清華大學
Tsinghua University

Deep Learning for Visual Understanding

Jiwen Lu

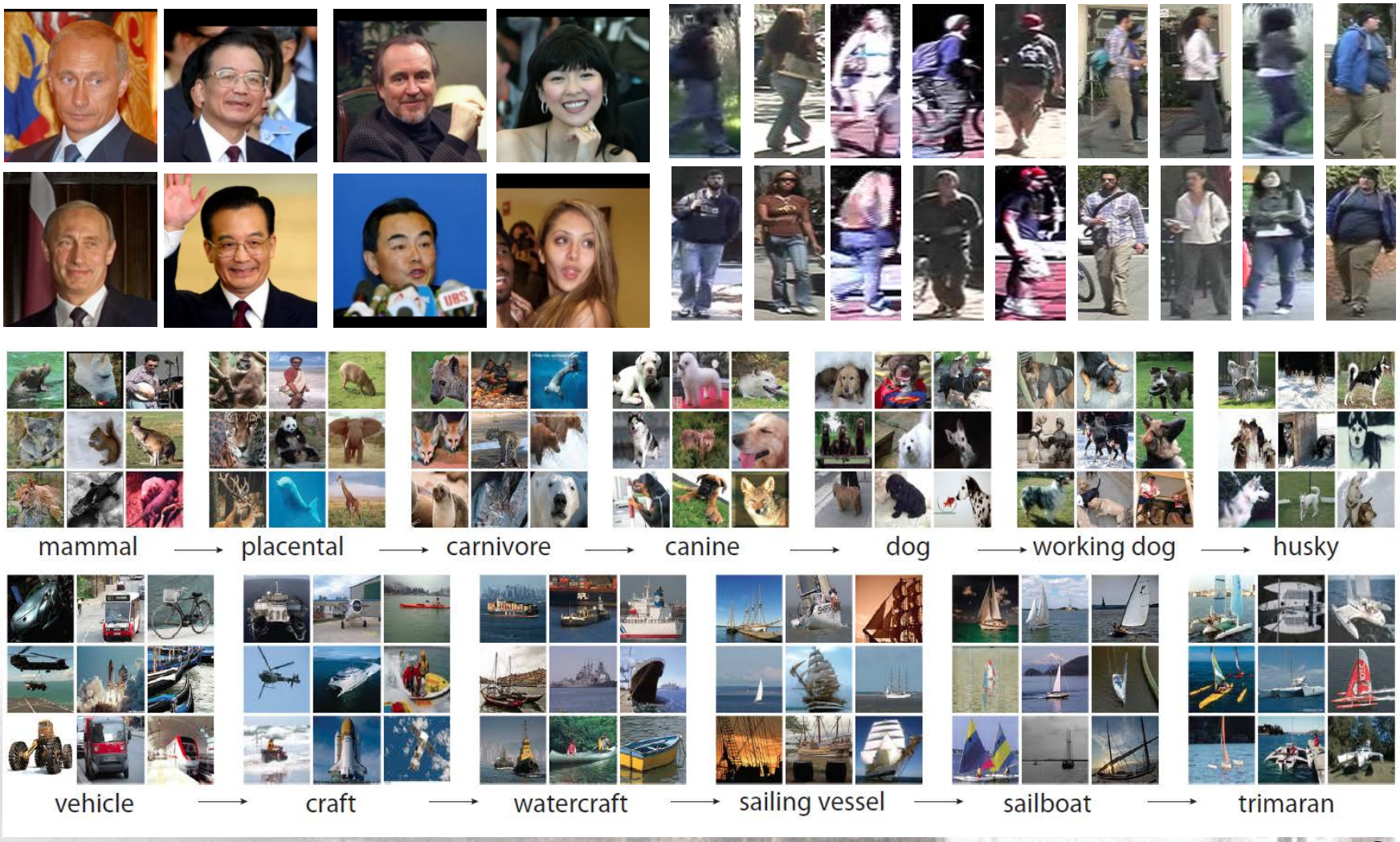
Department of Automation, Tsinghua University, China

http://ivg.au.tsinghua.edu.cn/Jiwen_Lu/

Visual Understanding

Tsinghua University

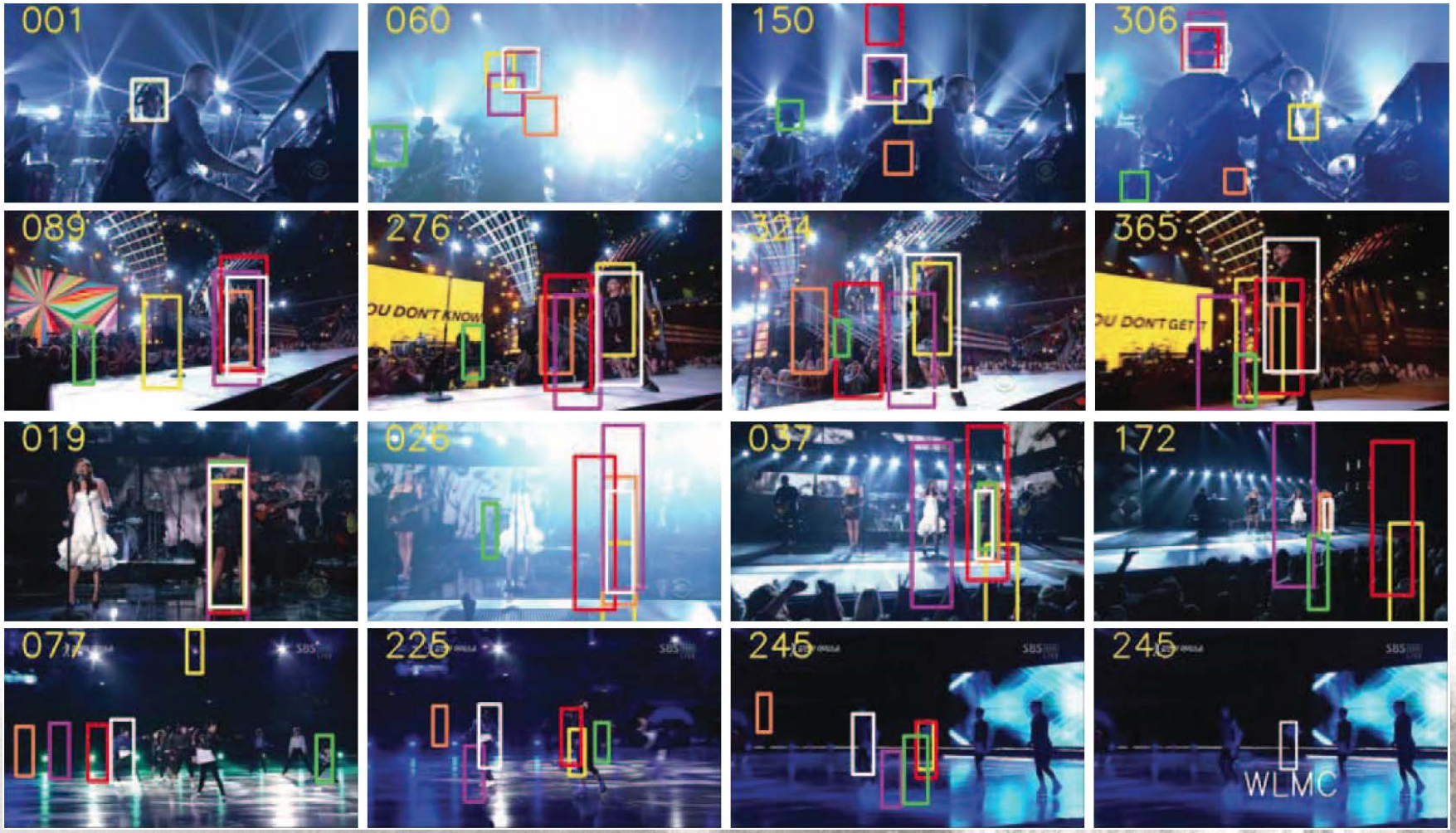
Visual Recognition



Visual Understanding

Tsinghua University

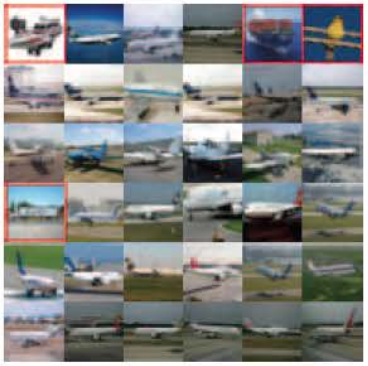
Visual Tracking



Visual Understanding

Tsinghua University

Visual Search



Precision: 90.00%

Precision: 72.22%

Precision: 55.56%

Precision: 44.44%

Precision: 94.44%

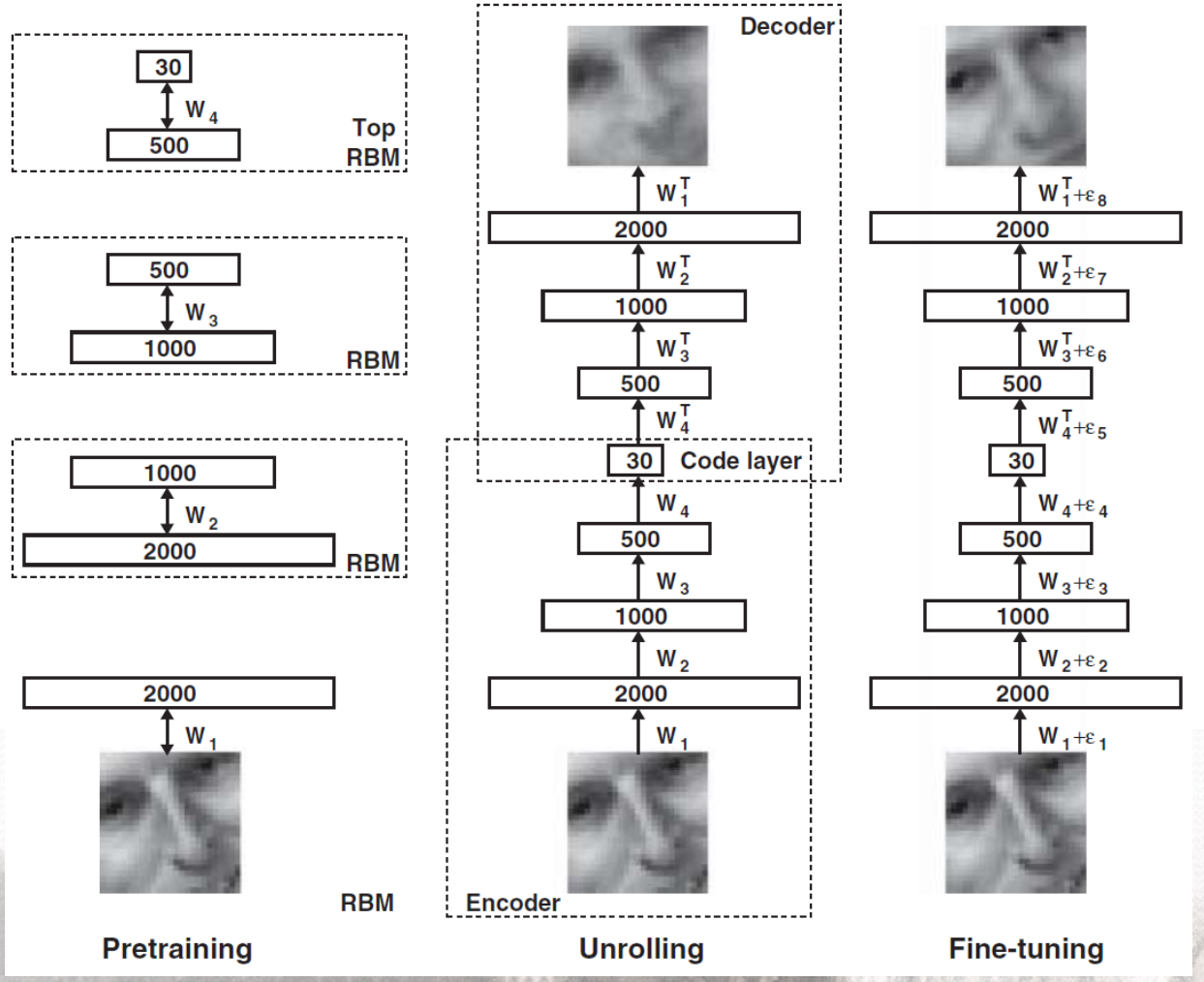
Precision: 66.67%

Precision: 55.56%

Precision: 36.11%

Deep Learning

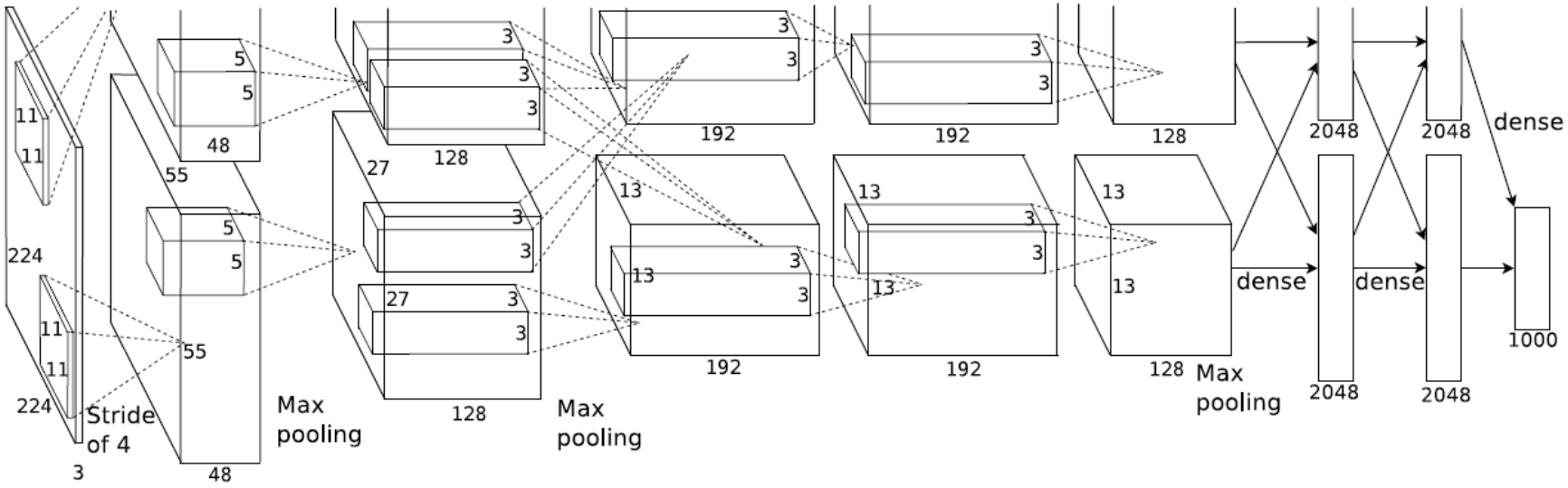
Deep Auto-Encoder Networks



[Hinton et al, Science 2006]

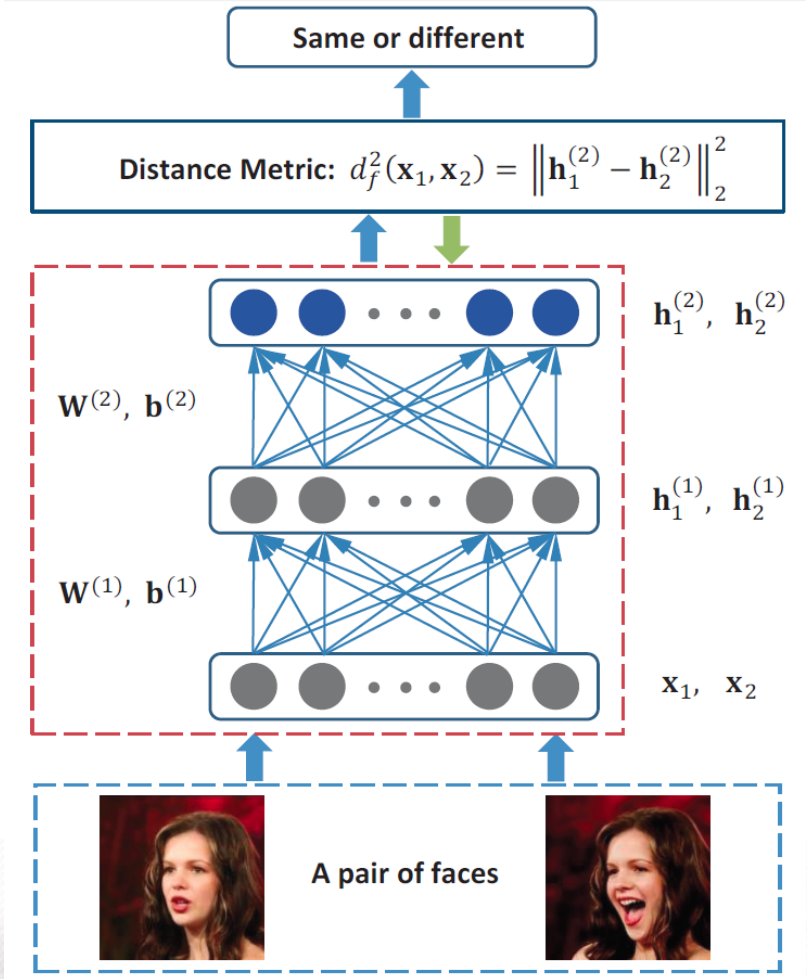
Deep Learning

Convolutional Neural Networks



[Alex et al, NIPS2012]

Deep Metric Learning



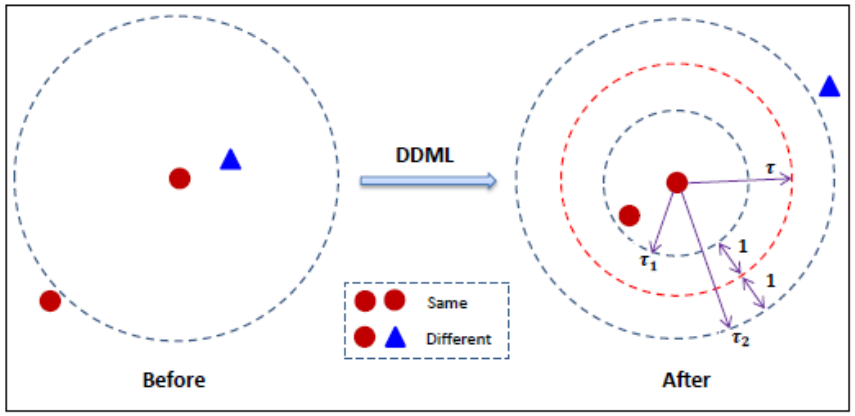
Final representation:

$$f(\mathbf{x}) = \mathbf{h}^{(M)} = s\left(\mathbf{W}^{(M)}\mathbf{h}^{(M-1)} + \mathbf{b}^{(M)}\right) \in \mathbb{R}^{p^{(M)}}$$

The distance of a pair is:

$$d_f^2(\mathbf{x}_i,\mathbf{x}_j) = \left\|f(\mathbf{x}_i) - f(\mathbf{x}_j)\right\|_2^2$$

Illustration at the top layer



$$\ell_{ij}\left(\tau - d_f^2(\mathbf{x}_i,\mathbf{x}_j)\right) > 1.$$

[1] Junlin Hu, **Jiwen Lu**, and Yap-Peng Tan, Discriminative deep metric learning for face verification in the wild, *CVPR*, pp. 1875-1882, 2014.

[2] **Jiwen Lu**, Junlin Hu, and Yap-Peng Tan, Discriminative deep metric learning for face and kinship verification, *IEEE Trans. on Image Processing*, vol. 26, no. 9, pp. 4269-4282, 2017.

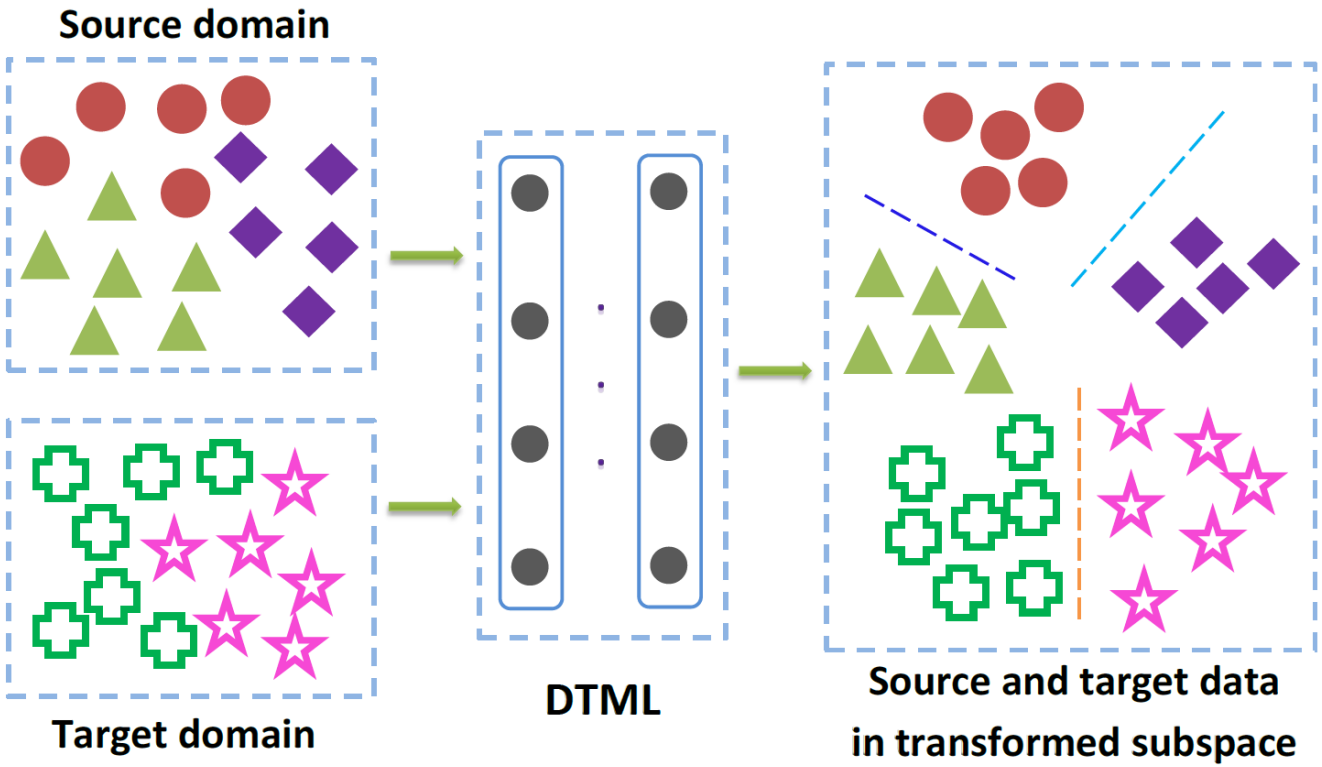
Deep Metric Learning

Comparison with State-of-the-Arts

Method	NoD	Accuracy
PCCA (SIFT) [25]	1	83.80 ± 0.40
CSML+SVM [26]	6	88.00 ± 0.37
PAF [39]	1	87.77 ± 0.51
STFRD+PMML [6]	8	89.35 ± 0.50
Fisher vector faces [28]	1	87.47 ± 1.49
DDML (SSIFT)	1	$\mathbf{87.83 \pm 0.93}$
DDML (combined)	6	$\mathbf{90.68 \pm 1.41}$

Verification rate (%) of different methods on LFW

Deep Metric Learning



Basic idea of the proposed DTML method.

[3] Junlin Hu, **Jiwen Lu**, and Yap-Peng Tan, Deep transfer metric learning, *CVPR*, pp. 325-333, 2015.
[4] Junlin Hu, **Jiwen Lu**, Yap-Peng Tan, and Jie Zhou, Deep transfer metric learning, *IEEE Trans. on Image Processing*, vol. 25, no. 12, pp. 5576-5588, 2016.

Deep Metric Learning

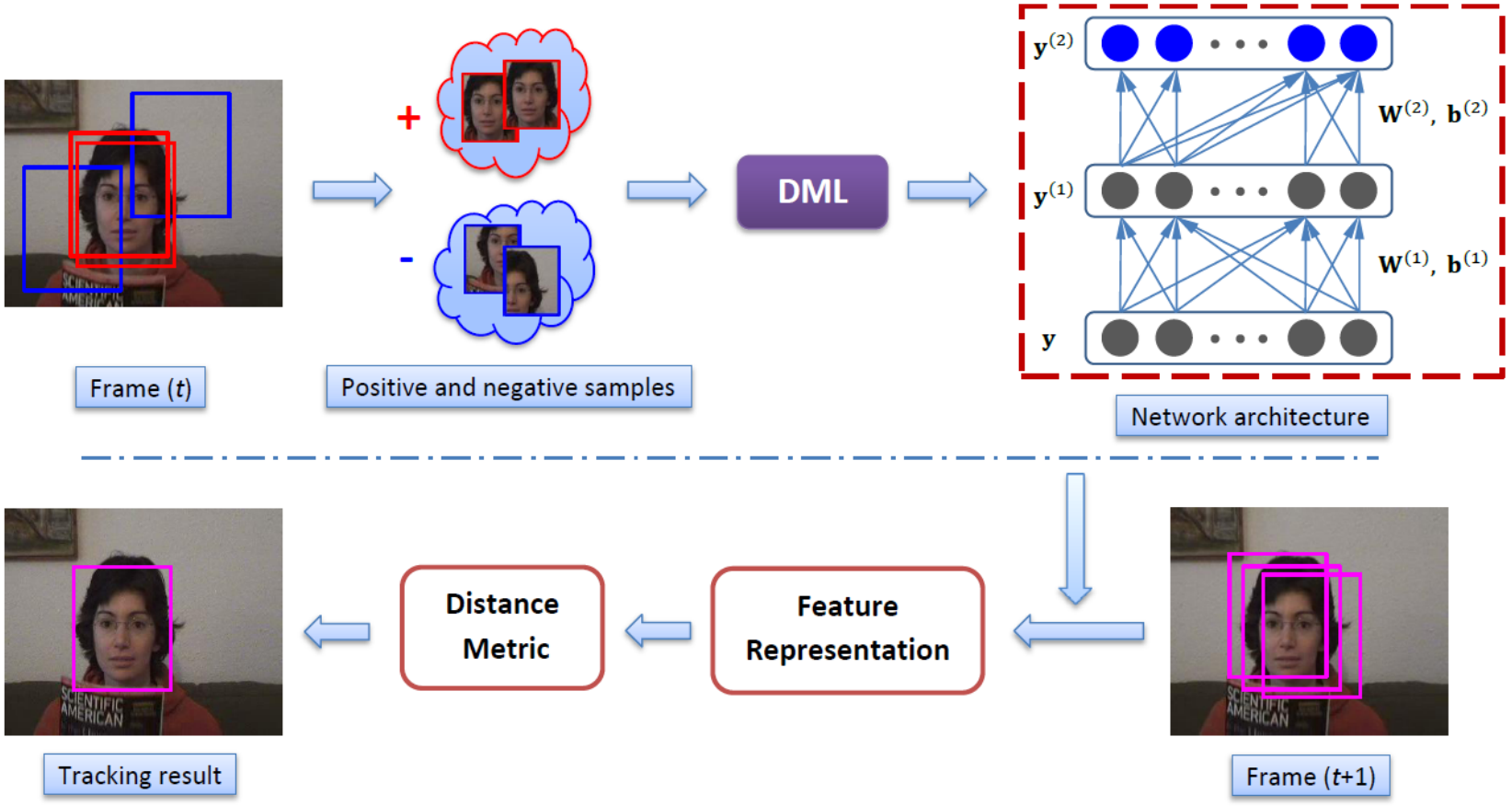
Cross-Dataset Person Re-identification

Method	Source	$r = 1$	$r = 5$	$r = 10$	$r = 30$
L_1	-	3.99	8.73	12.59	25.32
L_2	-	4.24	8.92	12.66	25.35
DDML [16]	i-LIDS	5.63	12.91	21.71	41.80
	CAVIAR	5.91	13.53	19.86	37.92
	3DPeS	6.67	17.16	23.87	41.65
DTML ($\beta = 0$)	i-LIDS	5.88	13.72	21.03	41.49
	CAVIAR	6.02	13.81	20.33	38.46
	3DPeS	7.20	18.04	25.96	43.80
DTML	i-LIDS	6.68	15.73	23.20	46.42
	CAVIAR	6.17	13.10	19.65	37.78
	3DPeS	8.51	**19.40**	**27.59**	**47.91**
DSTML	i-LIDS	6.11	16.01	23.51	45.35
	CAVIAR	6.61	16.93	24.40	41.55
	3DPeS	**8.58**	19.02	26.49	46.77

Top r matched results of different methods on the VIPeR dataset

Deep Metric Learning

Visual Tracking



Main procedure of our proposed DML tracker.

[5] Junlin Hu, **Jiwen Lu**, and Yap-Peng Tan, Deep metric learning for visual tracking, *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 26, no. 11, 2056-2068, 2016.

Deep Metric Learning

Quantitative Evaluation

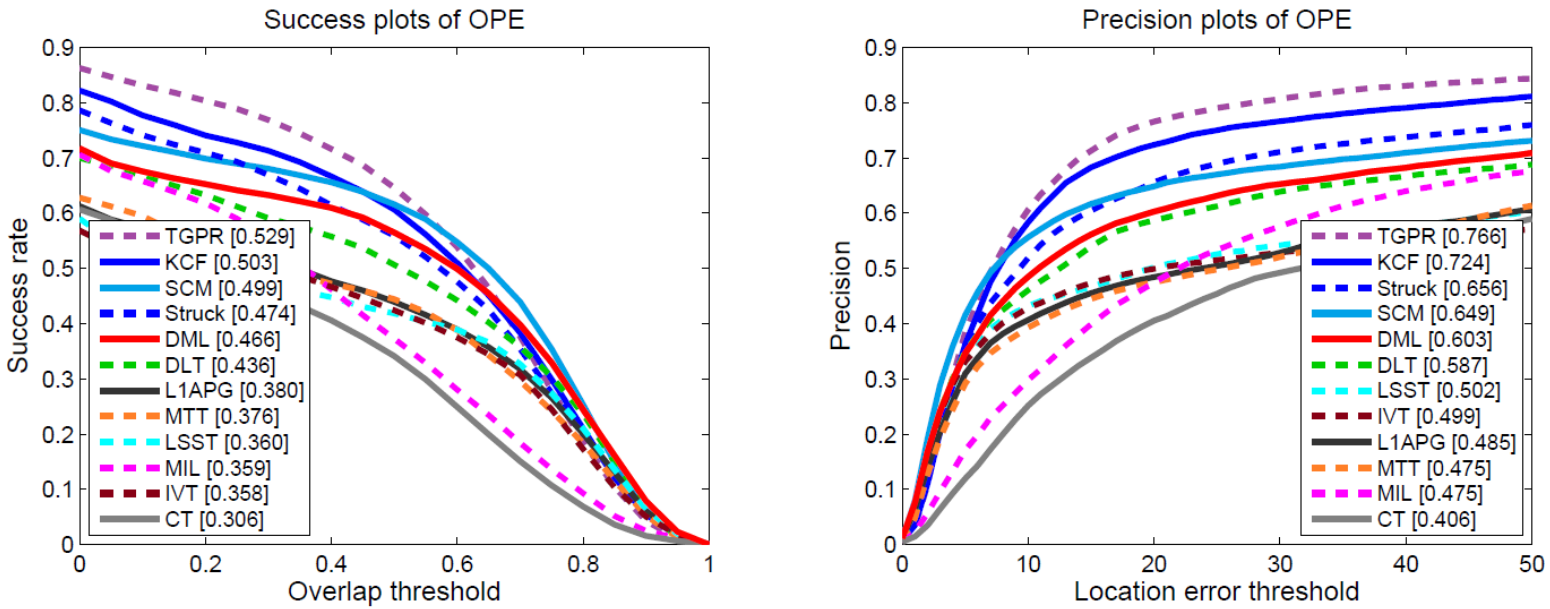
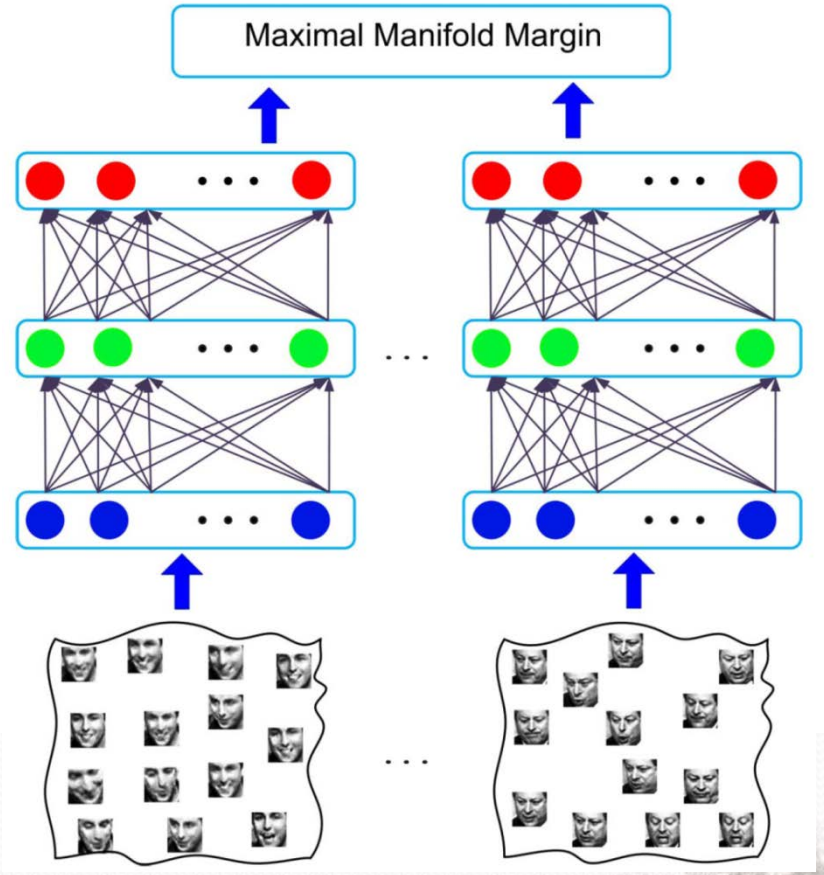


Fig. 2. The success plots and precision plots of the 12 trackers on the 51 sequences for comparing overall performance, respectively. The legend lists the performance score for each tracker. The proposed DML tracker (in red) is ranked fifth among these trackers in both the success and precision plots.

Deep Metric Learning

Image Set Classification



Our proposed multi-manifold deep metric learning framework.

[6] **Jiwen Lu**, Gang Wang, Weihong Deng, Pierre Moulin, and Jie Zhou, Multi-manifold deep metric learning for image set classification, *CVPR*, pp. 1137-1145, 2015.

Deep Metric Learning

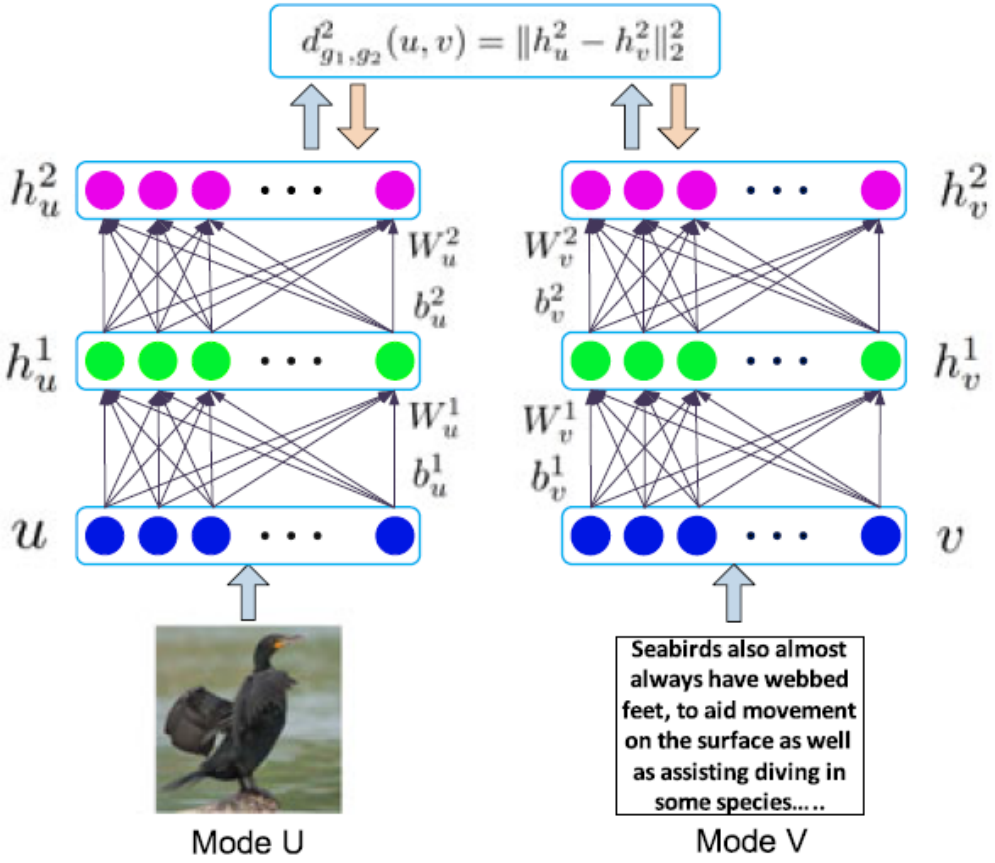
Image Set Classification

Method	Honda	Mobo	YTC	PubFig	ETH-80	Year
MSM [38]	92.5 ± 2.3	96.5 ± 2.0	61.7 ± 4.3	57.4 ± 1.7	75.5 ± 4.9	1998
DCC [16]	92.6 ± 2.5	88.9 ± 2.5	65.8 ± 4.5	45.5 ± 1.5	91.8 ± 3.7	2006
MMD [36]	92.1 ± 2.3	92.5 ± 2.9	67.7 ± 3.8	46.3 ± 1.5	86.5 ± 4.5	2008
MDA [34]	94.5 ± 3.2	94.4 ± 2.5	68.1 ± 4.3	48.6 ± 1.6	89.2 ± 3.7	2009
AHISD [2]	91.5 ± 1.8	94.1 ± 1.5	66.5 ± 4.5	62.1 ± 1.4	78.6 ± 4.7	2010
CHISD [2]	93.7 ± 1.9	95.8 ± 1.3	67.4 ± 4.7	64.5 ± 1.5	79.7 ± 4.3	2010
SANP [13]	95.3 ± 3.1	96.1 ± 1.5	68.3 ± 5.2	78.5 ± 1.4	80.5 ± 4.7	2011
CDL [35]	97.4 ± 1.3	92.5 ± 2.9	69.7 ± 4.5	65.5 ± 1.5	86.5 ± 3.7	2012
DFRV [5]	97.4 ± 1.9	94.4 ± 2.3	74.5 ± 4.5	74.5 ± 1.4	87.5 ± 2.7	2012
LMKML [27]	98.5 ± 2.5	94.5 ± 2.5	75.2 ± 3.9	72.5 ± 1.5	92.5 ± 4.5	2013
SSDML [40]	93.5 ± 2.8	95.1 ± 2.2	74.3 ± 4.5	65.5 ± 1.7	87.5 ± 4.7	2013
SFDL [26]	98.5 ± 1.5	96.5 ± 2.3	75.7 ± 3.4	78.5 ± 1.7	90.5 ± 4.7	2014
MMDML	$\mathbf{100.0 \pm 0.0}$	$\mathbf{97.8 \pm 1.0}$	$\mathbf{78.5 \pm 2.8}$	$\mathbf{82.5 \pm 1.2}$	$\mathbf{94.5 \pm 3.5}$	

Average classification rates of different methods on different datasets

Deep Metric Learning

Cross-Modal Matching



[7] Venice Erin Liong, **Jiwen Lu**, Yap-Peng Tan, and Jie Zhou, Deep coupled metric learning for cross-modal matching, *IEEE Trans. on Multimedia*, vol. 19, no. 6, pp. 1234-1244, 2017.

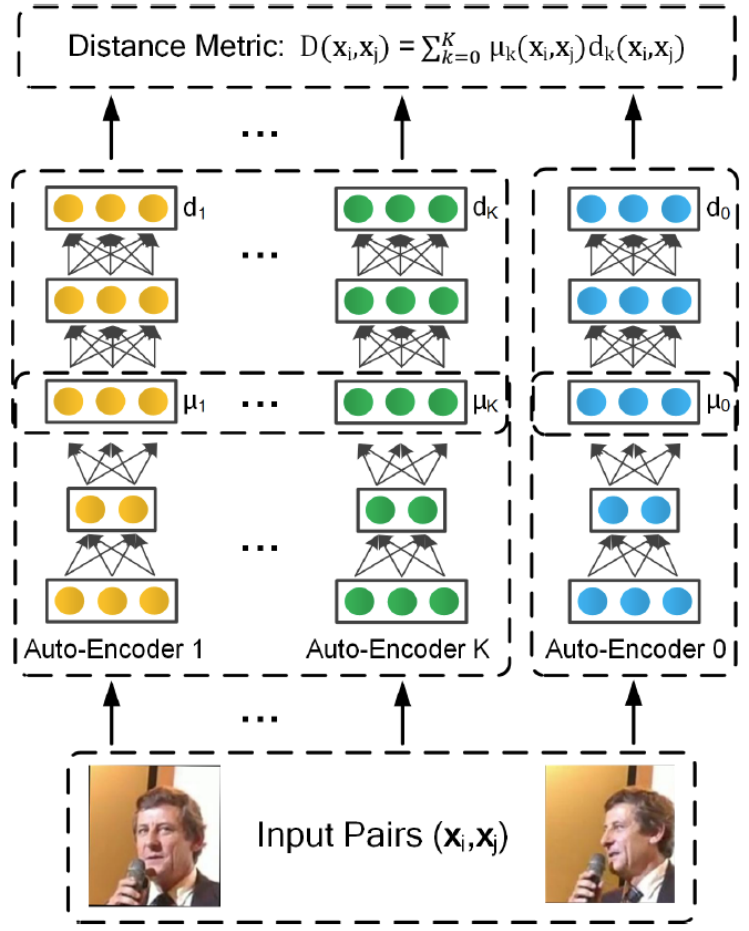
Deep Metric Learning

Cross-Modal Matching

Method	Image query	Text query	Average
PLS [20]	31.28	26.45	28.86
CCA [3]	35.42	32.50	33.96
GMMFA [45]	27.76	12.85	20.30
GMLDA [45]	18.01	14.09	16.05
MvDA [17]	16.17	12.00	14.09
SCM [4]	38.74	37.03	37.88
KCCA [21]	37.34	34.61	35.98
LCFS [5]	39.39	38.09	38.74
LRBS* [49]	44.41	37.70	41.06
RE-DNN* [25]	34.04	35.26	34.65
JFSSL* [48]	42.79	39.57	41.18
DCML	**55.36**	**53.81**	**54.59**

Deep Metric Learning

Face and Person Recognition

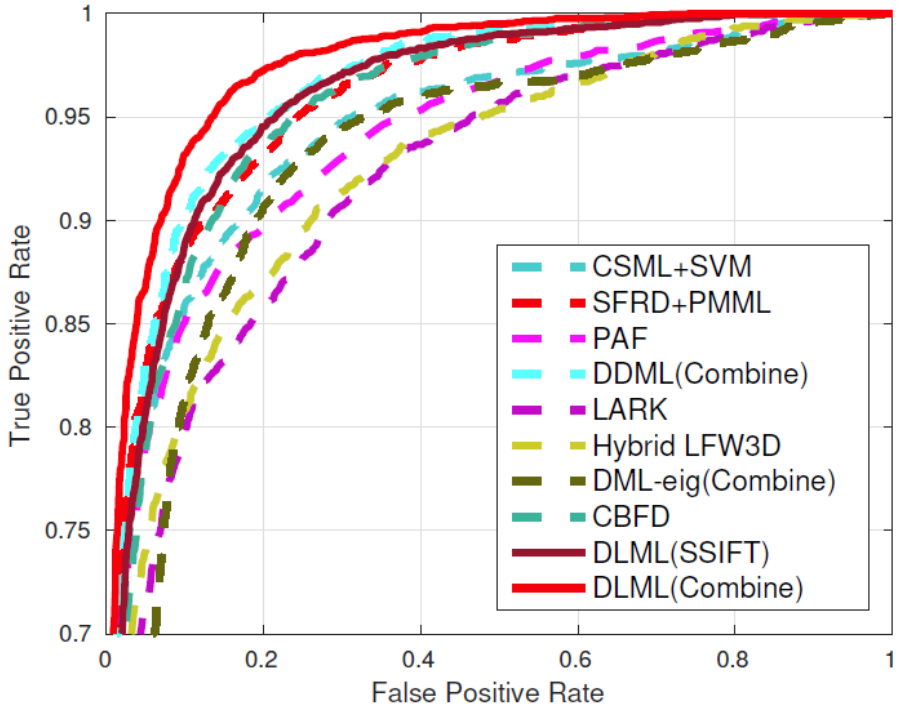


[8] Yueqi Duan, **Jiwen Lu**, Jianjiang Feng, and Jie Zhou, Deep localized metric learning, *IEEE Trans. on Circuits and Systems for Video Technology*, 2017, accepted.

Deep Metric Learning

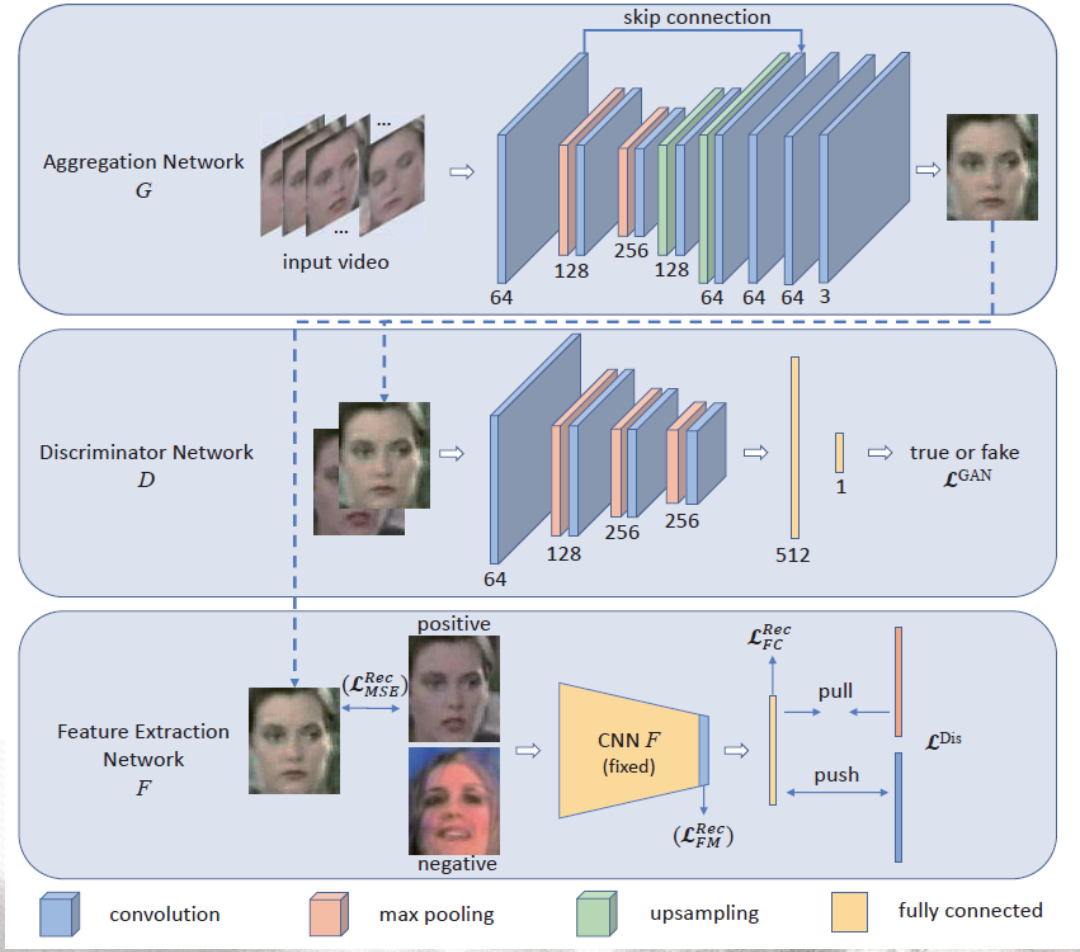
Face and Person Recognition

Method	Accuracy
Sub-SML [50]	89.90 ± 0.38
VMRS [51]	91.10 ± 0.59
Fisher vector [47]	87.47 ± 1.49
CSML+SVM [35]	80.00 ± 0.37
CBFD [30]	87.23 ± 1.68
STFRD+PMML [11]	89.35 ± 0.50
PAF [48]	87.77 ± 0.51
APEM (LBP) [49]	81.97 ± 1.90
APEM (SIFT) [49]	81.88 ± 0.94
APEM (Fusion) [49]	84.08 ± 1.20
DNLML-ISA (SSIFT) [34]	86.17 ± 0.40
DNLML-ISA [34]	88.50 ± 0.40
DDML (SSIFT) [3]	87.83 ± 0.93
DDML (Combine) [3]	90.68 ± 1.41
CDBN [46]	86.88 ± 0.62
CDBN+Hand-crafted [46]	87.77 ± 0.62
LM3L [45]	89.57 ± 1.53
DTML-AE [17]	88.23 ± 0.45
DLML (SSIFT)	89.95 ± 1.05
DLML (Combine)	$\mathbf{92.22 \pm 1.51}$



Deep Metric Learning

Adversarial Deep Metric Learning



[9] Yongming Rao, Ji Lin, **Jiwen Lu**, and Jie Zhou, Learning discriminative aggregation network for video-based face recognition, *ICCV,* pp. 3781-3790, 2017.

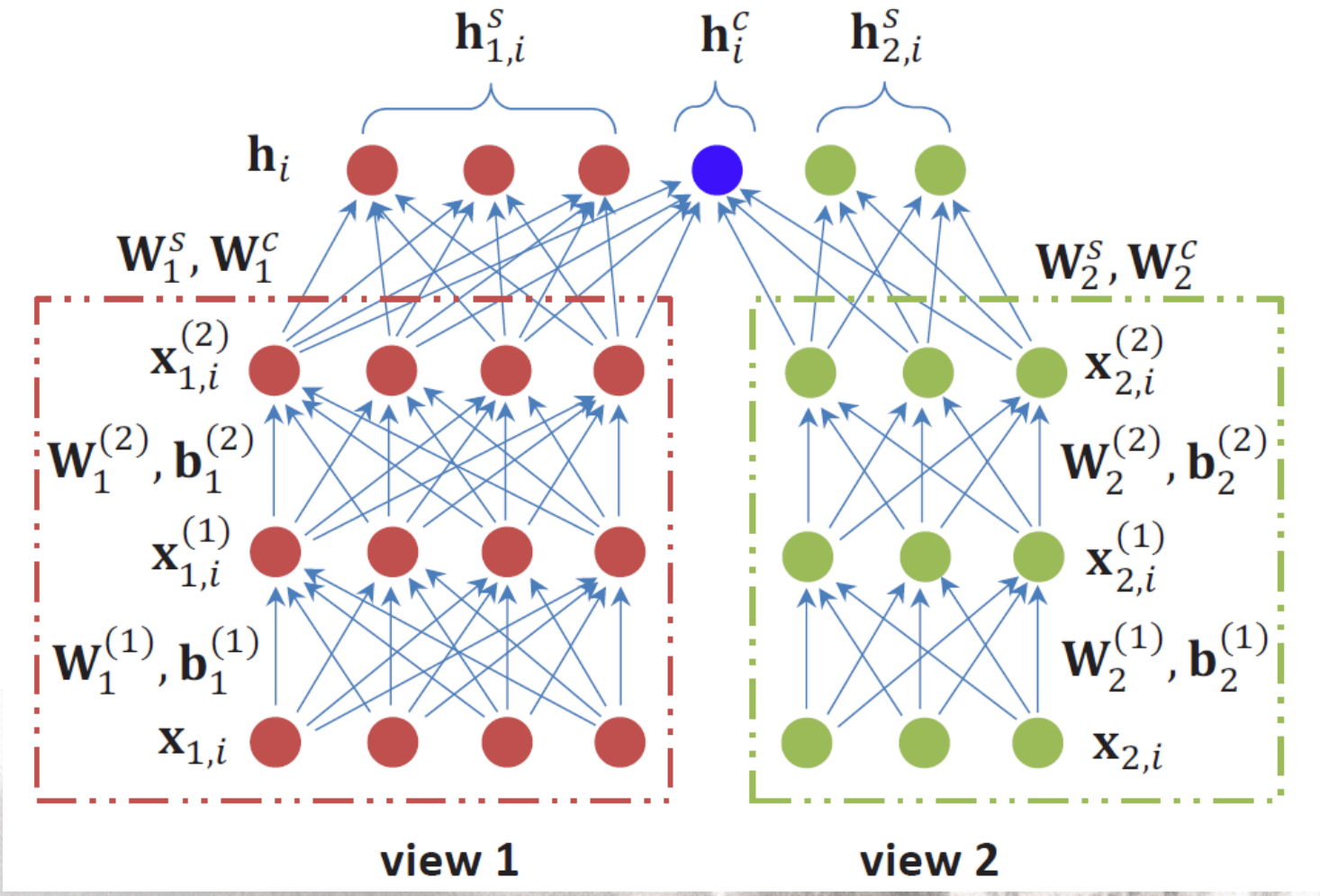
Deep Metric Learning

Adversarial Deep Metric Learning

Method	Control	Handheld
PittPatt	48.00	38.00
DeepO2P [21]	68.76	60.14
VGGFace	78.82	68.24
SPDNet [16]	80.12	72.83
GrNet [19]	80.52	72.76
CNN	90.78	78.67
Random+CNN	89.12	78.03
Hierarchical Pooling	89.83	78.23
DAN	92.06	80.33

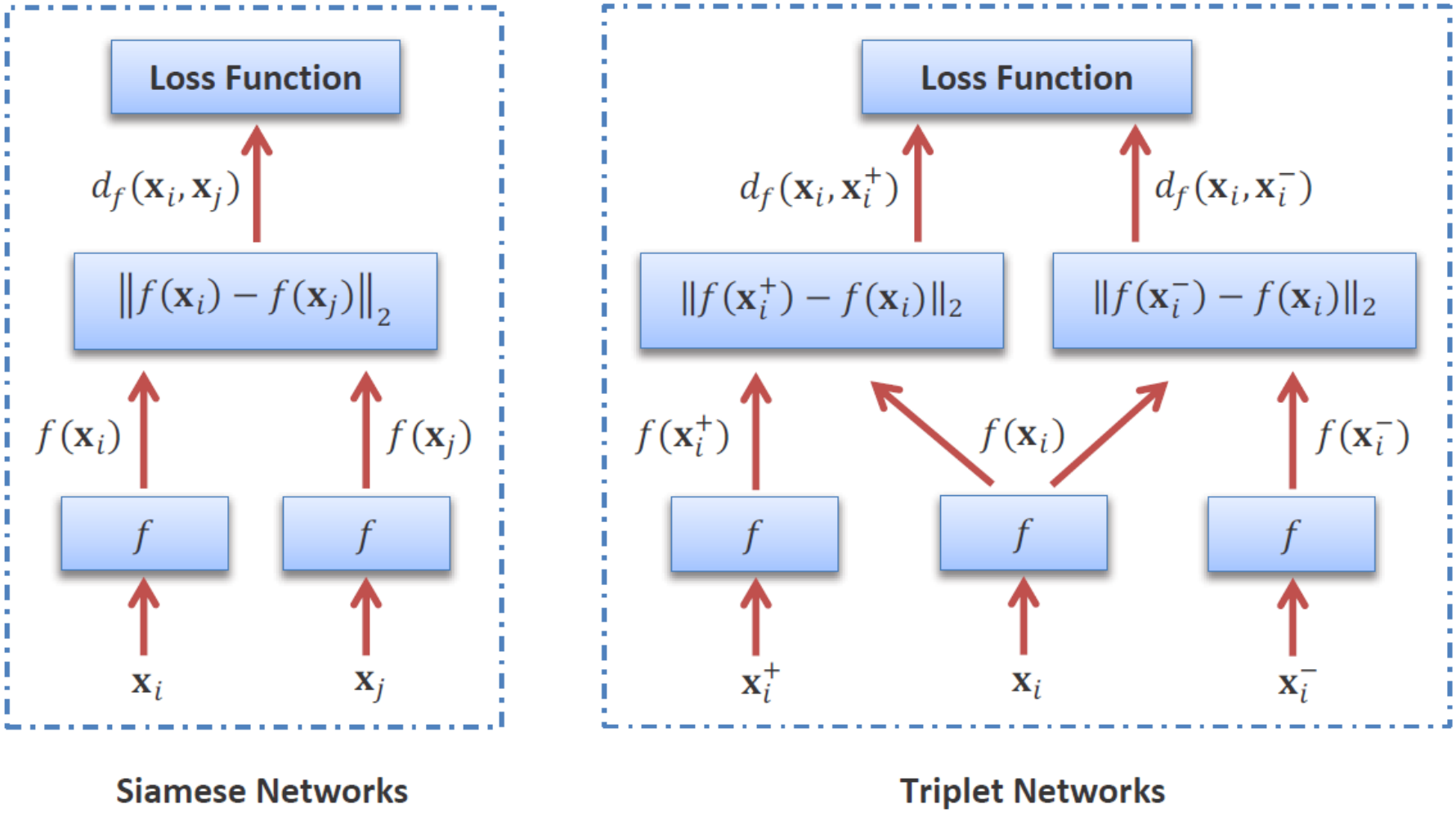
Deep Metric Learning

Multi-view Deep Metric Learning



[10] Junlin Hu, **Jiwen Lu**, and Yap-Peng Tan, Sharable and individual multi-view metric learning, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2017, accepted.

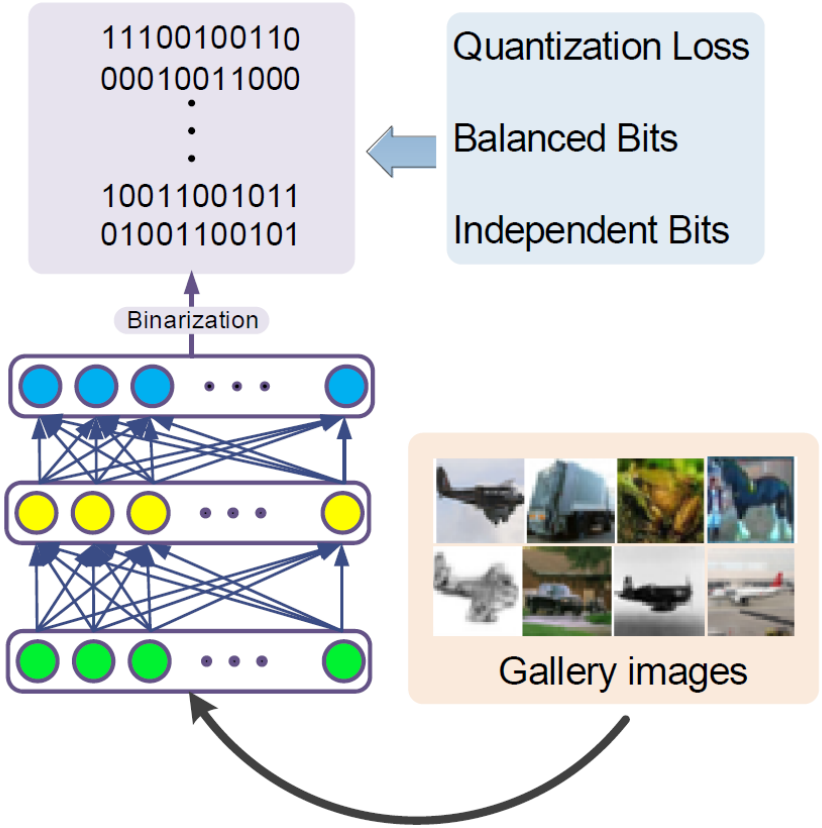
Deep Metric Learning



[11] **Jiwen Lu**, Junlin Hu, and Jie Zhou, Deep metric learning for visual understanding: an overview of recent advances, *IEEE Signal Processing Magazine*, vol. 34, no. 6, pp. 76-84, 2017.

Deep Hashing

Scalable Image Search



[12] Venice Erin Liong, **Jiwen Lu**, Gang Wang, Pierre Moulin, and Jie Zhou, Deep hashing for compact binary codes learning, *CVPR*, pp. 2475-2483, 2015.

[13] **Jiwen Lu**, Venice Erin Liong, and Jie Zhou, Deep hashing for scalable image search, *IEEE Trans. on Image Processing*, vol. 26, no. 5, pp. 2352-2367, 2017.

Deep Hashing

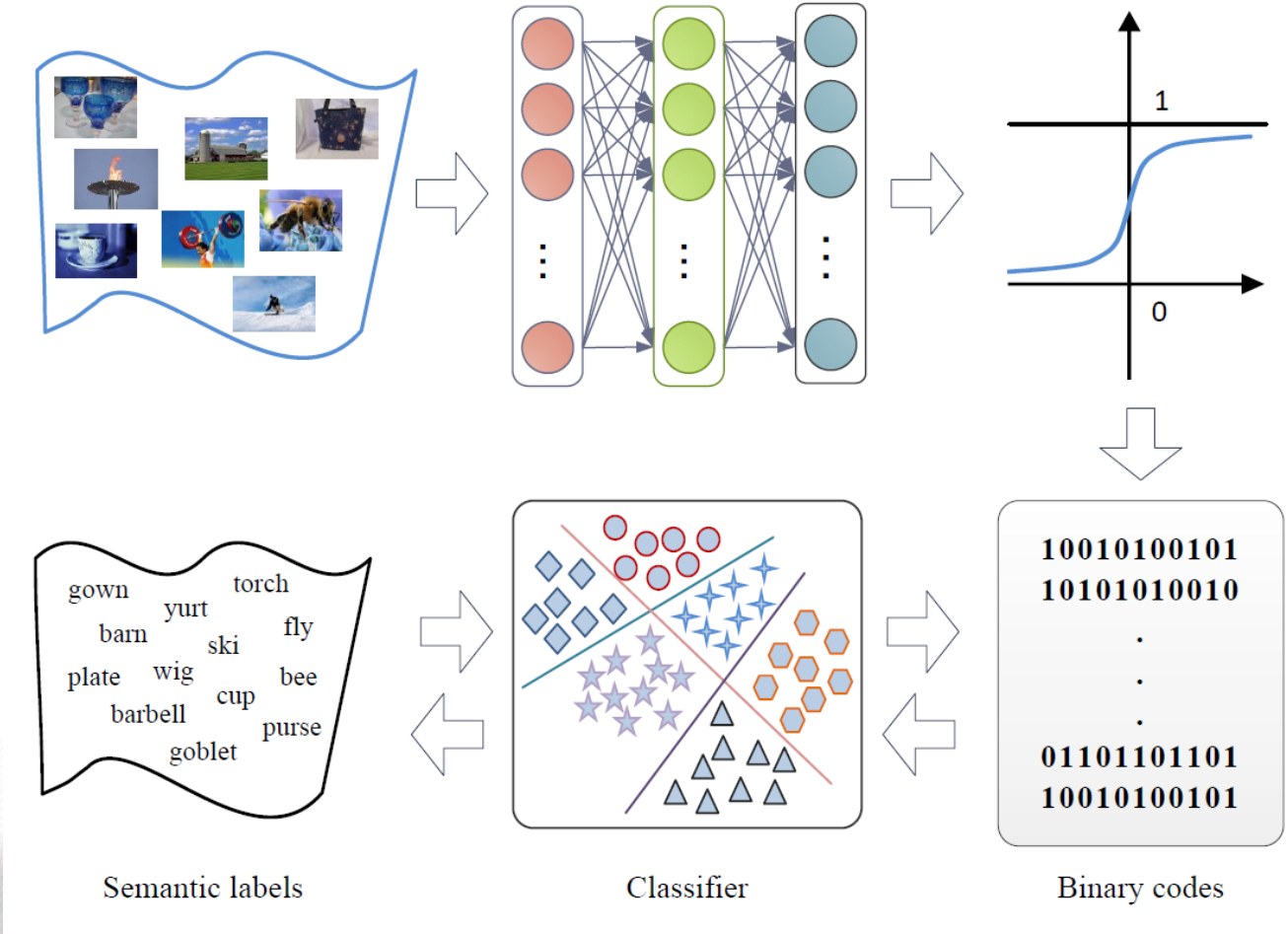
Scalable Image Search

Method	Hamming ranking (mAP, %)			precision (%) @ sample = 500			precision (%) @ r=2	
	16	32	64	16	32	64	16	32
PCA-ITQ [6]	15.67	16.20	16.64	22.46	25.30	27.09	22.60	14.99
KMH [8]	13.59	13.93	14.46	20.28	21.97	22.80	22.08	5.72
Spherical [9]	13.98	14.58	15.38	20.13	22.33	25.19	20.96	12.50
SH [35]	12.55	12.42	12.56	18.83	19.72	20.16	18.52	**20.60**
PCAH [33]	12.91	12.60	12.10	18.89	19.35	18.73	21.29	2.68
LSH [1]	12.55	13.76	15.07	16.21	19.10	22.25	16.73	7.07
DH	**16.17**	**16.62**	**16.96**	**23.79**	**26.00**	**27.70**	**23.33**	15.77
SPLH [33]	17.61	20.20	20.98	25.32	29.43	32.22	23.05	30.47
MLH [21]	18.37	20.49	21.89	24.43	29.60	33.01	23.52	28.72
BRE [15]	14.42	15.14	15.88	20.68	22.86	25.14	20.89	20.29
SDH	**18.80**	**20.83**	**22.51**	**26.32**	**30.42**	**33.60**	**23.26**	**31.48**

Results on CIFAR.

Deep Hashing

Scalable Image Search



[14] Zhixiang Chen, **Jiwen Lu**, Jianjiang Feng, and Jie Zhou, Nonlinear discrete hashing, *IEEE Trans. on* Multimedia, vol. 19, no. 1, pp. 123-135, 2017.

Deep Hashing

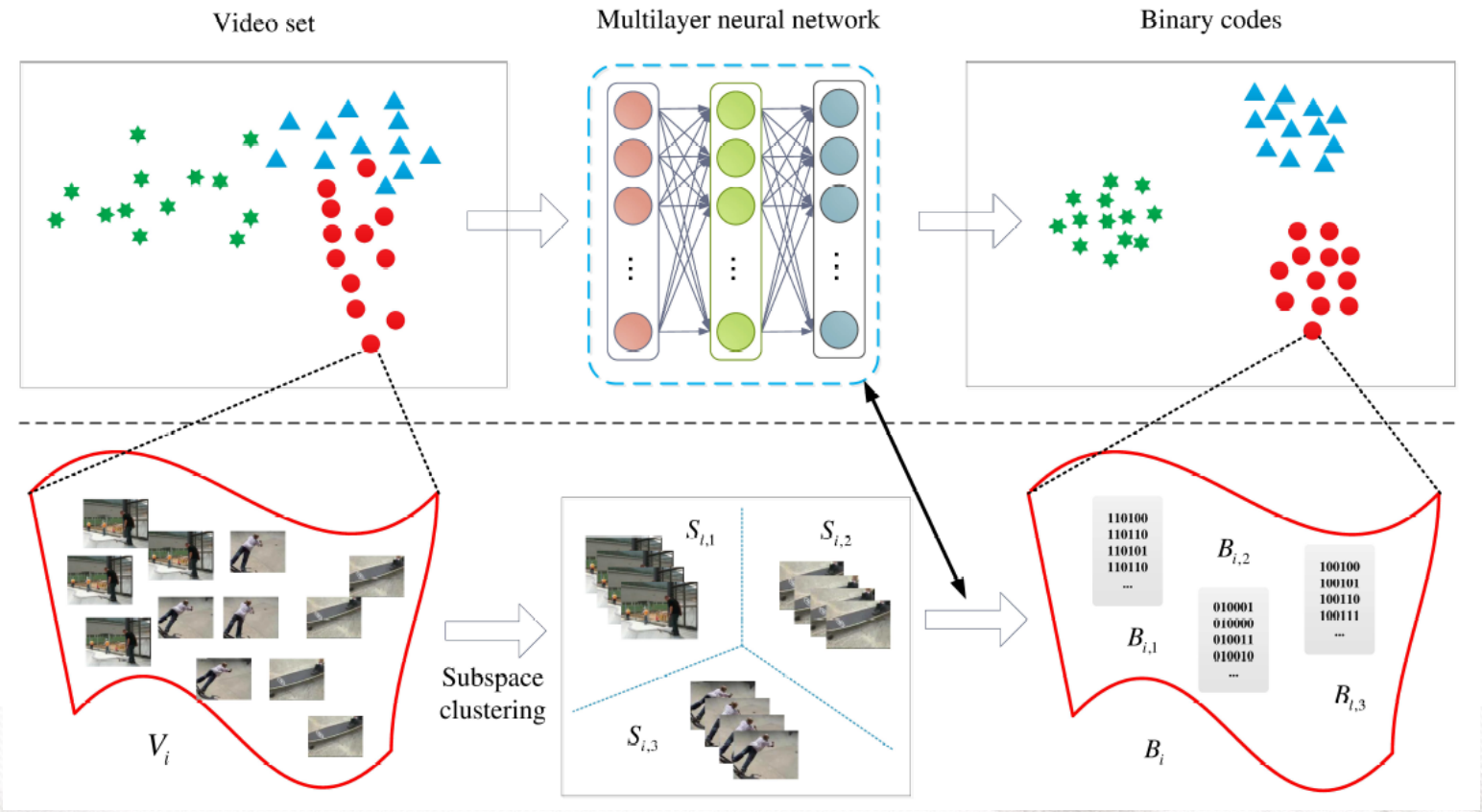
Scalable Image Search

Methods	Mean average precision(%)			Precision@500(%)			Precision@(radius==2)(%)		
	16	32	64	16	32	64	16	32	64
LSH [1]	12.63	13.70	14.62	15.32	17.23	19.36	16.67	6.35	0.1
SMLSH [8]	14.96	16.41	16.98	17.82	19.75	20.36	18.28	14.65	4.03
ITQ [10]	15.57	15.80	16.57	19.91	21.04	22.53	22.89	15.66	1.44
SPLH [18]	17.08	19.38	21.21	21.22	26.39	29.34	16.70	27.17	30.02
CCA-ITQ [10]	16.21	16.02	16.49	24.63	24.44	26.77	21.45	28.22	26.47
FastH [23]	27.94	33.09	36.55	37.74	43.13	46.84	37.76	34.42	11.64
SDH [24]	29.21	29.22	32.67	39.08	39.62	42.15	30.19	36.90	38.98
DeepH [25]	24.04	25.96	27.53	32.45	34.99	36.85	33.25	37.42	25.43
NDH	33.75	35.93	37.90	43.58	46.67	48.24	36.10	43.62	32.32

Results on CIFAR.

Deep Hashing

Scalable Video Search



[15] Zhixiang Chen, **Jiwen Lu**, Jianjiang Feng, and Jie Zhou, Nonlinear structural hashing for scalable video search, *IEEE Trans. on Circuits and Systems for Video Technology*, 2017, accepted.

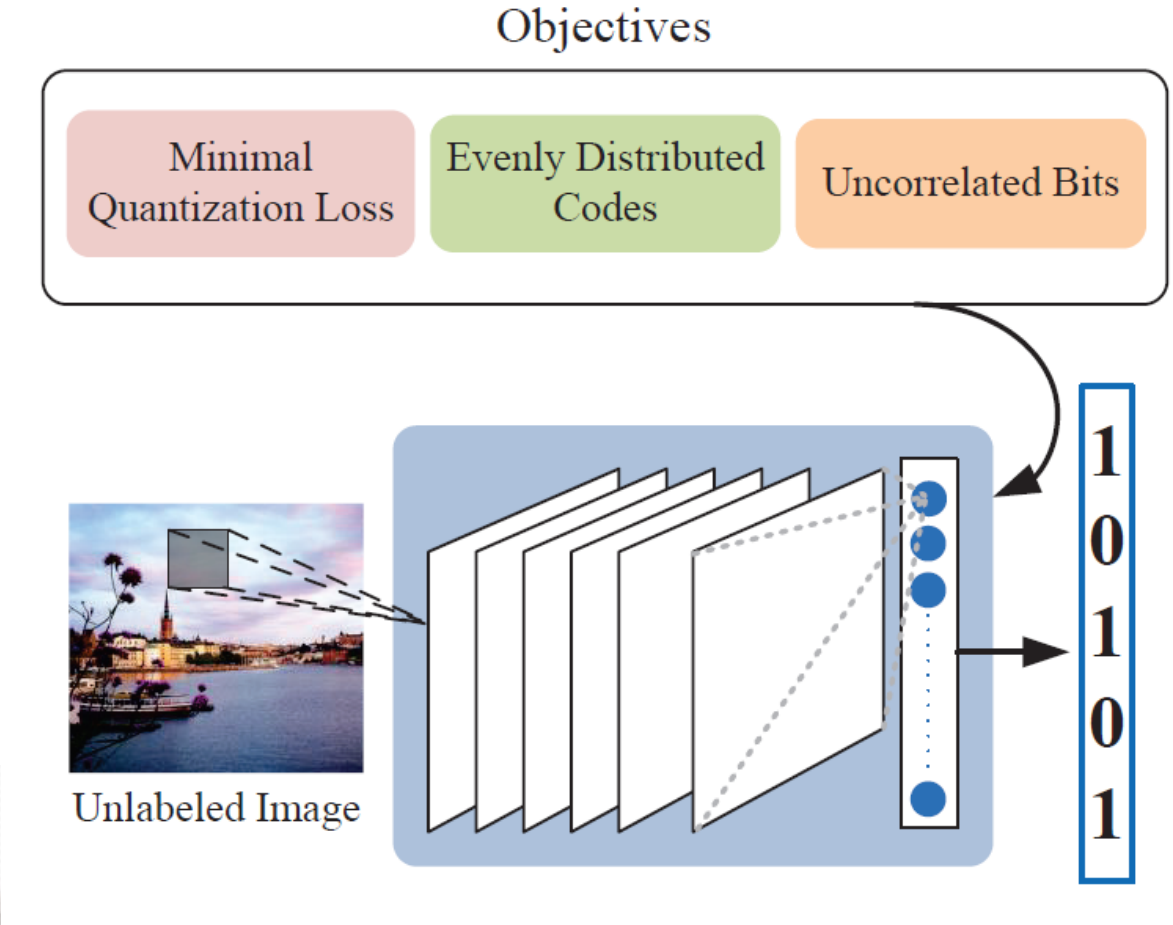
Deep Hashing

Scalable Video Search

Methods	length of binary code			
	16	32	64	96
ITQ [15]	4.63	5.12	4.94	5.41
DeepH [22]	4.42	4.95	5.53	5.99
CCA-ITQ [15]	19.95	19.99	19.95	19.79
KSH [31]	16.38	17.15	16.52	16.66
SDeepH [22]	32.43	33.65	32.42	32.77
SDH [33]	92.75	97.13	97.49	95.99
NSH	18.71	19.22	19.03	19.18

Deep Hashing

Image Matching



Main procedure of our proposed approach.

[16] Kevin Lin, **Jiwen Lu**, Chu-Song Chen, and Jie Zhou, Learning compact binary descriptors with unsupervised deep neural networks, *CVPR*, pp. 1183-1192, 2016.

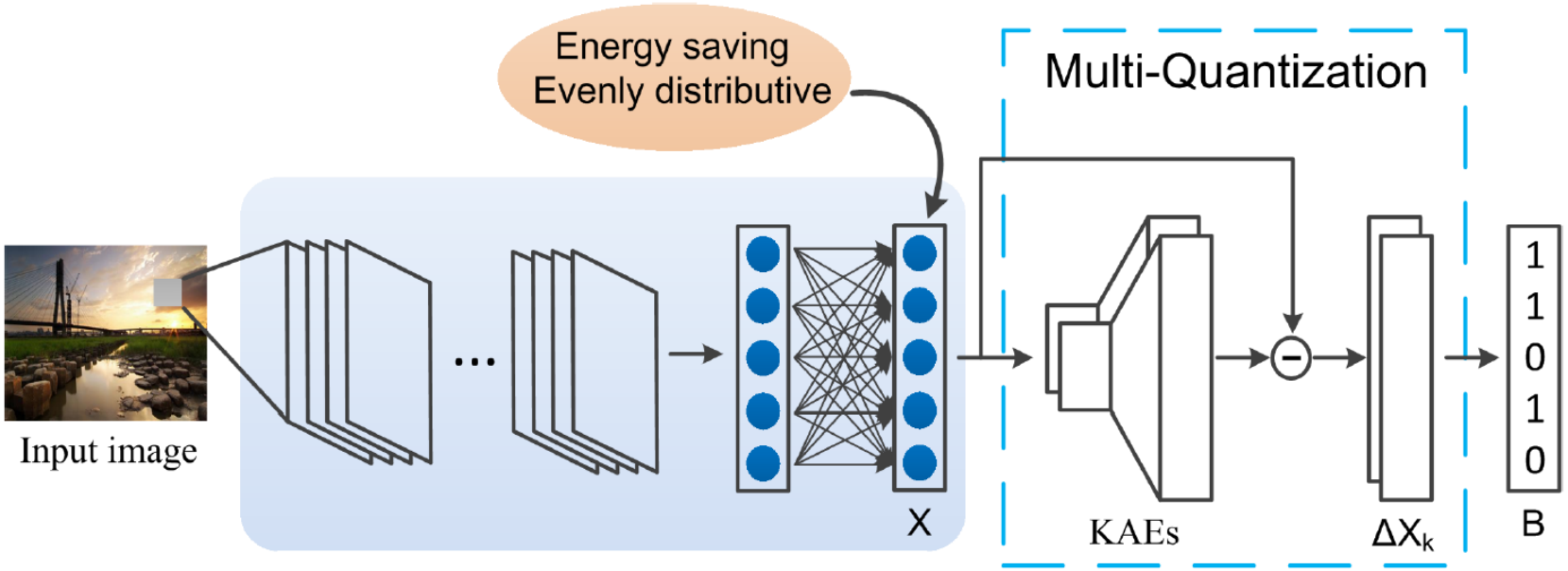
Deep Hashing

Image Matching

Train	Test	Real-valued	Binary						
		SIFT [26] 128 bytes	Boosted SSC [35] 16 bytes	BRISK [22] 64 bytes	ORB [33] 32 bytes	BRIEF [6] 32 bytes	LDAHash [38] 16 bytes	D-BRIEF [41] 4 bytes	DeepBit 32 bytes
Yosemite	Notredame	28.09	72.20	74.88	54.57	54.57	51.58	43.96	**29.60**
Yosemite	Liberty	36.27	71.59	79.36	59.15	59.15	49.66	53.39	**34.41**
Notredame	Yosemite	29.15	76.00	73.21	54.96	54.96	52.95	**46.22**	63.68
Notredame	Liberty	36.27	70.35	79.36	59.15	59.15	49.66	51.30	**32.06**
Liberty	Notredame	28.09	72.95	74.88	54.57	54.57	51.58	43.10	**26.66**
Liberty	Yosemite	29.15	77.99	73.21	54.96	54.96	52.95	**47.29**	57.61
Average 95% ERR		31.17	73.51	75.81	56.23	56.23	51.40	47.54	**40.67**

Deep Hashing

Image Matching



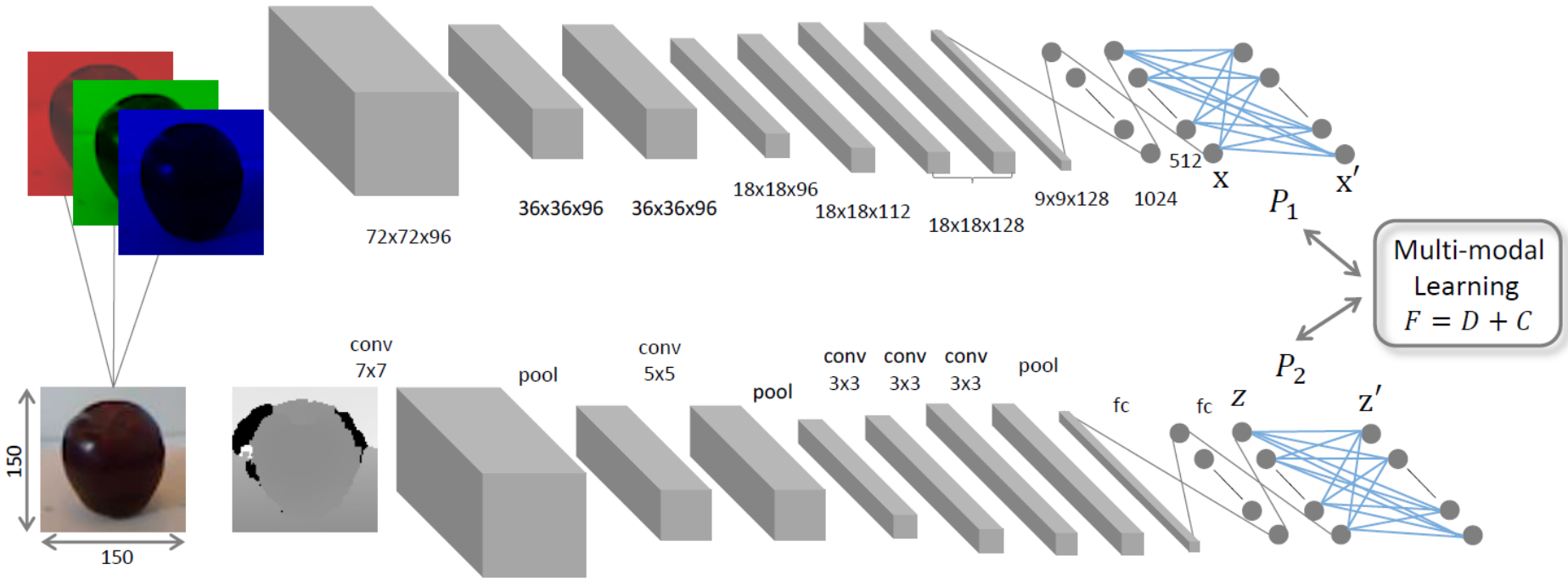
[17] Yueqi Duan, **Jiwen Lu**, Ziwei Wang, Jianjiang Feng, and Jie Zhou, Learning deep binary descriptor with multi-quantization, *CVPR*, pp. 1183-1192, 2017.

Deep Hashing

Image Matching

Train Test	Yosemite Noter Dame	Yosemite Liberty	Notre Dame Yosemite	Notre Dame Liberty	Liberty Notre Dame	Liberty Yosemite	Average ERR
SIFT [27] (128 bytes)	28.09	36.27	29.15	36.27	28.09	29.15	31.17
Boosted SSC [40] (16 bytes)	72.20	71.59	76.00	70.35	72.95	77.99	73.51
BRISK [25] (64 bytes)	74.88	79.36	73.21	79.36	74.88	73.21	75.81
BRIEF [6] (32 bytes)	54.57	59.15	**54.96**	59.15	54.57	**54.96**	56.23
DeepBit [26] (32 bytes)	29.60	34.41	63.68	32.06	26.66	57.61	40.67
LDAHash [42] (16 bytes)	51.58	49.66	52.95	49.66	51.58	52.95	51.40
D-BRIEF [47] (4 bytes)	43.96	53.39	46.22	51.30	43.10	47.29	47.54
BinBoost [45] (8 bytes)	14.54	21.67	18.96	20.49	16.90	22.88	19.24
RFD [10] (50-70 bytes)	11.68	19.40	14.50	19.35	13.23	16.99	15.86
DBD-MQ (32 bytes)	**27.20**	**33.11**	57.24	**31.10**	**25.78**	57.15	**38.59**

Multi-Modal Deep Learning



Our proposed multi-modal deep metric learning framework.

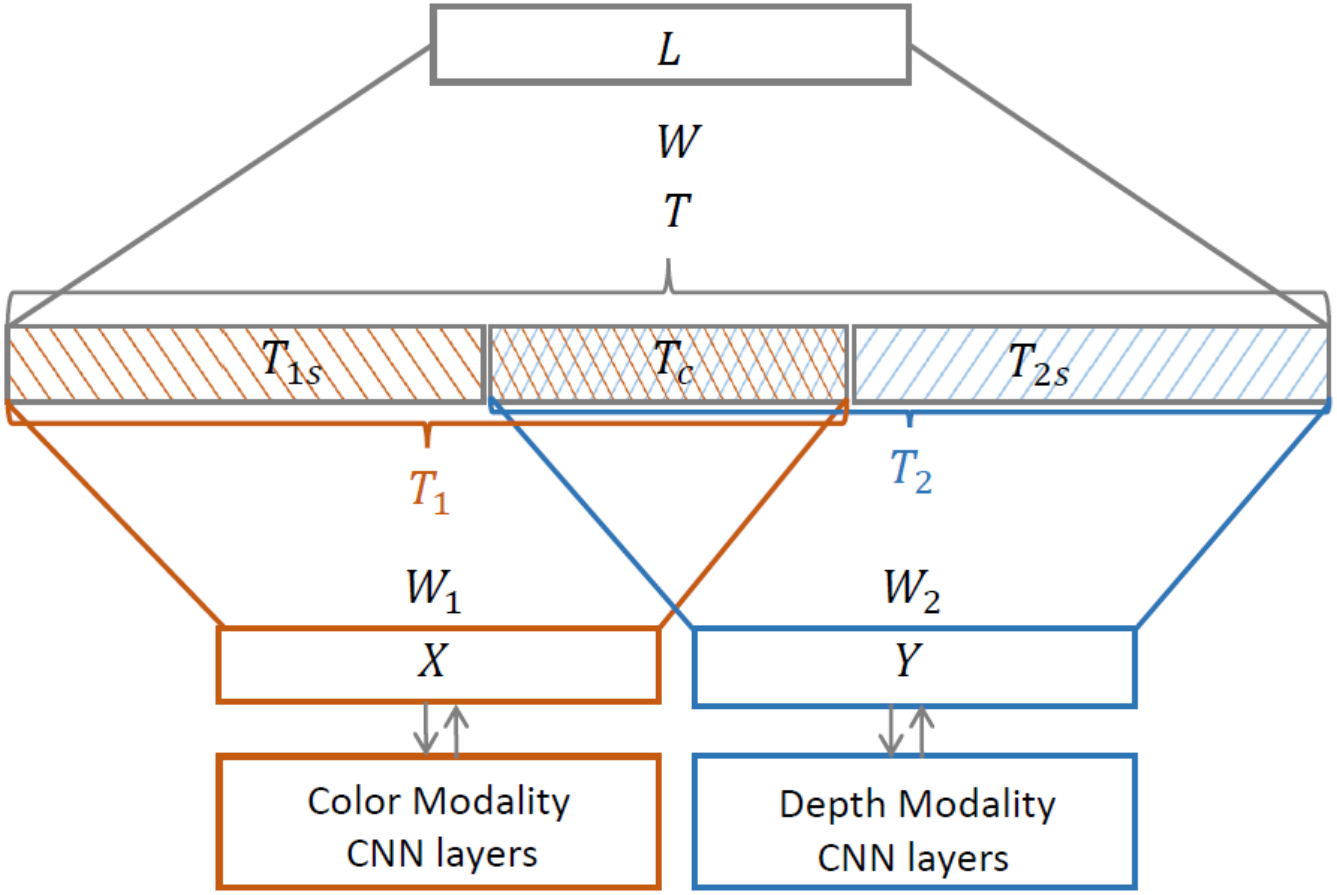
[18] Anran Wang, **Jiwen Lu**, Jianfei Cai, Tat-Jen Cham, and Gang Wang, Large-margin multi-modal deep learning for RGB-D object recognition, *IEEE Trans. on* Multimedia, vol. 17, no. 11, pp. 1887-1898, 2015.

Multi-Modal Deep Learning

RGB-D Object Recognition

Method	Accuracy (%)
Lai *et al.* [30]	81.9 ± 2.8
Blum *et al.* [4]	86.4 ± 2.3
Socher *et al.* [42]	86.8 ± 3.3
Bo *et al.* [7]	87.5 ± 2.9
Le *et al.* [31]	86.7 ± 2.7
Jhuo *et al.* [23]	89.6 ± 3.8
Ours	86.9 ± 2.6

Multi-Modal Deep Learning



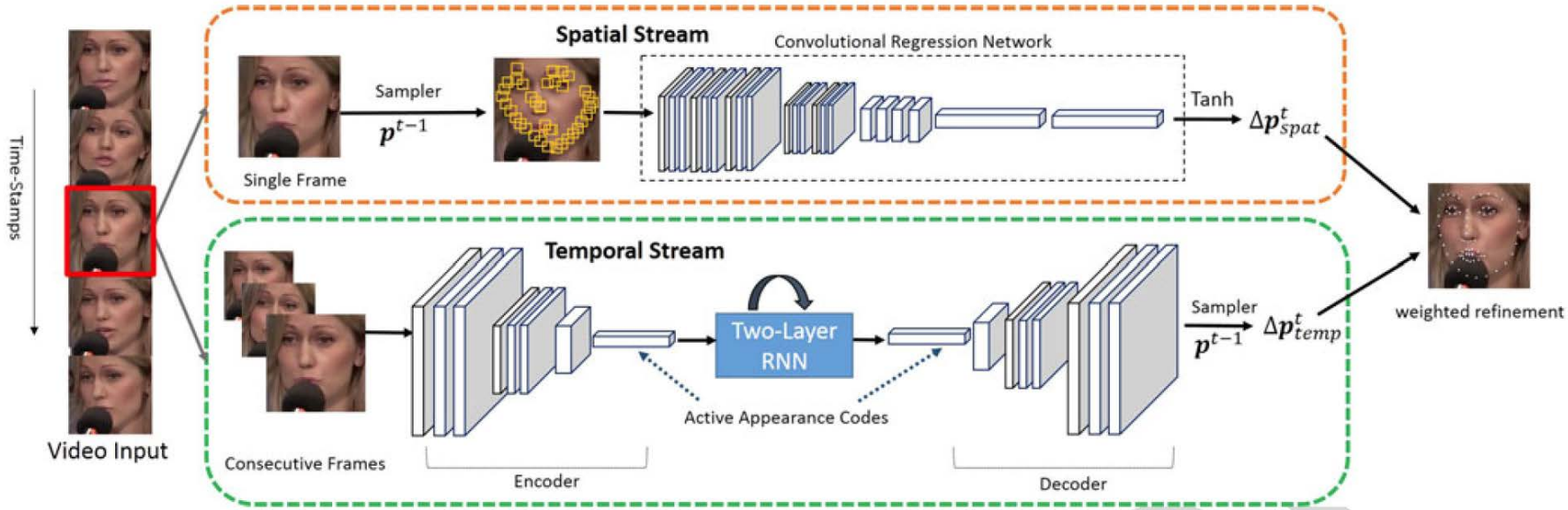
[19] Anran Wang, Jianfei Cai, **Jiwen Lu**, and Tat-Jen Cham, MMSS: Multi-modal sharable and specific feature learning for RGB-D object recognition, *ICCV*, pp. 1125-1133, 2015.

Multi-Modal Deep Learning

RGB-D Object Recognition

Method	Category (%)	Instance (%)
Lai *et al.* [18]	81.9 ± 2.8	73.9
Blum *et al.* [4]	86.4 ± 2.3	90.4
Socher *et al.* [29]	86.8 ± 3.3	-
Zhang *et al.* [34]	-	86.6
Bo *et al.* [6]	87.5 ± 2.9	92.8
Ours	88.5 ± 2.2	94.0

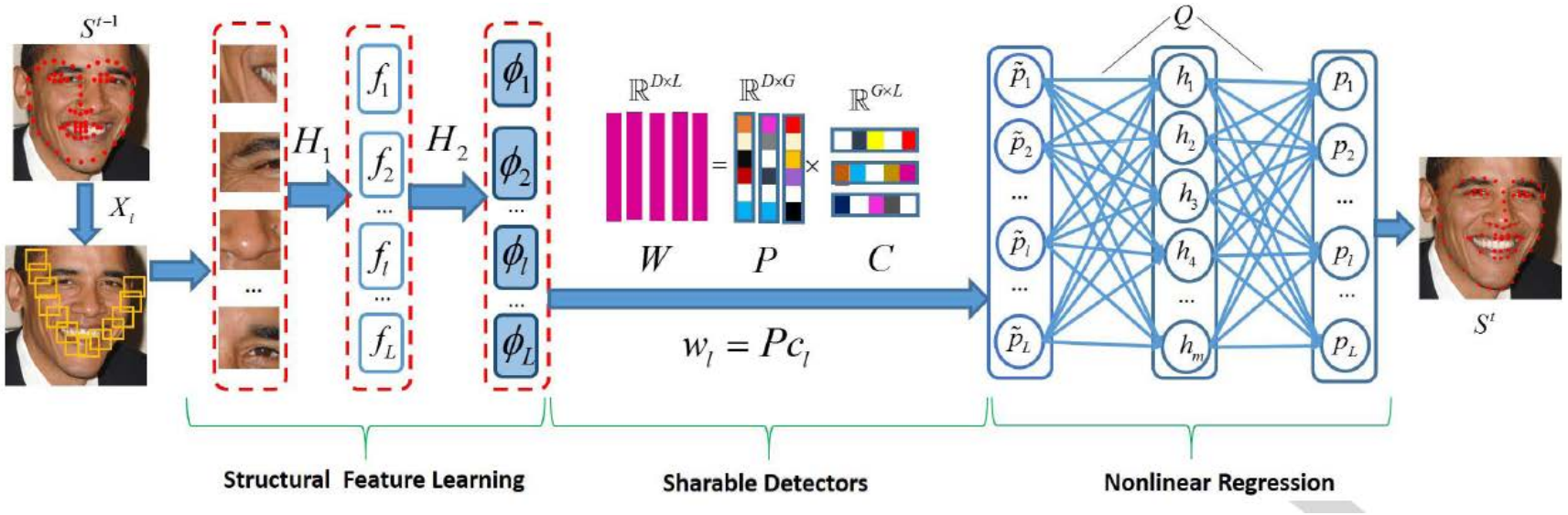
Multi-Modal Deep Learning



[20] Hao Liu, **Jiwen Lu**, Jianjiang Feng, and Jie Zhou, Two-stream transferable networks for video-based face alignment, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2017, accepted.

Deep Sharable Learning

Face Alignment



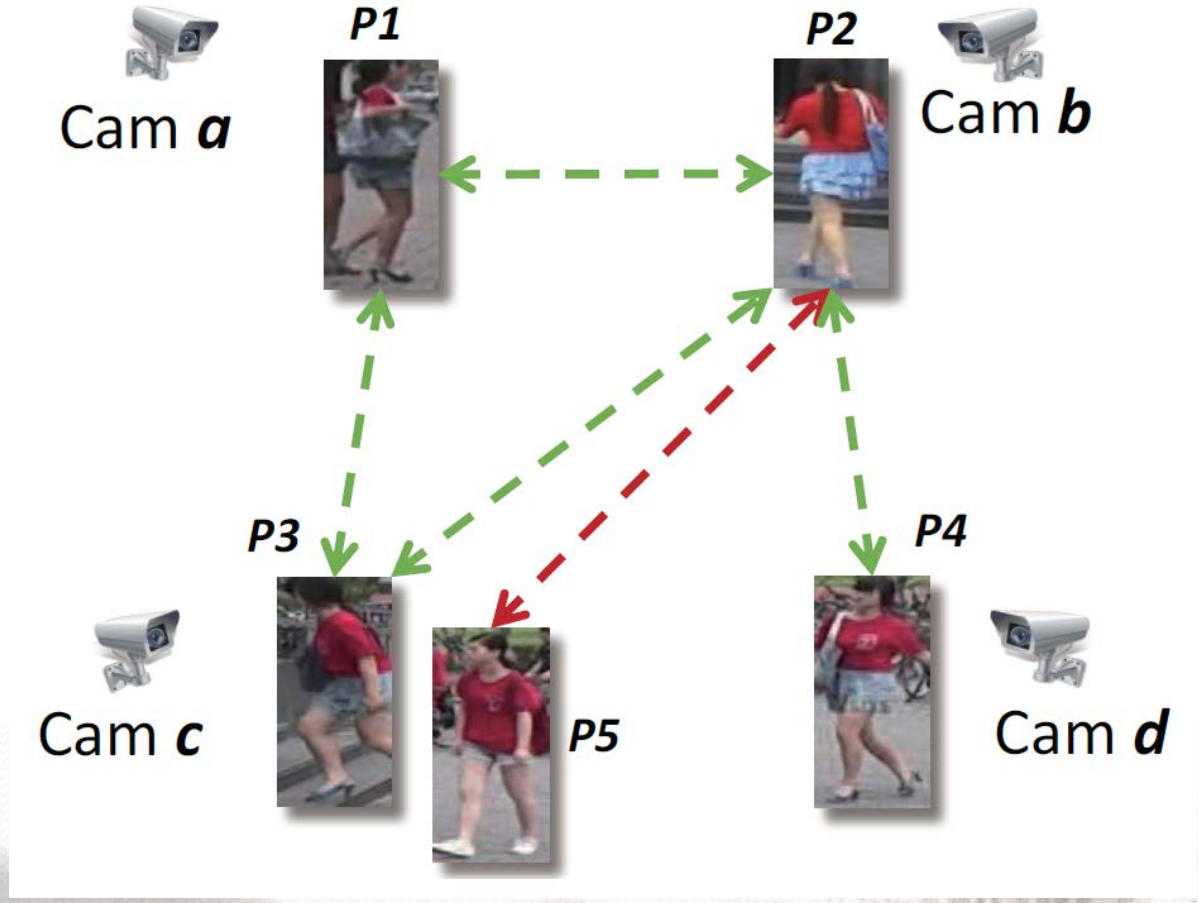
[21] Hao Liu, **Jiwen Lu**, Jianjiang Feng, and Jie Zhou, Learning deep sharable and structural detectors for face alignment, *IEEE Trans. on Image Processing*, vol. 26, no. 4, pp. 1666-1678, 2017.

Deep Sharable Learning

Face Alignment

Deep Sharable Learning

Multi-Camera Person Re-identification



[22] Ji Lin, Liangliang Ren, **Jiwen Lu**, Jianjiang Feng, and Jie Zhou, Consistent-aware deep learning for person re-identification in a camera network, *CVPR*, pp. 5771-5780, 2017.

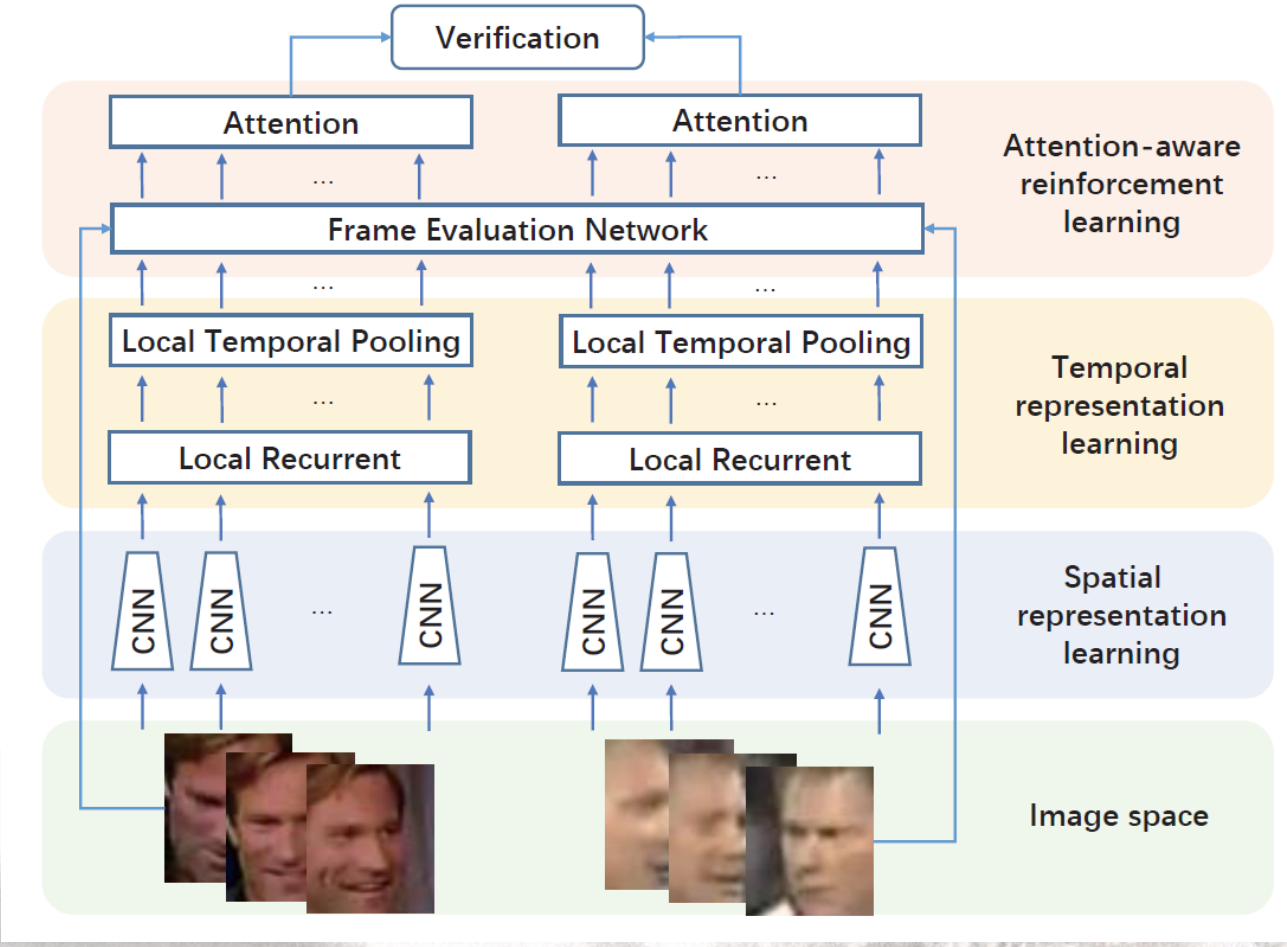
Deep Sharable Learning

Multi-Camera Person Re-identification

Method	Weight Acc.	Var.
BoW [47] - (SQ)	21.14	0.0321
Ours - Pretrained - (SQ)	27.29	0.0221
Ours - Contrastive - (SQ)	46.23	0.0084
Ours - Cosine - (SQ)	52.88	0.0049
Ours - CADL - (SQ)	**60.09**	**0.0111**
BoW [47] - (MQ)	27.76	0.0258
Ours - Pretrained - (MQ)	40.51	0.0189
Ours - Contrastive - (MQ)	59.16	0.0168
Ours - Cosine - (MQ)	72.02	0.0043
Ours - CADL - (MQ)	**81.15**	**0.0039**

Deep Reinforcement Learning

Video Face Recognition



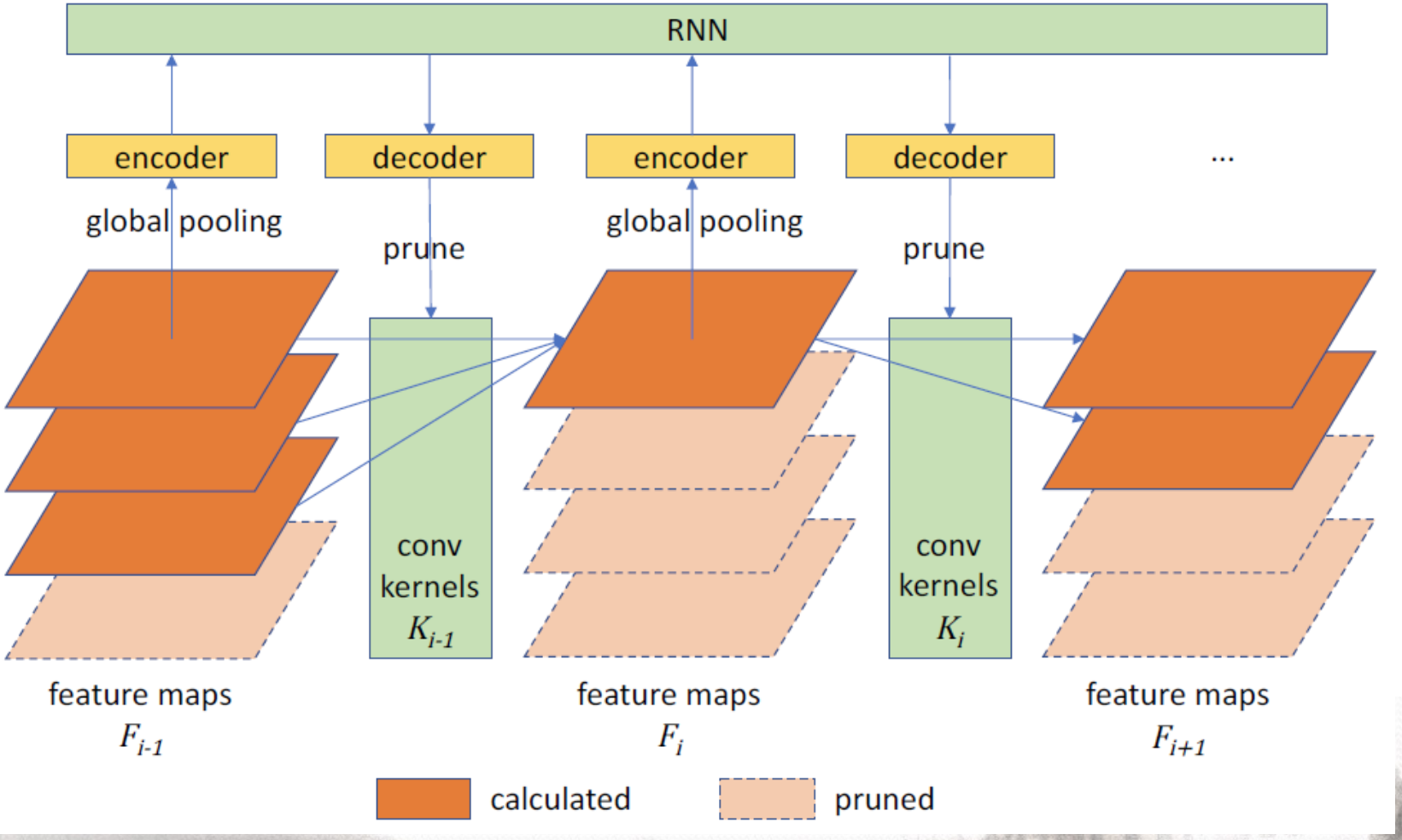
[23] Yongming Rao, **Jiwen Lu**, and Jie Zhou, Attention-aware deep reinforcement learning for video face recognition, *ICCV*, pp. 3931-3940, 2017.

Deep Reinforcement Learning

Video Face Recognition

Method	Accuracy	Year
LM3L [16]	81.3 ± 1.2	2014
DDML [15]	82.3 ± 1.2	2014
EigenPEP [24]	84.8 ± 1.4	2014
DeepFace-single [35]	91.4 ± 1.1	2015
DeepID2+ [34]	93.2 ± 0.2	2015
FaceNet [32]	95.12 ± 0.39	2015
Deep FR [31]	97.3	2015
NAN [43]	95.72 ± 0.64	2016
Wen *et al.* [40]	94.9	2016
TBE-CNN [10]	94.96 ± 0.31	2017
ADRL	95.96 ± 0.59	
ADRL-finetune	96.52 ± 0.54	

Deep Network Compression



[24] Ji Lin, Yongming Rao, **Jiwen Lu**, and Jie Zhou, Runtime neural pruning, *NIPS*, 2017.

Summary and Future Work

- Deep learning is very effective for many visual understanding tasks including visual recognition, visual tracking, and visual search.
- More visual cues can be exploited to help develop more elegant deep learning methods for visual understanding.
- Theoretical analysis for deep learning to better understand the principles is required to show how it improves various visual understanding tasks.

Thank you!