Brain-inspired Modeling and its Applications

Prof. Jie Yang

Shanghai Jiao Tong University

China

Part A:

Introduction of
Shanghai Jiao Tong University


交通大學

Shanghai JiaoTong University

1896 - 1904	Nan Yang Public School
1905 - 1906	Imperial Polytechnic College of the Commerce Ministry
1906 - 1911	Shanghai Industrial College of the Ministry of Posts and Telegraphs
1911 - 1912	Nan Yang College
1912 - 1921	Government Institute of Technology of the Communications Ministry
1921 - 1922	Nan Yang College of Chiao Tung
1922 - 1927	Nan Yang University of the Communications Ministry
1927 - 1928	First Chiao Tung University of the Communications Ministry
1928 - 1942	National Chiao Tung University(Main Campus in Shanghai)
1942 - 1946	National Chiao Tung University(Main Campus in Chong Qing)
1946 - 1949	National Chiao Tung University
1949 - 1957	Jiao Tong University
1957 - 1959	Jiao Tong University(Shanghai Campus)
1959 -	Shanghai Jiao Tong University

The Second oldest university in China

Ranking in China: Top 4, Top 2 (Engineering)

Total funding from NSFC,China: Top1

Shanghai JiaoTong University

- includes 5 campuses, area of about 4 km^2.
- 31 schools (departments)
- 63 undergraduate programs
- 250 masters-degree programs,
- 203 PhD programs,
- 28 post-Dr programs
- 11 state key laboratories / national eng. research centers.
- > 2,000 professors and associate professors.
- Approx. 40000 students: 18000 BSC students, 18000 MSC students, 4000 PhD students;

上海交通大学 SHANGHAI JIAO TONG UNIVERSITY

List of Schools in Engineering

- School of Naval Architecture, Ocean and Civil Eng.
- School of Mechanical Eng.
  - School of Nuclear Science and Eng.
- School of Electronic, Information and Electrical Eng.
  - School of Information Security Eng.
  - School of Software
  - School of Microelectronics
- School of Materials Science and Eng.
- School of Environmental Science and Eng.
- Univ. of Michigan-SJTU Joint Institute
- School of Aeronautics and Astronautics
- SJTU-ParisTech Elite Institute of Technology

List of Schools in Life Medical Sciences

- School of Life Sciences and Biotechnology
- School of Agriculture and Biology
- School of Medicine
- School of Pharmacy
- School of Biomedical Engineering

List of Schools in Humanities & Social Sciences

- School of Humanities
- Antai College of Economics and Management
- KoGuan Law School
- School of International and Public Affairs
- School of Media and Design
- School of Foreign Languages
- School of Marxism
- Department of Physical Education
- Shanghai Advanced Institute of Finance
- School of Entrepreneurship & Innovation
- School of Continuing Education
  - Continuing Education
  - E-Learning Lab
- School of International Education
- China Europe International Business School

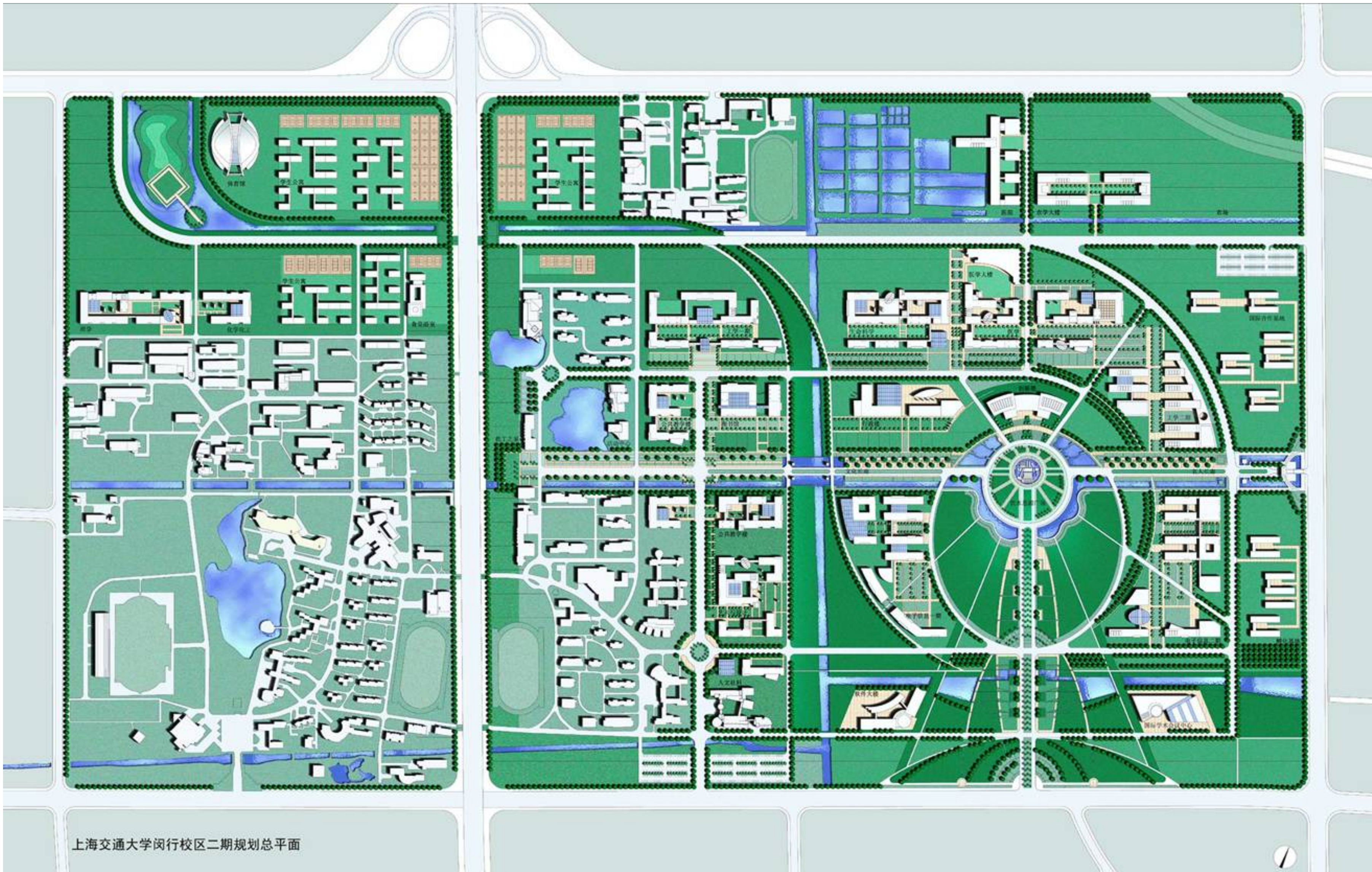
Campus of SJTU

Bao Yugang Library

Bao Zhaolong Library

上海交通大学 SHANGHAI JIAO TONG UNIVERSITY

New Campus of SJTU



New Campus of SJTU

School of Electronic, Information & Elect. Eng. (EIEE)

- Consists of 7 major disciplines, including:

  * Electrical Eng.

  * Electronic Science and Technology

  * Information and Communication Eng.

  * Control Science and Eng.

  * Computer Science and Technology

  * Software Eng.

  * Instrument Science and Technology
- 120 professors; 180 associate professors.
- > 700 PhD students, >2000 MSC students, >3500 BSC students.

Part 1

saliency detection

1. Introduction

- Motivation

Everyone knows what attention is…

——William James

- A computational approach to visual attention
- Fast selection for objects of interest in scenes

1. Introduction

- **Visual saliency** is the selective mechanism of human visual attention that concentrates on one aspect of the scene while ignoring other things.

Cognitive neuroscience

Neuropsychology

Psychology

Computer vision

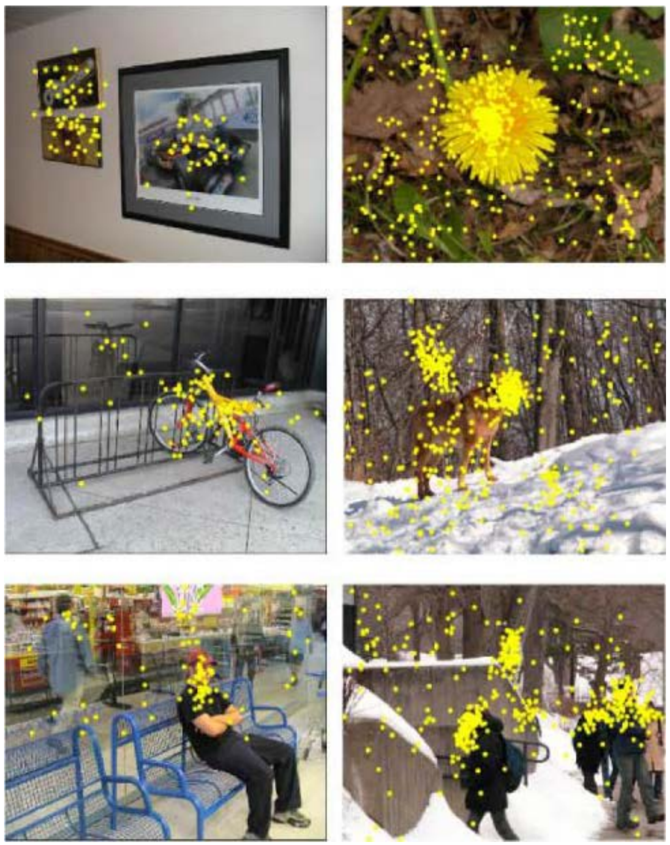
<u>Studied by multiple disciplines</u>

Applications

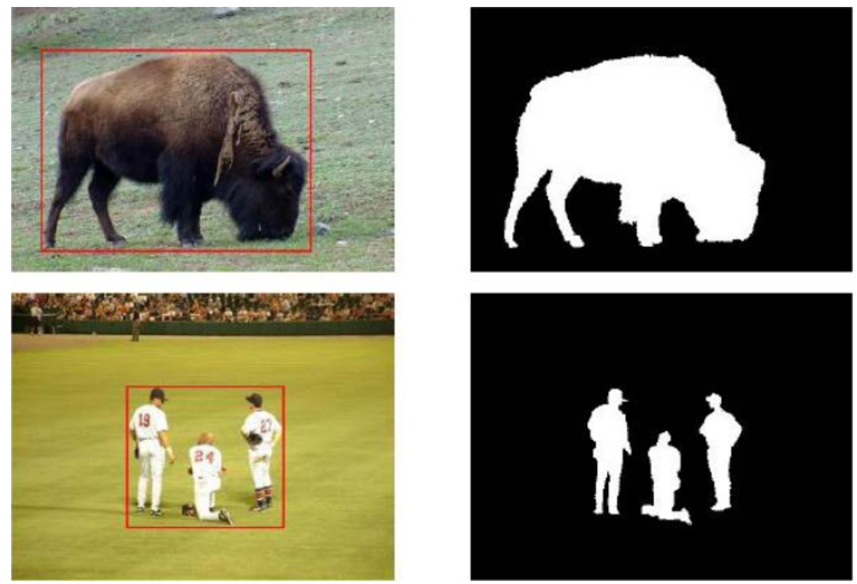
- object detection and recognition
- image compression
- video summarization,
- content-based image editing and image retrieval.

1. Introduction

- Two branches of saliency detection in computer vision:
  Eye fixation prediction v.s. **Salient object detection**



Eye fixation prediction



Salient object/region detection

上海交通大学 SHANGHAI JIAO TONG UNIVERSITY

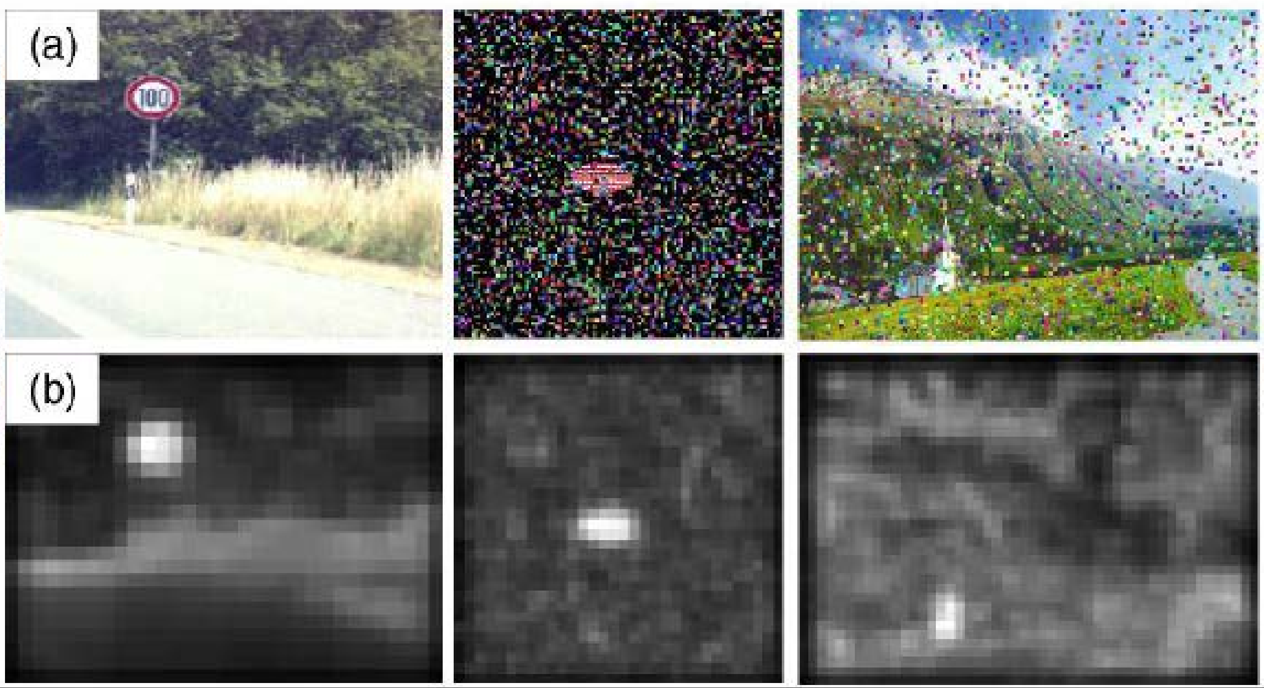
1. Introduction

- **Eye fixation prediction** becomes active after Itti *et al.*'s work (TPAMI 1998)....

A Model of Saliency-Based Visual Attention for Rapid Scene Analysis

Laurent Itti, Christof Koch, and Ernst Niebur

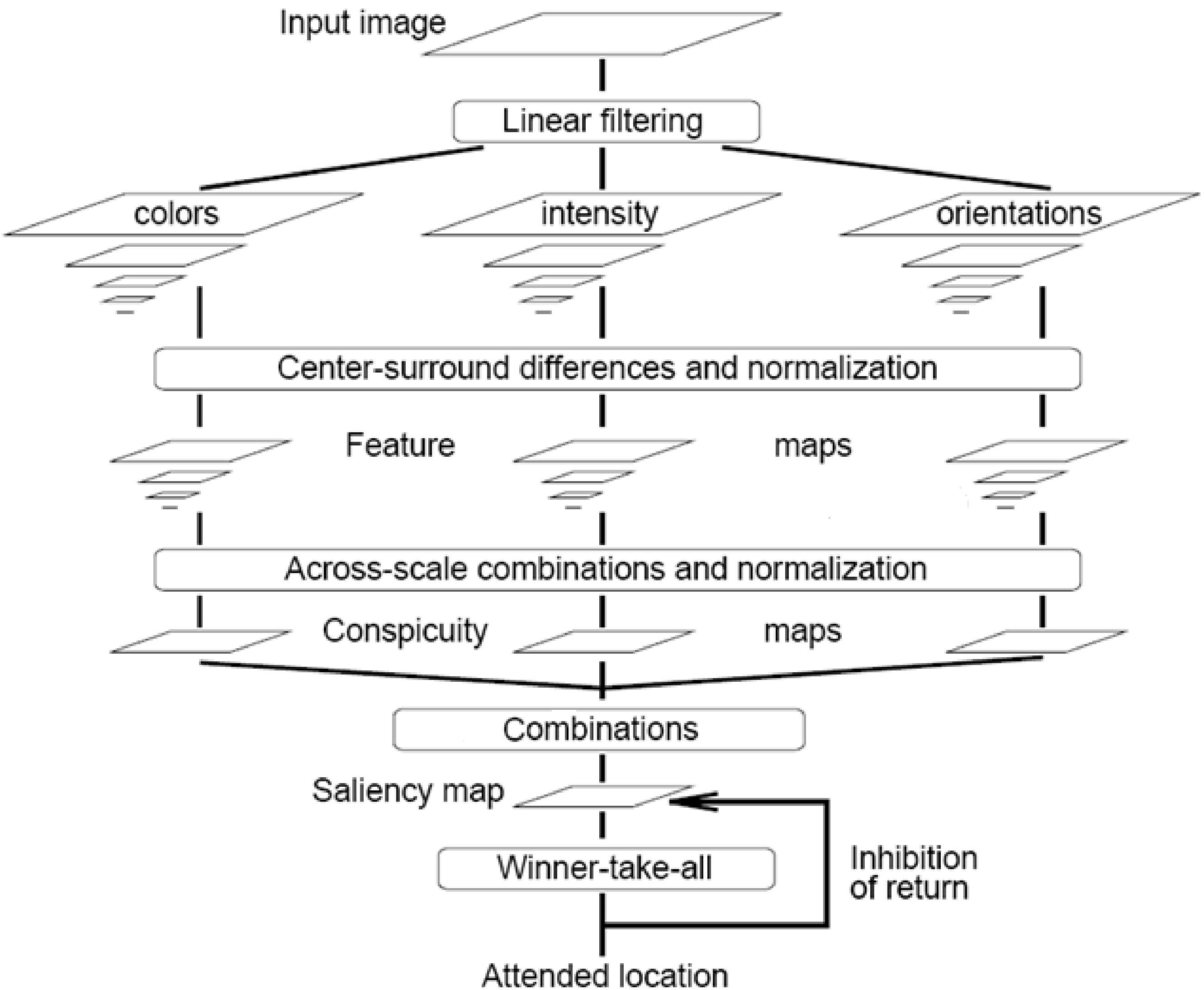
Abst neur Multi salie locat comp comp in de



ne d. he zed

Related Work

Feature Integration: Itti1998, Itti2000, Itti2005, Gao2008…



1. Introduction

- **Salient object detection** becomes active after Liu *et al.*'s work (CVPR 2007, TPAMI 2011)

Learning to Detect a Salient Object

Tie Liu, Zejian Yuan, Jian Sun, Jingdong Wang, Nanning Zheng, *Fellow, IEEE*, Xiaoou Tang, *Fellow, IEEE*, and Heung-Yeung Shum, *Fellow, IEEE*

Abstract—In this pap
where we separate th
surround histogram, a
field is learned to effec
salient object from seq
of thousands of carefu
them to demonstrate t

blem as a binary labeling task
g multiscale contrast, center-
obally. A conditional random
roposed approach to detect a
age database containing tens
ed a set of experiments over

Index Terms—Salien

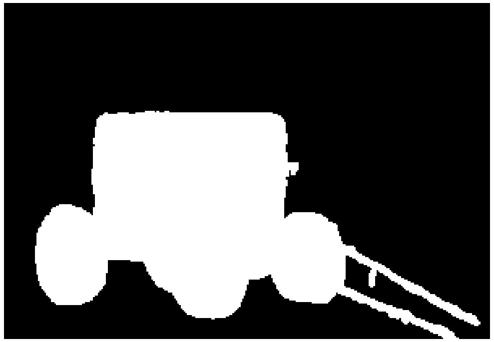
1. Introduction

- **Our work** focuses on **salient object detection**:

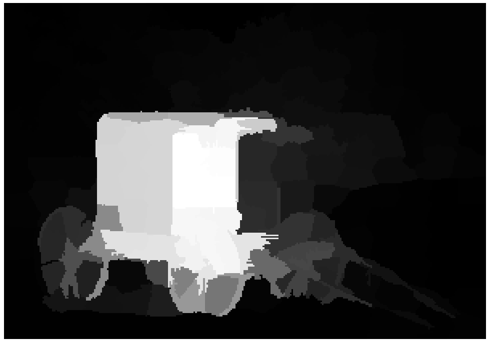
  -Automatically detect attention-grabbing objects in a scene;

  -Highlight entire objects uniformly and suppress irrelevant background in resulting saliency maps;

Input image



Ground truth



Our saliency map

1. Introduction

- **The motivation:**

Conventional pipeline

Image ⇨ Regions ⇨ Regional saliency

Results of state-of-the-art models

Input image RC SF LR HS DRFI MR

Regions for saliency computation

Superpixels for saliency computation

drawbacks

Only local color similarity considered;
object holism ignored;

1. Introduction

- **The motivation:**

Better grouping of a whole object → leads to → Better saliency estimation

Results of state-of-the-art models

Input image RC SF LR HS DRFI MR

NCS (ours) GT

1. Introduction

- **Novelties of this work:**

1. Apply the Normalized graph cut (Ncut) to salient region detection, and induce a saliency map by Ncut eigenvectors for better visual clustering;

2. Embed saliency detection in an adaptive multi-level merging scheme to discover cluster information conveyed by Ncut eigenvectors.

1. Introduction

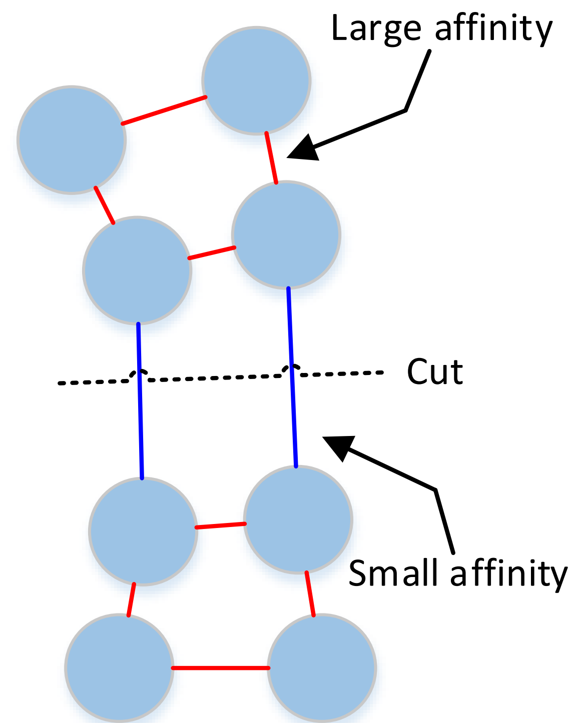
- **What is Normalized graph cut (Ncut)**?

Given a similarity graph G=(V,E) with affinity matrix **W** and a desired partition number k, Ncut finds a partition $\{A_1, A_2, \ldots, A_k\}$ of V which minimizes:

$$Ncut(A_1, \ldots, A_k) = \sum_{i=1}^{k} \frac{cut(A_i, \bar{A}_i)}{assoc(A_i, V)}$$

where: $cut(A, B) := \sum_{v_i \in A, v_j \in B} w_{ij}$

$$assoc(A_i, V) := \sum_{v_i \in A, v_j \in V} w_{ij}$$



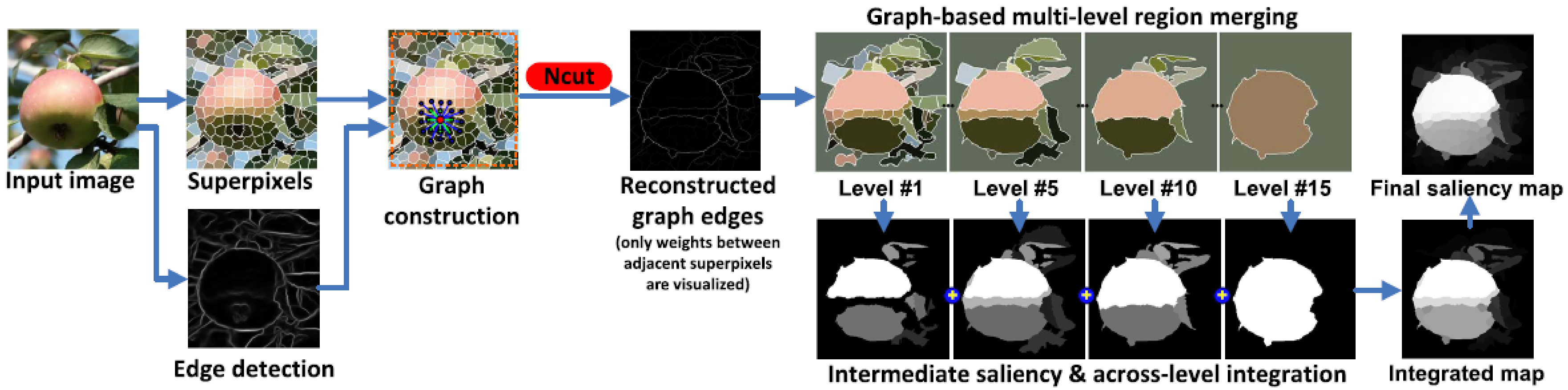
A toy graph

1. Introduction

- **Why use Ncut for saliency detection?**

1. Due to the normalization, Ncut biases cut of fairly large sets of vertices. Most salient objects are **perceptually large regions**, whereas **too small regions** often correspond to **noise** or **parts of an object**.
2. Ncut is a global, discriminative, and also non-parametric partition technique. Its approximated solution is efficient to achieve.
3. Ncut has not yet been used to inducing saliency maps.

2. Proposed Method



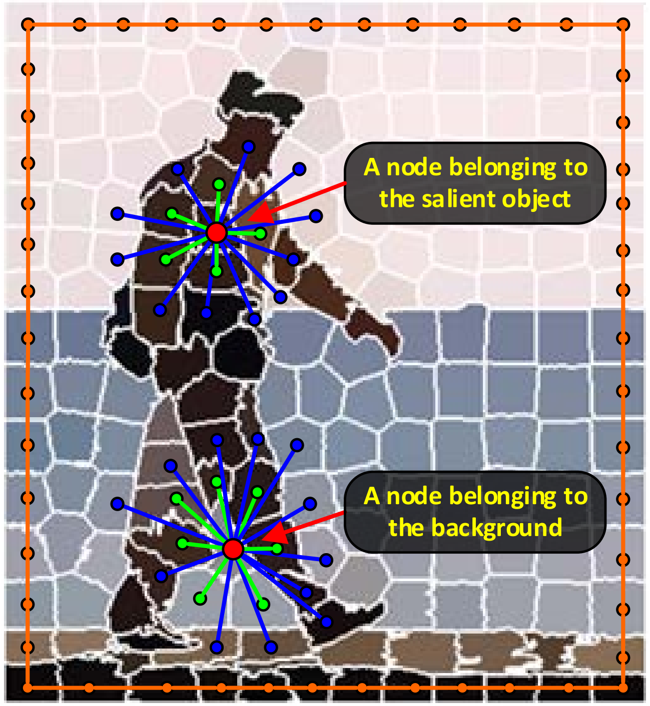
Overview of the proposed method

2. Proposed Method

- **Graph Construction for the Ncut:**

Input image

Superpixel graph



The 2-ring graph topology: green connections (neighbor superpixels)
+ blue connections (neighbors of neighbor superpixels)
+ brown connections (boundary superpixels)

2. Proposed Method

- **Graph Construction for the Ncut:**

Graph edge weight (affinity)

$$w_{ij} = \begin{cases} \exp\left(-\frac{d_{ij}^{app+edge}}{\sigma^2}\right) & \text{If } R_i, R_j \text{ are connected} \\ 0 & \text{Otherwise} \end{cases}$$

$$d_{ij}^{app+edge} = (1-\alpha)d_{ij}^{app} + \alpha d_{ij}^{edge}$$

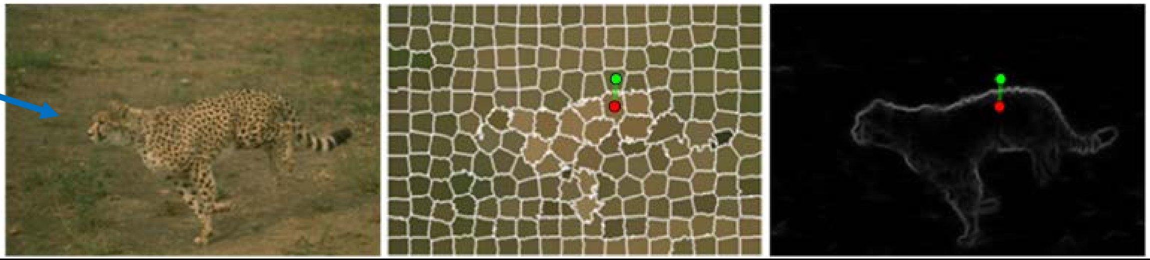
Superpixel color differences

$$d_{ij}^{app} = ||\mathbf{c}_i - \mathbf{c}_j||_2$$

Intervening edge magnitude

$$d_{ij}^{edge} = \max_{\mathbf{p}\in l(\mathbf{p}_i,\mathbf{p}_j)} \mathbb{E}(\mathbf{p})$$

Object and background have similar colors but different textures



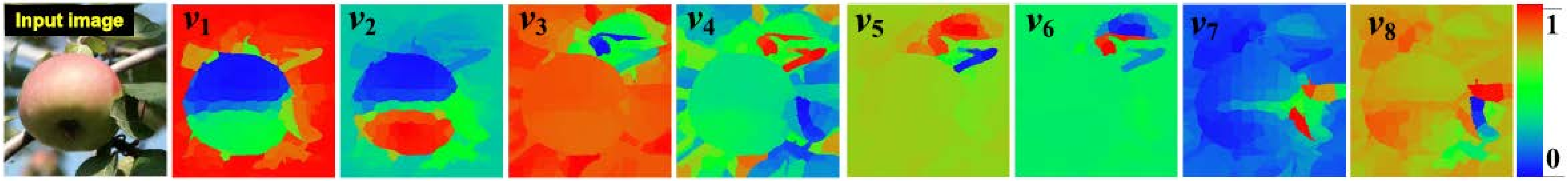
Intervening edge magnitude may help delineate object v.s. background!

2. Proposed Method

- **Apply the Ncut to Obtain Cluster Information**

1) Solve $(\mathbf{D}-\mathbf{W})\mathbf{v}=\lambda\mathbf{D}\mathbf{v}$ for generalized eigenvectors;

nvec (*nvec=8*) eigenvectors with smallest non-zero eigenvalues



2) Reconstruct the graph edge between two nodes:

$$e_{ij}=\sum_{k=1}^{nvec}\frac{1}{\sqrt{\lambda_k}}|\mathbf{v}_k(R_i^0)-\mathbf{v}_k(R_j^0)|$$

***Rationale**: eigenvectors are soft indicator vectors of Ncut. The reconstruction is indeed a measure of inter-cluster distance, i.e, the extent of the two nodes belonging to different clusters.*

上海交通大學 SHANGHAI JIAO TONG UNIVERSITY

2. Proposed Method

- **Graph-based Adaptive Merging of Vertices**

A multi-level adaptive merging scheme is proposed to generate regions for saliency computation:

1) Merging starts from initial superpixels $\{R_1^0, R_2^0, \ldots, R_N^0\}$
2) At level *l*, two regions R_i^l, R_j^l are fused if

$$D_{ij}^l \leq Th$$

$$D_{ij}^l = D(R_i^l, R_j^l) = \mathrm{mean}_{v_k \in R_i^l, v_m \in R_j^l, e_{km} \in E}\{e_{km}\}$$

3) At the next level *l+1*:

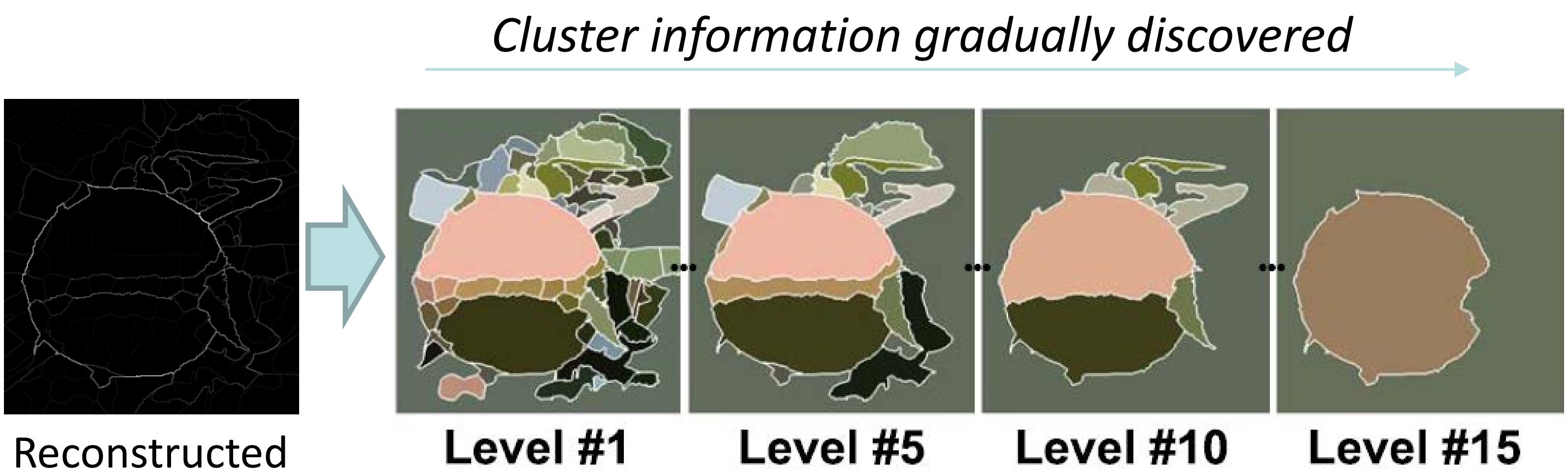
$$Th \leftarrow Th + \underline{T_s}$$

A fixed step during merging

4) The merging proceeds adaptively until the whole image becomes one region.

2. Proposed Method

- **Graph-based Adaptive Merging of Vertices**



Reconstructed graph edges from Ncut

2. Proposed Method

- **Regional Saliency Measures During Merging**

Consider saliency measures for a merged region R_i^l:

✓ Figure-Ground Contrast $S_{i,l}^{fg}$: the color contrast to boundary superpixels (boundary superpixels are pseudo-background)

✓ Center Bias $S_{i,l}^{cb}$: the closer to image center, the larger.

✓ Boundary Cropping $S_{i,l}^{bc}$: 0 for a region touching more than one image border, and 1 otherwise.

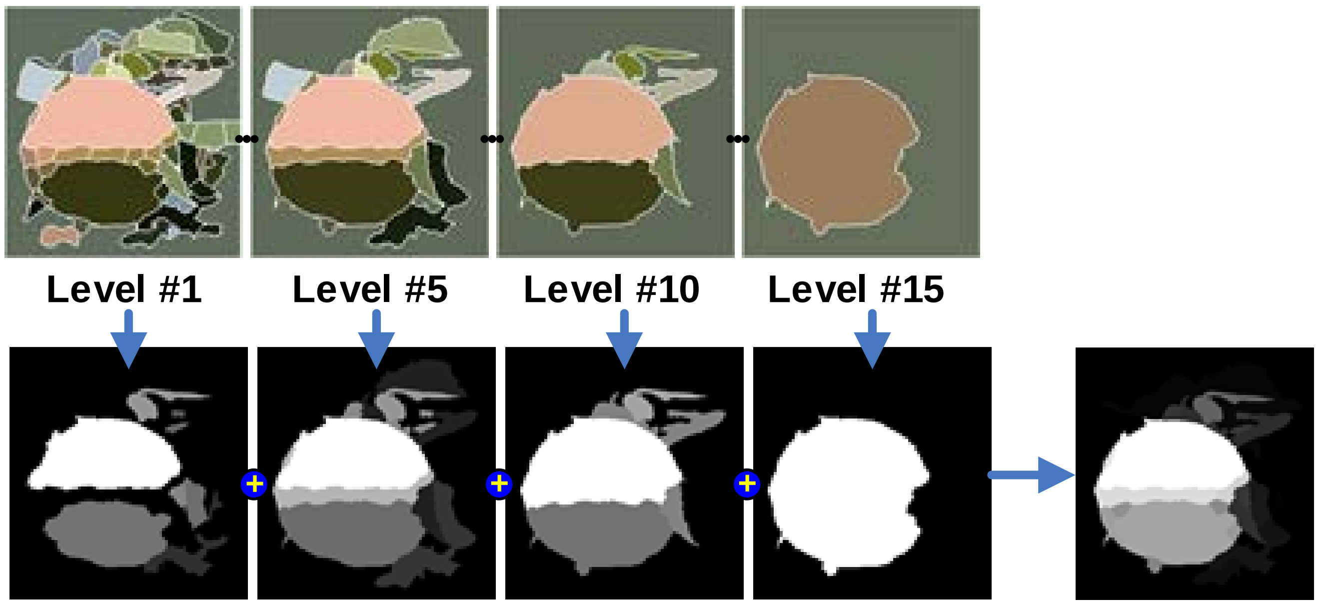
A combinatorial regional saliency score for R_i^l :

$$S_{i,l}^{final} = S_{i,l}^{fg} \cdot S_{i,l}^{cb} \cdot S_{i,l}^{bc}$$

2. Proposed Method

- **Across level integration**

Graph-based multi-level region merging



Intermediate saliency & across-level integration

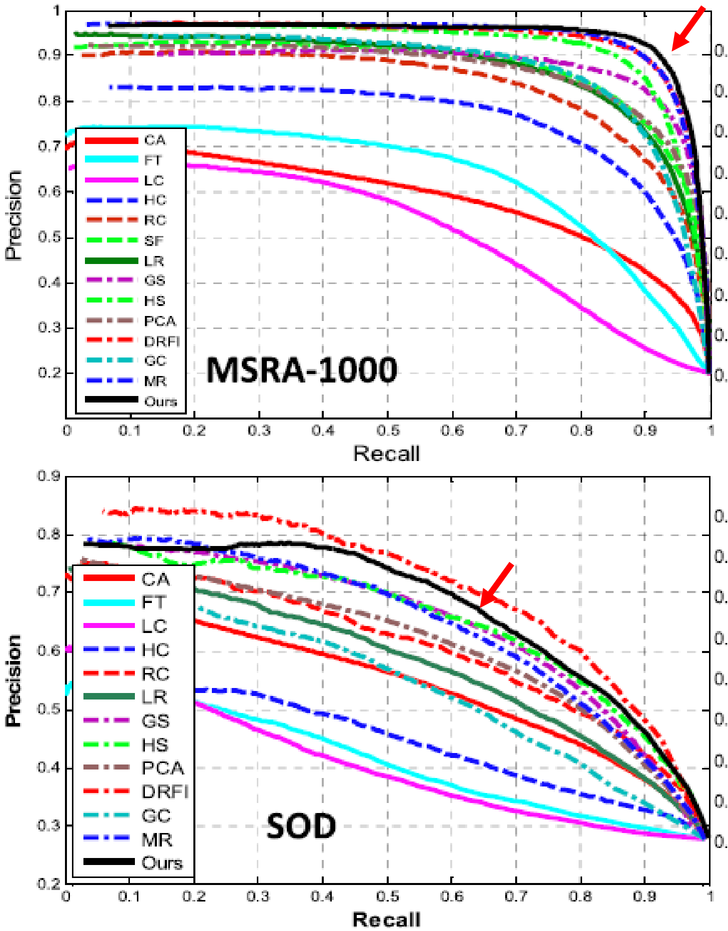
Integrated map

- **Post-smoothing by manifold ranking:**

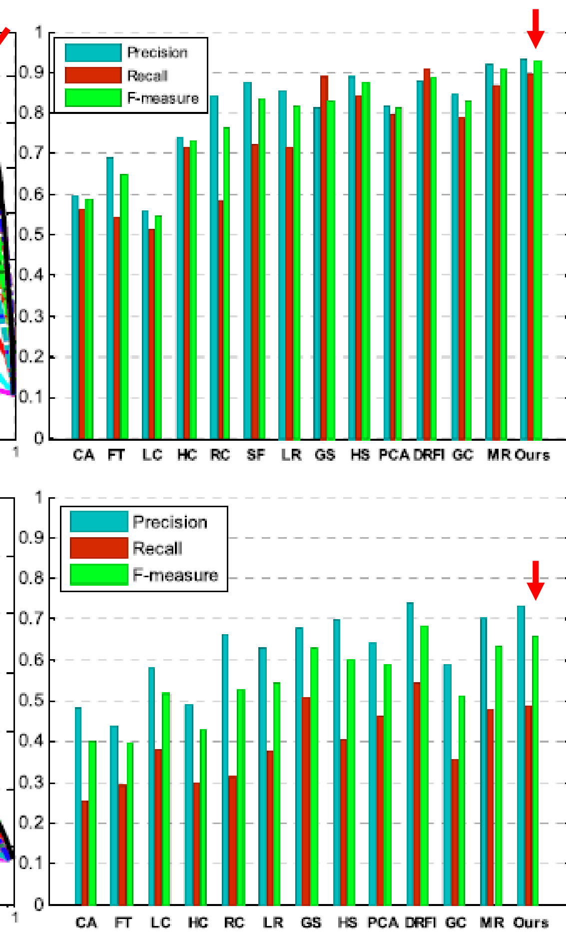
$$\underbrace{\mathbf{f}}_{\text{Smoothed saliency map}} = \underbrace{(\mathbf{D} - \beta\mathbf{W})^{-1}}_{\text{Manifold ranking}}\underbrace{\mathbf{s}}_{\text{Integrated saliency map}}$$

3. Experiments and Results

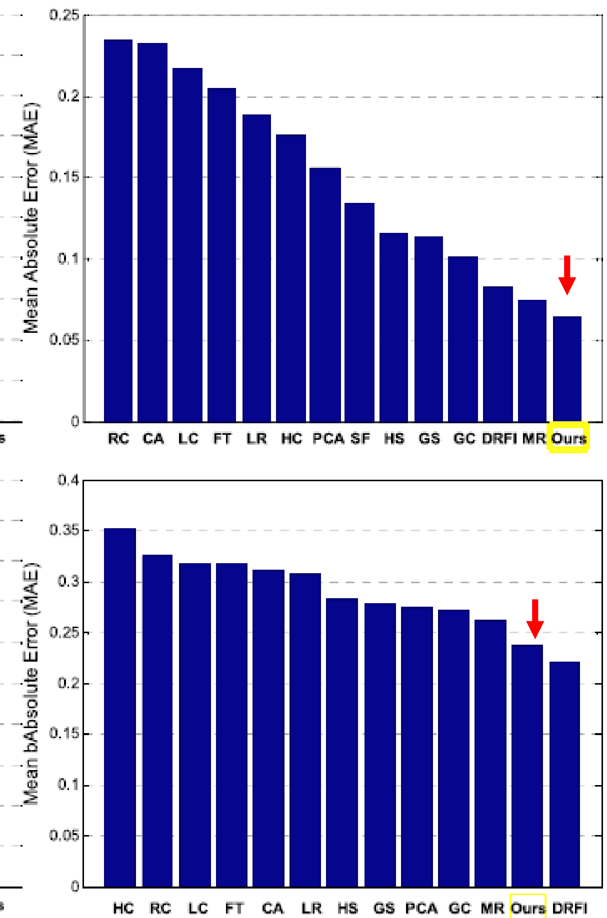
- **Quan. comparisons with 13 methods on 5 datasets**



Precision-recall curve (higher is better)



F-measure (higher is better)

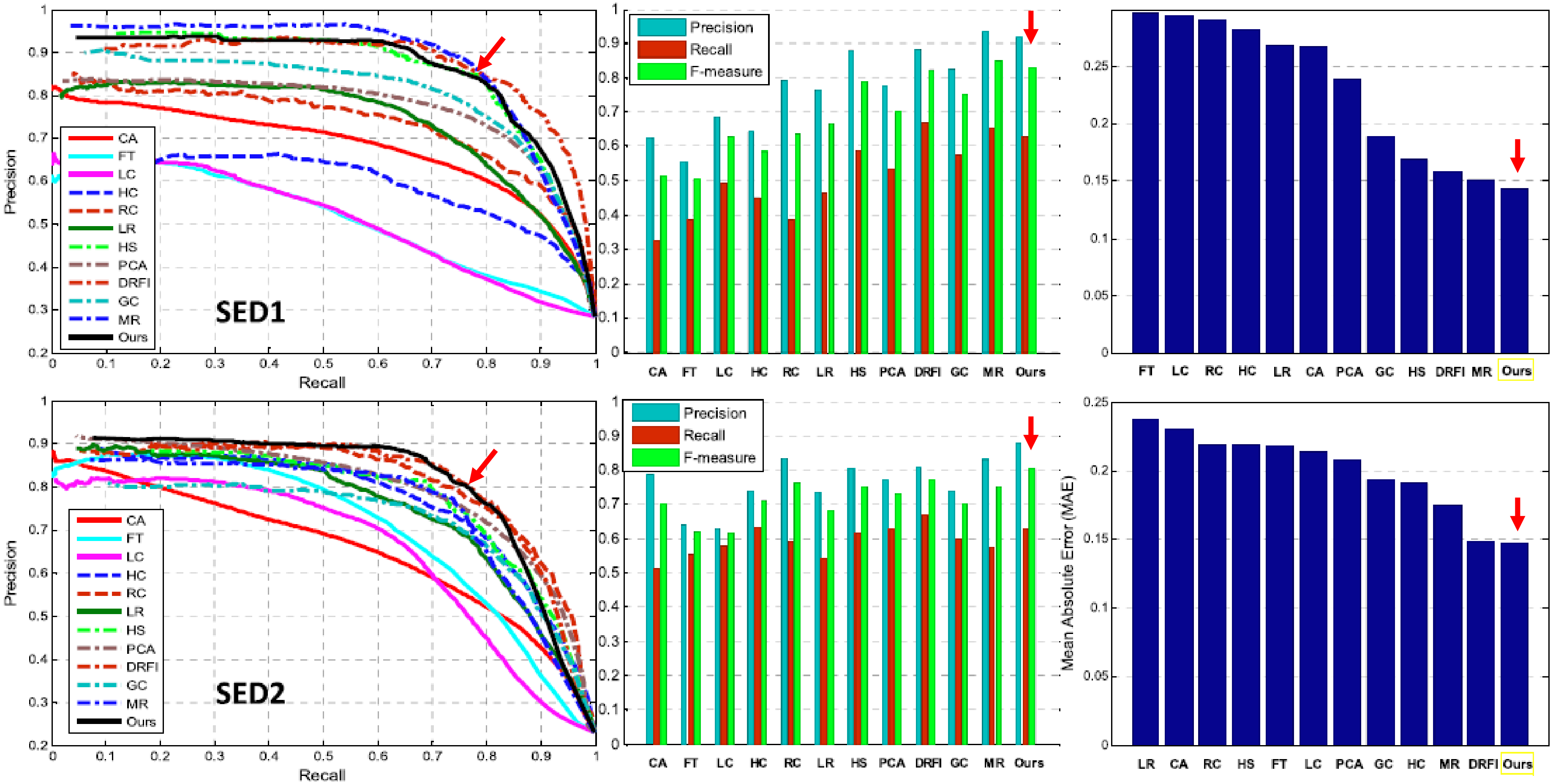


Mean absolute error (lower is better)

上海交通大学 SHANGHAI JIAO TONG UNIVERSITY

3. Experiments and Results

- **Quan. comparisons with 13 methods on 5 datasets**



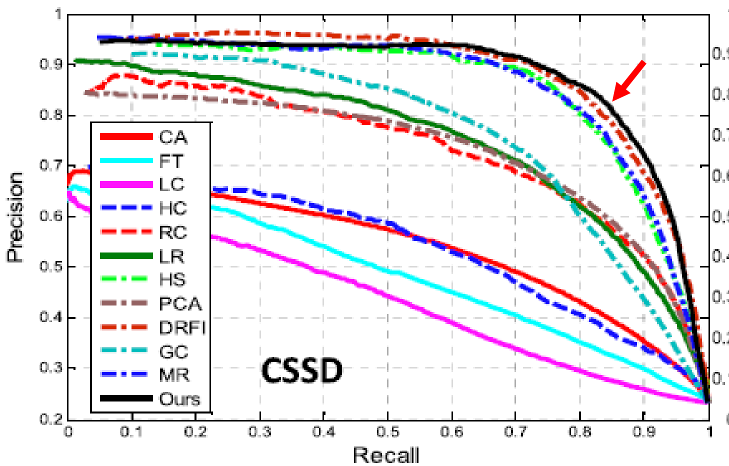
Precision-recall curve (higher is better)

F-measure (higher is better)

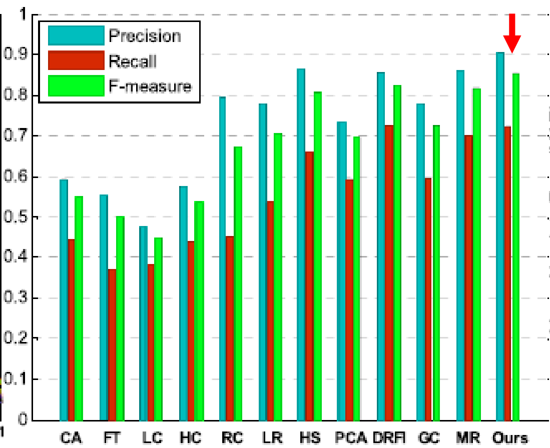
Mean absolute error (lower is better)

3. Experiments and Results

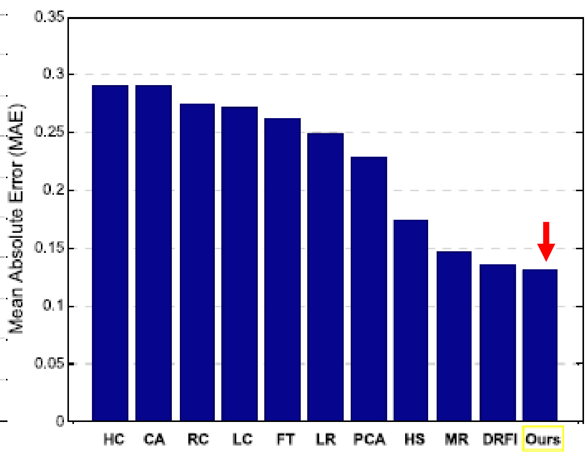
- Quan. comparisons with 13 methods on 5 datasets



Precision-recall curve (higher is better)



F-measure (higher is better)

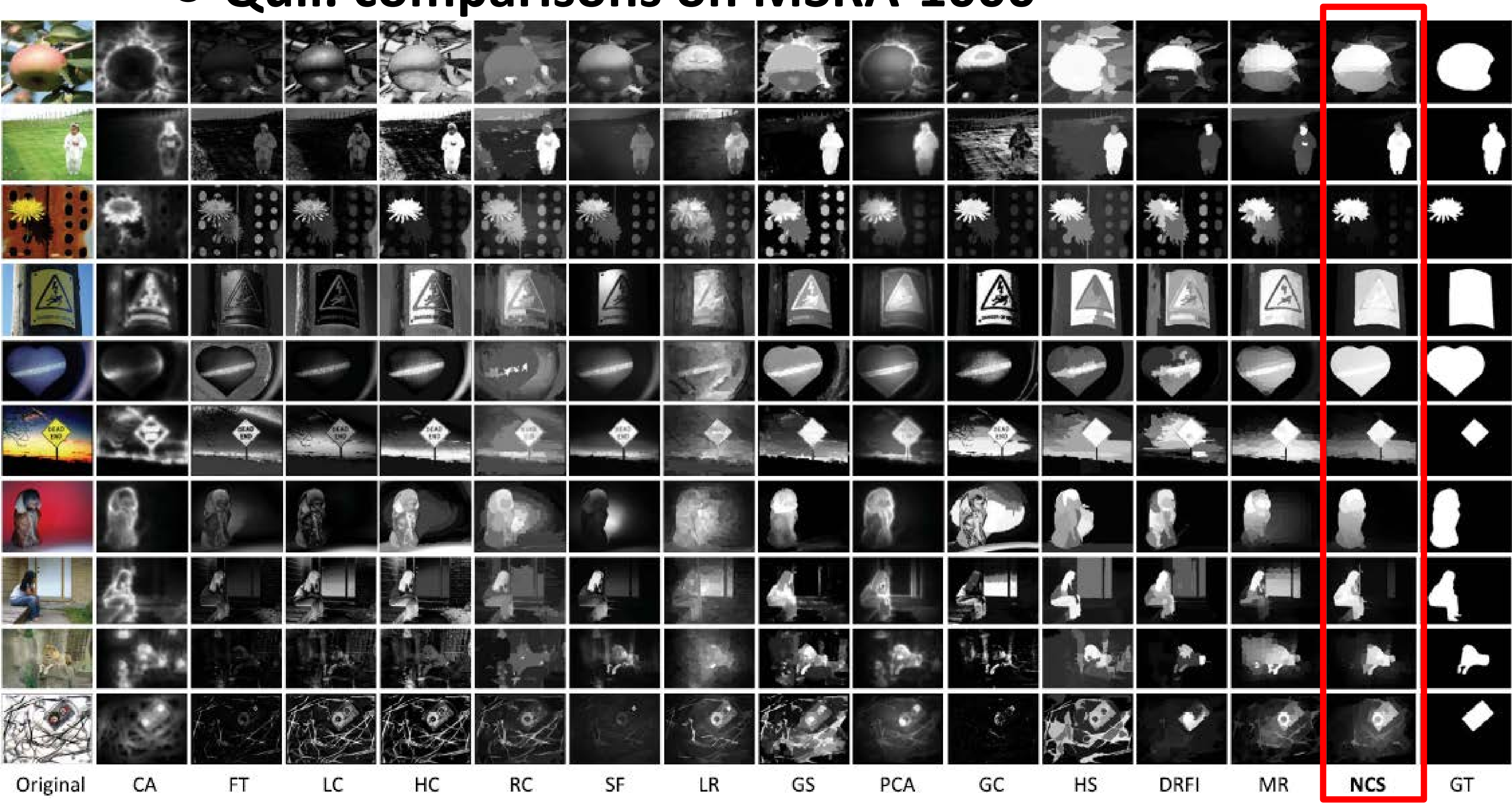


Mean absolute error (lower is better)

上海交通大學 SHANGHAI JIAO TONG UNIVERSITY

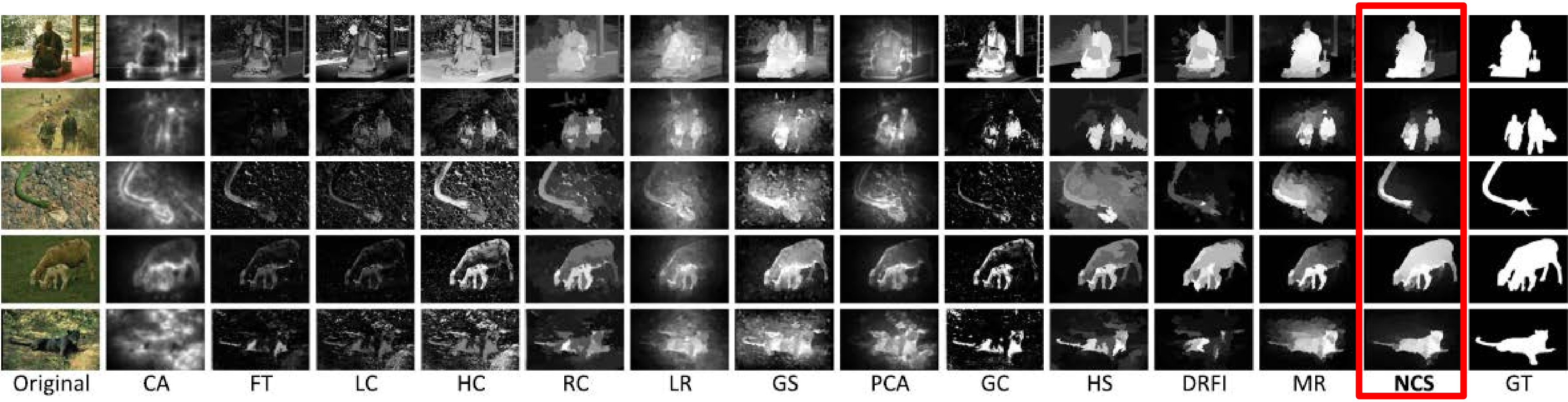
3. Experiments and Results

- Quli. comparisons on MSRA-1000



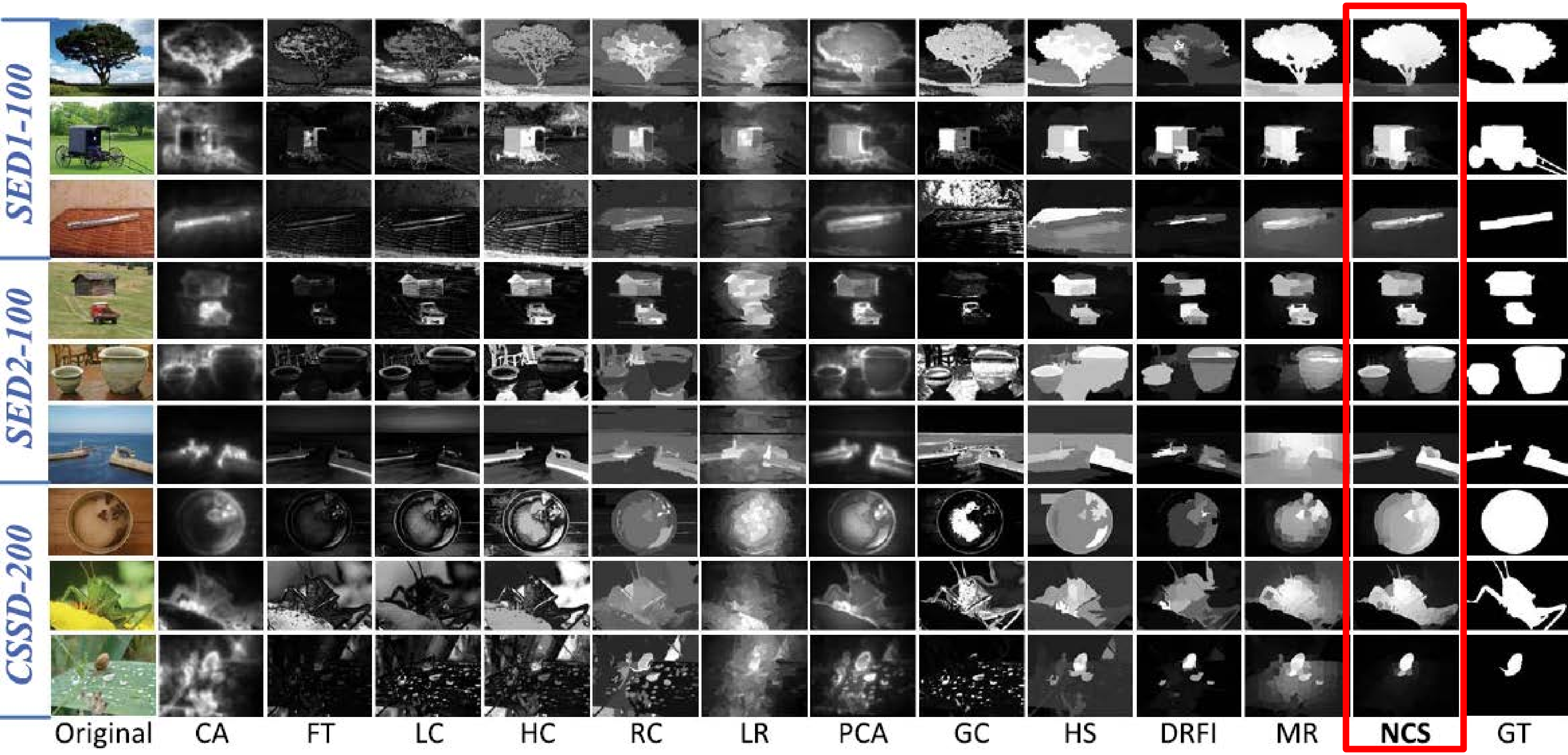
3. Experiments and Results

- **Quli. comparisons on SOD**



3. Experiments and Results

- Quli. comparisons on SED1, SED2 and CSSD



上海交通大學 SHANGHAI JIAO TONG UNIVERSITY

References

- Saliency Propagation From Simple To Difficult. *IEEE International Conference on Computer Vision and Pattern Recognition (**CVPR**),* 2015.
- Normalized Cut-Based Saliency Detection by Adaptive Multi-Level Region Merging**,** IEEE Transactions on Image Processing (**TIP**), 24(12):5671-5683,2015
- Saliency Detection by Fully Learning A Continuous Conditional Random Field, **IEEE TRANSACTIONS ON MULTIMEDIA (TMM)**. **DOI:** 10.1109/TMM.2017.2679898.
- Cross-modal Saliency Correlation for Image Annotation, **Neural Processing Letters**, p 1-13, March 2, 2016.
- Robust manifold-preserving diffusion-based saliency detection by adaptive weight construction **,** **NEUROCOMPUTING** **,** vol. 175: 336-347**,** JAN 29 2016
- Co-saliency detection via inter and intra saliency propagation; **SIGNAL PROCESSING-IMAGE COMMUNICATION**， 卷: 44 页: 69-83 出版年: MAY 2016

Shanghai Jiao Tong University

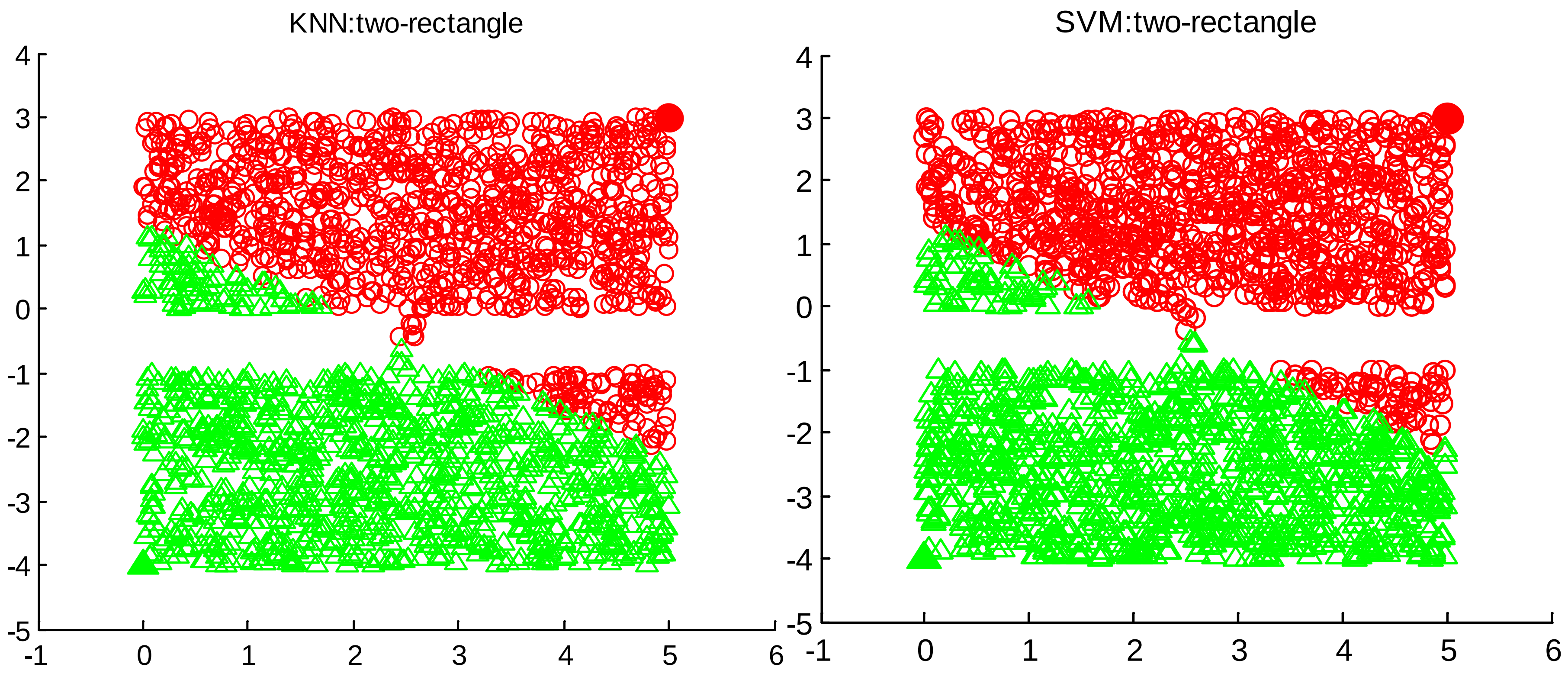
Part 2

Semi-Supervised Learning

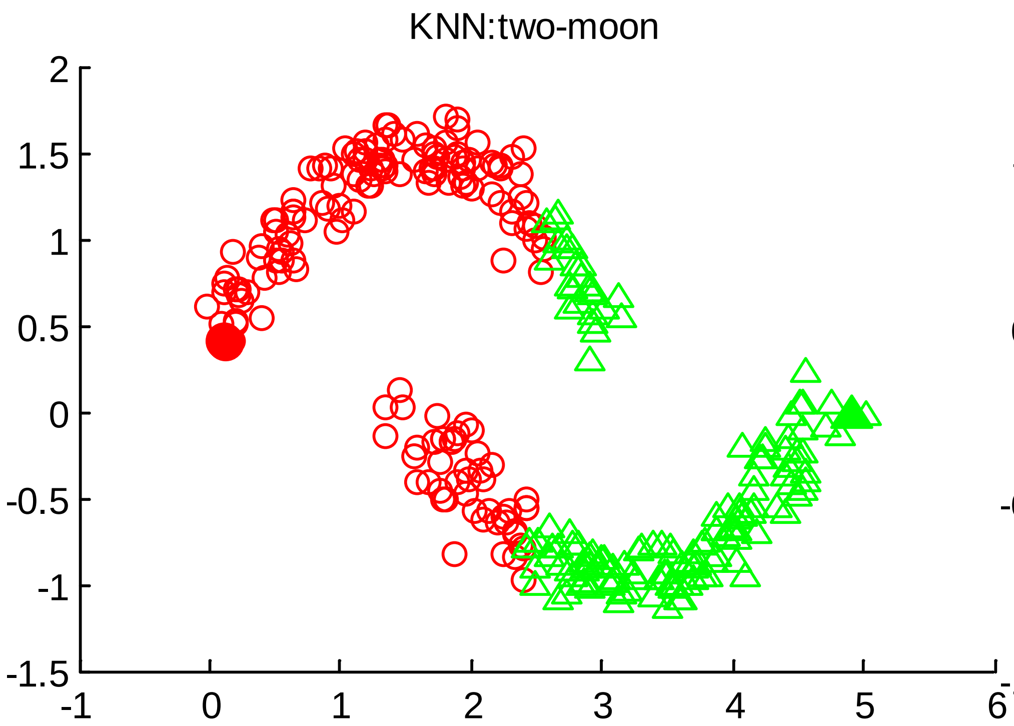
Problems of Supervised Learning

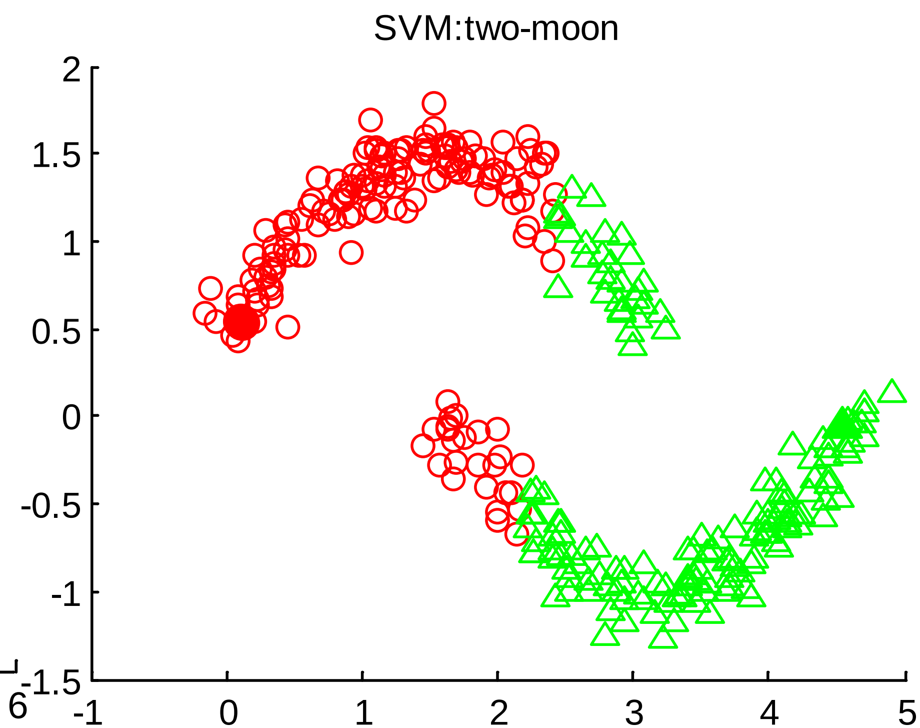
- High-quality labelled samples are often difficult to obtain
- Training instances are not uniformly sampled
- Sensitive to noise in training samples

Supervised learning results



Supervised learning results





Categories of SSL algorithms

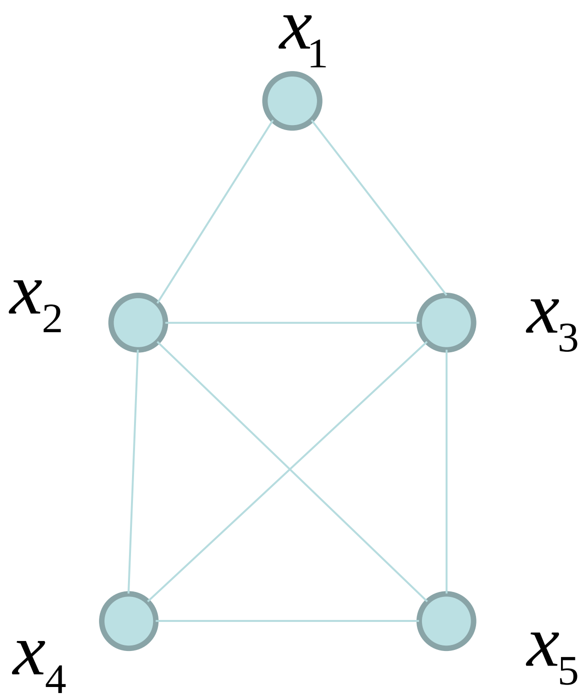
SSL

- probabilistic generative model
- semi-supervised SVM
- graph-based model √
- Co-training
- ...

Advantages of graph-based SSL:

1) mathematical background,
2) compact algebraic linear forms,
3) good results in computational biology, web mining, or text categorization, etc.

Basic Conceptions



Graph or Network: $G = (V, E)$

Adjacency Matrix: $W: \begin{cases} w_{ij} = 1, & if\ (i, j) \in E \\ w_{ij} = 0, & if\ (i, j) \notin E \end{cases}$

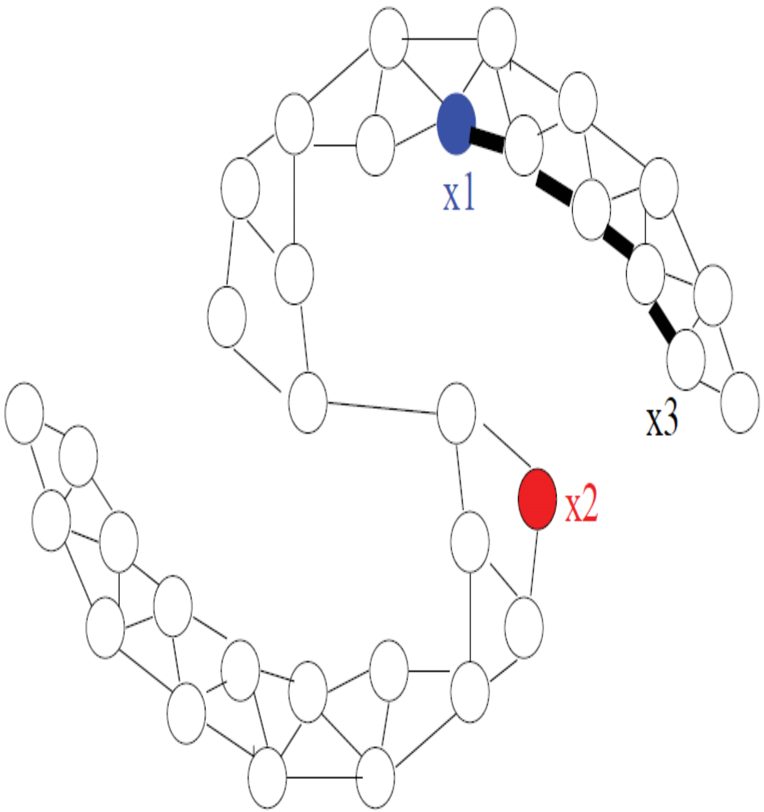
Degree Matrix: $D: \{d_{ii} = vol(i)\}$

Laplacian Matrix: $L = D - W$

Properties:

a) semi-positive definitive
b) multiplicity of λ_1 is equal to number of connected component
c) exist a group orthogonal eigenvectors

Graph Construction



Construct Graph $G = (V,E)$

vertices: $V = \{x_1, x_2, \ldots, x_n\}$

Edges: $E : \mathbf{W} = \{w_{ij} \geq 0\}$

Edge types:

1)Weight: $w_{ij} = \exp(\| x_i - x_j \|^2 / 2\sigma^2)$

2)Knn: $w_{ij} = 1, \ \ if \ \ x_j \in \{Knn \ of \ \ x_i\}$

3)$\mathcal{E}$nn: $w_{ij} = 1, \ \ if \ \| x_j - x_i \| < \varepsilon$

A Learning Model

- To learn a function f in graph meeting two constraints

$$S(\mathbf{f}) = \min \sum_{i \sim j} w_{ij} (f_i - f_j)^2$$

1) Smoothness:

$$C(\mathbf{f}) = \min \frac{1}{k} \sum_i (f_i - y_i)^2$$

2) Consistency:

$$F(\mathbf{f}) = S(\mathbf{f}) + C(\mathbf{f}) = \frac{1}{k} \sum_i (f_i - y_i)^2 + \gamma \mathbf{f}^T \mathbf{L} \mathbf{f}$$

- Final objective function

$$s.t. \quad <\mathbf{f}, \mathbf{e}> = 0, \quad \mathbf{e} = (1,1,1,\ldots,1)^T$$

LPDGL : Label Prediction via Deformed Graph Laplacian for Semi-supervised Learning

LPDGL: Deformed Graph Laplacian for Semi-supervised Learning

Motivation:

This paper introduces Deformed Graph Laplacian (DGL) and presen… for semi-superv… used in LPDG… neighbors locally… y properly dealin…

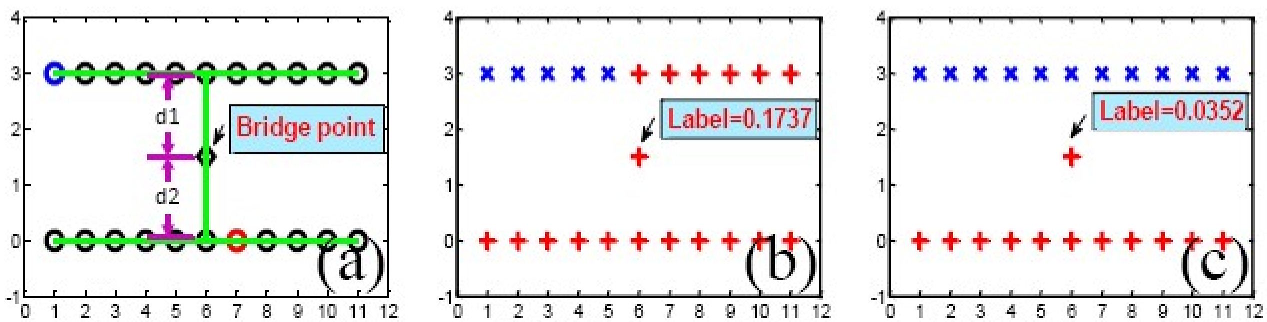


Fig. 1: The illustration of local smoothness constraint on *DoubleLine* dataset. A k-NN graph with $k = 2$ is built and the edges are shown as green lines in (a). (b) shows the result without incorporating the local smoothness, and (c) is the result produced by the proposed LPDGL. The labels of "bridge point" under two different simulations are highlighted in (b) and (c), respectively.

Reference: Deformed Graph Laplacian for Semi-supervised Learning, Chen Gong, Dacheng Tao, Keren Fu, Enmei Tu, Jie Yang, *accepted by TNNLS,* 2015.

Advantages:

- A novel local smoothness term is introduced "naturally", which is critical for our SSL model to better deal with ambiguous examples;
- LPDGL is able to achieve higher classification accuracy than some state-of-the-art methods for both transductive and inductive settings;
- LPDGL can be regarded as a unified framework of many popular SSL algorithms

Deformed graph Laplacian:

$$\tilde{\mathbf{L}}(\kappa) = \kappa(\mathbf{I} - \mathbf{W}/v) - \kappa^2(\mathbf{I} - \mathbf{D}/v).$$

The proposed regularizer:

$$\begin{aligned}
\mathbf{f}^T\tilde{\mathbf{L}}\mathbf{f} &= \mathbf{f}^T\left[\kappa(\mathbf{I} - \mathbf{W}/v) - \kappa^2(\mathbf{I} - \mathbf{D}/v)\right]\mathbf{f} \\
&= (\kappa - \kappa^2)\mathbf{f}^T\mathbf{f} - \frac{\kappa}{v}\mathbf{f}^T\mathbf{W}\mathbf{f} + \frac{\kappa^2}{v}\mathbf{f}^T(\mathbf{D} - \mathbf{W})\mathbf{f} + \frac{\kappa^2}{v}\mathbf{f}^T\mathbf{W}\mathbf{f} \\
&= (\kappa - \kappa^2)\mathbf{f}^T\mathbf{f} + \frac{\kappa^2}{v}\mathbf{f}^T\mathbf{L}\mathbf{f} + (\kappa - 1)\frac{\kappa}{v}\mathbf{f}^T\mathbf{W}\mathbf{f} - (\kappa - 1)\frac{\kappa}{v}\mathbf{f}^T\mathbf{D}\mathbf{f} \\
&\quad + (\kappa - 1)\frac{\kappa}{v}\mathbf{f}^T\mathbf{D}\mathbf{f} \\
&= \frac{\kappa}{v}\mathbf{f}^T\mathbf{L}\mathbf{f} + (\kappa - \kappa^2)\mathbf{f}^T(\mathbf{I} - \mathbf{D}/v)\mathbf{f} \\
&= \beta\mathbf{f}^T\mathbf{L}\mathbf{f} + \gamma\mathbf{f}^T(\mathbf{I} - \mathbf{D}/v)\mathbf{f},
\end{aligned}$$

Theatrical Analyses (Robustness)

Theorem 5: Let χ denote the input space, and $\forall \mathbf{x}_i, \mathbf{x}_j \in \chi, \|\mathbf{x}_i - \mathbf{x}_j\| \le \eta$. A k-NN graph is built with the edge weights represented by RBF kernel $\omega_{ij} = \exp\left(-\|\mathbf{x}_i - \mathbf{x}_j\|^2 / (2\sigma^2)\right)$. Under $\mathcal{N}(\varepsilon/2, \mathcal{X}, \|\cdot\|_2) < \infty$, the proposed LPDGL is $\left(\theta, 2\sqrt{\frac{nl}{\alpha}}\left(1+\sqrt{\frac{nl}{\alpha}}\right)\left[1-\exp\left(-\frac{\varepsilon^2+2\varepsilon\eta}{2\sigma^2}\right)\right]\right)$-robust.

Theatrical Analyses (Generalization)

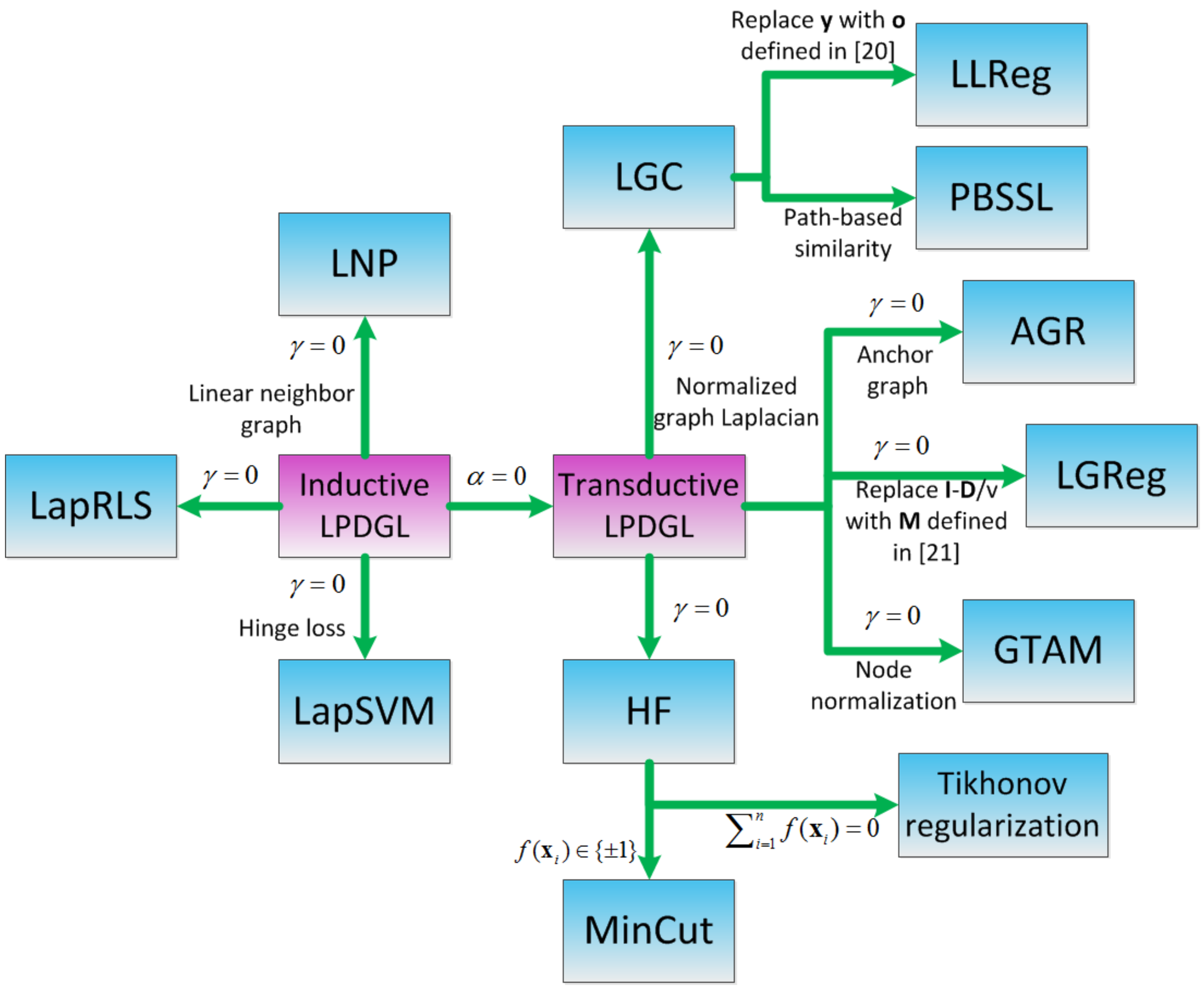
Theorem 7: Let $L(f,\Psi)=\frac{1}{2}\|\mathbf{y}-\mathbf{JKS}\|^2$ be the loss function of LPDGL, than

for any $\delta>0$, the generalization error of LPDGL is

$$\mathrm{Prob}\left[\left|\tilde{L}(\mathcal{A}_\Psi)-L_{emp}(\mathcal{A}_\Psi)\right|\geq 1-\delta\right]$$
$$\leq\left(2\sqrt{\frac{nl}{\alpha}}+\frac{2nl}{\alpha}\right)\left[1-\exp\left(-\frac{\varepsilon^2+2\varepsilon\eta}{2\sigma^2}\right)\right]+\left(\frac{l}{2}+\frac{nl}{\sqrt{\alpha}}+\frac{n^2l^2}{2\alpha}\right)\sqrt{\frac{2K\ln 2+2\ln(1/\delta)}{n}}.$$

LPDGL: Deformed Graph Laplacian for Semi-supervised Learning

Relationship with existing SSL algorithms:



Experimental Results

Synthetic Data

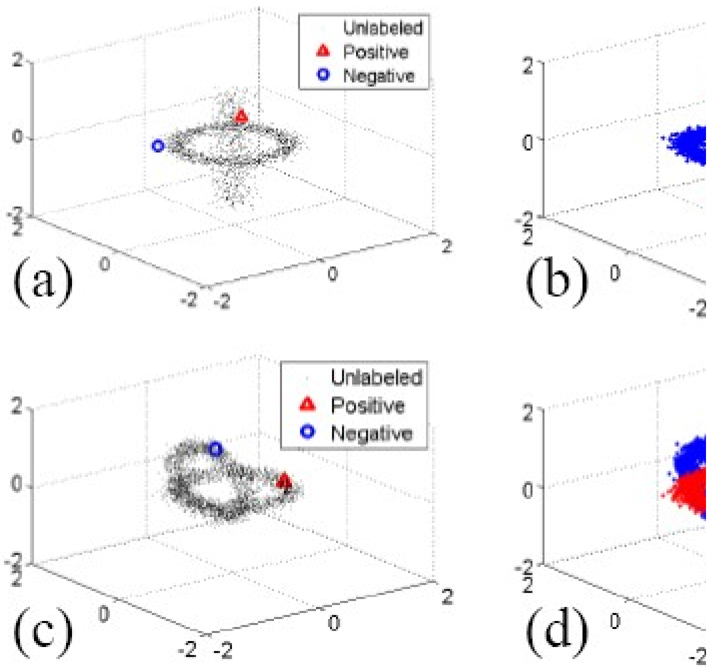


Fig. 3: Transduction on two 3D datasets: (
the initial states of *Cylinder&Ring* and *Knot*
which the red triangle denotes a positive e
blue circle represents a negative example. (b
transduction results of LPDGL on these two

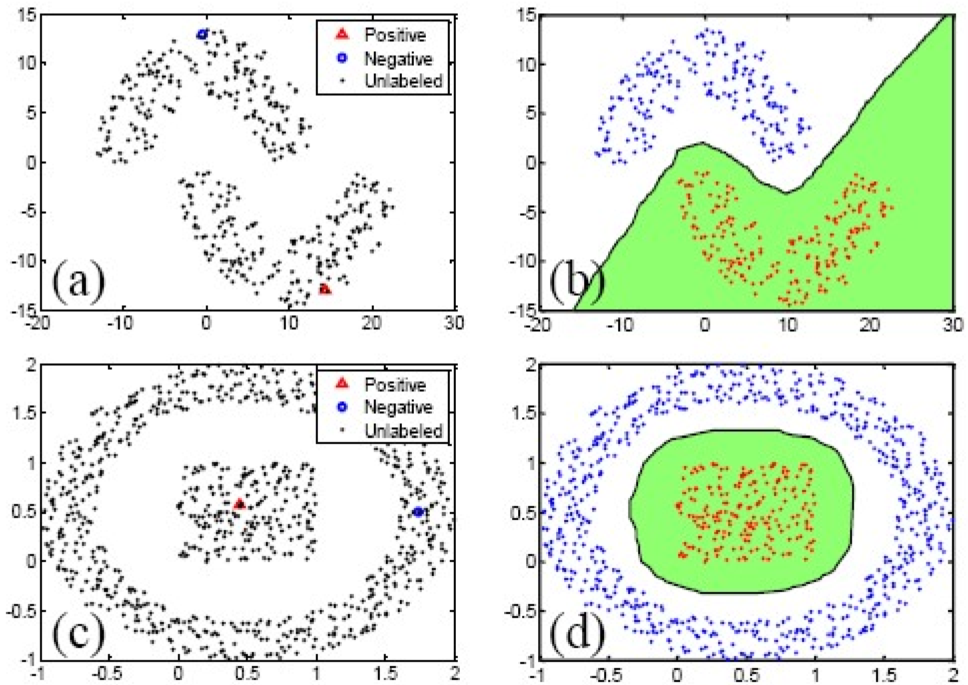


Fig. 4: Induction on *DoubleMoon* and *Square&Ring* datasets. (a) and (c) show the initial states with the marked labeled examples. (b) and (d) are induction results, in which the decision boundaries are plotted.

Experimental Results

UCI Data

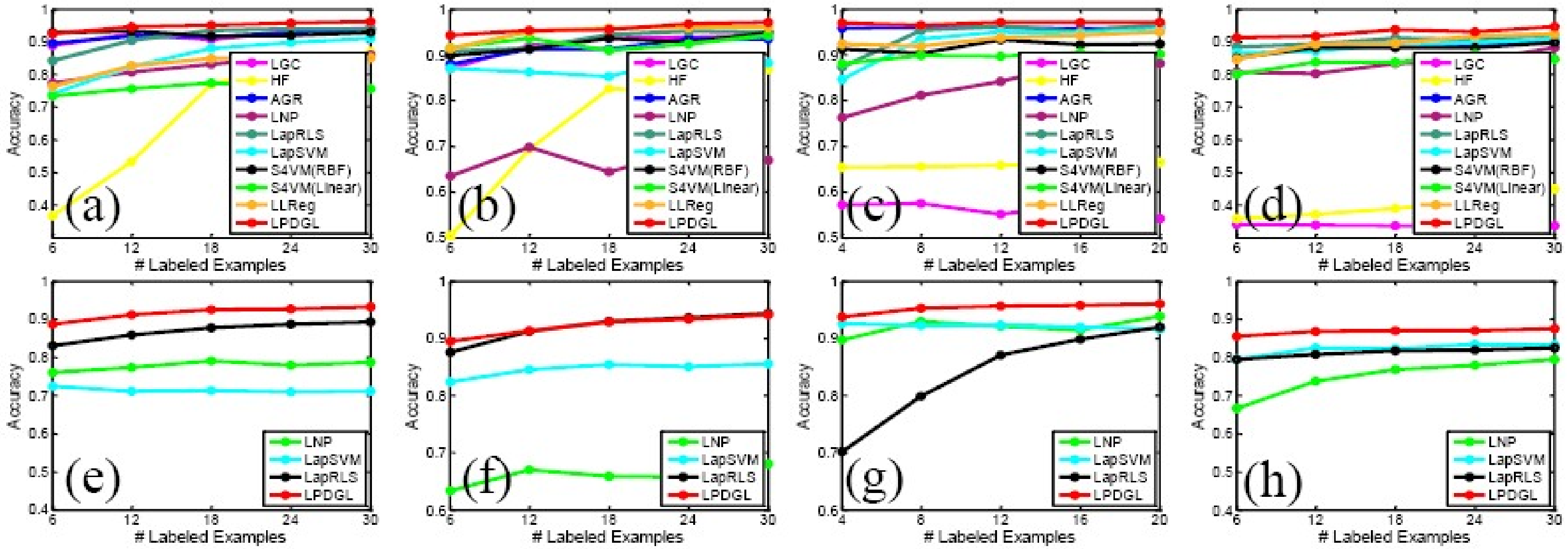


Fig. 5: Experimental results on four UCI datasets. (a) and (e) are *Iris*, (b) and (f) are *Wine*, (c) and (g) are *BreastCancer*, and (d) and (h) are *Seeds*. The sub-plots in the first row compare the transductive performance of the algorithms, and the sub-plots in the second row compare their inductive performance.

Experimental Results

USPS data

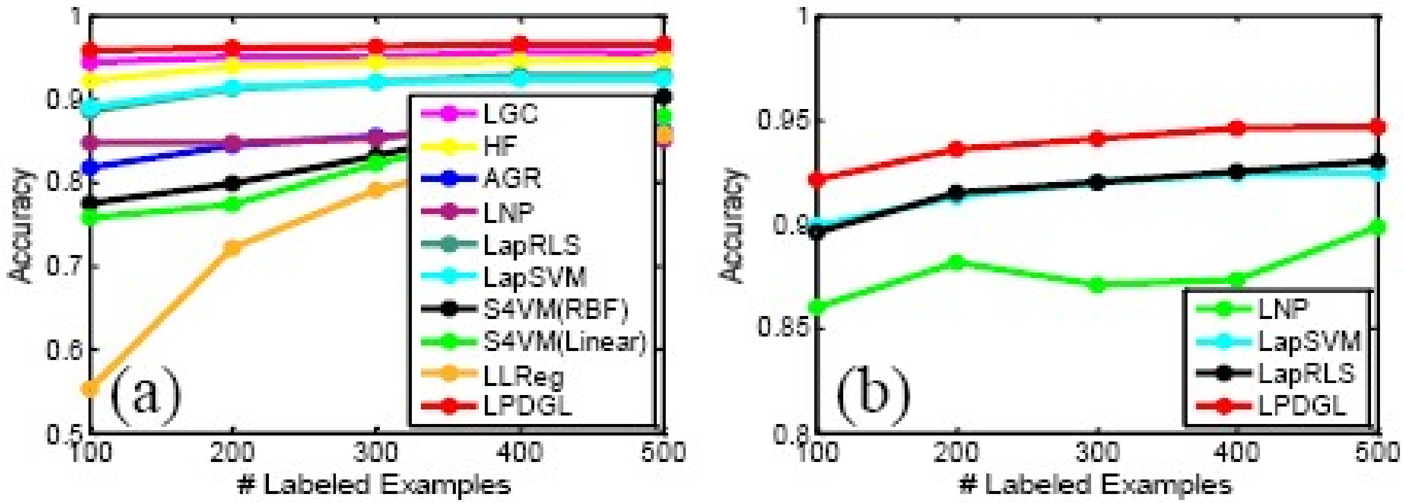


Fig. 7: Experimental results on *USPS* dataset. (a) shows the transductive results, and (b) shows the inductive results.

SHANGHAI JIAO TONG UNIVERSITY 上海交通大学

Experimental Results

Face recognition (Yale data)

Indiv

Indiv

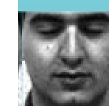


TABLE IV: Transductive comparison on *Yale* dataset

	$l = 30$	$l = 60$
LGC	0.66 ± 0.06	0.76 ± 0.02
HF	0.65 ± 0.04	0.79 ± 0.01
AGR	0.50 ± 0.03	0.64 ± 0.02
LNP	0.32 ± 0.05	0.34 ± 0.04
LapRLS	0.63 ± 0.05	0.71 ± 0.03
LapSVM	0.63 ± 0.05	0.72 ± 0.03
S4VM(Linear)	0.27 ± 0.07	0.52 ± 0.06
S4VM(RBF)	0.11 ± 0.02	0.23 ± 0.04
LLReg	0.65 ± 0.08	0.79 ± 0.09
LPDGL	0.67 ± 0.03	0.81 ± 0.01

TABLE V: Inductive comparison on *Yale* dataset

	$l = 30$	$l = 60$
LNP	0.10 ± 0.04	0.15 ± 0.05
LapSVM	0.69 ± 0.01	0.77 ± 0.01
LapRLS	0.68 ± 0.01	0.79 ± 0.01
LPDGL	0.69 ± 0.04	0.83 ± 0.03

Experimental Results

Face recognition (LFW data)

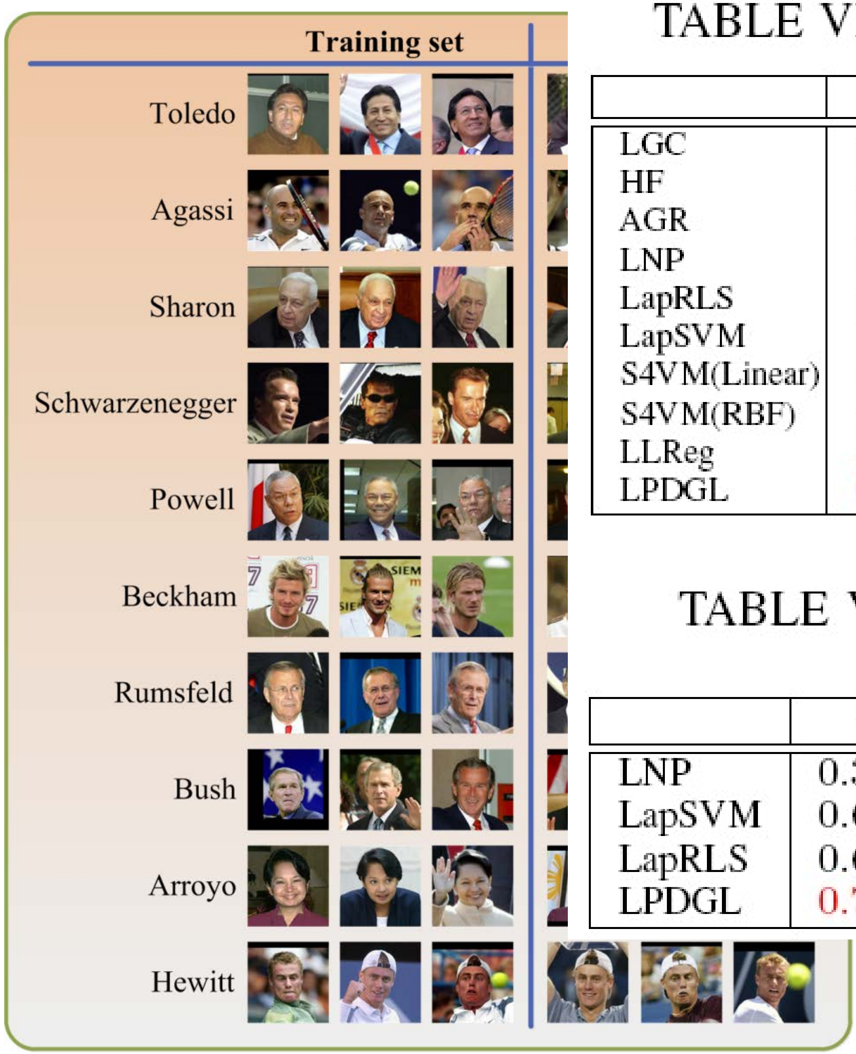


TABLE VI: Transductive comparison on *LFW* dataset

	$l = 50$	$l = 100$	$l = 150$	$l = 200$
LGC	0.50 ± 0.07	0.60 ± 0.05	0.65 ± 0.08	0.69 ± 0.06
HF	0.66 ± 0.03	0.78 ± 0.02	0.83 ± 0.01	0.87 ± 0.01
AGR	0.60 ± 0.03	0.71 ± 0.01	0.76 ± 0.02	0.80 ± 0.01
LNP	0.32 ± 0.07	0.38 ± 0.16	0.57 ± 0.12	0.59 ± 0.11
LapRLS	0.48 ± 0.03	0.62 ± 0.04	0.71 ± 0.03	0.75 ± 0.03
LapSVM	0.57 ± 0.02	0.70 ± 0.03	0.74 ± 0.03	0.76 ± 0.03
S4VM(Linear)	0.56 ± 0.05	0.68 ± 0.03	0.73 ± 0.03	0.77 ± 0.02
S4VM(RBF)	0.45 ± 0.06	0.61 ± 0.02	0.70 ± 0.02	0.73 ± 0.02
LLReg	0.52 ± 0.04	0.69 ± 0.02	0.86 ± 0.02	0.88 ± 0.01
LPDGL	0.71 ± 0.02	0.81 ± 0.02	0.86 ± 0.01	0.90 ± 0.01

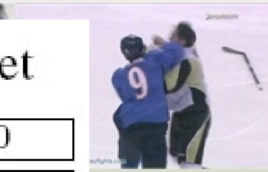
TABLE VII: Inductive comparison on *LFW* dataset

	$l = 50$	$l = 100$	$l = 150$	$l = 200$
LNP	0.30 ± 0.07	0.38 ± 0.09	0.45 ± 0.13	0.45 ± 0.09
LapSVM	0.65 ± 0.01	0.69 ± 0.03	0.75 ± 0.02	0.76 ± 0.01
LapRLS	0.67 ± 0.04	0.73 ± 0.02	0.78 ± 0.01	0.79 ± 0.01
LPDGL	0.70 ± 0.03	0.78 ± 0.03	0.80 ± 0.02	0.83 ± 0.02

Experimental Results

Fight detection (HockeyFight data)

Fight:



Non-fig

TABLE VIII: Transductive results on *HockeyFight* dataset

	$l = 40$	$l = 80$	$l = 120$	$l = 160$
LGC	0.80 ± 0.03	0.82 ± 0.02	0.83 ± 0.02	0.84 ± 0.01
HF	0.80 ± 0.02	0.84 ± 0.01	0.86 ± 0.01	0.87 ± 0.01
AGR	0.79 ± 0.02	0.82 ± 0.01	0.83 ± 0.01	0.83 ± 0.01
LNP	0.61 ± 0.08	0.65 ± 0.10	0.65 ± 0.09	0.67 ± 0.11
LapRLS	0.72 ± 0.02	0.76 ± 0.01	0.79 ± 0.01	0.79 ± 0.01
LapSVM	0.67 ± 0.03	0.66 ± 0.02	0.70 ± 0.02	0.71 ± 0.01
S4VM(Linear)	0.80 ± 0.05	0.84 ± 0.02	0.84 ± 0.03	0.86 ± 0.01
S4VM(RBF)	0.81 ± 0.03	0.84 ± 0.01	0.86 ± 0.01	0.87 ± 0.01
LLReg	0.78 ± 0.04	0.79 ± 0.01	0.82 ± 0.01	0.82 ± 0.01
LPDGL	0.81 ± 0.03	0.85 ± 0.01	0.87 ± 0.01	0.88 ± 0.01

TABLE IX: Inductive results on *HockeyFight* dataset

	$l = 40$	$l = 80$	$l = 120$	$l = 160$
LNP	0.58 ± 0.12	0.58 ± 0.08	0.58 ± 0.10	0.59 ± 0.11
LapSVM	0.59 ± 0.02	0.61 ± 0.01	0.61 ± 0.01	0.65 ± 0.01
LapRLS	0.70 ± 0.01	0.73 ± 0.01	0.73 ± 0.01	0.74 ± 0.01
LPDGL	0.71 ± 0.02	0.73 ± 0.03	0.74 ± 0.02	0.75 ± 0.01

上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

Experimental Results

Parametric Sensitivity

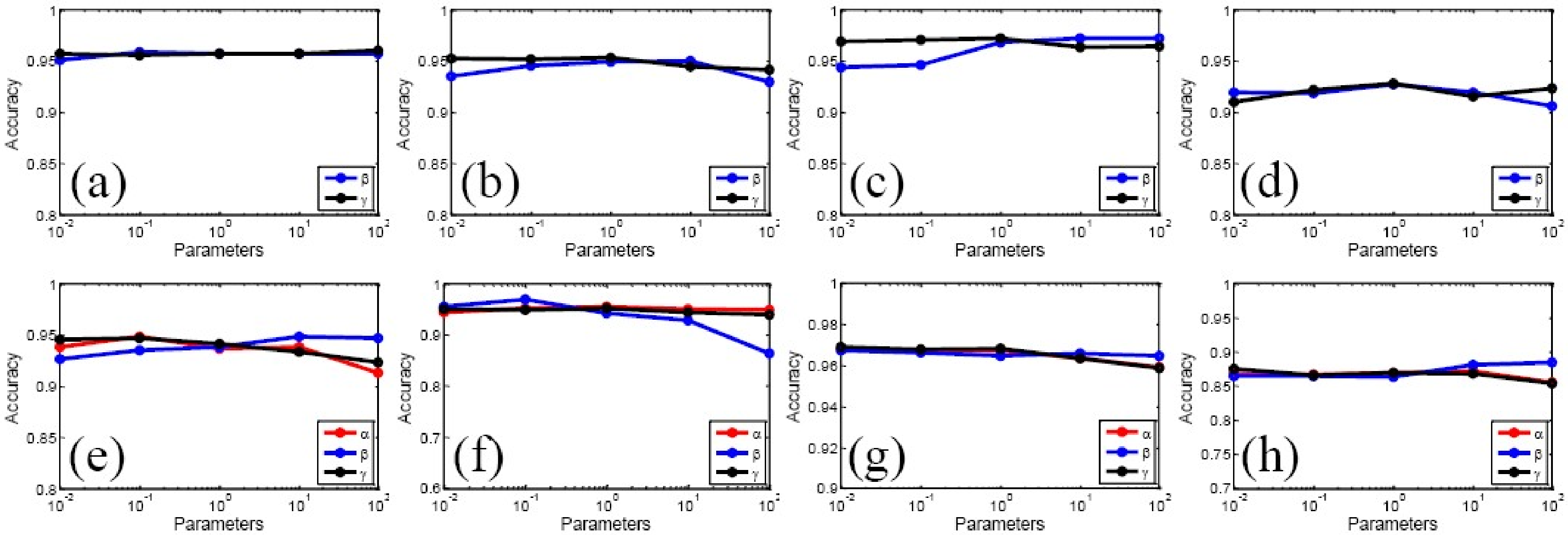


Fig. 6: Empirical studies on the parametric sensitivity of LPDGL. (a) and (e) are *Iris*, (b) and (f) are *Wine*, (c) and (g) are *BreastCancer*, and (d) and (h) are *Seeds*. The sub-plots in the first row show the transductive results, and the sub-plots in the second row display the inductive results.

Conclusion:

The performance of LPDGL is not sensitive to the choice of parameters.

Conclusion

1. Given a few labelled samples, semi-supervised learning generally performs much better than supervised learning, like SVM

2. semi-supervised learning algorithms are more robust to noise

Multi-modal Curriculum Learning for Semi-supervised Image Classification

Background—Multi-modal learning

- In practical applications, data is often obtained from multiple sources rather than a single source.
- Multi-modal learning (MML) is therefore proposed to explicitly fuse the complementary information from different modalities to achieve improved performance.
- MML algorithms can be classified into three groups (arXiv 15): co-training (COLT 98), multiple kernel learning (JMLR 04), and subspace learning (NIPS 12).

Background—Curriculum learning

- Curriculum learning (ICML 09) aims to improve the learning performance by designing suitable curriculums from simple to difficult for the stepwise learner.
- curriculum learning is able to boost the convergence speed of the training process as well as find a better local minima than the existing solvers for non-convex problems
- The existing curriculum learning algorithms can be divided into two categories: self-paced learning (NIPS 10; MM 13; NIPS 14), and teaching-to-learn and learning-to-teach (CVPR 15; TNNLS 16; AAAI 16).

Motivation- Why curriculum learning?

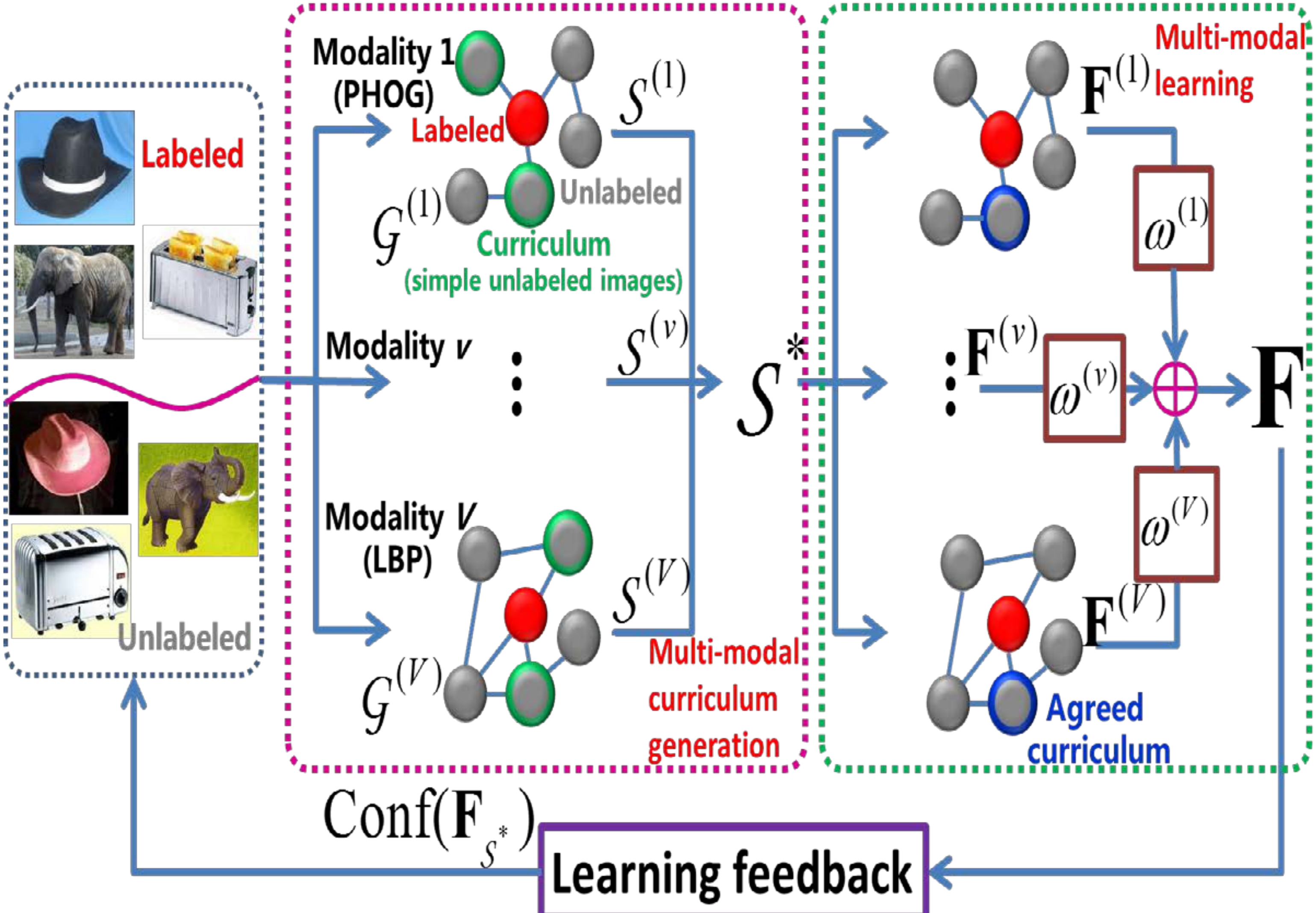
- Existing SSL methods often yield unsatisfactory results, as they are very likely to make incorrect classifications on “outliers” or “bridge examples”. This is because existing methods treat all the unlabeled images equally without considering the difficulty or reliability of their classification.
- We assume that different images have different levels of difficulty and utilize curriculum learning to re-organize the learning sequence, so that the unlabeled images are logically classified from simple to difficult.
- the previously attained simple knowledge to facilitate the subsequent classification of complex images.

Motivation - Why multi-modal learning?

- An image can usually be characterized by different feature descriptors.
- We regard each type of features as one modality and develop “Multi-Modal Curriculum Learning” (MMCL) to guide the learning process. As a result, the consistency and complementarity of various features can be fully exploited.

Our MMCL strategy is very similar to the human’s acquisition of knowledge during the various stages from childhood to adulthood, during which time an individual gains knowledge from many teachers of different subjects.

Algorithm--Framework



Single-modal Curriculum Generation

The **reliability** and **discriminability** of every unlabeled image are investigated by the "teacher" to make a selection.

Reliability:

- A curriculum $\mathcal{S}$ is reliable w.r.t. the labeled set $\mathcal{L}$ if the conditional entropy $H(y_{\mathcal{S}}|y_{\mathcal{L}})$ is small.
- Small $H(y_{\mathcal{S}}|y_{\mathcal{L}})$ suggests that the curriculum set $\mathcal{S}$ comes as no "surprise" to the labeled set $\mathcal{L}$.

Therefore

$$
\begin{aligned}
&\min_{\mathcal{S}\subset\mathcal{U}} H(\mathbf{y}_{\mathcal{S}}|\mathbf{y}_{\mathcal{L}}) \\
\Leftrightarrow &\min_{\mathcal{S}\subset\mathcal{U}} H(\mathbf{y}_{\mathcal{S}\cup\mathcal{L}}) - H(\mathbf{y}_{\mathcal{L}}) \\
\Leftrightarrow &\min_{\mathcal{S}\subset\mathcal{U}} \left(\frac{s+l}{2}(1+\ln 2\pi) + \frac{1}{2}\ln\left|\Sigma_{\mathcal{S}\cup\mathcal{L},\mathcal{S}\cup\mathcal{L}}\right|\right) \\
&\qquad - \left(\frac{l}{2}(1+\ln 2\pi) + \frac{1}{2}\ln\left|\Sigma_{\mathcal{L},\mathcal{L}}\right|\right) \\
\Leftrightarrow &\min_{\mathcal{S}\subset\mathcal{U}} \frac{s}{2}(1+\ln 2\pi) + \frac{1}{2}\ln\frac{\left|\Sigma_{\mathcal{S}\cup\mathcal{L},\mathcal{S}\cup\mathcal{L}}\right|}{\left|\Sigma_{\mathcal{L},\mathcal{L}}\right|},
\end{aligned}
$$

Considering that

$$
\frac{\left|\Sigma_{\mathcal{S}\cup\mathcal{L},\mathcal{S}\cup\mathcal{L}}\right|}{\left|\Sigma_{\mathcal{L},\mathcal{L}}\right|} = \frac{\left|\Sigma_{\mathcal{L},\mathcal{L}}\right|\left|\Sigma_{\mathcal{S},\mathcal{S}} - \Sigma_{\mathcal{S},\mathcal{L}}\Sigma_{\mathcal{L},\mathcal{L}}^{-1}\Sigma_{\mathcal{L},\mathcal{S}}\right|}{\left|\Sigma_{\mathcal{L},\mathcal{L}}\right|} = \left|\Sigma_{\mathcal{S}|\mathcal{L}}\right|
$$

The optimization problem regarding reliability is

$$
\min_{\mathcal{S}\subseteq\mathcal{U}} \operatorname{tr}\left(\Sigma_{\mathcal{S},\mathcal{S}} - \Sigma_{\mathcal{S},\mathcal{L}}\Sigma_{\mathcal{L},\mathcal{L}}^{-1}\Sigma_{\mathcal{L},\mathcal{S}}\right)
$$

Discriminability:

- A curriculum is discriminable if the included images are significantly inclined to certain classes.
- The tendency of an image $\boldsymbol{x}_i$ belonging to a class $\mathcal{C}_j$ is modeled by the average commute time between $\boldsymbol{x}_i$ and all the images in $\mathcal{C}_j$.

average commute time: $\bar{T}(\mathrm{x}_i, \mathcal{C}_j) = \frac{1}{n_{\mathcal{C}_j}} \sum_{\mathrm{x}_{i'} \in \mathcal{C}_j} T(\mathrm{x}_i, \mathrm{x}_{i'})$ where

$$T(\mathrm{x}_i, \mathrm{x}_{i'}) = \sum_{k=1}^{n} h(\lambda_k)\left(u_{ki} - u_{ki'}\right)^2$$ (PAMI 07)

Therefore, $\boldsymbol{x}_i$ is discriminable if the gap $M(\mathrm{x}_i) = \bar{T}(\mathrm{x}_i, \mathcal{C}_2) - \bar{T}(\mathrm{x}_i, \mathcal{C}_1)$ is large.

Here $\mathcal{C}_1$ and $\mathcal{C}_2$ are the two closest classes to $\boldsymbol{x}_i$ measured by average commute time.

The simplest curriculum in view of discriminability is found by solving

$$\min_{\mathcal{S}=\{\mathrm{x}_{i_k} \in \mathcal{U}\}_{k=1}^{s}} \sum_{k=1}^{s} 1/M(\mathrm{x}_{i_k})$$

By combining reliability and discriminability, we arrive at the following optimization problem:

$$\min_{\mathcal{S}=\{\mathrm{x}_{i_k}\in\mathcal{U}\}_{k=1}^{s}} \mathrm{tr}\left(\Sigma_{\mathcal{S},\mathcal{S}}-\Sigma_{\mathcal{S},\mathcal{L}}\Sigma_{\mathcal{L},\mathcal{L}}^{-1}\Sigma_{\mathcal{L},\mathcal{S}}\right)+\sum_{k=1}^{s}1/M(\mathrm{x}_{i_k})$$

To make it tractable, we introduce a binary selection matrix $\mathbf{S}\in\{1,0\}^{b\times s}$

The element $\boldsymbol{S}_{ij}=1$ means that the i-th image is selected as the j-th element in the curriculum $\mathcal{S}$.

The optimization problem can be reformulated to the following matrix form:

$$\begin{aligned}\min_{\mathbf{S}}\ & \mathrm{tr}\left(\mathbf{S}^{\top}\Sigma_{\mathcal{B},\mathcal{B}}\mathbf{S}-\mathbf{S}^{\top}\Sigma_{\mathcal{B},\mathcal{L}}\Sigma_{\mathcal{L},\mathcal{L}}^{-1}\Sigma_{\mathcal{L},\mathcal{B}}\mathbf{S}\right)\\ & +\mathrm{tr}\left(\mathbf{S}^{\top}\mathbf{M}\mathbf{S}\right),\\ \text{s.t.}\ & \mathbf{S}\in\{1,0\}^{b\times s},\ \mathbf{S}^{\top}\mathbf{S}=\mathbf{I}_{s\times s},\end{aligned}$$

The orthogonality constraint ensures that every image is selected only once

上海交通大学 SHANGHAI JIAO TONG UNIVERSITY

Multi-modal Curriculum Generation

High level idea: force the V teachers to reach a consensus on selecting the optimal curriculum $\mathcal{S}^*$

$$\begin{aligned}\min_{\mathbf{S}^{(1)},\cdots,\mathbf{S}^{(V)},\mathbf{S}^*} \ & \sum_{v=1}^{V} \mathrm{tr}\left(\mathbf{S}^{(v)\top}\mathbf{R}^{(v)}\mathbf{S}^{(v)}\right)+\beta\sum_{v=1}^{V}\left\|\mathbf{S}^{(v)}-\mathbf{S}^*\right\|_F^2 \\ \text{s.t.} \ & \mathbf{S}^* \in \{1,0\}^{b\times s}, \ \mathbf{S}^{*\top}\mathbf{S}^* = \mathbf{I}_{s\times s}, \\ & \mathbf{S}^{(v)} \in \{1,0\}^{b\times s}, \ \mathbf{S}^{(v)\top}\mathbf{S}^{(v)} = \mathbf{I}_{s\times s}, \ \text{for } v=1,\cdots,V.\end{aligned}$$

$$\mathrm{R} = \Sigma_{\mathcal{B},\mathcal{B}} - \Sigma_{\mathcal{B},\mathcal{L}}\Sigma_{\mathcal{L},\mathcal{L}}^{-1}\Sigma_{\mathcal{L},\mathcal{B}} + \mathrm{M}$$

Algorithm 1 The algorithm for solving $\mathbf{S}^{(v)}$-subproblem ($\boxed{10}$)

Input: $\mathbf{R}^{(v)}$, $\mathbf{S}^{*}$, $\mathbf{S}^{(v)} \in St$, $\mathbf{\Lambda}^{(v)} = \mathbf{O}$, $\sigma^{(v)} = 1$, $\rho = 1.2$, β, $iter = 0$

repeat

  // Compute $\mathbf{T}^{(v)}$

  $\mathbf{T}_{ij}^{(v)} = \max(0, \ \mathbf{S}_{ij}^{(v)} + \mathbf{\Lambda}_{ij}^{(v)}/\sigma^{(v)})$;

  // Update $\mathbf{S}^{(v)}$ by using Eq. ($\boxed{11}$)

  $\mathbf{S}^{(v)} := \mathrm{Proj}_{St}\left[\mathbf{S}^{(v)} - \tau \nabla_{\mathbf{S}^{(v)}} L\left(\mathbf{S}^{(v)}, \mathbf{\Lambda}^{(v)}, \mathbf{T}^{(v)}, \sigma^{(v)}\right)\right]$;

  // Update variables

  $\mathbf{\Lambda}_{ij}^{(v)} := \max\left(0, \mathbf{\Lambda}_{ij}^{(v)} - \sigma^{(v)} \mathbf{S}_{ij}^{(v)}\right)$;

  $\sigma^{(v)} := \min(\rho\sigma^{(v)}, 10^{10})$; $iter := iter + 1$;

until Convergence

Output: $\mathbf{S}^{(v)}$ that minimizes Eq. ($\boxed{10}$)

Multi-modal Classification with Feedback

We employ the label propagation algorithm (ICML 03) as the learner because it is naturally incremental and does not require retraining with the arrival of a new curriculum.

$$\mathbf{F}_i^{(v)[t]} = \begin{cases} \mathbf{P}_i^{(v)} \mathbf{F}^{[t-1]}, & \mathbf{x}_i \in (\mathcal{S}^{*[1]} \cup \cdots \cup \mathcal{S}^{*[t-1]}) \cup \mathcal{S}^{*[t]} \\ \mathbf{F}_i^{[0]}, & \mathbf{x}_i \in \mathcal{L}^{[0]} \cup (\mathcal{U}^{[0]} - \mathcal{S}^{*[1]} \cup \cdots \cup \mathcal{S}^{*[t]}) \end{cases}$$

the integrated label matrix is computed by:

$$\mathbf{F}^{[t]} = \sum_{v=1}^{V} \omega^{(v)[t]} \mathbf{F}^{(v)[t]}$$

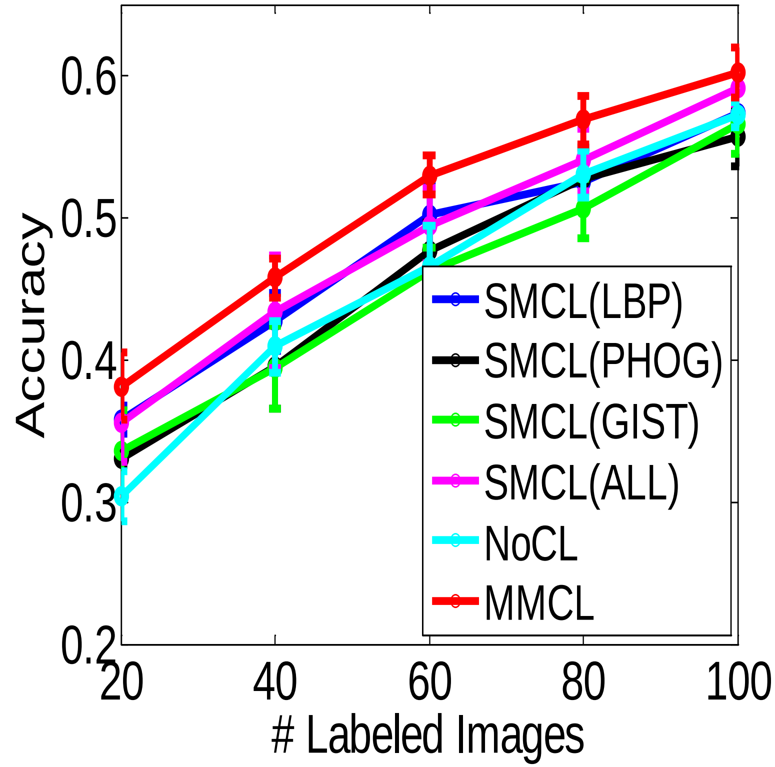
$$\omega^{(v)[t]} = \frac{exp\left(-\left\|\mathbf{S}^{(v)[t]} - \mathbf{S}^{*[t]}\right\|_{\mathrm{F}}^{2}\right)}{\sum_{v=1}^{V} exp\left(-\left\|\mathbf{S}^{(v)[t]} - \mathbf{S}^{*[t]}\right\|_{\mathrm{F}}^{2}\right)}$$

Experiments

All the images in the adopted datasets are represented by the 72-dimensional PHOG, 512-dimensional GIST, and 256-dimensional LBP. (Totally 3 modalities)

We first validate the motivation of our MMCL algorithm on a small database, and then compare MMCL with several state-of-the-art methods on eight practical image datasets.

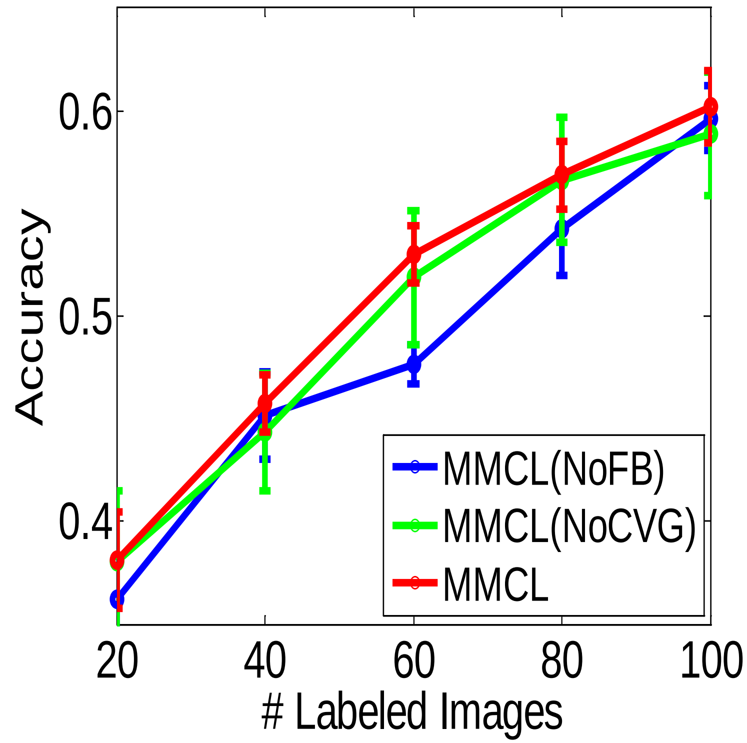
Algorithm validation (1)



Two arguments are demonstrated:

1) curriculum learning is critical to improving classification performance;
2) MMCL is superior to single-modal curriculum learning (SMCL).

Algorithm validation (2)

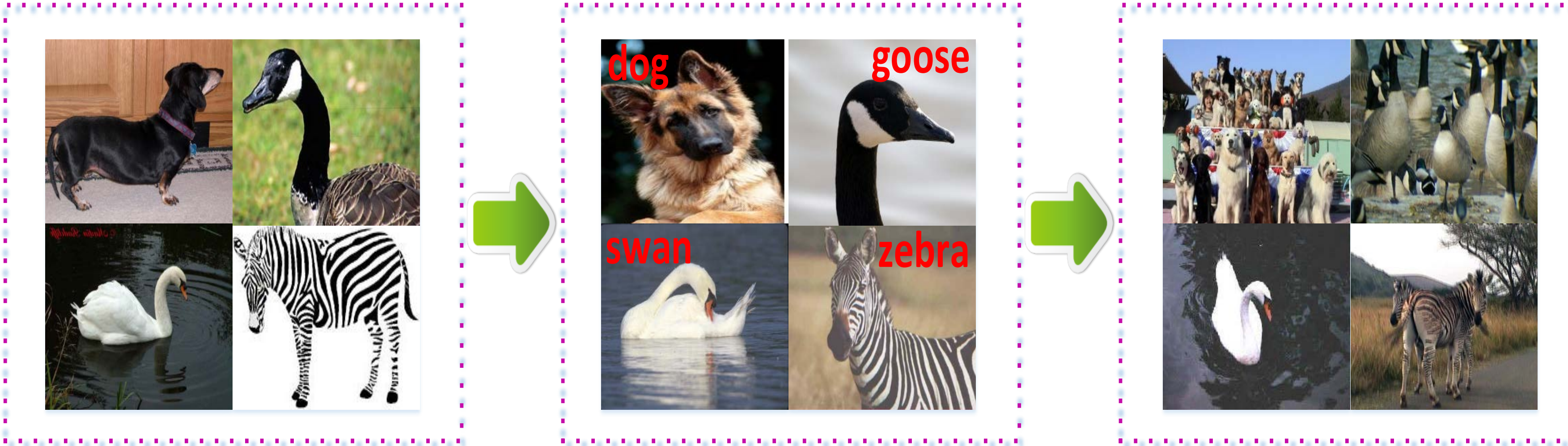


The effectiveness of two key steps in our MMCL is demonstrated:

1) the establishment of learning feedback;
2) the convergence of propagations

Algorithm validation (3)

We visualize the curriculum images selected by our MMCL during the entire teaching and learning process.



$t = 1 \sim 2$ $t = 8 \sim 9$ $t = 11 \sim 14$

The introduced teachers in MMCL can accurately evaluate the difficulty level of every unlabelled image, and effectively organize the entire propagation process so that all the images are classified from simple to difficult.

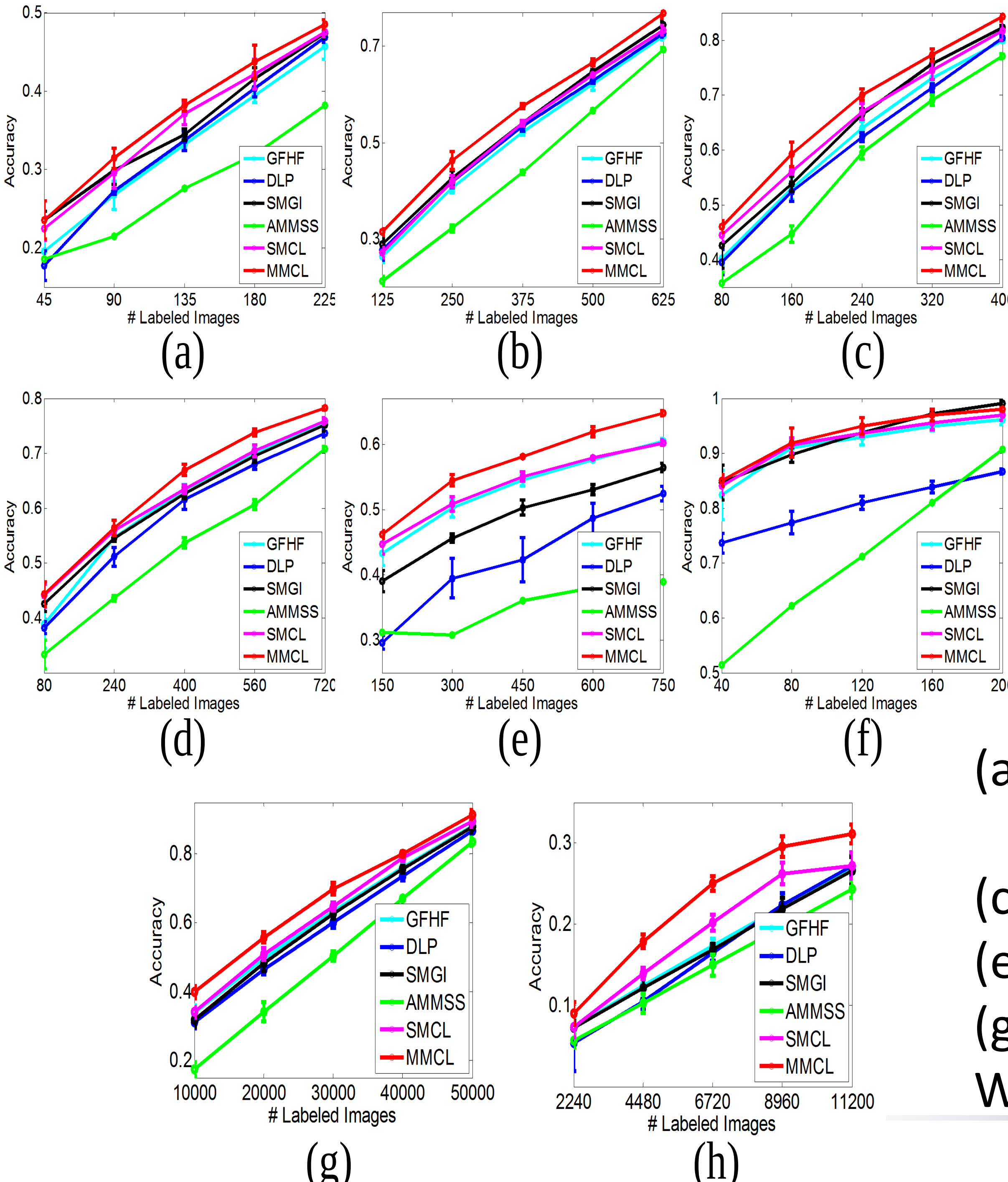
Comparison with other algorithms

Datasets:

	CaltechAnimal	*Architecture*	*MSRC*	*UIUC*	*Scene15*	*ORLFace*	*CIFAR100*	*NUS-WIDE*
# classes	9	25	20	8	15	40	100	112
# images	720	1000	589	1579	4485	400	60000	47254

Baselines:

GFHF: The Gaussian Field and Harmonic Functions (ICML 03)
DLP: Dynamic Label Propagation (ICCV 13)
AMMSS: Adaptive Multi-Modal Semi-Supervised classifier (ICCV 13)
SMGI: Sparse Multiple Graph Integration (SMGI) (TNNLS 13)
SMCL: Single Modal Curriculum Learning



(a) CaltechAnimal (b) is Architecture
(c) MSRC (d) UIUC
(e) Scene15 (f) ORLFace
(g) CIFAR100 (h) NUS-WIDE

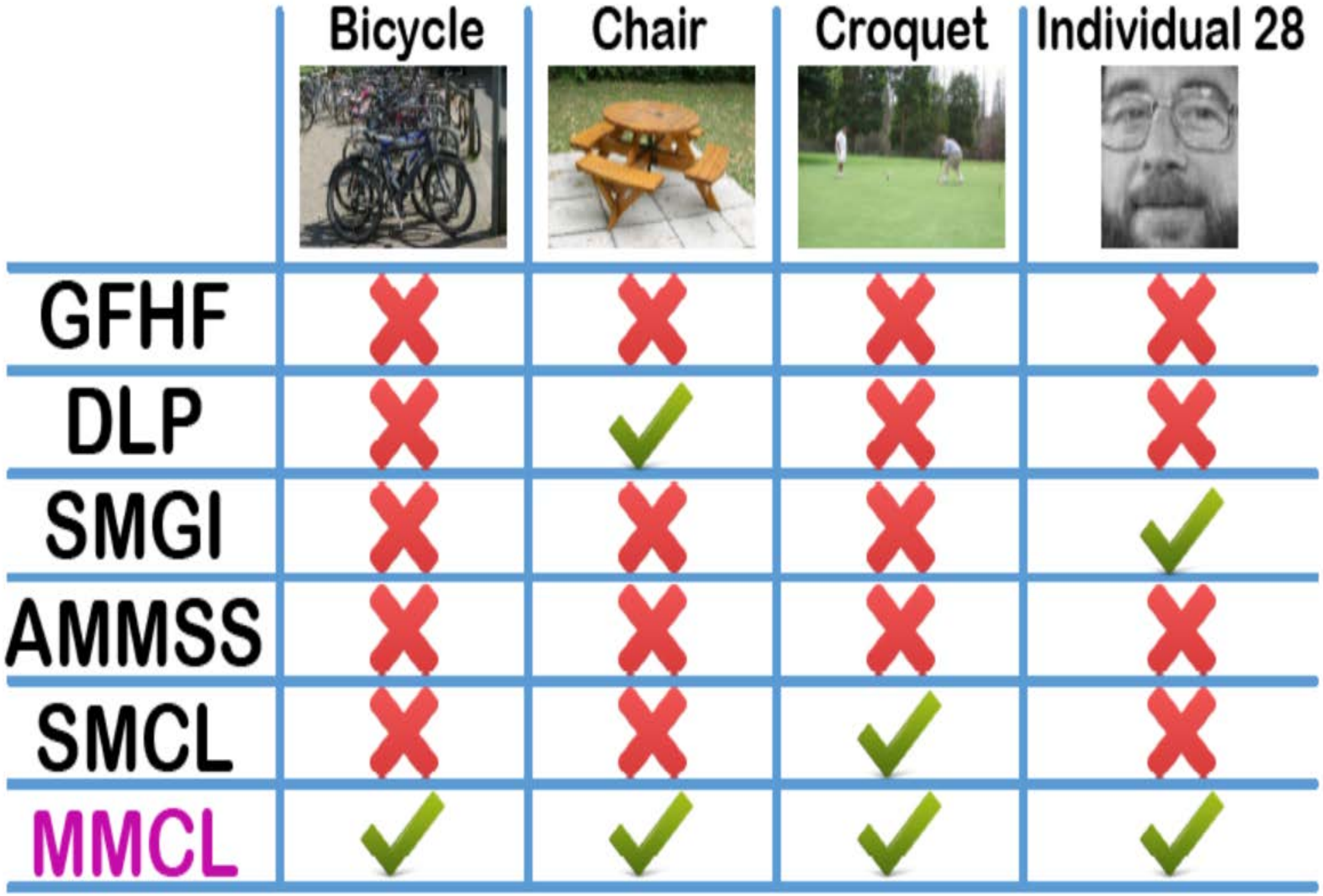


Fig. 6. Classification results of the compared methods on several visually challenging images. The red crosses represent "incorrect classifications" while the green ticks denote "correct classifications".

References

- Saliency Propagation From Simple To Difficult. *IEEE International Conference on Computer Vision and Pattern Recognition (**CVPR**),* 2015.
- Teaching-to-Learn and Learning-to-Teach for Multi-label Propagation, ***AAAI***, 2016. **(oral)**
- Label Propagation via Teaching-to-Learn and Learning-to-Teach, *IEEE Transactions on Neural Networks and Learning Systems (**TNNLS**)*, 2016.
- Multi-modal curriculum learning for semi-supervised image classification. IEEE Transactions on Image Processing (**TIP**), v 25, n 7, p 3249-3260, July 2016.
- Normalized Cut-Based Saliency Detection by Adaptive Multi-Level Region Merging, IEEE Transactions on Image Processing (**TIP**), 24(12):5671-5683,2015
- Deformed Graph Laplacian for Semi-supervised Learning, IEEE Transactions on Neural Networks and Learning Systems ***(TNNLS)***, v 26, n 10, p 2261-2274, October 1, 2015.
- Fick's Law Assisted Propagation for Semi-Supervised Learning, IEEE Transactions on Neural Networks and Learning Systems ***(TNNLS)***, v 26, n 9, p 2148-2162, September 1, 2015.

Thanks!