IAPR ICPR 2018

INTERNATIONAL CONFERENCE ON PATTERN RECOGNITION

In Beijing, China — August 20th-24th 2018

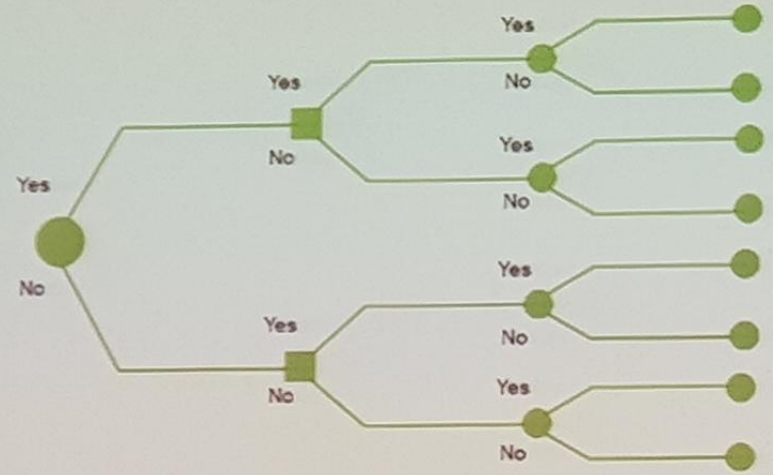
http://icpr2018.org

Non-NN Style Deep Learning

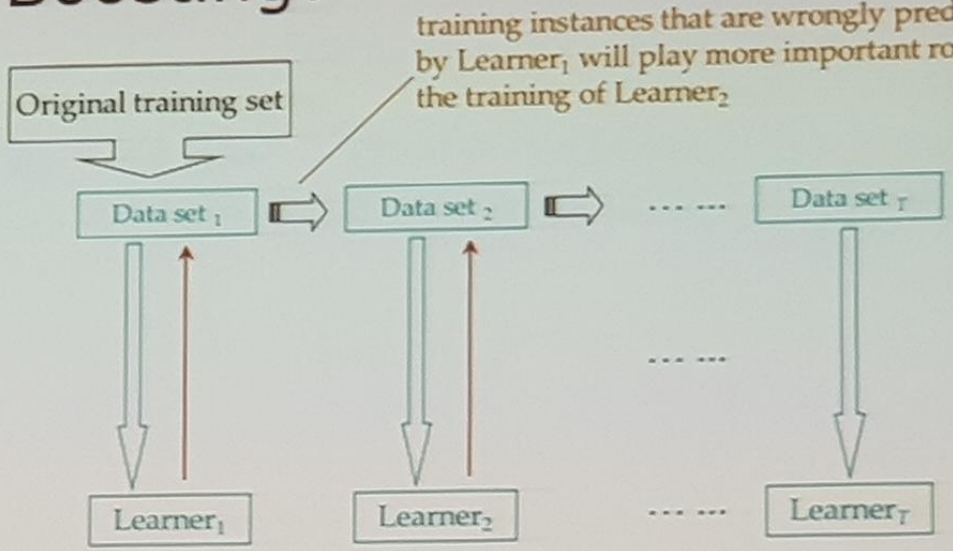
Zhi-Hua Zhou, Nanjing University

Deep Forest, https://arxiv.org/abs/1702.08835

Decision trees ?



Boosting?



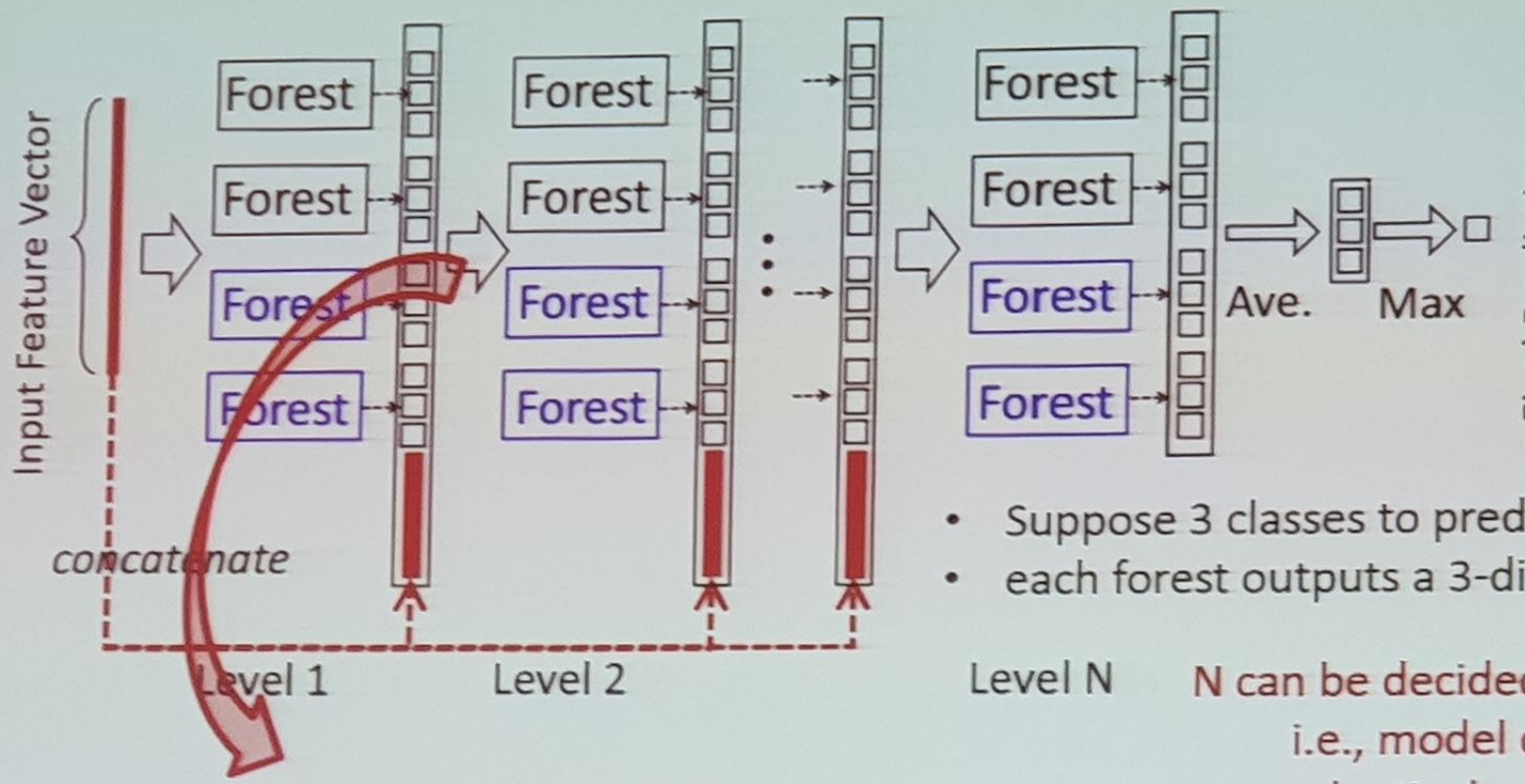
layer-by-layer processing, but ...

- **insufficient complexity**
- **always on original features**

- still, **insufficient complexity**
- **always on original features**

Cascade Forest structure

http://lamda.nju.edu



- Suppose 3 classes to predict
- each forest outputs a 3-dim **class vector**

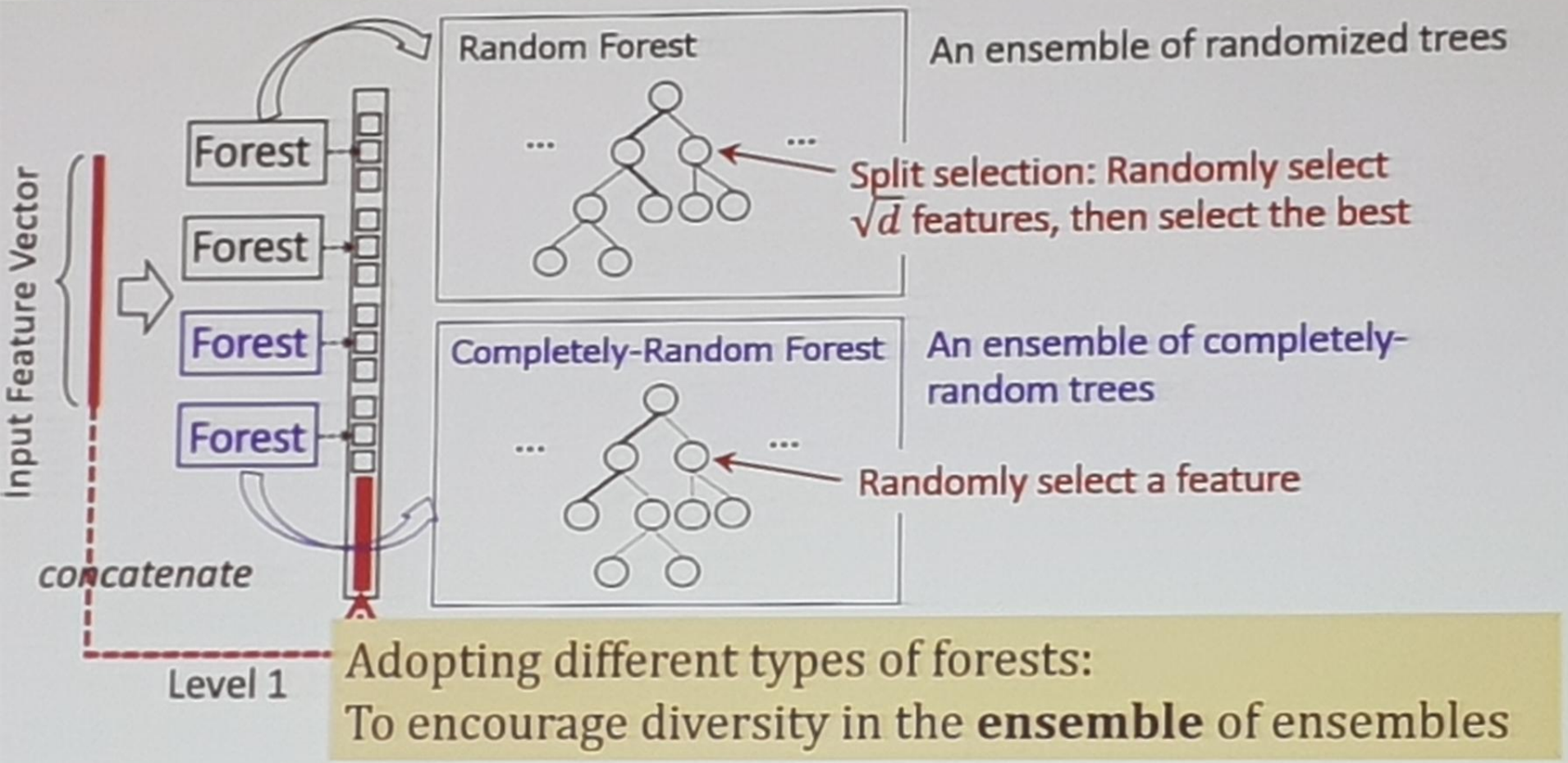
N can be decided by CV
i.e., model complexity
adaptively decided

Passing the output of one level as input to another level:

- Related to *Stacking* [Wolpert, NNJ 1992; Breiman, MLJ 1996], a famous ensemble method
- Stacking usually one or two levels, as it is easy to overfit with more than two levels; could not enable a deep model by itself

Ensemble of ensembles

http://lamda.nju.edu.cn

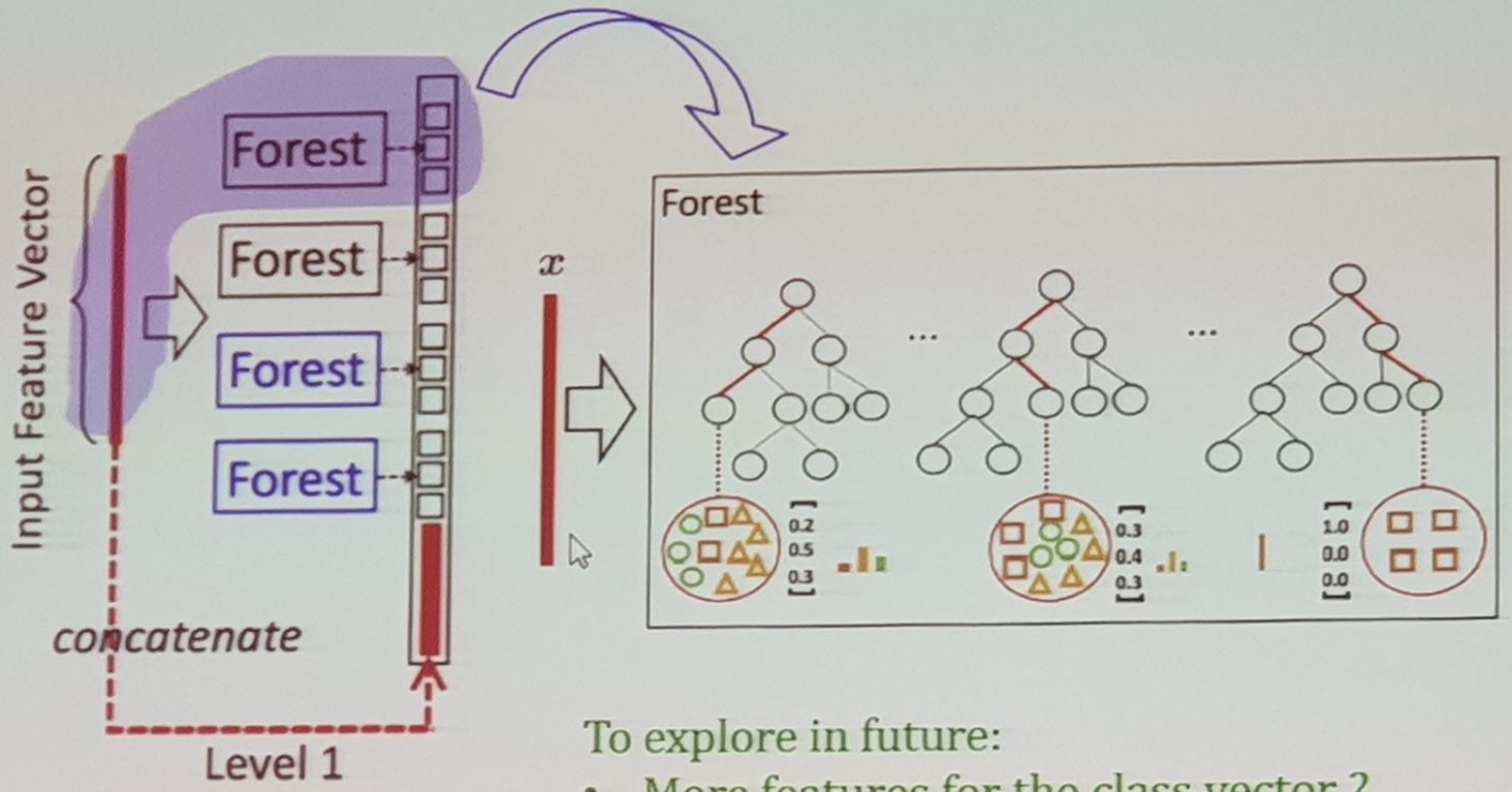


Adopting different types of forests:
To encourage diversity in the **ensemble** of ensembles

To explore in future:

- Completely-random trees also offer the possibility of using unlabeled data

Generation of class vectors



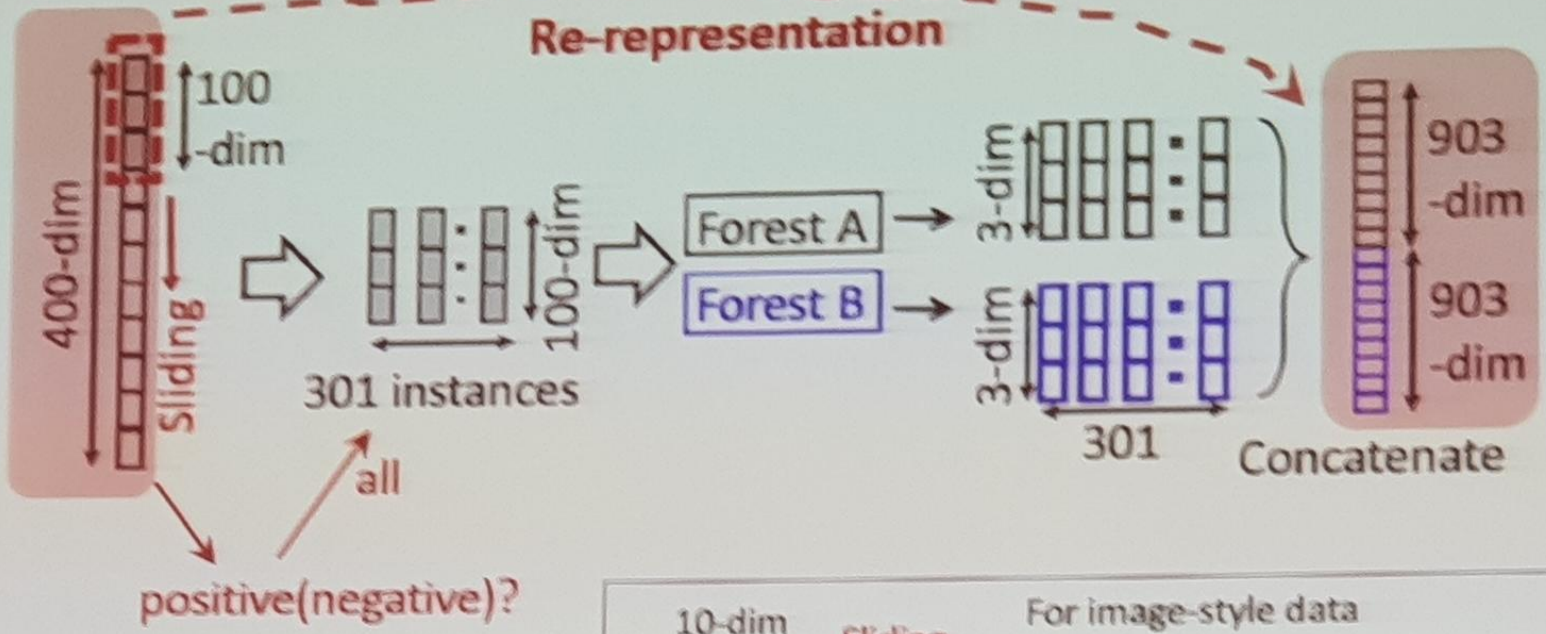
To explore in future:

- More features for the class vector ?

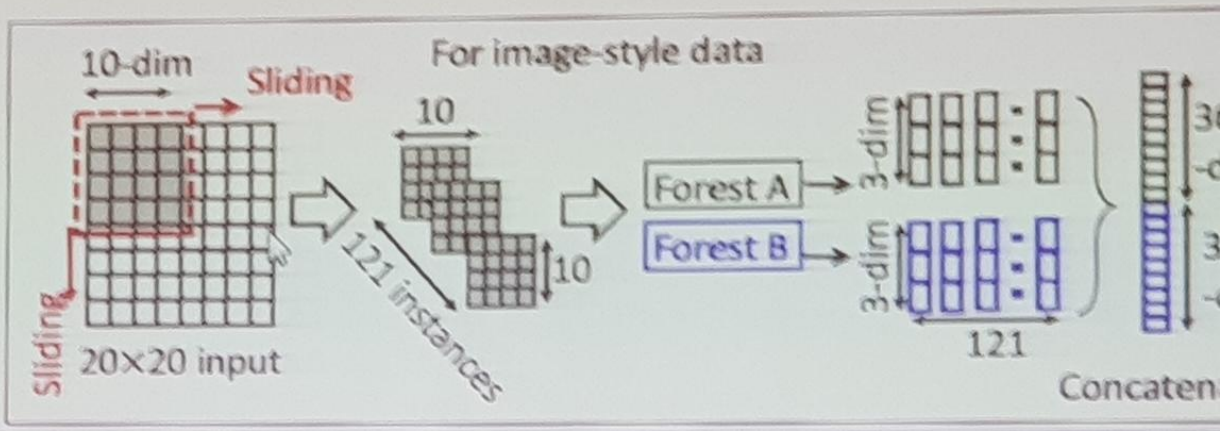
  ... such as parents nodes (prior distribution), sibling nodes (complement distribution), decision patch encoding, ...

Sliding window scanning

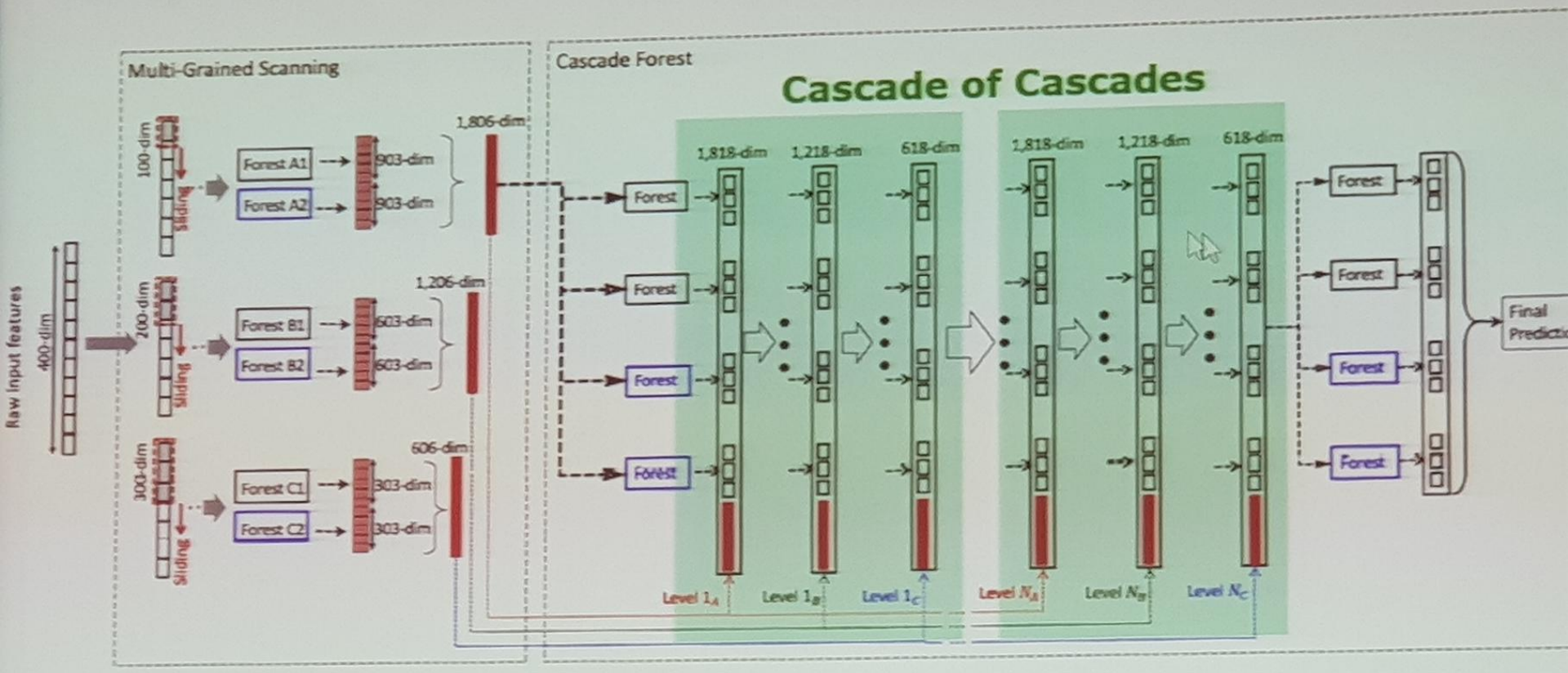
Inspired by: CNNs/RNNs exploit spatial/sequential relationships



Related to *Flipping Output* [Breiman, MLJ 2000], an output manipulation approach to encourage ensemble diversity



Overall architecture



For grained scanning:

- 500 trees per forest
- Tree growth: till pure leaf, or depth =100
- Sliding window size $\lfloor d/16 \rfloor$, $\lfloor d/8 \rfloor$, $\lfloor d/4 \rfloor$

For cascade:

- 500 trees per forest
- Tree growth: till pure leaf

Deep forest results

- Non-differentiable building blocks, not rely on BP
- Much less hyper-parameters than DNNs → easier to train
- Model complexity decided upon data → applicable to small data
- Performance competitive to DNNs on a broad range of tasks

Hyper-parameters

Table 1: Summary of hyper-parameters and default settings. Boldfont highlights hyper-parameters with relatively larger influence; "?" indicates default value unknown, or generally requiring different settings for different tasks.

Deep neural networks (e.g., convolutional neural networks)	gcForest
Type of activation functions: Sigmoid, ReLU, tanh, linear, etc.	Type of forests: Completely-random tree forest, random forest, etc.
Architecture configurations: **No. Hidden layers**: ? **No. Nodes in hidden layer**: ? **No. Feature maps**: ? **Kernel size**: ?	Forest in multi-grained scanning: **No. Forests**: {2} **No. Trees in each forest**: {500} Tree growth: till pure leaf, or reach depth 100 **Sliding window size**: {⌊d/16⌋, ⌊d/8⌋, ⌊d/4⌋} Forest in cascade: **No. Forests**: {8} **No. Trees in each forest**: {500} Tree growth: till pure leaf
Optimization configurations: **Learning rate**: ? Dropout: {0.25/0.50} **Momentum**: ? **L1/L2 weight regularization penalty**: ? Weight initialization: Uniform, glorot_normal, glorot_uniform, etc. Batch size: {32/64/128}	

In Experiments:

- **gcForest uses the same hyper-parameters for all data**
- **DNNs carefully tune per dataset**

Experimental results

Image categorization (MNIST)

gcForest	**99.26%**
LeNet-5	99.05%
Deep Belief Net	98.75% [Hinton *et al.*, 2006]
SVM (rbf kernel)	98.60%
Random Forest	96.80%

Face recognition (ORL)

	5 image	7 images	9 images
gcForest	**91.00%**	**96.67%**	**97.50%**
Random Forest	91.00%	93.33%	95.00%
CNN	86.50%	91.67%	95.00%
SVM (rbf kernel)	80.50%	82.50%	85.00%
*k*NN	76.00%	83.33%	92.50%

Experimental results

Music classification (GTZAN)

gcForest	**65.67%**
CNN	59.20%
MLP	58.00%
Random Forest	50.33%
Logistic Regression	50.00%
SVM (rbf kernel)	18.33%

Sentiment classification (IMDB)

gcForest	**89.16%**
CNN	89.02% [Kim, 2014]
MLP	88.04%
Logistic Regression	88.62%
SVM (linear kernel)	87.56%
Random Forest	85.32%

Hand movement recognition (sEMG)

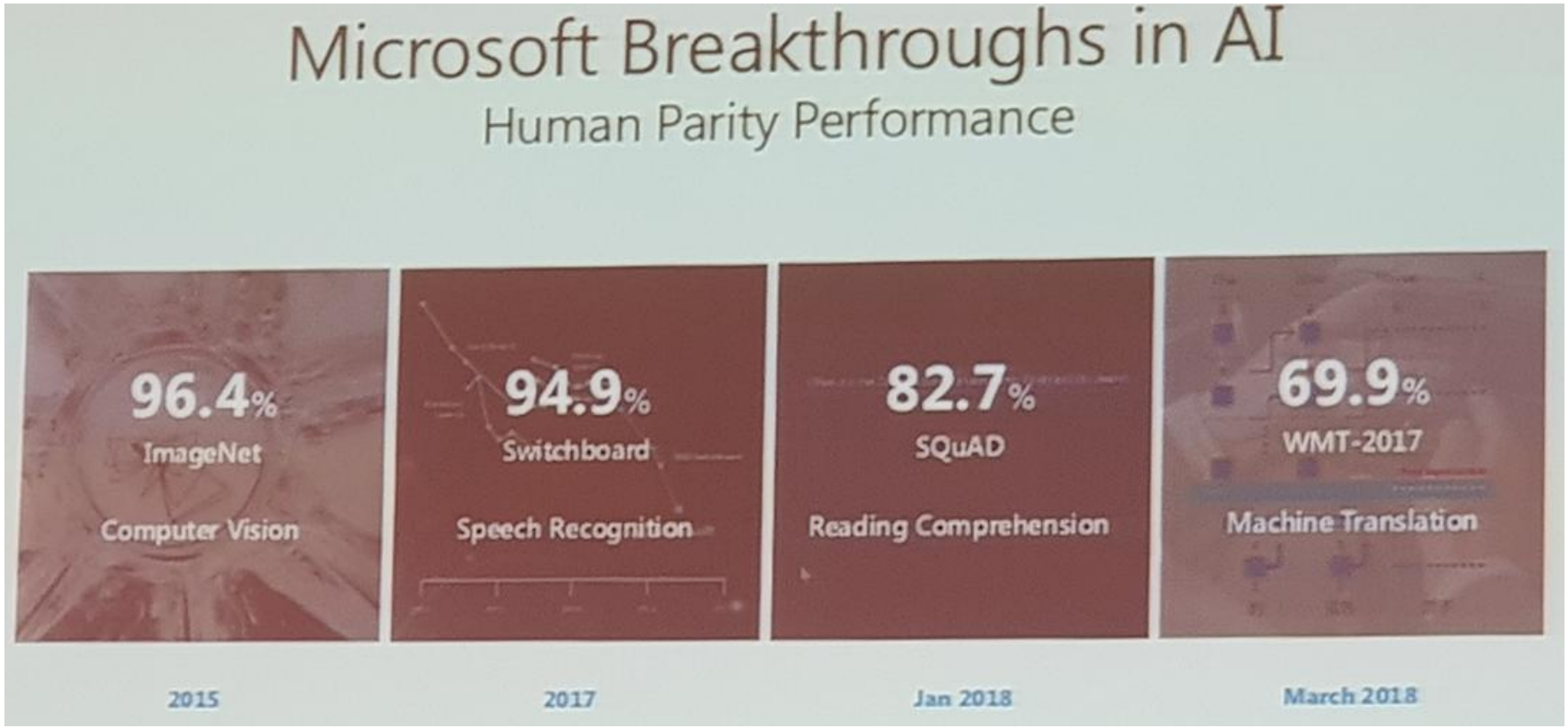
gcForest	**71.30%**
LSTM	45.37%
MLP	38.52%
Random Forest	29.62%
SVM (rbf kernel)	29.62%
Logistic Regression	23.33%

Low-dimensional data (features: 16, 14, 8)

	LETTER	ADULT	YEAST
gcForest	**97.40%**	**86.40%**	**63.45%**
Random Forest	96.50%	85.49%	61.66%
MLP	95.70%	85.25%	55.60%

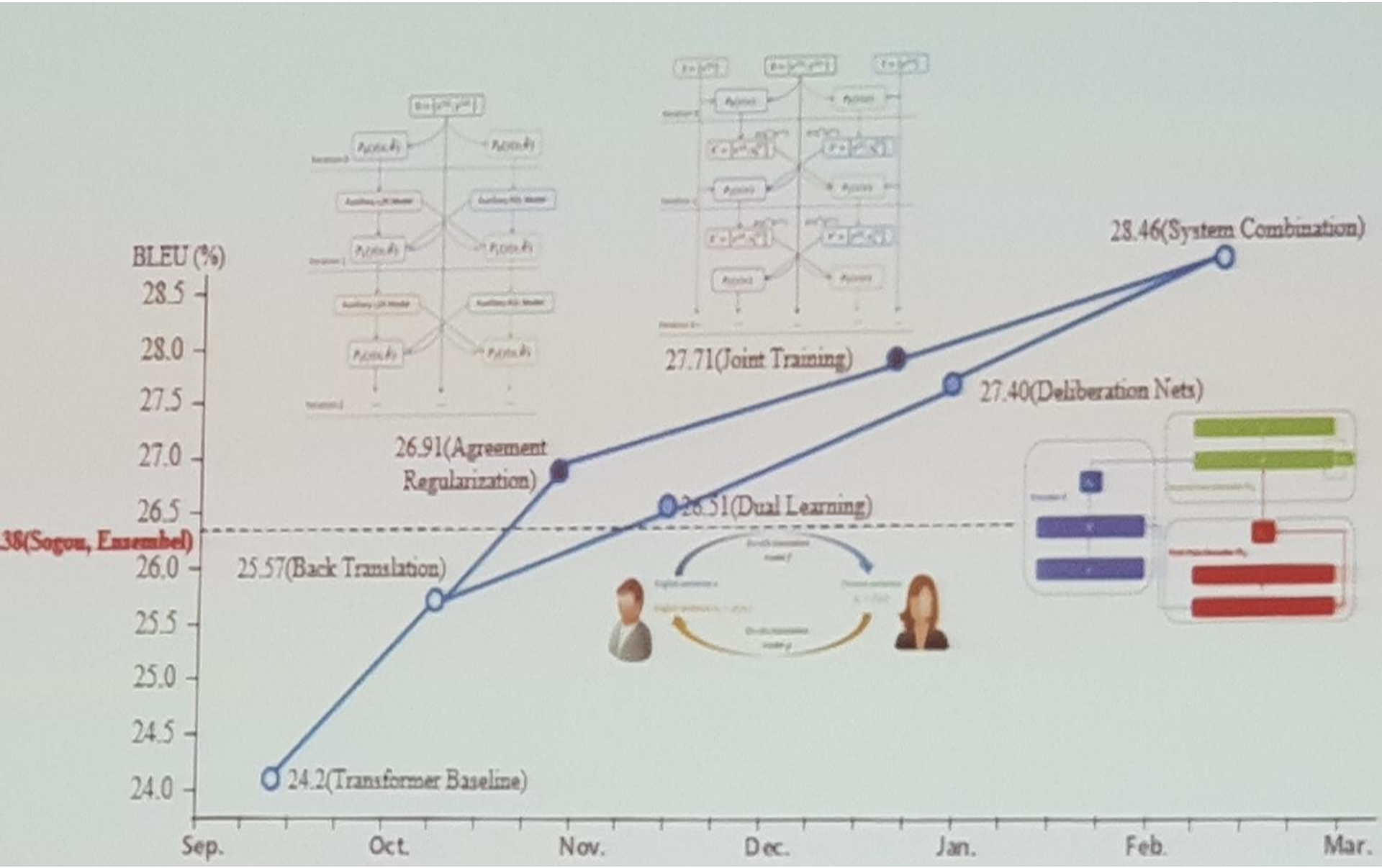
Microsoft super-human recognition

Jianchang Mao (Corporate Vice President of Bing Ads Marketplace & Serving, Artificial Intelligence & Research division at Microsoft**)**

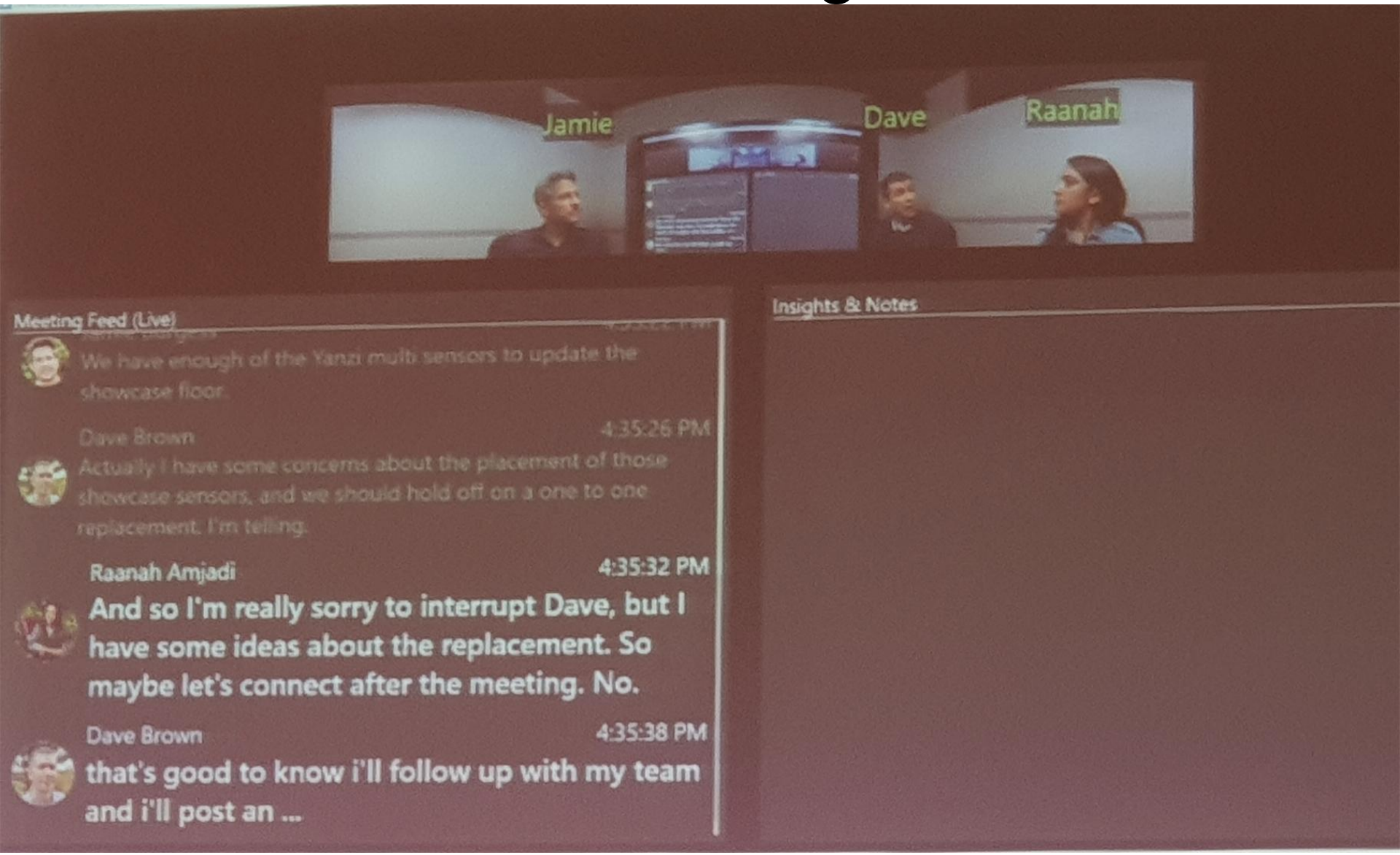


Chinese-to-English news translation

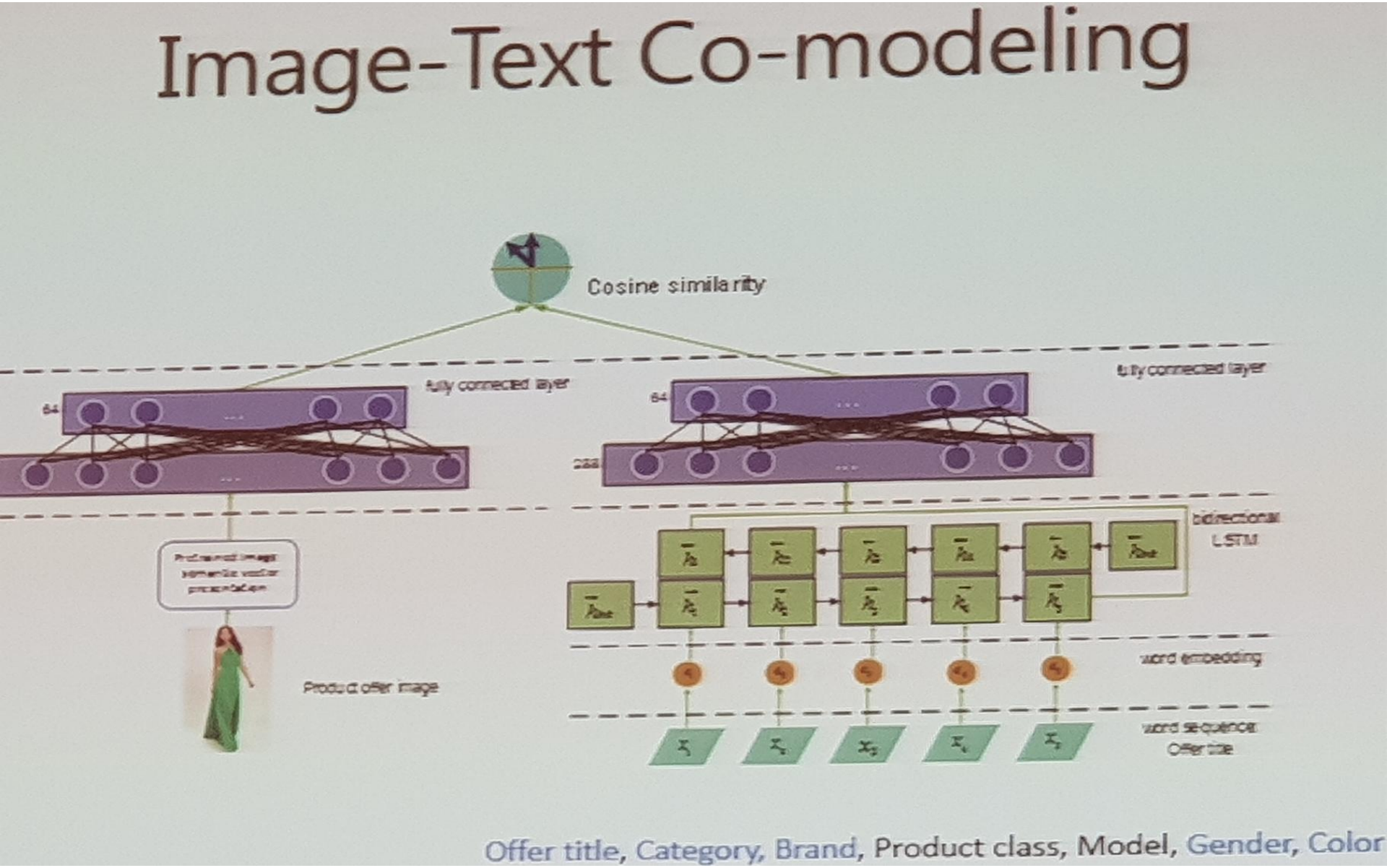
https://arxiv.org/abs/1803.05567



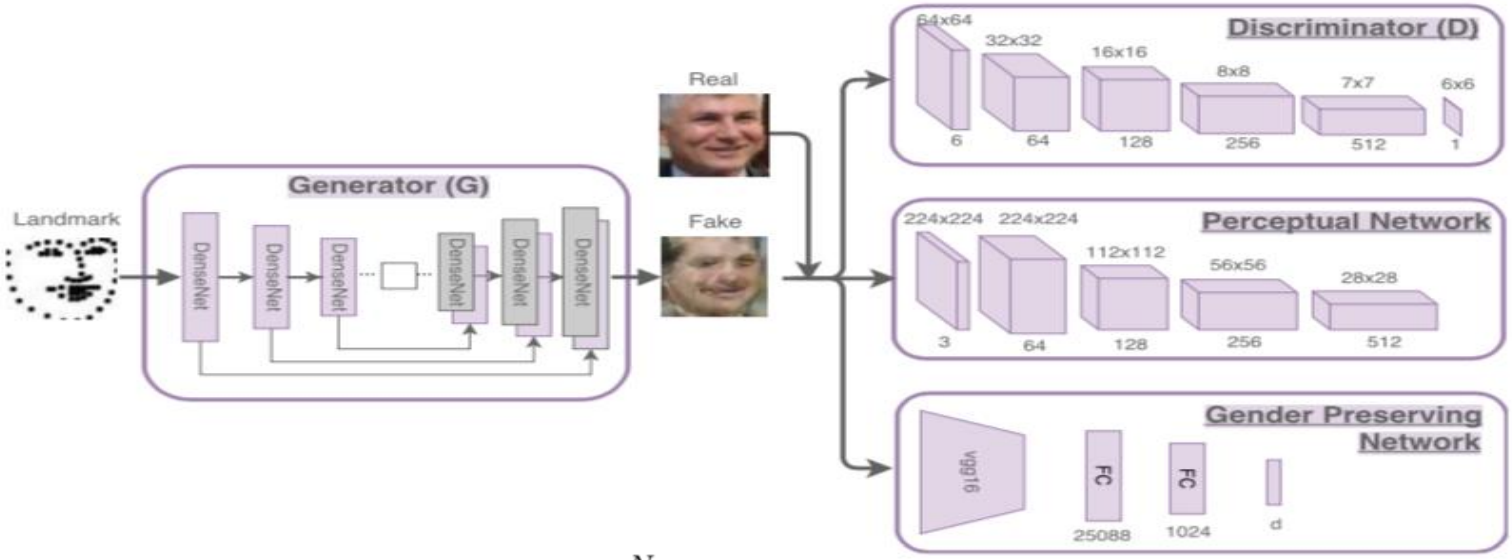
Modern meeting demo



Visual shopping assistant in Bing



GP-GAN: Gender Preserving GAN for Synthesizing Faces from Landmarks



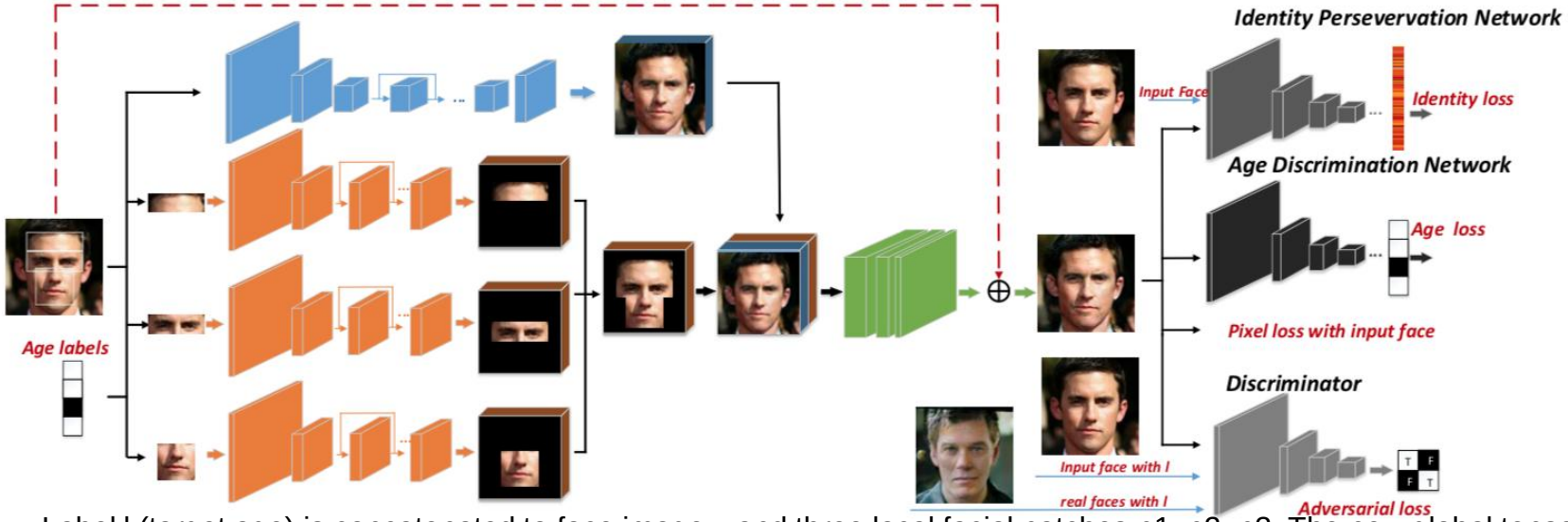
$$L = L_A + \lambda_P L_P + \lambda_C L_C + \lambda_1 L_1$$

- **Adversarial loss** $L_A = -\frac{1}{N}\sum_{i=1}^{N} log(D(\hat{x}_i))$
- **Perceptual loss** $L_P = ||V(\hat{x}) - V(x)||_1$
- **Gender preserving loss** $L_C = -\frac{1}{N}\sum_{i}(C(x_i)\log(C(\hat{x}_i)) + (1 - C(x_i))\log(C(\hat{x}_i))$
- **L1 loss** $L_1 = ||G(\hat{x}) - x||_1$

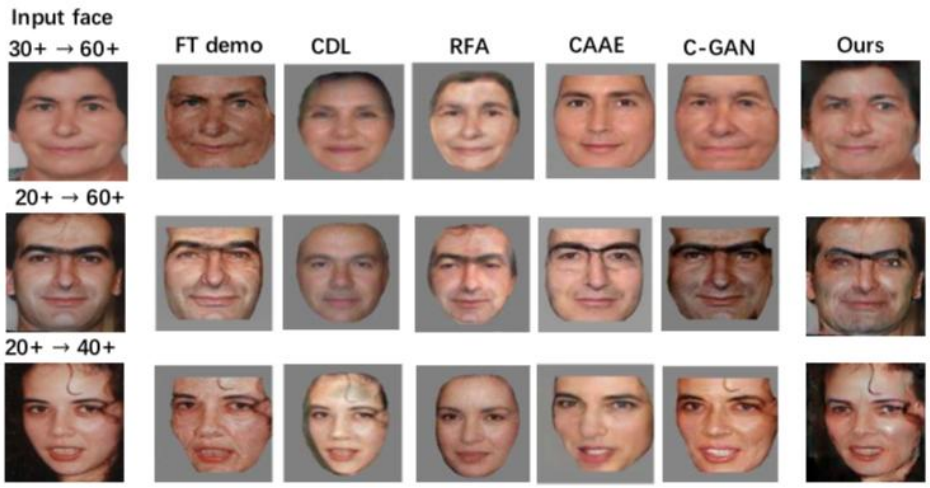
QUANTITATIVE COMPARISON OF GENDER RECOGNITION ACCURACY (%) FOR VARIOUS METHODS.

	LM (D)	LM (A)	CycleGAN	CVAE	BEGAN	CGAN	GP-GAN (UNet+GP-Loss)	GP-GAN (UDeNet,No GP Loss)	GP-GAN (UDeNet+GP-Loss)
LFW	78.0 ± 1.9	79.8 ± 2.4	81.8 ± 1.1	80.3 ± 2.0	84.4 ± 1.9	86.3 ± 2.5	91.1 ± 1.1	91.7 ± 1.6	93.1 ± 1.2
CASIA	61.0 ± 11.8	61.7 ± 13.6	64.8 ± 3.3	62.0 ± 4.1	67.8 ± 5.0	70.4 ± 5.5	73.2 ±3.9	76.7 ± 4.3	78.4 ± 4.1

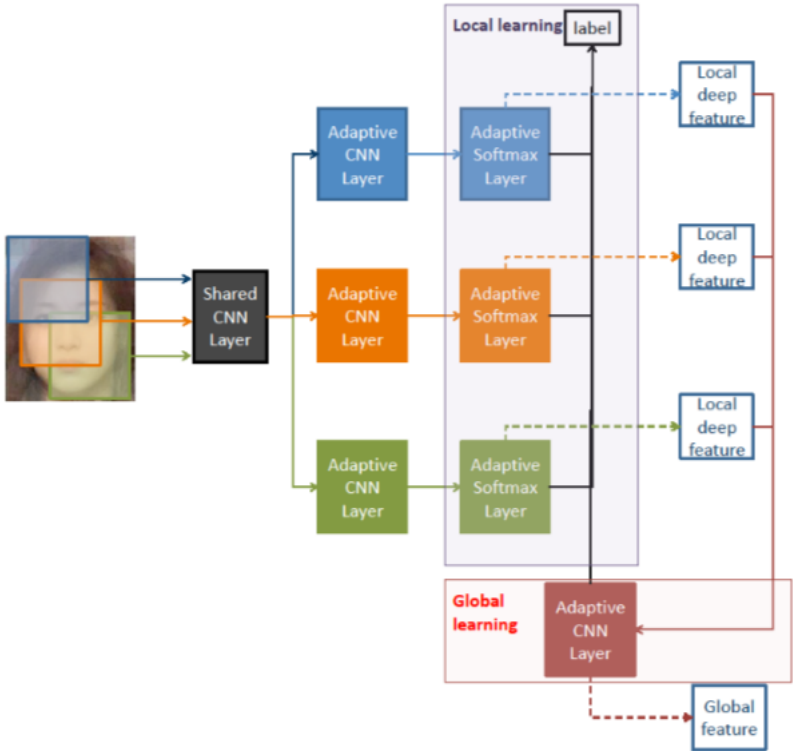
Global and Local Consistent Age Generative Adversarial Networks

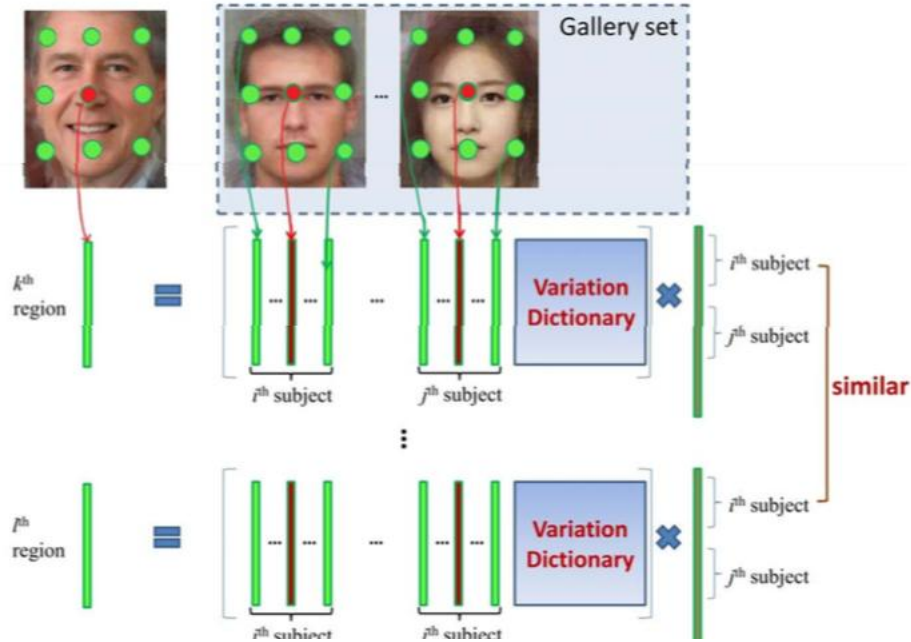


Label l (target age) is concatenated to face image x and three local facial patches p1, p2, p3. The new global tensor [x,l] is fed to global network, while the three new tensors [p1,l], [p2,l], [p3,l] are fed to their corresponding local networks. The outputs of the four networks are fused and fed into three convolutional layers to synthesize the residual face. Besides the discri[illegible]rk are also used to strengthen our GLCA-GAN.



Adaptive convolution local and global learning for class-level joint representation of face recognition with single sample per person





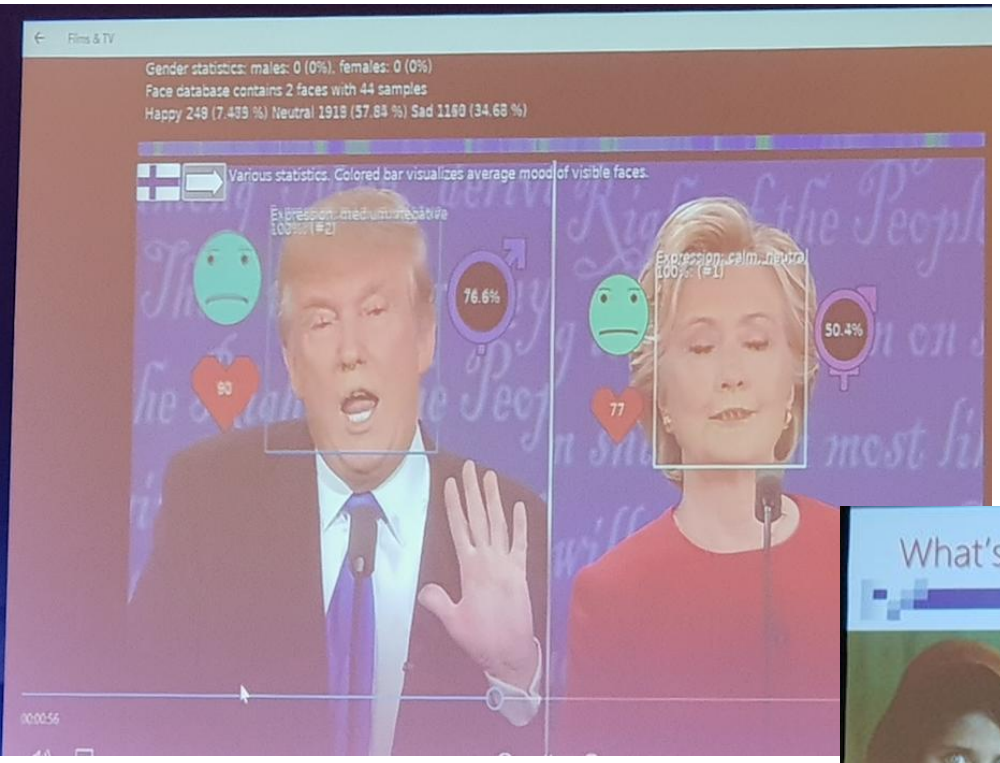
$$arg\ \min\left\{\sum_{k=1}^{K}\left\|\boldsymbol{y}^k-\boldsymbol{G}_i^k\alpha_i^k-\boldsymbol{D}^k\boldsymbol{\alpha}_v^k\right\|_2^2/\left\|\alpha_i^k;\boldsymbol{\alpha}_v^k\right\|_2^2\right\}$$

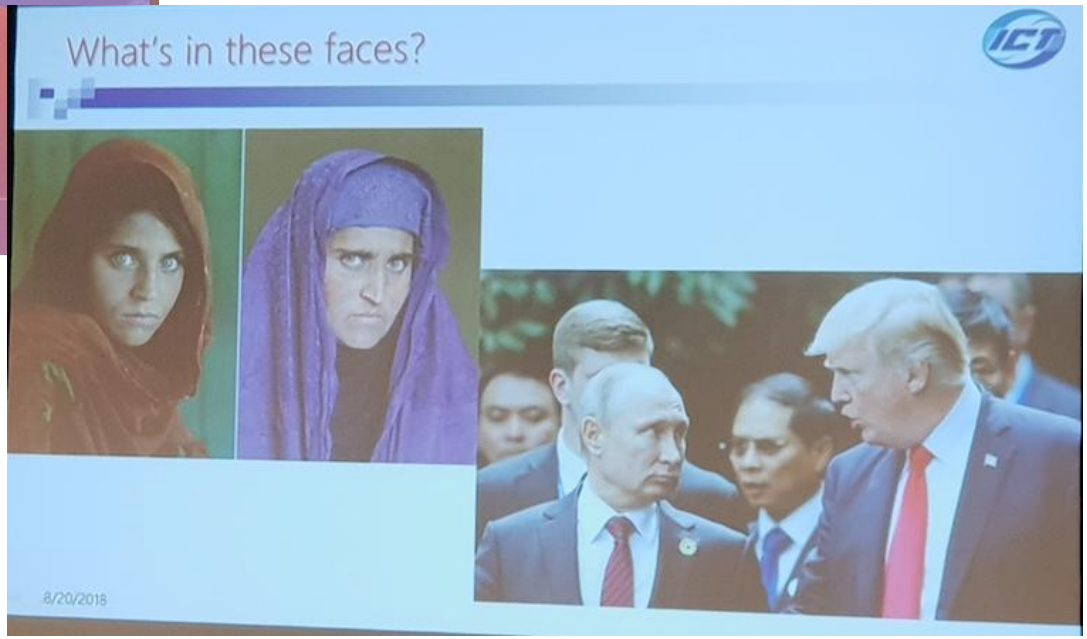
Table.2. Recognition accuracy (%) on LFW database with 50 gallery subjects.

Method	SRC [6]	DMMA	ESRC [7]	PRP [10]
Accuracy	20.4	17.8	27.3	33.1
Method	LGR [3]	DeepID [13]	Center-Fea [24]	CJR-RACF
Accuracy	30.4	70.7	90.6	**94.6**

Facial emotion recognition

- **Matti Pietikäinen (LBP): face analysis for emotion AI**
- **Xilin Chen: challenges in facial expression understanding**





Dynamic Ensemble Active Learning: A Non-Stationary Bandit with Expert Advice

N experts - ensemble of active learning criteria; K arms - available points in the pool.
Each expert (criterion) n provides a probability vector encoding preference $\boldsymbol{\xi}_t^n \in [0,1]^K$
Multi-armed Bandit with Expert Advice: choose the best expert giving advice vectors

$$\max_n \sum_{t=1}^{T} y_t^n - \mathbb{E}(\sum_{t=1}^{T} y_t^\pi) \qquad y_t^n = \sum_{k=1}^{K} \xi_t^{n,k} \times r_t^k$$

Non-stationary case: expected reward y_t^n of each expert n can change at any time step t
Proposed a non-stationary MAB with expert advice algorithm that has performance guarantees, and validate its practical application to active learning.

Temporal uncertainty set:

$$\mathcal{V} = \left\{ y \in [0,1]^{N \times K} : \sum_{t=1}^{T-1} \sup_n |y_t^n - y_{t+1}^n| \le V_T \right\}$$

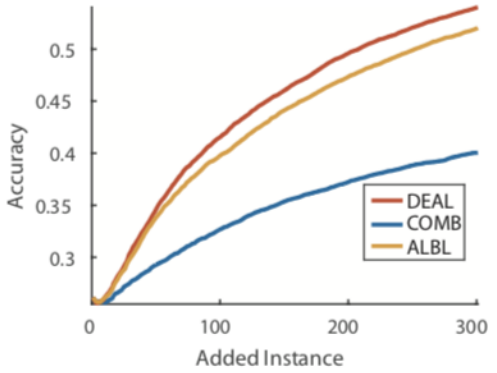
Theorem 1. *Let π be the REXP4 policy with a batch size $\Delta_T = \lceil (A \log N)^{1/3} (T/V_T)^{2/3} \rceil$ and $\gamma = \min\{1, \sqrt{\frac{A \log N}{(e-1)\Delta_T}}\}$. Then, there is some constant C such that for every $T \ge 1, K \ge 2, N \ge 2$, and $V_T \in [A^{-1}, A^{-1}T]$*

$$R^\pi(\mathcal{V}, T) \le C(A \log N \cdot V_T)^{1/3} T^{2/3} \qquad (4)$$

where $A = \min\{N, K\}$ indicates the smaller number of experts or arms.

Dynamic Regret:

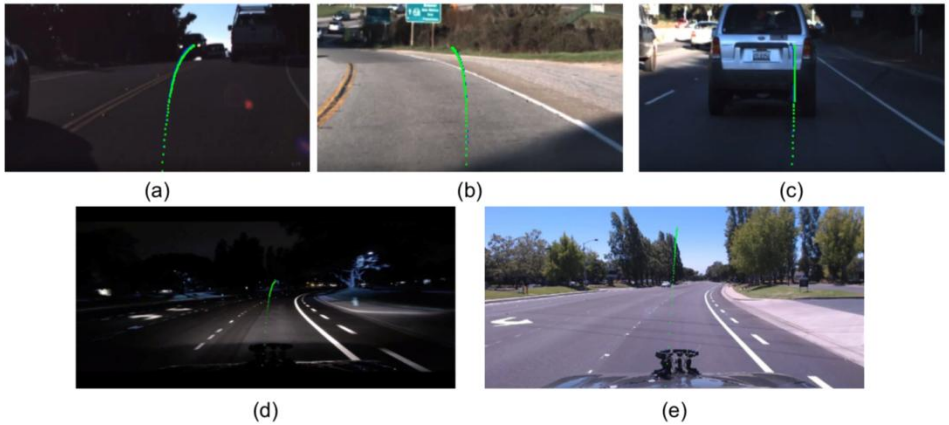
$$R^\pi(\mathcal{V}, T) = \sup_{y \in \mathcal{V}} \{ \sum_{t=1}^{T} y_t^* - \mathbb{E}^\pi [\sum_{t=1}^{T} y_t^\pi] \}$$

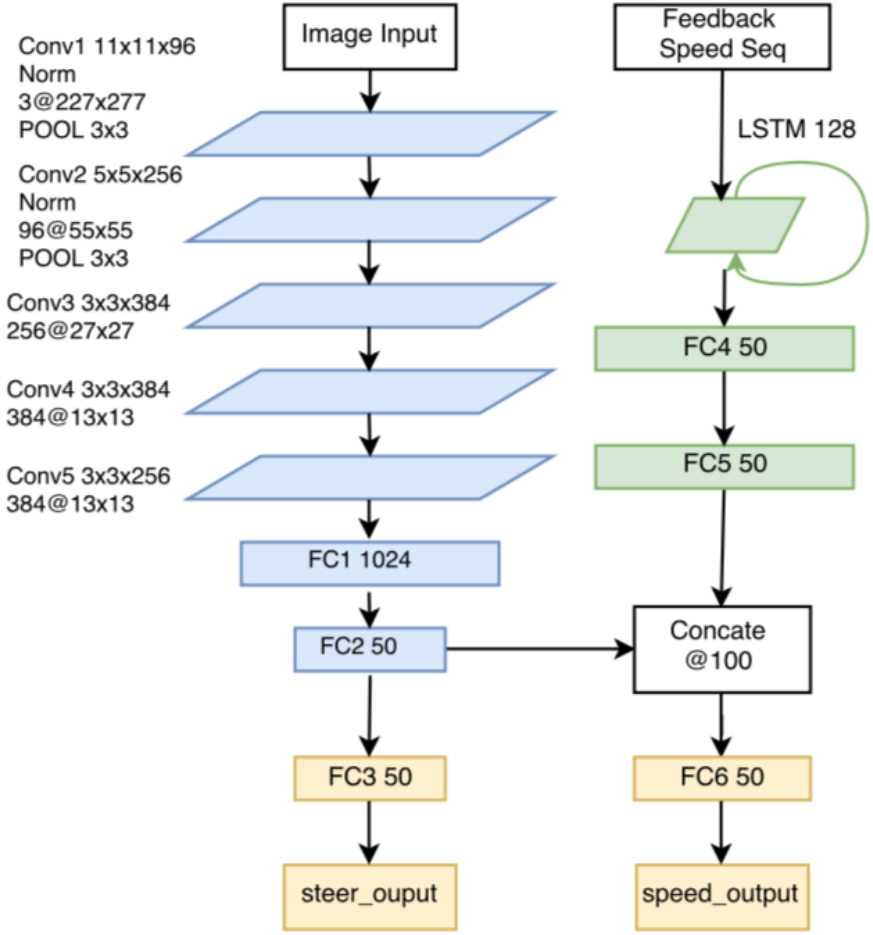


(d) letter

End-to-end Multi-Modal Multi-Task Vehicle Control for Self-Driving Cars

Udacity/SAIC datasets





Method	Angle (MAE in *degree*)
Nvidia's PilotNet [6]	4.26
Cg Network [28]	4.18
Base Steering Model	2.84
Discrete Speed Command Network	1.85
Multi-modal Multi-task Network	1.26

Dataset	Speed (MAE in m/s)
Udacity [12]	0.19
SAIC	0.45

Kernel-Weighted Graph Convolutional Network for Traffic Forecasting

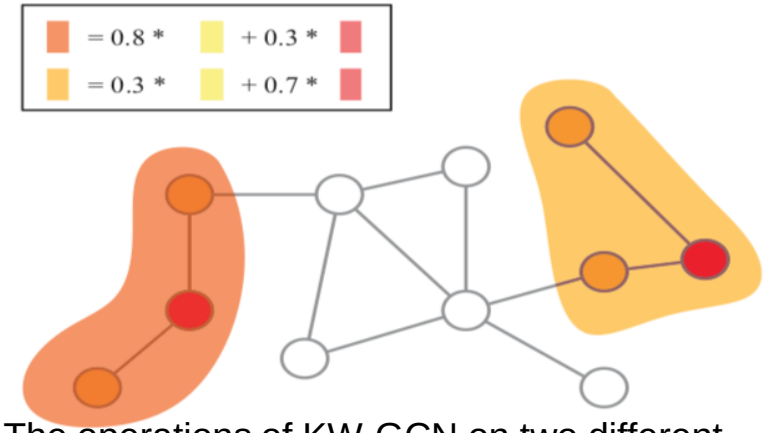
(a) (b)

Fig. 3. (a) The heatmap of taxis flow. (b) The graph representation of BeijingIF

KW-GCN introduces multiple candidate convolution kernels and kernel weighted parameters to fit an appropriate kernel which changes with locations

$$Y_{nq} = f\left(\sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{r=1}^{c} A_{rn}\mathcal{W}_{kjq}^{r}X_{j}^{S(n,k)}\right)$$

Location-dependent kernel weighted parameters



The operations of KW-GCN on two different nodes of a graph. Red nodes are central node and brown ones are the neighbors. Two candidate convolutional kernels, indicated as pink and yellow are used and the kernel size is three. The number 0.8, 0.3 and 0.7 are the kernel weighted parameters.

Quantum Walks

Edwin Hancock (PR Editor-in-Chief)

Use qubits to represent state of walk. Evolution governed by Schrodinger's equation

$$\frac{d}{dt}|\psi_t>=-iL|\psi_t>$$

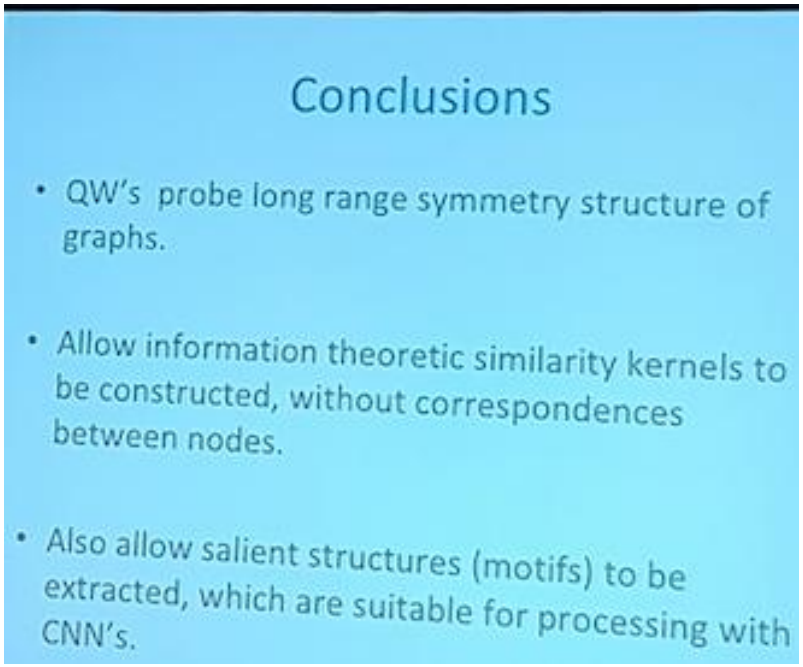
$$|\psi_t>=\exp[-iLt]|\psi_0>$$

Interference in quantum walks used to develop alternatives to classical random walks algorithm

- Quantum walks hit exponentially faster than classical ones
- Reversible, non ergodic, no limiting distribution
- Admit interference of different feasible walks

Tasks:

1. Distinguish between graphs that are cospectral with respect to standard matrix representations (Pattern Recognition 42 (2009) 1988—2002)
2. Measure graph similarity with the quantum Jensen-Shannon divergence (PHYSICAL REVIEW E 91, 022815 (2015))



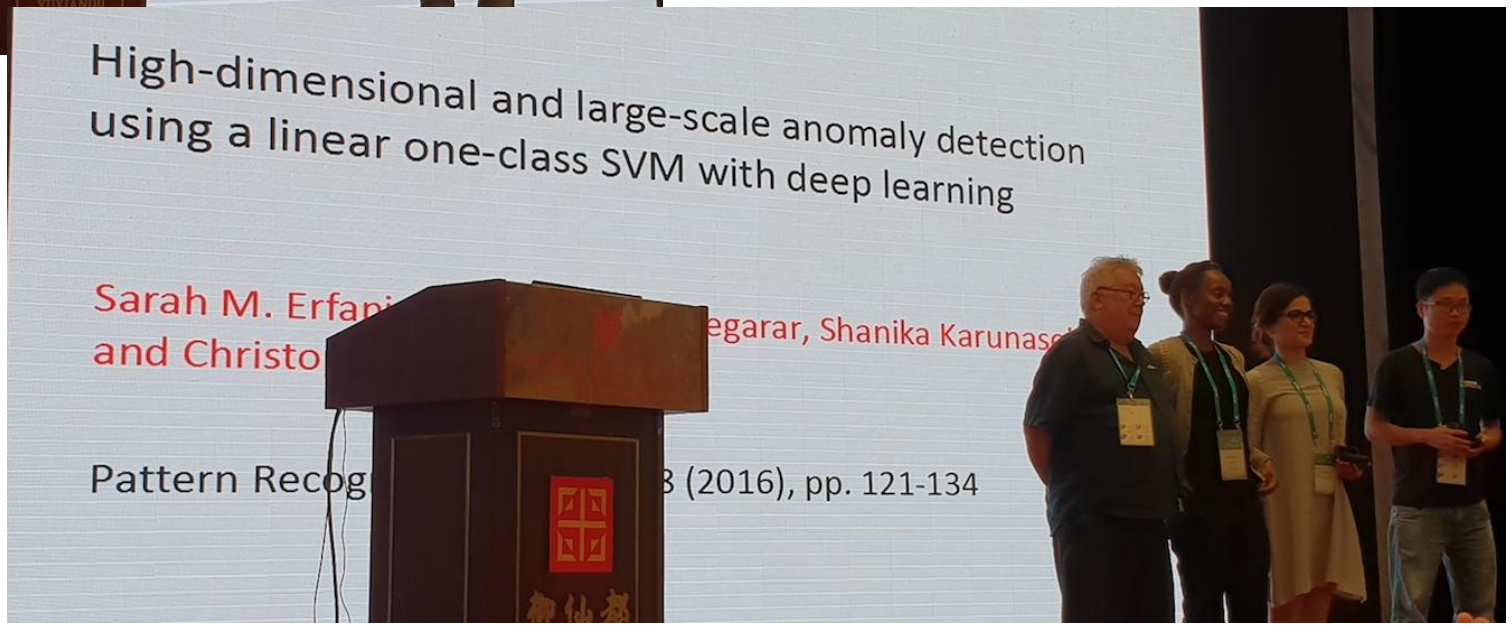
Best Scientific Papers

- Zhihong Zhang, Da Zhou, Chuanyu Xu, Beizhan, Wang; Dong Wang; Guijun REN, Edwin Hancock, Lu Bai and Lixin Cui, **Depth-Based Subgraph Convolutional Neural Networks**
- Jiyuan Zhang, Gang Zeng and Hongbin Zha, **Scalable Monocular SLAM by Fusing and Connecting Line Segments with Inverse Depth Filter**
- Jiedong Hao, Jing Zhang, Wei Wang and Tieniu Tan, **DeepFirearm: Learning Discriminative Feature Representation for Fine-Grained Firearm Retrieval**
- Qin Xu, Yifan Sun, Yali Li and Shengjin Wang, **Attend and Align: Improving Deep Representations with Feature Alignment Layer for Person Retrieval**
- Pau Riba, Andreas Fischer, Josep Llados and Alicia Fornés, **Learning Graph Distances with Message Passing Neural Networks**
- Ashis Kumar Dhara,Erik Arvids, Markus Fahlström, Johan Wikström, Elna-Marie Larsson and Robin Strand, **Interactive Segmentation of Glioblastoma for Post-Surgical Treatment Follow-Up**

Best Student Papers

- Shan Xu, Xiao Zhang and Shizhong Liao, **A Linear Incremental Nystrom Method for Online Kernel Learning**
- Doris Antensteiner, Svorad Ttolc and Thoma Pock, **Variational Fusion of Light Field and Photometric Stereo for Precise 3D Sensing within a Multi-Line Scan Framework**
- Pramuditha Perera, Abavisani Mahdi, and Vishal Patel, **In2I : Unsupervised Multi-Image-to-Image Translation Using Generative Adversarial Networks**
- Vishwanath Sindagi, Vishal Patel, and Xing Di, **GP-GAN: Gender Preserving GAN for Synthesizing Faces from Landmarks**
- Shailza Jolly, Brian Iwana, Ryohei Kuroki and Seiichi Uchida, **How Do Convolutional Neural Networks Learn Design?**
- Daniel Lopez-Martinez, Ke Peng, Sarah Steele, Arielle Lee, David Borsook and Rosalind Picard, **Multi-task Multiple Kernel Machines for Personalized Pain Recognition from Functional Near-Infrared Spectroscopy Brain Signals**

Best papers in “Pattern Recognition”



今儿个是俺后的六十大寿
寿比南山不老松
福如東海長流水
壽
壽