中国科学院大学

University of Chinese Academy of Sciences

Deep Learning

Applications in Image and Videos

Xinfeng Zhang（张新峰）

School of Computer Science and Technology

University of Chinese Academy of Sciences

Email: xfzhang@ucas.ac.cn

计算机科学与技术学院

SCHOOL OF COMPUTER SCIENCE AND TECHNOLOGY

提纲

- 图像/视频处理
- 图像/视频压缩
- 传统的计算机视觉处理
- 图像分类
- 目标检测
- 图像分割
- 图像回归
- 中英文术语对照

中国科学院大学
University of Chinese Academy of Sciences

1

图像/视频处理

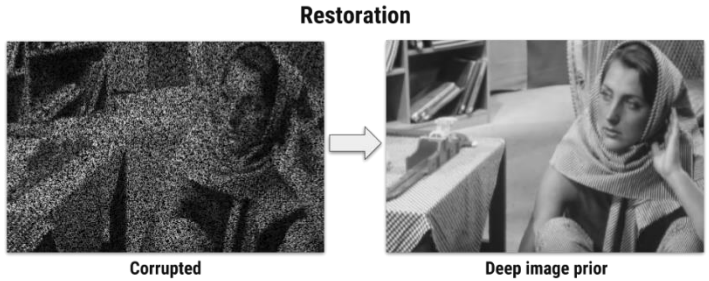
图像处理问题

- 图像复原、增强和质量评价

超分辨率

Super-Resolution

弱光照增强



去噪

动态范围增强

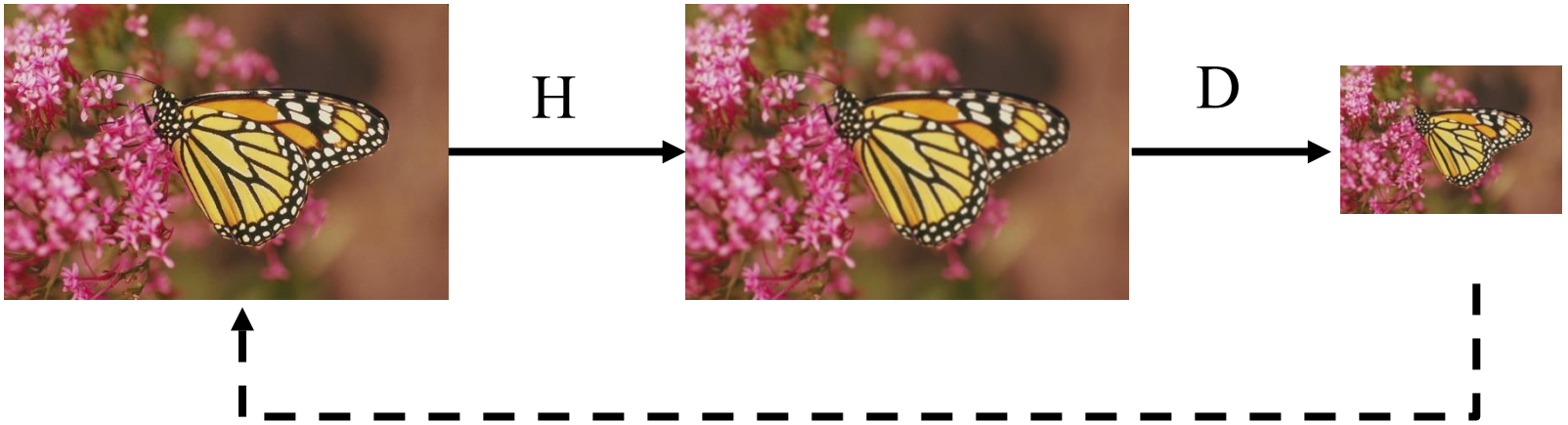
质量评价

超分辨率问题

❑ 传统超分辨率问题

$$arg\min_{\mathbf{X}} \|\mathrm{DH}\mathbf{X} - \mathbf{Y}\|_2^2 + \lambda\|\mathbf{X}\|$$

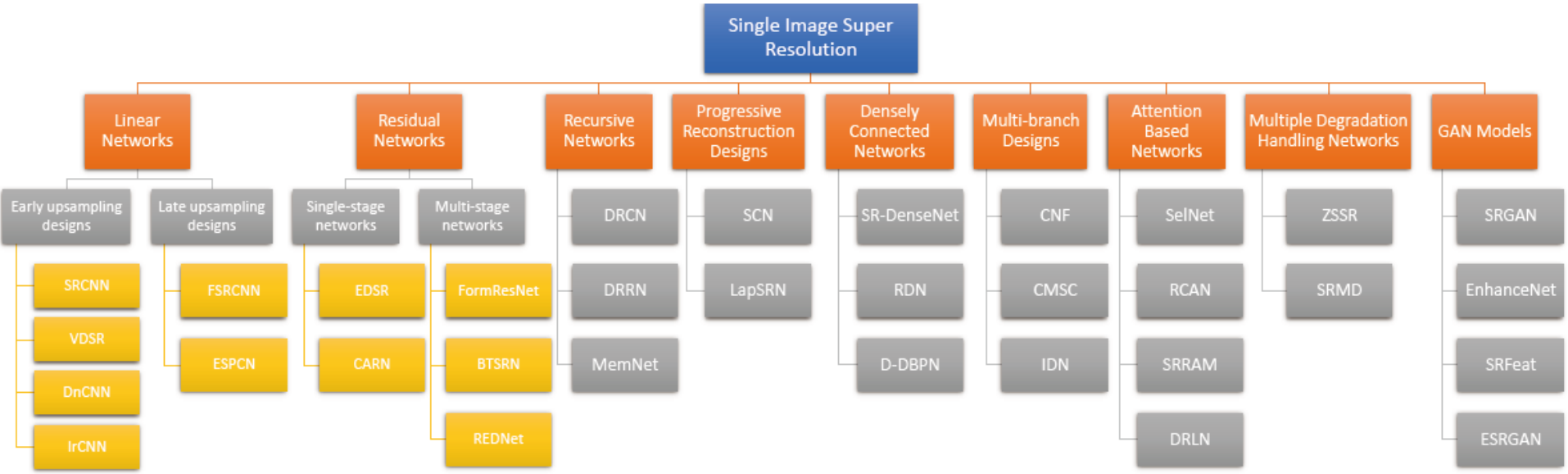
传统方法主要集中在图像先验，包含基于重构的方法、基于字典稀疏表示等



基于深度学习的超分辨率技术

❑ 研究现状：

– 主要是卷积神经网络

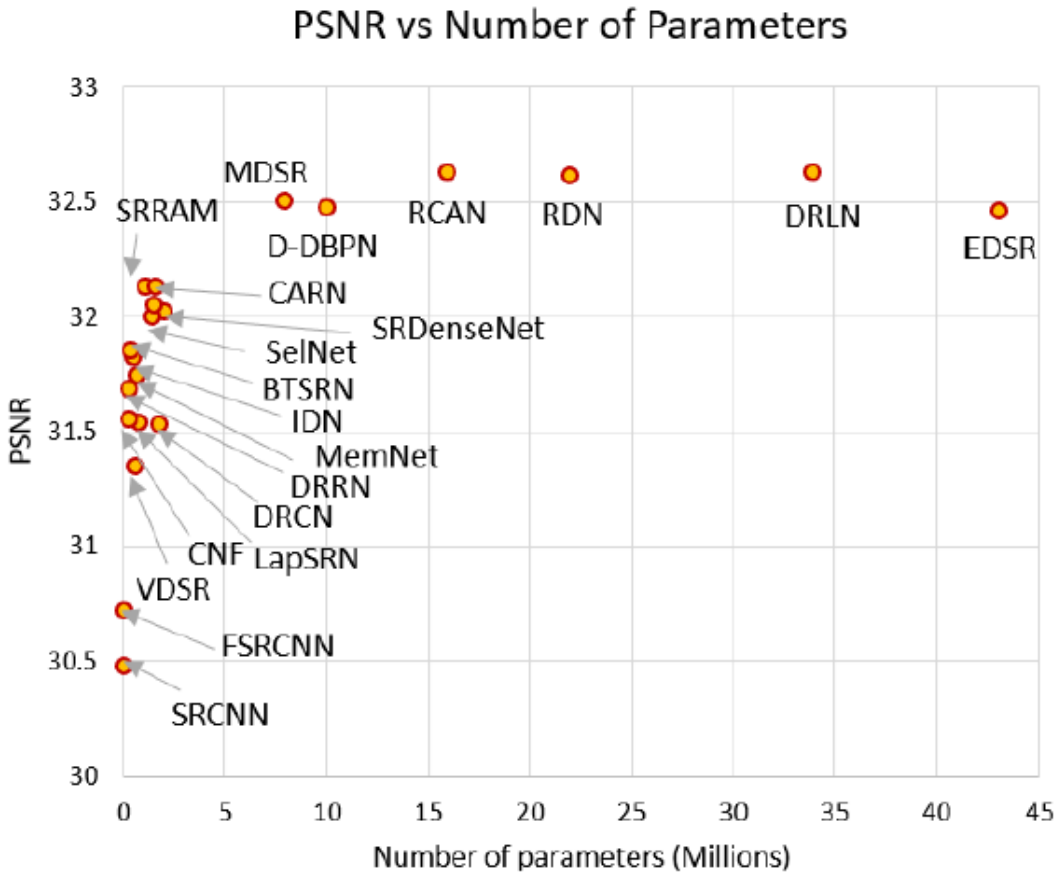


Anwar, Saeed, Salman Khan, and Nick Barnes. "A deep journey into super-resolution: A survey." arXiv preprint arXiv:1904.07523 (2019).

基于深度学习的超分辨率技术

❑ 研究现状：

– 主要性能

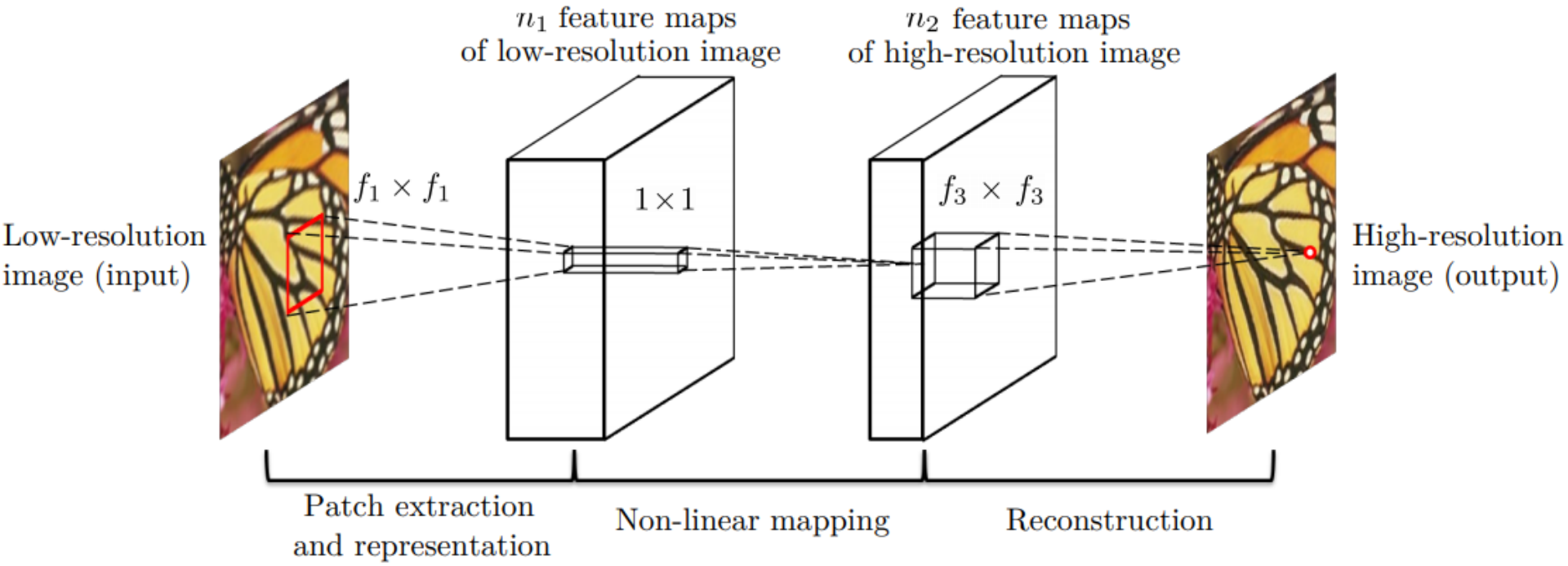


Anwar, Saeed, Salman Khan, and Nick Barnes. "A deep journey into super-resolution: A survey." arXiv preprint arXiv:1904.07523 (2019).

基于深度学习的超分辨率技术

❑ SRCNN

– 先将低分辨率图像使用双三次插值放大至目标尺寸（如放大至2倍、3倍、4倍），此时仍然称放大至目标尺寸后的图像为低分辨率图像(Low-resolution image)，即图中的输入(input)；



Dong, Chao, Chen Change Loy, Kaiming He, and Xiaoou Tang. "Learning a deep convolutional network for image super-resolution." In *European conference on computer vision*, pp. 184-199. Springer, Cham, 2014.

基于深度学习的超分辨率技术

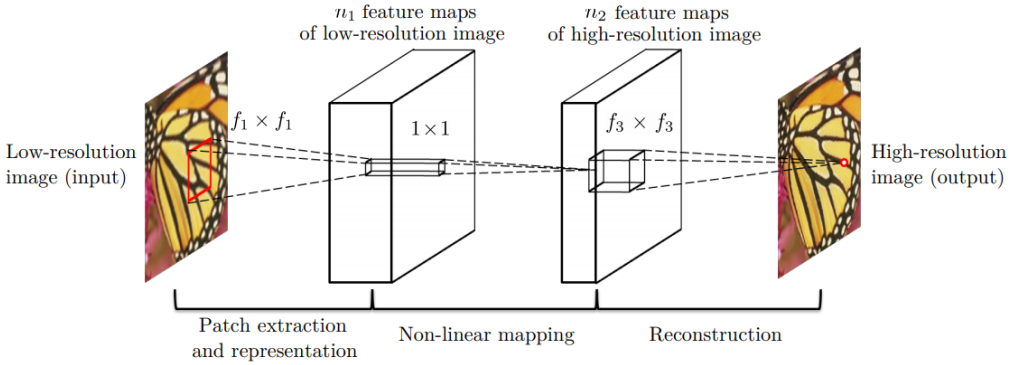
❑ SRCNN

– 先将低分辨率图像使用双三次插值放大至目标尺寸（如放大至2倍、3倍、4倍），此时仍然称放大至目标尺寸后的图像为低分辨率图像(Low-resolution image)，即图中的输入(input)；

Patch extraction and representation: 从低分辨率图像中提取重叠块，将每一个块通过卷积网络表达为高维向量

Nonlinear Mapping: 将每一个块的高维向量经过非线性映射为另外一个高维向量（维度增加）

Reconstruction: 将高维向量累积成高分辨率图像



Dong, Chao, Chen Change Loy, Kaiming He, and Xiaoou Tang. "Learning a deep convolutional network for image super-resolution." In *European conference on computer vision*, pp. 184-199. Springer, Cham, 2014.

基于深度学习的超分辨率技术

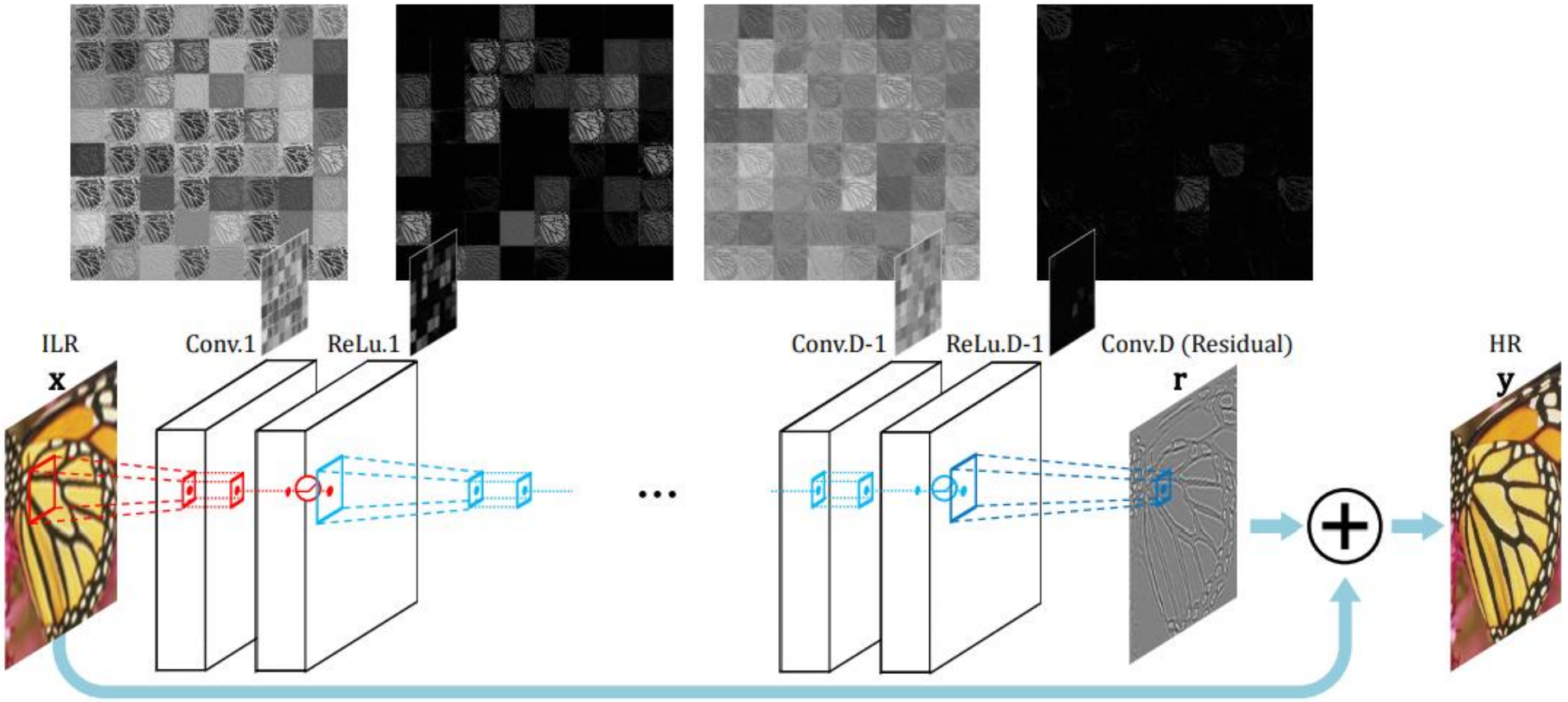
❑ VDSR

- SRCNN存在三个问题需要
  - 依赖于小图像区域的内容
  - 训练收敛太慢
  - 网络只对于某一个比例有效
- VDSR的改进
  - 增加了感受野，在处理大图像上有优势，由SRCNN的13*13变为41*41
  - 采用残差图像进行训练，收敛速度变快
  - 考虑多个尺度，一个卷积网络可以处理多尺度问题

Kim, Jiwon, Jung Kwon Lee, and Kyoung Mu Lee. "Accurate image super-resolution using very deep convolutional networks." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1646-1654. 2016.

基于深度学习的超分辨率技术

❑ VDSR

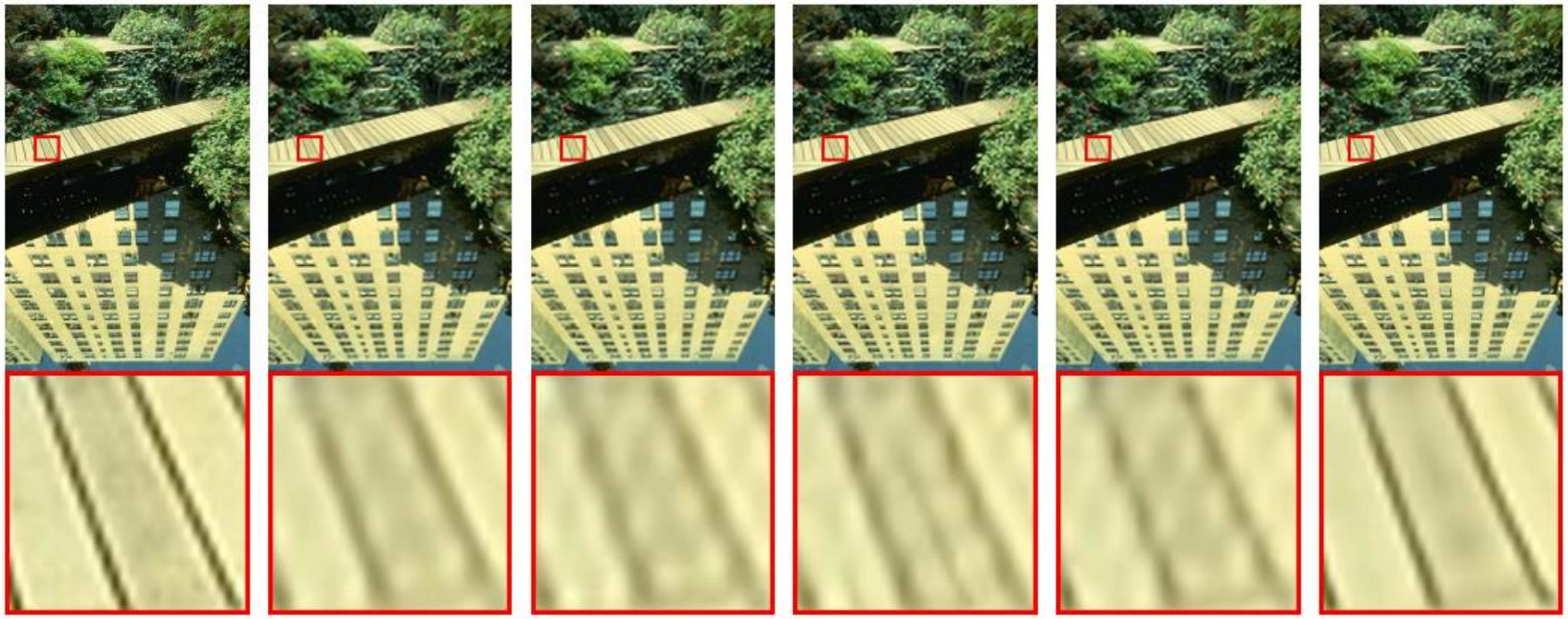


Kim, Jiwon, Jung Kwon Lee, and Kyoung Mu Lee. "Accurate image super-resolution using very deep convolutional networks." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1646-1654. 2016.

基于深度学习的超分辨率技术

❑ VDSR性能

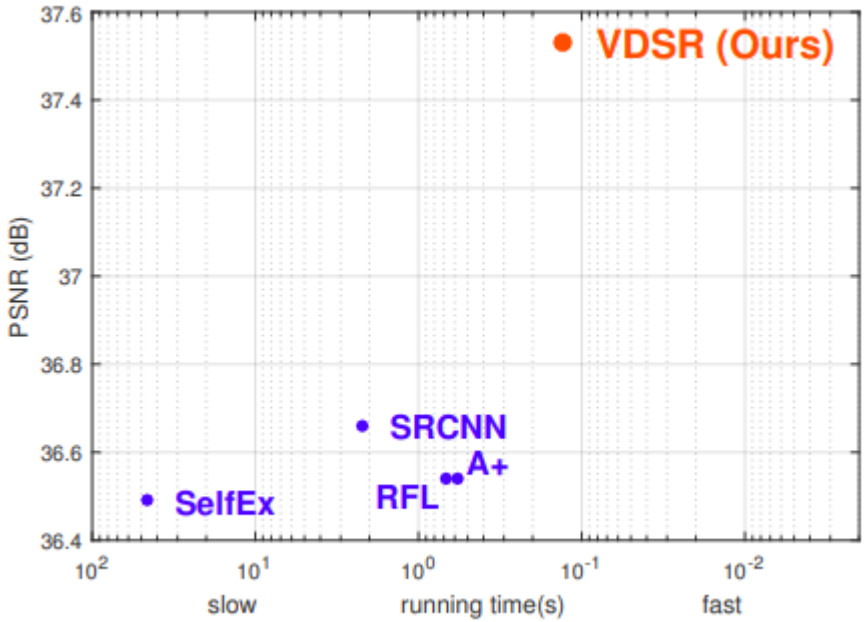
– 分辨率扩大为原来的3倍



Kim, Jiwon, Jung Kwon Lee, and Kyoung Mu Lee. "Accurate image super-resolution using very deep convolutional networks." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1646-1654. 2016.

基于深度学习的超分辨率技术

❑ VDSR性能



Kim, Jiwon, Jung Kwon Lee, and Kyoung Mu Lee. "Accurate image super-resolution using very deep convolutional networks." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1646-1654. 2016.

图像去噪

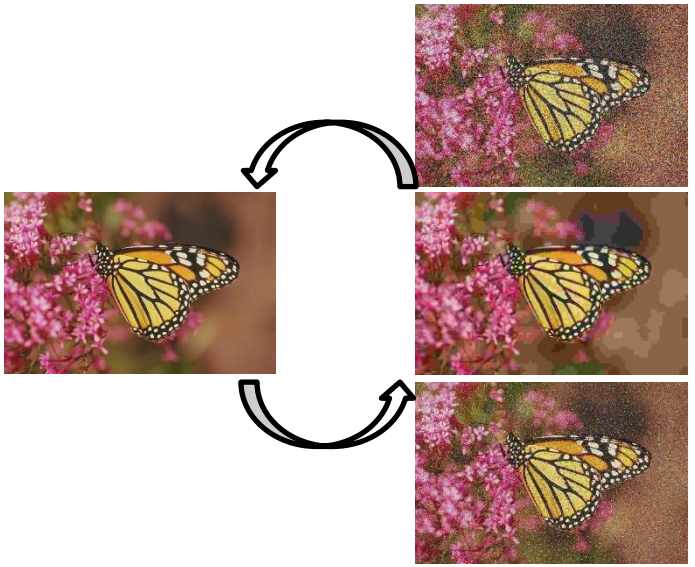
❑ 问题描述

– 图像在获取、处理和传输过程中都会引入噪声，去噪问题就是从带噪声的实际图像中恢复出原始图像

$$arg\min_{\mathbf{X}}\|\mathbf{X}-\mathbf{Y}\|_2^2+\lambda\|\mathbf{T}(\mathbf{X})\| \qquad \mathbf{Y}=\mathbf{X}+\boldsymbol{n}$$

– 去噪方法

- 滤波类
- 稀疏表达类
- 外部先验
- 聚类低秩
- 深度学习



基于深度学习的图像去噪

❑ 研究现状

– 这篇综述介绍了超过200篇关于深度学习去噪的论文

Table 1: CNN/NN for AWNI denoising.

References	Methods	Applications	Key words (remarks)
Zhang et al. (2017) [222]	CNN	Gaussian image denoising, super-resolution and JPEG deblocking	CNN with residual learning, and BN for image denoising
Wang et al. (2017) [190]	CNN	Gaussian image denoising	CNN with dilated convolutions, and BN for image denoising
Bae et al. (2017) [12]	CNN	Gaussian image denoisng, super-resolution	CNN with wavelet domain, and residual learning (RL) for image restoration
Jin et al. (2017) [90]	CNN	Medical (X-ray) image restoration	Improved Unet from iterative shrinkage idea for medical image restoration
Tai et al. (2017) [174]	CNN	Gaussian image denoisng, super-resolution and JPEG deblocking	CNN with recursive unit, gate unit for image restoration
Anwar et al. (2017) [11]	CNN	Gaussian image denoisng	CNN with fully connected layer, RL and dilated convolutions for image denoising
McCann et al. (2017)] [91]	CNN	Inverse problems (i.e. denoising, deconvolution, super-resolution)	CNN for inverse problems
Ye et al. (2018) [209]	CNN	Inverse problems(i.e. Gaussian image denoising, super-resoluion)	Signal processing ideas guide CNN for inverse problems
Yuan et al. (2018) [214]	CNN	Hyper-spectral image denoising	CNN with multiscale, multilevel features techniques for hyper-spectral image denoising
Jiang et al. (2018) [87]	CNN	Gaussian image denoising	Multi-channel CNN for image denoising
Chang et al. (2018) [25]	CNN	Hyper-spectral image (HSI) denoising, HIS restoration	CNN consolidated spectral and spatial coins for hyper-spectral image denoising
Jeon et al. (2018) [82]	CNN	Speckle noise reduction from digital holographic images	Speckle noise reduction of digital holographic image from Multi-scale CNN
Gholizadeh-Ansari et al. (2018) [56]	CNN	Low-dose CT image denoising, X-ray image denosing	CNN with dilated convolutions for low-dose CT image denoising
Uchida et al. (2018) [185]	CNN	Non-blind image denoising	CNN with residual learning for non-blind image denoising
Xiao et al. (2018) [195]	CNN	Stripe noise reduction of infrared cloud images	CNN with skip connection for infrared-cloud-image denoising
Chen et al. (2018) [30]	CNN	Gaussian image denoisng, blind denoising	CNN based on RL and perceptual loss for edge enhancement
Yu et al. (2018) [213]	CNN	Seismic, random, linear and multiple noise reduction of images	A survey on deep learning for three applications
Yu et al. (2018) [212]	CNN	Optical coherence tomography (OCT) image denoising	GAN with dense skip connection for optical coherence tomography image denoising
Li et al. (2018) [107]	CNN	Ground-roll noise reduction	An overview of deep learning techniques on ground-roll noise
Abbasi et al. (2018) [2]	CNN	OCT image denoising	Fully CNN with multiple inputs, and RL for OCT image denoising
Zarshenas et al. (2018) [218]	CNN	Gaussian noisy image denoising	Deep CNN with internal and external residual learning for image denoising
Chen et al. (2018) [26]	CNN	Gaussian noisy image denoising	CNN with recursive operations for image denoising
Panda et al. (2018) [149]	CNN	Gaussian noisy image denoising	CNN with exponential linear units, and dilated convolutions for image denoising
Sheremet et al. (2018) [163]	CNN	Image denoising from info-communication systems	CNN on image denoising from info-communication systems
Chen et al. (2018) [27]	CNN	Aerial-image denoising	CNN with multi-scale technique, and RL for aerial-image denoising
Pardasani et al. (2018) [150]	CNN	Gaussian, poisson or any additive-white noise reduction	CNN with BN for image denoising
Couturier et al. (2018) [37]	NN	Gaussian and multiplicative speckle noise reduction	Encoder-decoder network with multiple skip connections for image denoising
Park et al. (2018) [151]	CNN	Gaussian noisy image denoising	CNN with dilated convolutions for image denoising
Priyanka et al. (2019) [155]	CNN	Gaussian noisy image denoisng	CNN with symmetric network architecture for image denoising
Lian et al. (2019) [171]	CNN	Poisson-noise-image denoising	CNN with multi scale, and multiple skip connections for Poisson image denoising
Tripathi et al. (2018) [184]	CNN	Gaussian noisy image denoising	GAN for image denoising
Zheng et al. (2019) [234]	CNN	Gaussian noisy image denoising	CNN for image denoising
Remez et al. (2018) [158]	CNN	Gaussian and Poisson image denoising	CNN for image denoising

Tian, Chunwei, Lunke Fei, Wenxian Zheng, Yong Xu, Wangmeng Zuo, and Chia-Wen Lin. "Deep Learning on Image Denoising: An overview." arXiv preprint arXiv:1912.13171 (2019).

基于深度学习的图像去噪

❑ 研究现状

– 这篇综述介绍了超过200篇关于深度学习去噪的论文

Table 4: CNNs for real noisy image denoising.

References	Methods	Applications	Key words (remarks)
Tao et al. (2019) [177]	CNN	Real noisy image denoising, low-light image enhancement	CNN with ReLU, and RL for real noisy image denoising
Chen et al. (2018) [28]	CNN	Real noisy image denoising, blind denoising	GAN for real noisy image denoising
Han et al. (2018) [66]	CNN	CT image reconstruction	U-Net with skip connection for CT image reconstruction
Chen et al. (2018) [29]	CNN	Real noisy image denoising	CNNs with anisotropic parallax analysis for real noisy image denoising
Jian et al. (2018) [86]	CNN	Low-light remote sense image denoising	CNN for image denoising
Khoroushadi et al. (2019) [96]	CNN	Medical image denoising, CT image denoising	CNN for image denoising
Jiang et al. (2018) [88]	CNN	Low-light image enhancement	CNN with symmetric pathways for low-light image enhancement
Godard et al. (2018) [57]	CNN	Real noisy image denoising	CNN with recurrent connections for real noisy image denoising
Zhao et al. (2019) [232]	CNN	Real noisy image denoising	CNN with recurrent connections for real noisy image denoising
Anwar et al. (2019) [10]	CNN	Real noisy image denoising	CNN with RL, attention mechanism for real noisy image denoising
Jaroensri et al. (2019) [191]	CNN	Real noisy image denoising	CNN for real noisy image denoising
Green et al. (2018) [60]	CNN	CT image denoising, real noisy image denoising	CNN for real noisy image denoising
Brooks et al. (2019) [21]	CNN	Real noisy image denoising	CNN with image processing pipeline for real noisy image denoising
Tian et al. (2020) [182]	CNN	Gaussian image denoising and real noisy image denoising	CNN with BRN, RL and dilated convolutions for image denosing
Tian et al. (2020) [181]	CNN	Gaussian image denoising, blind denoising and real noisy image denoising	CNN with attention mechanism and sparse method for image denoising

Table 6: Deep learning techniques for blind denoising.

References	Methods	Applications	Key words (remarks)
Zhang et al. (2018) [224]	CNN	Blind denoising	CNN with varying noise level for blind denoising
Kenzo et al. (2018) [79]	CNN	Blind denoising	CNN with soft shrinkage for blind denoising
Soltanayev et al. (2018) [168]	CNN	Blind denoising	CNN for unpaired noisy images
Yang et al. (2017) [206]	CNN	Blind denoising	CNNs with RL for blind denoising
Zhang et al. (2018) [220]	CNN	Blind denoising, random noise	CNN with RL for blind denoising
Si et al. (2018) [165]	CNN	Blind denoising, random noise	CNN for image denoising
Majumdar et al. (2018) [88]	NN	Blind denoising	Auto-encoder for blind denoising
Abiko et al. (2019) [57]	CNN	Blind denoising, complex noisy image denoising	cascaded CNNs for blind denoising
Cha et al. (2019) [206]	CNN	Blind denoising	GAN for blind image denoising

Table 7: Deep learning techniques for hybrid noisy image denoising.

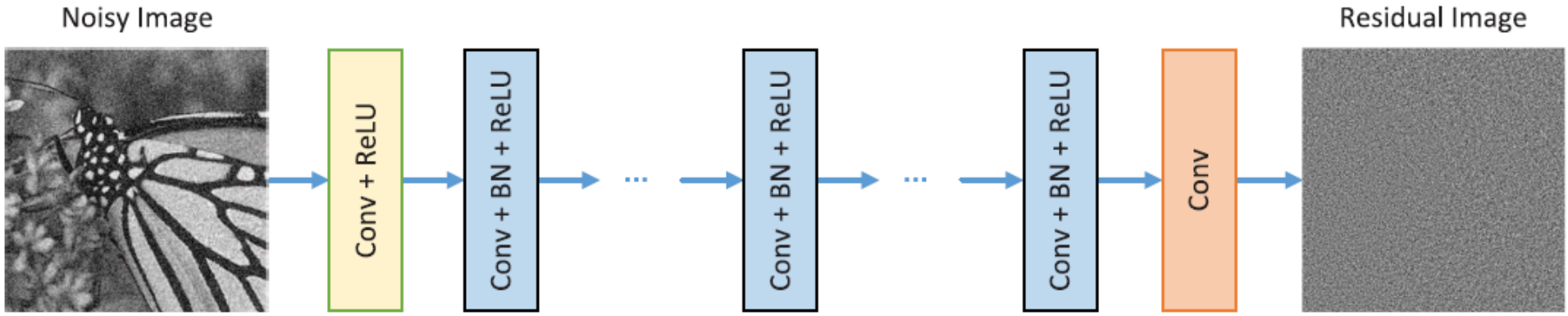
References	Methods	Applications	Key words (remarks)
Li et al. (2018) [111]	CNN	Noise, blur kernel, JPEG compression	The combination of CNN and warped guidance for multiple degradations
Zhang et al. (2018) [225]	CNN	Noise, blur kernel, low-resolution image	CNN for multiple degradations
Kokkinos et al. (2019) [97]	CNN	Image demosaicking and denoising	Residual CNN with iterative algorithm for image demosaicking and denoising

Tian, Chunwei, Lunke Fei, Wenxian Zheng, Yong Xu, Wangmeng Zuo, and Chia-Wen Lin. "Deep Learning on Image Denoising: An overview." arXiv preprint arXiv:1912.13171 (2019).

基于深度学习的图像去噪

❑ DnCNN

- 面向高斯噪声的端到端深度卷积去噪网络
- 使用residual learning和batch normalization
- 单一DnCNN模型处理盲的高斯去噪问题、JPEG deblocking

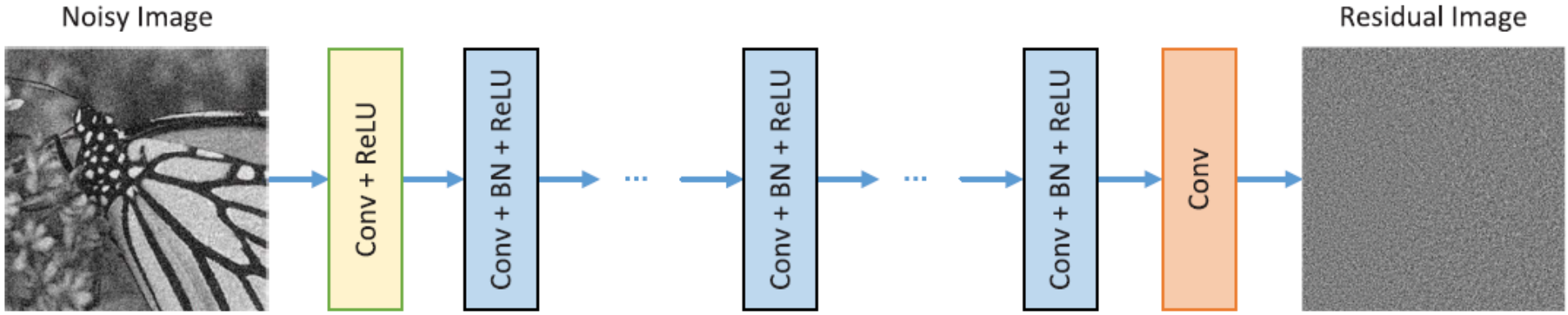


Zhang, Kai, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. "Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising." *IEEE Transactions on Image Processing* 26, no. 7 (2017): 3142-3155.

基于深度学习的图像去噪

❑ DnCNN

- 目标函数： $\ell(\Theta) = \frac{1}{2N}\sum_{i=1}^{N} \|\mathcal{R}(\mathbf{y}_i;\Theta) - (\mathbf{y}_i - \mathbf{x}_i)\|_F^2$
- 复原图像： $\boldsymbol{x}_i = \boldsymbol{y}_i - \mathcal{R}(\boldsymbol{y}_i)$
- 这里与传统的residual learning不同，DnCNN并非每隔两层就加一个shortcut connection，而是将网络的输出直接改成residual image（残差图片）



Zhang, Kai, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. "Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising." *IEEE Transactions on Image Processing* 26, no. 7 (2017): 3142-3155.

基于深度学习的图像去噪

❑ DnCNN性能分析

– DnCNN-S和DnCNN-B对应特定噪声强度和盲的噪声强度

THE PSNR(dB) RESULTS OF DIFFERENT METHODS ON 12 WIDELY USED TESTING IMAGES

Images	C.man	House	Peppers	Starfish	Monar.	Airpl.	Parrot	Lena	Barbara	Boat	Man	Couple	Average
Noise Level	$\sigma = 15$												
BM3D [2]	31.91	34.93	32.69	31.14	31.85	31.07	31.37	34.26	33.10	32.13	31.92	32.10	32.372
WNNM [13]	32.17	**35.13**	32.99	31.82	32.71	31.39	31.62	34.27	**33.60**	32.27	32.11	32.17	32.696
EPLL [33]	31.85	34.17	32.64	31.13	32.10	31.19	31.42	33.92	31.38	31.93	32.00	31.93	32.138
CSF [14]	31.95	34.39	32.85	31.55	32.33	31.33	31.37	34.06	31.92	32.01	32.08	31.98	32.318
TNRD [16]	32.19	34.53	33.04	31.75	32.56	31.46	31.63	34.24	32.13	32.14	32.23	32.11	32.502
DnCNN-S	**32.61**	34.97	**33.30**	**32.20**	**33.09**	**31.70**	**31.83**	**34.62**	32.64	**32.42**	**32.46**	**32.47**	**32.859**
DnCNN-B	32.10	34.93	33.15	32.02	32.94	31.56	31.63	34.56	32.09	32.35	32.41	32.41	32.680
Noise Level	$\sigma = 25$												
BM3D [2]	29.45	32.85	30.16	28.56	29.25	28.42	28.93	32.07	30.71	29.90	29.61	29.71	29.969
WNNM [13]	29.64	**33.22**	30.42	29.03	29.84	28.69	29.15	32.24	**31.24**	30.03	29.76	29.82	30.257
EPLL [33]	29.26	32.17	30.17	28.51	29.39	28.61	28.95	31.73	28.61	29.74	29.66	29.53	29.692
MLP [24]	29.61	32.56	30.30	28.82	29.61	28.82	29.25	32.25	29.54	29.97	29.88	29.73	30.027
CSF [14]	29.48	32.39	30.32	28.80	29.62	28.72	28.90	31.79	29.03	29.76	29.71	29.53	29.837
TNRD [16]	29.72	32.53	30.57	29.02	29.85	28.88	29.18	32.00	29.41	29.91	29.87	29.71	30.055
DnCNN-S	**30.18**	33.06	**30.87**	**29.41**	**30.28**	**29.13**	**29.43**	**32.44**	30.00	**30.21**	**30.10**	**30.12**	**30.436**
DnCNN-B	29.94	33.05	30.84	29.34	30.25	29.09	29.35	32.42	29.69	30.20	30.09	30.10	30.362
Noise Level	$\sigma = 50$												
BM3D [2]	26.13	29.69	26.68	25.04	25.82	25.10	25.90	29.05	27.22	26.78	26.81	26.46	26.722
WNNM [13]	26.45	**30.33**	26.95	25.44	26.32	25.42	26.14	29.25	**27.79**	26.97	26.94	26.64	27.052
EPLL [33]	26.10	29.12	26.80	25.12	25.94	25.31	25.95	28.68	24.83	26.74	26.79	26.30	26.471
MLP [24]	26.37	29.64	26.68	25.43	26.26	25.56	26.12	29.32	25.24	27.03	27.06	26.67	26.783
TNRD [16]	26.62	29.48	27.10	25.42	26.31	25.59	26.16	28.93	25.70	26.94	26.98	26.50	26.812
DnCNN-S	**27.03**	30.00	27.32	25.70	26.78	25.87	**26.48**	**29.39**	26.22	27.20	**27.24**	26.90	27.178
DnCNN-B	**27.03**	30.02	**27.39**	**25.72**	**26.83**	**25.89**	**26.48**	29.38	26.38	**27.23**	27.23	**26.91**	**27.206**

Zhang, Kai, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. "Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising." *IEEE Transactions on Image Processing* 26, no. 7 (2017): 3142-3155.

基于深度学习的图像去噪

❑ DnCNN性能分析

– DnCNN-S和DnCNN-B对应特定噪声强度和盲的噪声强度

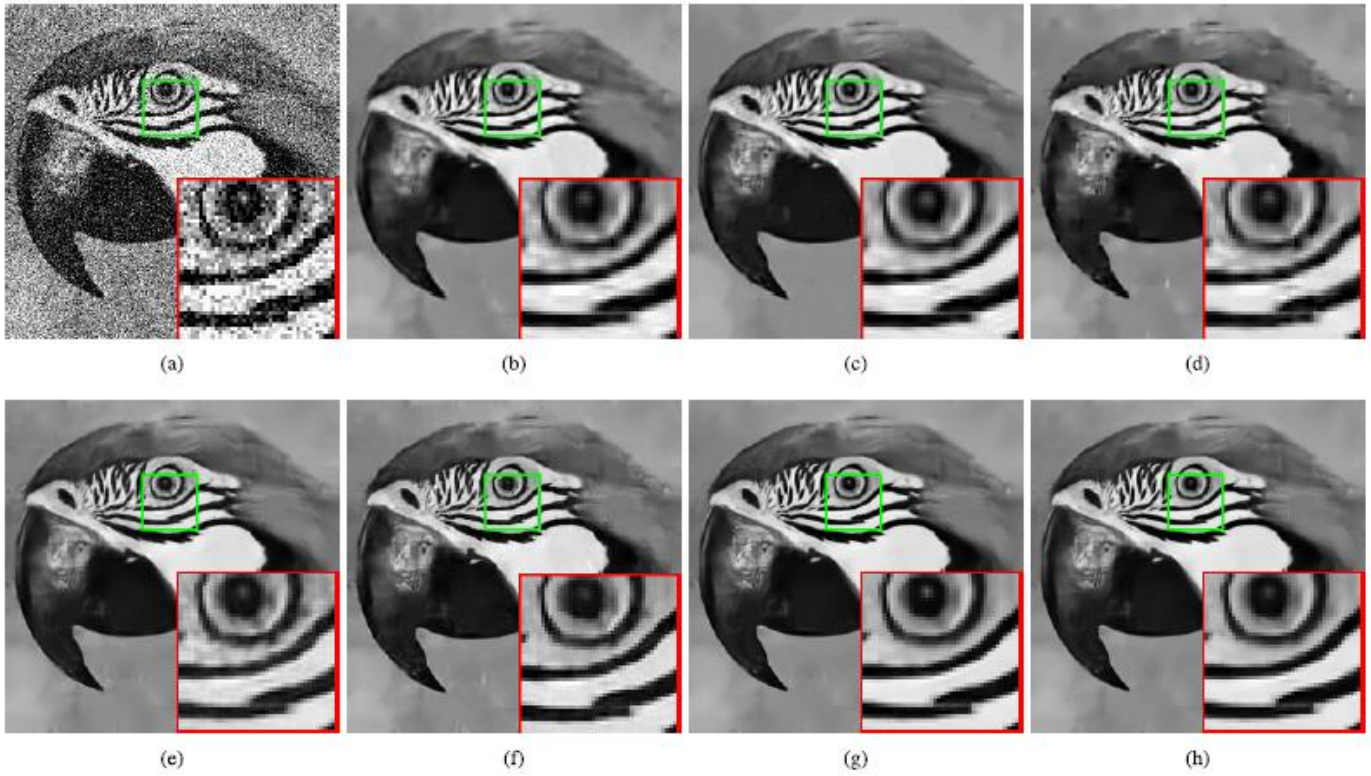


Fig. 5. Denoising results of the image "parrot" with noise level 50. (a) Noisy / 15.00dB. (b) BM3D / 25.90dB. (c) WNNM / 26.14dB. (d) EPLL / 25.95dB. (e) MLP / 26.12dB. (f) TNRD / 26.16dB. (g) DnCNN-S / 26.48dB. (h) DnCNN-B / 26.48dB.

Zhang, Kai, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. "Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising." *IEEE Transactions on Image Processing* 26, no. 7 (2017): 3142-3155.

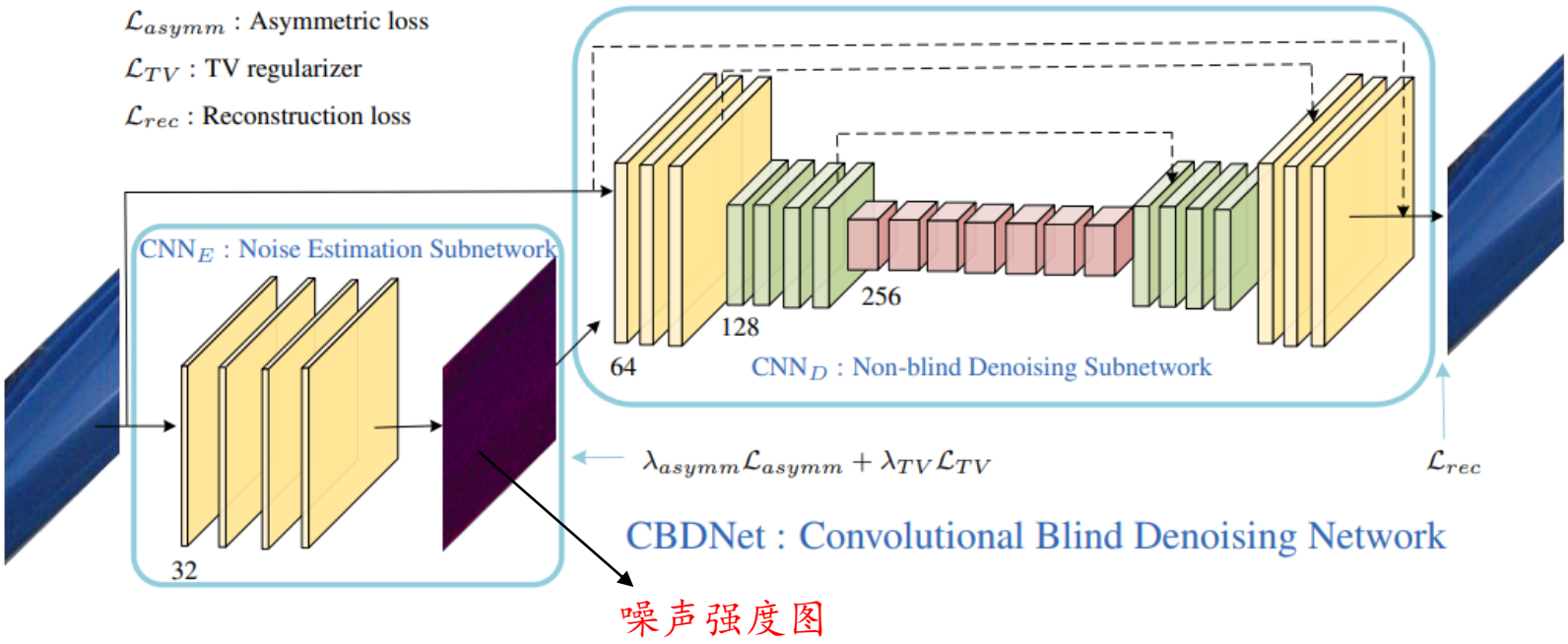
基于深度学习的图像去噪

❑ CBDNet

– 为了解决盲去噪的鲁棒性，增加了噪声估计子网络

• CBDNet包括噪声估计子网CNN_E和非盲去噪子网CNN_D

– 更加真实的噪声模型，以及引入非对称损失函数

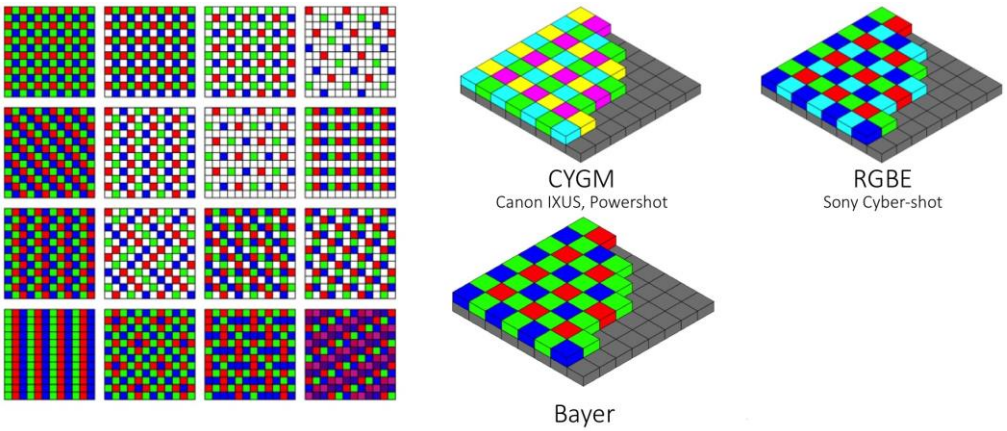


Guo S, Yan Z, Zhang K, et al. Toward convolutional blind denoising of real photographs[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 1712-1722.

基于深度学习的图像去噪

❑ CBDNet

- 独立同分布的高斯噪声模型不能准确反映实际噪声，因为实际的噪声是通常是复杂且信号依赖的
- 提出了多种噪声混合的模型，更加接近真实图像噪声
  - 成像过程中的噪声由泊松分布和高斯分布联合构成：$\mathbf{n}(\mathbf{L}) = \mathbf{n}_s(\mathbf{L}) + \mathbf{n}_c$
  - 图像处理过程中的噪声
    - Demosaicing和Gamma校正：$\boldsymbol{y} = f(\mathbf{DM}(\mathbf{L} + \mathbf{n}(\mathrm{L})))$



Guo S, Yan Z, Zhang K, et al. Toward convolutional blind denoising of real photographs[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 1712-1722.

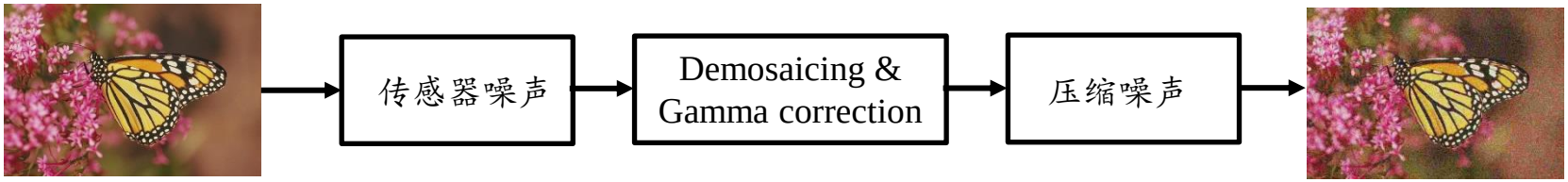
基于深度学习的图像去噪

❑ CBDNet

– 独立同分布的高斯噪声模型不能准确反映实际噪声，因为实际的噪声是通常是复杂且信号依赖的

– 提出了多种噪声混合的模型，更加接近真实图像噪声

- 成像过程中的噪声由泊松分布和高斯分布联合构成：$\mathbf{n}(\mathbf{L}) = \mathbf{n}_s(\mathbf{L}) + \mathbf{n}_c$
- 图像处理过程中的噪声
  - Demosaicing和Gamma校正：$\boldsymbol{y} = f(\mathbf{DM}(\mathbf{L} + \mathbf{n}(\mathrm{L})))$
  - JPEG压缩：$\boldsymbol{y} = \mathrm{JPEG}(f(\mathbf{DM}(\mathbf{L} + \mathbf{n}(\mathrm{L}))))$



Guo S, Yan Z, Zhang K, et al. Toward convolutional blind denoising of real photographs[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 1712-1722.

基于深度学习的图像去噪

❑ CBDNet

– 非对称损失函数：根据实际效果，当估计的噪声强度高于实际噪声强度时，去噪具有较强的鲁棒性，可以得到较好的图像质量；当估计的噪声强度低于实际噪声强度时，图像中会残留较多噪声，质量不高

– 因此，训练过程中目标函数引入非对称约束

$$\mathcal{L}_{asymm} = \sum_{\cdot} |\alpha - \mathbb{I}_{(\hat{\sigma}(y_i)-\sigma(y_i))<0}| \cdot (\hat{\sigma}(y_i) - \sigma(y_i))^2$$

其中，$\hat{\sigma}(y_i)$是对第i个像素估计的噪声，$\sigma(y_i)$是第i个像素的真实噪声，如果e<0,$\mathbb{I}_e$=1,否则为0, $0 < \alpha < 0.5$

– 总的优化目标函数： $\mathcal{L} = \mathcal{L}_{rec} + \lambda_{asymm}\mathcal{L}_{asymm} + \lambda_{TV}\mathcal{L}_{TV}$

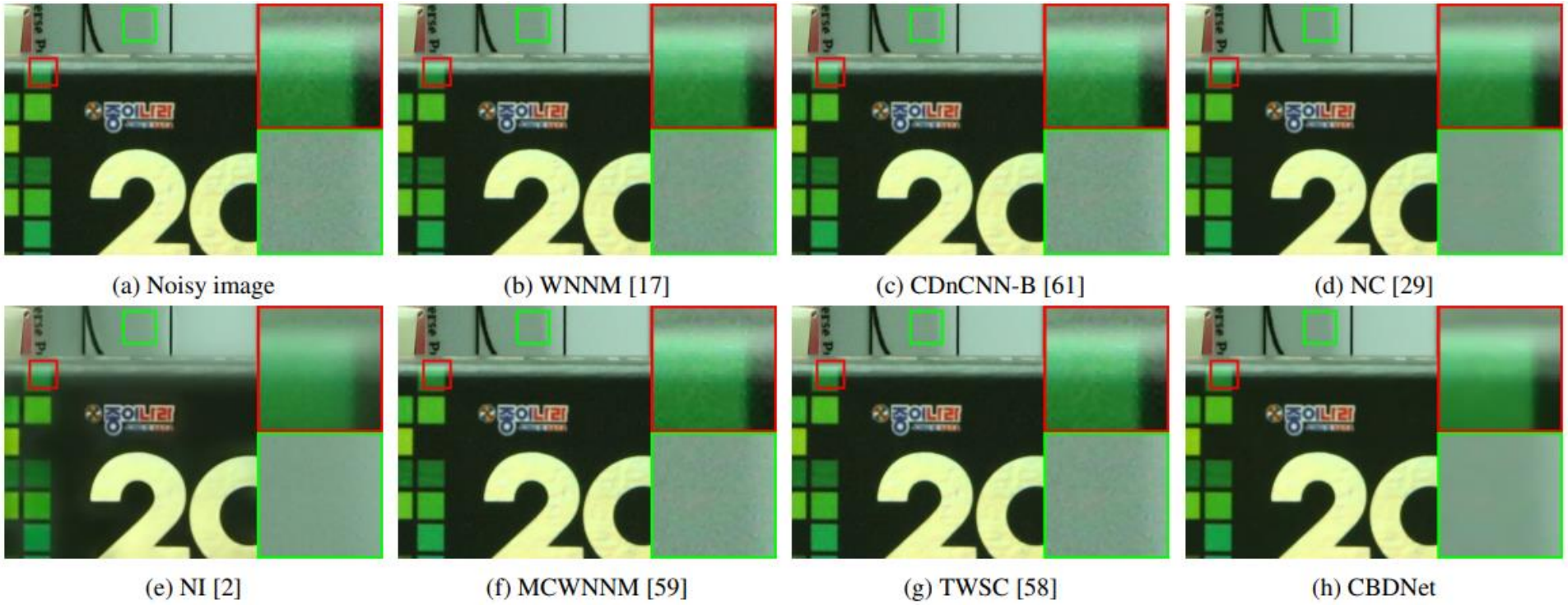
$$\mathcal{L}_{rec} = \|\hat{\mathbf{x}} - \mathbf{x}\|_2^2 \qquad \mathcal{L}_{TV} = \|\nabla_h\hat{\sigma}(\mathbf{y})\|_2^2 + \|\nabla_v\hat{\sigma}(\mathbf{y})\|_2^2$$

Guo S, Yan Z, Zhang K, et al. Toward convolutional blind denoising of real photographs[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 1712-1722.

基于深度学习的图像去噪

❑ CBDNet性能分析

– 在实际的带噪图像上的结果



(a) Noisy image (b) WNNM [17] (c) CDnCNN-B [61] (d) NC [29]

(e) NI [2] (f) MCWNNM [59] (g) TWSC [58] (h) CBDNet

Guo S, Yan Z, Zhang K, et al. Toward convolutional blind denoising of real photographs[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 1712-1722.

图像增强

- **华为P20手持超级夜景**
  - 多种图像增强技术：图像的选帧、图像的降噪、图像的增强、重影的消除、对比度调整

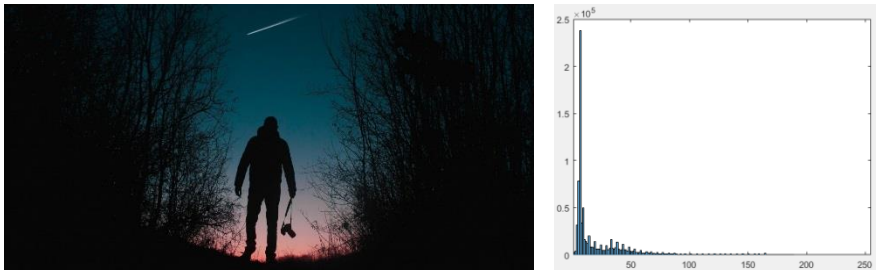
弱光照增强

❑ 问题描述

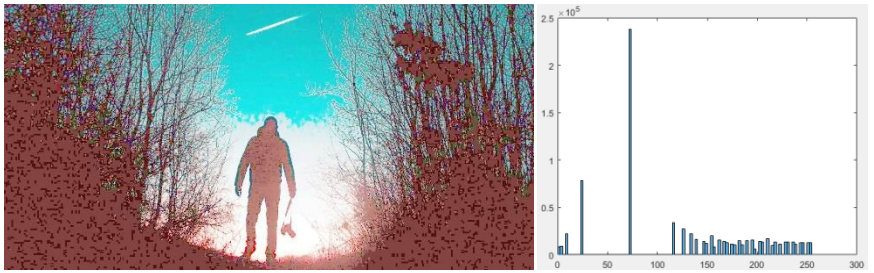
– 解决暗光拍照由于光线不足，导致的欠曝光或者对比度不足的问题

❑ 解决手段

– 根本上是调整图像的对比度



低光照图像



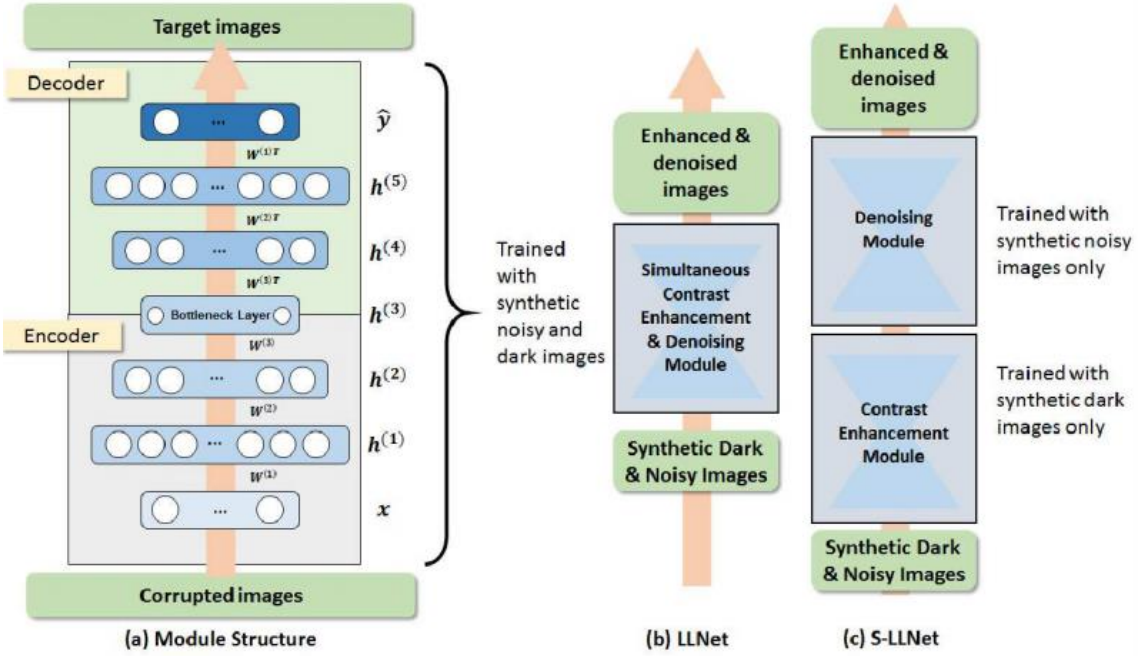
直方图均衡化调整图像

基于深度学习的夜景增强

❑ LLNet (Low-light Net)

– 探索了两种类型的网络结构

- (a) LLNet，同时学习对比度增强和去噪；
- (b) S-LLNet，使用两个模块分阶段执行对比度增强和去噪。



Lore, K. G., Akintayo, A., & Sarkar, S. (2017). LLNet: A deep autoencoder approach to natural low-light image enhancement. Pattern Recognition, 61, 650-662.

基于深度学习的夜景增强

❑ LLNet (Low-light Net)

– 提出了一种训练数据生成方法（即伽马校正和添加高斯噪声）来模拟低光环境

– 在真实拍摄到的低光照图像上进行了实验，证明了用合成数据训练的模型的有效性

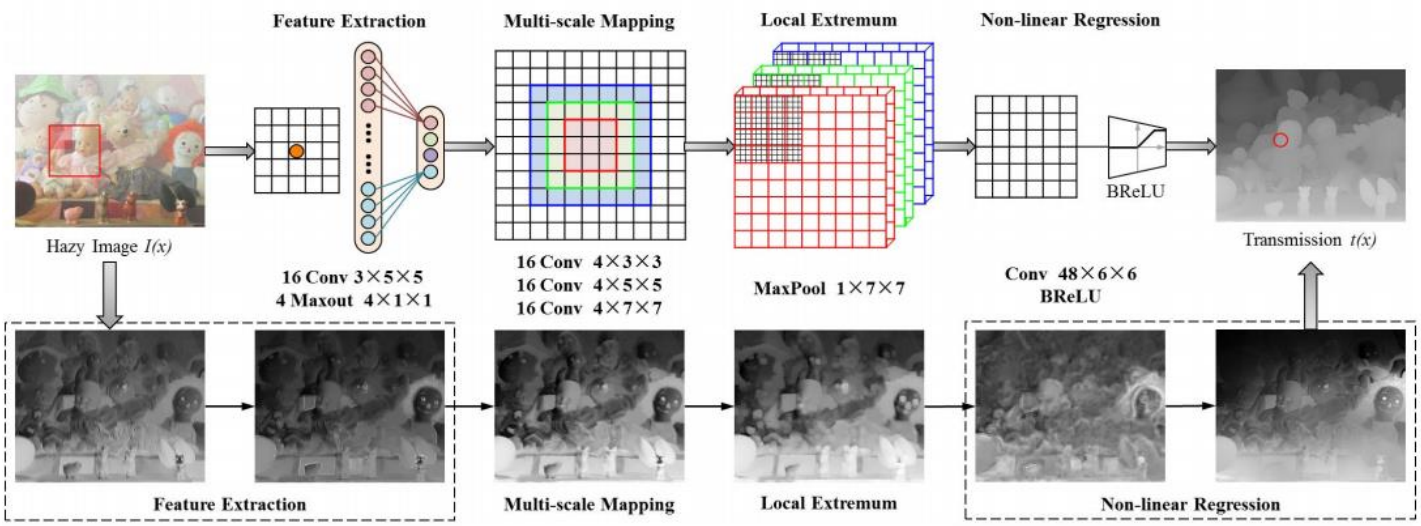
Lore, K. G., Akintayo, A., & Sarkar, S. (2017). LLNet: A deep autoencoder approach to natural low-light image enhancement. Pattern Recognition, 61, 650-662.

基于深度学习的去雾

❑ DehazeNet

– 去雨去雾的研究可以认为是图像增强，在恶劣天气环境下如何拍摄高质量图像视频上有很多应用，2009年何凯明的去雾论文获得CVPR最佳论文[1]

– DehazeNet 是一个end-to-end的去雾网络架构



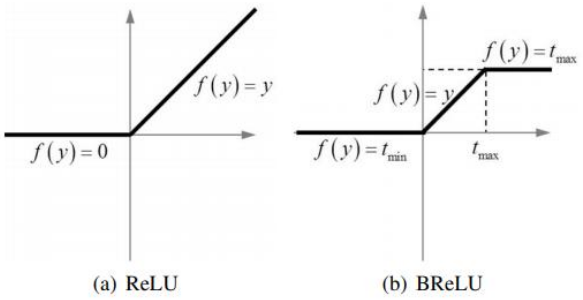
[1] Kaiming He, Jian Sun and Xiaoou Tang, "Single image haze removal using dark channel prior," 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, 2009, pp. 1956-1963.

[2] Cai, B., Xu, X., Jia, K., Qing, C., & Tao, D. (2016). Dehazenet: An end-to-end system for single image haze removal. IEEE Transactions on Image Processing, 25(11), 5187-5198.

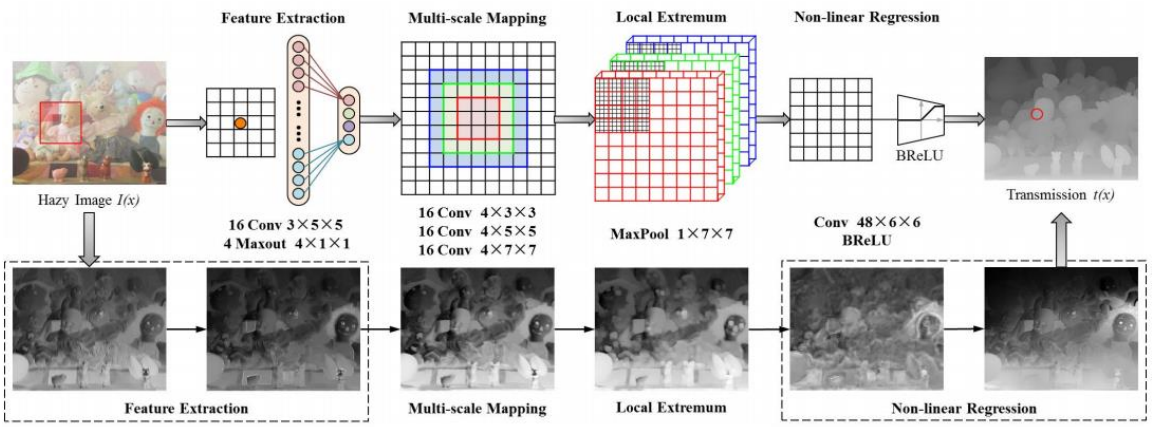
基于深度学习的去雾

❑ DehazeNet

- 由4个模块构成：feature extraction, multi-scale mapping, local extremum，non-linear regression
- 提出了新的激活函数BReLU（Bilateral Rectified Linear Unit）



(a) ReLU (b) BReLU



基于深度学习的去雾

❑ DehazeNet性能

– 仿真的数值结果

Metric	ATM [39]	BCCR [11]	FVR [38]	DCP [9]	CAP[2][18]	RF [17]	DehazeNet
MSE	0.0689	0.0243	0.0155	0.0172	0.0075 (0.0068)	0.0070	**0.0062**
SSIM	0.9890	0.9963	0.9973	0.9981	0.9991 (0.9990)	0.9989	**0.9993**
PSNR	60.8612	65.2794	66.5450	66.7392	70.0029 (70.6581)	70.0099	**70.9767**
WSNR	7.8492	12.6230	13.7236	13.8508	16.9873 (17.7839)	17.1180	**18.0996**

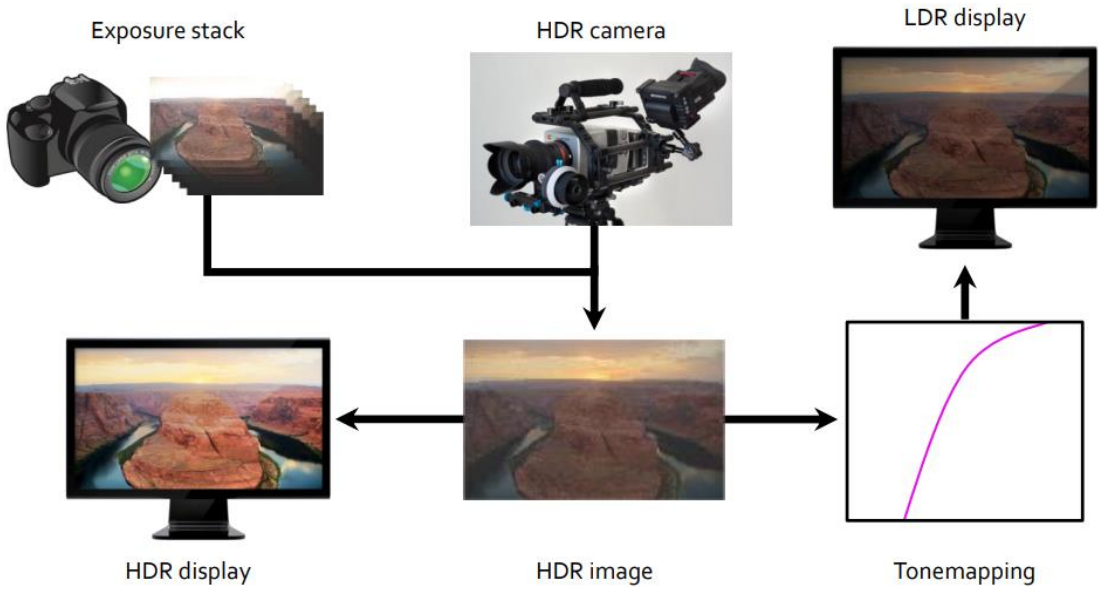
– 真实图像上的主观结果

(a) Hazy image (b) ATM (c) BCCR (d) FVR (e) DCP (f) CAP (g) RF (h) DehazeNet

动态范围增强

❑ 标准动态图像到高动态图像的转化

- 和超分辨率应用类似，显示设备动态范围与视频/图像动态范围不匹配的矛盾
- Tone Mapping技术实现高动态视频/图像内容在低动态显示设备上显示
- Inverse Tone Mapping技术实现低动态视频/图像内容在高动态显示设备上显示



基于深度学习的动态范围增强

❑ Deep Revers Tone Mapping

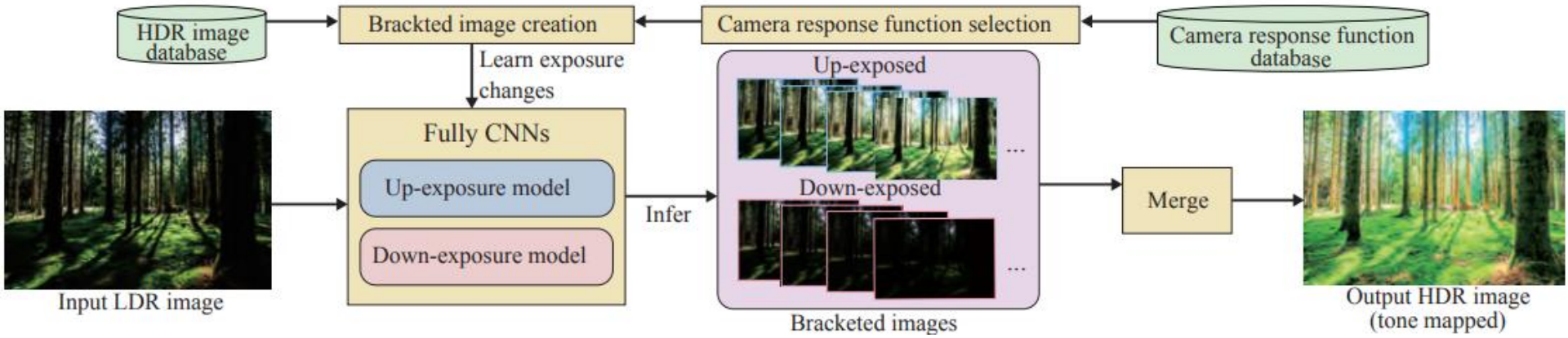
- 第一个通过深度学习框架获得的从LDR到HDR的方法，目标实现8 bit的LDR图像到32bit的HDR图像转化
- 难点：
  - 即使LDR的曝光度不同，相同的HDR图片应具有输出一致性
  - HDR像素含有比LDR像素更多的变化，意味着训练数据的急剧增加
  - LDR数据集很多，但是HDR的数据集就很少了
  - 即使loss上微小的变化也会影响输出结果，使得训练的模型不稳定

Endo, Y., Kanamori, Y., & Mitani, J. (2017). Deep reverse tone mapping. ACM Trans. Graph., 36(6), 177-1.

基于深度学习的动态范围增强

❑ Deep Revers Tone Mapping提出的方案

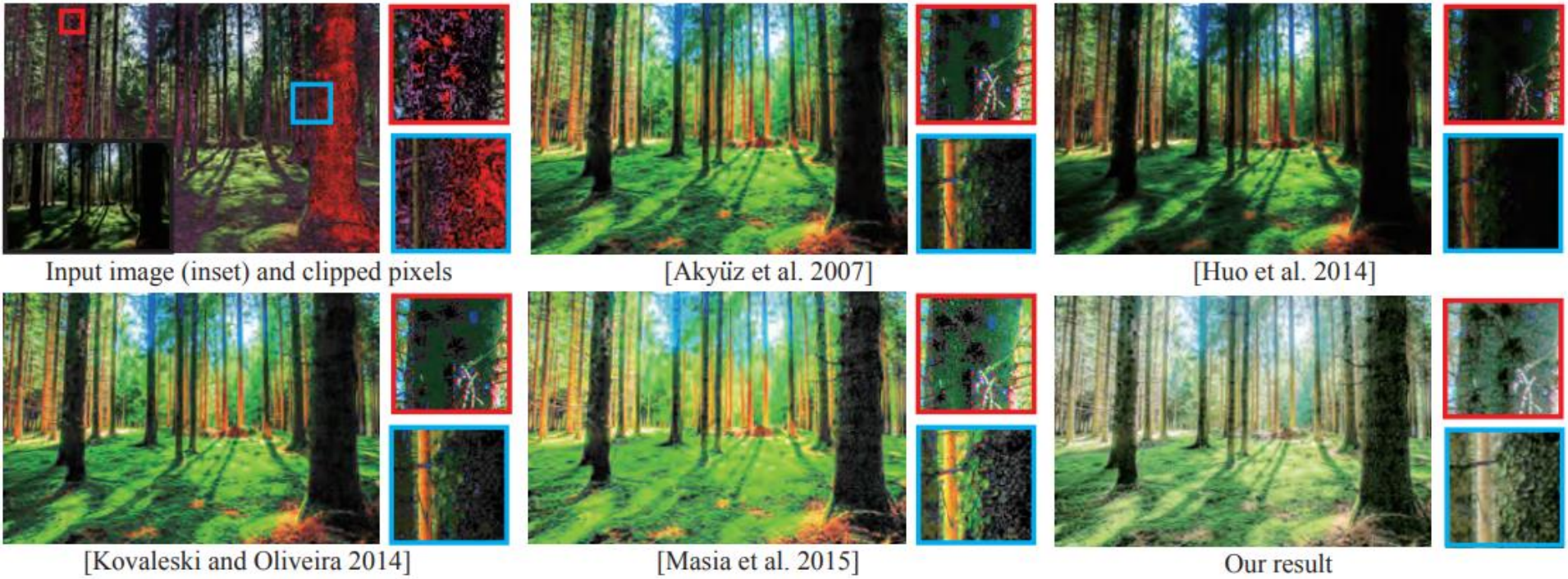
– 合成多曝光图像，通过多个不同曝光的LDR图像来推断HDR



Endo, Y., Kanamori, Y., & Mitani, J. (2017). Deep reverse tone mapping. ACM Trans. Graph., 36(6), 177-1.

基于深度学习的动态范围增强

❑ 性能对比



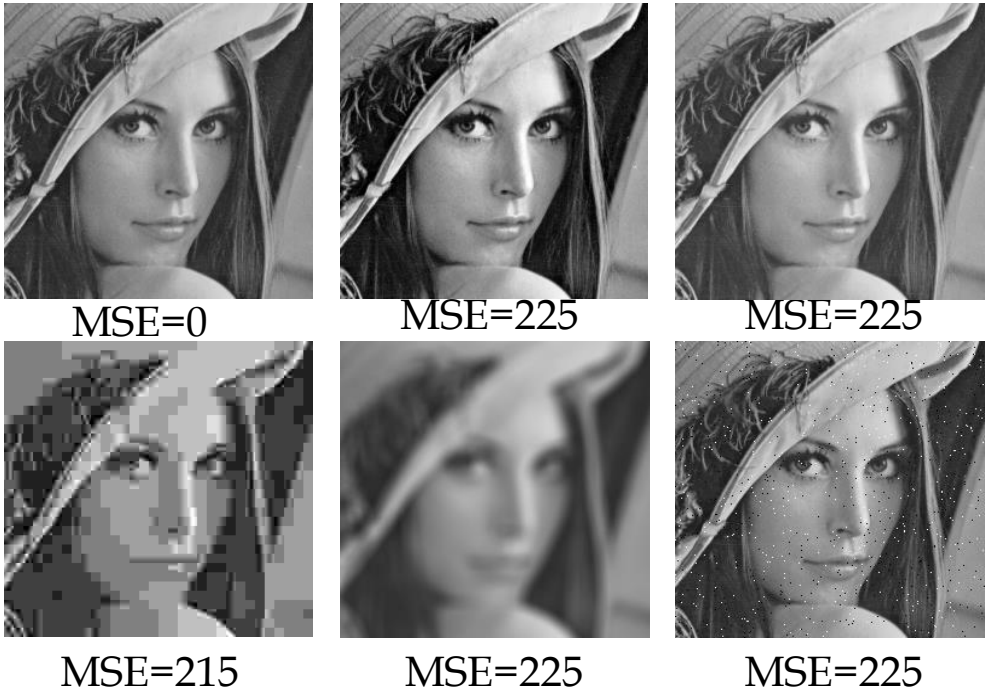
Endo, Y., Kanamori, Y., & Mitani, J. (2017). Deep reverse tone mapping. ACM Trans. Graph., 36(6), 177-1.

图像质量评价

❑ 常见质量评价: PSNR, SSIM

- $p = 10\log_{10}\frac{v_{max}^2}{MSE}$
- $s = \frac{(2\mu_X\mu_Y+C_1)(2\sigma_{XY}+C_2)}{(\mu_X^2+\mu_Y^2+C_1)(\sigma_X^2+\sigma_Y^2+C_2)}$

❑ 质量评价的挑战

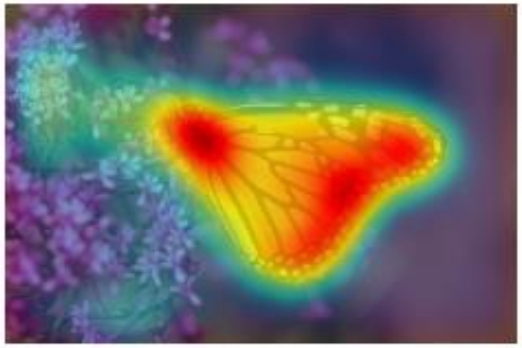


图像质量评价

❑ 传统质量评价问题研究

- 研究视觉敏感的图像结构特征表达
- 研究人眼视觉系统特性，比如显著性、掩蔽效应、JND （just-noticeable difference）






图像质量评价

❑ 图像质量评价研究目标

– 设计的评价算法和人标注的质量相关度尽可能的高

❑ 图像/视频质量评价数据库

– 目前针对深度学习质量评价严重问题是数据库不足，常用的数据增强方法不适用

– 常见数据库

SUMMARY OF IQA DATABASES WITH COMPRESSION DISTORTIONS.

Datasets	No. of Ref. Images	No. of Dist. Types	No. of Dist. Images	No. of JPEG images/Levels	No. of JPEG 2000 images/Levels	Subjective Score Types
TID2008	25	17	1700	4	4	MOS
TID2013	25	24	3000	5	5	MOS
Toyama	14	2	182	6	7	MOS
CSIQ	30	6	900	5	5	DMOS
CIDIQ	23	6	690	5	5	MOS

图像质量评价

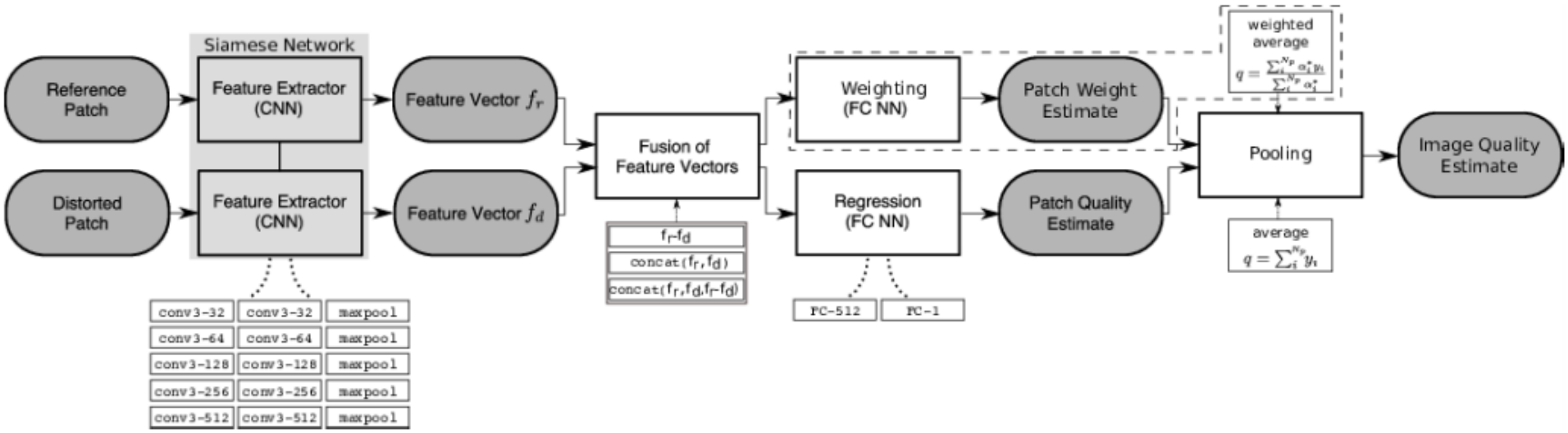
❑ 图像质量评价研究方法

- 全参考质量评价方法（full reference quality assessment）
- 半参考质量评价方法（Reduced reference quality assessment）
- 无参考质量评价方法（No reference quality assessment）

基于深度学习的图像质量评价

❑ DIQaM

- 既支持无参考质量评价又支持全参考质量评价
- 联合训练图像局部质量和局部权重（即局部质量对全局质量的重要性）
  - 全参考质量评价网络结构



Bosse, Sebastian, Dominique Maniry, Klaus-Robert Müller, Thomas Wiegand, and Wojciech Samek. "Deep neural networks for no-reference and full-reference image quality assessment." IEEE Transactions on Image Processing 27, no. 1 (2017): 206-219.

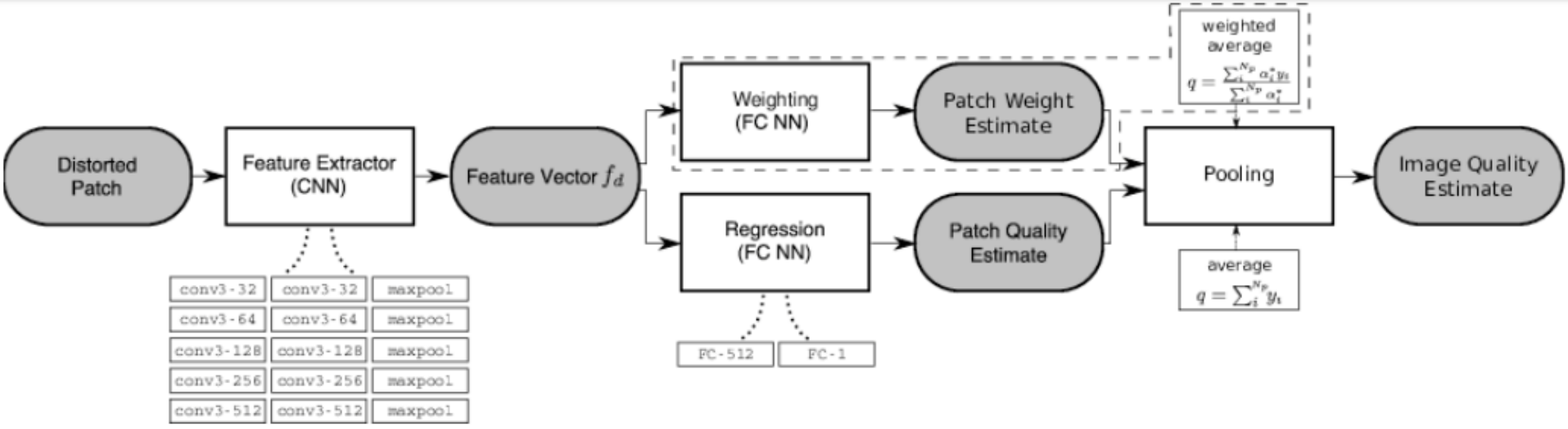
基于深度学习的图像质量评价

❑ DIQaM

– 既支持无参考质量评价又支持全参考质量评价

– 联合训练图像局部质量和局部权重（即局部质量对全局质量的重要性）

- 无参考质量评价网络结构



Bosse, Sebastian, Dominique Maniry, Klaus-Robert Müller, Thomas Wiegand, and Wojciech Samek. "Deep neural networks for no-reference and full-reference image quality assessment." IEEE Transactions on Image Processing 27, no. 1 (2017): 206-219.

基于深度学习的图像质量评价

❑ DIQaM性能分析

	IQM	LIVE		TID2013	
		LCC	SROCC	LCC	SROCC
Full-Reference	PSNR	0.872	0.876	0.675	0.687
	SSIM [8]	0.945	0.948	0.790	0.742
	$FSIM_C$ [10]	0.960	0.963	0.877	0.851
	GMSD [11]	0.956	0.958	-	-
	DOG-SSIM [15]	0.963	0.961	0.919	0.907
	DeepSim [14]	0.968	**0.974**	0.872	0.846
	DIQaM-FR (proposed)	0.977	0.966	0.880	0.859
	WaDIQaM-FR (proposed)	**0.980**	0.970	**0.946**	**0.940**
No-Reference	BLIINDS-II[19]	0.916	0.912	0.628	0.536
	DIIVINE [18]	0.923	0.925	0.654	0.549
	BRISQUE [20]	0.942	0.939	0.651	0.573
	NIQE [21]	0.915	0.914	0.426	0.317
	BIECON [28]	0.962	0.961	-	-
	FRIQUEE [22]	0.930	0.950	-	-
	CORNIA [24]	0.935	0.942	0.613	0.549
	CNN [27]	0.956	0.956	-	-
	SOM [25]	0.962	**0.964**	-	-
	DIQaM-NR (proposed)	**0.972**	0.960	**0.855**	**0.835**
	WaDIQaM-NR (proposed)	0.963	0.954	0.787	0.761

Bosse, Sebastian, Dominique Maniry, Klaus-Robert Müller, Thomas Wiegand, and Wojciech Samek. "Deep neural networks for no-reference and full-reference image quality assessment." IEEE Transactions on Image Processing 27, no. 1 (2017): 206-219.

基于深度学习的图像质量评价

❑ NIMA

- 使用AVA和TID2013数据集训练
  - AVA包含大约255,000个图像，每一张都被大约200个熟练的摄影师评分过，分数范围 [1,10]，10是最高分
  - TID2013 Dataset，有3,000张图片。其中25张原始图片，每张有24种修改版，每个修改版有5个强度级别。(25*24*5=3000)。图片的label是平均分和标准差
  - 数据增强
    - 考虑到美学评分涉及取景和布局会影响分数的，所以random crop不可取
    - 作者认为左右flip的方式是可行的

Talebi, Hossein, and Peyman Milanfar. "NIMA: Neural image assessment." IEEE Transactions on Image Processing 27, no. 8 (2018): 3998-4011.

基于深度学习的图像质量评价

❑ NIMA

- 预测目标是主观分数的分布，不是打分的均值
- 基于图像分类架构，尝试了VGG16，Inception-v2，MobileNet

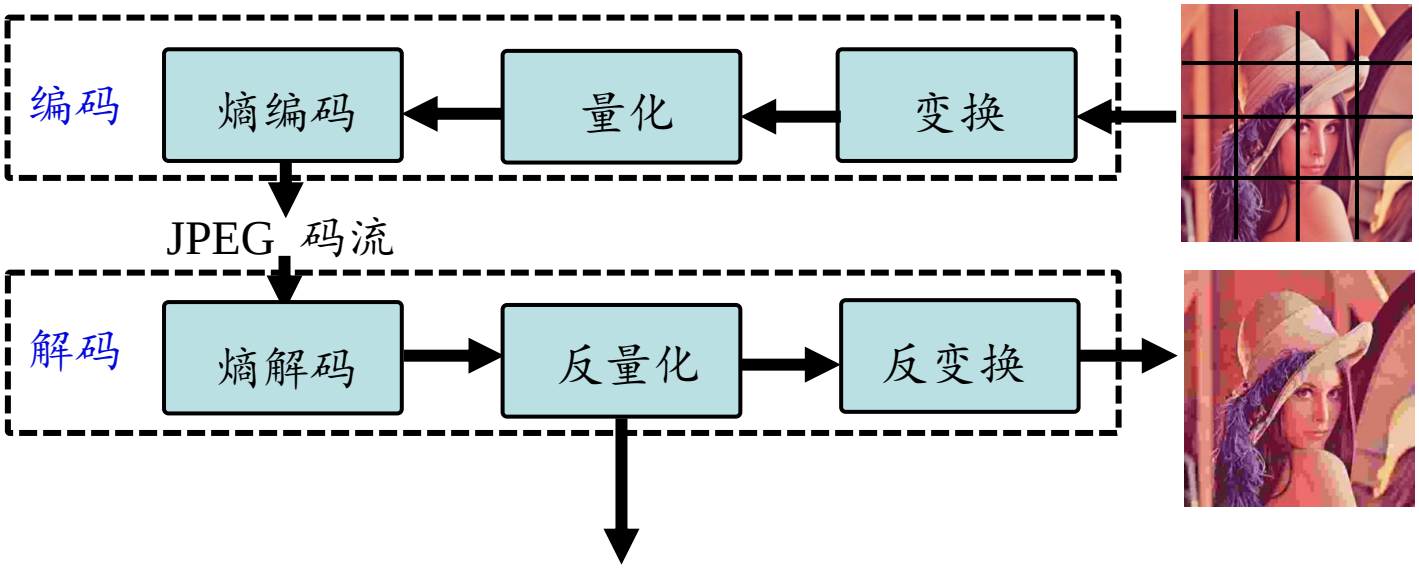
Model	*LCC (mean)*	*SRCC (mean)*	*LCC (std.dev)*	*SRCC (std.dev)*	*EMD*
Kim et al. [16]	0.80	0.80	–	–	–
Moorthy et al. [39]	0.89	0.88	–	–	–
Mittal et al. [40]	0.92	0.89	–	–	–
Saad et al. [41]	0.91	0.88	–	–	–
Kottayil et al. [42]	0.89	0.88	–	–	–
Xu et al. [35]	0.96	0.95	–	–	–
Bianco et al. [7]	0.96	0.96	–	–	–
NIMA(MobileNet)	0.782	0.698	0.209	0.181	0.105
NIMA(VGG16)	0.941	0.944	0.538	0.557	0.054
NIMA(Inception-v2)	0.827	0.750	0.470	0.468	0.064

中国科学院大学
University of Chinese Academy of Sciences

2 图像/视频压缩

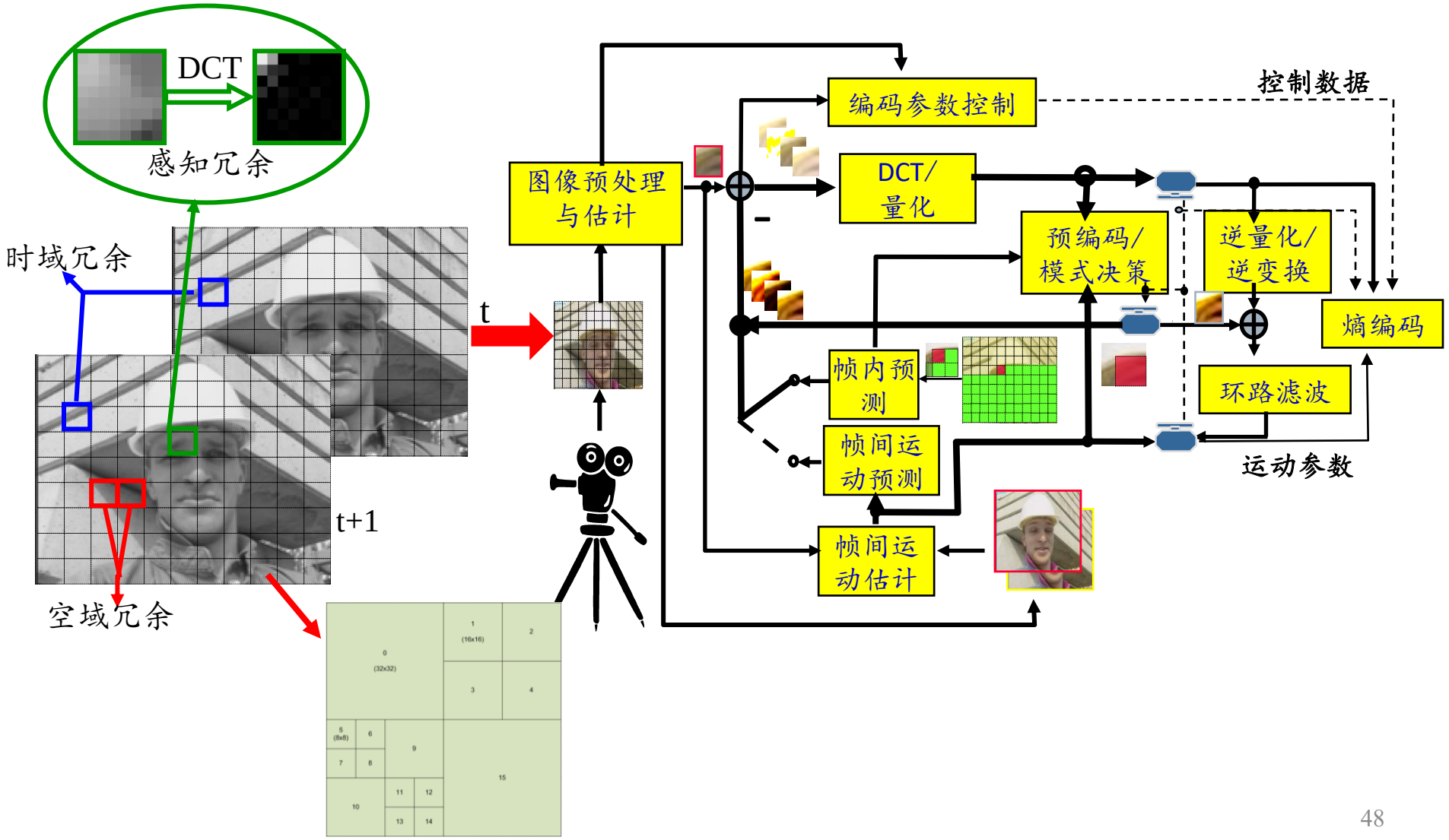
编码框架

❑ JPEG压缩



编码框架

❑ 视频编码


DCT
感知冗余
时域冗余
t
t+1
空域冗余
图像预处理与估计
编码参数控制
控制数据
DCT/量化
预编码/模式决策
逆量化/逆变换
熵编码
帧内预测
环路滤波
帧间运动预测
帧间运动估计
运动参数

视频编码标准

❑ 视频编码标准

- 规范化的码流格式，使得不同的解码器或者芯片可以播放不同来源的视频

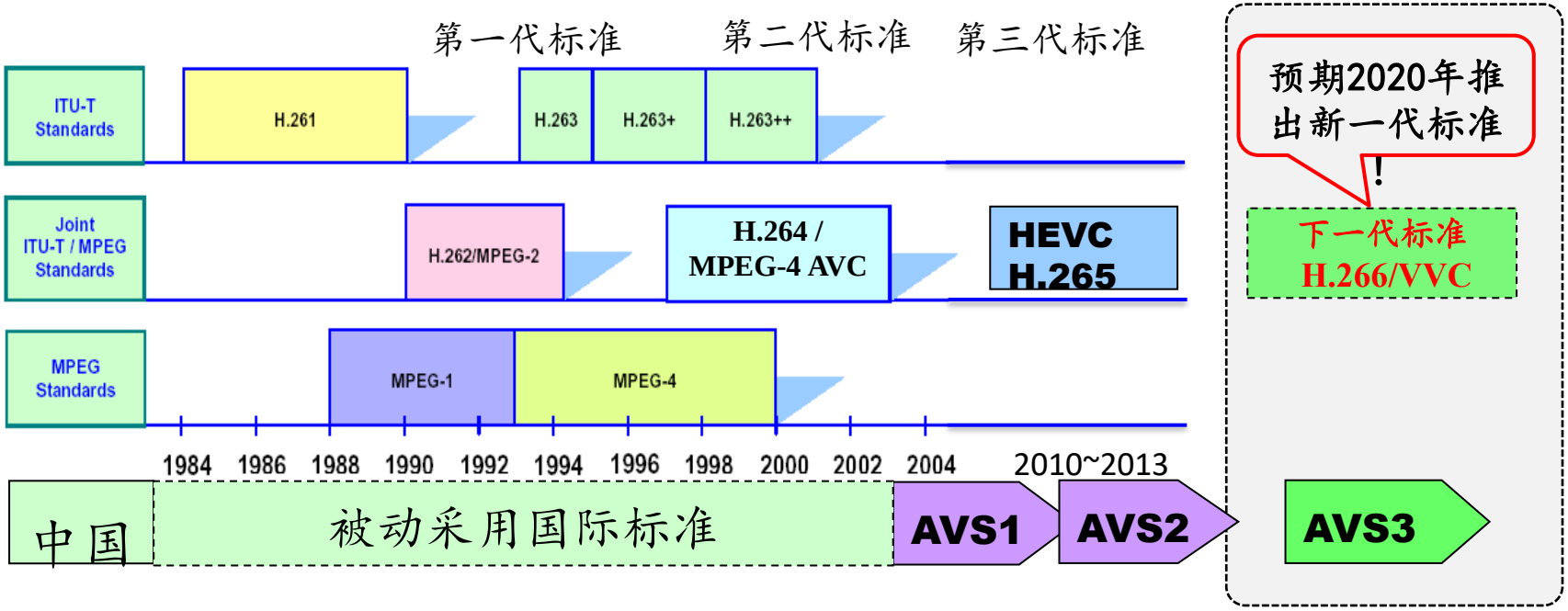
❑ 标准化组织

- 国际组织：ISO/IEC MPEG 、 ITU-T VCEG等
  - MPEG-2, H.264/AVC, H.265/HEVC, H.266/VVC
- 行业组织：AOM (Alliance for Open Media)
  - 代表公司：Google，Netflix, Facebook 等
  - AV1
- 中国音视频标准化工作组: AVS
  - AVS-1/2/3

视频编码标准

❑ 为什么中国做自主知识产权的视频编码标准

- 2002年DVD专利事件, 中国DVD、机顶盒、MP3、电视机从世界首位迅(2亿台)速滑落，我国公司需支付40亿美元/年的专利使用费,其中每个终端MPEG收取2.5美元，事件后我国已经没有自主品牌的DVD产业
- AVS每个终端1元



预测编码

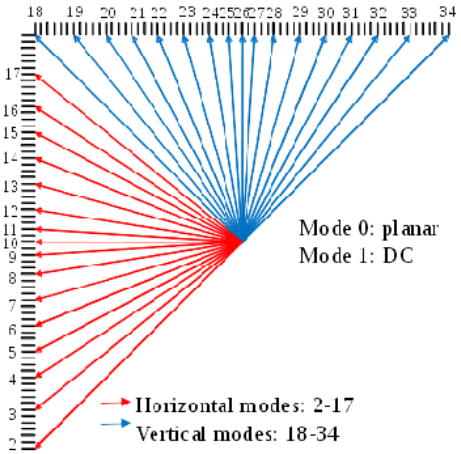
❑ 帧内预测编码

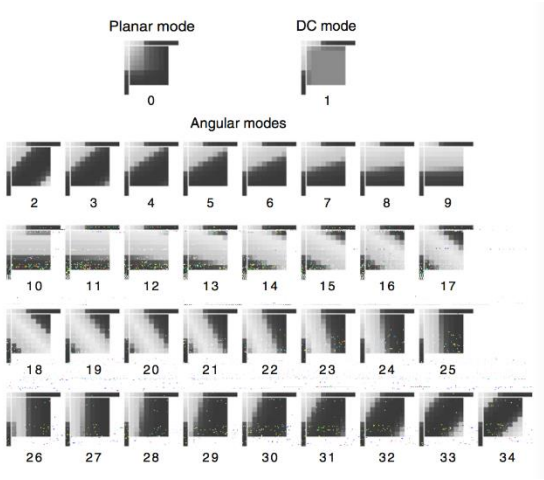
– 基于当前帧内已经编码重构的块预测当前编码块，将当前块和预测块差值进行编码

❑ 从H.264/AVC到H.265/HEVC,再到VVC

– H.264/AVC有9种预测模式

– H.265/HEVC有35种预测模式





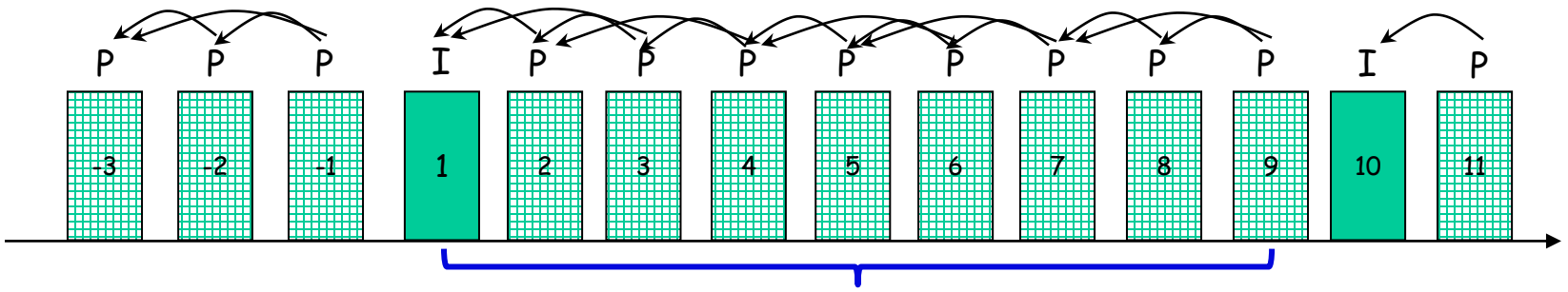
预测编码

❑ 帧间预测

- 基于一个或多个已编码帧预测，构造预测块，将当前块和预测块的残差进行编码
- 编码顺序和显示顺序

编码(传输)顺序:

1(I), 2(P), 3(P), 4(P), 5(P), 6(P), 7(P), 8(P), 9(P), 10(I)



GOP: Group of Pictures

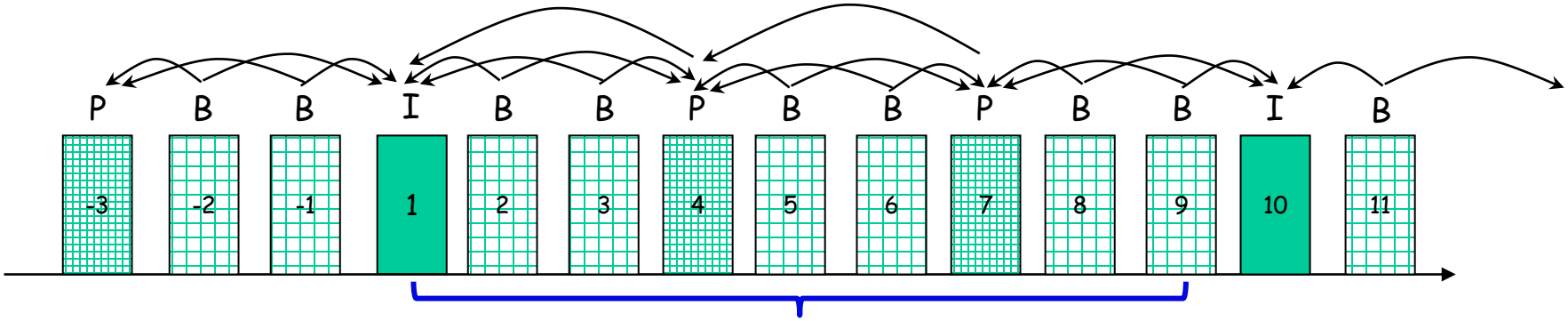
预测编码

❑ 帧间预测

- 基于一个或多个已编码帧预测，构造预测块，将当前块和预测块的残差进行编码
- 编码顺序和显示顺序

编码(传输)顺序:

1(I), 4(P), 2(B), 3(B), 7(P), 5(B), 6(B), 10(I), 8(B), 9(B)

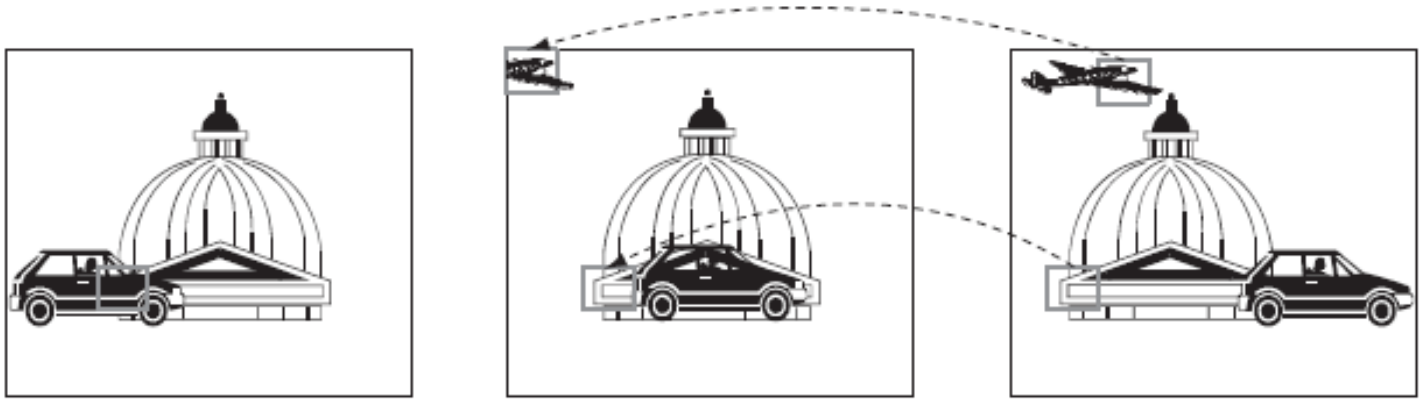


GOP: Group of Pictures

预测编码

❑ 帧间预测

- 前向预测(P frame)和双向预测（B frame）
- 运动估计

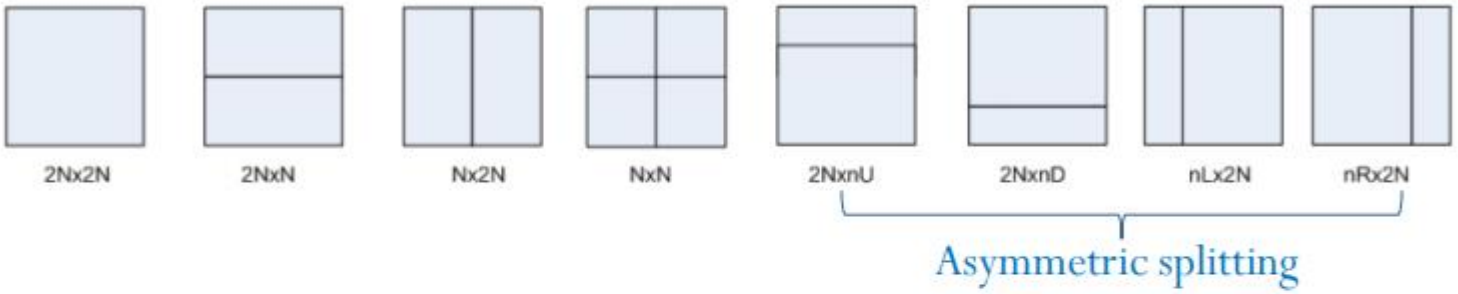


Frame *t*-1　　Frame *t*　　Frame *t*+1

预测编码

❑ 帧间预测

- 前向预测(P frame)和双向预测（B frame）
- 预测块划分
- 2Nx2N, NxN, 2NxN, Nx2N, 2NxnU, 2NxnD, nLx2N, nRx2N



预测编码

❑ 帧间预测

- 前向预测(P frame)和双向预测（B frame）
- 预测块划分
- 预测精度: 整数像素→1/2像素→1/4像素→1/8像素

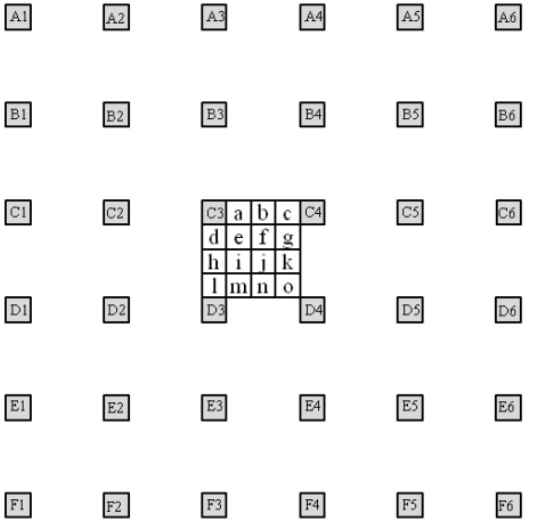
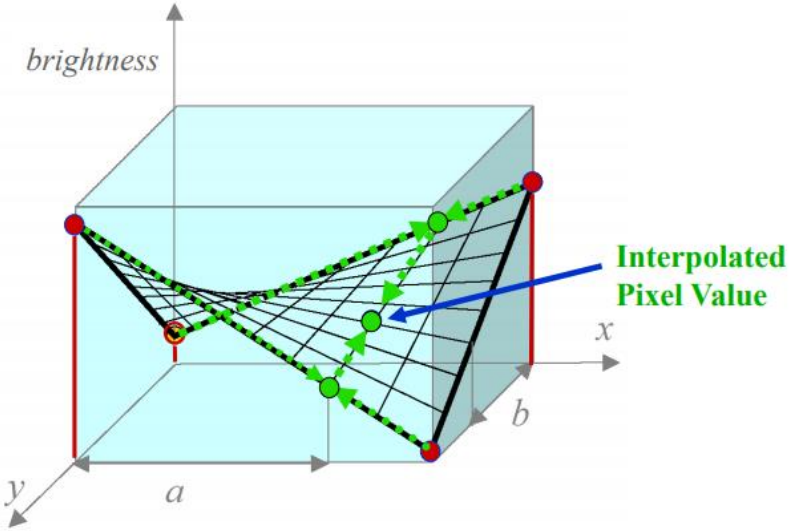


Fig. 1. Integer samples (shaded blocks with upper-case letters) and fractional sample positions (white blocks with lower-case letters).



预测编码

❑ 编码中分像素估计

- 固定滤波器系数
  - H.264/AVC: 6-tap 1/2 Pel（1,-5,20,20,-5,1）、1/4 Pel bilinear interpolation
  - HEVC: 8-tap
    - 1/4 : {-1, 4, -10, 58, 17, -5, 1, 0}
    - 1/2 : {-1, 4, -11, 40, 40, -11, 4, -1}
    - 3/4 : {0, 1, -5, 17, 58, -10, 4, -1}

变换编码

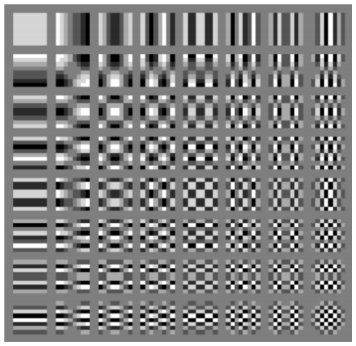
❑ 分布的特性的改变

- 能力集中特性
- 分布方差变化

198	202	194	179	180	184	196	168
187	196	192	181	182	185	189	174
188	185	193	179	188	188	187	170
184	188	182	187	183	186	195	174
194	193	189	187	180	183	181	181
193	195	193	192	170	189	187	181
181	185	183	180	175	184	185	176
195	185	177	178	170	179	195	175

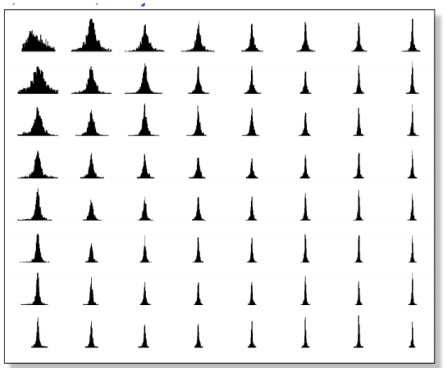
离散余弦变换 (DCT)

1480	26	9.5	8.9	-26.4	15.1	-8.1	0.3
11	8.3	-8.2	3.8	-8.4	-6.0	-2.8	10.6
-5.5	4.5	9.0	5.3	-8.0	4.0	-5.1	4.9
10.7	9.8	4.9	-8.3	-2.1	-1.9	2.8	-8.1
1.6	1.4	8.2	4.3	3.4	4.1	-7.9	1.0
-4.5	-5.0	-6.4	4.1	-4.4	1.8	-3.2	2.1
5.9	5.8	2.4	2.8	-2.0	5.9	3.2	1.1
-3.0	2.5	-1.0	0.7	4.1	-6.1	6.0	5.7



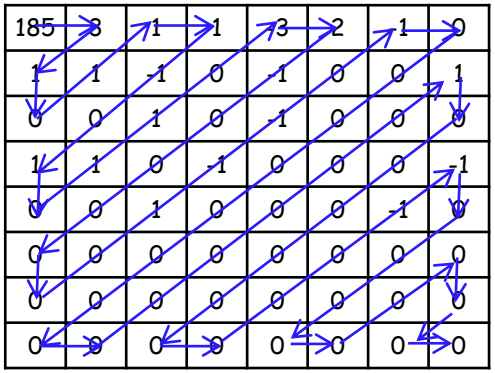
量化

185	3	1	1	-3	2	-1	0
1	1	-1	0	-1	0	0	1
0	0	1	0	-1	0	0	0
1	1	0	-1	0	0	0	-1
0	0	1	0	0	0	-1	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0



熵编码

- 根据信号统计特性，为每个输入符号分配一个唯一的码字
  - 无损压缩过程
  - 可变长编码
    - 为高概率符号分配短码字，低概率符号分配长码字， 例如JPEG中Huffman编码, H.264/AVC中的CAVLC (Context-adaptive variable-length coding)
  - 基于上下文的二进制算术编码(CABAC)



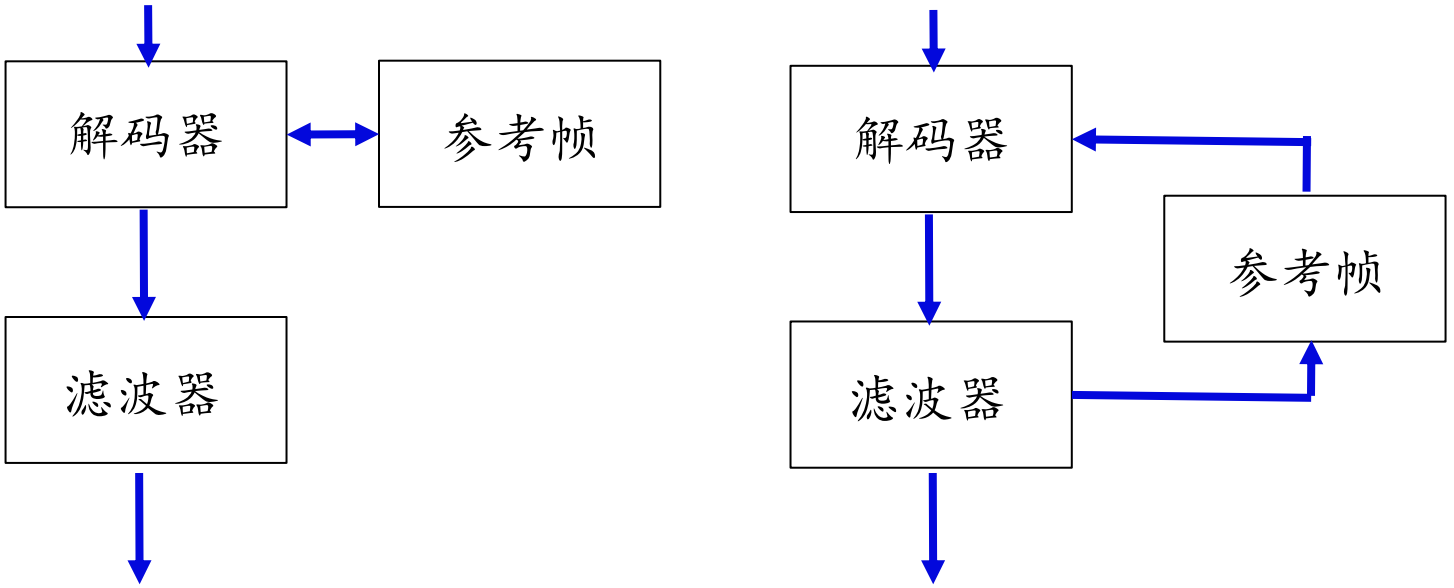
DC:185
AC: (*run*, *level*)
(0,3), (0,1), (1,1), (0,1), (0,1), (0,-1), (1,1), (1,1), (0,1), (1,-3), (0,2), (0,-1), (6,1), (0,-1), (1,-1), (14,1), (9,-1), (0,-1), EOB

CAVLC
- 编码非零系数个数TotalCoeffs和拖尾系数个数TrailingOnes
- 编码level
- 编码run

- 标准附录表9-5~~~9-10, 根据上下文选择不同的表

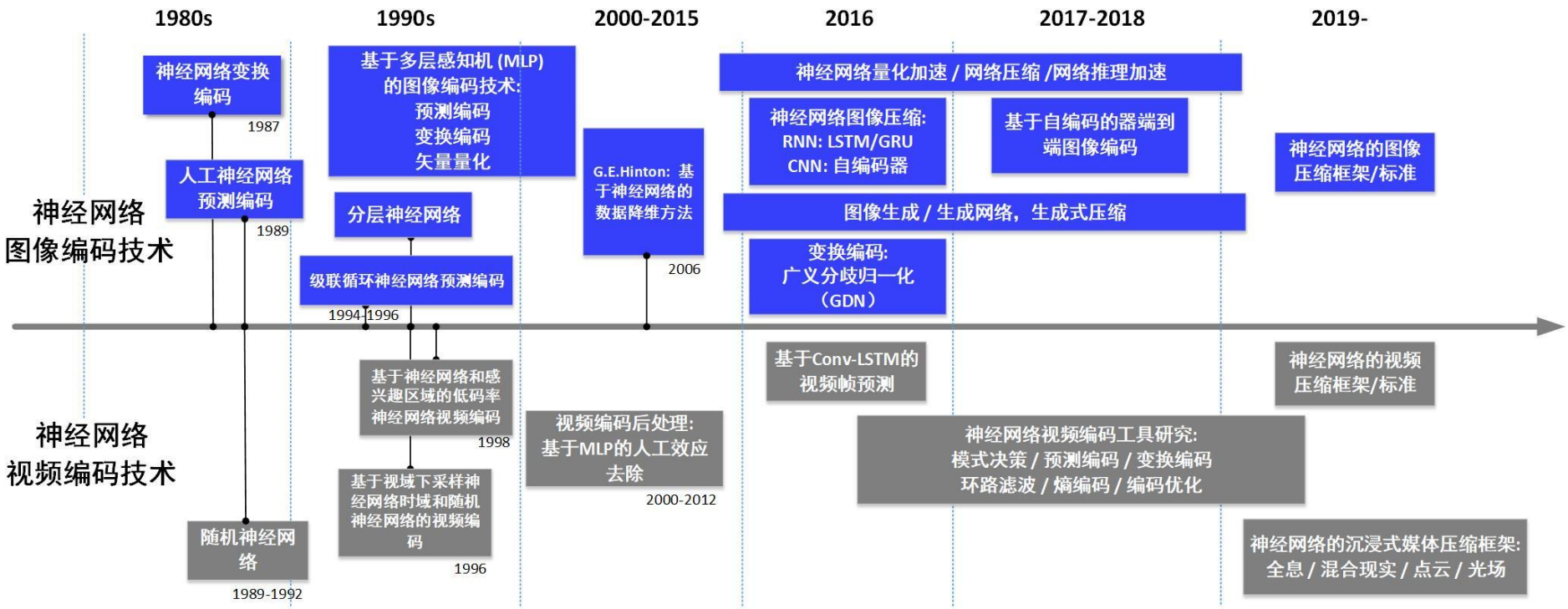
环路滤波

- **H.263+首先引入环路滤波，HEVC拓展多个环路滤波**
  - 降低压缩噪声，提高当前编码帧质量
  - 提高预测精度，从而提高编码效率



编码框架

❑ 基于神经网络的图像和视频编码发展

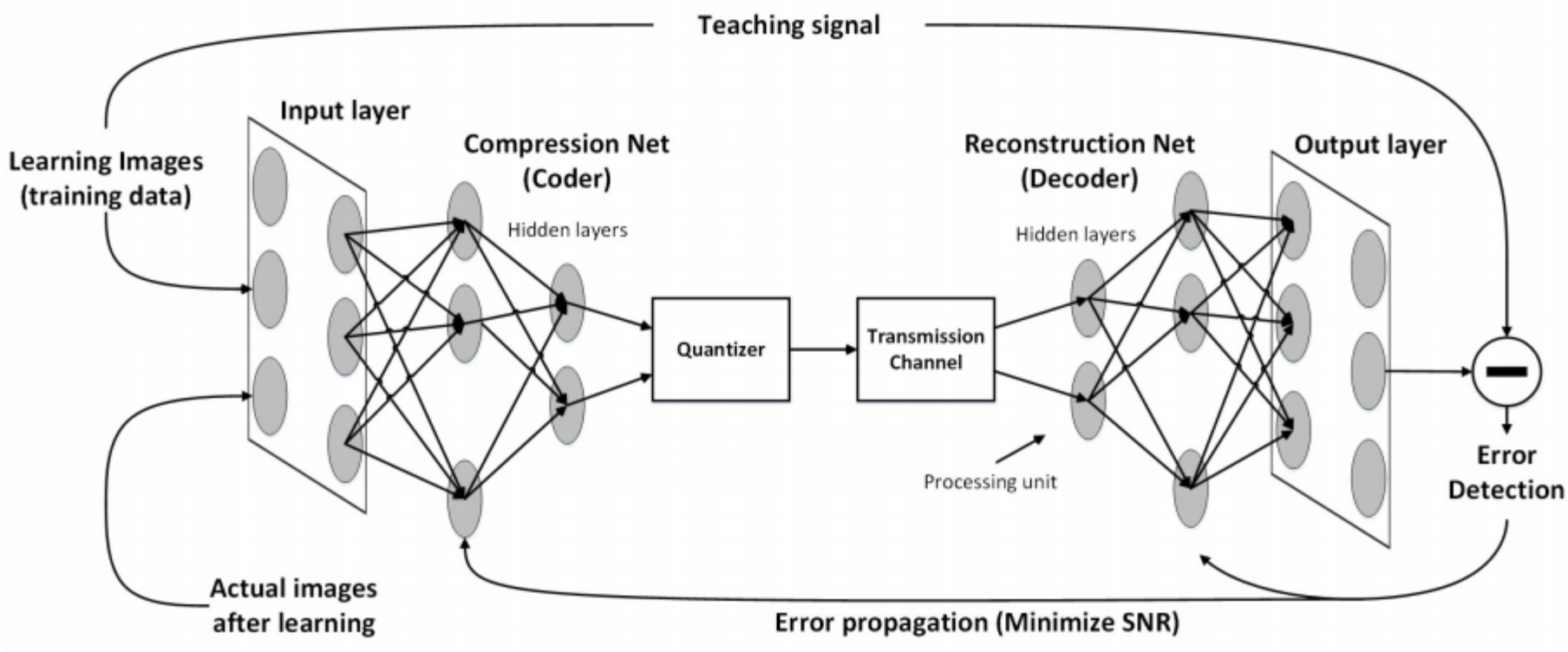


Ma, Siwei, Xinfeng Zhang, Chuanmin Jia, Zhenghui Zhao, Shiqi Wang, and Shanshe Wanga. "Image and video compression with neural networks: A review." IEEE Transactions on Circuits and Systems for Video Technology (2019).

基于神经网络的图像压缩

❑ 基于自编码器的压缩

- 分块压缩
- 不同子网络对应不同的结构特性的图像块

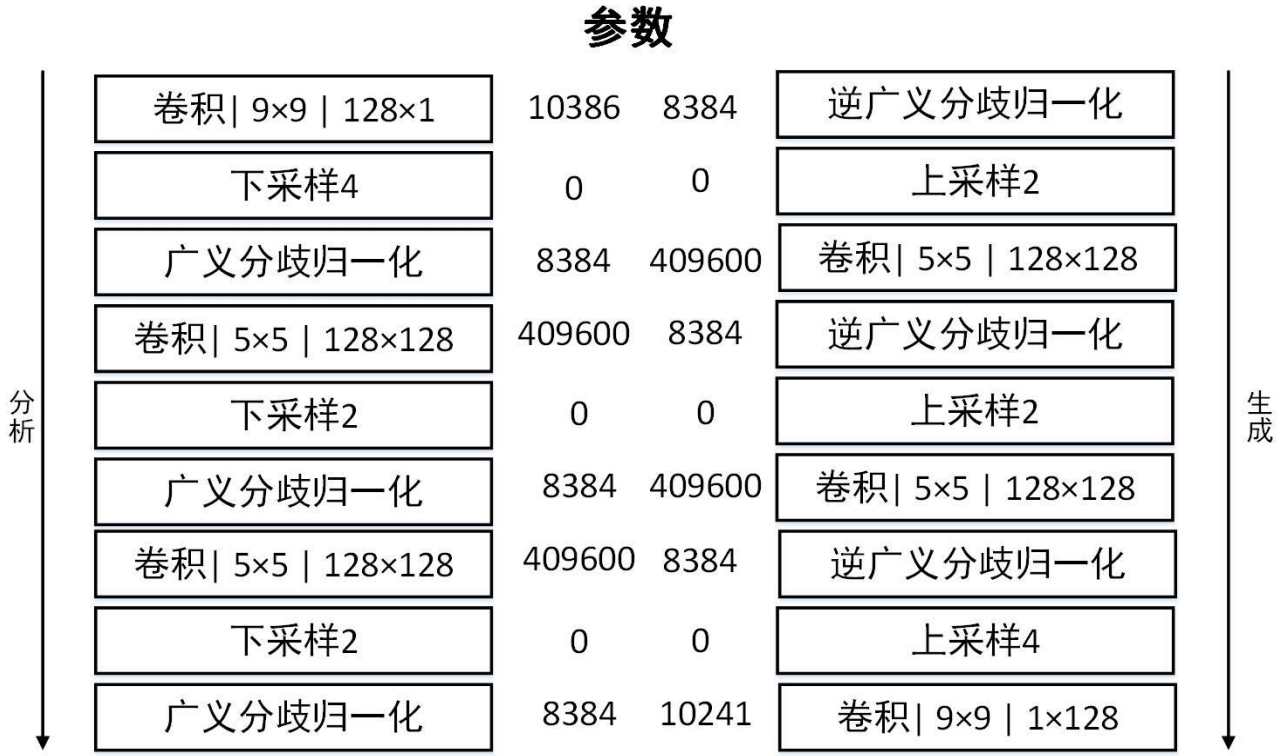


N. Sonehara, M. Kawato, S. Miyake, and K. Nakane, “Image data compression using a neural network model,” in Proc. IJCNN, vol. 2, 1989, pp. 35–41.

基于卷积神经网络的图像压缩

❑ 卷积网络压缩框架

– 包含分析和生成两个模块，对应了编码和解码功能

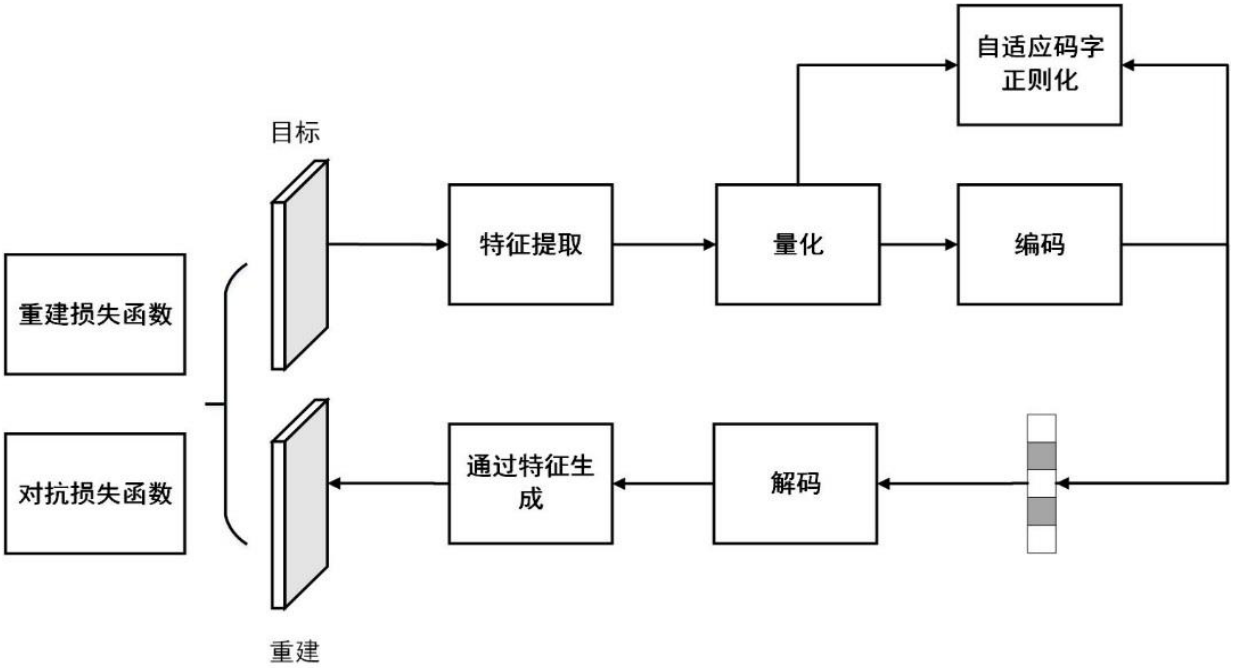


J. Balle, V. Laparra, and E. P. Simoncelli, “End-to-end optimized image compression,” arXiv preprint arXiv:1611.01704, 2016.

基于对抗生成网络的图像压缩

❑ 压缩框架

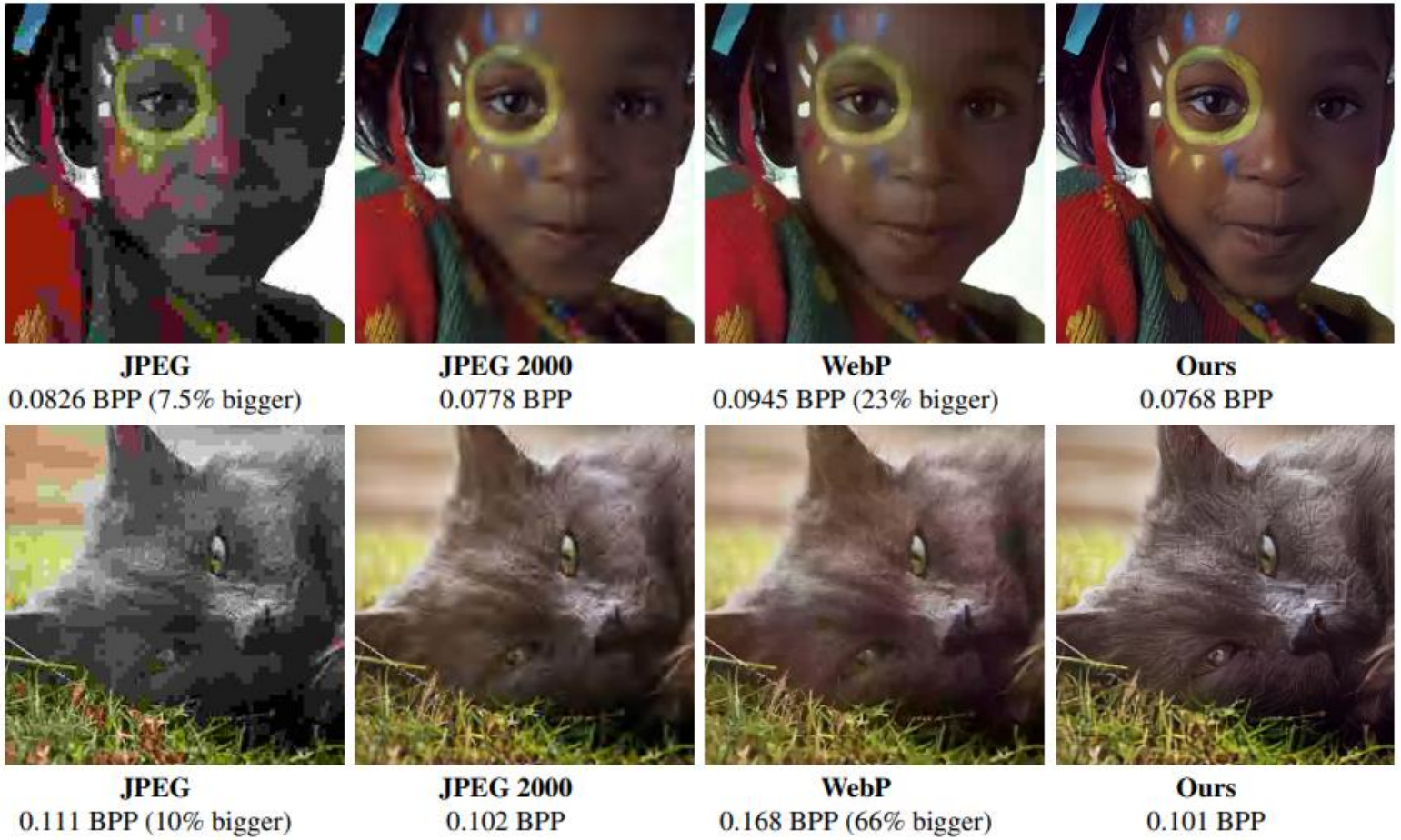
- WaveOne公司提出 http://www.wave.one/
- 实际是特征的压缩，通过特征生成图像
- 在率失真目标函数中引入对抗损失函数进行端到端训练



Rippel, Oren, and Lubomir Bourdev. "Real-time adaptive image compression.“ In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 2922-2930. JMLR. org, 2017.

基于对抗生成网络的图像压缩

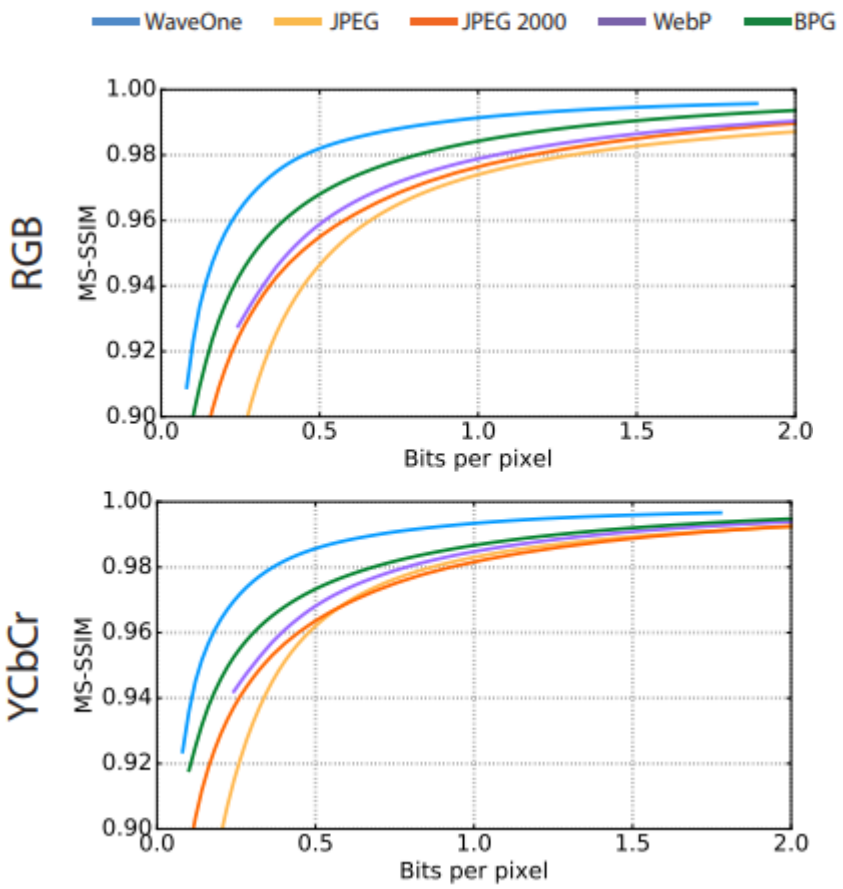
❑ 压缩性能



Rippel, Oren, and Lubomir Bourdev. "Real-time adaptive image compression.“ In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 2922-2930. JMLR. org, 2017.

基于对抗生成网络的图像压缩

❑ 压缩性能

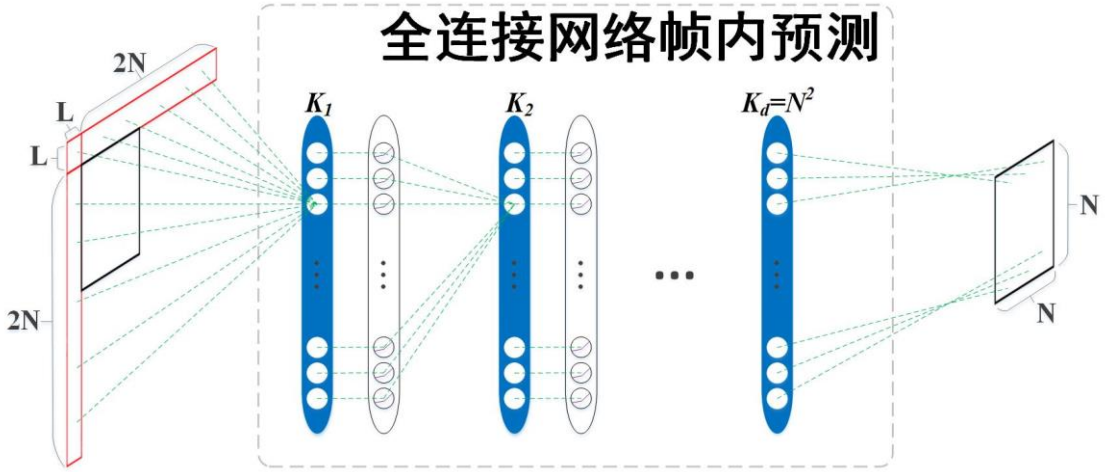


Rippel, Oren, and Lubomir Bourdev. "Real-time adaptive image compression.“ In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 2922-2930. JMLR. org, 2017.

混合编码框架下的深度学习视频编码

❑ 全连接网络帧内预测模式

- 与HEVC已有的35种模式进行率失真优化决策
- 将左侧、上方和左上的重构像素作为输入，通过网络得到当前块预测



Li, Jiahao, Bin Li, Jizheng Xu, Ruiqin Xiong, and Wen Gao. "Fully connected network-based intra prediction for image coding." IEEE Transactions on Image Processing 27, no. 7 (2018): 3236-3247.

混合编码框架下的深度学习视频编码

❑ 全连接网络帧内预测模式

- 与HEVC已有的35种模式进行率失真优化决策
- 将左侧、上方和左上的重构像素作为输入，通过网络得到当前块预测

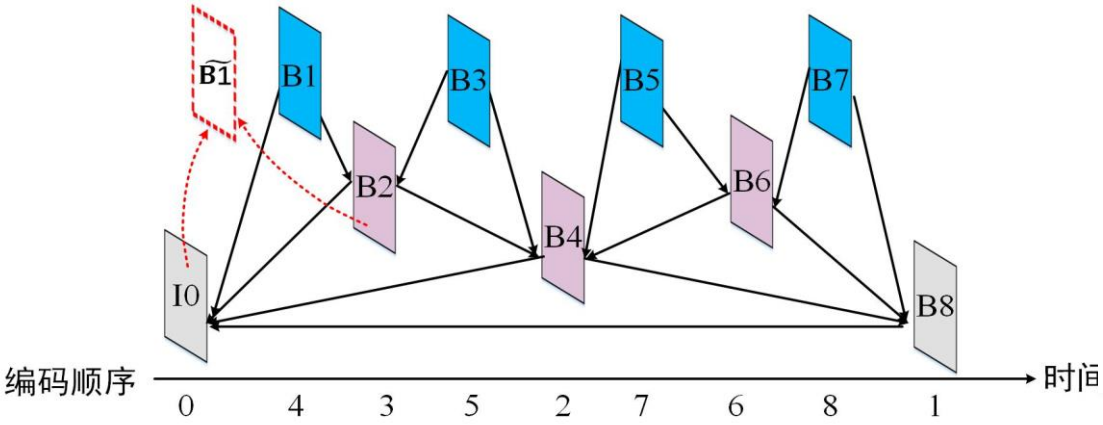
Sequences	IPFCN vs. HM-16.9			
	IPFCN-S	IPFCN-D	IPFCN-S-L	IPFCN-D-L
Class A	-3.8 %	-4.4 %	-3.0%	-3.7%
Class B	-2.8 %	-3.2 %	-2.2%	-2.8%
Class C	-1.9 %	-2.1 %	-1.6%	-1.9%
Class D	-1.7 %	-1.8 %	-1.4%	-1.7%
Class E	-3.9 %	-4.5 %	-3.0%	-3.5%
Overall	-2.6 %	-3.0 %	-2.1%	-2.5%
Encode Time	4930%	13052%	285%	483%
Decode Time	26572%	28927%	923%	1141%

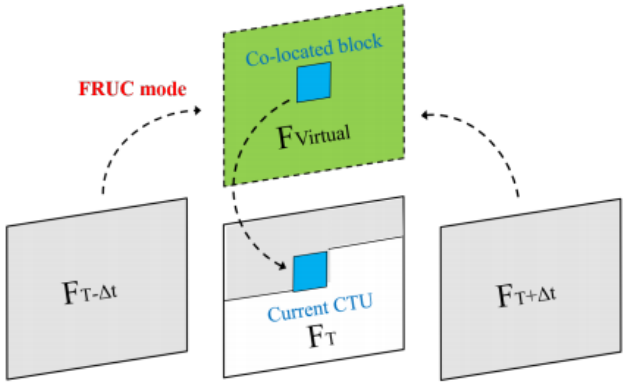
Li, Jiahao, Bin Li, Jizheng Xu, Ruiqin Xiong, and Wen Gao. "Fully connected network-based intra prediction for image coding." IEEE Transactions on Image Processing 27, no. 7 (2018): 3236-3247.

混合编码框架下的深度学习视频编码

❑ 基于帧率提升网络的帧间预测编码

- 通过网络生成虚拟参考帧
- 通过视频增强网络提高生成质量
- 虚拟参考帧参与帧间运动估计的竞争，通过率失真代价进行选择





Zhao, Lei, Shiqi Wang, Xinfeng Zhang, Shanshe Wang, Siwei Ma, and Wen Gao. "Enhanced motion-compensated video coding with deep virtual reference frame generation." IEEE Transactions on Image Processing 28, no. 10 (2019): 4832-4844.

混合编码框架下的深度学习视频编码

❑ 基于帧率提升网络的帧间预测编码性能

– 解码时间增加了70多倍

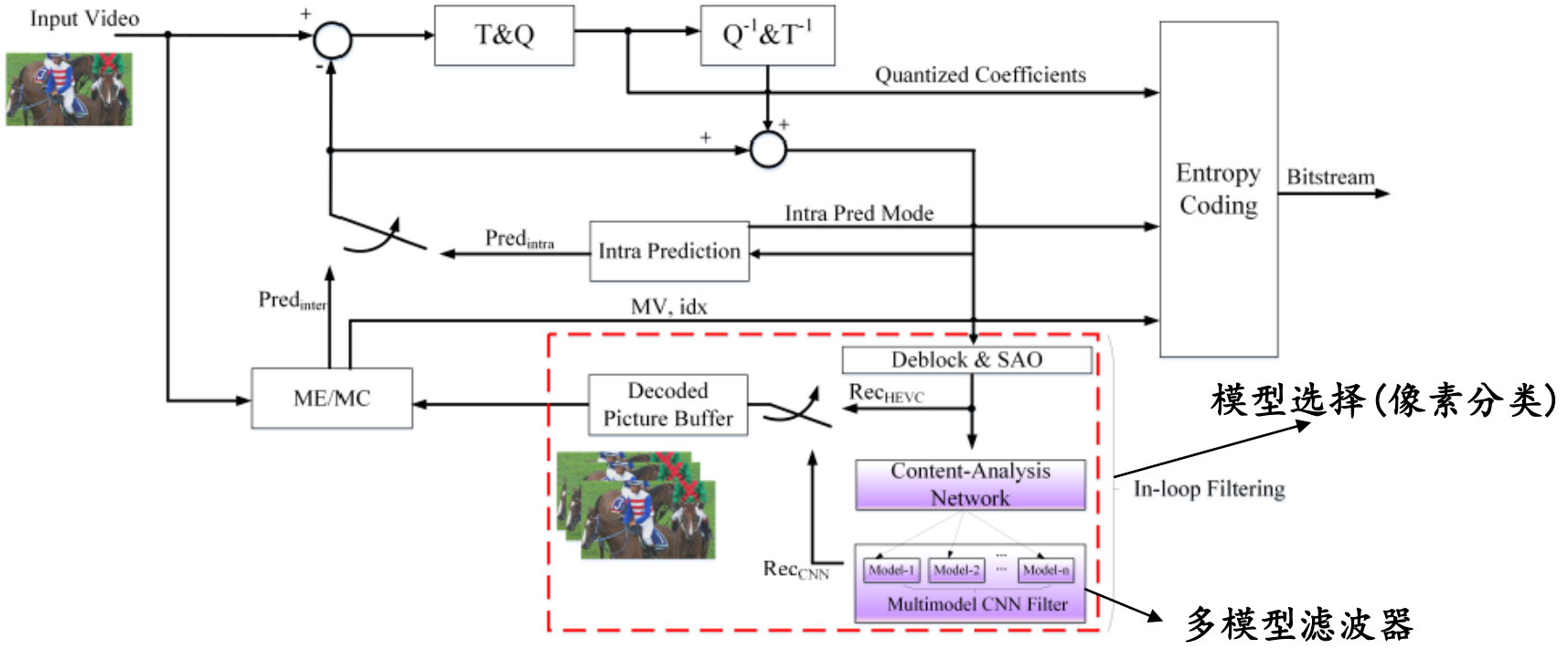
Sequences		BD-rate performance with normal QPs setting			BD-rate performance with larger QPs setting		
		Y	U	V	Y	U	V
Class A	Traffic	-5.5%	-2.6%	-1.6%	-7.2%	-3.7%	-3.1%
	PeopleOnStreet	-7.9%	-1.4%	-3.8%	-10.1%	-4.4%	-6.9%
Class B	Kimono	-2.5%	3.9%	5.0%	-2.8%	4.2%	5.6%
	ParkScene	-3.3%	2.3%	2.8%	-4.8%	2.8%	3.7%
	Cactus	-5.2%	-2.8%	-2.9%	-6.5%	-3.7%	-4.6%
	BasketballDrive	-2.7%	-0.1%	-1.8%	-3.8%	-1.3%	-3.6%
	BQTerrace	-3.2%	-0.6%	1.2%	-3.4%	-1.5%	-1.3%
Class C	BasketballDrill	-3.0%	-2.7%	-2.7%	-4.4%	-4.0%	-4.5%
	BQMall	-5.8%	-2.7%	-3.8%	-7.8%	-4.1%	-6.3%
	PartyScene	-5.2%	-3.3%	-3.2%	-5.6%	-2.9%	-3.4%
	RaceHorsesC	-3.5%	-2.4%	-3.3%	-3.3%	-2.5%	-3.5%
Class D	BasketballPass	-5.4%	-1.1%	-5.0%	-7.4%	-1.4%	-6.8%
	BQSquare	-8.7%	0.4%	-1.7%	-8.2%	1.2%	-1.4%
	BlowingBubbles	-5.1%	-3.0%	-2.9%	-5.4%	-2.9%	-2.9%
	RaceHorses	-3.8%	-2.2%	-3.1%	-5.6%	-3.2%	-4.7%
Class E	FourPeople	-5.9%	-3.7%	-4.1%	-6.0%	-4.5%	-5.3%
	Johnny	-2.8%	-1.3%	-1.8%	-3.1%	-2.1%	-2.7%
	KristenAndSara	-3.9%	-2.1%	-1.6%	-4.6%	-3.5%	-3.2%
Average		**-4.5%**	**-1.3%**	**-1.8%**	**-5.6%**	**-2.1%**	**-3.0%**

Zhao, Lei, Shiqi Wang, Xinfeng Zhang, Shanshe Wang, Siwei Ma, and Wen Gao. "Enhanced motion-compensated video coding with deep virtual reference frame generation." IEEE Transactions on Image Processing 28, no. 10 (2019): 4832-4844.

混合编码框架下的深度学习视频编码

❑ 多模型CNN环路滤波

– CNN像素分类器和多个CNN环路滤波器

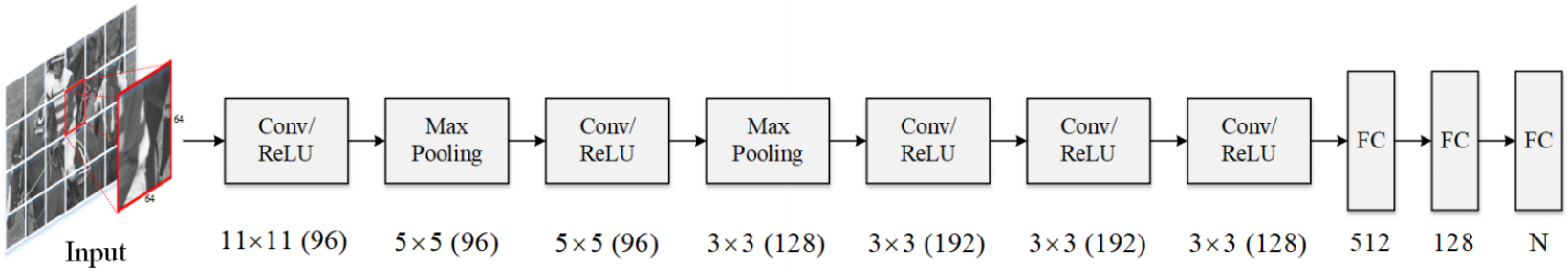


[Jia 2019a] C. Jia, S. Wang, X. Zhang, S. Wang, J. Liu, S. Pu, and S. Ma. "Content-Aware Convolutional Neural Network for In-loop Filtering in High Efficiency Video Coding." *IEEE Transactions on Image Processing* (2019).

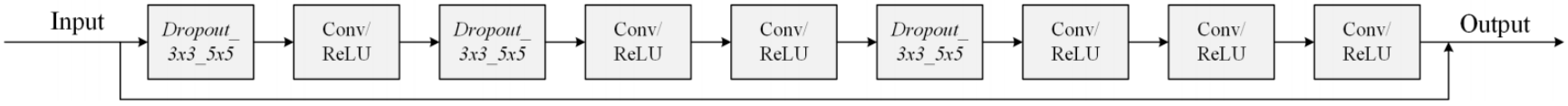
混合编码框架下的深度学习视频编码

❑ 多模型CNN环路滤波

– CNN像素分类器和多个CNN环路滤波器



模型选择分类器网络结构(轻量级Alexnet[Krizhevsky2012])



CNN滤波器网络结构

C. Jia, S. Wang, X. Zhang, S. Wang, J. Liu, S. Pu, and S. Ma. "Content-Aware Convolutional Neural Network for In-loop Filtering in High Efficiency Video Coding." *IEEE Transactions on Image Processing* (2019).

混合编码框架下的深度学习视频编码

❑ 多模型CNN环路滤波性能

- 参考软件 HM16.9，QP = 22, 27, 32, 37
- 随着模型数量增多性能提升，N=6/8时性能基本稳定

Sequences	AI			LDB			LDP			RA		
	Y	U	V	Y	U	V	Y	U	V	Y	U	V
N=1	-3.0%	-2.8%	-3.4%	-3.9%	-2.1%	-2.3%	-3.7%	-1.0%	-1.6%	-3.9%	-2.1%	-2.3%
N=2	-3.3%	-3.2%	-4.2%	-4.9%	-2.8%	-3.3%	-4.3%	-1.1%	-1.6%	-4.9%	-3.1%	-3.7%
N=4	-3.7%	-3.1%	-3.8%	-5.4%	-2.5%	-3.0%	-4.6%	-0.9%	-1.1%	-5.4%	-2.9%	-3.4%
N=6	-3.9%	-3.5%	-4.2%	-5.9%	-2.4%	-3.0%	-4.7%	-0.7%	-1.3%	-5.8%	-2.6%	-3.3%

混合编码框架下的深度学习视频编码

❑ 多模型CNN环路滤波性能

- 参考软件 HM16.9，QP = 22, 27, 32, 37
- 随着模型数量增多性能提升，N=6/8时性能基本稳定

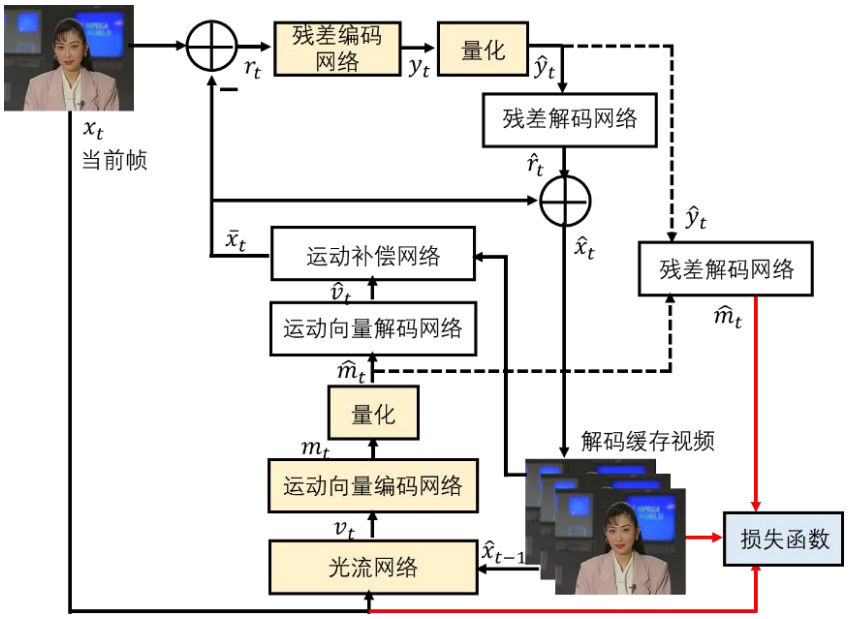
Sequences	AI			LDB			LDP			RA		
	Y	U	V	Y	U	V	Y	U	V	Y	U	V
N=1	-3.0%	-2.8%	-3.4%	-3.9%	-2.1%	-2.3%	-3.7%	-1.0%	-1.6%	-3.9%	-2.1%	-2.3%
N=2	-3.3%	-3.2%	-4.2%	-4.9%	-2.8%	-3.3%	-4.3%	-1.1%	-1.6%	-4.9%	-3.1%	-3.7%
N=4	-3.7%	-3.1%	-3.8%	-5.4%	-2.5%	-3.0%	-4.6%	-0.9%	-1.1%	-5.4%	-2.9%	-3.4%
N=6	-3.9%	-3.5%	-4.2%	-5.9%	-2.4%	-3.0%	-4.7%	-0.7%	-1.3%	-5.8%	-2.6%	-3.3%

深度学习端到端视频编码框架

❑ 多网络模型的端到端视频编码框架

– 采用卷积神经网络估计运动光流场、对依照光流场产生的预测图像进行滤波得到高质量的运动补偿视频帧以及采用非线性神经网络进行残差变换

– 该方案在通测序列上部分码率范围内取得了超过very fast配置下x265的性能，但还是明显低于HEVC的最优性能配置



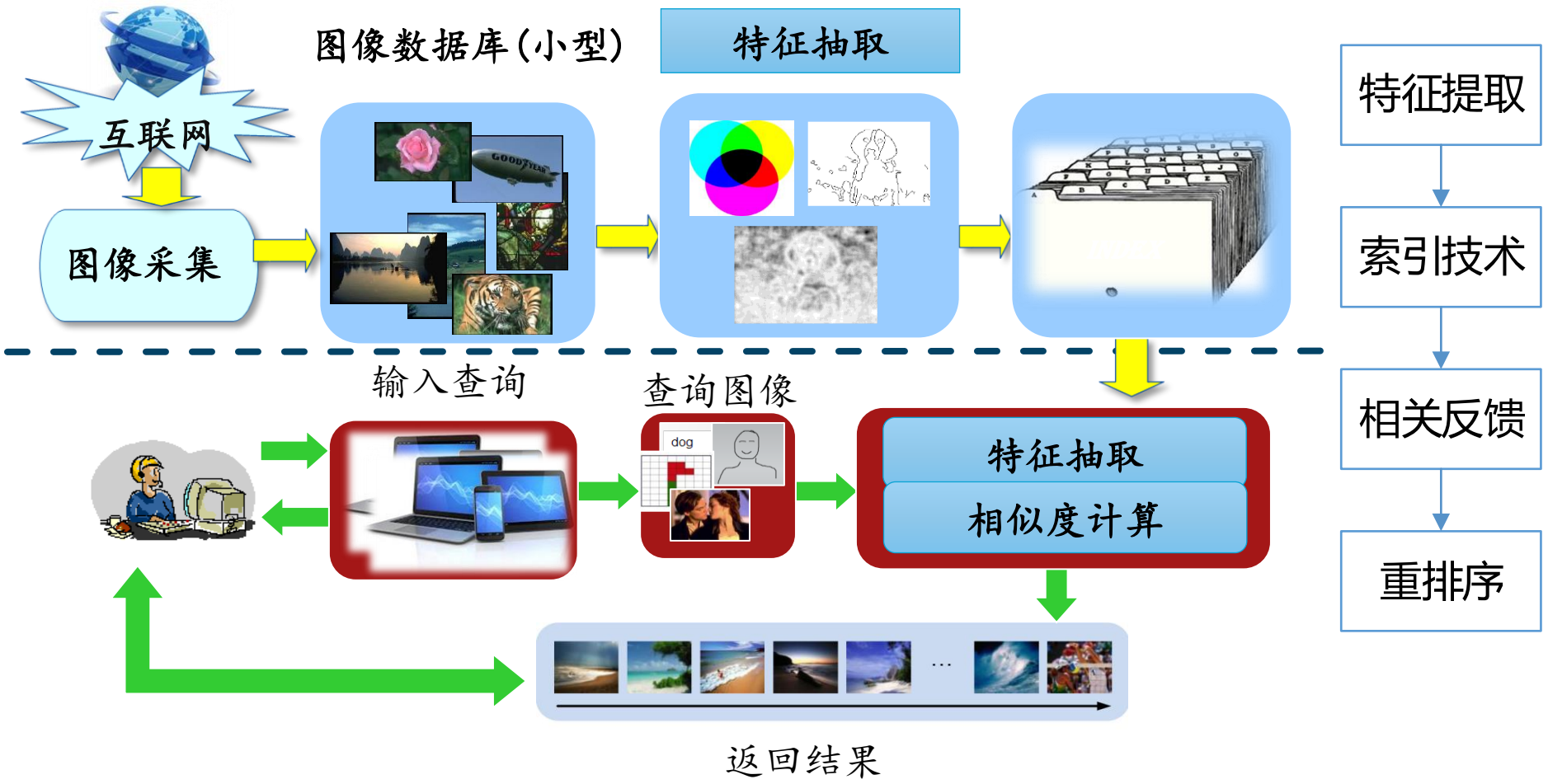
Lu G., Ouyang W., Xu D., et al., “DVC: An End-to-end Deep Video Compression Framework, ” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, 2019

中国科学院大学
University of Chinese Academy of Sciences

3 传统的计算机视觉处理

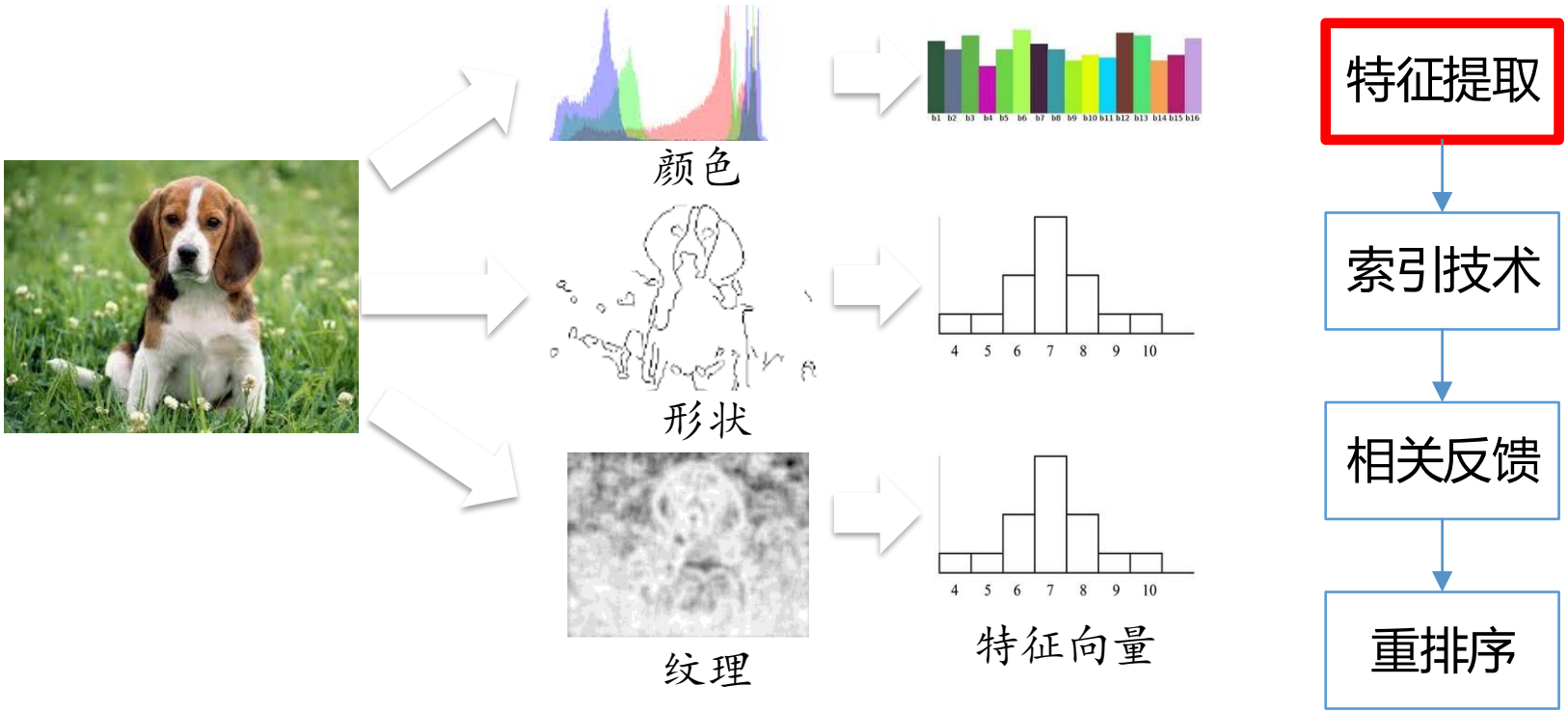
早期的计算机视觉处理（1990-2003）

❑ 处理流程



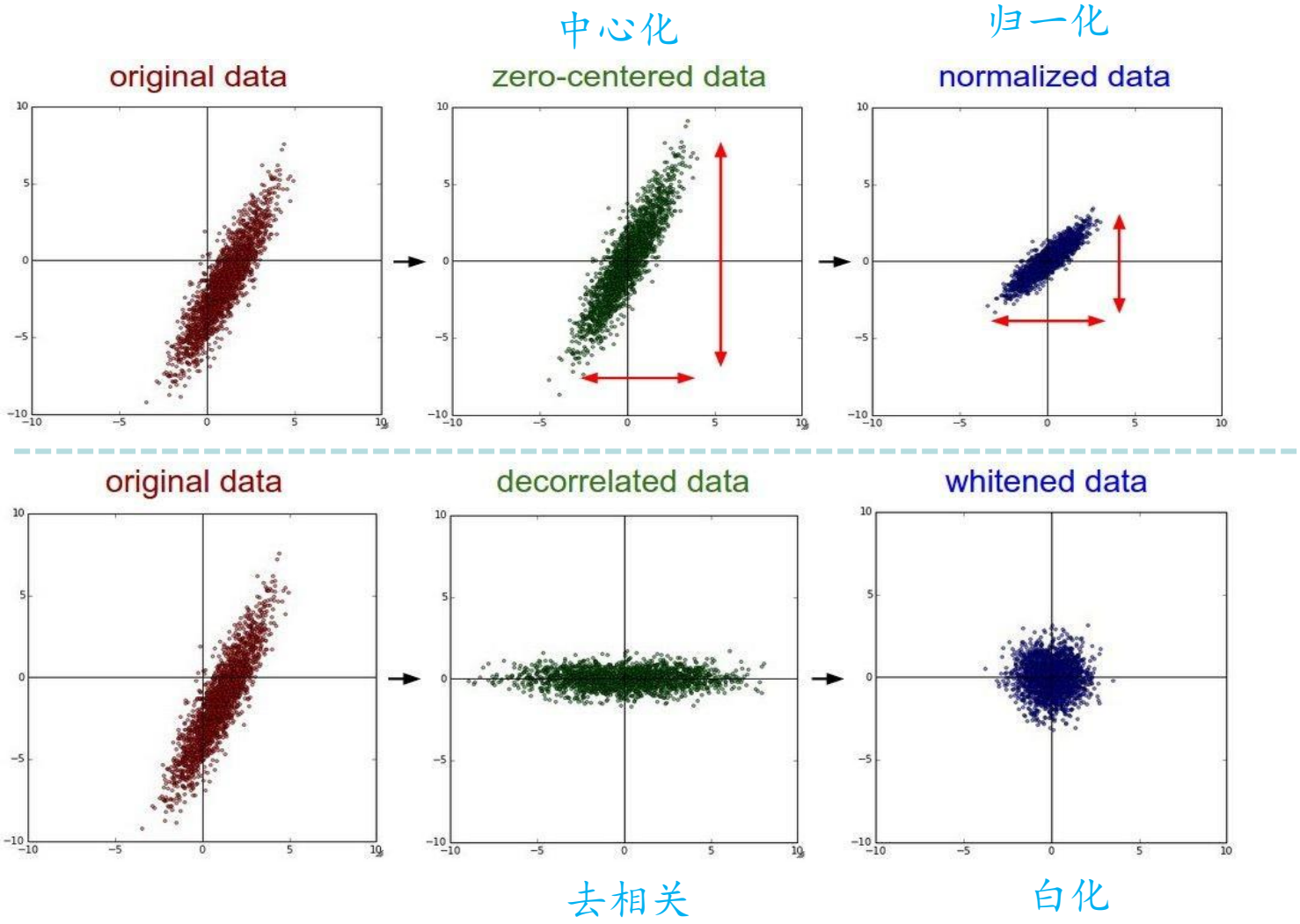
早期的计算机视觉处理（1990-2003）

- 全局特征提取：用全局的视觉底层特性统计量表示图像



早期的计算机视觉处理（1990-2003）

❑ 简单特征变换



特征提取

索引技术

相关反馈

重排序

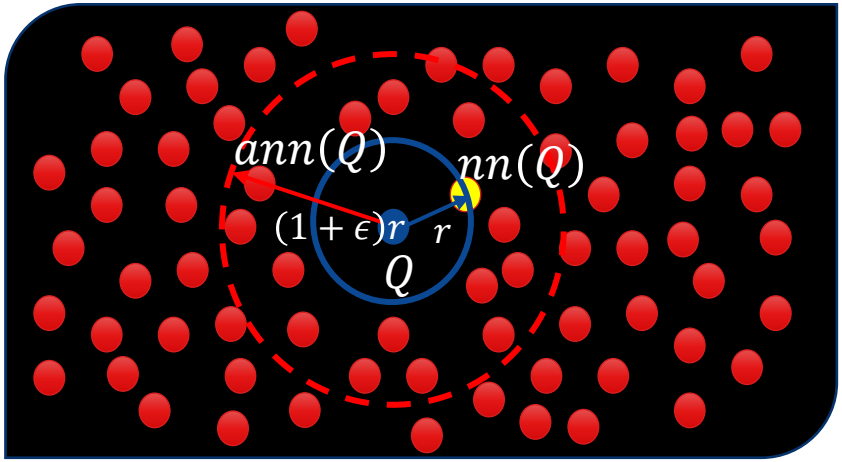
早期的计算机视觉处理（1990-2003）

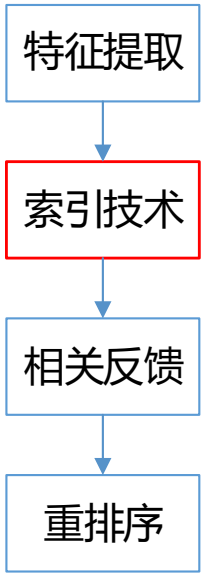
❑ 索引技术

穷举搜索：效率太低，时间复杂度太高

改进方式：牺牲精度，寻找近似的最近邻居

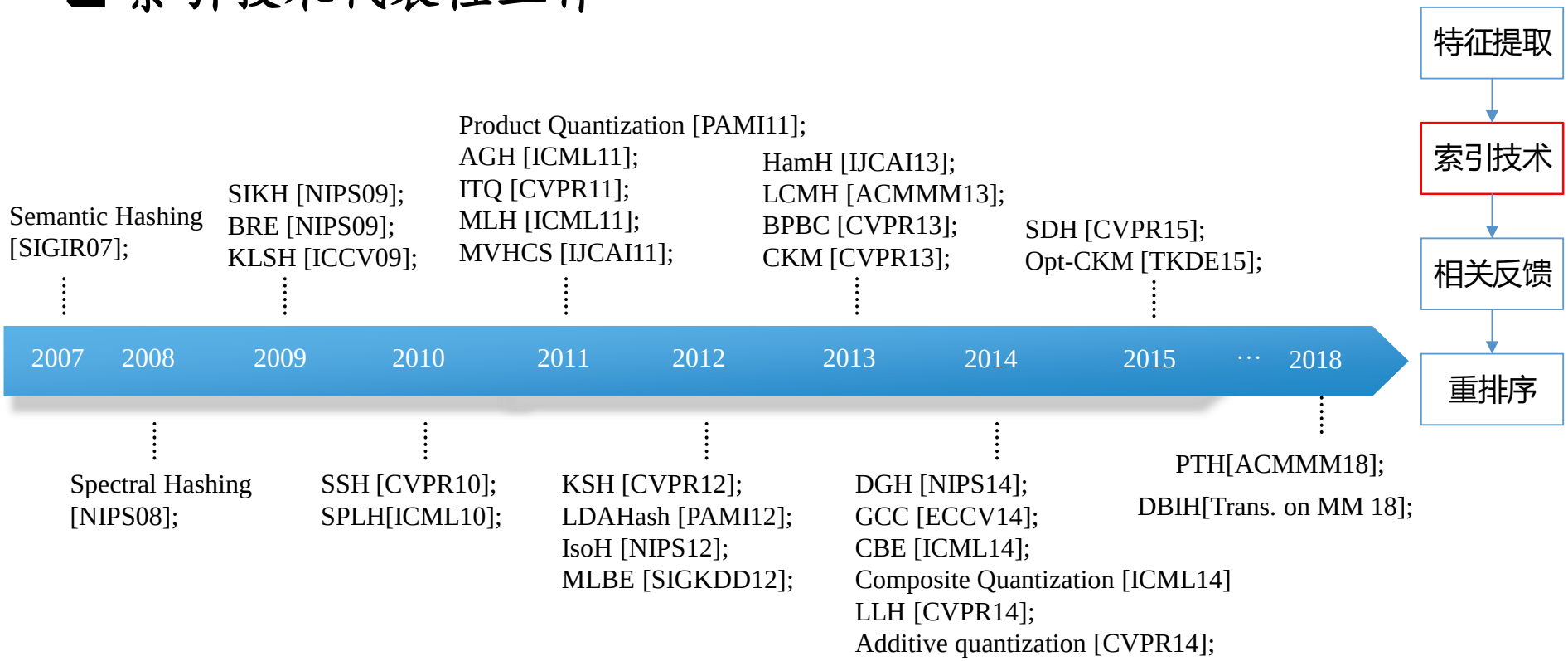
常用方法：KD-Tree, LSH (Locality Sensitive Hashing)





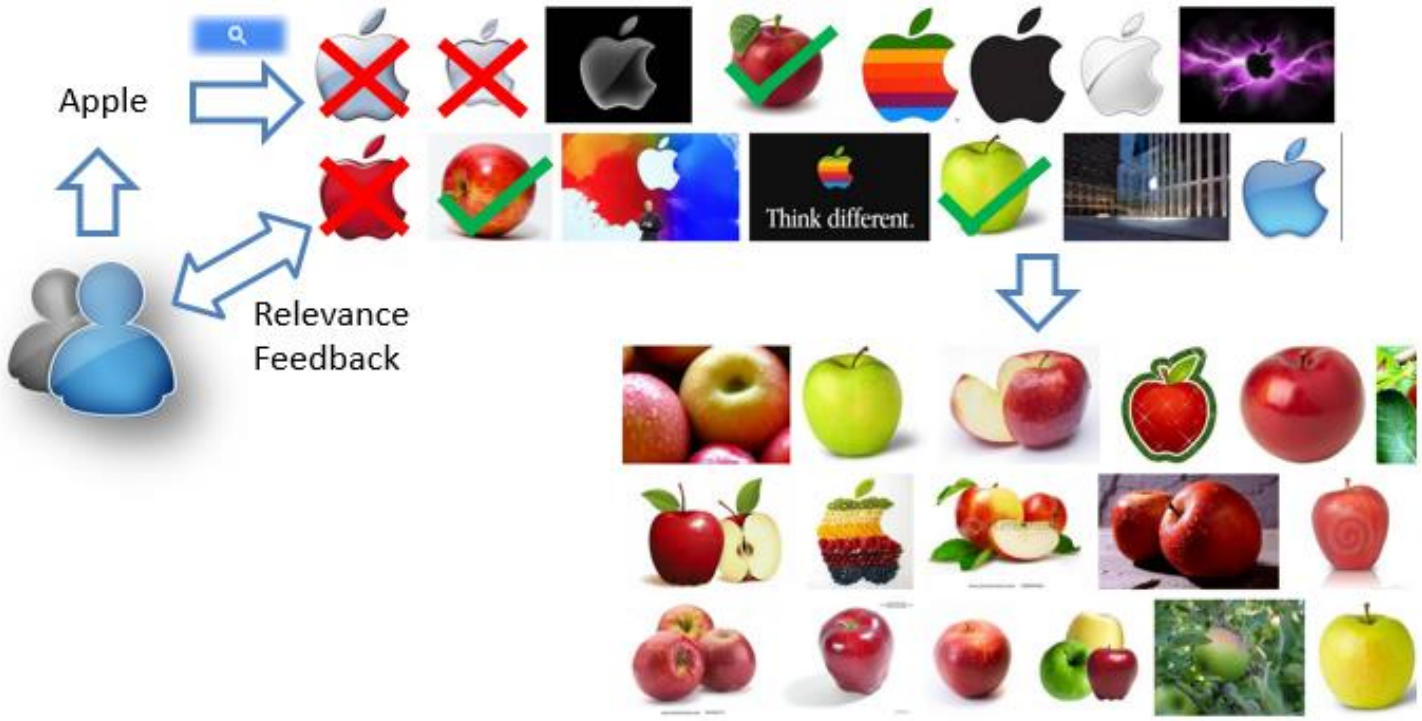
早期的计算机视觉处理（1990-2003）

❑ 索引技术代表性工作



早期的计算机视觉处理（1990-2003）

❑ 相关反馈



特征提取

索引技术

相关反馈

重排序

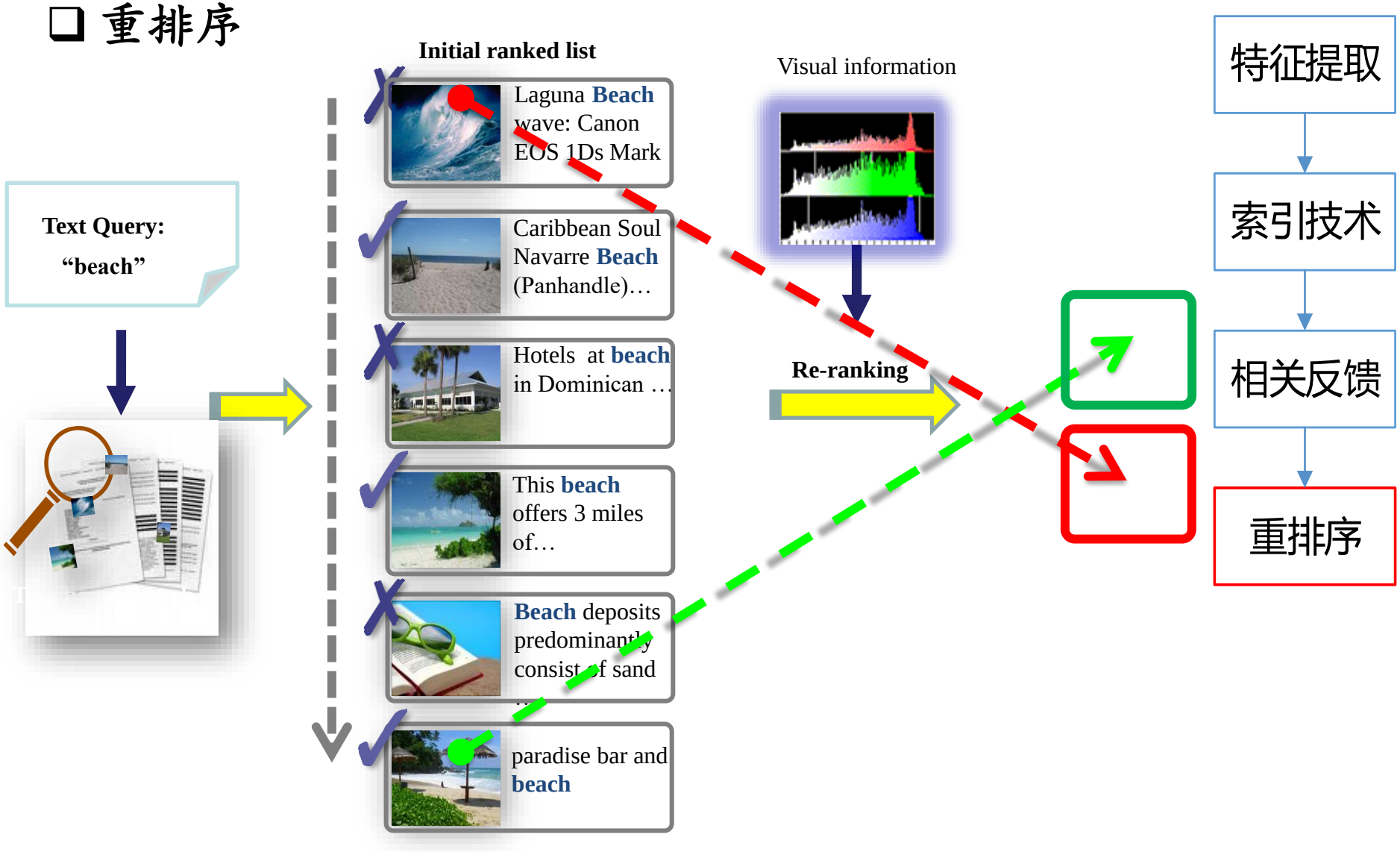
- 反馈类型

Explicit feedback： 反馈正例或者负例

Implicit feedback: 根据可观察的行为推断用户意图

早期的计算机视觉处理（1990-2003）

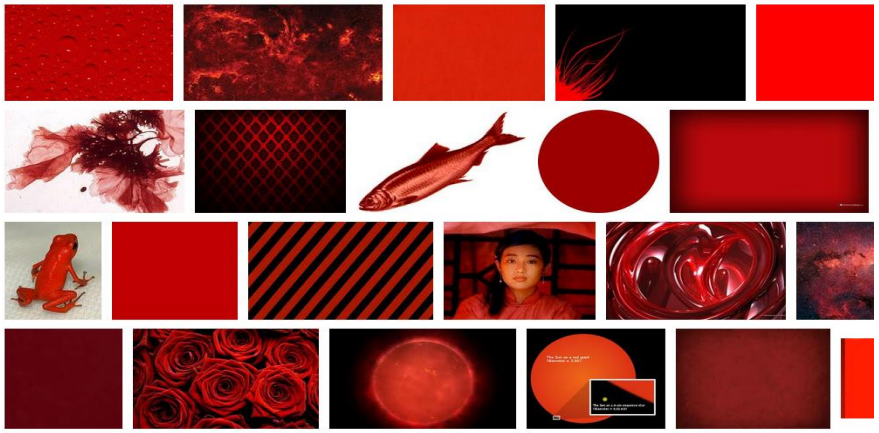
❑ 重排序



早期的计算机视觉处理（1990-2003）

- 早期图像检索技术的问题：全局特征丢掉了图像细节

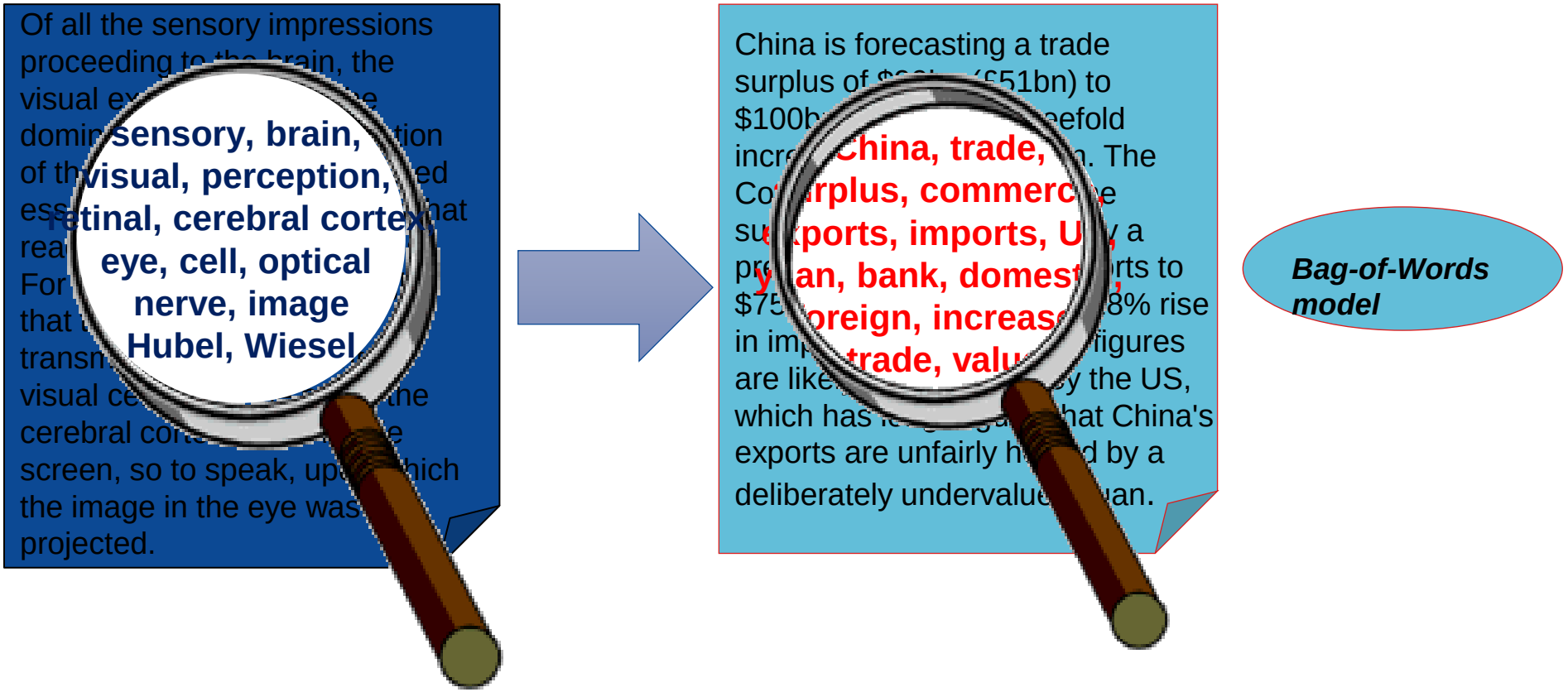
正确匹配



错误匹配

中期的计算机视觉处理（2004-2012）

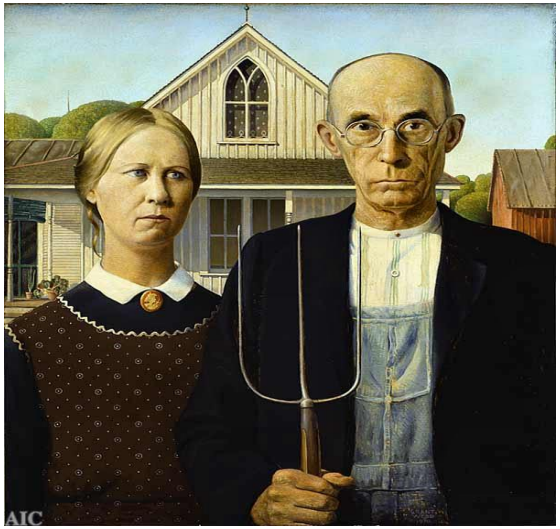
- 文本搜索的经典模型：词袋模型（ Bag-of-Words ）



中期的计算机视觉处理（2004-2012）

- 图像能被表示为视觉词袋(Bag-of-Visual Words)吗？

图像

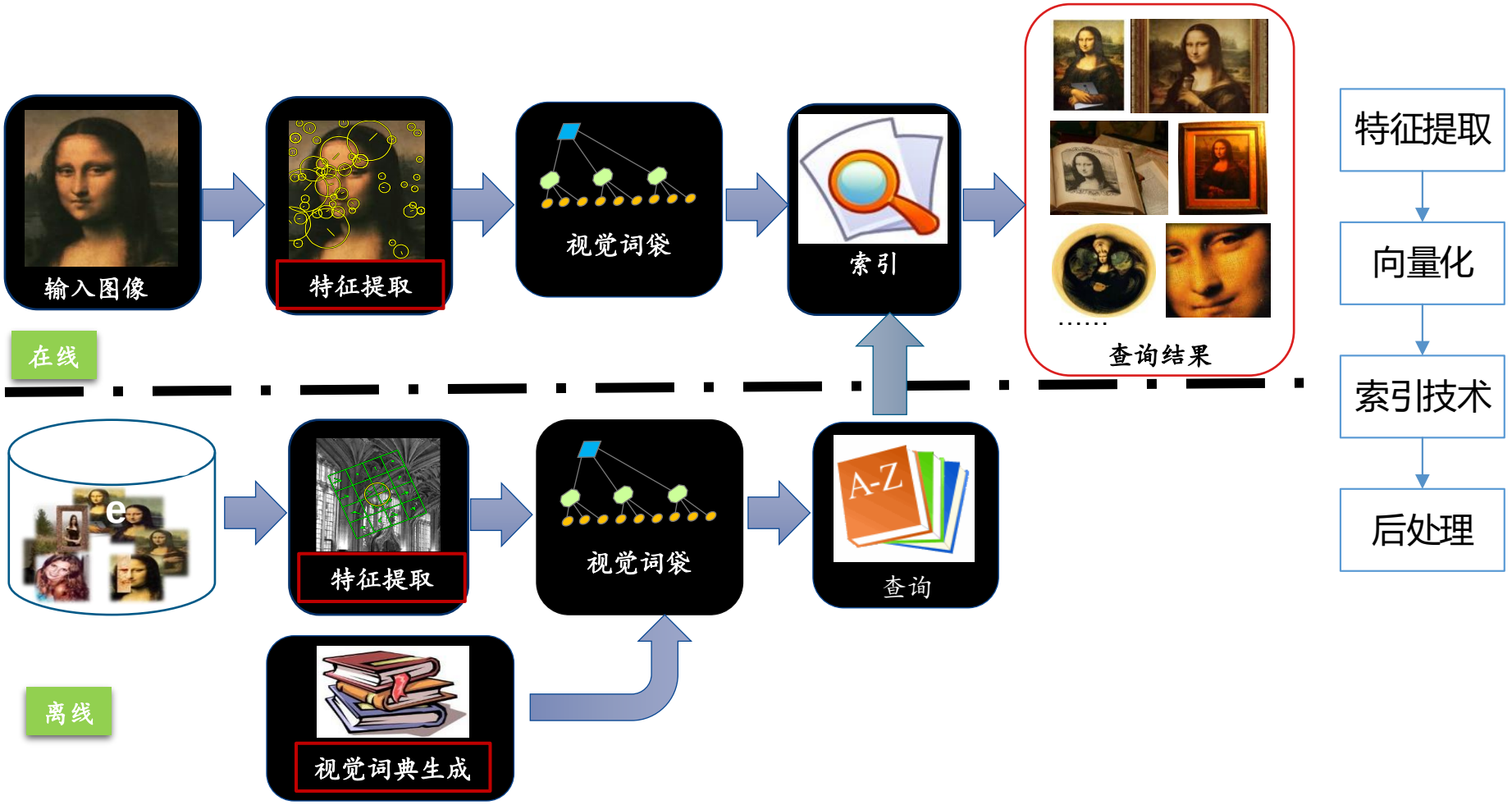


?

视觉词袋

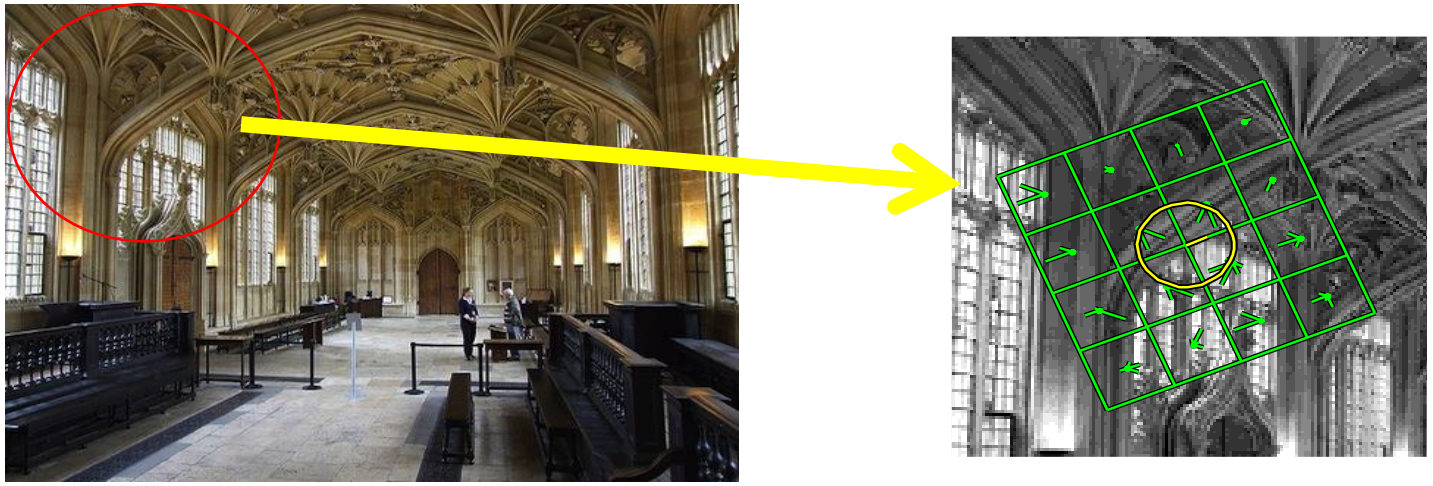
中期的计算机视觉处理（2004-2012）

❑ 中期图像识别框架



中期的计算机视觉处理（2004-2012）

❑ 局部特征：图像区块的向量



特征提取

向量化

索引技术

后处理

问题 1: 图像的哪个区块，即区块坐标是?

答：特征检测器（Feature Detector）

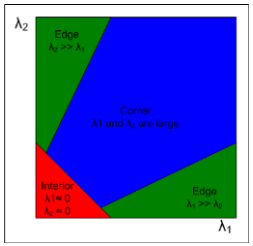
问题2: 怎么把图像区块表示为向量?

答：特征描述器（Feature Descriptor）

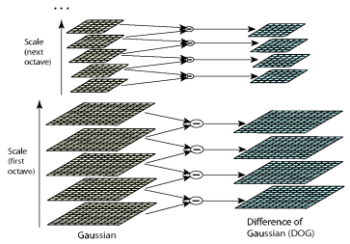
中期的计算机视觉处理（2004-2012）

❑ 局部检测器：检测图像区块的兴趣点（具有区分意义的点）

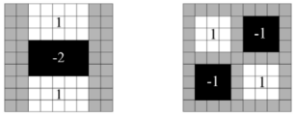
- Harris, DoG, SURF, Harris-Affine, Hessian-Affine, and MSER…



Harris

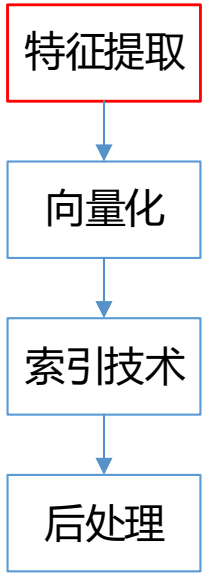


DoG



SURF

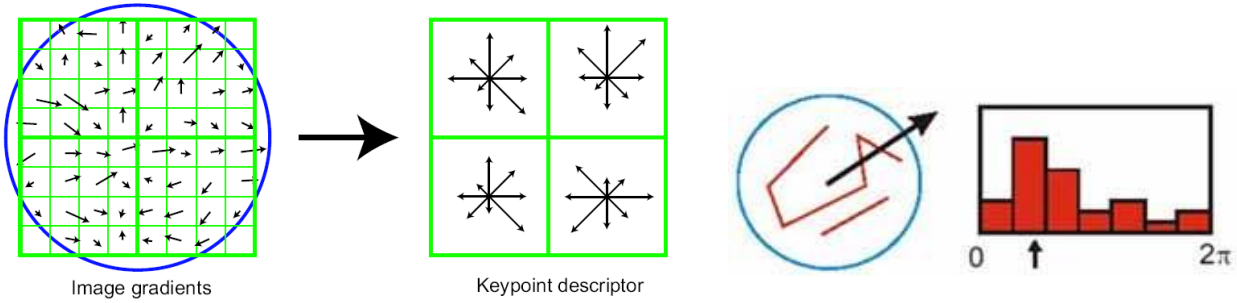
Name	Invariant	Keys
Harris	Rotation	Harris Matrix
Harris-Laplace	Rotation, Scale	Laplacian-of-Gaussian operator
Hession-Laplace	Rotation, Scale	scale selection accuracy is higher than in the case of the Harris-Laplace and DoG in the case of blob-like
Hession-Affine	Rotation, Scale, Affine	Hession-Laplace+affine adaptation
Harris-Affine	Rotation, Scale, Affine	Harris-Laplace + affine adaptation



中期的计算机视觉处理（2004-2012）

❑ 局部描述器：描述图像区块的视觉内容

- SIFT, PCA-SIFT, GLOH, Shape Context, ORB, COGE….

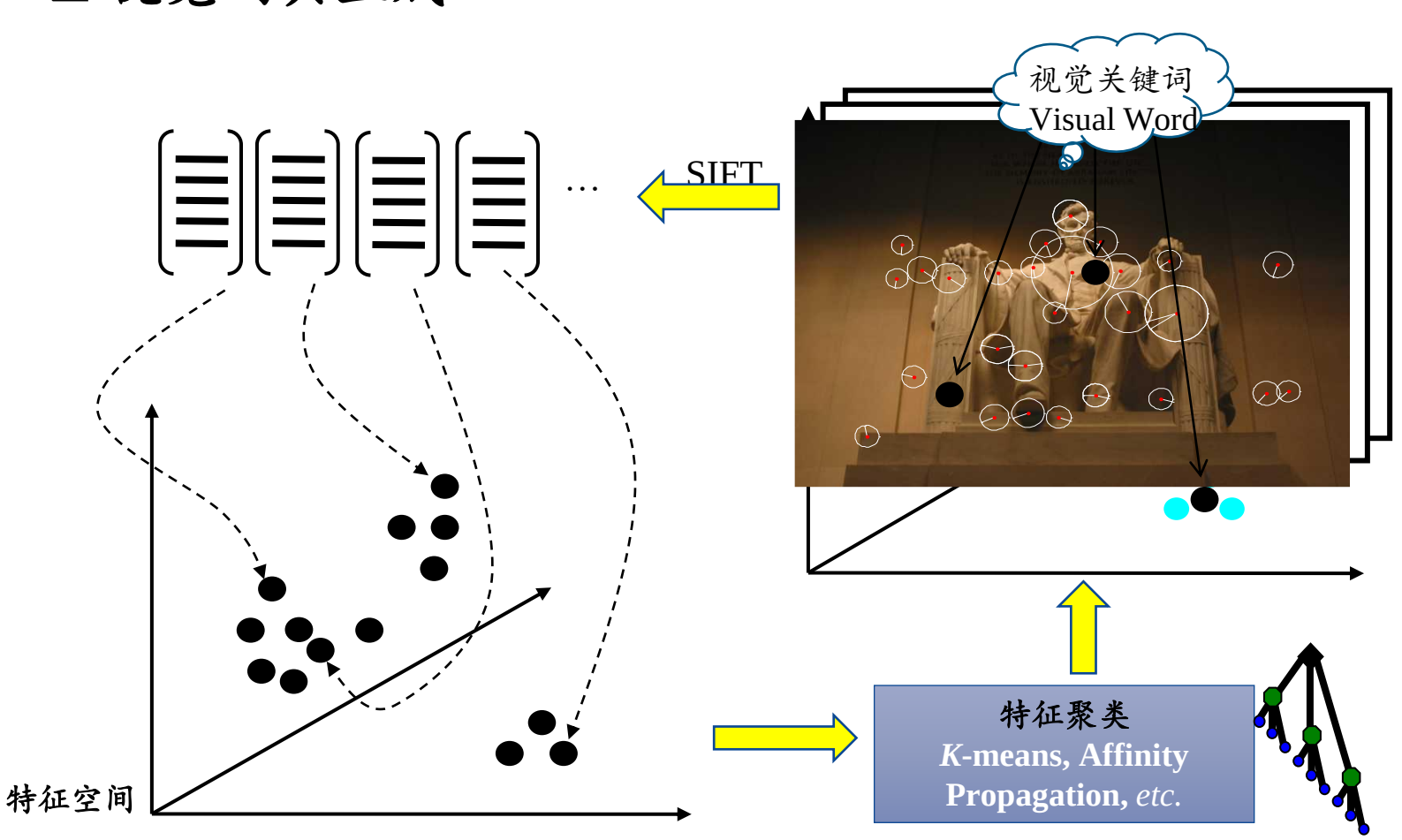


特征提取 → 向量化 → 索引技术 → 后处理

Name	Keys
Gradient location-orientation histogram (GLOH)	3 bins in radial direction and 8 in angular direction, which results in 17 location bins, quantized, reduced with PCA.
Shape context	Edges are extracted by the Canny detector. Location is quantized with the radius.
PCA-SIFT	Image gradient vector in a 39*39 region, reduced to 36 with PCA.

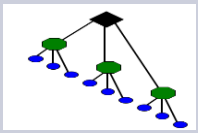
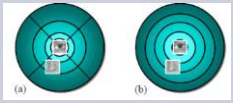
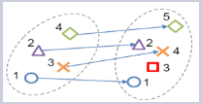
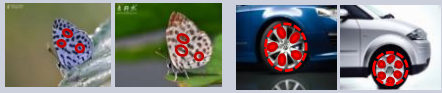
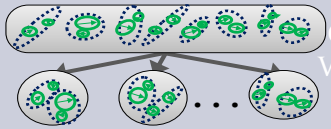
中期的计算机视觉处理（2004-2012）

❑ 视觉词典生成



中期的计算机视觉处理（2004-2012）

❑ 视觉词典生成-相关工作

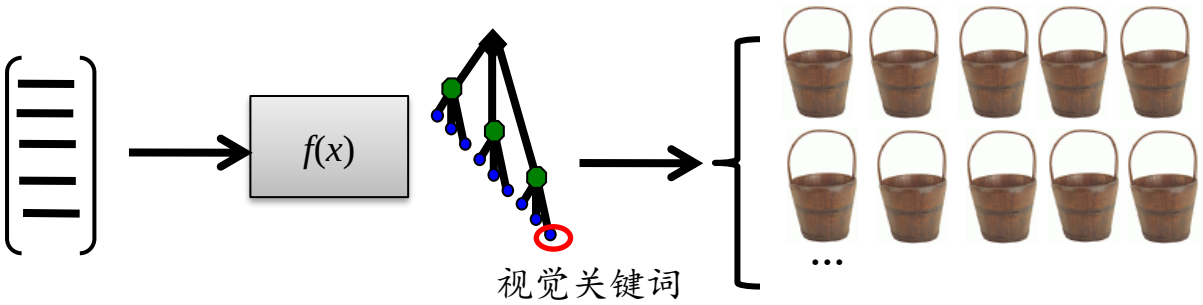
Related Work	Task	Visual word proposed/utilized	Properties
Video Google [Sivic *et al.*, ICCV 03] **Vocabulary Tree** [Nister *et al.*, CVPR06]	image, object, and scene retrieval	 Common Visual word	• Noisy • Low descriptive power • Large-scale
Higher-order Visual Word [Liu *et al.*, CVPR 08] **Collocating Pattern** [Yuan *et al.*, CVPR 07]	object recognition	 Combined spatial information	• Descriptive, spatial info. • Time consuming • Small-scale
Bundled Features [Wu *et al.*, CVPR 09]	near-duplicated image retrieval	 Grouped visual words	• Weak spatial info. • Time consuming • Large quantization error
Visual Synset [Zheng *et al.*, CVPR 08]	object recognition	 Visual Synset	• Descriptive, spatial info • Time consuming • Small-scale
Descriptive Visual words, Descriptive Visual Phrases [Zhang, Tian *et al.*, MM09]	image retrieval, object recognition		• Descriptive, spatial info. • Large-scale • Large quantization error
Contextual Visual Vocabulary [Zhang, Tian *et al.*, MM10]	Image retrieval, image search re-ranking	 Contextual Visual Word	• Descriptive, spatial and semantic info. • Low quantization error • Large-scale

特征提取 → 向量化 → 索引技术 → 后处理

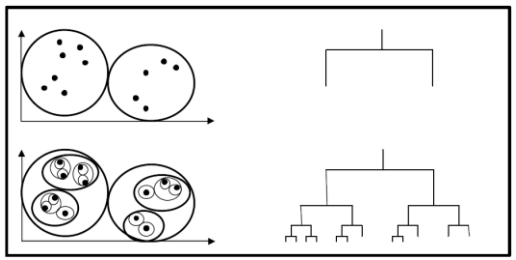
中期的计算机视觉处理（2004-2012）

❑ 局部特征->视觉关键词（Visual Word）

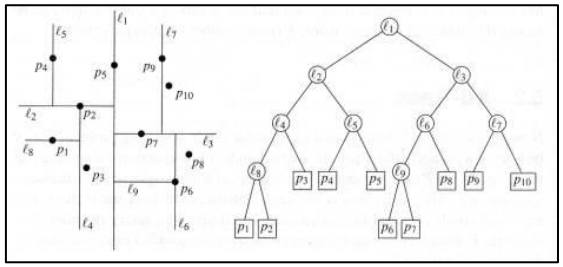
- **局部特征转化为视觉关键词（即特征量化，Feature Quantization）：**对每一个局部特征，查找其在视觉词典里距离最近的视觉关键词，把局部特征的向量转化为该视觉关键词在词典中的序号



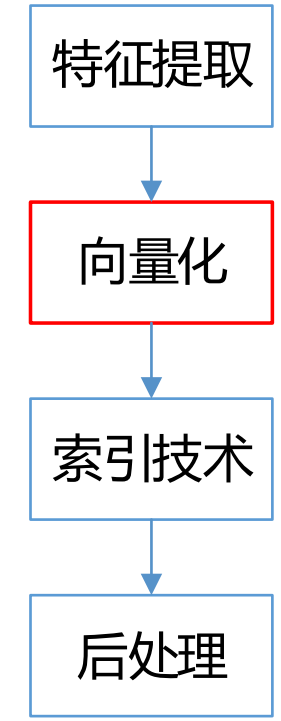
- **常用特征量化技术：**Hierarchical 1-NN，KD-tree



Hierarchical 1-NN



KD-tree



中期的计算机视觉处理（2004-2012）

❑ 基于视觉关键词的图像表示

特征提取

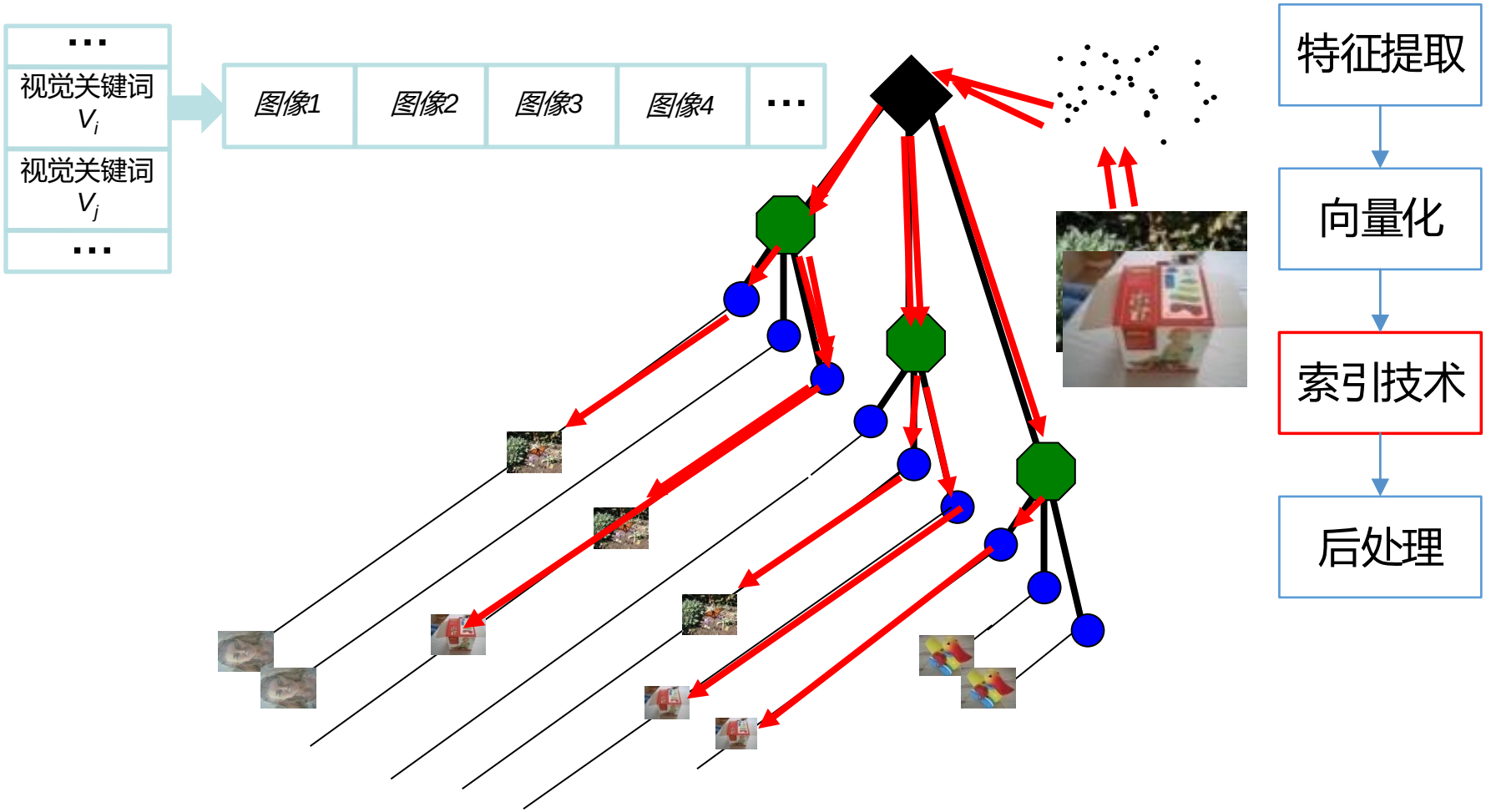
向量化

索引技术

后处理

中期的计算机视觉处理（2004-2012）

❑ 倒排索引

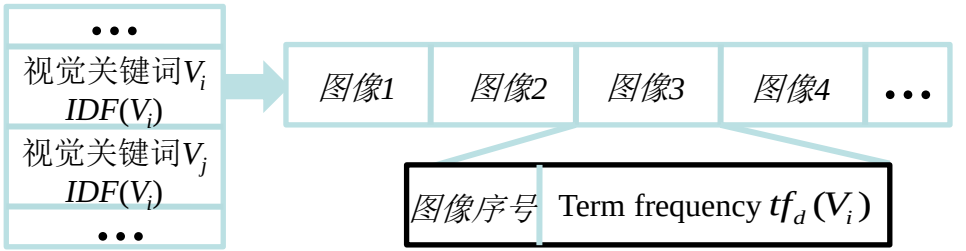


Nister, Scalable Recognition with a Vocabulary Tree, *CVPR 06*.

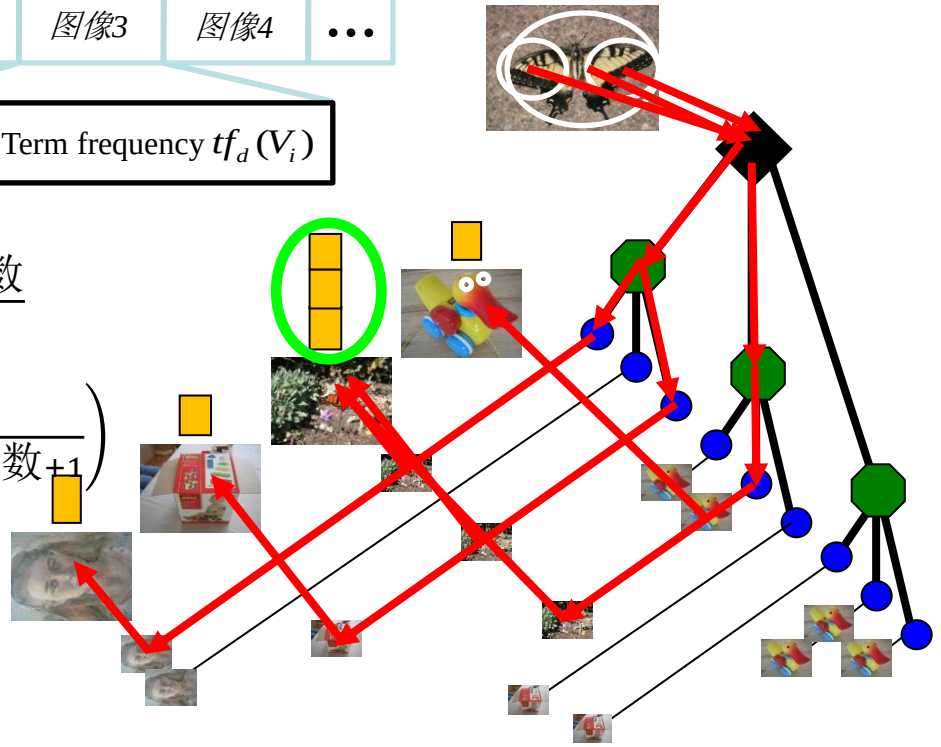
中期的计算机视觉处理（2004-2012）

❑ 排序

- TF-IDF 加权（Term frequency - inverse document frequency）



- $TF = \frac{\text{某文档中词条}v_i\text{出现次数}}{\text{该文档的所有词条}}$
- $IDF = log\left(\frac{\text{文档总数}}{\text{包含词条}v_i\text{的文档数}+1}\right)$



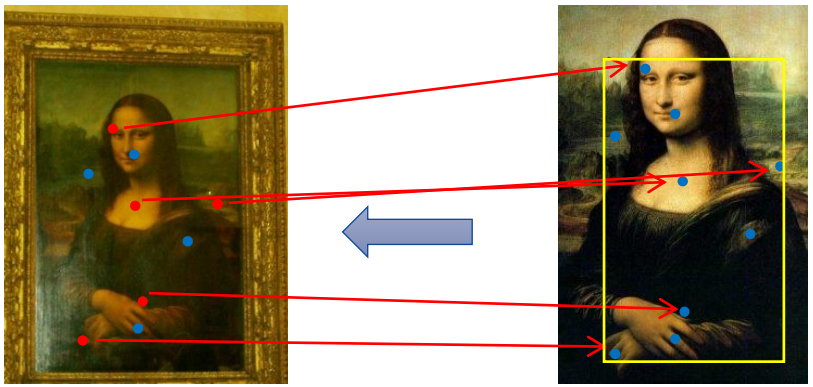
特征提取 → 向量化 → 索引技术 → 后处理

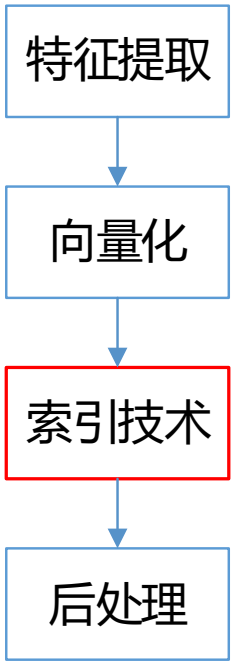
Nister, Scalable Recognition with a Vocabulary Tree, *CVPR 06*.

中期的计算机视觉处理（2004-2012）

❑ 查询扩展

- 目标：使原有查询项含有更多的局部特征，再进行扩展查询





- 后处理技术：局部几何验证（Local Geometric Verification），乘积量化(Product Quantization) 等等

Chum, et al. Total recall: Automatic query expansion with a generative feature model for object retrieval. In *Proc. ICCV*, 2007.

为什么使用深度学习

测量空间 → 特征空间 → 类别空间

核心任务

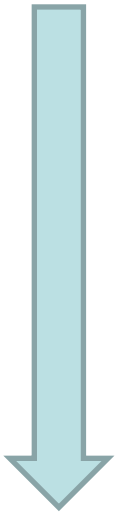
- 传统方法
  - 人工特征提取+分类器

- 深度学习
  - 手工地选取特征是一件非常费力、启发式（需要专业知识）的方法，能不能选取好很大程度上靠经验和运气，而且它的调节需要大量的时间。

深度学习在计算机视觉领域的应用

❑ 计算机视觉领域主要应用

- 图像分类（物体识别）：整幅图像的分类或识别
- 物体检测：检测图像中物体的位置进而识别物体
- 图像分割：对图像中的特定物体按边缘进行分割
- 图像回归：预测图像中物体组成部分的坐标



细化

中国科学院大学
University of Chinese Academy of Sciences

4 图像分类

图像识别的目标

- 计算机观察的图像

我们看到的

0	3	2	5	4	7	6	9	8
3	0	1	2	3	4	5	6	7
2	1	0	3	2	5	4	7	6
5	2	3	0	1	2	3	4	5
4	3	2	1	0	3	2	5	4
7	4	5	2	3	0	1	2	3
6	5	4	3	2	1	0	3	2
9	6	7	4	5	2	3	0	1
8	7	6	5	4	3	2	1	0

计算机看到的

图像识别面临的挑战

❑ 语义鸿沟(Semantic Gap)现象

- Semantic Gap: the gap between low-level visual features and high-level concepts（图像的底层视觉特性和高层语义概念之间的鸿沟）
- 例1：相似的视觉特性(color, texture, shape，…) ,不同的语义概念

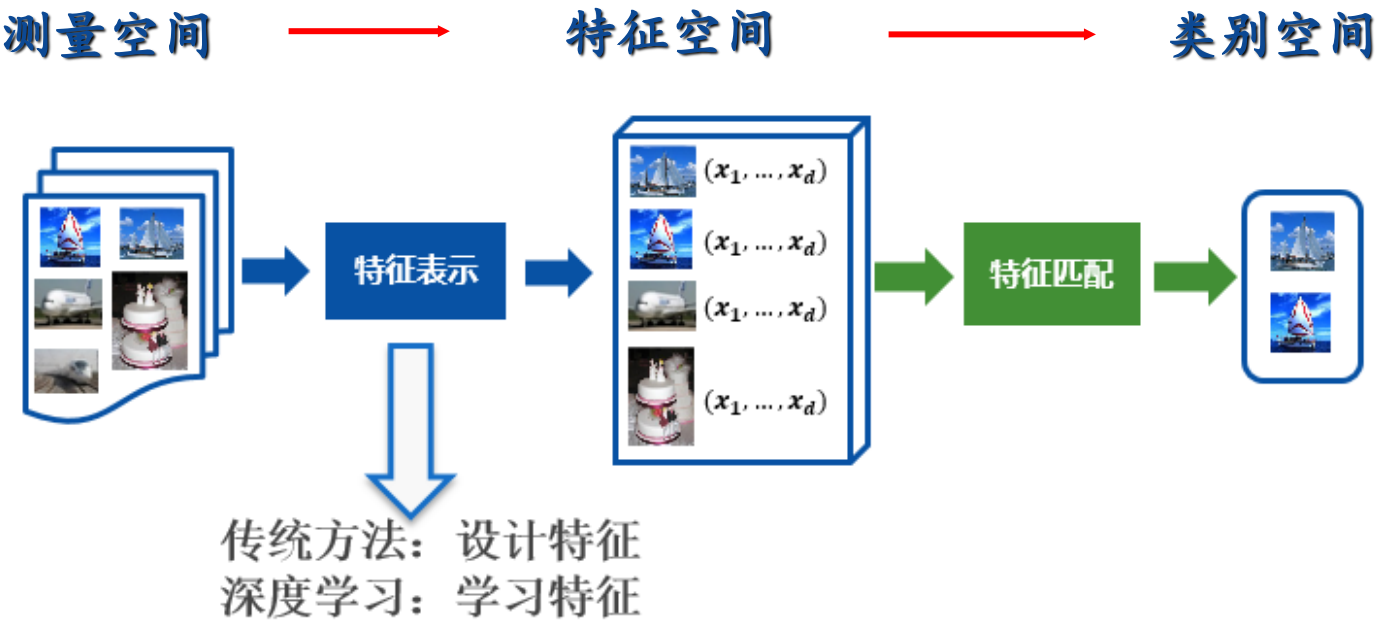
图像识别面临的挑战

❑ 语义鸿沟(Semantic Gap)现象

– 例2：不相同的视觉特性(color, texture, shape，…) ,相同的语义概念

图像识别基本框架

❑ 图像识别框架



图像识别基本框架（场景识别、目标识别、人脸识别、…）

常见数据集

- MNIST（Mixed National Institute of Standards and Technology）
- CIFAR（Canada Institute For Advanced Research）
- Places2
- Cats vs Dogs
- ImageNet
- PASCAL VOC

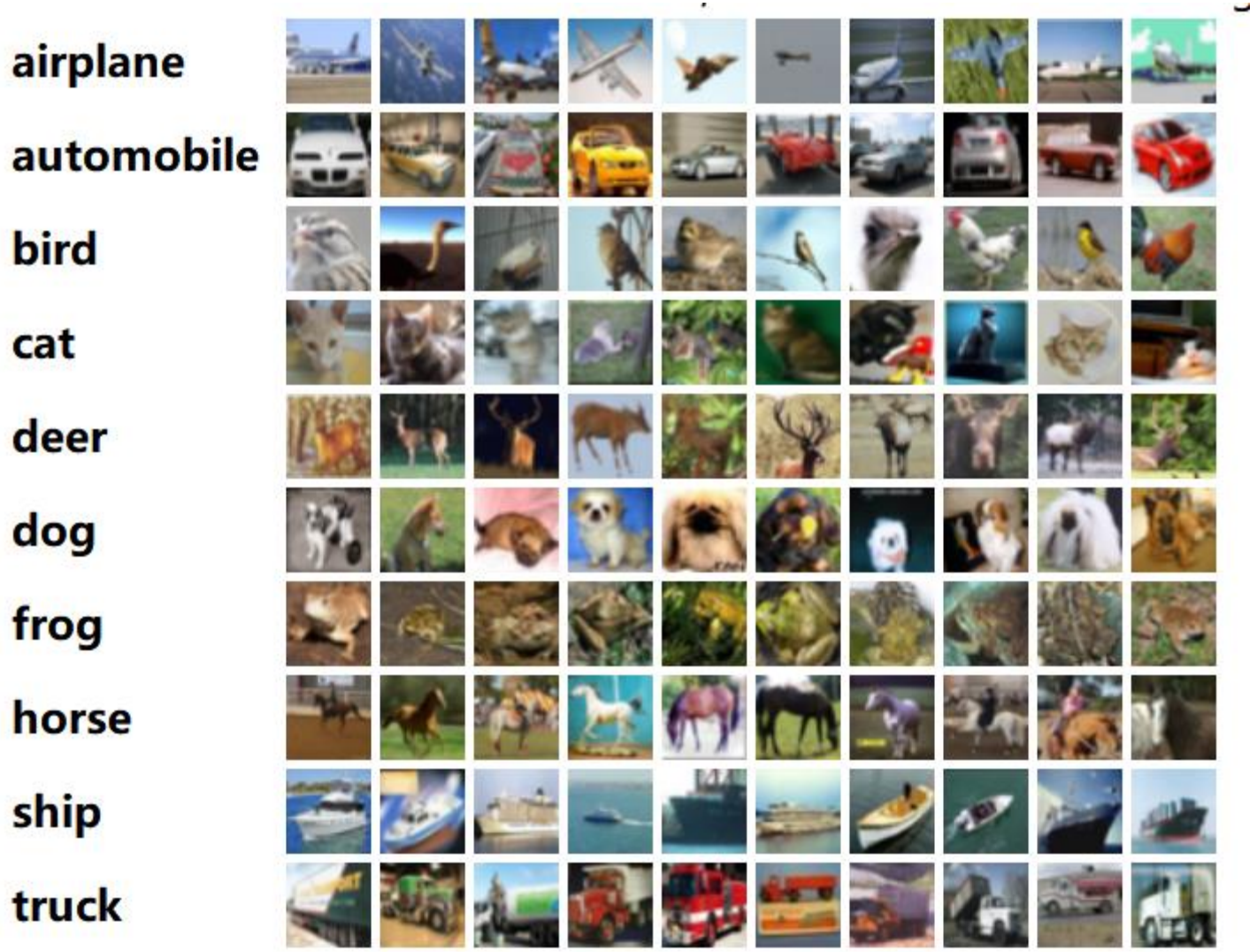
MNIST

- MNIST是一个大型的手写数字数据库，广泛用于机器学习领域的训练和测试
- MNIST包含60000个训练集，10000个测试集，每张图都进行了尺度归一化和数字居中处理，固定尺寸大小为28*28。
- 由纽约大学的Yann LeCun整理。
- 主页：http://yann.lecun.com/exdb/mnist/

CIFAR

- CIFAR是由加拿大先进技术研究院的AlexKrizhevsky, Vinod Nair和Geoffrey Hinton收集而成的小图片数据集，包含CIFAR-10和CIFAR-100两个数据集。
- CIFAR-10由60000张32*32的RGB彩色图片构成，共10个分类。50000张训练，10000张测试（交叉验证）
- CIFAR-100由60000张图像构成，包含100个类别，每个类别600张图像，其中500张用于训练，100张用于测试。其中这100个类别又组成了20个大的类别，每个图像包含小类别和大类别两个标签
- 主页：http://www.cs.toronto.edu/~kriz/cifar.html

CIFAR-10


airplane
automobile
bird
cat
deer
dog
frog
horse
ship
truck

CIFAR-100

Superclass	Classes
aquatic mammals	beaver, dolphin, otter, seal, whale
fish	aquarium fish, flatfish, ray, shark, trout
flowers	orchids, poppies, roses, sunflowers, tulips
food containers	bottles, bowls, cans, cups, plates
fruit and vegetables	apples, mushrooms, oranges, pears, sweet peppers
household electrical devices	clock, computer keyboard, lamp, telephone, television
household furniture	bed, chair, couch, table, wardrobe
insects	bee, beetle, butterfly, caterpillar, cockroach
large carnivores	bear, leopard, lion, tiger, wolf
large man-made outdoor things	bridge, castle, house, road, skyscraper
large natural outdoor scenes	cloud, forest, mountain, plain, sea
large omnivores and herbivores	camel, cattle, chimpanzee, elephant, kangaroo
medium-sized mammals	fox, porcupine, possum, raccoon, skunk
non-insect invertebrates	crab, lobster, snail, spider, worm
people	baby, boy, girl, man, woman
reptiles	crocodile, dinosaur, lizard, snake, turtle
small mammals	hamster, mouse, rabbit, shrew, squirrel
trees	maple, oak, palm, pine, willow
vehicles 1	bicycle, bus, motorcycle, pickup truck, train
vehicles 2	lawn-mower, rocket, streetcar, tank, tractor

Places2

- Places2 是由MIT开发的一个场景图像数据集，可用于以场景和环境为应用内容的视觉认知任务
- 包含 1千万张 图片，400多个不同类型的场景环境，每一类有5000-30000张图片
- 主页：http://places2.csail.mit.edu/index.html
- 论文：http://places2.csail.mit.edu/PAMI_places.pdf

Places2

Fig. 1. Image samples from various categories of the Places Database (two samples per category). The dataset contains three macro-classes: Indoor, Nature, and Urban.

Cats vs Dogs

- 猫狗分类数据集，共25000张图片，猫、狗各12500张。
- 下载链接：https://www.kaggle.com/c/dogs-vs-cats/data




ImageNet

- 美国斯坦福的计算机科学家李飞飞建立的，能够从图片识别物体
- 是目前世界上图像识别最大的数据库，目前已经包含14197122张图像，21841 indexed synsets
- 举办多届竞赛——ILSVRC比赛，全称是ImageNet Large-Scale Visual Recognition Challenge。使用的数据集是ImageNet的一个子集，总共有1000类，每类大约有1000张图像。具体地，有大约1.2 million的训练集，5万验证集，15万测试集
- 主页：http://www.image-net.org/

ILSVRC历届比赛情况

年	网络/队名	val top-1	val top-5	test top-5	备注
2012	AlexNet	38.1%	16.4%	16.42%	5 CNNs
2012	**AlexNet**	36.7%	15.4%	15.32%	7CNNs。用了2011年的数据
2013	OverFeat			14.18%	7 fast models
2013	OverFeat			13.6%	赛后。7 big models
2013	ZFNet			13.51%	ZFNet论文上的结果是14.8
2013	Clarifai			11.74%	
2013	**Clarifai**			11.20%	用了2011年的数据

ILSVRC历届比赛情况

2014	VGG			7.32%	7 nets, dense eval
2014	VGG（亚军）	23.7%	6.8%	6.8%	赛后。2 nets
2014	**GoogleNet v1**			6.67%	7 nets, 144 crops
	GoogleNet v2	20.1%	4.9%	4.82%	赛后。6 nets, 144 crops
	GoogleNet v3	17.2%	3.58%		赛后。4 nets, 144 crops
	GoogleNet v4	16.5%	3.1%	3.08%	赛后。v4+Inception-Res-v2
2015	**ResNet**			3.57%	6 models
2016	**Trimps-Soushen**			2.99%	公安三所
2016	ResNeXt（亚军）			3.03%	加州大学圣地亚哥分校
2017	**SENet**			2.25%	Momenta 与牛津大学

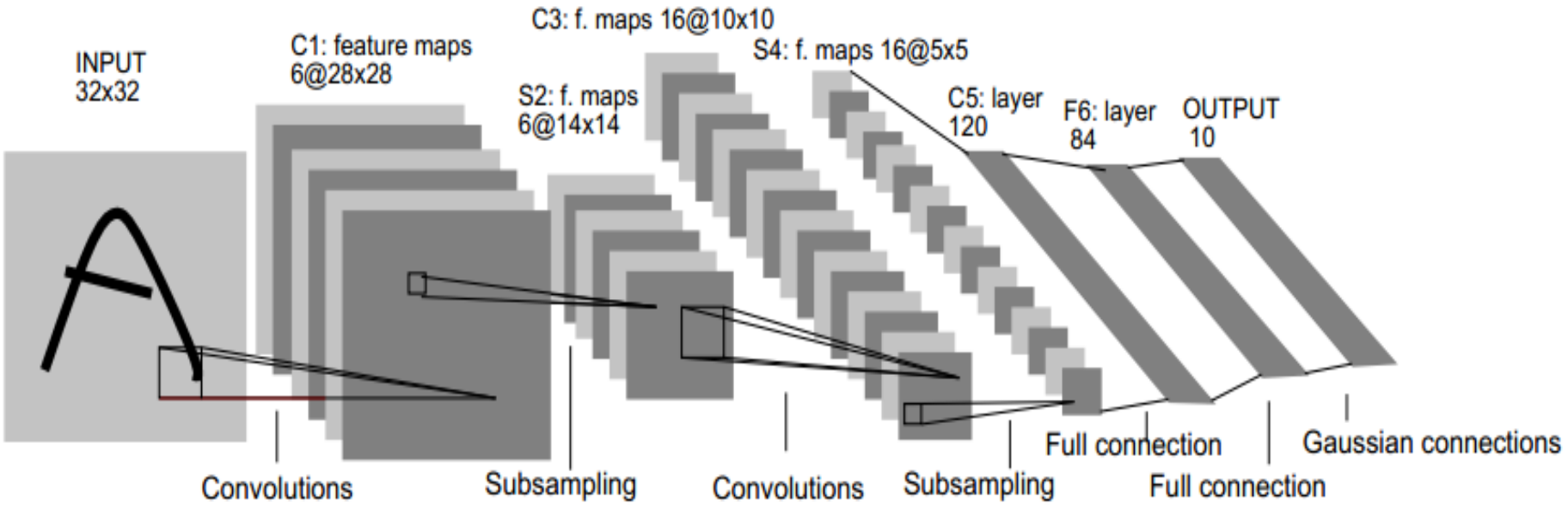
PASCAL VOC

- PASCAL VOC 数据集是视觉对象的分类识别和检测的一个基准测试集，提供了检测算法和学习性能的标准图像注释数据集和标准的评估系统
- 图像包含VOC2007（430M），VOC2012（1.9G）两个下载版本。
- 下载链接：http://pjreddie.com/projects/pascal-voc-dataset-mirror/

图像分类模型

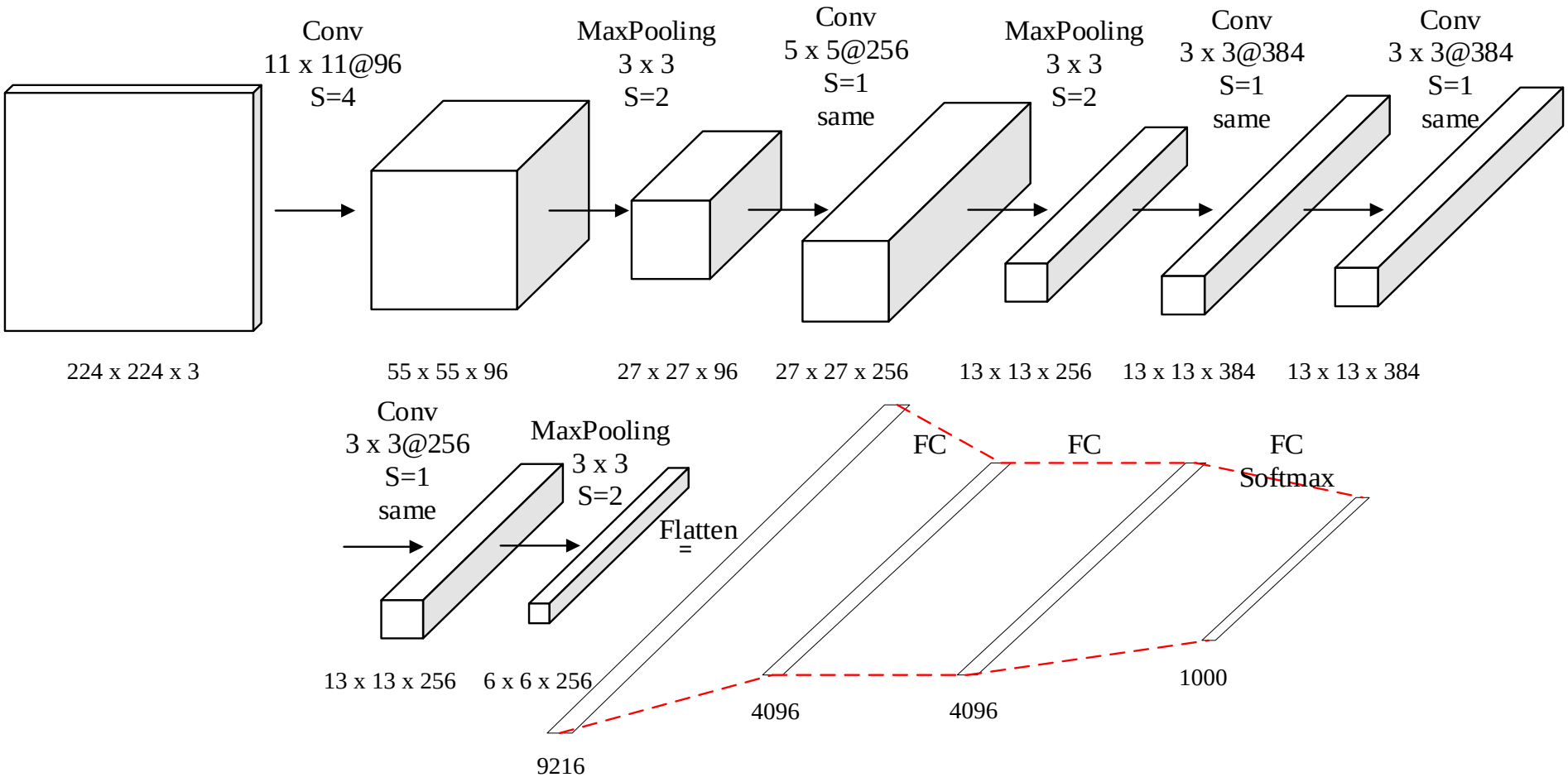
- LeNet-5
- AlexNet
- VGGNet
- Inception Net
- ResNet

LeNet-5

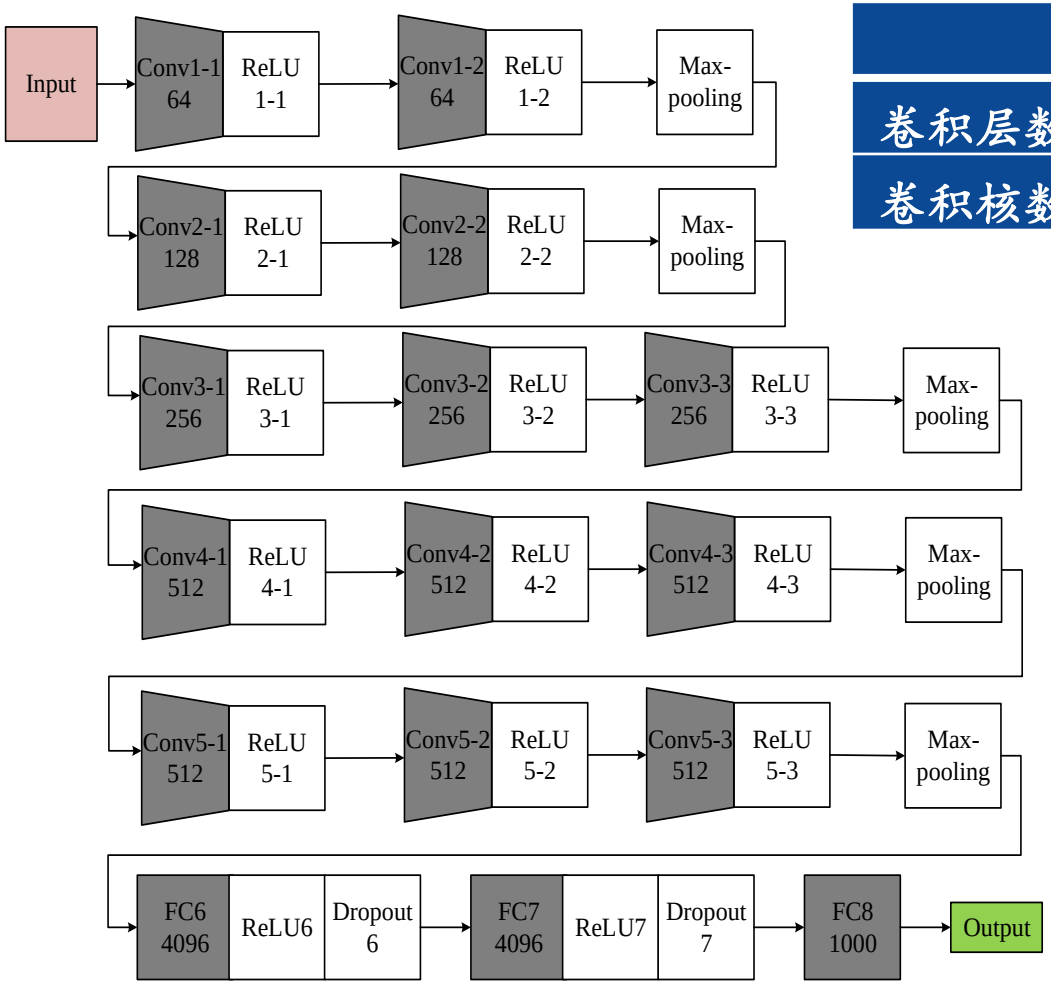


Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. Proceedings of the IEEE, November 1998.

AlexNet


Conv
11 x 11@96
S=4
MaxPooling
3 x 3
S=2
Conv
5 x 5@256
S=1
same
MaxPooling
3 x 3
S=2
Conv
3 x 3@384
S=1
same
Conv
3 x 3@384
S=1
same
224 x 224 x 3
55 x 55 x 96
27 x 27 x 96
27 x 27 x 256
13 x 13 x 256
13 x 13 x 384
13 x 13 x 384
Conv
3 x 3@256
S=1
same
MaxPooling
3 x 3
S=2
Flatten
=
FC
FC
FC
Softmax
13 x 13 x 256
6 x 6 x 256
9216
4096
4096
1000

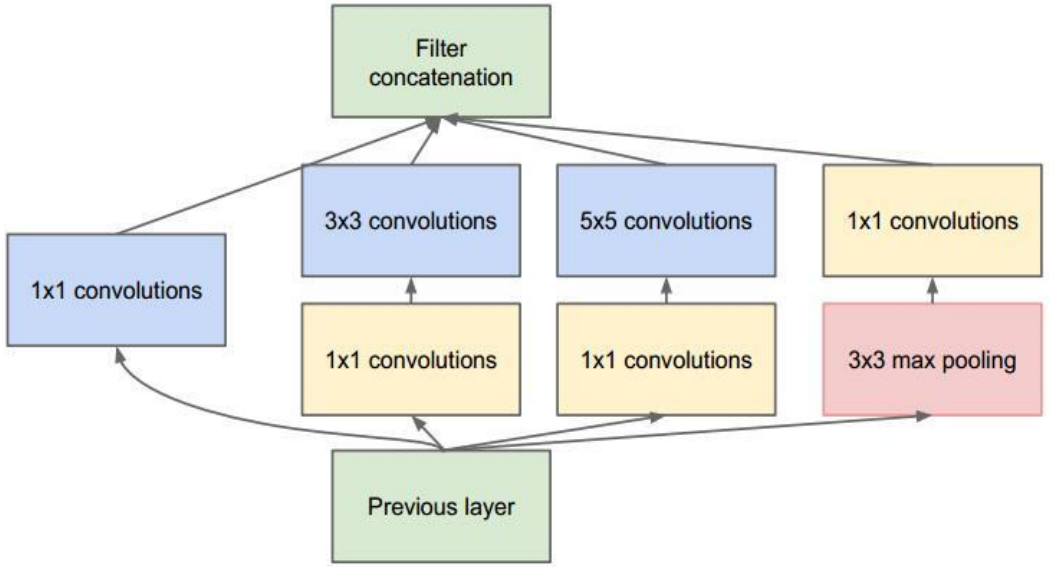
VGG-16



	Conv1	Conv2	Conv3	Conv4	Conv5
卷积层数	2	2	3	3	3
卷积核数	64	128	256	512	512

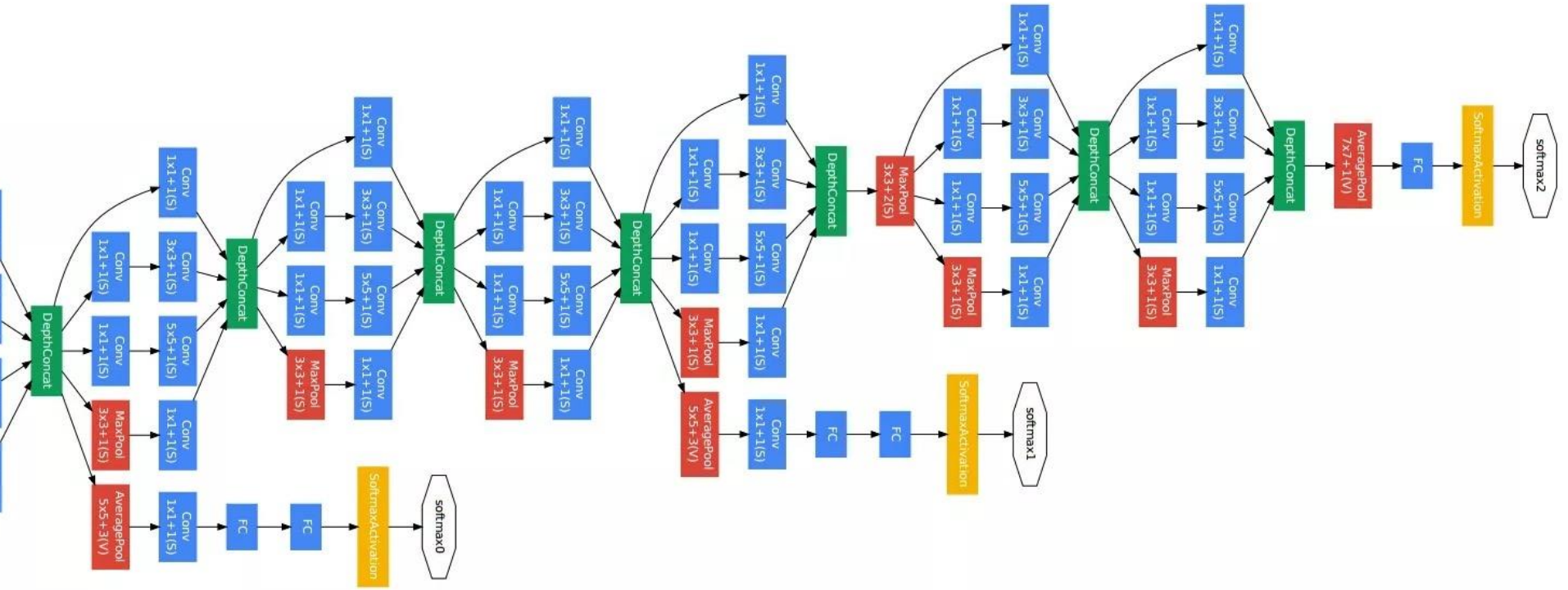
Inception Net

- 深度：层数更深，采用了22层，在不同深度处增加了两个loss来避免上述提到的梯度消失问题
- 宽度：Inception Module包含4个分支，在卷积核3x3、5x5之前、max pooling之后分别加上了1x1的卷积核，起到了降低特征图厚度的作用

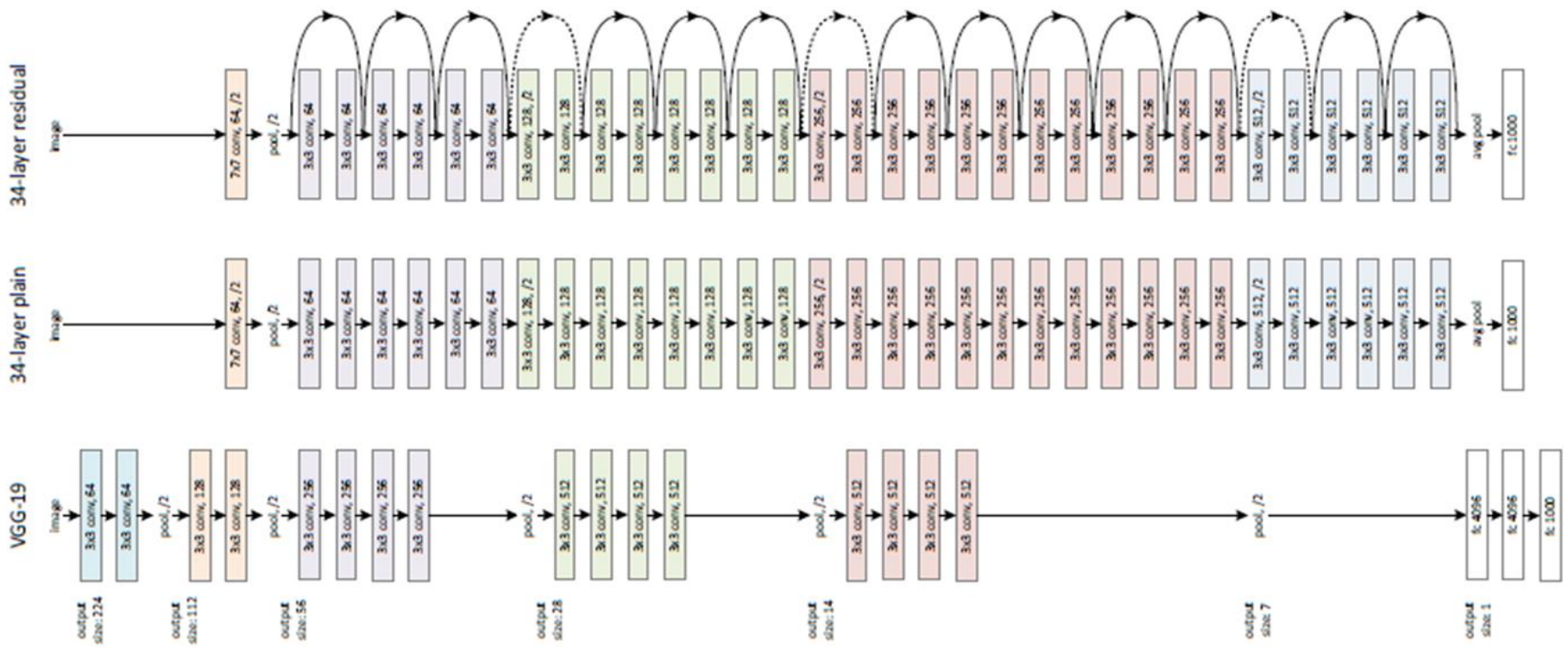


Inception Module

Inception Net



ResNet


VGG-19
34-layer plain
34-layer residual
image
3x3 conv, 64
pool, /2
output size: 224
output size: 112
output size: 56
output size: 28
output size: 14
output size: 7
output size: 1
fc 4096
fc 1000
7x7 conv, 64, /2
3x3 conv, 128
3x3 conv, 256
3x3 conv, 512
avg pool

中国科学院大学
University of Chinese Academy of Sciences

5 目标检测

常见目标检测数据集

- **PASCAL VOC**
- **ImageNet**
- **MS COCO**
- **KITTI**

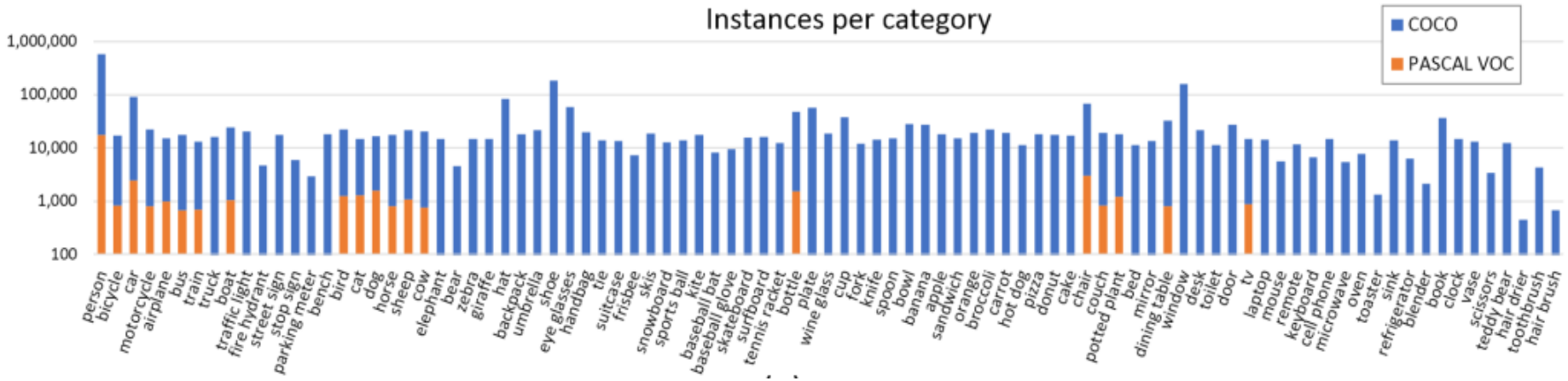
MS COCO

- MS COCO（Common Objects in COntext）数据集是Microsoft公司建立的。
- 这个数据集用于多种竞赛：图像标题生成，目标检测，关键点检测和物体分割。
- 对于目标检测任务，COCO共包含80个类别，每年大赛的训练和验证数据集包含超过120,000个图片，超过40,000个测试图片。
- 主页：http://cocodataset.org/
- 论文： https://arxiv.org/pdf/1405.0312.pdf

MS COCO

(a) Iconic object images (b) Iconic scene images (c) Non-iconic images

Fig. 2: Example of (a) iconic object images, (b) iconic scene images, and (c) non-iconic images.



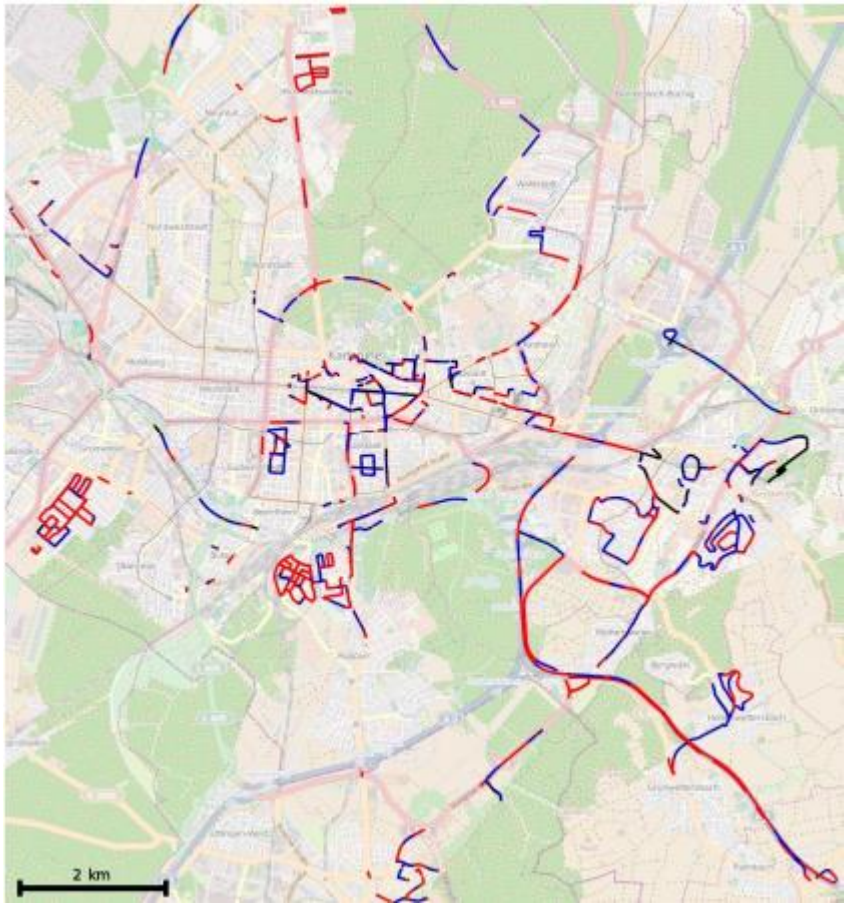
KITTI

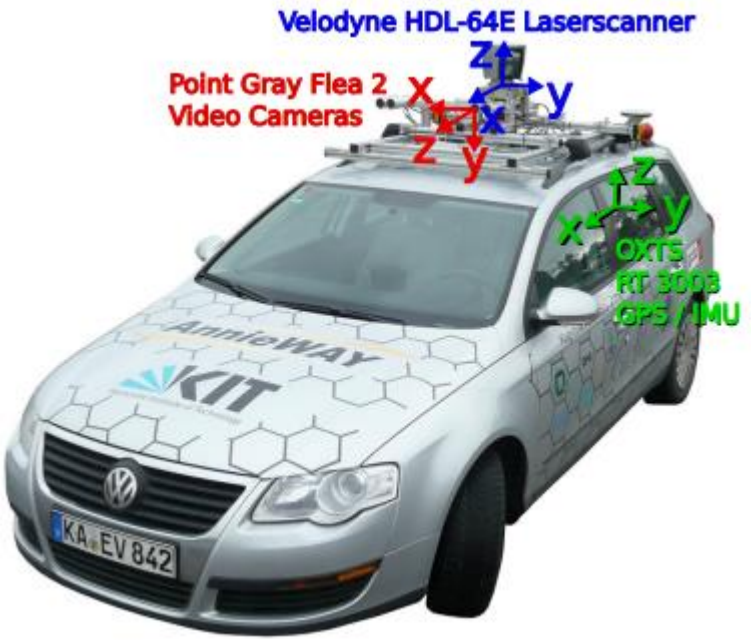
- KITTI数据集由德国卡尔斯鲁厄理工学院和丰田美国技术研究院联合创办，是目前国际上最大的自动驾驶场景下的计算机视觉算法评测数据集
- 该数据集用于评测立体图像(stereo)，光流(optical flow)，视觉测距(visual odometry)，3D物体检测(object detection)和3D跟踪(tracking)等计算机视觉技术在车载环境下的性能
- KITTI包含市区、乡村和高速公路等场景采集的真实图像数据，每张图像中最多达15辆车和30个行人，还有各种程度的遮挡与截断

KITTI

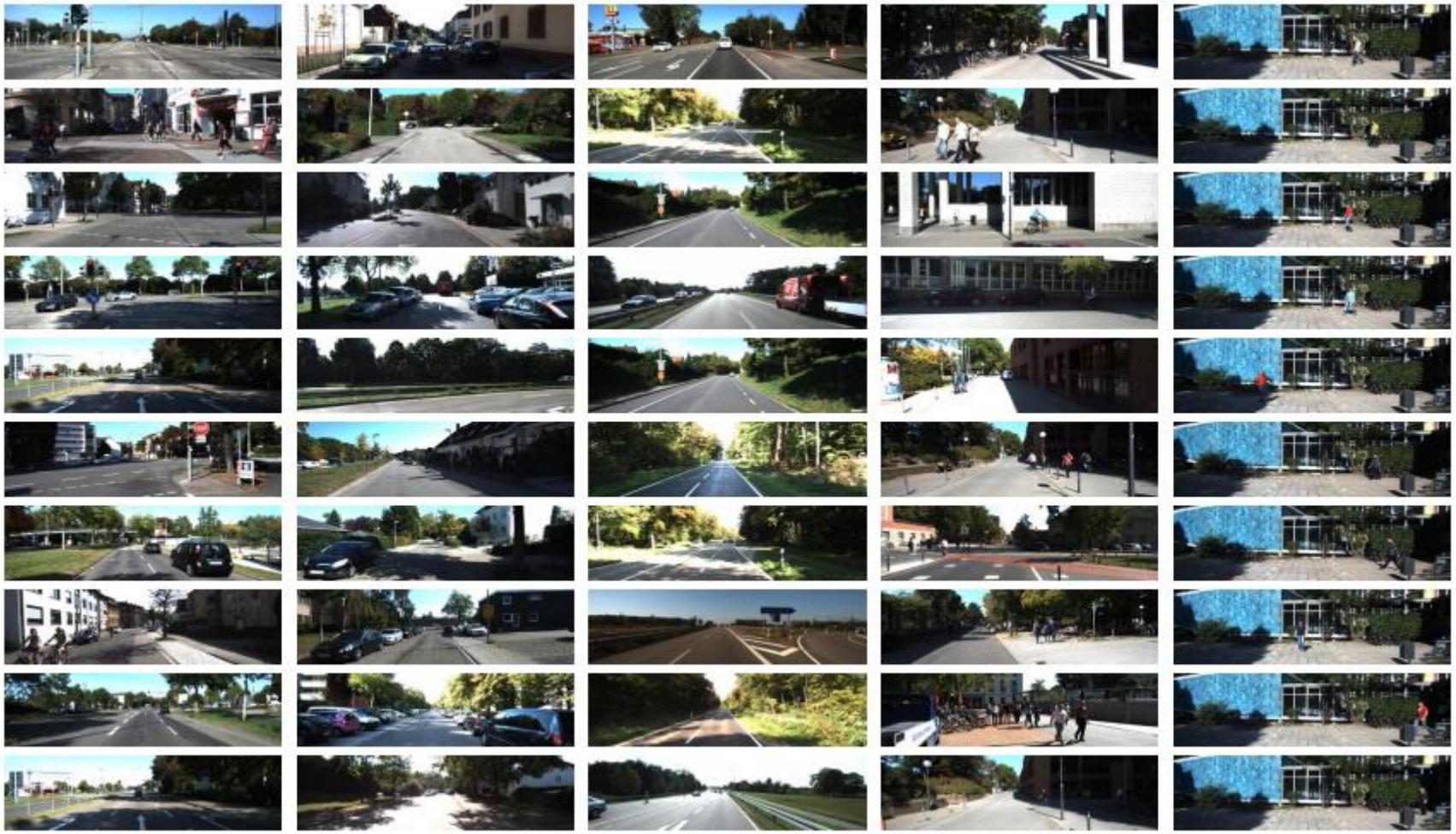
- 整个数据集由389对立体图像和光流图，39.2 km视觉测距序列以及超过200k 3D标注物体的图像组成，以10Hz的频率采样及同步
- 总体上看，原始数据集被分为Road, City, Residential, Campus 和 Person五大类
- 对于3D物体检测，label细分为car, van, truck, pedestrian, pedestrian(sitting), cyclist, tram以及misc组成
- 论文
  - http://www.cvlibs.net/publications/Geiger2013IJRR.pdf
  - http://www.cvlibs.net/publications/Geiger2012CVPR.pdf

KITTI


2 km

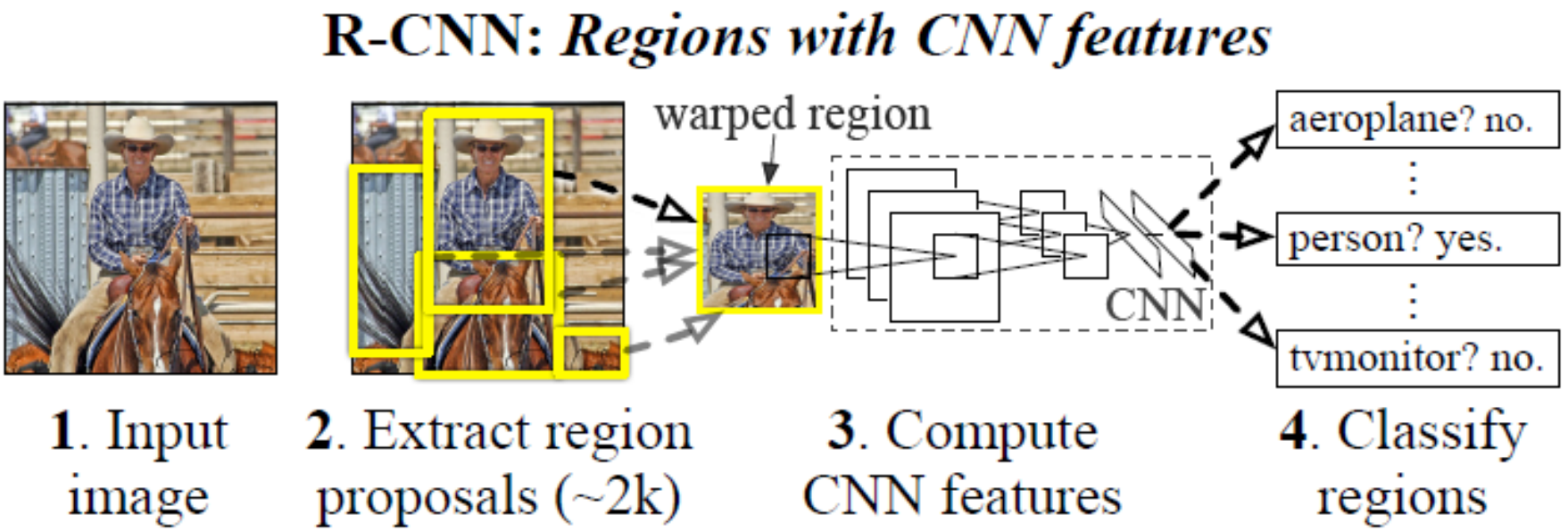

Velodyne HDL-64E Laserscanner
Point Gray Flea 2
Video Cameras
OXTS
RT 3003
GPS / IMU
KIT
KA EV 842

KITTI



R-CNN系列

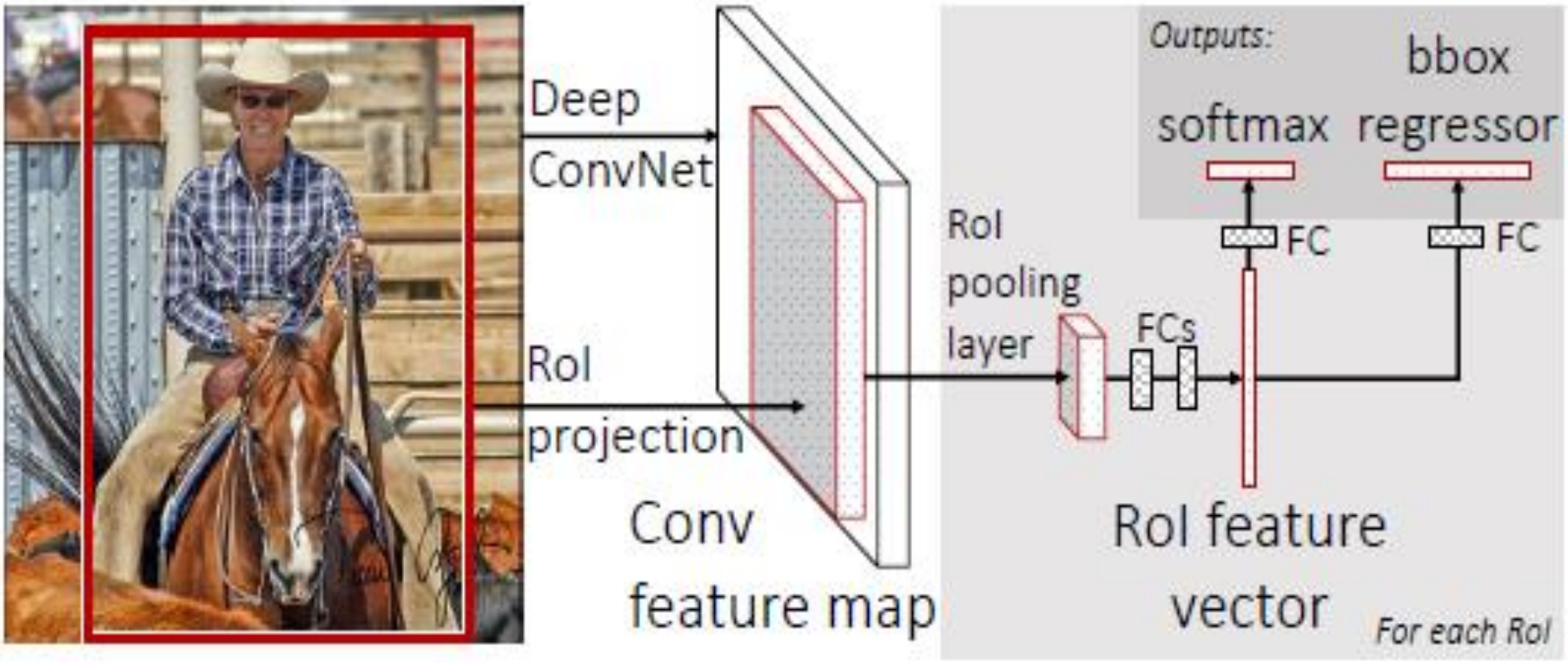
❑ R-CNN



Ross B. Girshick, Jeff Donahue, Trevor Darrell, Jitendra Malik. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. CVPR 2014: 580-587

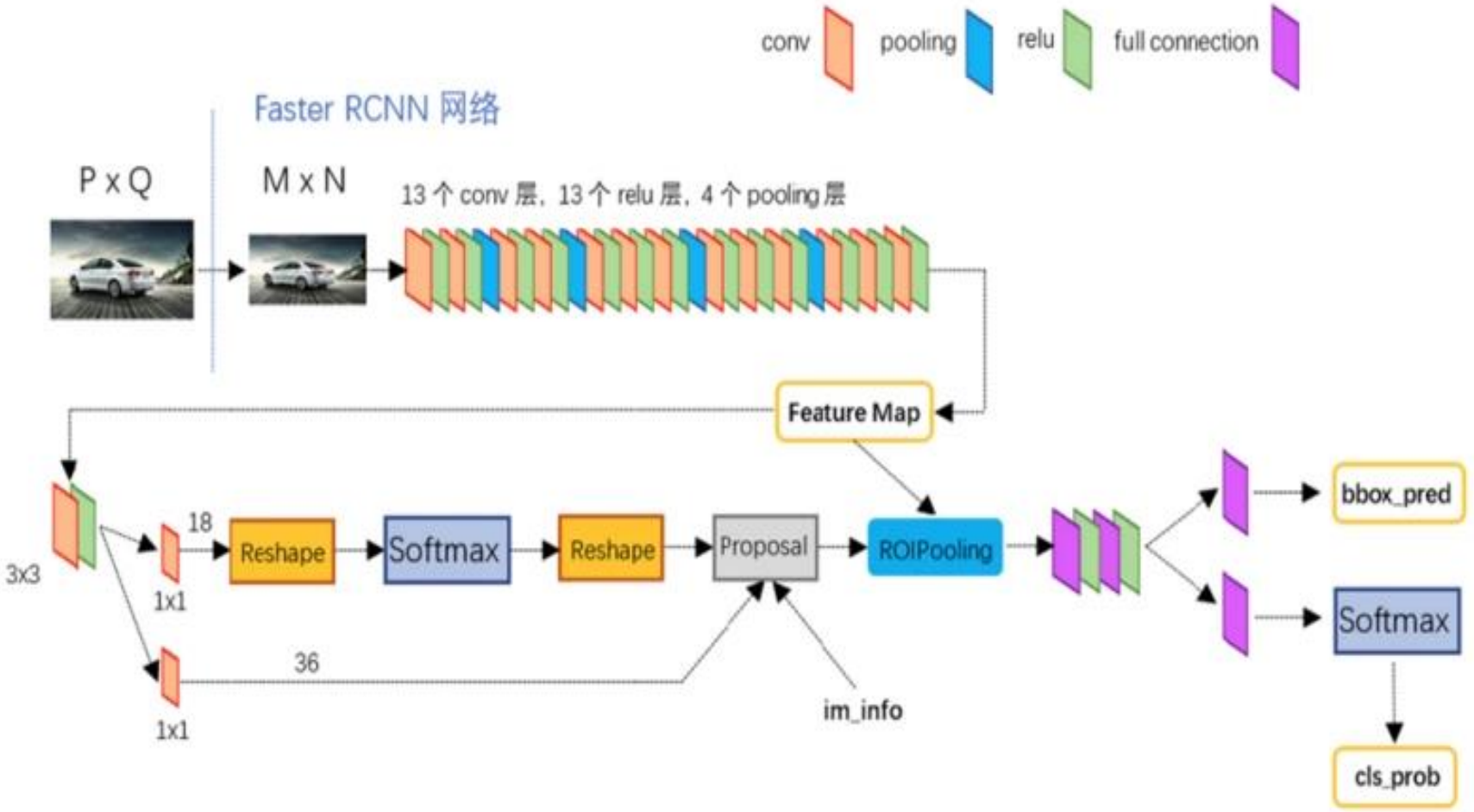
R-CNN系列

- **Fast R-CNN**



R-CNN系列

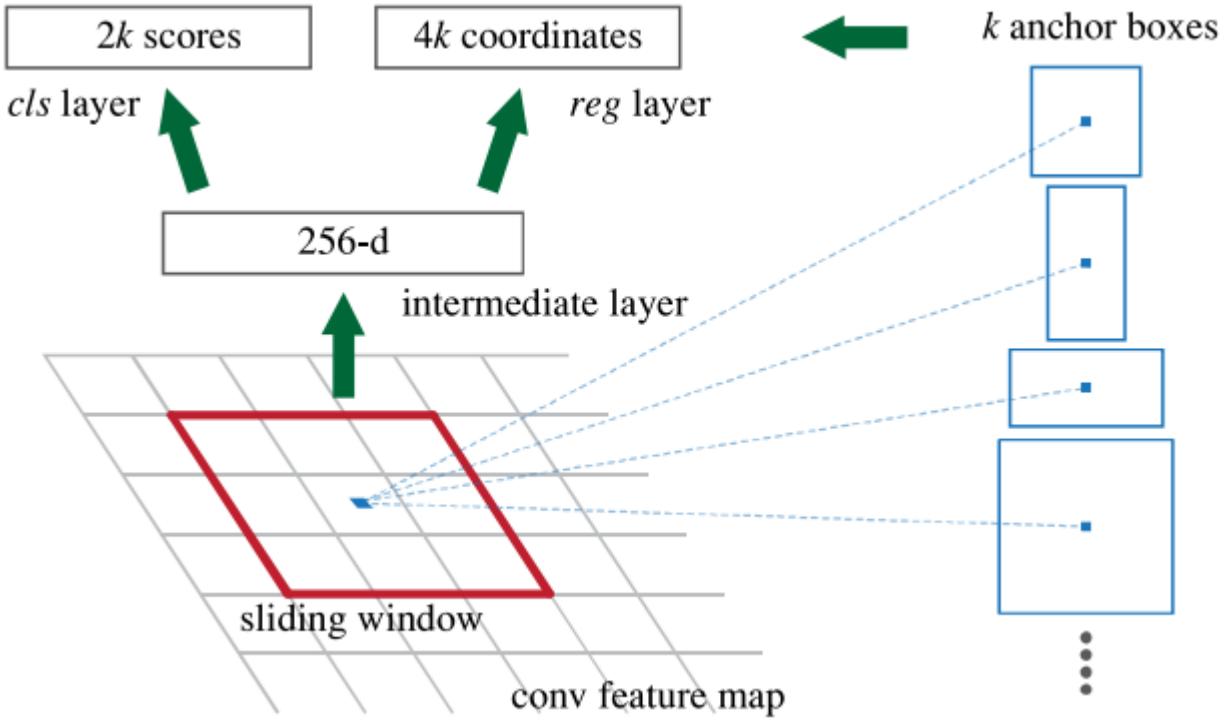
❑ Faster R-CNN



R-CNN系列

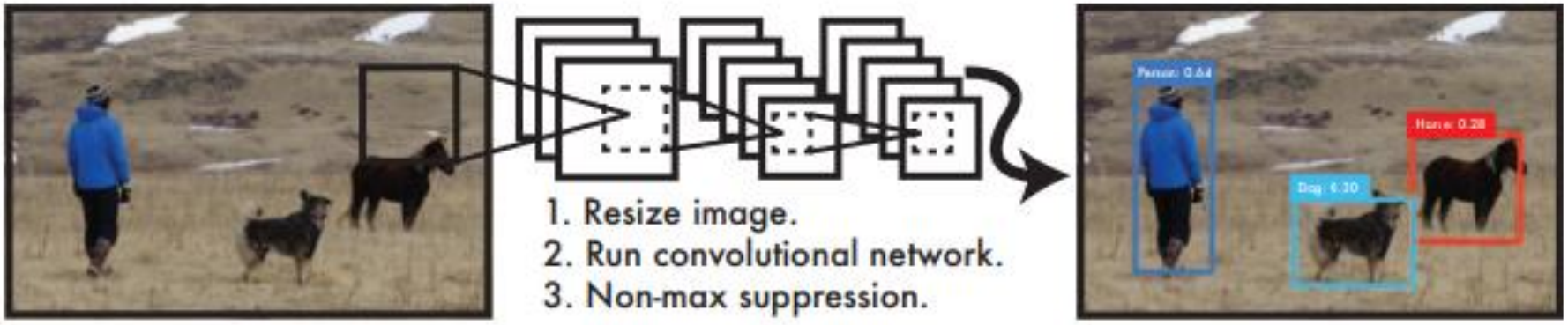
❑ RPN

- Anchor内是某个object 的概率
- Anchor边界框回归输出(Δx, Δy, Δw, Δh)



YOLO系列

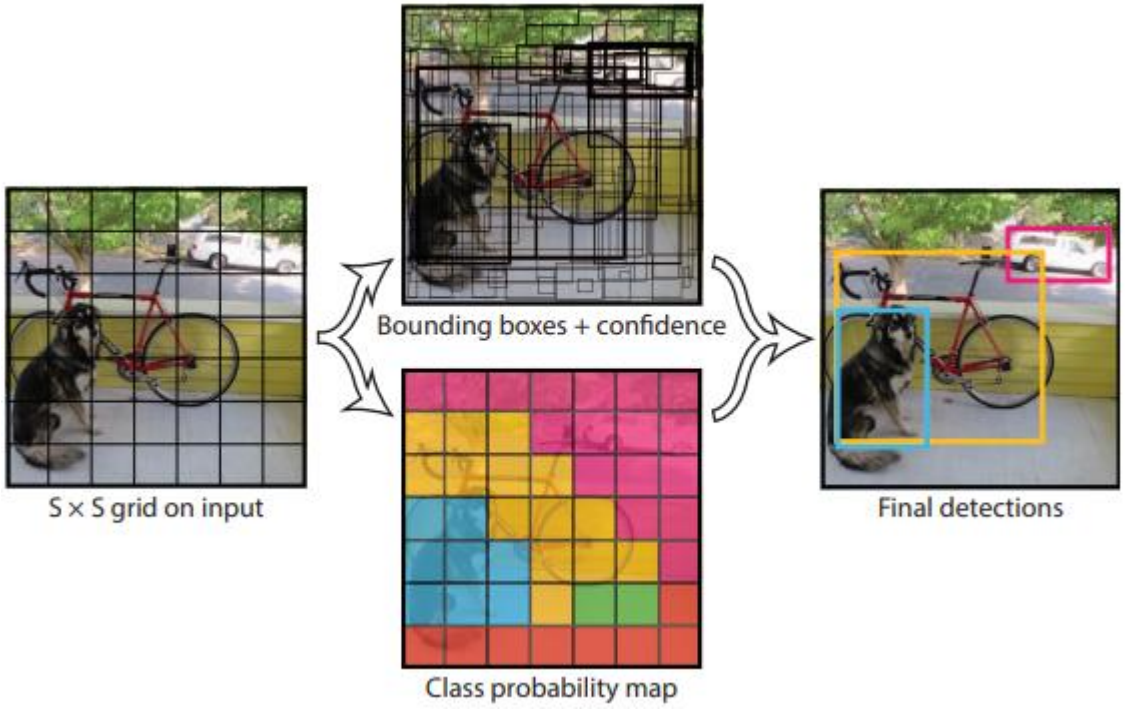
❑ 目标检测和识别



The YOLO Detection System. Processing images with YOLO is simple and straightforward. Our system (1) resizes the input image to 448 × 448, (2) runs a single convolutional network on the image, and (3) thresholds the resulting detections by the model's confidence

YOLO系列

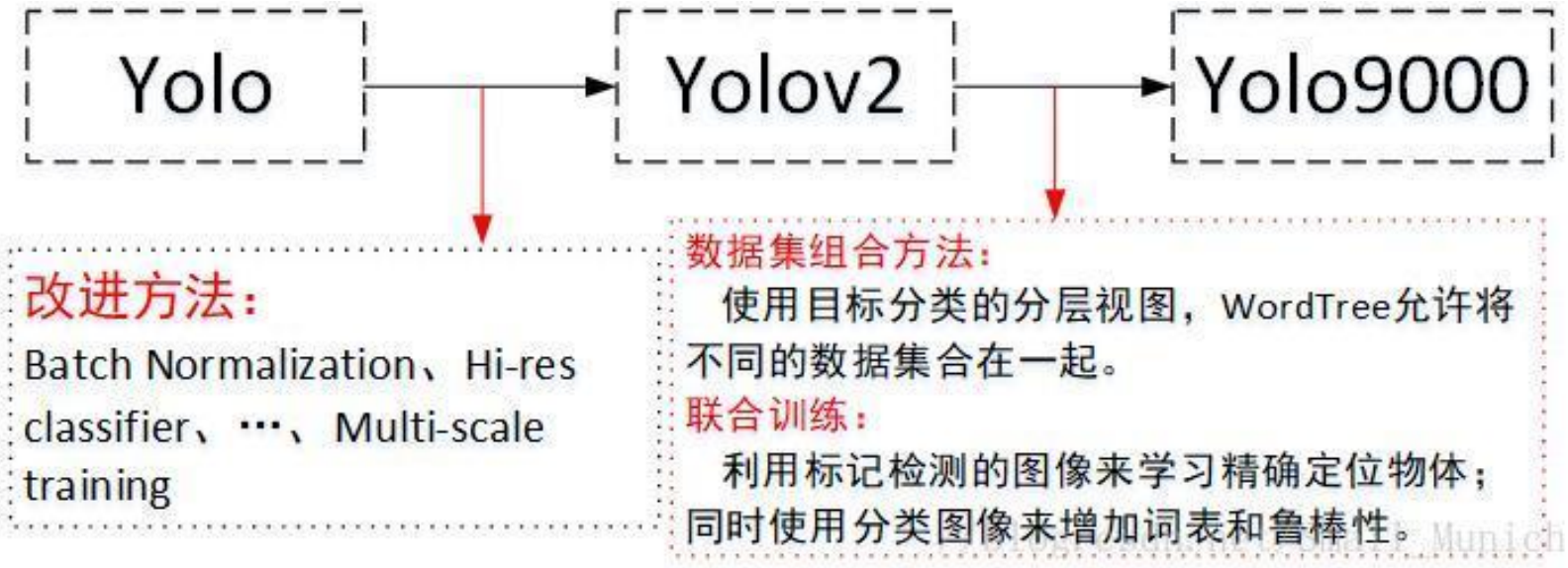
❑ 目标检测和识别



The Model. Our system models detection as a regression problem. It divides the image into an S×S grid and for each grid cell predicts B bounding boxes, confidence for those boxes, and C class probabilities. These predictions are encoded as an S×S× (B∗5+ C) tensor.

YOLO系列

❑ YOLO2和YOLO9000



Joseph Redmon, Ali Farhadi. YOLO9000: Better, Faster, Stronger. CVPR 2017: 6517-6525

目标检测算法性能对比

-	One-stage系列	Two-stage系列
代表性算法	YOLOv1、SSD、YOLOv2、YOLOv3	R-CNN、SPPNet、Fast R-CNN、Faster R-CNN
检测精度	低	高
检测速度	快	慢

One-stage系列中的类别不平衡

- One-stage系列目标检测算法早期会生成大量的bbox。而一幅常规的图片中，没有几个object。这意味着，绝大多数的bbox属于背景（background）
- One-stage系列目标检测算法的detector直接在前面生成的“类别极不平衡”的bbox中就进行难度极大的细分类，意图直接输出bbox和标签（分类结果）
- 而原有交叉熵损失(CE)作为分类任务的损失函数，无法抗衡“类别极不平衡”，容易导致分类器训练失败。因此，One-stage目标检测算法虽然保住了检测速度，却丧失了检测精度

Focal Loss

- **类别不平衡是one-stage detector在精度上逊于two-stage detector的病结所在。**
- **原先训练回归任务使用的交叉熵损失改为焦点损失 (focal loss)：**
  - 以二分类为例：$\mathrm{CE}(p_t) = -\log(p_t)$ $\quad p_t = \begin{cases} p & \text{if } y = 1 \\ 1-p & \text{otherwise,} \end{cases}$

$$\mathrm{CE}(p_t) = -\alpha_t \log(p_t) \longrightarrow \mathrm{FL}(p_t) = -\alpha_t (1-p_t)^{\gamma} \log(p_t)$$

loss量级	量大的类别 (如background)	量少的类别
被正确分类时的loss	大幅↓	稍微↓
被错误分类时的loss	适当↓	近乎保持不变

Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, Piotr Dollár. Focal Loss for Dense Object Detection. ICCV 2017: 2999-3007

RetinaNet

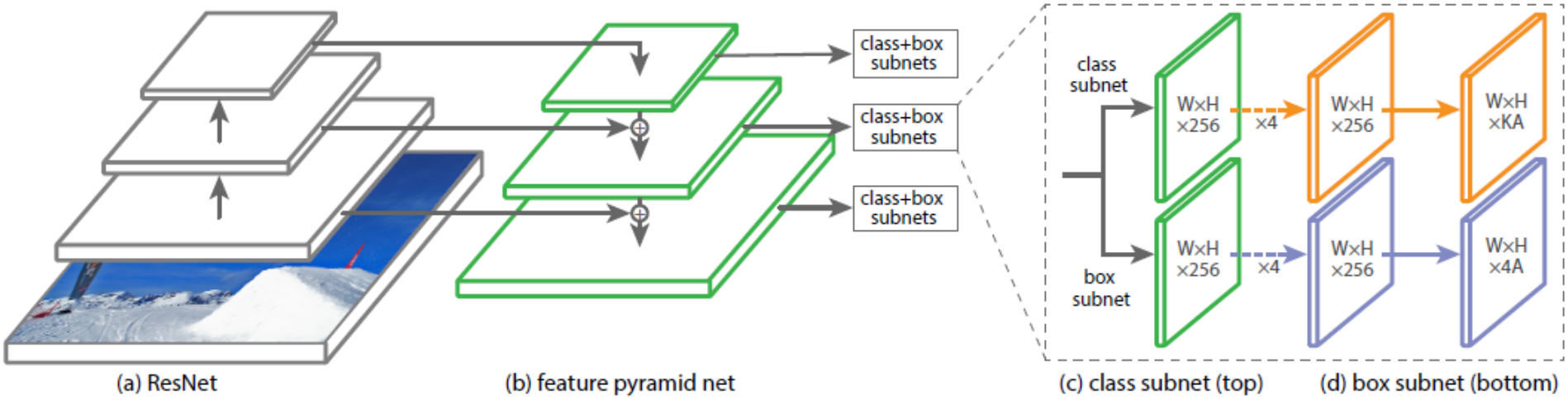
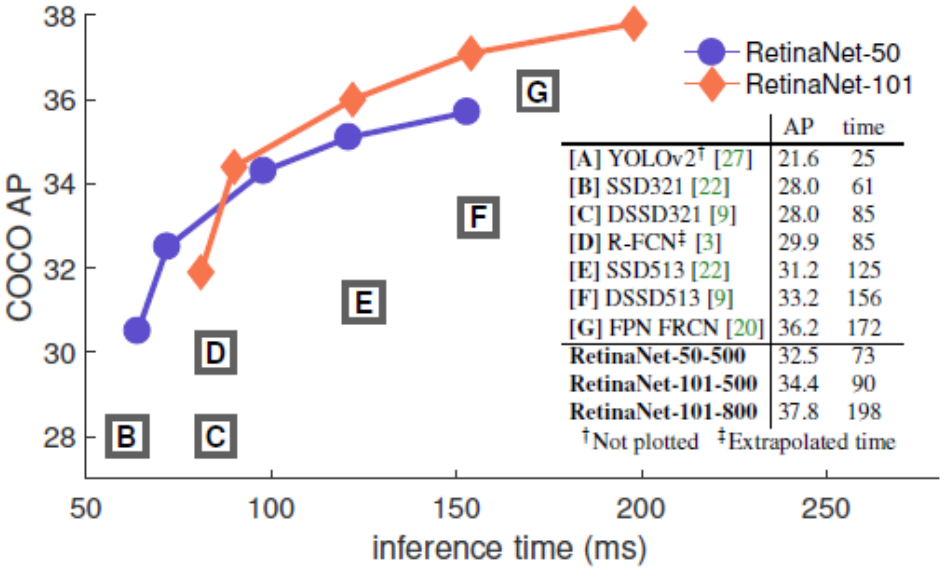


Figure 3. The one-stage **RetinaNet** network architecture uses a Feature Pyramid Network (FPN) [20] backbone on top of a feedforward ResNet architecture [16] (a) to generate a rich, multi-scale convolutional feature pyramid (b). To this backbone RetinaNet attaches two subnetworks, one for classifying anchor boxes (c) and one for regressing from anchor boxes to ground-truth object boxes (d). The network design is intentionally simple, which enables this work to focus on a novel focal loss function that eliminates the accuracy gap between our one-stage detector and state-of-the-art two-stage detectors like Faster R-CNN with FPN [20] while running at faster speeds.

Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, Piotr Dollár. Focal Loss for Dense Object Detection. ICCV 2017: 2999-3007

RetinaNet



	AP	time
[A] YOLOv2† [27]	21.6	25
[B] SSD321 [22]	28.0	61
[C] DSSD321 [9]	28.0	85
[D] R-FCN‡ [3]	29.9	85
[E] SSD513 [22]	31.2	125
[F] DSSD513 [9]	33.2	156
[G] FPN FRCN [20]	36.2	172
RetinaNet-50-500	32.5	73
RetinaNet-101-500	34.4	90
RetinaNet-101-800	37.8	198

†Not plotted ‡Extrapolated time

Figure 2. Speed (ms) versus accuracy (AP) on COCO `test-dev`. Enabled by the focal loss, our simple one-stage *RetinaNet* detector outperforms all previous one-stage and two-stage detectors, including the best reported Faster R-CNN [28] system from [20]. We show variants of RetinaNet with ResNet-50-FPN (blue circles) and ResNet-101-FPN (orange diamonds) at five scales (400-800 pixels). Ignoring the low-accuracy regime (AP<25), RetinaNet forms an upper envelope of all current detectors, and an improved variant (not shown) achieves 40.8 AP. Details are given in §5.

Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, Piotr Dollár. Focal Loss for Dense Object Detection. ICCV 2017: 2999-3007

RetinaNet

	backbone	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Two-stage methods							
Faster R-CNN+++ [16]	ResNet-101-C4	34.9	55.7	37.4	15.6	38.7	50.9
Faster R-CNN w FPN [20]	ResNet-101-FPN	36.2	59.1	39.0	18.2	39.0	48.2
Faster R-CNN by G-RMI [17]	Inception-ResNet-v2 [34]	34.7	55.5	36.7	13.5	38.1	52.0
Faster R-CNN w TDM [32]	Inception-ResNet-v2-TDM	36.8	57.7	39.2	16.2	39.8	**52.1**
One-stage methods							
YOLOv2 [27]	DarkNet-19 [27]	21.6	44.0	19.2	5.0	22.4	35.5
SSD513 [22, 9]	ResNet-101-SSD	31.2	50.4	33.3	10.2	34.5	49.8
DSSD513 [9]	ResNet-101-DSSD	33.2	53.3	35.2	13.0	35.4	51.1
RetinaNet (ours)	ResNet-101-FPN	39.1	59.1	42.3	21.8	42.7	50.2
RetinaNet (ours)	ResNeXt-101-FPN	**40.8**	**61.1**	**44.1**	**24.1**	**44.2**	51.2

Table 2. **Object detection** *single-model* results (bounding box AP), *vs.* state-of-the-art on COCO `test-dev`. We show results for our RetinaNet-101-800 model, trained with scale jitter and for 1.5× longer than the same model from Table 1e. Our model achieves top results, outperforming both one-stage and two-stage models. For a detailed breakdown of speed versus accuracy see Table 1e and Figure 2.

Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, Piotr Dollár. Focal Loss for Dense Object Detection. ICCV 2017: 2999-3007

Anchor-free+RetinaNet

- 提出了基于无anchor机制的特征选择（Feature Selective Anchor-Free，FSAF)模块，它是一个简单高效的One-stage组件，其可以结合特征金字塔嵌入到单阶段检测器中
- FSAF的通用解释是将在线特征选择应用于与anchor无关的分支的训练上。即无anchor的分支添加到特征金字塔的每一层，从而可以以任意层次对box进行编码解码
- 训练过程中，将每个实例动态的放置在最适合的特征层次上。在进行inference时，FSAF可以结合带anchor的分支并行的输出预测结果

Chenchen Zhu, Yihui He, Marios Savvides. Feature Selective Anchor-Free Module for Single-Shot Object Detection. CoRR abs/1903.00621 (2019)

Anchor-free+RetinaNet

- FSAF主要包含无anchor分支的实现及在线特征选择两部分
- FSAF结合RetinaNet在MS COCO数据集上实现更高的准确率及速度，获得了44.6%的mAP，超过了当前存在的单阶段检测网络

Chenchen Zhu, Yihui He, Marios Savvides. Feature Selective Anchor-Free Module for Single-Shot Object Detection. CoRR abs/1903.00621 (2019)

Anchor-free+RetinaNet

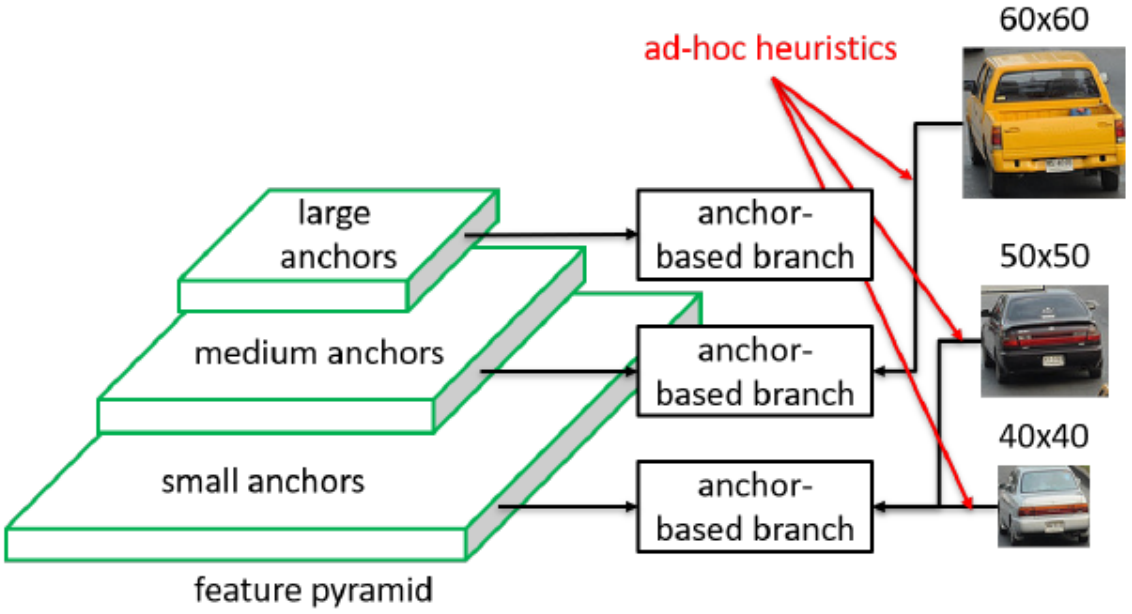


Figure 2: Selected feature level in anchor-based branches may not be optimal.

Chenchen Zhu, Yihui He, Marios Savvides. Feature Selective Anchor-Free Module for Single-Shot Object Detection. CoRR abs/1903.00621 (2019)

Anchor-free+RetinaNet

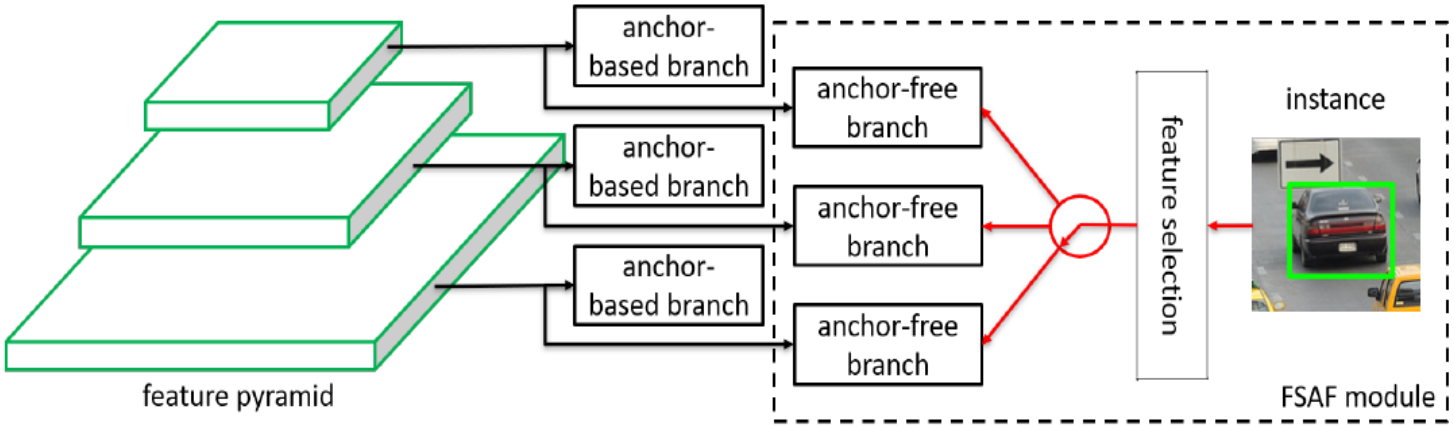


Figure 3: Overview of our FSAF module plugged into conventional anchor-based detection methods. During training, each instance is assigned to a pyramid level via feature selection for setting up supervision signals.

Chenchen Zhu, Yihui He, Marios Savvides. Feature Selective Anchor-Free Module for Single-Shot Object Detection. CoRR abs/1903.00621 (2019)

Anchor-free+RetinaNet

❑ FSAF

- 如何创建anchor-free分支
- 如何产生anchor-free分支的监督信号
- 如何为每个实例动态选择特征
- 如何联合训练anchor分支及anchor-free分支

Chenchen Zhu, Yihui He, Marios Savvides. Feature Selective Anchor-Free Module for Single-Shot Object Detection. CoRR abs/1903.00621 (2019)

Anchor-free+RetinaNet

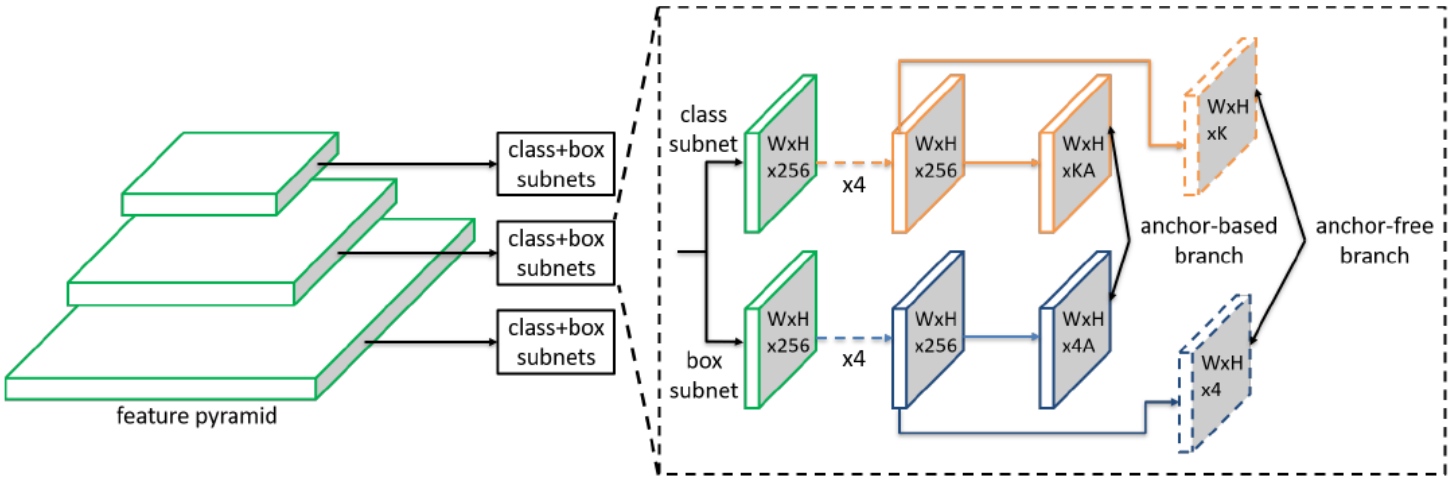


Figure 4: Network architecture of RetinaNet with our FSAF module. The FSAF module only introduces two additional conv layers (dashed feature maps) per pyramid level, keeping the architecture fully convolutional.

Chenchen Zhu, Yihui He, Marios Savvides. Feature Selective Anchor-Free Module for Single-Shot Object Detection. CoRR abs/1903.00621 (2019)

Anchor-free+RetinaNet

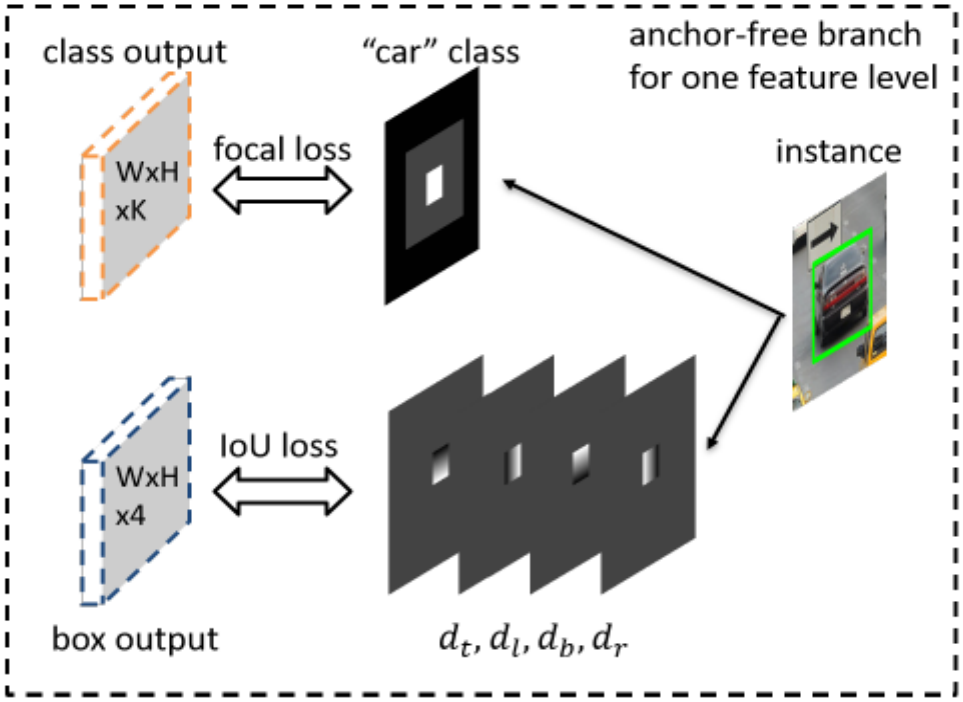


Figure 5: Supervision signals for an instance in one feature level of the anchor-free branches. We use focal loss for classification and IoU loss for box regression.

Chenchen Zhu, Yihui He, Marios Savvides. Feature Selective Anchor-Free Module for Single-Shot Object Detection. CoRR abs/1903.00621 (2019)

Anchor-free+RetinaNet

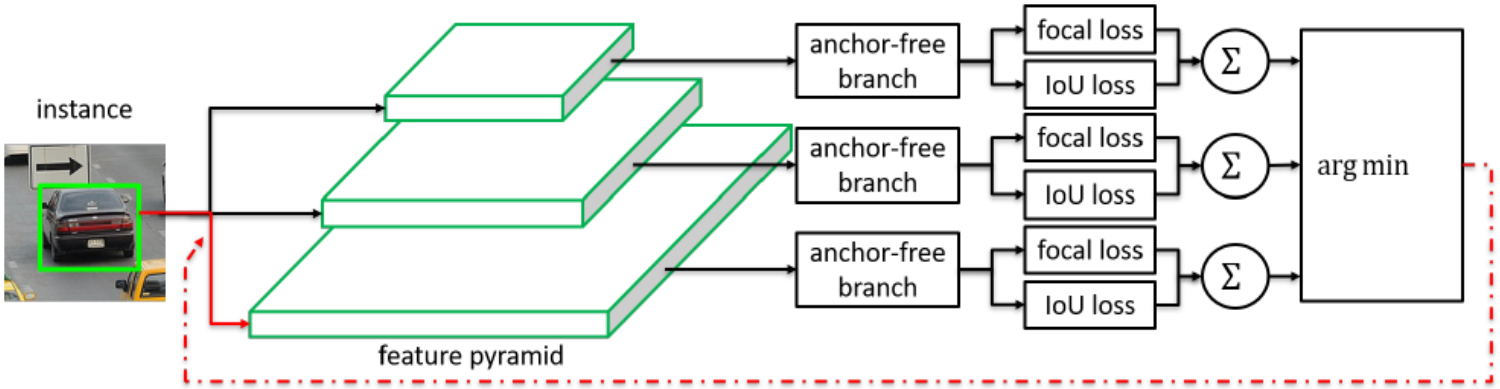


Figure 6: Online feature selection mechanism. Each instance is passing through all levels of anchor-free branches to compute the averaged classification (focal) loss and regression (IoU) loss over effective regions. Then the level with minimal summation of two losses is selected to set up the supervision signals for that instance.

Chenchen Zhu, Yihui He, Marios Savvides. Feature Selective Anchor-Free Module for Single-Shot Object Detection. CoRR abs/1903.00621 (2019)

Anchor-free+RetinaNet

Backbone	Method	AP	AP_{50}	Runtime (ms/im)
R-50	RetinaNet	35.7	54.7	131
	Ours(FSAF)	35.9	55.0	107
	Ours(AB+FSAF)	37.2	57.2	138
R-101	RetinaNet	37.7	57.2	172
	Ours(FSAF)	37.9	58.0	148
	Ours(AB+FSAF)	39.3	59.2	180
X-101	RetinaNet	39.8	59.5	356
	Ours(FSAF)	41.0	61.5	288
	Ours(AB+FSAF)	41.6	62.4	362

Table 2: Detection accuracy and inference latency with different backbone networks on the COCO `minival`. **AB**: Anchor-based branches. **R**: ResNet. **X**: ResNeXt.

Chenchen Zhu, Yihui He, Marios Savvides. Feature Selective Anchor-Free Module for Single-Shot Object Detection. CoRR abs/1903.00621 (2019)

Anchor-free+RetinaNet

Figure 7: More qualitative comparison examples between anchor-based RetinaNet (top, Table 1 1st entry) and our detector with additional FSAF module (bottom, Table 1 5th entry). Both are using ResNet-50 as backbone. Our FSAF module helps finding more challenging objects.

Chenchen Zhu, Yihui He, Marios Savvides. Feature Selective Anchor-Free Module for Single-Shot Object Detection. CoRR abs/1903.00621 (2019)

Anchor-free+RetinaNet

Method	Backbone	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Multi-shot detectors							
CoupleNet [42]	ResNet-101	34.4	54.8	37.2	13.4	38.1	50.8
Faster R-CNN+++ [28]		34.9	55.7	37.4	15.6	38.7	50.9
Faster R-CNN w/ FPN [21]		36.2	59.1	39.0	18.2	39.0	48.2
Regionlets [35]		39.3	59.8	n/a	21.7	43.7	50.9
Fitness NMS [31]		41.8	60.9	44.9	21.5	45.0	57.5
Cascade R-CNN [3]		42.8	62.1	46.3	23.7	45.5	55.2
Deformable R-FCN [4]	Aligned-Inception-ResNet	37.5	58.0	n/a	19.4	40.1	52.5
Soft-NMS [2]		40.9	62.8	n/a	23.3	43.6	53.3
Deformable R-FCN + SNIP [30]	DPN-98	45.7	67.3	51.1	29.3	48.8	57.1
Single-shot detectors							
YOLOv2 [27]	DarkNet-19	21.6	44.0	19.2	5.0	22.4	35.5
SSD513 [24]	ResNet-101	31.2	50.4	33.3	10.2	34.5	49.8
DSSD513 [8]		33.2	53.3	35.2	13.0	35.4	51.1
RefineDet512 [37] (single-scale)		36.4	57.5	39.5	16.6	39.9	51.4
RefineDet [37] (multi-scale)		41.8	62.9	45.7	25.6	45.1	**54.1**
RetinaNet800 [22]		39.1	59.1	42.3	21.8	42.7	50.2
GHM800 [18]		39.9	60.8	42.5	20.3	43.6	**54.1**
Ours800 (single-scale)		40.9	61.5	44.0	24.0	44.2	51.3
Ours (multi-scale)		**42.8**	**63.1**	**46.5**	**27.8**	**45.5**	53.2
CornerNet511 [17] (single-scale)	Hourglass-104	40.5	56.5	43.1	19.4	42.7	53.9
CornerNet [17] (multi-scale)		42.1	57.8	45.3	20.8	44.8	56.7
GHM800 [18]	ResNeXt-101	41.6	62.8	44.2	22.3	45.1	**55.3**
Ours800 (single-scale)		42.9	63.8	46.3	26.6	46.2	52.7
Ours (multi-scale)		**44.6**	**65.2**	**48.6**	**29.7**	**47.1**	54.6

Table 3: Object detection results of our best *single* model with the FSAF module vs. state-of-the-art single-shot and multi-shot detectors on the COCO `test-dev`.

Chenchen Zhu, Yihui He, Marios Savvides. Feature Selective Anchor-Free Module for Single-Shot Object Detection. CoRR abs/1903.00621 (2019)

CornerNet

❑ Anchor-based方法的缺点

- 首先，我们通常需要一组非常大的anchor boxes，例如： 在DSSD中超过4万，在RetinaNet中超过10万
- 其次，anchor boxes的使用引入了许多超参数和设计选择。 这些包括多少个box，大小和宽高比。 这些选择主要是通过ad-hoc启发式方法进行的，并且当与多尺度架构相结合时可能会变得更加复杂

Hei Law, Jia Deng.CornerNet: Detecting Objects as Paired Keypoints. ECCV (14) 2018: 765-781

CornerNet

❑ CornerNet的思路

- 将一个目标物体检测为一对关键点——边界框的左上角和右下角。
- 使用单个卷积网络来预测同一物体类别的所有实例的左上角的热图，所有右下角的热图，以及每个检测到的角点的嵌入向量

Hei Law, Jia Deng.CornerNet: Detecting Objects as Paired Keypoints. ECCV (14) 2018: 765-781

CornerNet

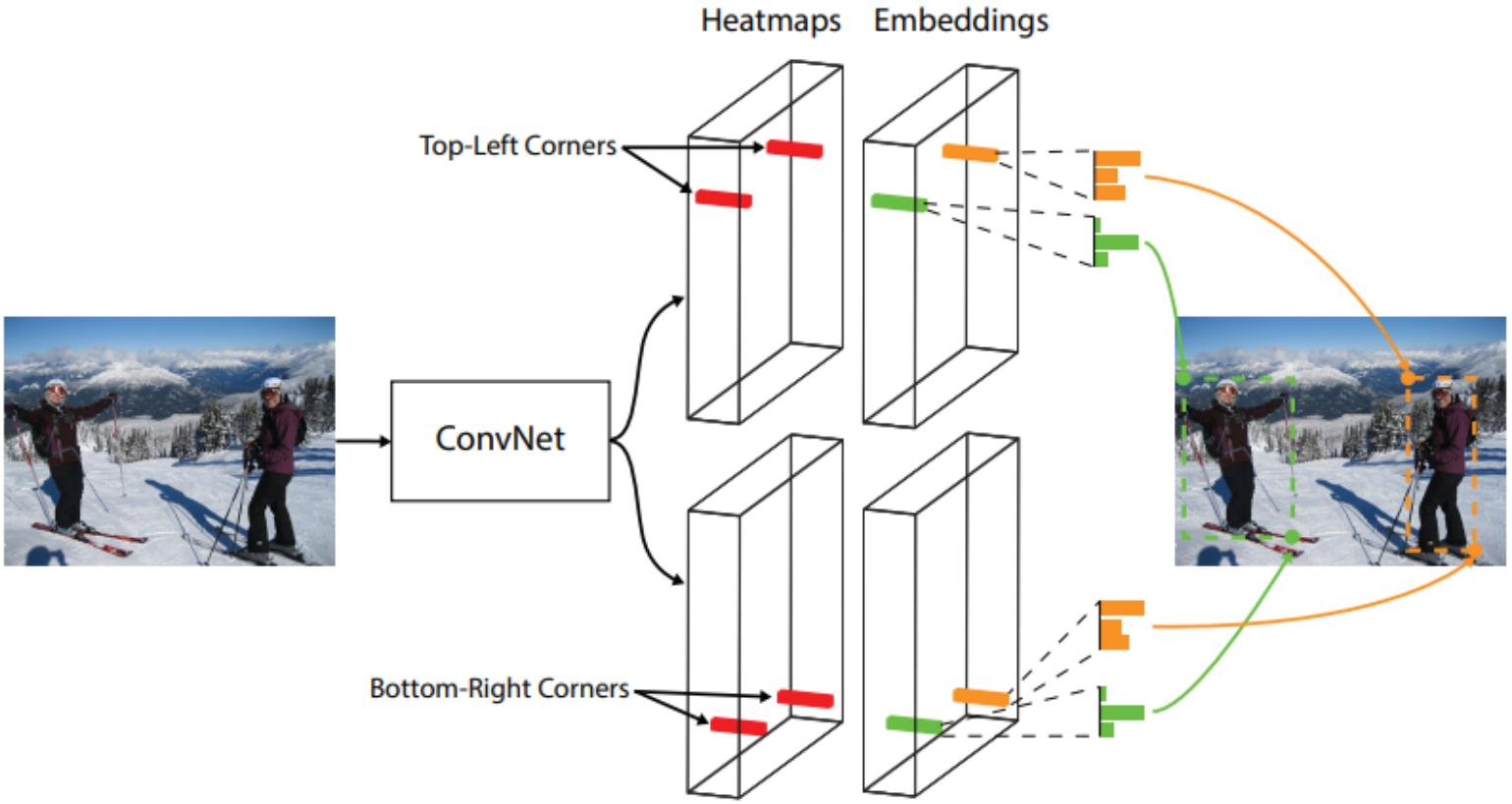


Fig. 1 We detect an object as a pair of bounding box corners grouped together. A convolutional network outputs a heatmap for all top-left corners, a heatmap for all bottom-right corners, and an embedding vector for each detected corner. The network is trained to predict similar embeddings for corners that belong to the same object.

Hei Law, Jia Deng.CornerNet: Detecting Objects as Paired Keypoints. ECCV (14) 2018: 765-781

CornerNet

Fig. 2 Often there is no local evidence to determine the location of a bounding box corner. We address this issue by proposing a new type of pooling layer.

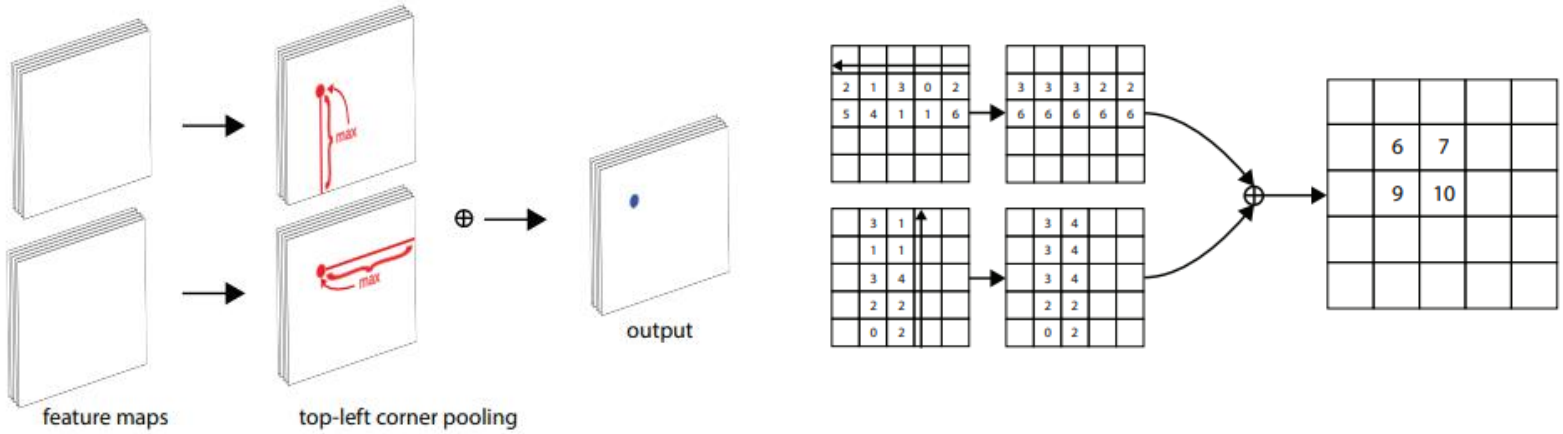


Fig. 3 Corner pooling: for each channel, we take the maximum values *(red dots)* in two directions *(red lines)*, each from a separate feature map, and add the two maximums together *(blue dot)*.

Hei Law, Jia Deng.CornerNet: Detecting Objects as Paired Keypoints. ECCV (14) 2018: 765-781

CornerNet

Table 7 CornerNet versus others on MS COCO test-dev. CornerNet outperforms all one-stage detectors and achieves results competitive to two-stage detectors

Method	Backbone	AP	AP^{50}	AP^{75}	AP^s	AP^m	AP^l	AR^1	AR^{10}	AR^{100}	AR^s	AR^m	AR^l
Two-stage detectors													
DeNet (Tychsen-Smith and Petersson, 2017a)	ResNet-101	33.8	53.4	36.1	12.3	36.1	50.8	29.6	42.6	43.5	19.2	46.9	64.3
CoupleNet (Zhu et al., 2017)	ResNet-101	34.4	54.8	37.2	13.4	38.1	50.8	30.0	45.0	46.4	20.7	53.1	68.5
Faster R-CNN by G-RMI (Huang et al., 2017)	Inception-ResNet-v2 (Szegedy et al., 2017)	34.7	55.5	36.7	13.5	38.1	52.0	-	-	-	-	-	-
Faster R-CNN+++ (He et al., 2016)	ResNet-101	34.9	55.7	37.4	15.6	38.7	50.9	-	-	-	-	-	-
Faster R-CNN w/ FPN (Lin et al., 2016)	ResNet-101	36.2	59.1	39.0	18.2	39.0	48.2	-	-	-	-	-	-
Faster R-CNN w/ TDM (Shrivastava et al., 2016)	Inception-ResNet-v2	36.8	57.7	39.2	16.2	39.8	52.1	31.6	49.3	51.9	28.1	56.6	71.1
D-FCN (Dai et al., 2017)	Aligned-Inception-ResNet	37.5	58.0	-	19.4	40.1	52.5	-	-	-	-	-	-
Regionlets (Xu et al., 2017)	ResNet-101	39.3	59.8	-	21.7	43.7	50.9	-	-	-	-	-	-
Mask R-CNN (He et al., 2017)	ResNeXt-101	39.8	62.3	43.4	22.1	43.2	51.2	-	-	-	-	-	-
Soft-NMS (Bodla et al., 2017)	Aligned-Inception-ResNet	40.9	62.8	-	23.3	43.6	53.3	-	-	-	-	-	-
LH R-CNN (Li et al., 2017)	ResNet-101	41.5	-	-	25.2	45.3	53.1	-	-	-	-	-	-
Fitness-NMS (Tychsen-Smith and Petersson, 2017b)	ResNet-101	41.8	60.9	44.9	21.5	45.0	57.5	-	-	-	-	-	-
Cascade R-CNN (Cai and Vasconcelos, 2017)	ResNet-101	42.8	62.1	46.3	23.7	45.5	55.2	-	-	-	-	-	-
D-RFCN + SNIP (Singh and Davis, 2017)	DPN-98 (Chen et al., 2017)	45.7	67.3	51.1	29.3	48.8	57.1	-	-	-	-	-	-
One-stage detectors													
YOLOv2 (Redmon and Farhadi, 2016)	DarkNet-19	21.6	44.0	19.2	5.0	22.4	35.5	20.7	31.6	33.3	9.8	36.5	54.4
DSOD300 (Shen et al., 2017a)	DS/64-192-48-1	29.3	47.3	30.6	9.4	31.5	47.0	27.3	40.7	43.0	16.7	47.1	65.0
GRP-DSOD320 (Shen et al., 2017b)	DS/64-192-48-1	30.0	47.9	31.8	10.9	33.6	46.3	28.0	42.1	44.5	18.8	49.1	65.0
SSD513 (Liu et al., 2016)	ResNet-101	31.2	50.4	33.3	10.2	34.5	49.8	28.3	42.1	44.4	17.6	49.2	65.8
DSSD513 (Fu et al., 2017)	ResNet-101	33.2	53.3	35.2	13.0	35.4	51.1	28.9	43.5	46.2	21.8	49.1	66.4
RefineDet512 (single scale) (Zhang et al., 2017)	ResNet-101	36.4	57.5	39.5	16.6	39.9	51.4	-	-	-	-	-	-
RetinaNet800 (Lin et al., 2017)	ResNet-101	39.1	59.1	42.3	21.8	42.7	50.2	-	-	-	-	-	-
RefineDet512 (multi scale) (Zhang et al., 2017)	ResNet-101	41.8	62.9	45.7	25.6	45.1	54.1	-	-	-	-	-	-
CornerNet511 (single scale)	Hourglass-104	40.6	56.4	43.2	19.1	42.8	54.3	35.3	54.7	59.4	37.4	62.4	77.2
CornerNet511 (multi scale)	Hourglass-104	42.2	57.8	45.2	20.7	44.8	56.6	36.6	55.9	60.3	39.5	63.2	77.3

Hei Law, Jia Deng.CornerNet: Detecting Objects as Paired Keypoints. ECCV (14) 2018: 765-781

6 图像分割

常见数据集

- PASCAL VOC 2012
- MS COCO
- Automatic Portrait Segmentation for Image Stylization
- SYNTHIA

常见数据集

Name and Reference	Purpose	Year	Classes	Data	Resolution	Sequence	Synthetic/Real	Samples (training)	Samples (validation)	Samples (test)
PASCAL VOC 2012 Segmentation [27]	Generic	2012	21	2D	Variable	✗	R	1464	1449	Private
PASCAL-Context [28]	Generic	2014	540 (59)	2D	Variable	✗	R	10103	N/A	9637
PASCAL-Part [29]	Generic-Part	2014	20	2D	Variable	✗	R	10103	N/A	9637
SBD [30]	Generic	2011	21	2D	Variable	✗	R	8498	2857	N/A
Microsoft COCO [31]	Generic	2014	+80	2D	Variable	✗	R	82783	40504	81434
SYNTHIA [32]	Urban (Driving)	2016	11	2D	960 × 720	✗	S	13407	N/A	N/A
Cityscapes (fine) [33]	Urban	2015	30 (8)	2D	2048 × 1024	✓	R	2975	500	1525
Cityscapes (coarse) [33]	Urban	2015	30 (8)	2D	2048 × 1024	✓	R	22973	500	N/A
CamVid [34]	Urban (Driving)	2009	32	2D	960 × 720	✓	R	701	N/A	N/A
CamVid-Sturgess [35]	Urban (Driving)	2009	11	2D	960 × 720	✓	R	367	100	233
KITTI-Layout [36, 37]	Urban/Driving	2012	3	2D	Variable	✗	R	323	N/A	N/A
KITTI-Ros [38]	Urban/Driving	2015	11	2D	Variable	✗	R	170	N/A	46
KITTI-Zhang [39]	Urban/Driving	2015	10	2D/3D	1226 × 370	✗	R	140	N/A	112
Stanford background [40]	Outdoor	2009	8	2D	320 × 240	✗	R	725	N/A	N/A
SiftFlow [41]	Outdoor	2011	33	2D	256 × 256	✗	R	2688	N/A	N/A
Youtube-Objects-Jain [42]	Objects	2014	10	2D	480 × 360	✓	R	10167	N/A	N/A
Adobe's Portrait Segmentation [26]	Portrait	2016	2	2D	600 × 800	✗	R	1500	300	N/A
MINC [43]	Materials	2015	23	2D	Variable	✗	R	7061	2500	5000
DAVIS [44, 45]	Generic	2016	4	2D	480p	✓	R	4219	2023	2180
NYUDv2 [46]	Indoor	2012	40	2.5D	480 × 640	✗	R	795	654	N/A
SUN3D [47]	Indoor	2013	–	2.5D	640 × 480	✓	R	19640	N/A	N/A
SUNRGBD [48]	Indoor	2015	37	2.5D	Variable	✗	R	2666	2619	5050
RGB-D Object Dataset [49]	Household objects	2011	51	2.5D	640 × 480	✓	R	207920	N/A	N/A
ShapeNet Part [50]	Object/Part	2016	16/50	3D	N/A	✗	S	31,963	N/A	N/A
Stanford 2D-3D-S [51]	Indoor	2017	13	2D/2.5D/3D	1080 × 1080	✓	R	70469	N/A	N/A
3D Mesh [52]	Object/Part	2009	19	3D	N/A	✗	S	380	N/A	N/A
Sydney Urban Objects Dataset [53]	Urban (Objects)	2013	26	3D	N/A	✗	R	41	N/A	N/A
Large-Scale Point Cloud Classification Benchmark [54]	Urban/Nature	2016	8	3D	N/A	✗	R	15	N/A	15

Alberto Garcia-Garcia, Sergio Orts-Escolano, Sergiu Oprea, Victor Villena-Martinez, José García Rodríguez. A Review on Deep Learning Techniques Applied to Semantic Segmentation. CoRR abs/1704.06857 (2017)

Automatic Portrait Segmentation for Image Stylization

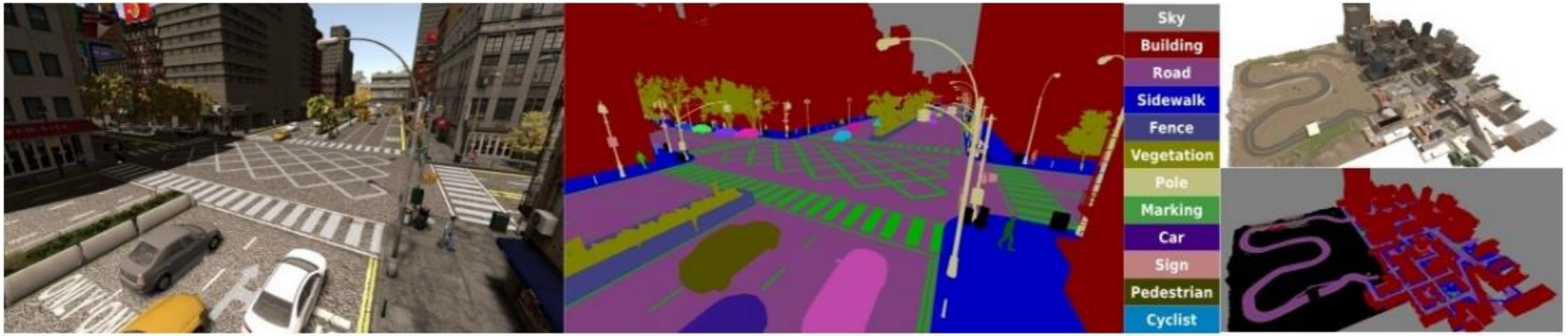
- 人体肖像分割数据库(Automatic Portrait Segmentation for Image Stylization, APSIS)
  - http://xiaoyongshen.me/webpage_portrait/index.html

(a) Input (b) Segmentation (c) Stylization (d) Depth-of-field (e) Cartoon

SYNTHIA

- 计算机合成的城市道路驾驶环境的像素级标注的数据集
- 是为了在自动驾驶或城市场景规划等研究领域中的场景理解而提出的
- 提供了13个类别物体（分别为混合的、天空、建筑、道路、人行道、栅栏、植被、杆、车、信号标志、行人、骑自行车的人、车道）细粒度的像素级别的标注
- http://synthia-dataset.net

SYNTHIA

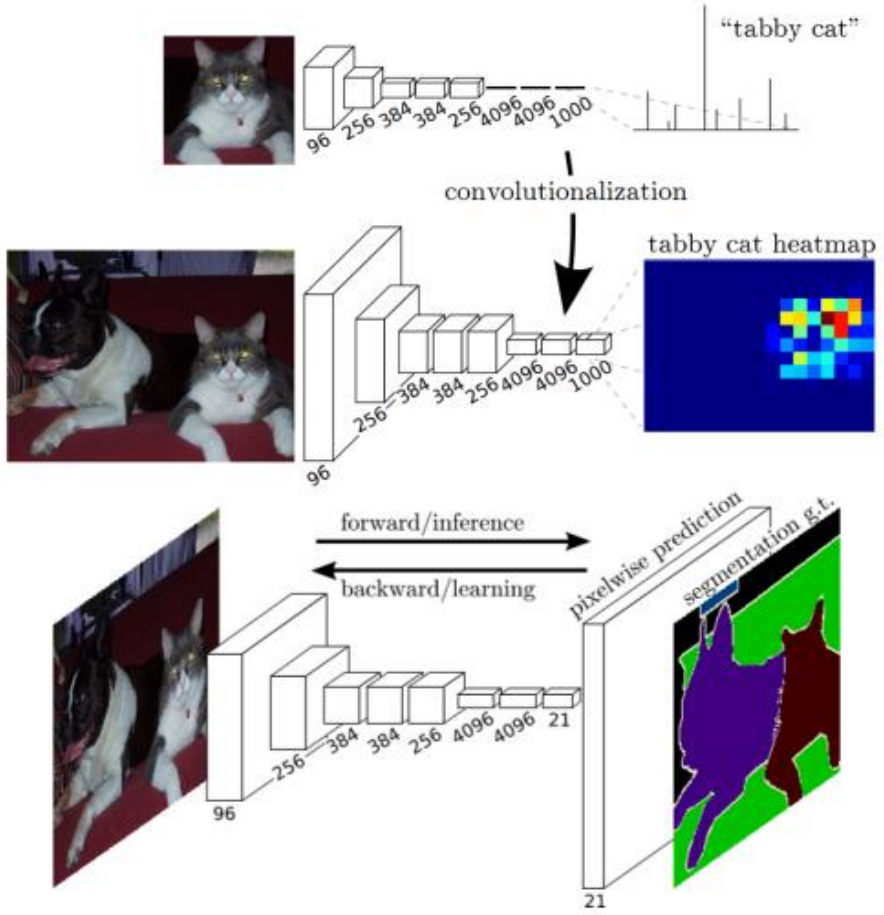

Sky
Building
Road
Sidewalk
Fence
Vegetation
Pole
Marking
Car
Sign
Pedestrian
Cyclist

图像分割方法概貌

		Targets								
Name and Reference	Architecture	Accuracy	Efficiency	Training	Instance	Sequences	Multi-modal	3D	Source Code	Contribution(s)
Fully Convolutional Network [65]	VGG-16(FCN)	⋆	⋆	⋆	✗	✗	✗	✗	✓	Forerunner
SegNet [66]	VGG-16 + Decoder	⋆⋆⋆	⋆⋆	⋆	✗	✗	✗	✗	✓	Encoder-decoder
Bayesian SegNet [67]	SegNet	⋆⋆⋆	⋆	⋆	✗	✗	✗	✗	✓	Uncertainty modeling
DeepLab [68] [69]	VGG-16/ResNet-101	⋆⋆⋆	⋆	⋆	✗	✗	✗	✗	✓	Standalone CRF, atrous convolutions
MINC-CNN [43]	GoogLeNet(FCN)	⋆	⋆	⋆	✗	✗	✗	✗	✓	Patchwise CNN, Standalone CRF
CRFasRNN [70]	FCN-8s	⋆	⋆⋆	⋆⋆⋆	✗	✗	✗	✗	✓	CRF reformulated as RNN
Dilation [71]	VGG-16	⋆⋆⋆	⋆	⋆	✗	✗	✗	✗	✓	Dilated convolutions
ENet [72]	ENet bottleneck	⋆⋆	⋆⋆⋆	⋆	✗	✗	✗	✗	✓	Bottleneck module for efficiency
Multi-scale-CNN-Raj [73]	VGG-16(FCN)	⋆⋆⋆	⋆	⋆	✗	✗	✗	✗	✗	Multi-scale architecture
Multi-scale-CNN-Eigen [74]	Custom	⋆⋆⋆	⋆	⋆	✗	✗	✗	✗	✓	Multi-scale sequential refinement
Multi-scale-CNN-Roy [75]	Multi-scale-CNN-Eigen	⋆⋆⋆	⋆	⋆	✗	✗	⋆⋆	✗	✗	Multi-scale coarse-to-fine refinement
Multi-scale-CNN-Bian [76]	FCN	⋆⋆	⋆	⋆⋆	✗	✗	✗	✗	✗	Independently trained multi-scale FCNs
ParseNet [77]	VGG-16	⋆⋆⋆	⋆	⋆	✗	✗	✗	✗	✓	Global context feature fusion
ReSeg [78]	VGG-16 + ReNet	⋆⋆	⋆	⋆	✗	✗	✗	✗	✓	Extension of ReNet to semantic segmentation
LSTM-CF [79]	Fast R-CNN + DeepMask	⋆⋆⋆	⋆	⋆	✗	✗	✗	✗	✓	Fusion of contextual information from multiple sources
2D-LSTM [80]	MDRNN	⋆⋆	⋆⋆	⋆	✗	✗	✗	✗	✗	Image context modelling
rCNN [81]	MDRNN	⋆⋆⋆	⋆⋆	⋆	✗	✗	✗	✗	✓	Different input sizes, image context
DAG-RNN [82]	Elman network	⋆⋆⋆	⋆	⋆	✗	✗	✗	✗	✓	Graph image structure for context modelling
SDS [10]	R-CNN + Box CNN	⋆⋆⋆	⋆	⋆	⋆⋆	✗	✗	✗	✓	Simultaneous detection and segmentation
DeepMask [83]	VGG-A	⋆⋆⋆	⋆	⋆	⋆⋆	✗	✗	✗	✓	Proposals generation for segmentation
SharpMask [84]	DeepMask	⋆⋆⋆	⋆	⋆	⋆⋆⋆	✗	✗	✗	✓	Top-down refinement module
MultiPathNet [85]	Fast R-CNN + DeepMask	⋆⋆⋆	⋆	⋆	⋆⋆⋆	✗	✗	✗	✓	Multi path information flow through network
Huang-3DCNN [86]	Own 3DCNN	⋆	⋆	⋆	✗	✗	✗	⋆⋆⋆	✗	3DCNN for voxelized point clouds
PointNet [87]	Own MLP-based	⋆⋆	⋆	⋆	✗	✗	✗	⋆⋆⋆	✓	Segmentation of unordered point sets
Clockwork Convnet [88]	FCN	⋆⋆	⋆⋆	⋆	✗	⋆⋆⋆	✗	✗	✓	Clockwork scheduling for sequences
3DCNN-Zhang	Own 3DCNN	⋆⋆	⋆	⋆	✗	⋆⋆⋆	✗	✗	✓	3D convolutions and graph cut for sequences
End2End Vox2Vox [89]	C3D	⋆⋆	⋆	⋆	✗	⋆⋆⋆	✗	✗	✗	3D convolutions/deconvolutions for sequences

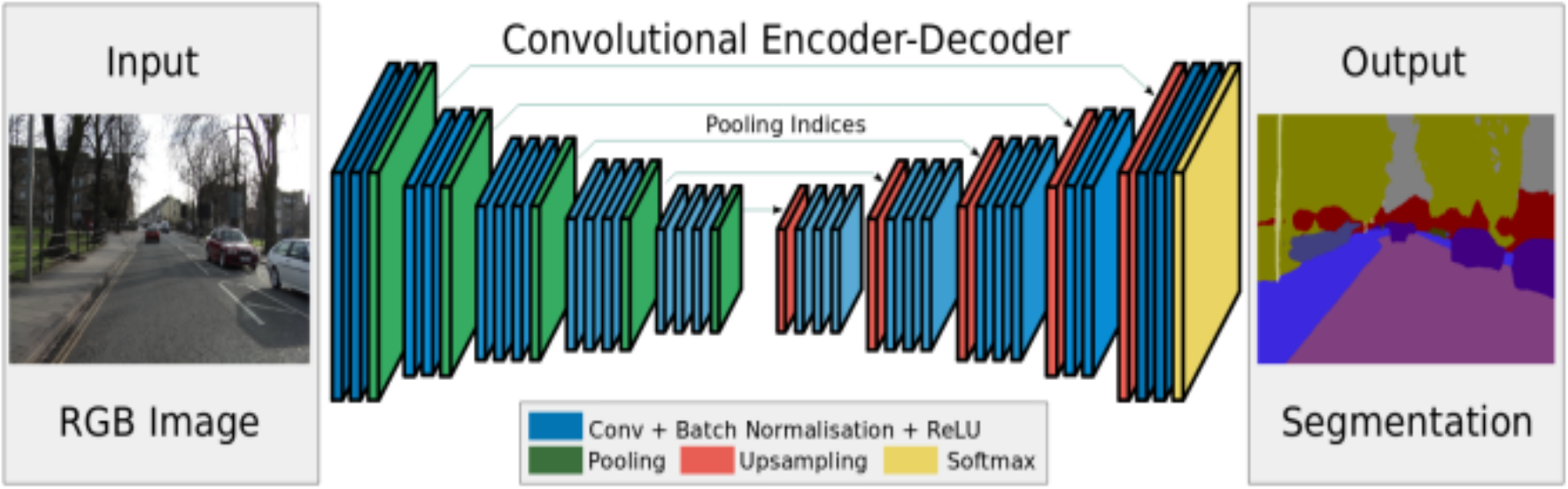
Alberto Garcia-Garcia, Sergio Orts-Escolano, Sergiu Oprea, Victor Villena-Martinez, José García Rodríguez. A Review on Deep Learning Techniques Applied to Semantic Segmentation. CoRR abs/1704.06857 (2017)

图像分割方法FCN



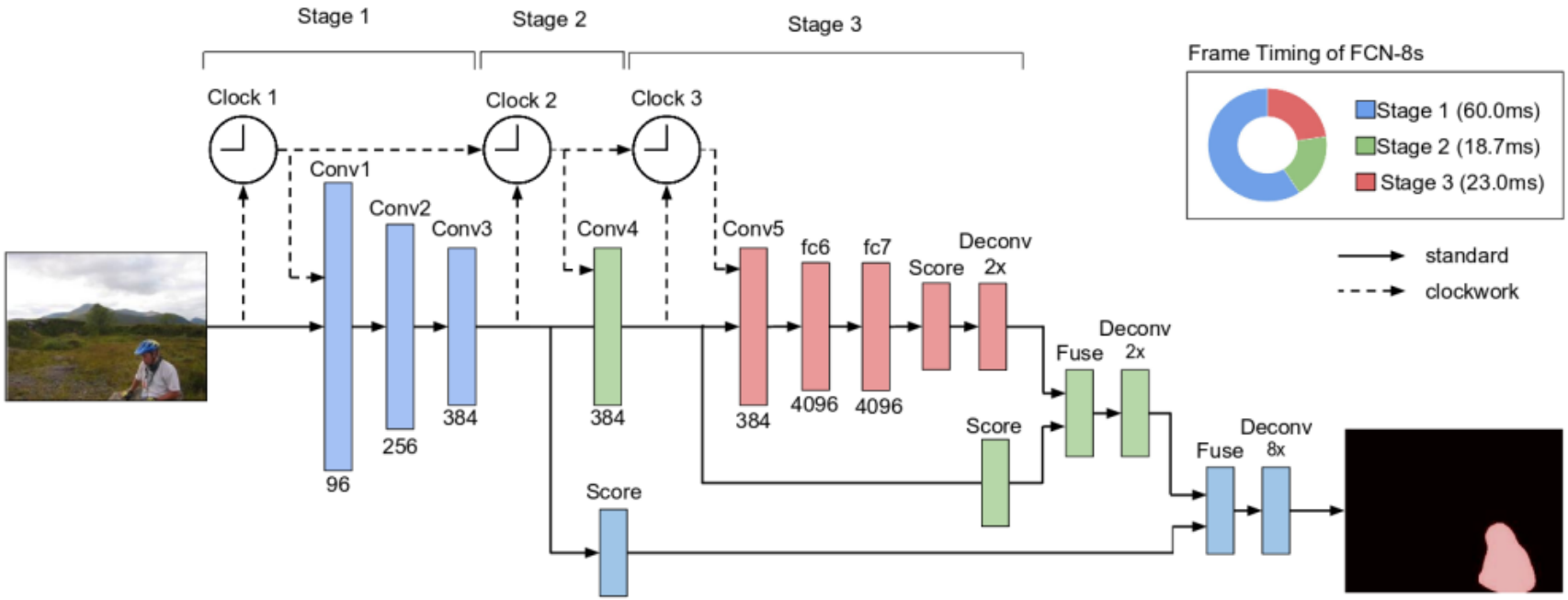
J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp.3431–3440.

图像分割方法SegNet



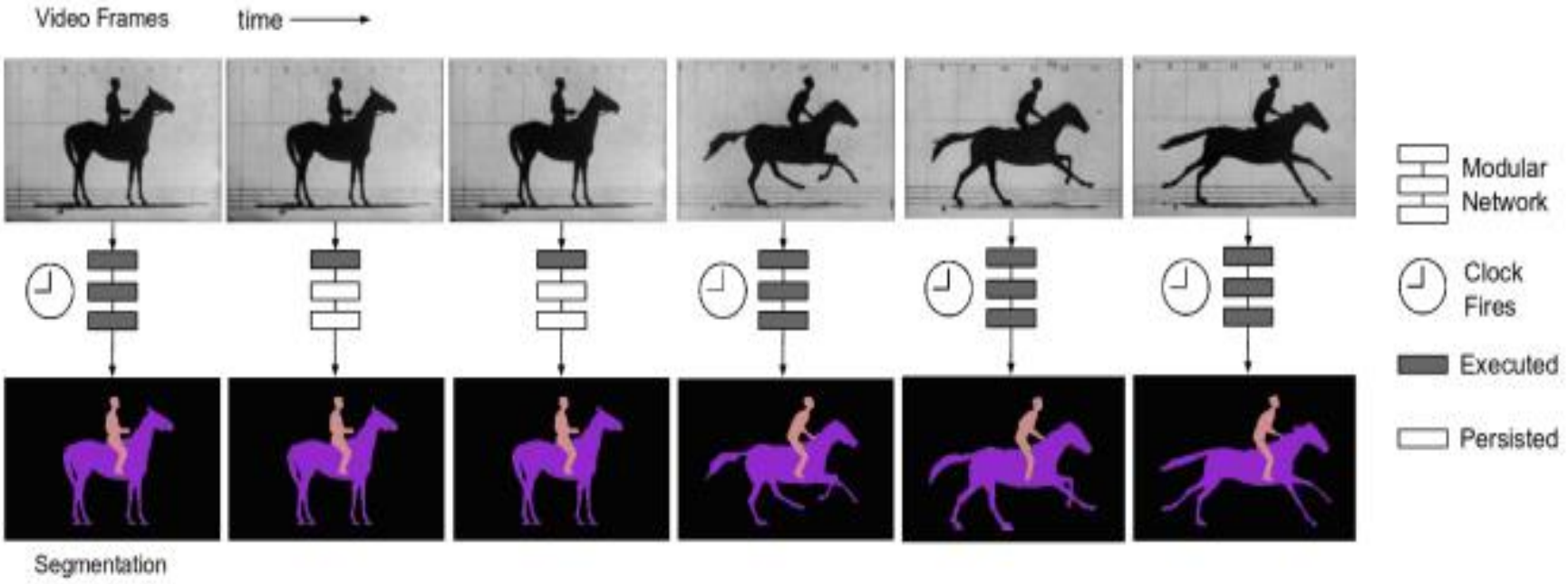
J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp.3431–3440.

图像分割方法ClockFCN



E. Shelhamer, K. Rakelly, J. Hoffman, and T. Darrell. Clockwork convnets for video semantic segmentation, in Computer Vision–ECCV 2016 Workshops. Springer, 2016, pp. 852–868.

图像分割方法ClockFCN



E. Shelhamer, K. Rakelly, J. Hoffman, and T. Darrell. Clockwork convnets for video semantic segmentation, in Computer Vision–ECCV 2016 Workshops. Springer, 2016, pp. 852–868.

7

图像回归

人体姿态数据集

- **LSP：**
  - 地址：http://sam.johnson.io/research/lsp.html
  - 样本数：2K
  - 关节点个数：14
  - 全身，单人
- **FLIC**
  - 地址：https://bensapp.github.io/flic-dataset.html
  - 样本数：2W
  - 关节点个数：9
  - 全身，单人

人体姿态数据集

❑ MPII

- 地址：http://human-pose.mpi-inf.mpg.de/
- 样本数：25K
- 关节点个数：16
- 全身，单人/多人，40K people，410 human activities

❑ MSCOCO

- 地址：http://cocodataset.org/#download
- 样本数：>= 30W
- 关节点个数：18
- 全身，多人，keypoints on 10W people

人体姿态数据集

❑ AI Challenge

- 地址：https://challenger.ai/competition/keypoint/subject
- 样本数：21W Training, 3W Validation, 3W Testing
- 关节点个数：14
- 全身，多人，38W people

人体姿态估计

Figure 1. Besides extreme variability in articulations, many of the joints are barely visible. We can guess the location of the right arm in the left image only because we see the rest of the pose and anticipate the motion or activity of the person. Similarly, the left body half of the person on the right is not visible at all. These are examples of the need for *holistic reasoning*. We believe that DNNs can naturally provide such type of reasoning.

Alexander Toshev, Christian Szegedy. DeepPose: Human Pose Estimation via Deep Neural Networks. CVPR 2014: 1653-1660

人体姿态估计

Figure 2. Left: schematic view of the DNN-based pose regression. We visualize the network layers with their corresponding dimensions, where convolutional layers are in blue, while fully connected ones are in green. We do not show the parameter free layers. Right: at stage s, a refining regressor is applied on a sub image to refine a prediction from the previous stage.

Alexander Toshev, Christian Szegedy. DeepPose: Human Pose Estimation via Deep Neural Networks. CVPR 2014: 1653-1660

人脸数据集

数据库	描述	用途	获取方法
WebFace	10k+人，约500K张图片	非限制场景	链接
FaceScrub	530人，约100k张图片	非限制场景	链接
YouTube Face	1,595个人 3,425段视频	非限制场景、视频	链接
LFW	5k+人脸，超过10K张图片	标准的人脸识别数据集	链接
MultiPIE	337个人的不同姿态、表情、光照的人脸图像，共750k+人脸图像	限制场景人脸识别	链接 需购买
MegaFace	690k不同的人的1000k人脸图像	新的人脸识别评测集合	链接
IJB-A		人脸识别，人脸检测	链接
CAS-PEAL	1040个人的30k+张人脸图像，主要包含姿态、表情、光照变化	限制场景下人脸识别	链接
Pubfig	200个人的58k+人脸图像	非限制场景下的人脸识别	链接

https://blog.csdn.net/chenriwei2/article/details/50631212

人脸检测

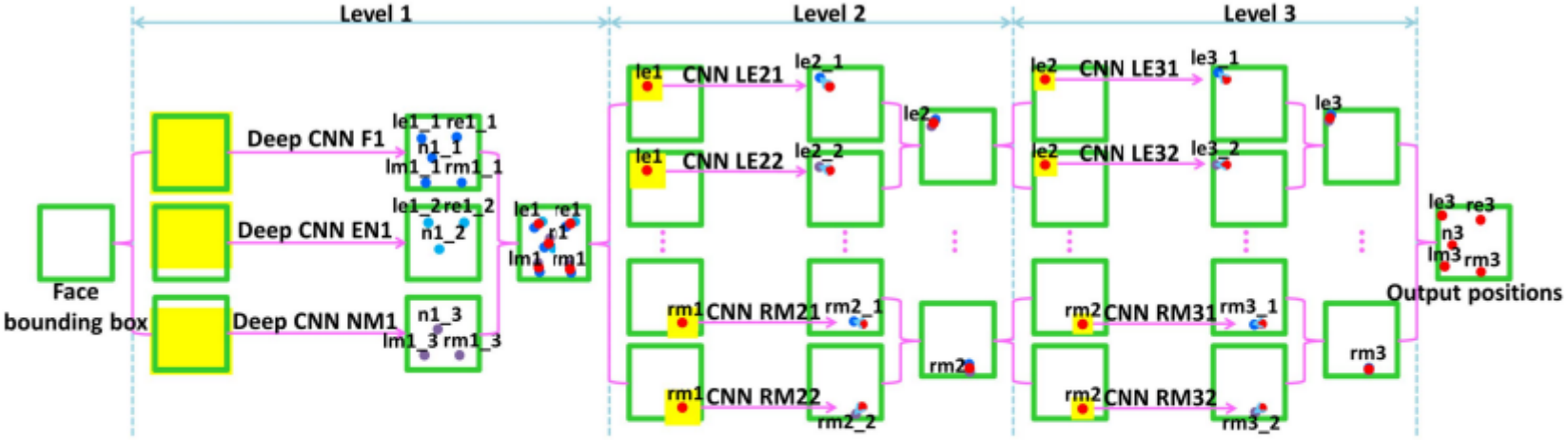


Figure 2: Three-level cascaded convolutional networks. The input is the face region returned by a face detector. The three networks at level 1 are denoted as F1, EN1, and NM1. Networks at level 2 are denoted as LE21, LE22, RE21, RE22, N21, N22, LM21, LM22, RM21, and RM22. Both LE21 and LE22 predict the left eye center, and so forth. Networks at level 3 are denoted as LE31, LE32, RE31, RE32, N31, N32, LM31, LM32, RM31, and RM32. Green square is the face bounding box given by the face detector. Yellow shaded areas are the input regions of networks. Red dots are the final predictions at each level. Dots in other colors are predictions given by individual networks.

Yi Sun, Xiaogang Wang, Xiaoou Tang: Deep Convolutional Network Cascade for Facial Point Detection. CVPR 2013: 3476-3483

人脸检测

Figure 1: Examples of facial point detection. First row: initial detection with our first level of convolutional networks. It achieves good estimation with global context information even if some facial components are invisible or ambiguous in appearance. Second row: finely tuned results with our second and third levels of networks. The accuracy is improved. Third row: result from [5]. It is more restricted with shape templates learned from the training set and not accurate under some unusual poses and expressions.

Yi Sun, Xiaogang Wang, Xiaoou Tang: Deep Convolutional Network Cascade for Facial Point Detection. CVPR 2013: 3476-3483

8 中英文术语对照

中英文术语对照

- 计算机视觉：Computer Vision
- 语义鸿沟：Sematic gap
- 中心化： Center
- 归一化：Normalize
- 去相关：decorrelate
- 白化：Whiten
- 局部敏感哈希：Locality Sensitive Hashing
- 视觉词袋：Bag-of-Visual Words
- 特征量化：Feature Quantization
- 词频-逆文本频率：Term frequency - Inverse Document Frequency
- 局部几何验证：Local Geometric Verification
- 乘积量化：Product Quantization

谢谢！