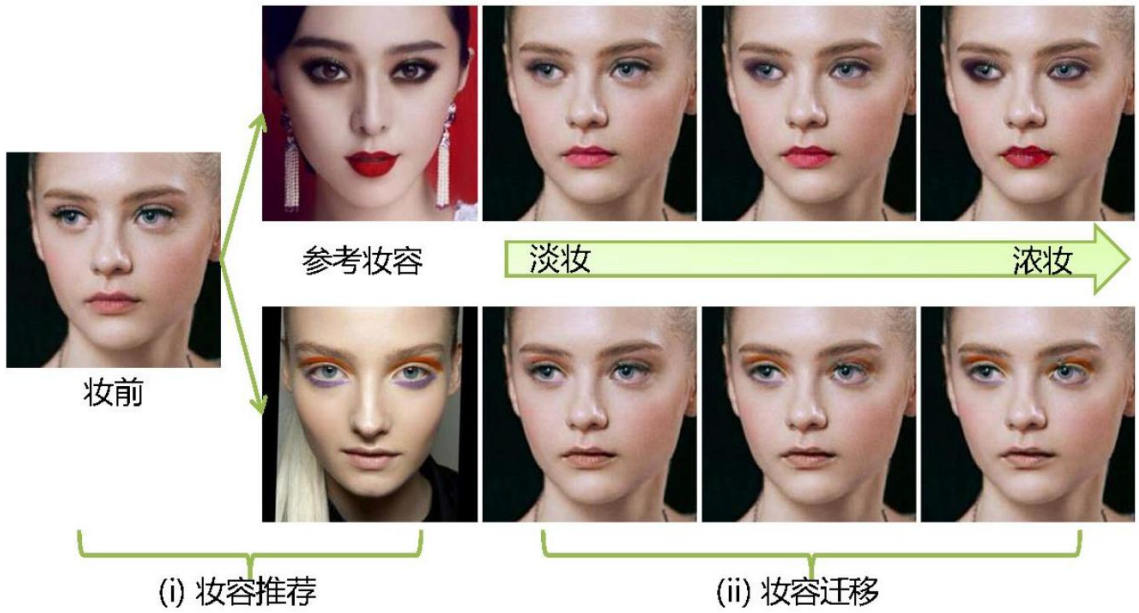


两个目标

- 为每一个妆前的脸推荐若干最合适的妆容图；
- 将参考妆容的粉底、眼影和唇彩迁移到化妆前的脸上。同时，妆容迁移的强度是可控制的。

基于上下文融合的人像妆容迁移

问题提出

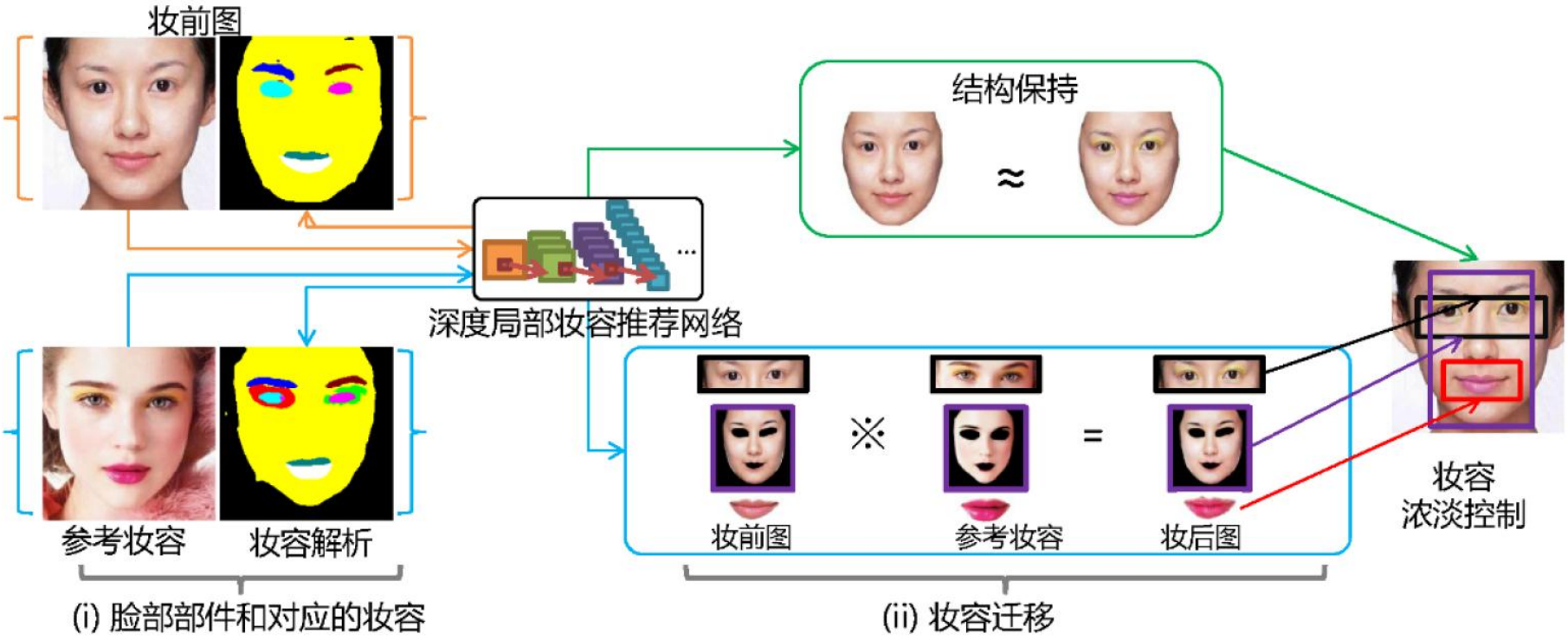


五个标准

- 全面的化妆功能
- 妆容定制
- 局部化
- 自然
- 妆容强度可控

基于上下文融合的人像妆容迁移

解决方案



妆容推荐 → 人脸解析 → 眼影变形 → 妆容迁移

基于上下文融合的人像妆容迁移

解决方案：妆容推荐

基于上下文融合的人像妆容迁移

解决方案：人脸解析

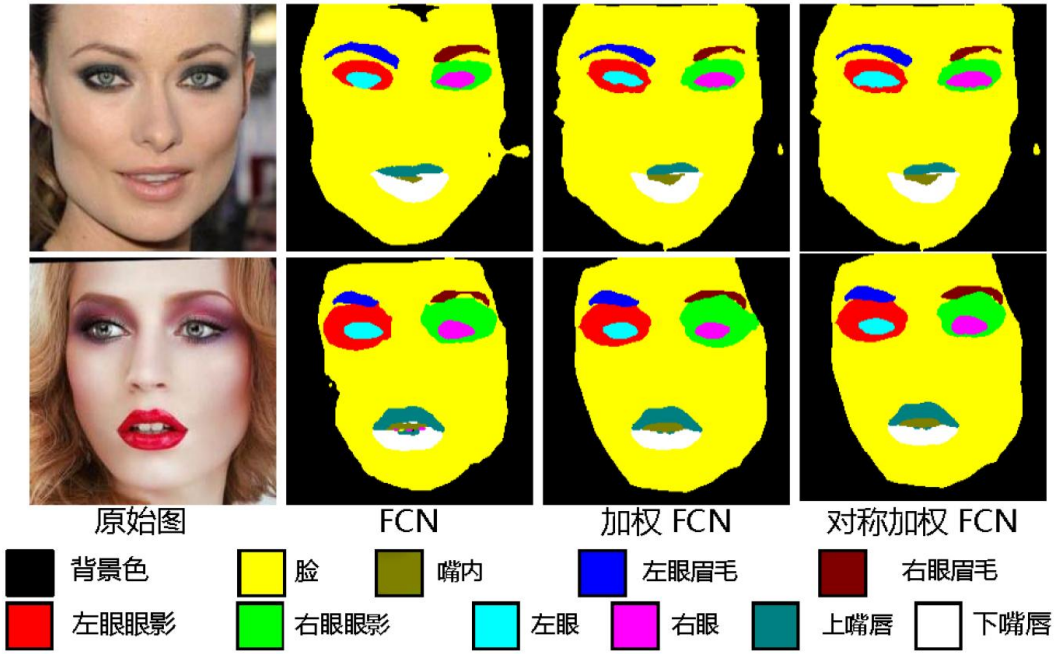


图 4-4 两组图像的人脸解析效果图

对称加权 FCN

$$\ell(x;\theta) = \sum_{ij} \ell'\left(y_{ij}, p(x_{ij};\theta)\right) \cdot w(y_{ij})$$

基于上下文融合的人像妆容迁移

解决方案：眼影变形 （左眼）

妆前图

Parsing

参考图

Parsing

Warping

基于上下文融合的人像妆容迁移

解决方案：妆容迁移

妆容迁移

结构保持

Total Variance Smooth

$$A^* = \underset{A \in \mathrm{R}^{H \times W \times C}}{\operatorname{argmin}} \lambda_s R_s(A) + \lambda_e\left(R_l(A) + R_r(A)\right) + \lambda_f R_f(A) + \lambda_l\left(R_{up}(A) + R_{low}(A)\right) + R_{V^\beta}(A)$$

脸型结构　眼影　粉底　唇彩　Total Variance Smooth

基于上下文融合的人像妆容迁移

解决方案：妆容迁移 – 结构保持

两个眼影迁移的例子

$$A^* = \underset{A \in \mathrm{R}^{H \times W \times C}}{\operatorname{argmin}} \lambda_s R_s(A) + \lambda_e\left(R_l(A) + R_r(A)\right) + \lambda_f R_f(A) + \lambda_l\left(R_{up}(A) + R_{low}(A)\right) + R_{V^\beta}(A)$$

Structure Preservation(or Eye shadow transfer):

$$A^* = \underset{A \in \mathrm{R}^{H \times W \times C}}{\operatorname{argmin}} R_l(A) + \underset{A \in \mathrm{R}^{H \times W \times C}}{\operatorname{argmin}} \left\| P\left(\Omega^l\left(A(s'_b)\right)\right) - P\left(\Omega^l\left(R(s'_r)\right)\right) \right\|_2^2$$

基于上下文融合的人像妆容迁移

解决方案：妆容迁移 – 风格保持

两个粉底和唇彩迁移的例子

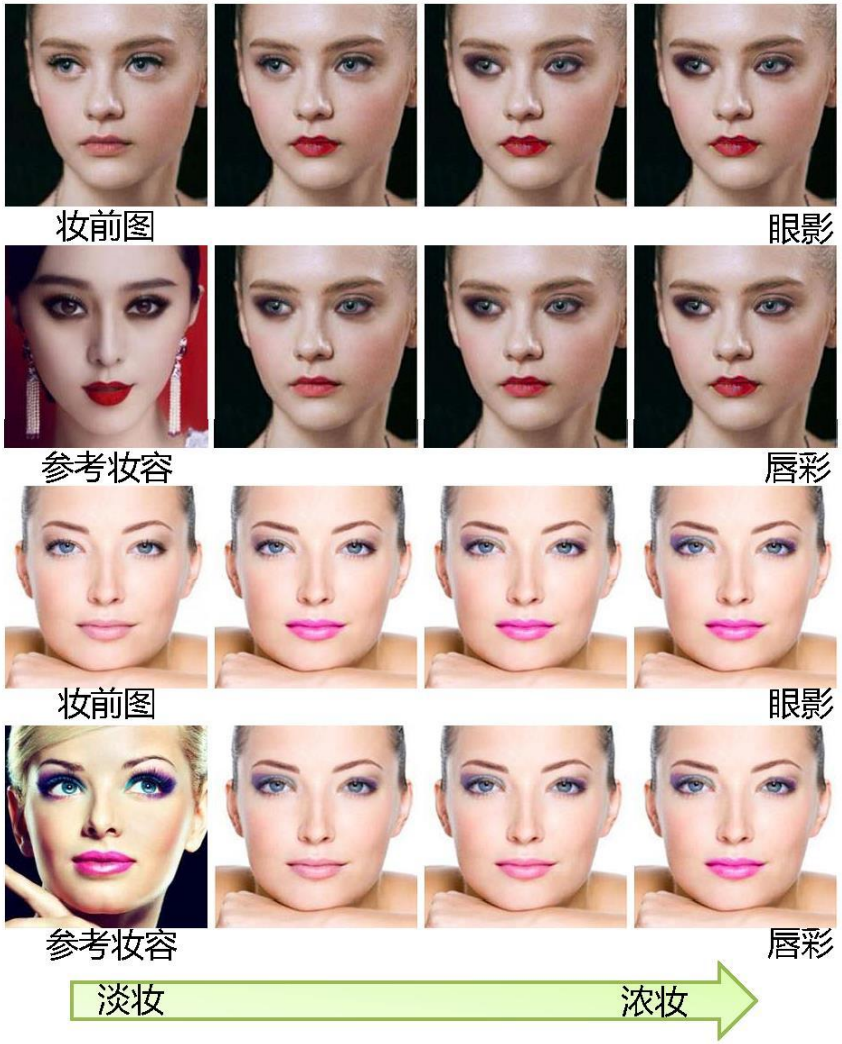
$$A^* = \underset{A \in \mathrm{R}^{H \times W \times C}}{\operatorname{argmin}} \lambda_s R_s(A) + \lambda_e\left(R_l(A) + R_r(A)\right) + \lambda_f R_f(A) + \lambda_l\left(R_{up}(A) + R_{low}(A)\right) + R_{V^\beta}(A)$$

Lip gloss and Foundation Transfer:

$$A^* = \underset{A \in \mathrm{R}^{H \times W \times C}}{\operatorname{argmin}} R_f(A) = \underset{A \in \mathrm{R}^{H \times W \times C}}{\operatorname{argmin}} \sum_{l=1}^{L} \left\| \Omega_{ij}^l\left(A(s_b')\right) - \Omega_{ij}^l\left(\mathrm{R}(s_r')\right) \right\|_2^2$$

基于上下文融合的人像妆容迁移

实验结果 – 可控妆容迁移示意图

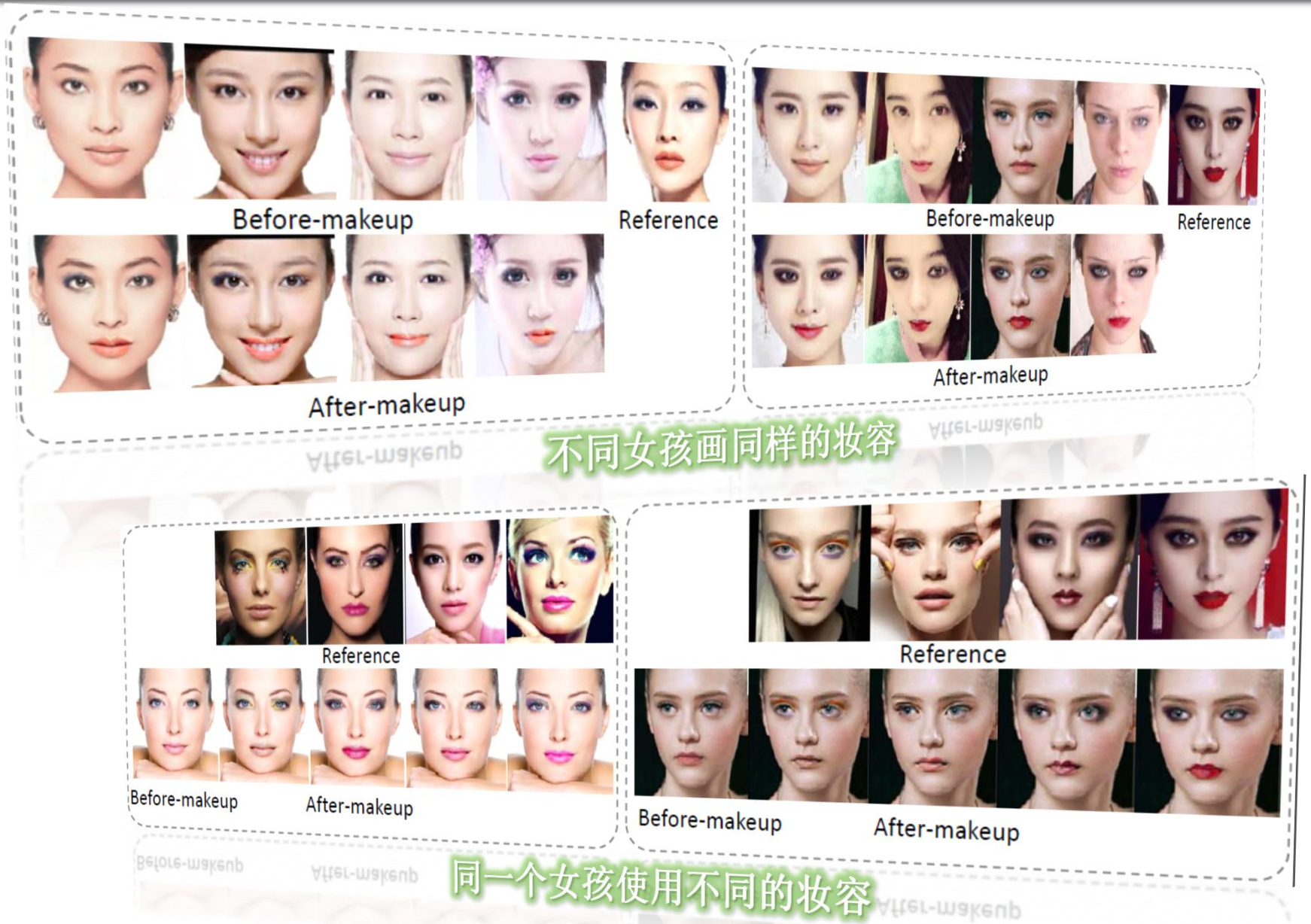


基于上下文融合的人像妆容迁移

实验结果 – 定性和定量对比结果

	好很多	更好	相同	更差	差很多
Guo 等人[136]	9.7%	55.9%	22.4%	11.1%	1.0%
NeuralStyle-CC[184]	82.7%	14.0%	3.24%	0.15%	0%
NeuralStyle-CS[184]	82.8%	14.9%	2.06%	0.29%	0%

基于上下文融合的人像妆容迁移



基于上下文融合的人像妆容迁移

失败案例分析

人脸解析极为重要！

妆前图　参考图　解析结果　妆后图

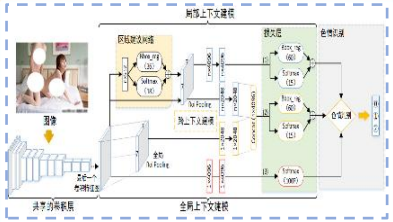
基于上下文融合的人像妆容迁移

本章贡献

在本章的工作中，我们提出了一种新颖的深度局部妆容迁移网络（Deep Localized Makeup Transfer Network，DLMTN）去自动地从一个带妆人脸图像上将妆容迁移到素颜图。总的来说，本章有如下三个贡献：

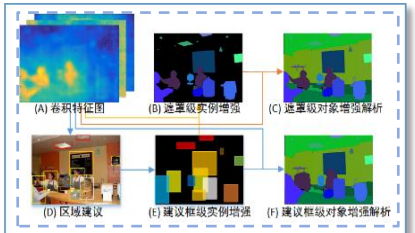
1. 提出一个统一的深度学习生成框架，实现从人脸推荐、人脸解析、人脸全局和局部特征提取和特征迁移等多项功能。基于这个功能，我们实现了一个有趣的应用——人像妆容迁移。
2. 充分利用人脸全局和局部的上下文信息，在保持人脸形状的基础上，实现了全局粉底和局部眼影、局部唇彩的特征迁移。特征迁移的过程是一个生成学习的过程，利用迭代学习的方法，我们的网络较好地处理了多种不同区域的不同特征的融合和生成。
3. 在特征迁移的过程中，通过对不同上下文信息的分离处理，在权重的控制下，可以定制不同局部和整体特征的强度，实现不同妆容的浓淡调节。

关键技术与进展



基于多上下文语义的成人内容识别

成人内容识别是图像内容识别的一个具体应用，它是互联网应用健康发展的基础工作。

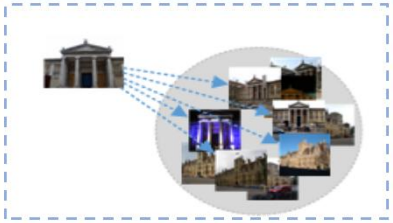


基于局部语义增强的场景解析

场景解析是计算机视觉的一个重要任务，它在自动驾驶、卫星遥感图像分析、图像搜索、机器人导航、室内三维建模等多个领域都有较广泛的潜在应用。

基于上下文融合的人像妆容迁移

化妆已成为目前大众日常生活的一部分，这给基于人脸的身份验证系统带来了巨大挑战。研究一种自动化妆系统对理解化妆后人脸身份验证具有重要意义。



基于层次化语义哈希的图像检索

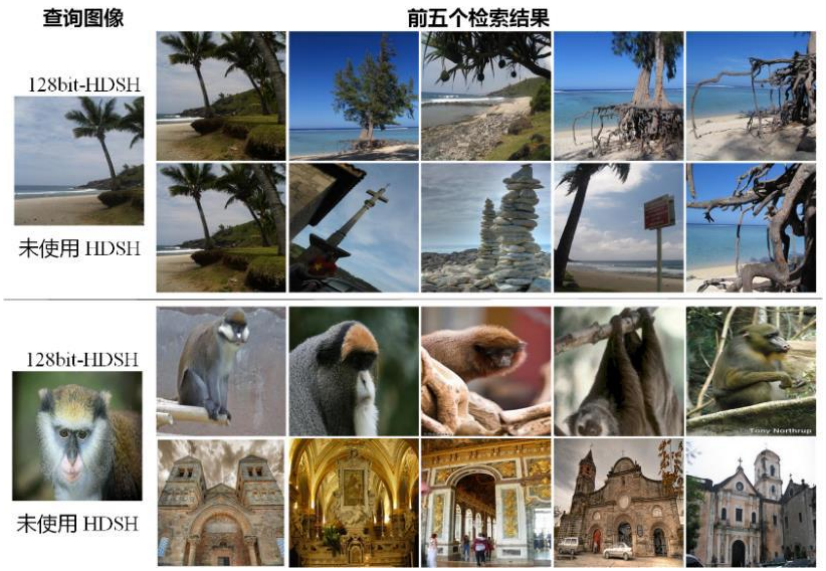
基于内容的图像检索可以帮助我们从海量的数据中找到符合我们需求的样本。它被广泛应用到搜索引擎、电子商务以及各种Web 2.0的教学系统中。

基于层次化语义哈希的图像检索

问题提出

基于内容的图像检索是计算机视觉任务的一个重要组成部分，它可以帮助我们从海量的数据中找到符合我们需求的样本。它被广泛应用到搜索引擎、电子商务以及各种Web 2.0的教学系统中。

然而，搜索的**准确率**和**效率**，逐渐成为大数据环境下的制约因素。



平均检索时间
0.83毫秒
138毫秒
333毫秒
45558毫秒

图5-1 *Holidays*和*Imagenet*数据集上使用HDSH前后图像检索结果对比图

基于层次化语义哈希的图像检索

解决方案

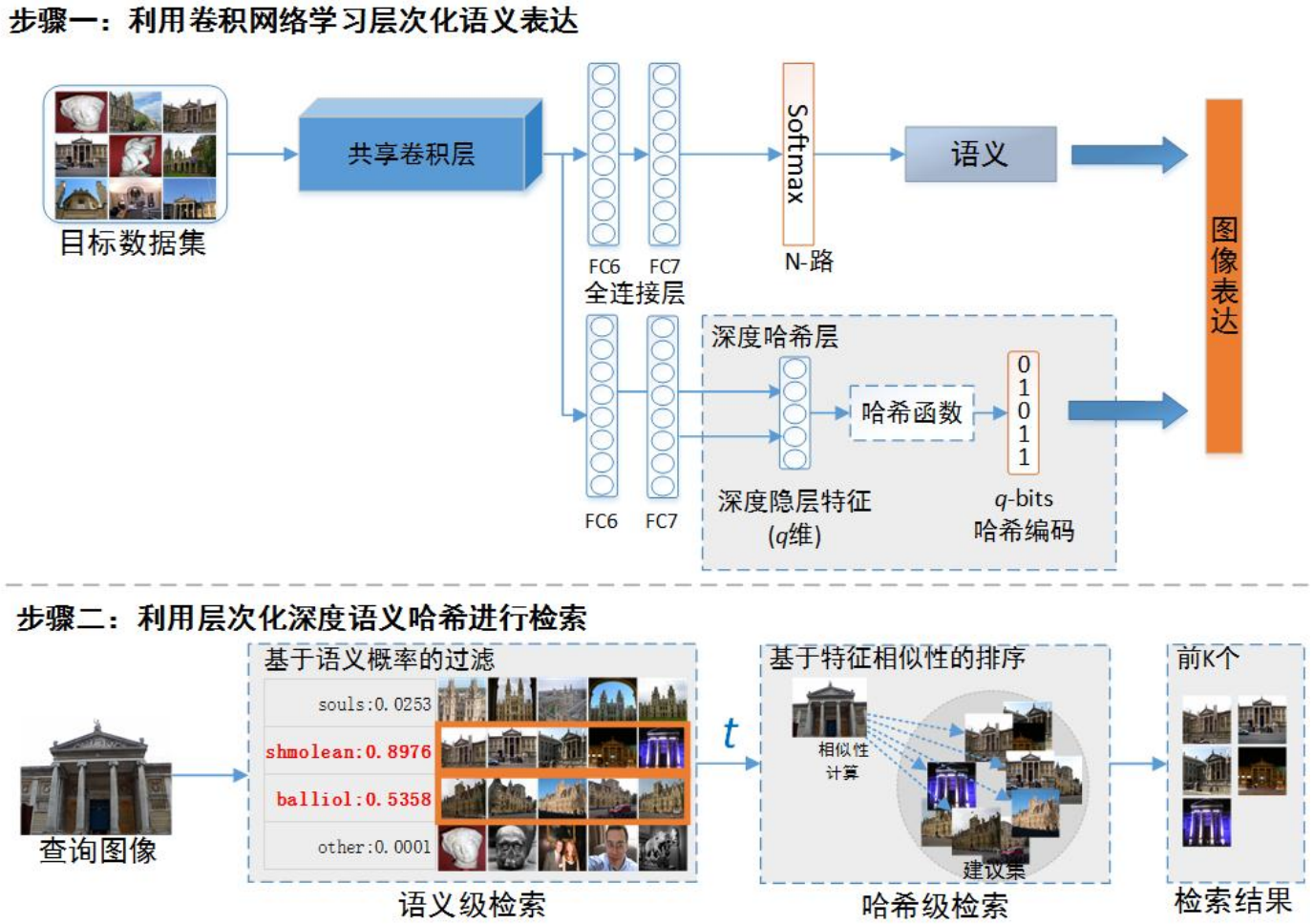


图5-3 基于层次化深度语义哈希的图像检索体系结构

基于层次化语义哈希的图像检索

解决方案

基于概率的语义级相似性

$$R_{II}\big(I(a),I(b)\big)=\big[P\big(I(a),I(b)\big)\big]_t$$

$$x=\begin{cases}[x]_t, & x>t\\ 0, & otherwise\end{cases}\qquad \text{其中，}\ x=P\big(I(a),I(b)\big)$$

哈希级相似性

$$H=h(x)=\begin{cases}1, & f(x_i)-Avg_i^q\big(f(x_i)\big)>0\\ 0, & f(x_i)-Avg_i^q\big(f(x_i)\big)<0\end{cases}$$

$$d(q,i)=Dist\big(I_q,I_i\big)=||H_q-H_i||$$

基于层次化语义哈希的图像检索

解决方案

语义级和哈希级融合的相似性

$$Sim(a,b) = \sum_{i}^{C} [P(I(a), I(b))]_t \times (1 - d(a,b))$$

基于概率的语义级相似性

哈希级相似性

检索过程

1. 计算查询图像q和目标图像集中每一个图像b的语义相关性$\boldsymbol{R}_{II}(I(q), I(b))$，如果语义相关性$\boldsymbol{R}_{II} = 0$，则丢弃图像$b$，反之，如果语义相关性$\boldsymbol{R}_{II} \neq 0$，则将图像b加入到候选图像集中。
2. 在候选建议集P上计算哈希级相似性。

基于层次化语义哈希的图像检索

解决方案 – 效率分析

假设：计算查询图像I_q和一个待查询图像I_i之间的相似性的时间复杂度为$\mathcal{O}(1)$

则，计算整个数据集的时间复杂度将为$\mathcal{O}(n)$

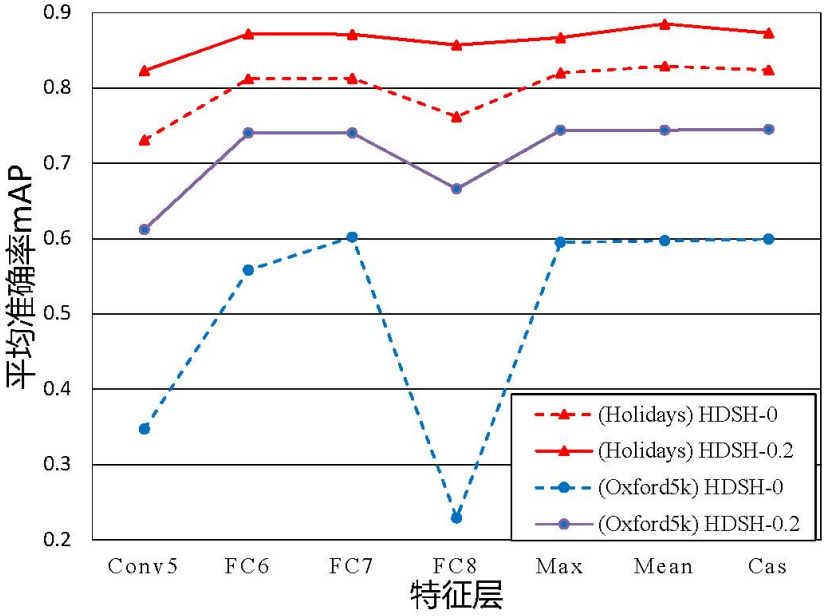
令：候选建议集P的大小m，假设每个类的数据分布是均匀的，则 $\frac{n}{C} = \frac{m}{C^+}$（$m \ll$ n ）

因此，总的加速比$Rate_{up} = \frac{\mathcal{O}(n)}{\mathcal{O}(m)} = \frac{C}{C^+}$

由于 一个查询图像具有相关性的语义类别数量C^+通常少于10（大多数时候甚至少于5），因此，例如，在*Holidays*数据集上的加速比大约为$\frac{500}{5\sim10} = 50\sim100$倍。

基于层次化语义哈希的图像检索

实验结果



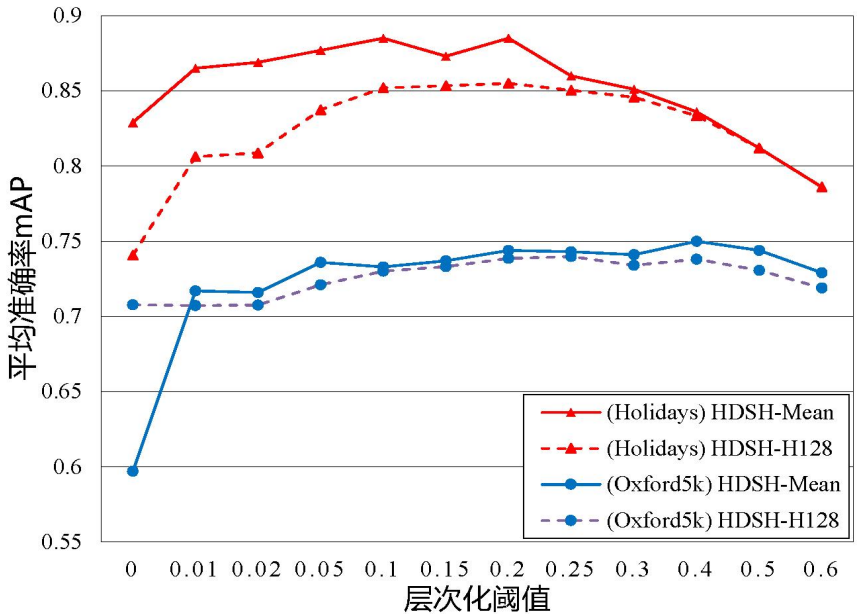


图 5- 4 *Holidays*和*Oxford5k*数据集上不同特征的性能比较

基于层次化语义哈希的图像检索

实验结果

Method	Holidays	Oxford5k	Oxf105k
SIFT-based methods			
BoW 200k-D [12]	0.54	0.364	–
Improved FV [28]	0.626	0.414	–
VLADintra [1]	0.653	0.558	–
LCS+RN [5]	0.658	0.517	0.456
CVLAD [40]	0.827	0.514	-
HE+MA+PGM [19]	**0.892**	**0.737**	–
CNN-based methods			
Neural Codes [2]	0.793	0.545	0.512
MOP-CNN [8]	0.808	–	–
LFDN [24]	0.840	0.581	54.2
CNNaug-ss [30]	0.843	0.68	–
Spatial Pooling [31]	**0.896**	**0.843**	**0.795**
DHRS [41]	0.858	0.712	0.603
HDSH-Mean-0	0.829	0.597	0.523
HDSH-Mean-0.2	**0.885**	**0.744**	**0.712**

表5- 1(左) 未压缩特征的性能比较

Method	D	Holidays	Ox5k	Ox105k
LCS+RN [5]	16	0.323	0.27	0.222
Neural Codes [2]	16	0.609	0.418	0.354
HDSH-H16(0.2)	16	0.815	0.722	0.665
Neural Codes [2]	32	0.729	0.515	0.467
HDSH-H32(0.2)	32	0.858	0.723	0.665
Neural Codes [2]	64	0.777	0.548	0.508
HDSH-H64(0.2)	64	0.856	0.737	0.671
FV + T [10]	128	0.617	0.433	–
VLADintra [1]	128	0.625	0.448	–
LCS+RN [5]	128	0.335	0.322	0.262
Neural Codes [2]	128	0.789	0.557	0.523
LFDN [24]	128	0.836	0.558	52.9
HDSH-H128(0.2)	128	**0.855**	**0.739**	**0.676**
Neural Codes [2]	256	0.789	0.557	0.524
DHRS [41]	256	0.818	0.574	0.488
Spatial Pooling [31]	256	0.742	0.533	0.511
HDSH-H256(0.2)	256	**0.858**	**0.754**	**0.688**
MOP-CNN [8]	512	0.784	–	–
Neural Codes [2]	512	0.789	0.557	0.522
DHRS [41]	512	0.838	0.672	0.563
HDSH-H512(0.2)	512	**0.86**	**0.768**	**0.693**

表5- 3(右) 压缩特征的性能比较

基于层次化语义哈希的图像检索

实验结果 – 大数据集性能对比

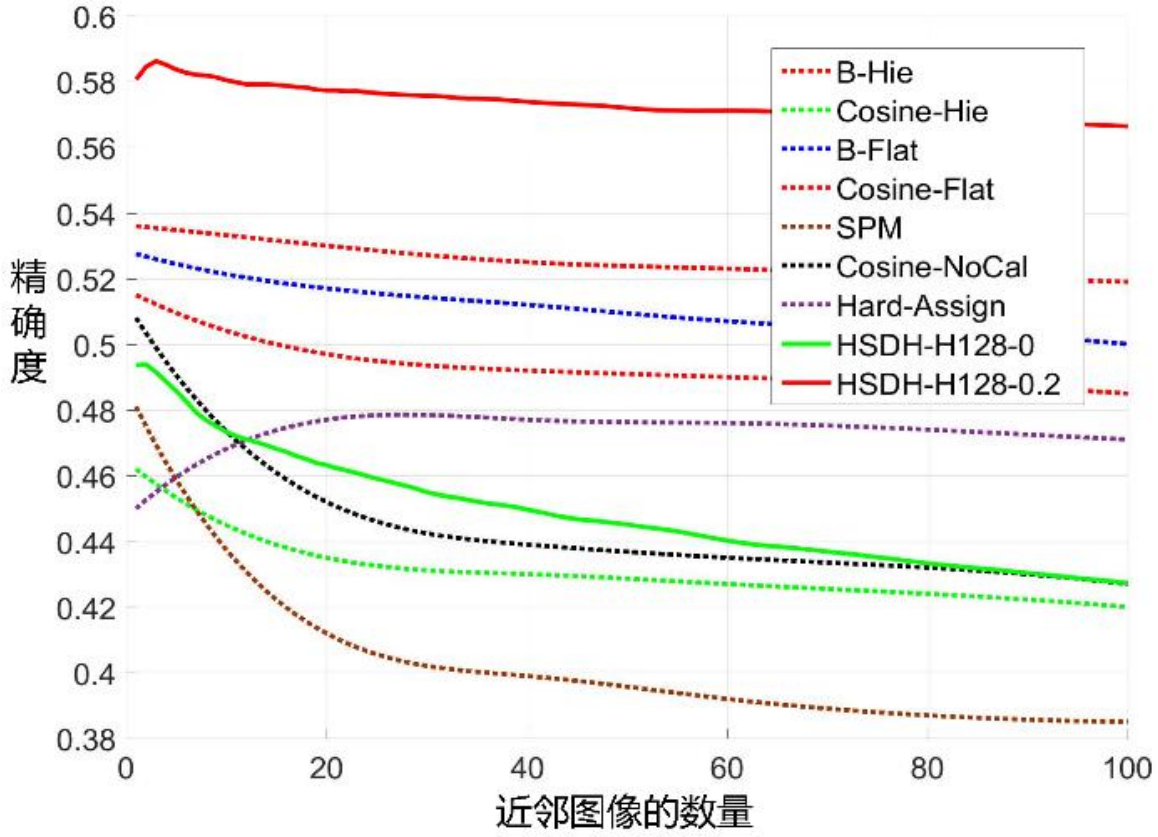
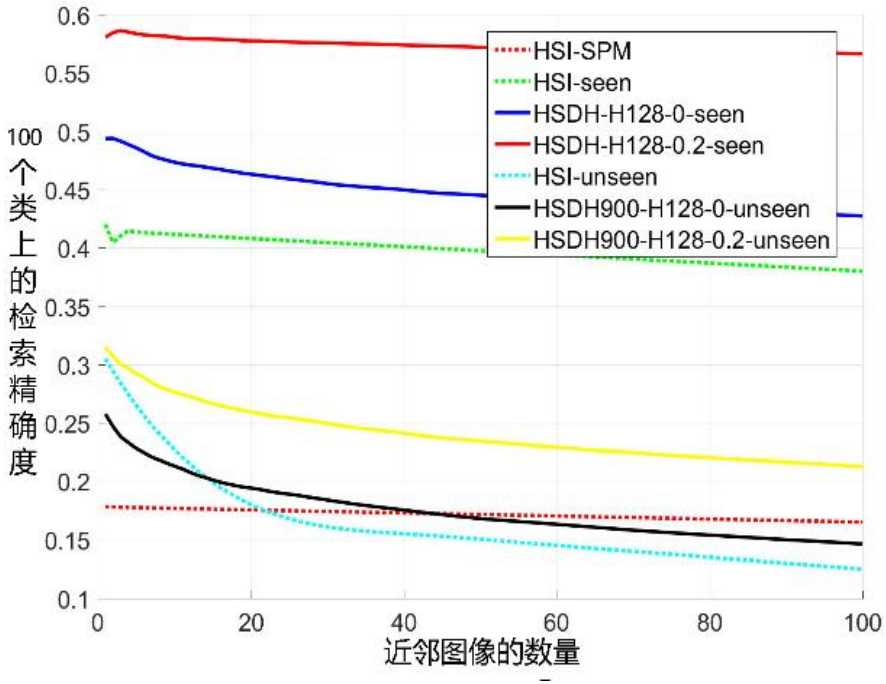


图 5- 5 *Imagenet*数据集上检索精度对比图

基于层次化语义哈希的图像检索

实验结果 – 跨类泛化性能分析



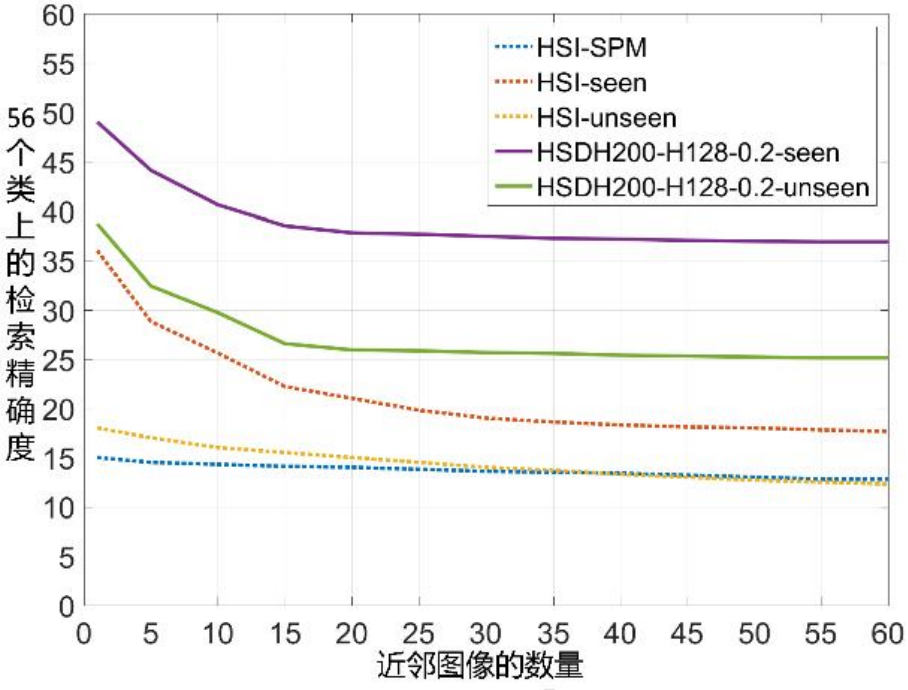


图 5 7 Imagenet（左）和Caltech256（右）数据集上的跨类检索性能对比

基于层次化语义哈希的图像检索

实验结果 – 效率分析

表 5-4 所有数据集上的检索时间对比（单位：毫秒）

方法	维度	Holidays	Oxford5k	Oxford105k	Caltech256	Imagenet
类别数		500	12	12	257	1000
HDSH-mean-0	4K	138	693	3504	1121	45558
HDSH-mean-0.2	4K	0.83	167	666	13	333
加速比		167.3	4.14	5.26	87	134.9
HDSH-H128-0	128	8.1	114	613	198	8900
HDSH-H128-0.2	128	0.15	28	140	5.4	54
加速比		54	4.15	4.37	36.7	165.1

基于层次化语义哈希的图像检索

本章贡献

在本章中，我们提出了一种新颖的快速图像检索算法——层次化深度语义哈希（Hierarchical Deep Semantic Hashing, HDSH），该算法有效地解决了基于内容的图像检索中最重要的两个问题：**检索精度**和**检索效率**。我们的工作主要有四个方面的贡献：

1. 提出了一个完整的端对端的特征学习和提取框架，它可以同时输出基于概率的语义级特征和哈希级特征。
2. 通过组合语义级相似性和哈希级相似性，HDSH不仅为计算图像的距离提供了强大的先验知识，提高了检索的性能，也大幅缩小了检索空间，使算法可以被应用到大规模的数据集上。
3. 大量的实验证明，HDSH算法具有较强的稳定性，即使特征维度降低到一个较小的尺度，例如128bit的哈希编码，它仍然具有较强的判别能力。
4. 我们还提出了一种简单但是有效数据扩展方法，用来解决数据集样本和样本类别不平衡问题。

Part
04

结论分析与总结

/ 工作总结

/ 研究展望

/ 致谢

工作总结

针对成人识别任务中样本多样性问题	基于高层语义的细到粗策略 多上下文联合决策
针对场景解析任务中难对象识别问题和额外背景类造成的误判问题	基于局部语义的特征增强策略 基于语义的黑洞填充策略
针对人像妆容迁移任务中上下文融合时语义保持困难的问题	对称加权交叉熵损失 基于迭代的全局和局部上下文融合网络
针对大数据环境下搜索空间太大引起的图像搜索效率降低的问题	基于概率的层次化语义相似性过滤策略

研究展望

Understanding Scene

场景理解

Generative
Adversarial Network

生成对抗网络

Predictive Learning

预测学习

華中科技大學
HUAZHONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Thank you !

- 感谢导师凌贺飞教授的教诲
- 感谢实验室李平老师和全体成员的互相帮助与扶持
- 感谢中科院信工所刘偲副研究员、操晓春研究员和孙瑶副研究员的指导和帮助
- 感谢云南开放大学各位领导的关心
- 感谢父母、妻子和儿子的支持
- 感谢论文的评阅专家及答辩委员会的各位专家

请各位专家指正

华中科技大学 计算机科学与技术学院

博士生：欧新宇　　导　师：凌贺飞 教授