

# Deep Learning Solutions for Visual World Understanding

Jiashi FENG

Learning and Vision Group

NUS

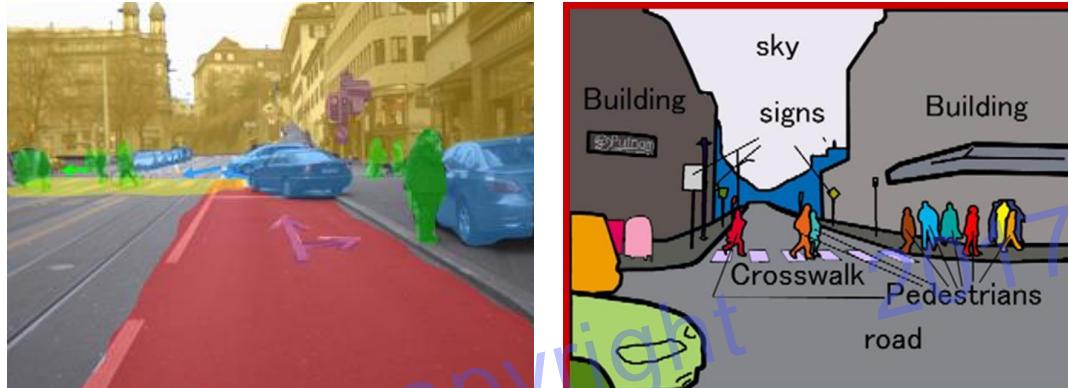
copyright 2017 by MEDIVIEWS™

# Intelligent Image/video Processing

- Image/video processing



- Scene Understanding



copyright  
MEDEVIEW

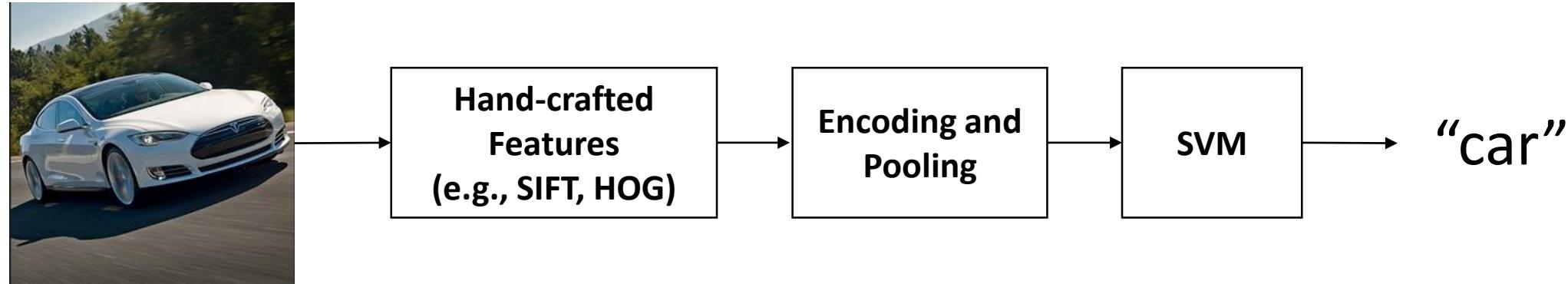
- Face analysis



- Human analysis

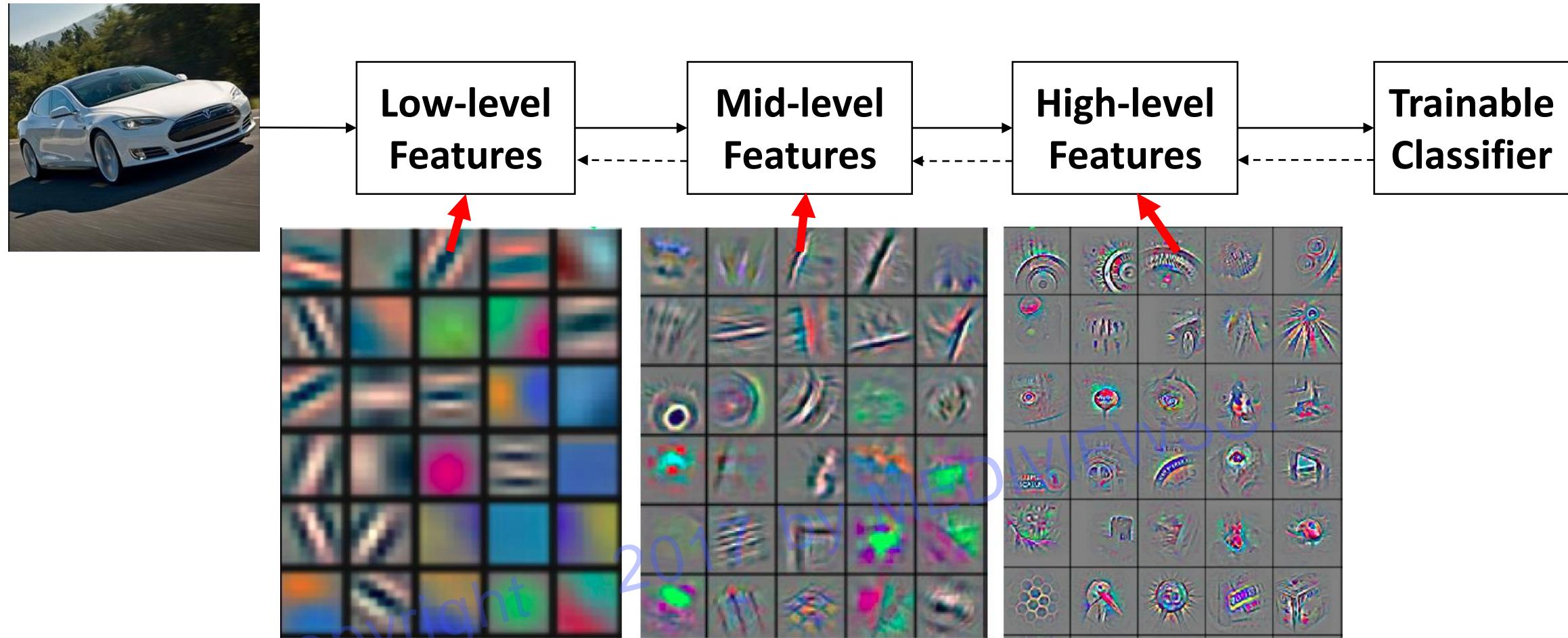


# Traditional Approaches



- Heavily relies on expert knowledge
- Same level of abstraction for simple and complicated images.

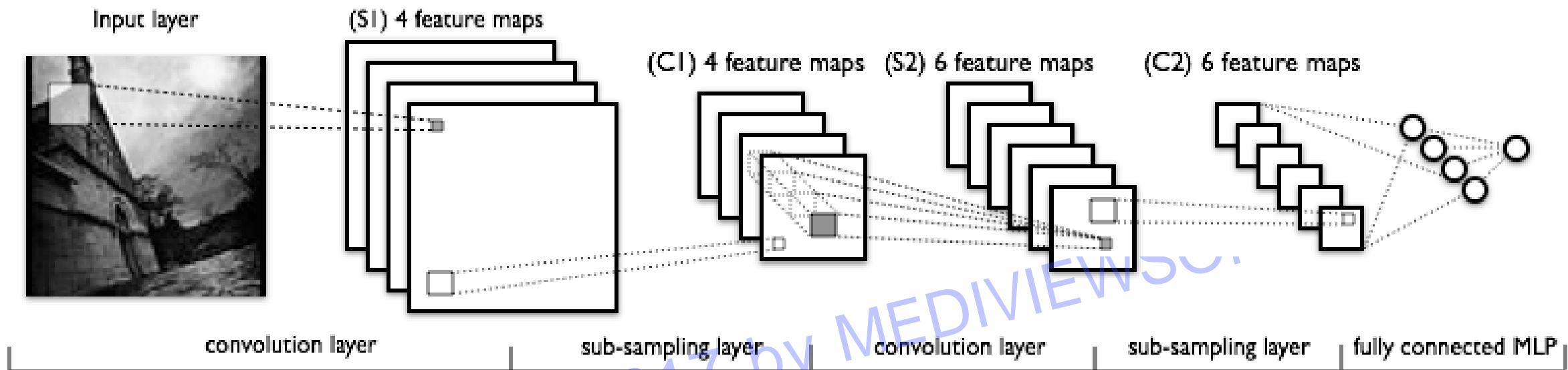
# End-to-end Hierarchical Representation Learning Approaches



Feature visualization of convolutional net trained on ImageNet from [Zeiler & Fergus 2013]

# Convolutional Neural Networks

- Each filter is shared over the image plane (convolution)
- Each output sees a large input context (pooling)



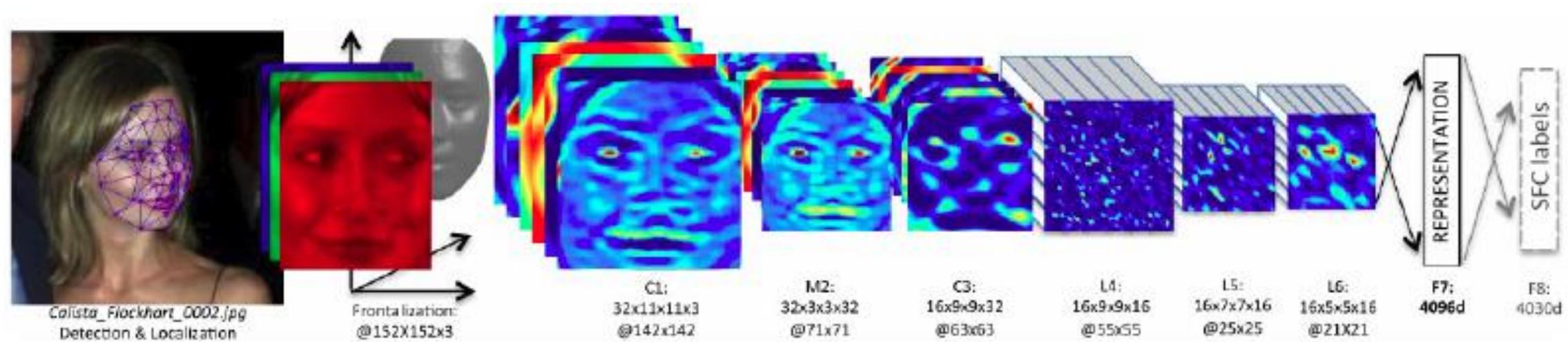
copyright

2017 by MEDIVIEWSU.

[LeCun et al. NIPS 1989]

# ConvNets for Visual Understanding

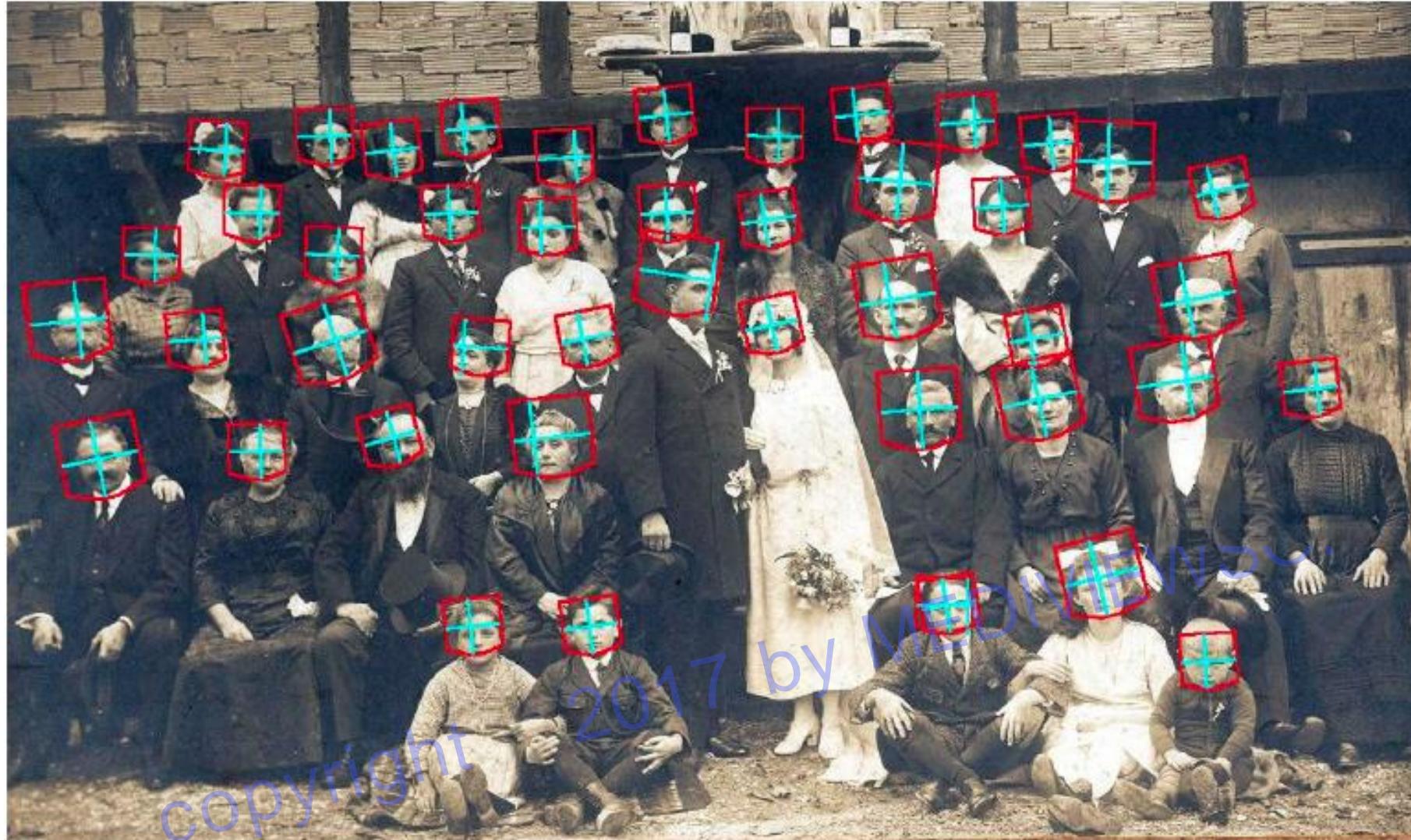
- Face recognition



- Deep face
- Alignment + ConvNet feature learning + metric learning
- Deployed at Facebook for auto tagging

# ConvNets for Visual Understanding

- Face detection & pose estimation



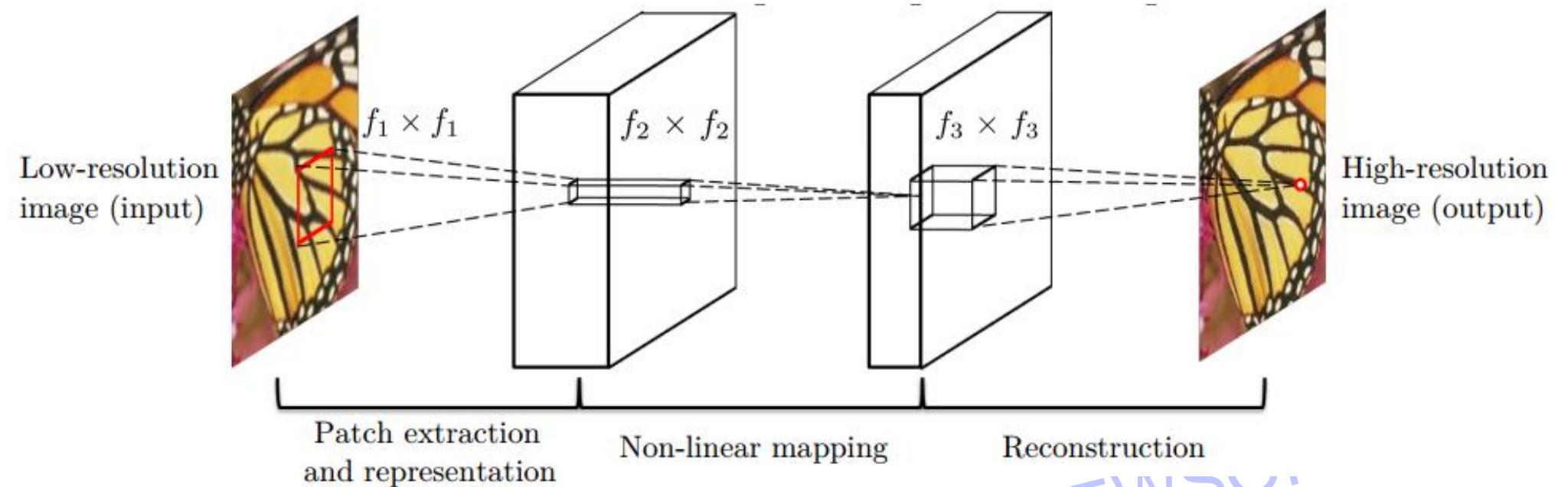
# ConvNets for Visual Understanding

## - Scene parsing



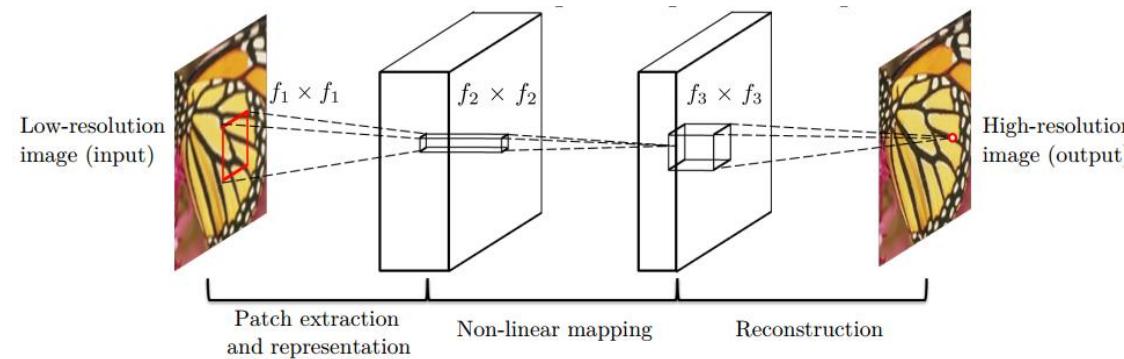
- No post-processing
- Frame by frame
- ConvNet runs at 50 ms/frame on Virtex-6 FPGA hardware

# Application I – Deep Image Super-resolution

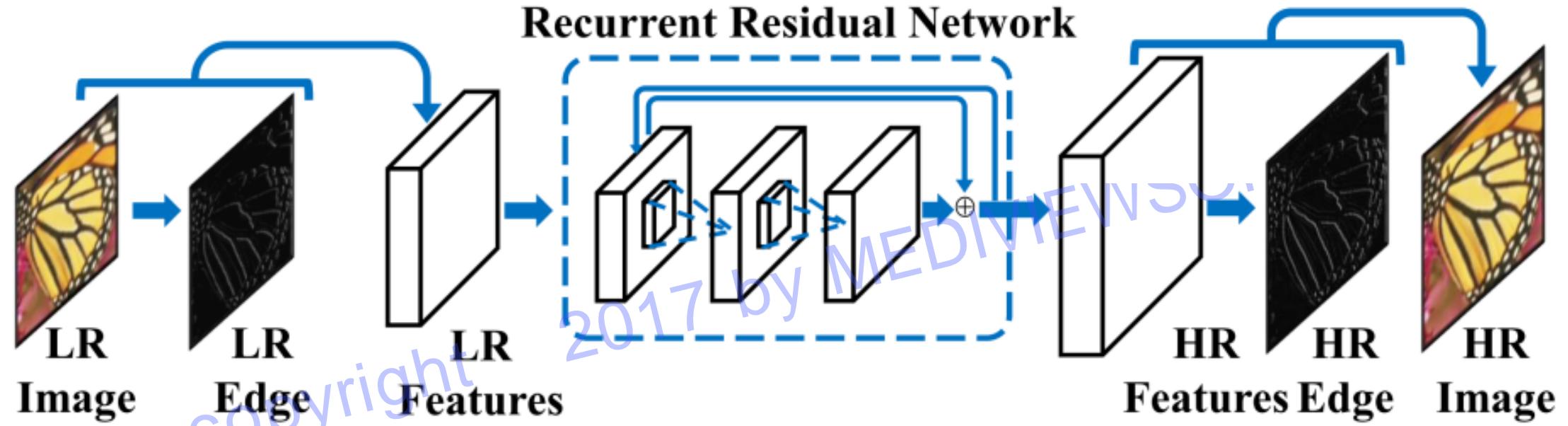


copyright 2017 by MEDIVIEWS™

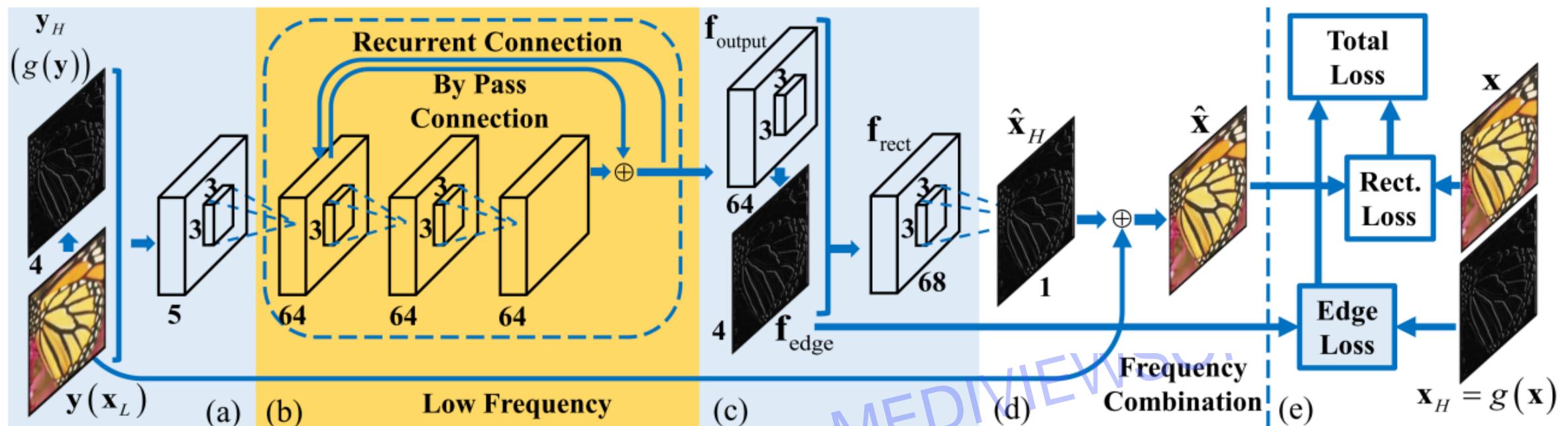
# Application I – Deep Image Super-resolution



- Exploit the edge-prior
- Recurrently super-resolve images

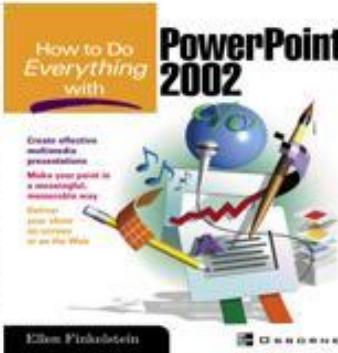
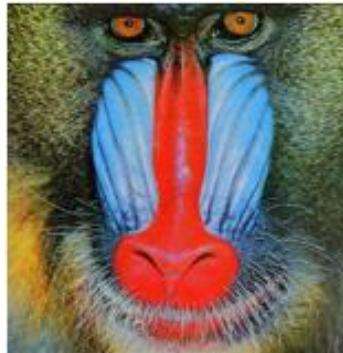


# Recurrent Residue ConvNet for Image SR



copyright 2017 by MEDIVIEW

# Deep Image SR Results



2017 by MEDIVIEWS

copyright

# Deep Image SR Results



(a) High-res



(b) A+



(c) SRCNN



(d) TSE-SR



(e) CSCN

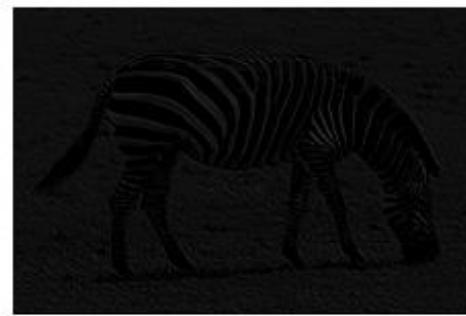


(f) DEGREE

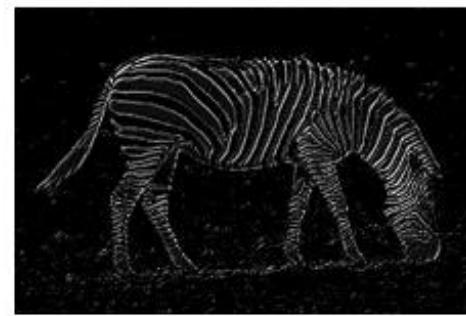
# Learned Features at Different Layers



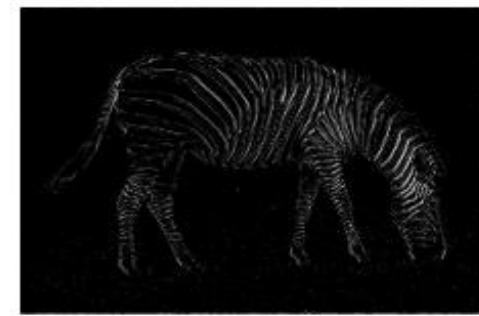
(a) LR



(b) 1L



(c) 1R



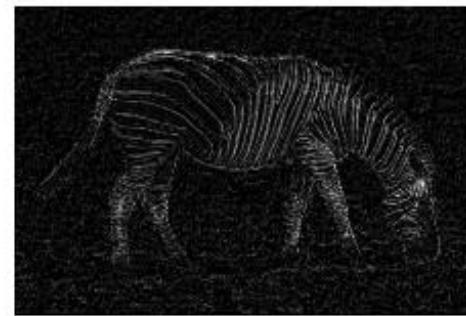
(d) 2R



(e) 3R



(f) 4R



(g) 5R



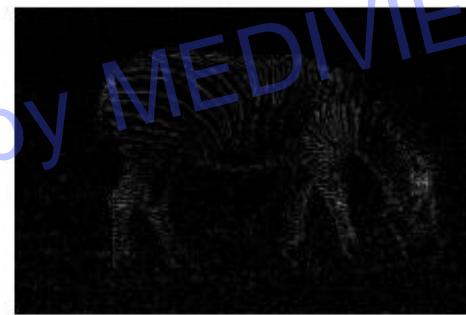
(h) 6R



(i) 7R



(j) 8R



(k) 9R



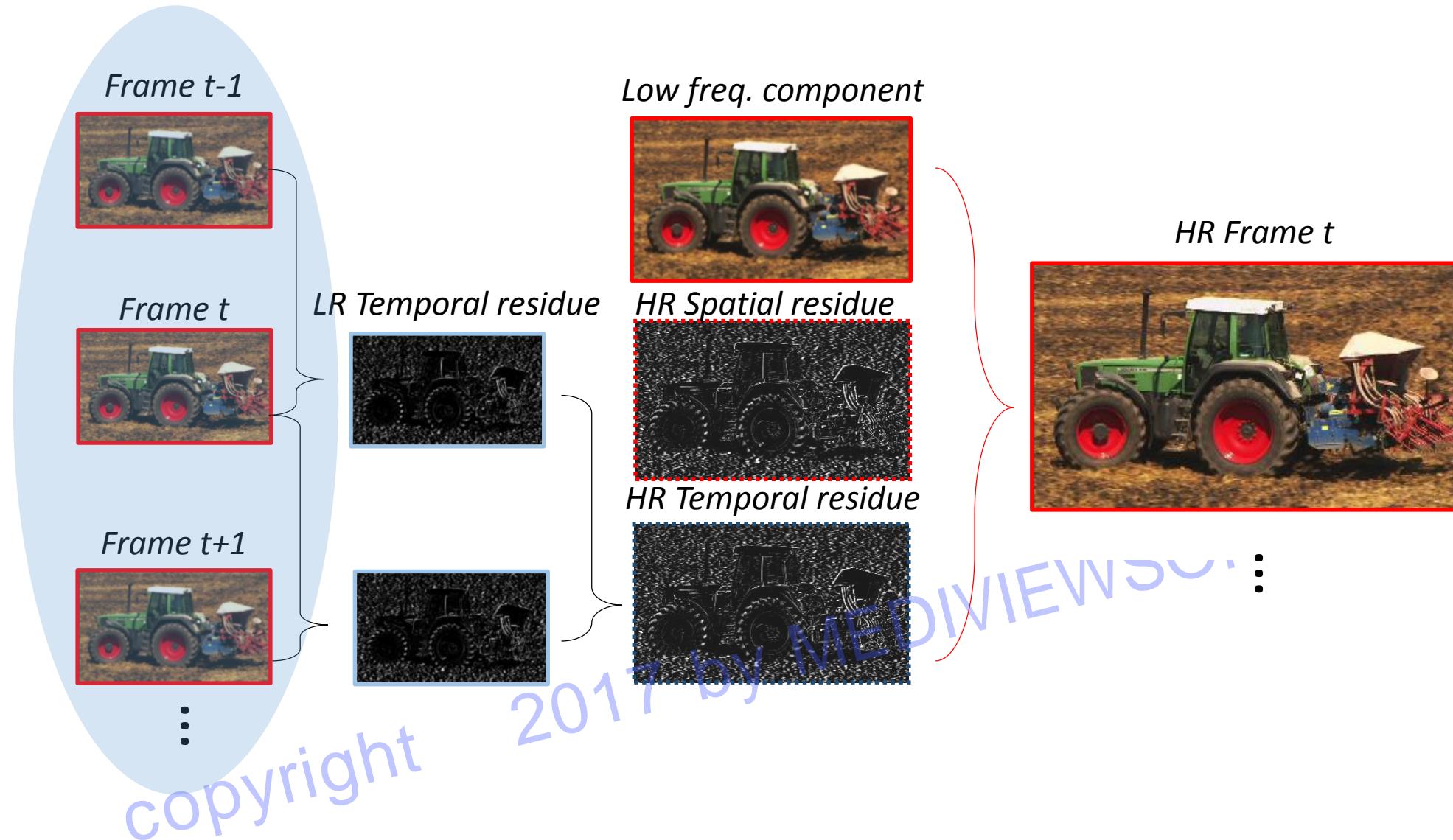
(l) Result

# Deep Image SR Results

Dataset		Set5			Set14			BSD100		
Method	Metric	$\times 2$	$\times 3$	$\times 4$	$\times 2$	$\times 3$	$\times 4$	$\times 2$	$\times 3$	$\times 4$
Bicubic	PSNR	33.66	30.39	28.42	30.13	27.47	25.95	29.55	27.20	25.96
	SSIM	0.9096	0.8682	0.8105	0.8665	0.7722	0.7011	0.8425	0.7382	0.6672
ScSR	PSNR	35.78	31.34	29.07	31.64	28.19	26.40	30.77	27.72	26.61
	SSIM	0.9485	0.8869	0.8263	0.8990	0.7977	0.7218	0.8744	0.7647	0.6983
A+	PSNR	36.56	32.60	30.30	32.14	29.07	27.28	30.78	28.18	26.77
	SSIM	0.9544	0.9088	0.8604	0.9025	0.8171	0.7484	0.8773	0.7808	0.7085
TSE-SR	PSNR	36.47	32.62	30.24	32.21	29.14	27.38	31.18	28.30	26.85
	SSIM	0.9535	0.9092	0.8609	0.9033	0.8194	0.7514	0.8855	0.7843	0.7108
JSB-NE	PSNR	36.59	32.32	30.08	32.34	28.98	27.22	31.22	28.14	26.71
	SSIM	0.9538	0.9042	0.8508	0.9058	0.8105	0.7393	0.8869	0.7742	0.6978
CNN	PSNR	36.34	32.39	30.09	32.18	29.00	27.20	31.11	28.20	26.70
	SSIM	0.9521	0.9033	0.8530	0.9039	0.8145	0.7413	0.8835	0.7794	0.7018
CNN-L	PSNR	36.66	32.75	30.49	32.45	29.30	27.50	31.36	28.41	26.90
	SSIM	0.9542	0.9090	0.8628	0.9067	0.8215	0.7513	0.8879	0.7863	0.7103
CSCN	PSNR	36.88	33.10	30.86	32.50	29.42	27.64	31.40	28.50	27.03
	SSIM	0.9547	0.9144	0.8732	0.9069	0.8238	0.7573	0.8884	0.7885	0.7161
CSCN-MV	PSNR	37.14	33.26	31.04	32.71	29.55	27.76	31.54	28.58	27.11
	SSIM	0.9567	0.9167	0.8775	0.9095	0.8271	0.762	0.8908	0.791	0.7191
DEGREE-1	PSNR	37.29	33.29	30.88	32.87	29.53	27.69	31.66	28.59	27.06
	SSIM	0.9574	0.9164	0.8726	0.9103	0.8265	0.7574	0.8962	0.7916	0.7177
DEGREE-2	PSNR	37.40	33.39	31.03	32.96	29.61	27.73	31.73	28.63	27.07
	SSIM	0.9580	0.9182	0.8761	0.9115	0.8275	0.7597	0.8937	0.7921	0.7177
DEGREE-MV	PSNR	37.61	33.70	31.30	33.11	29.77	27.92	31.84	28.76	27.18
	SSIM	0.9589	0.9212	0.8807	0.9129	0.8309	0.7637	0.8951	0.7956	0.7207
Gain	PSNR	0.47	0.45	0.26	0.40	0.22	0.16	0.30	0.18	0.07
	SSIM	0.0022	0.0045	0.0032	0.0034	0.0038	0.0017	0.0043	0.0046	0.0016

CSCN: Wang et al. CVPR 2015  
 CNN, CNN-L: Dong et al. TPAMI 2015  
 A+: Timofte et al. ACCV 2014  
 JSB-NE: Song et al. ICASSP 2016

# Deep Video SR



# Spatial-Temporal Recurrent Residual Networks for Video Super-Resolution

Supplementary Material

copyright 2017 by MEDVIEW

# Deep Video SR Results (4x)



(i) VE



(j) 3DSKR



(k) BRCN



(l) STR-ResNet



(m) Details of VE



(n) Details of 3DSKR



(o) Details of BRCN



(p) Details of STR-ResNet

# Deep Video SR Results (4x)

	Bicubic	A+	SRCNN	VE	3DKR	DRAFT (5)	DRAFT (31)	BR CN	STR-ResNet
Tractor	31.10	32.07	32.13	31.27	32.27	31.73	30.34	33.23	<b>33.85</b>
Sunflower	37.85	38.87	38.69	37.55	37.57	35.62	36.43	39.28	<b>40.02</b>
Blue sky	28.77	30.02	30.16	29.19	29.74	30.34	30.92	31.40	<b>32.23</b>
Station2	33.35	34.26	34.38	33.36	34.80	32.99	33.22	35.20	<b>35.63</b>
Pedestrian	33.51	34.43	34.55	33.60	33.91	33.40	31.78	34.95	<b>35.22</b>
Rush_hour	38.17	39.15	38.90	37.96	37.49	36.93	36.22	39.86	<b>40.30</b>
Average	33.79	34.80	34.80	33.82	34.30	33.50	33.15	35.65	<b>36.21</b>

Degradation: blurred with  $9 \times 9$  Gaussian filters with blur level 1.6

Sequences are from: <https://media.xiph.org/video/derf/>

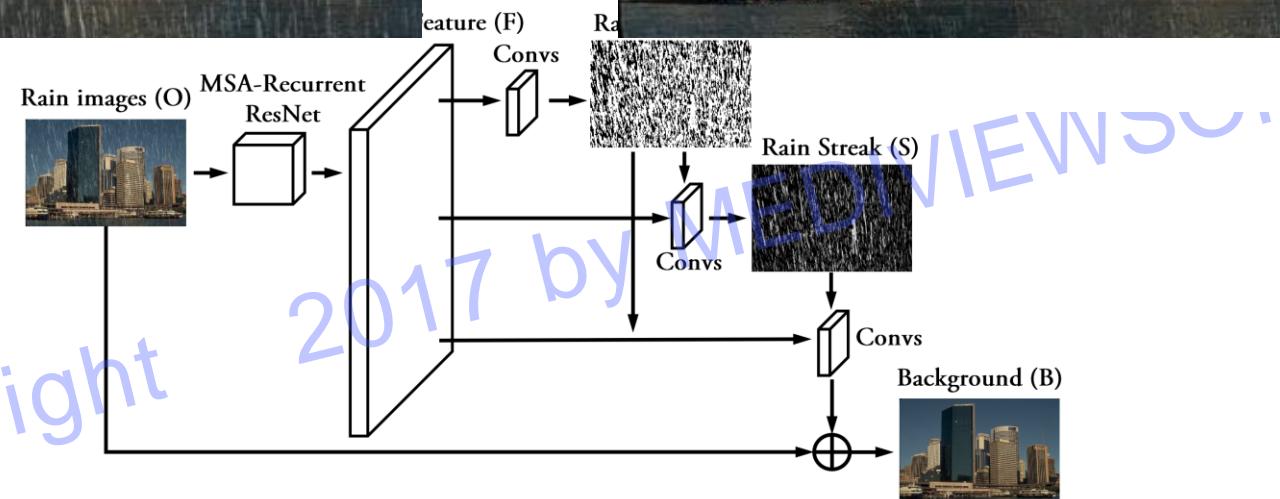
[VE] A commercial software: <http://www.infognition.com/videoenhancer/>.

[3DSKR] H. Takeda, P. Milanfar, M. Protter, and M. Elad. Super-resolution without explicit subpixel motion estimation. *TIP*. 18(9):1958–1975, 2009.

[DRAFT] R. Liao, X. Tao, R. Li, Z. Ma, and J. Jia. Video super-resolution via deep draft-ensemble learning. June 2015.

[BRCN] Y. Huang, W. Wang, and L. Wang. Bidirectional recurrent convolutional networks for multi-frame super-resolution. In *NIPS*. 2016.

# Other Image Processing Applications



copyright 2017 by NEDVIEWS.

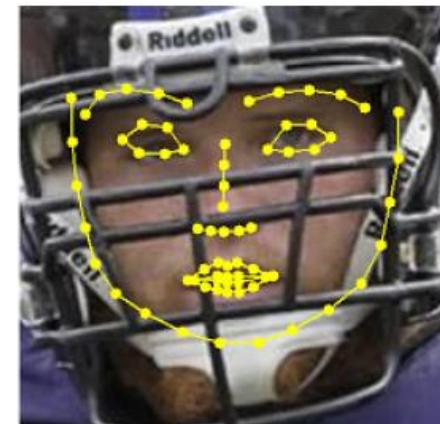
# Application II - Extreme Face Analysis



copyright 2017 by MEDIVIEWS

# Extreme Face Analysis

- Two most challenging issues in face analysis
  - Heavy occlusions
  - Aging issues



Landmark detection



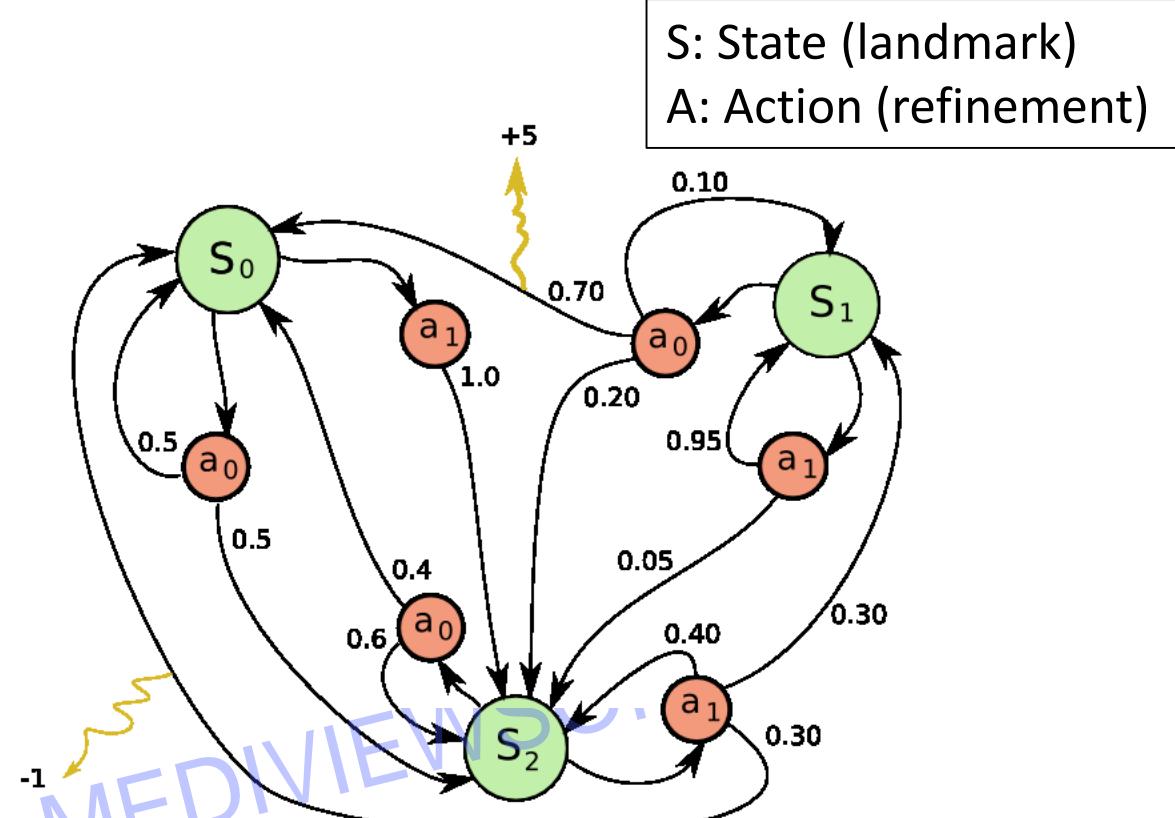
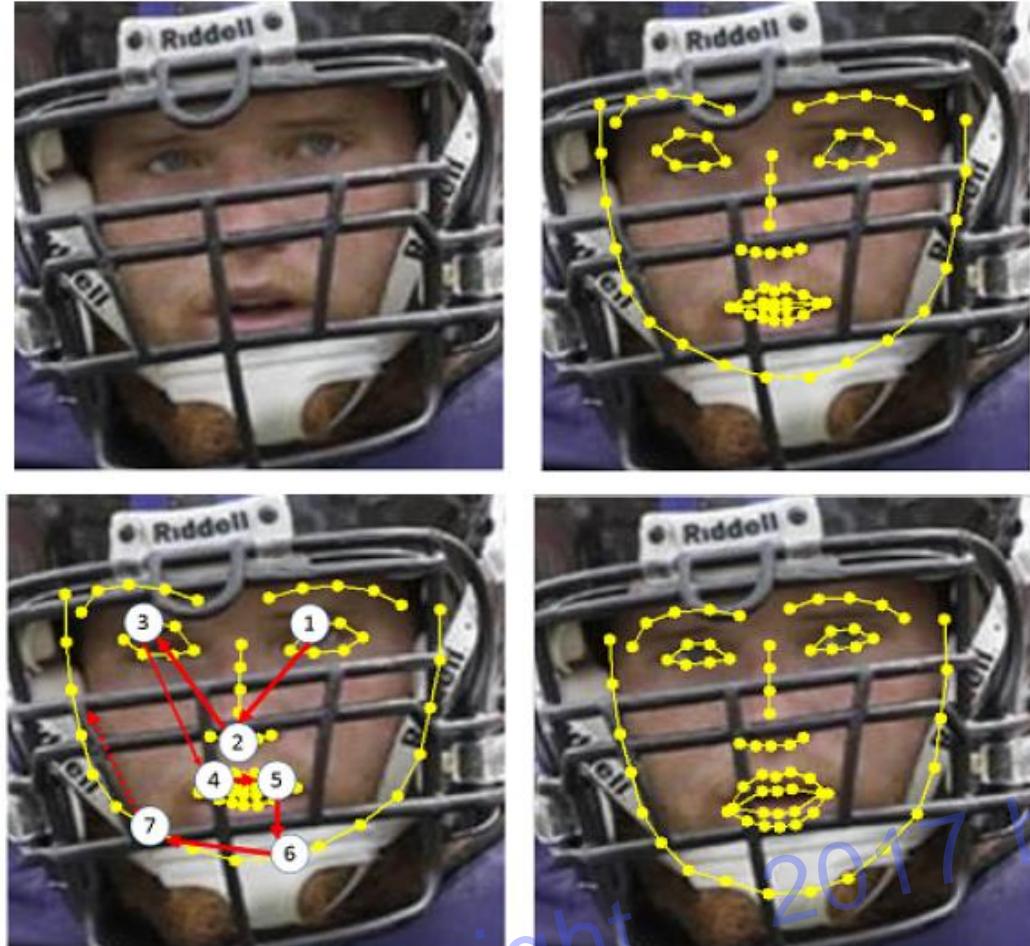
Same person?

Yes: Mickey Rourke

copyright

2017 by MEDIVIEW

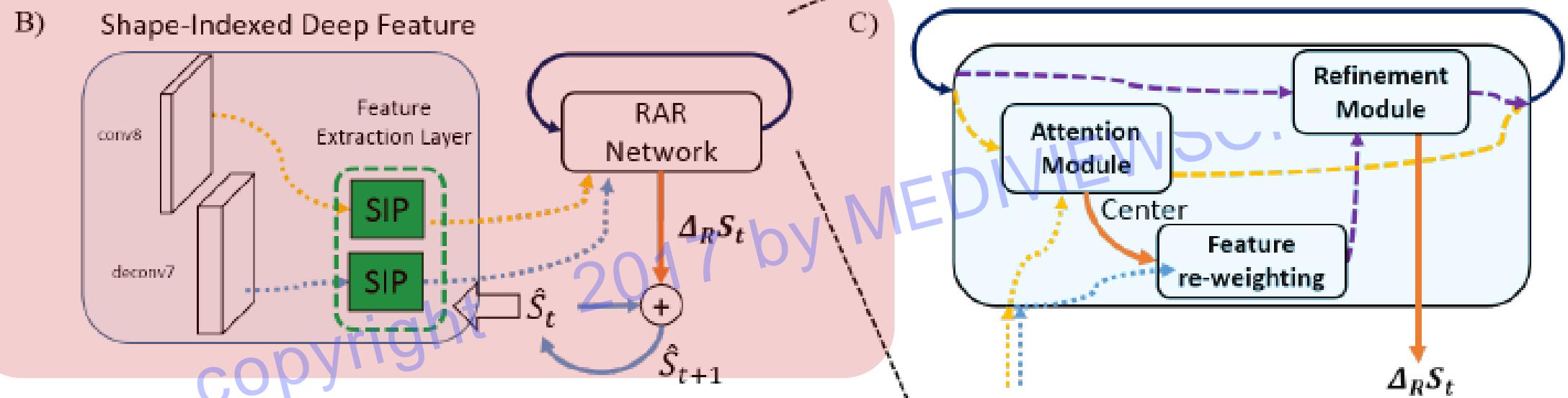
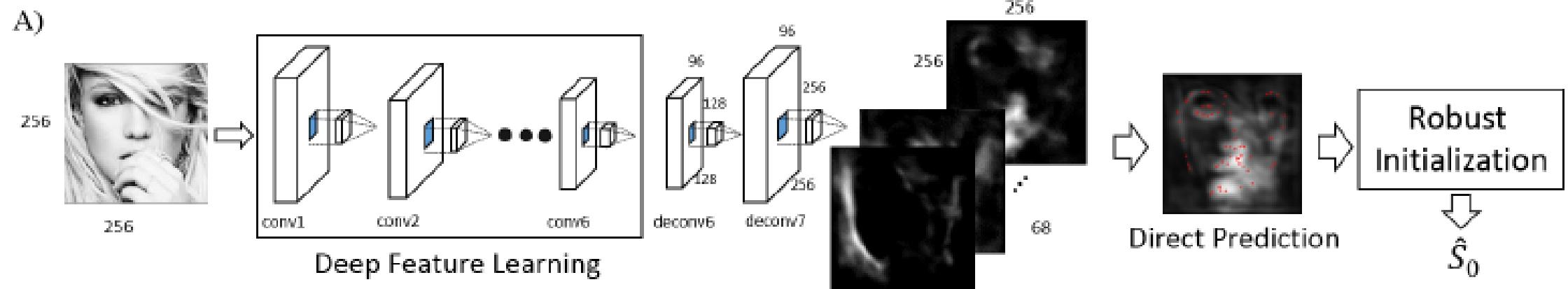
# Occlusion-Robust Face Alignment



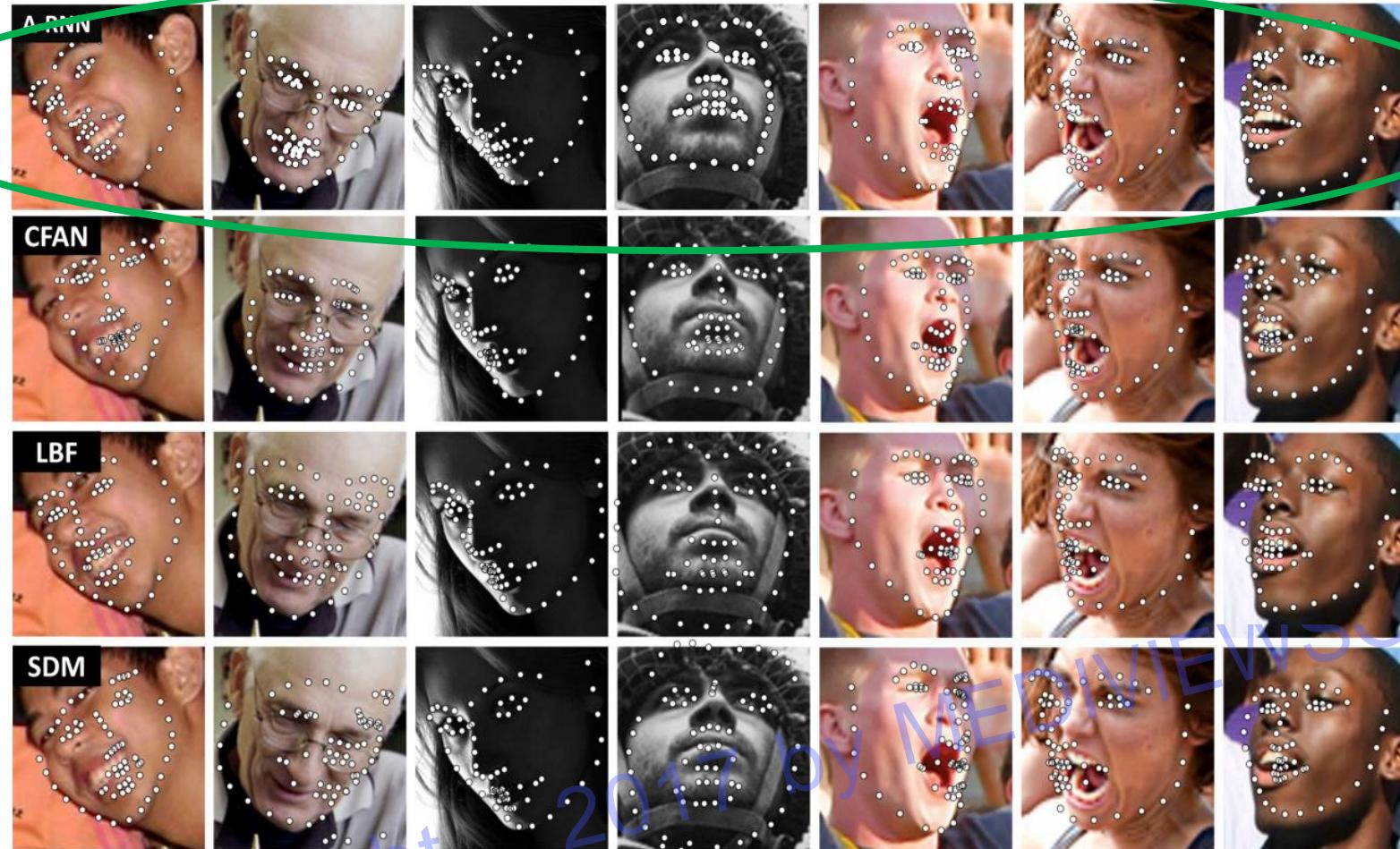
- Robust to heavy occlusions
- Efficient – encouraging good detection in early steps

copyright 2017 by MEDIVIEW

# Recurrent Attentive Networks



# Robust Face Alignment Results



copyright 2017 by MEDIVIEWDS.

*CFAN: Coarse-to-fine auto-encoder networks (cfan) for real-time face alignment. ECCV 2014*

*LBF: Face alignment at 3000 fps via regressing local binary features. ECCV 2014*

*SDM: Supervised descent method and its applications to face alignment. CVPR 2013*

# Robust Face Alignment Results

Methods	300-W Dataset		
	Common	Challenging	Full
Zhu et.al [2012]	8.22	18.33	12.0
RCPR [Burgos,2013]	6.18	17.26	8.35
SDM [Xiong,2013]	5.57	15.40	7.50
LBF [Ren,2014]	4.95	11.98	6.32
LBF Fast [Ren,2014]	5.38	15.50	7.37
CFSS [Zhu, 2015]	4.73	9.98	5.76
cGPRT [Lee, 2015]	---	---	5.71
Linkface	4.80	8.60	5.54
Ours (RAR)	4.12	8.35	4.94

Zhu: Face detection, pose estimation, and landmark localization in the wild. CVPR 2012

PCPR: Robust face landmark estimation under occlusion. ICCV 2013

LBF: Face alignment at 3000 fps via regressing local binary features. ECCV 2014

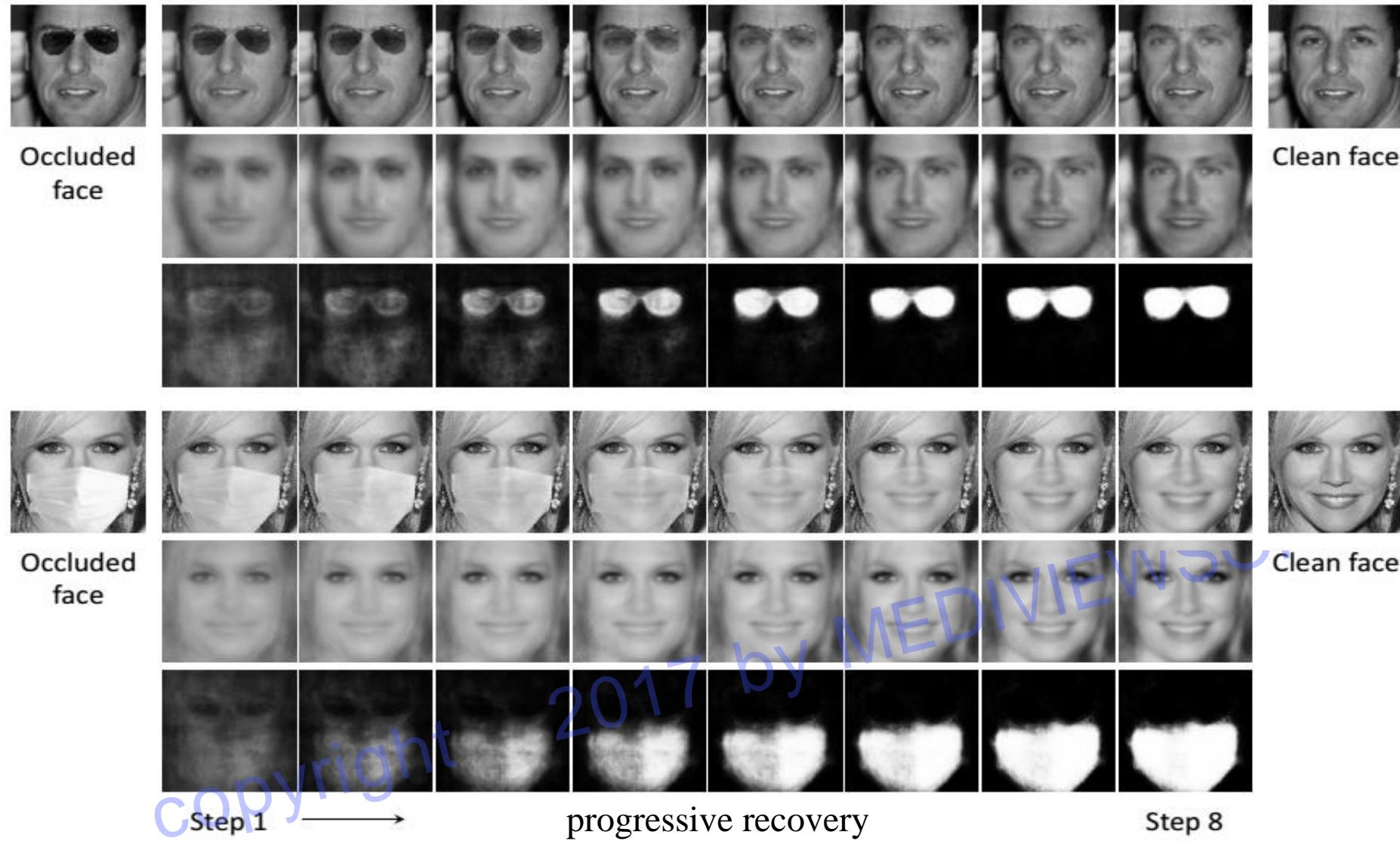
SDM: Supervised descent method and its applications to face alignment. CVPR 2013

CFSS: Face alignment by coarse-to-fine shape searching. CVPR 2015

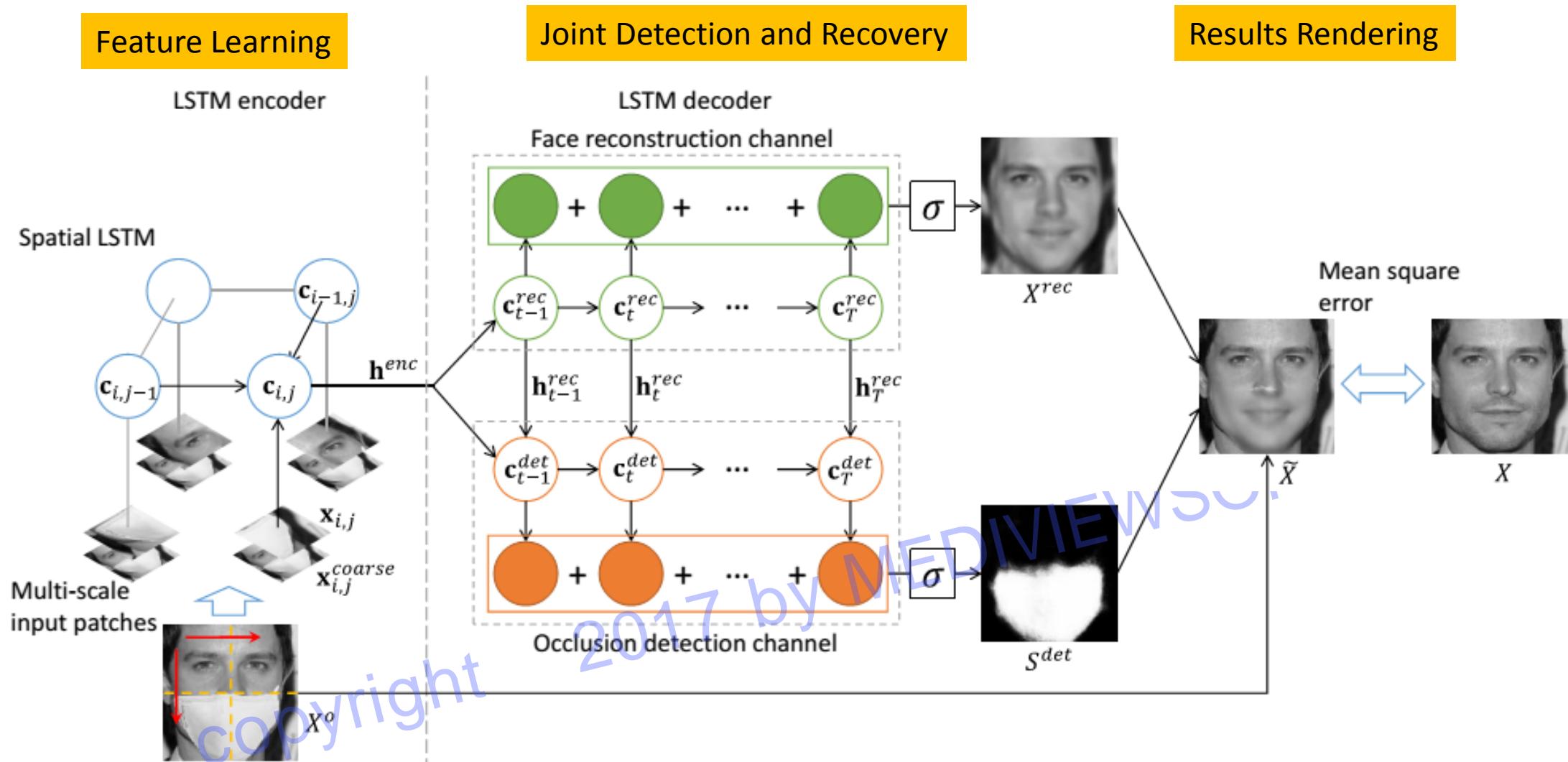
cGPRT: Face alignment using cascade Gaussian process regression trees. CVPR 2015

copyright

# Face De-occlusion



# Face De-occlusion Networks



# Face De-occlusion Results

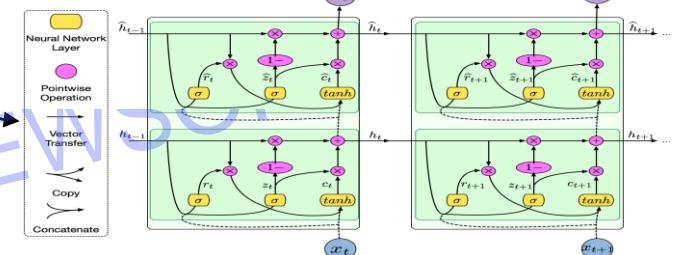
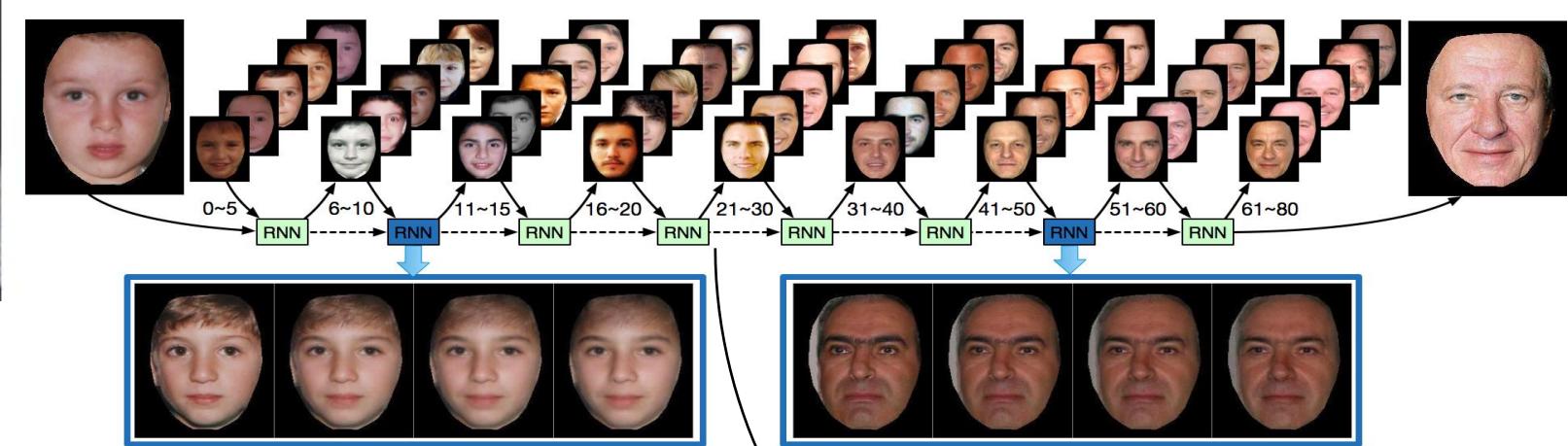


copyright 2017 by MEDIVIEW

# Face Aging Networks

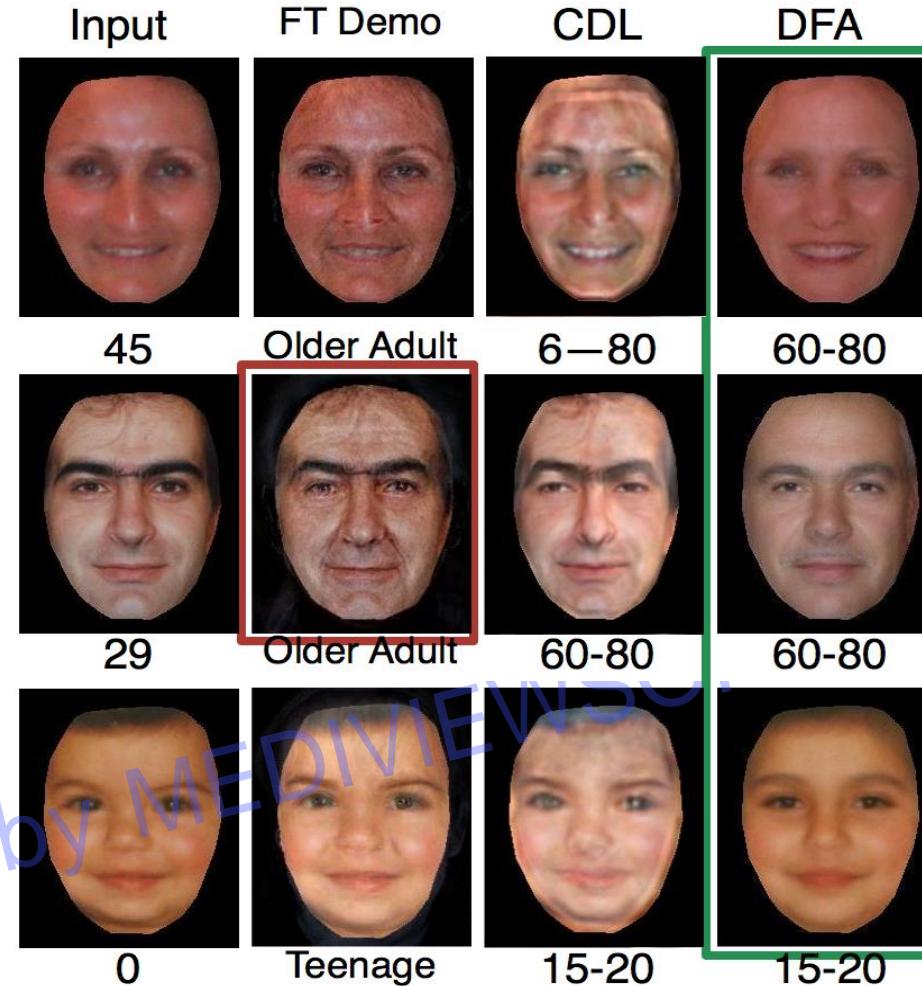
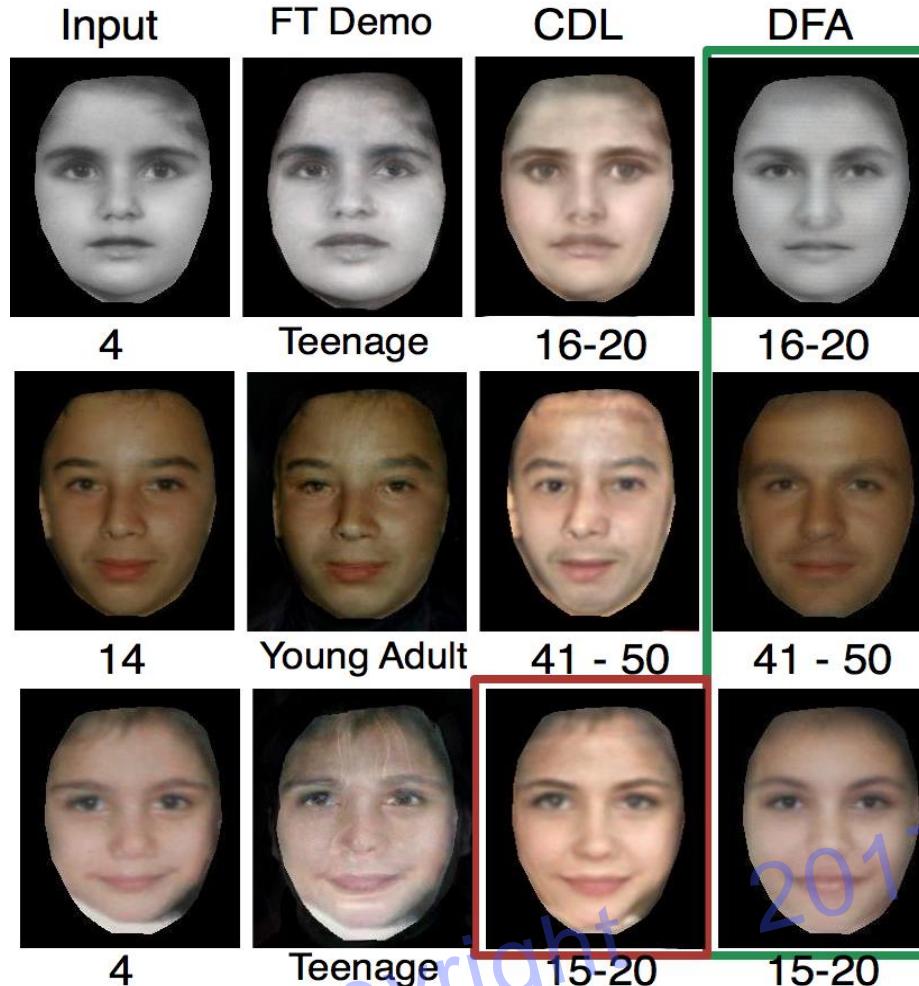


*Same person?*



copyright

# Face Aging Results

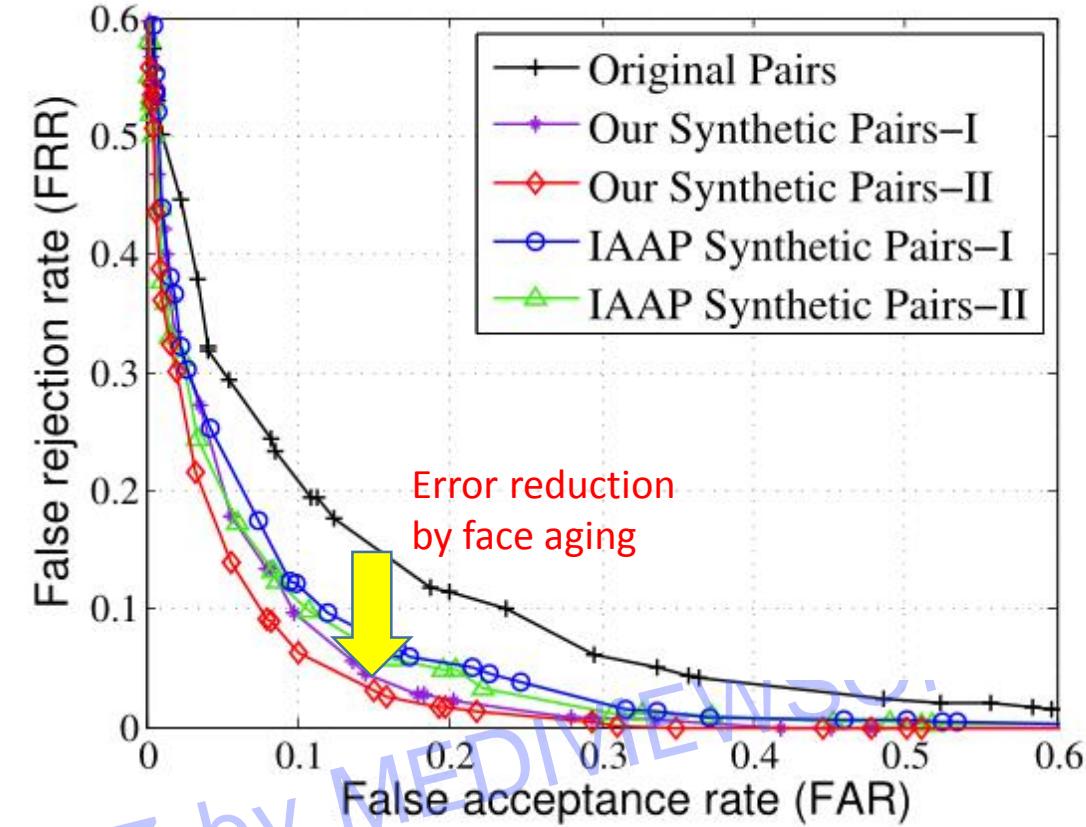


copyright 2017 by MEDIVIEW3D

# Cross-age Face Verification Results

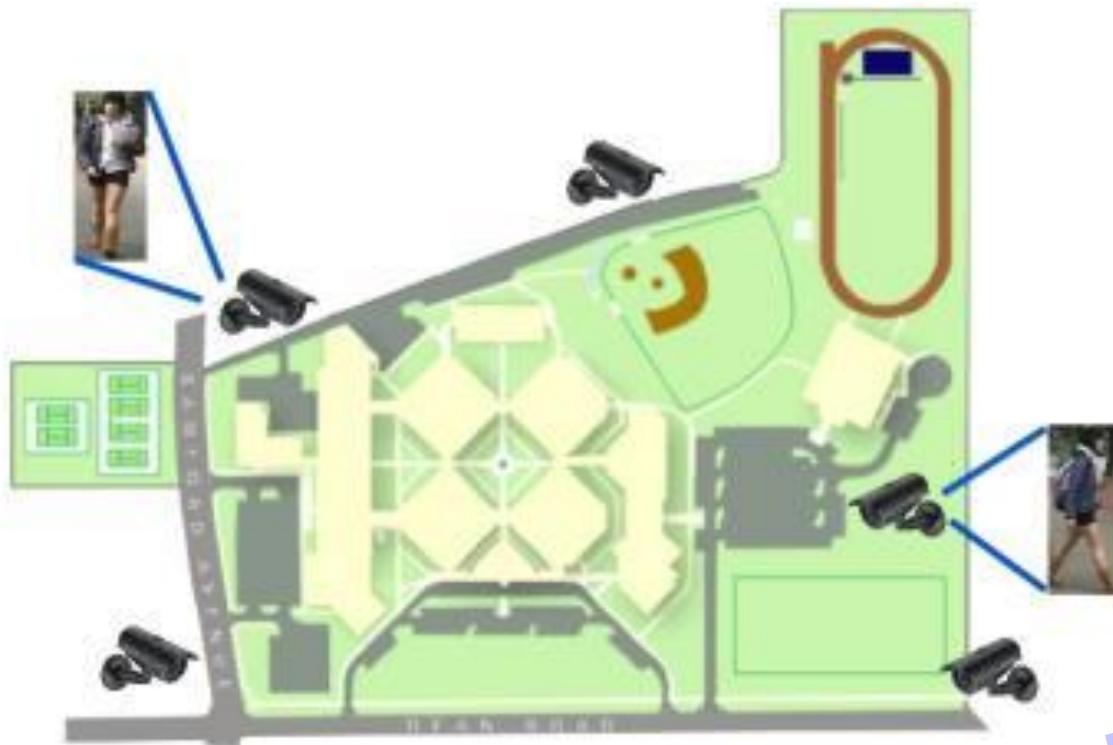


(a) Pair setting.



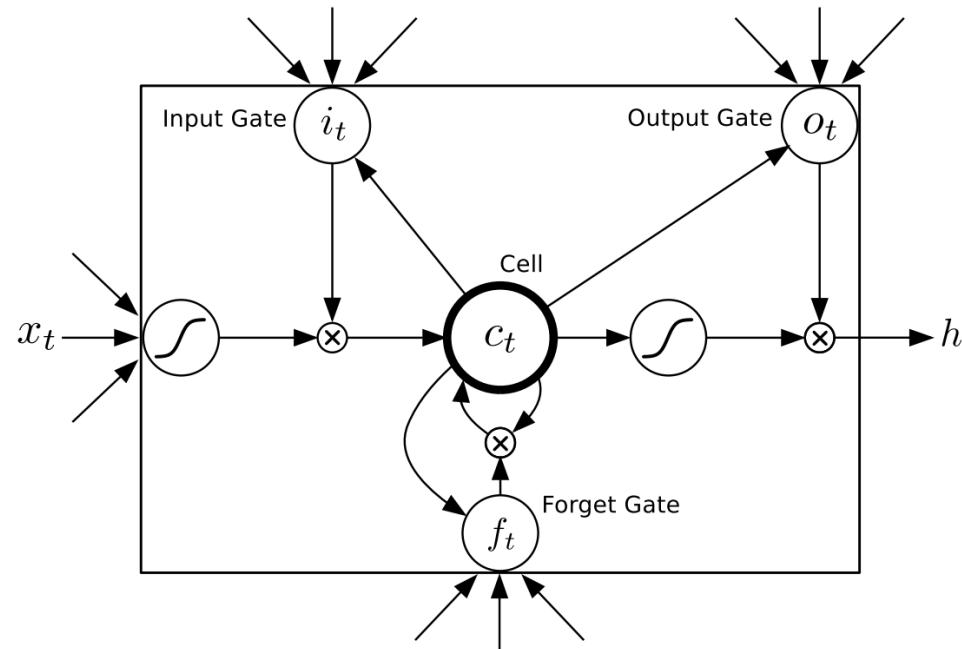
(b) FAR-FRR curve.

# Application III: Human Re-identification

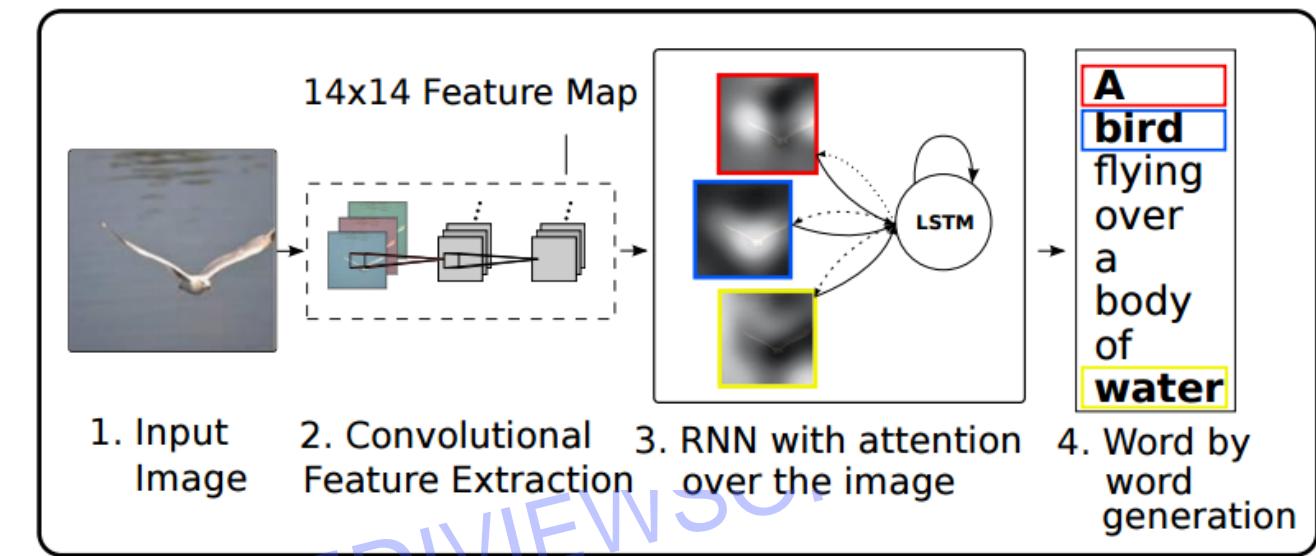


copyright 2017 by MEDIVIEW

# Attention Networks



Long Short-Term Memory (LSTM)



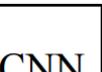
Attention Networks

# Attentive Comparison Networks

Image pairs



Feature extraction



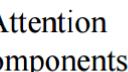
$X$

Sharing parameters



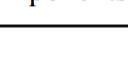
$X'$

Comparative attention generation



$H$

Sharing parameters

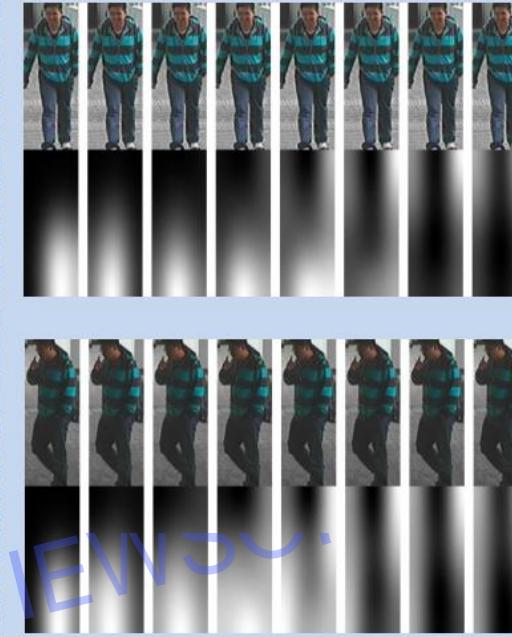


$H'$

Distance computation



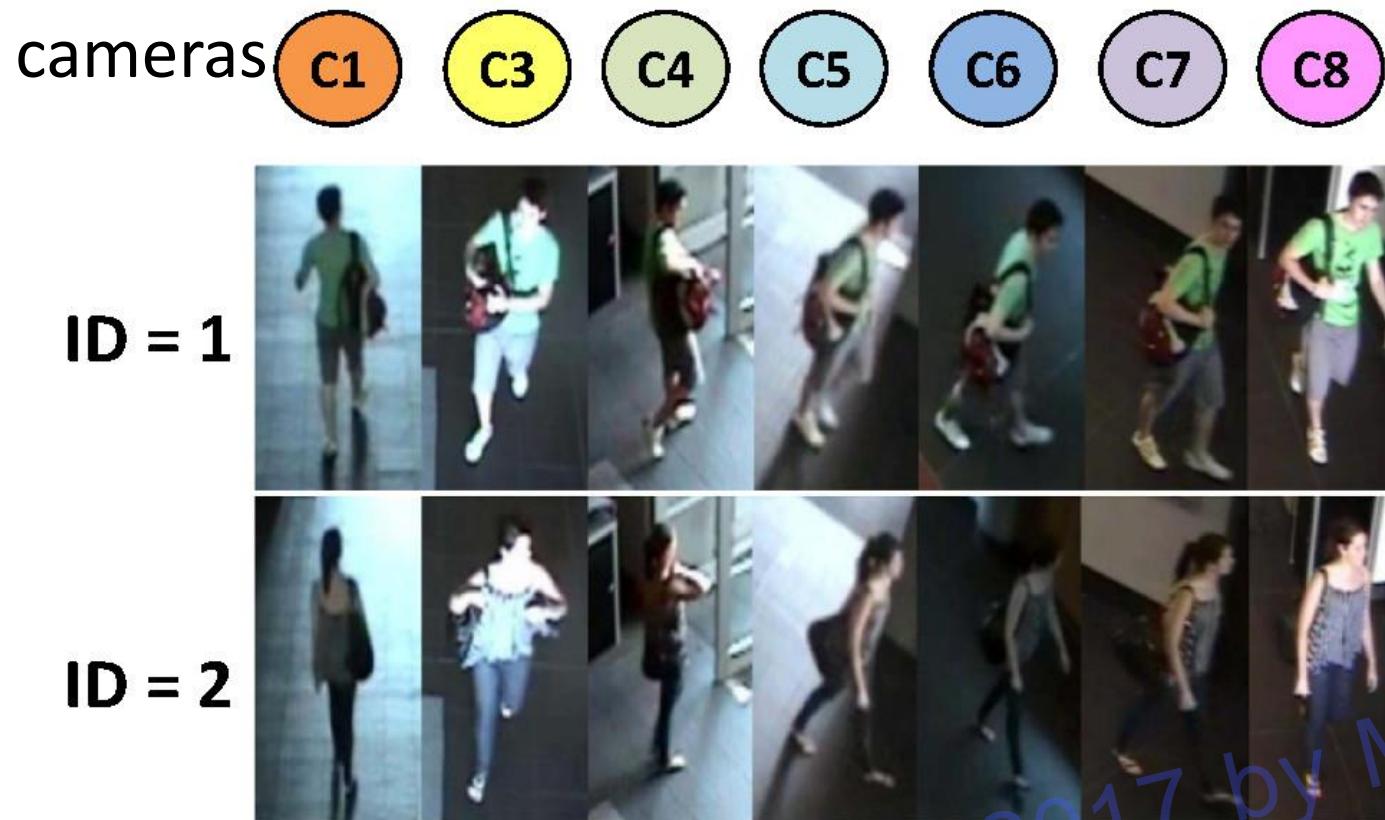
Person 1



Recurrent comparison with attention

copyright 2017 by MEDIVIEWS

# Person Re-ID Results



Results on CUHK01					
Model	Rank1	Rank5	Rank10	CMC	
CNN (extra data)	70.45	94.12	98.33	2.11	
CNN (CUHK03 pre-train)	70.12	91.33	95.44	2.33	
Ours	81	97	1	1.77	
SOTA	65.00	89	94	---	

Results on CUHK03					
Model	Rank1	Rank5	Rank10	CMC	
CNN	52.33	83.02	90.12	3.58	
Ours	63	88	93	3.49	
SOTA	44.96	77.5	83	---	

\* SOTA = State-of-the-art

copyright 2017 by MEDIVIEWSDX

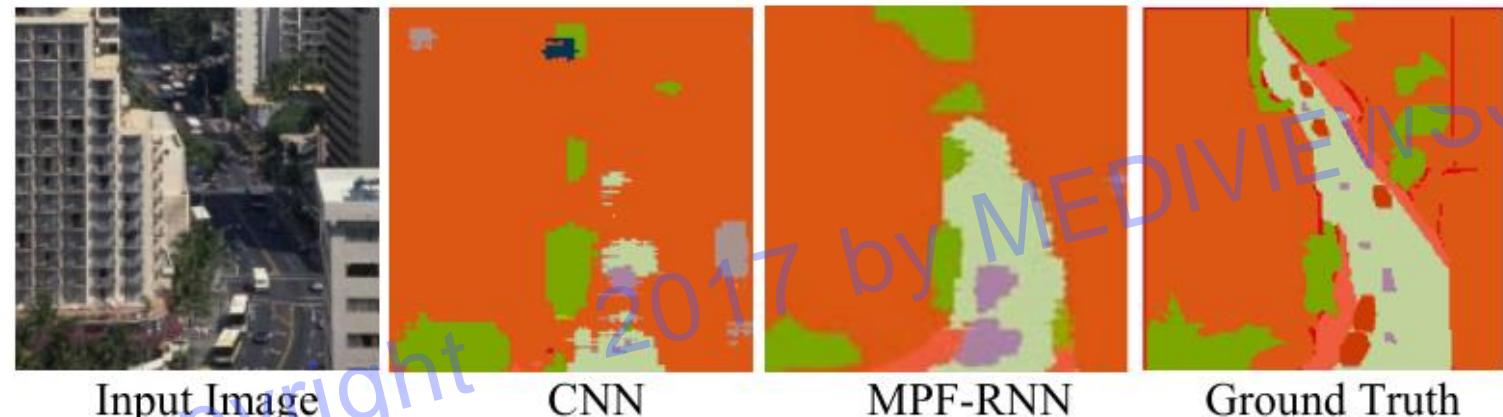
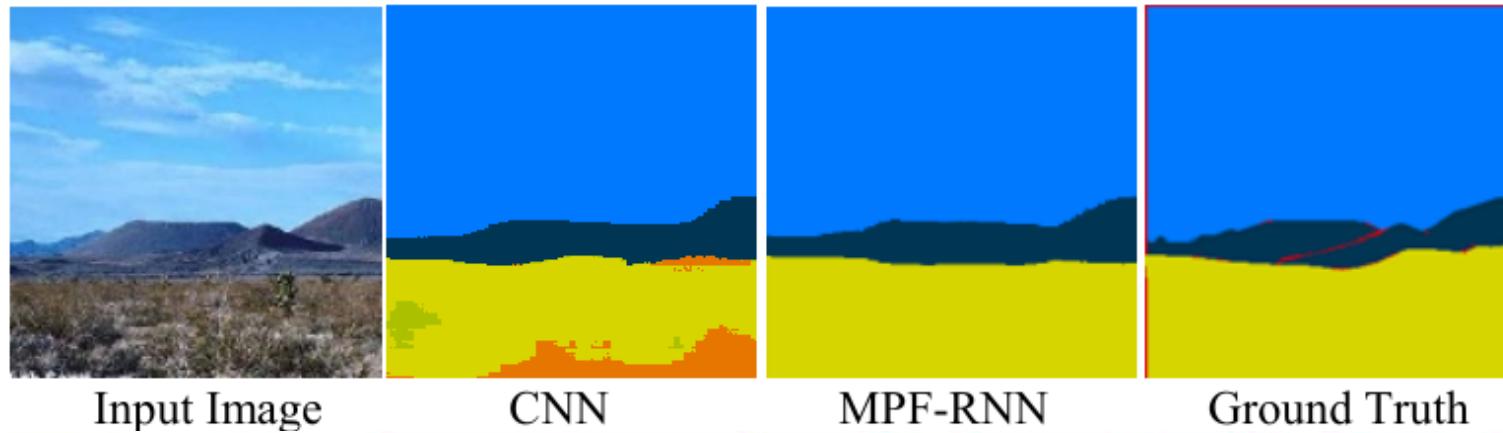
# Application IV - Scene Understanding

- Predict category label for each pixel.
- Challenging: joint segmentation, classification and detection.



# Scene Understanding

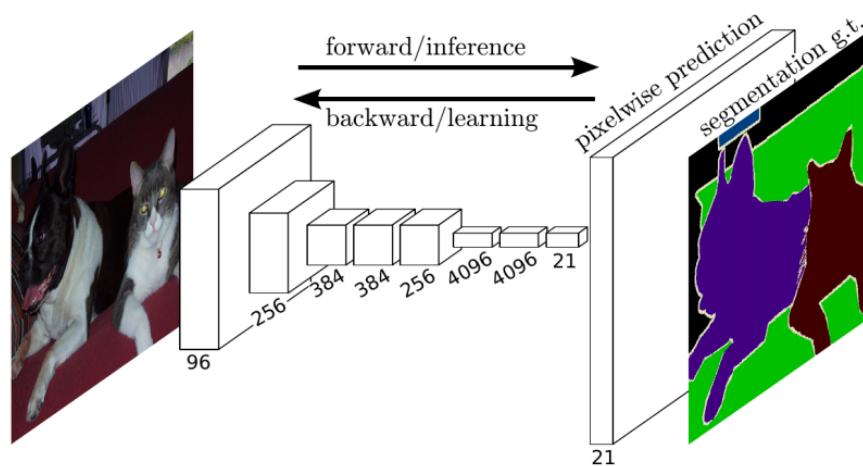
- Context is important for distinguishing confusing pixels



copyright  
2017 by MEDIVIEW

# Multi-path Feedback Networks

- Context is important
- However, CNN has limited Receptive Fields even at top layers



A typical CNN model

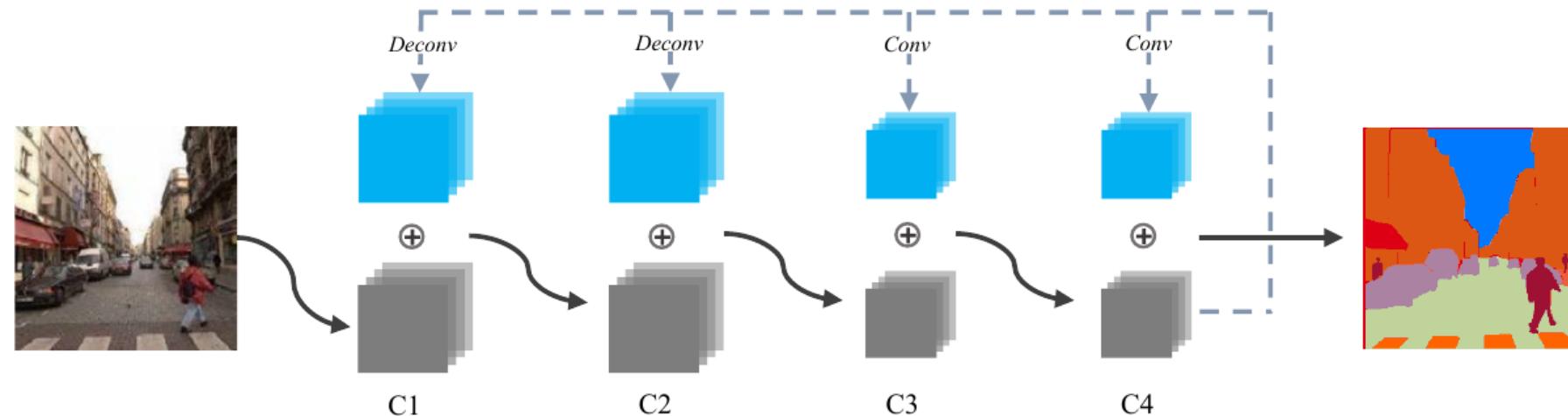


receptive field

copyright 2017 by MEDIVIEWS

# Multi-path Feedback Networks

- Architecture

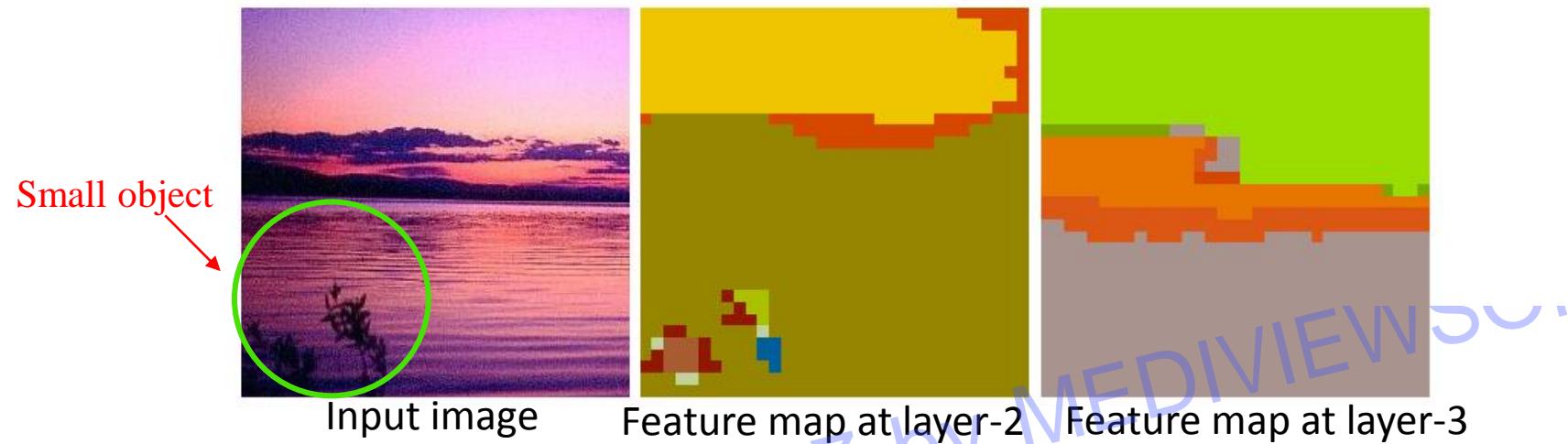


- Multi-path feedback: propagate context information at top layers to multiple bottom layers
- Multi-step fusion: aggregate results from multiple steps to parse small objects better

copyright 2017 by MEDIEVENS

# Multi-path Feedback Networks

- Multi-step Loss
  - Output features capturing context from smaller RF
  - More discriminative for small objects



copyright

2017 by MEDIVIEWS.

# Scene Parsing Results



\*SIFTFLOW Dataset

<sup>+</sup>Dag-recurrent neural networks for scene labeling. ArXiv 2015

copyright  
2017 by

# Conclusions



- Deep learning (deep neural networks) is revolutionizing visual signals processing.
- Big visual data is very valuable.
- Computational resource is important.

copyright 2017 by MEDIVIEW

# Thanks!

<http://www.lv-nus.org>

copyright 2017 by MEDIVIEWS®.