Noise/Loss Modeling Principle

Deyu Meng

Xi'an Jiaotong University

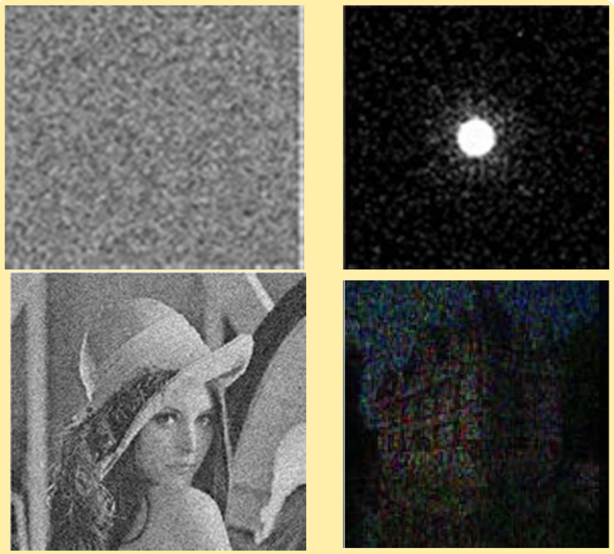
dymeng@mail.xjtu.edu.cn

http://gr.xjtu.edu.cn/web/dymeng

Signal Recovery

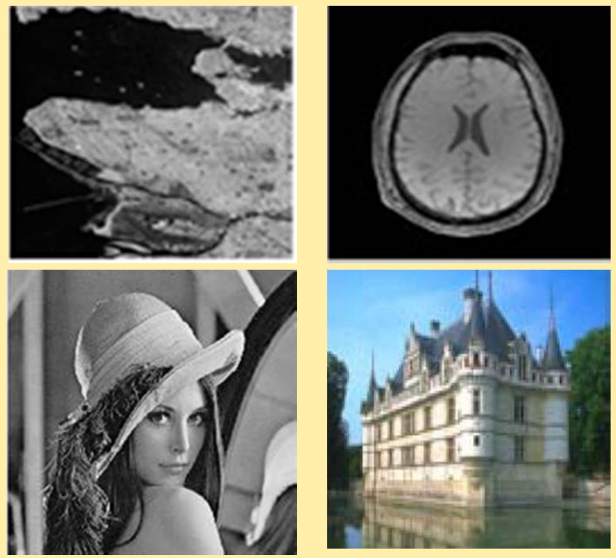
The inverse problem: $X = f(W) + \varepsilon$

Observation X



- Compressive sensing
- Image super-resolution
- Image denoising
- Image deblurring
- ……

Recovery f(W)



A General Machine learning Framework

$$\min_{f \in \Omega} \quad \mathcal{L}(f(\mathbf{W}), \mathbf{X}) \;+\; \mathrm{R}(\mathbf{W})$$

Learning machine

Loss/likelihood term

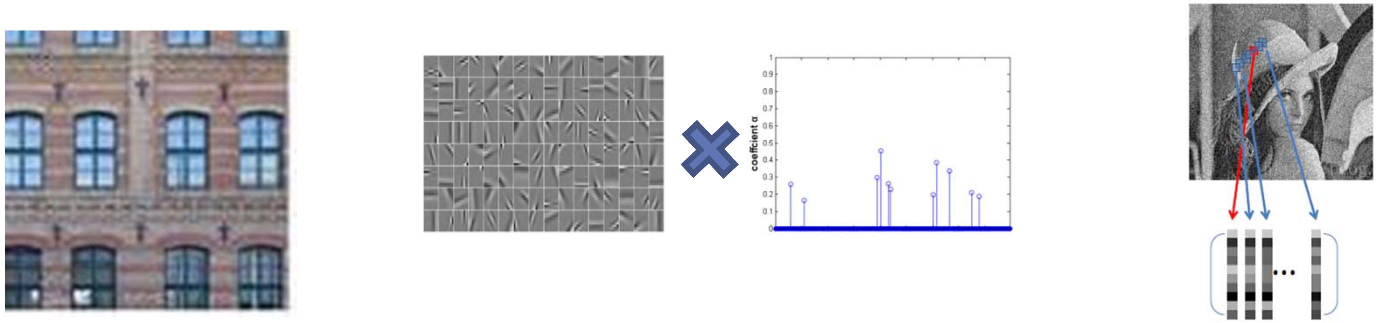
Regularization/prior term

A General Machine learning Framework

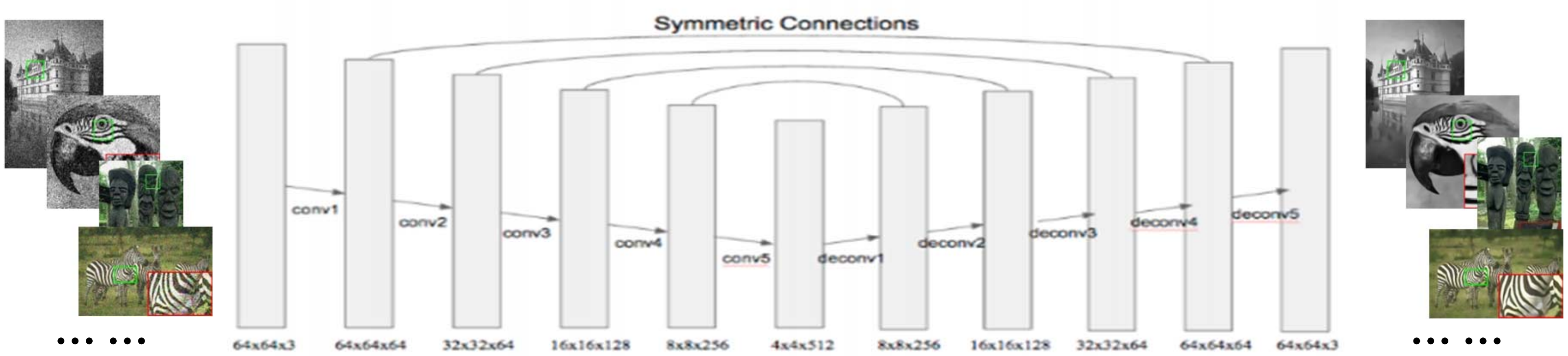
$$\min_{f\in\Omega} \quad \mathcal{L}(f(\mathbf{W}),\mathbf{X}) \quad + \quad \mathrm{R}(\mathbf{W})$$

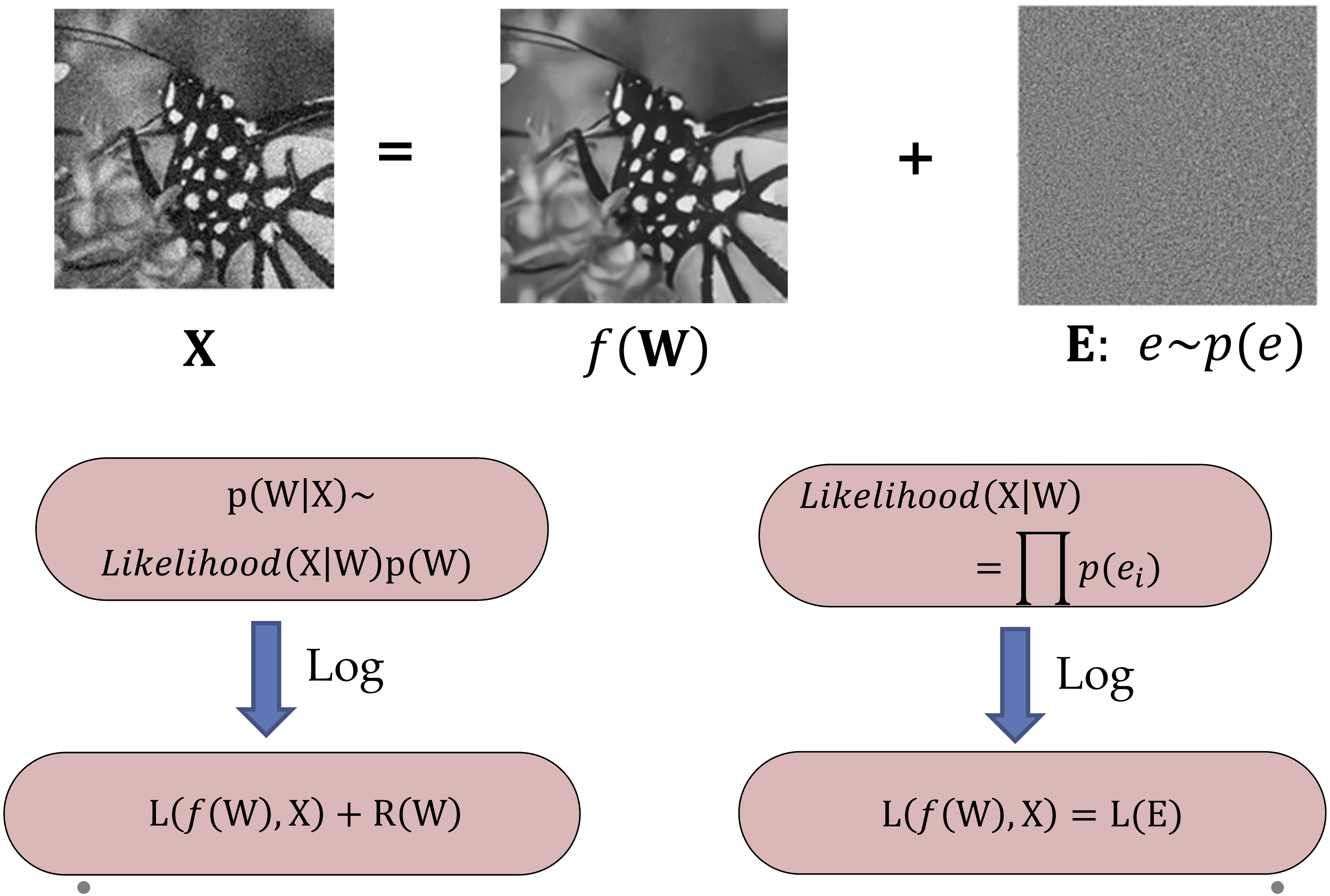
Low rank | **Sparsity** | **Non-local Similarity**

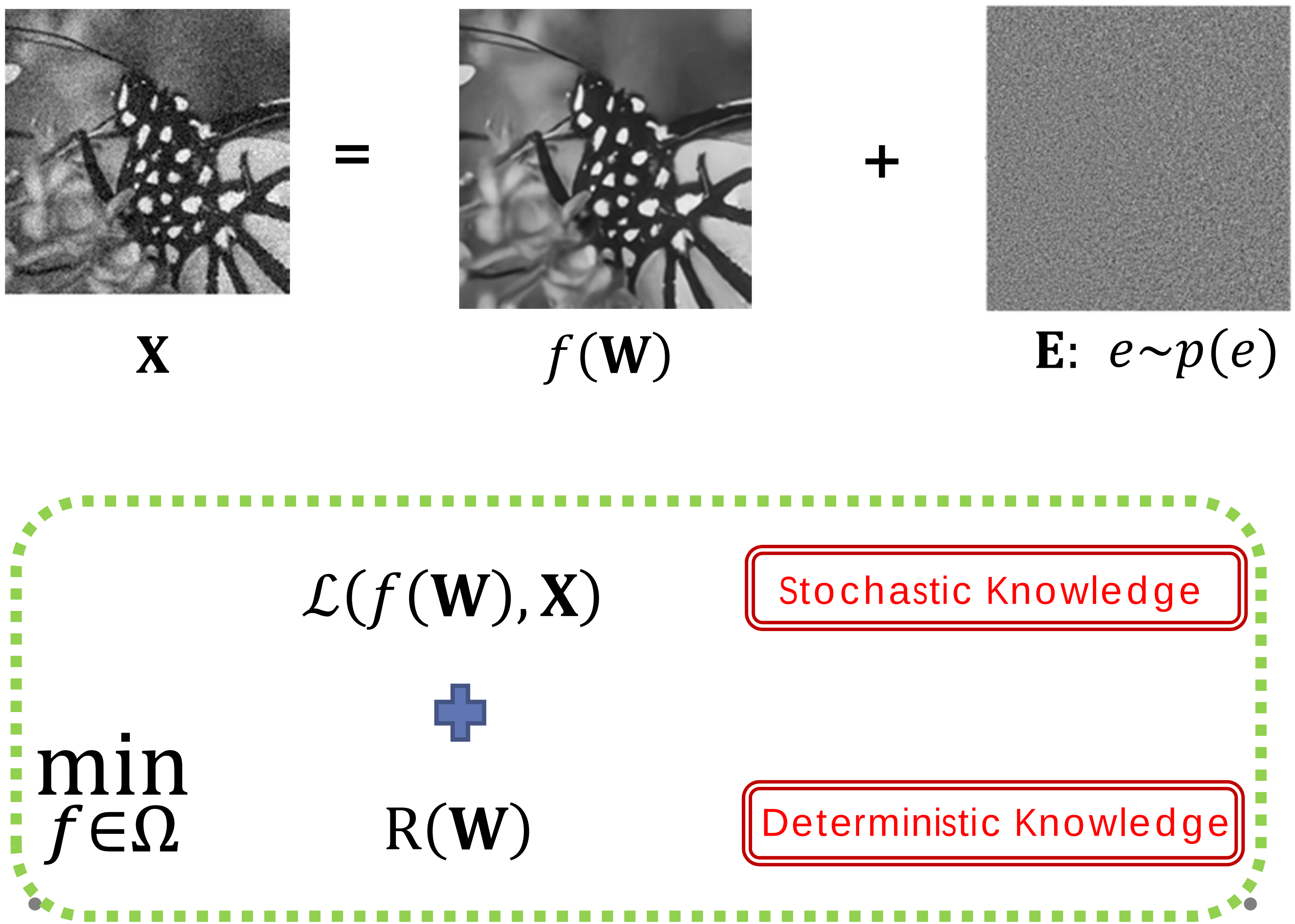
$\mathbf{W}$



$\boldsymbol{f}$




$\mathbf{X}$
=
$f(\mathbf{W})$
+
$\mathbf{E}$: $e \sim p(e)$
$\mathrm{p(W|X)} \sim$
$Likelihood\mathrm{(X|W)p(W)}$
Log
$\mathrm{L}(f\mathrm{(W),X)} + \mathrm{R(W)}$
$Likelihood\mathrm{(X|W)}$
$= \prod p(e_i)$
Log
$\mathrm{L}(f\mathrm{(W),X)} = \mathrm{L(E)}$



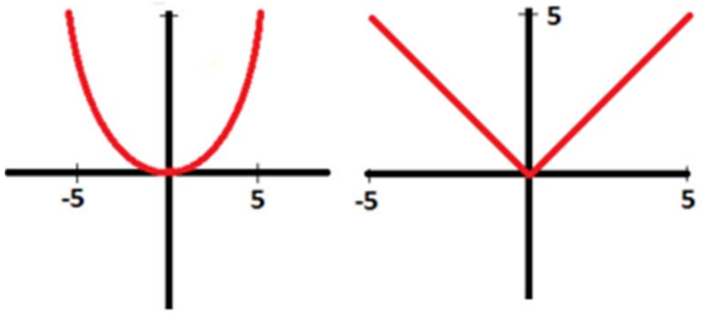
$\mathbf{X}$ = $f(\mathbf{W})$ + $\mathbf{E}$: $e \sim p(e)$
$\min_{f \in \Omega}$
$\mathcal{L}(f(\mathbf{W}), \mathbf{X})$
Stochastic Knowledge
$\mathrm{R}(\mathbf{W})$
Deterministic Knowledge

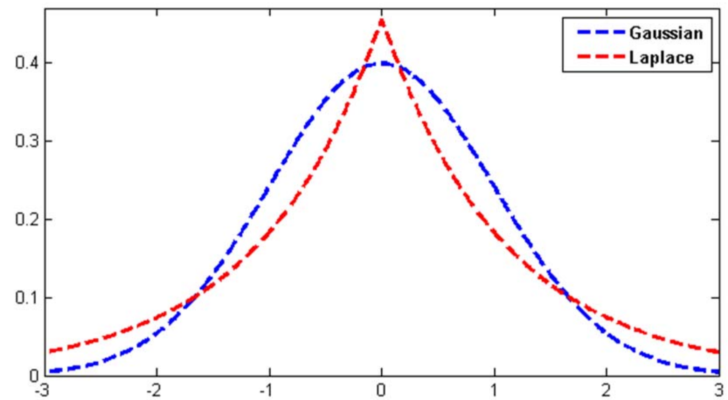
$\mathcal{L}(f(\mathbf{W}), \mathbf{X})$

Loss function

$\mathbf{E}$

noise distribution





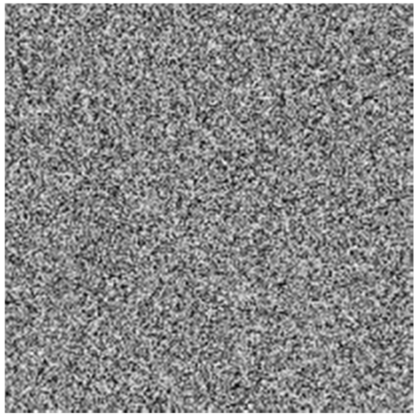
$\|D - F(W)\|_2$

$e \sim N(e; 0, \sigma^2)$

$\|D - F(W)\|_1$

$e \sim \mathrm{Laplace}(e; 0, b)$

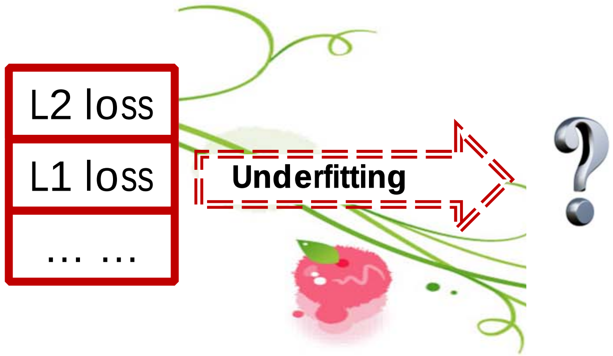
What we assume noise as:



But the real noise is:

Robust Problem

$$\min_{f\in\Omega} \boxed{\mathcal{L}(f(\mathbf{W}),\mathbf{X})} + \mathrm{R}(\mathbf{W})$$


L2 loss
L1 loss
… …
Underfitting

Noise Modeling Principle

- Assume the noise distribution follows a parametric model **p(e,θ)**
- Learn θ from data

We have only observations, while not noises

E needs to be estimated under known model parameter W

⇕

W needs to be calculated under fixed loss function L_θ

Noise Modeling Principle

- Assume the noise distribution follows a parametric model **p(e,θ)**
- Learn θ from data

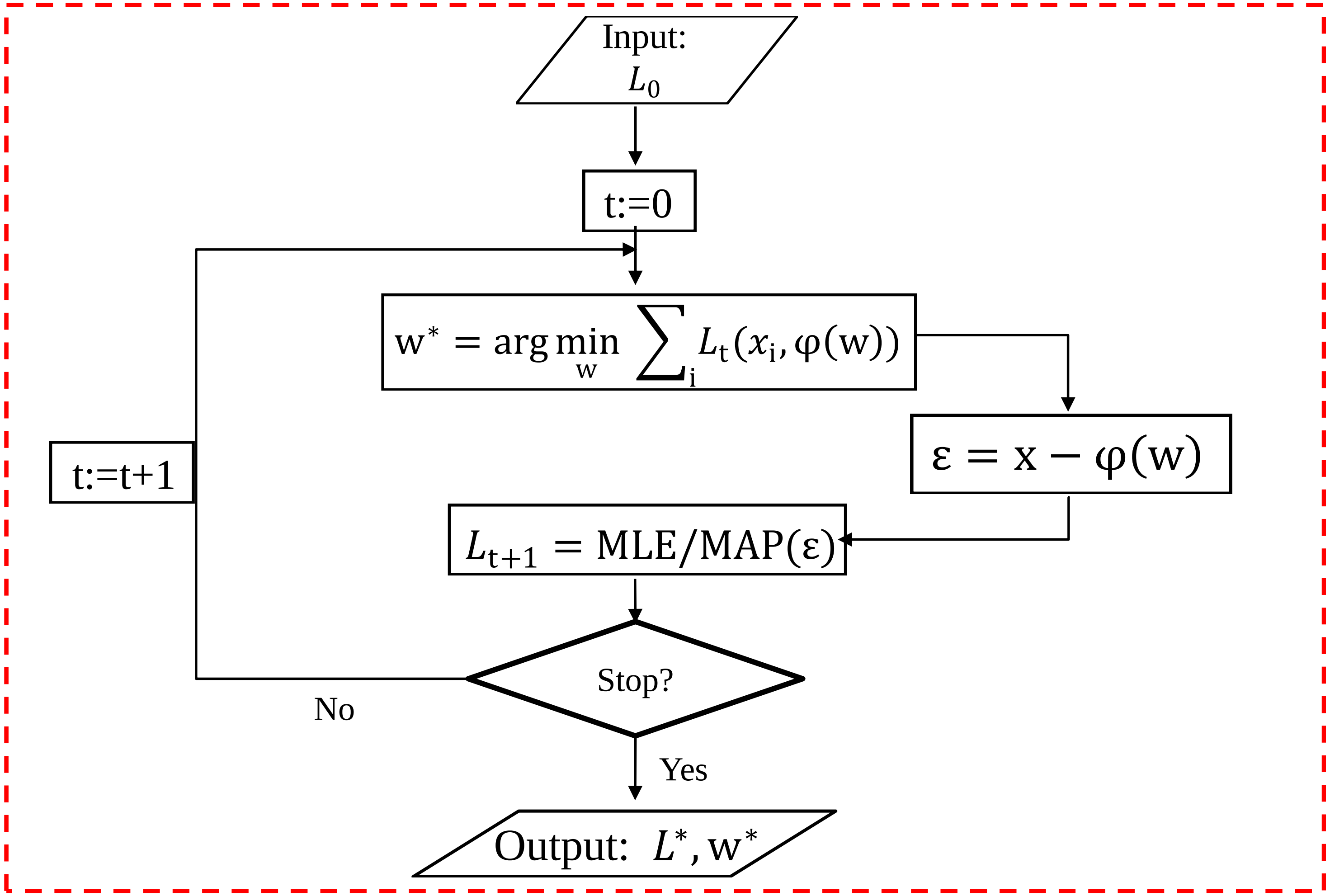
Step 1: Given an initial model parameter(s) W and parametric noise distribution $e \sim p(e, \theta)$;

Step 2: Estimate the loss function/noise distribution $\theta^* = \arg\max_\theta \sum_i \log p(e_i|\theta)$;

Step 3: Optimize the model parameter W under fixed loss function L_{θ^*}.

Loss Modeling Principle

- Assume the loss function $L_{\boldsymbol{\theta}}$ containing certain parameters θ
- Learn L_θ from data



L2 loss

L1 loss

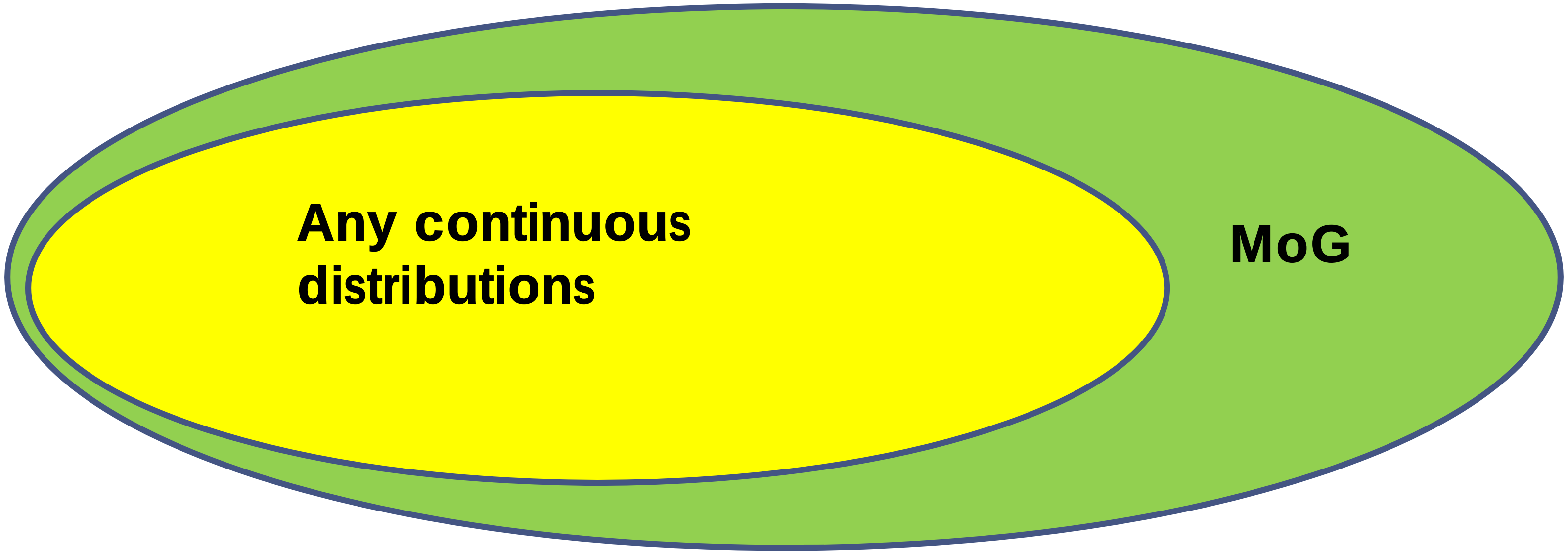
… …

Underfitting

$$e \sim \sum_k \pi_k N(e|0, \sigma_k{}^2)$$

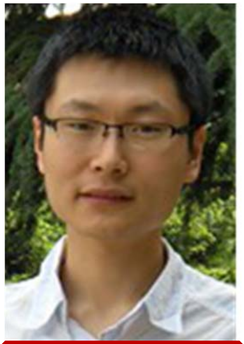
DY Meng, D Fernando, ICCV 2013; Q, Zhao, DY Meng, et al., ICML, 2014

Universal approximation property of MoG



(Maz'ya and Schmidt, 1996)

MoG Noise Modeling



Zhao Qian

Fernando de la Torre

Step 1: Given an initial model parameter(s) W and noise distribution $e \sim \sum_k \pi_k N(e; 0, \sigma_k{}^2)$

Step 2: Estimate the loss function/noise distribution

$$\{\boldsymbol{\pi}^*, \boldsymbol{\sigma}^*\} = \arg\max_{\boldsymbol{\pi},\boldsymbol{\sigma}} \sum_i \log \sum_k \pi_k N(e_i; 0, \sigma_k{}^2);$$

$$\pi_k{}^* = \frac{1}{N}\sum_i \gamma_{i,k}, \sigma_k{}^* = \left(\frac{\sum_i \gamma_{i,k} e_i{}^2}{\sum_i \gamma_{i,k}}\right)^{1/2}, \gamma_{i,k} = \frac{\pi_k N(e_i; 0, \sigma_k{}^2)}{\sum_k \pi_k N(e_i; 0, \sigma_k{}^2)};$$

Step 3: Optimize the model parameter W under fixed loss function

$$W^* = \arg\min_W \left\| H(\boldsymbol{\pi}^*, \boldsymbol{\sigma}^*) \odot \left(D - f(W)\right)\right\|_2.$$

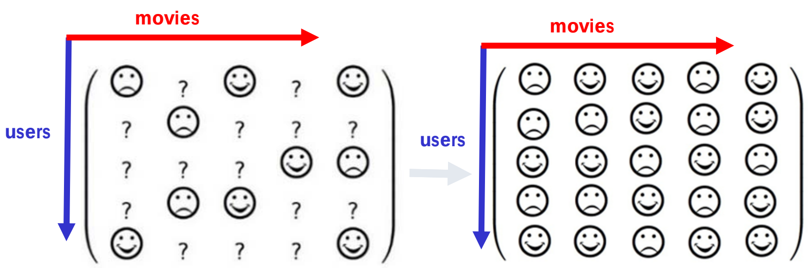
Low-rank Data Structure

LRMF

$$X = UV^T;\ X \in R^{d\times n}, U \in R^{d\times k}, V \in R^{n\times k}; k \ll d, n$$

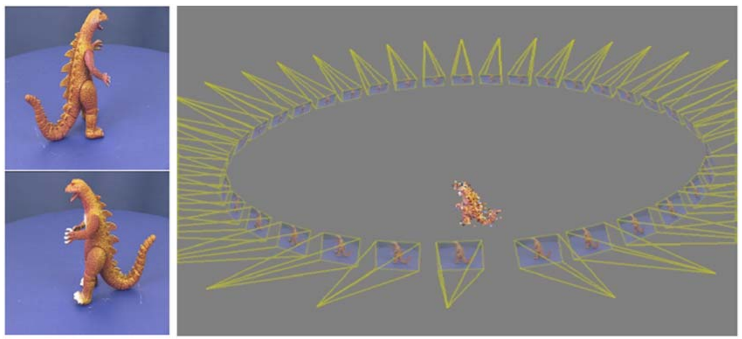
$$\min_{X}\ L(X, UV^T)$$

Recommend-ation system

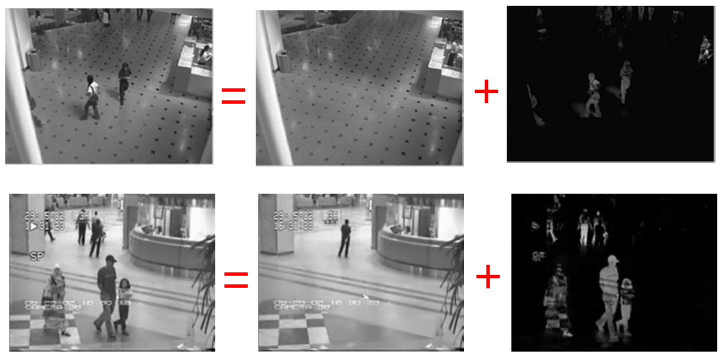


Face recognition

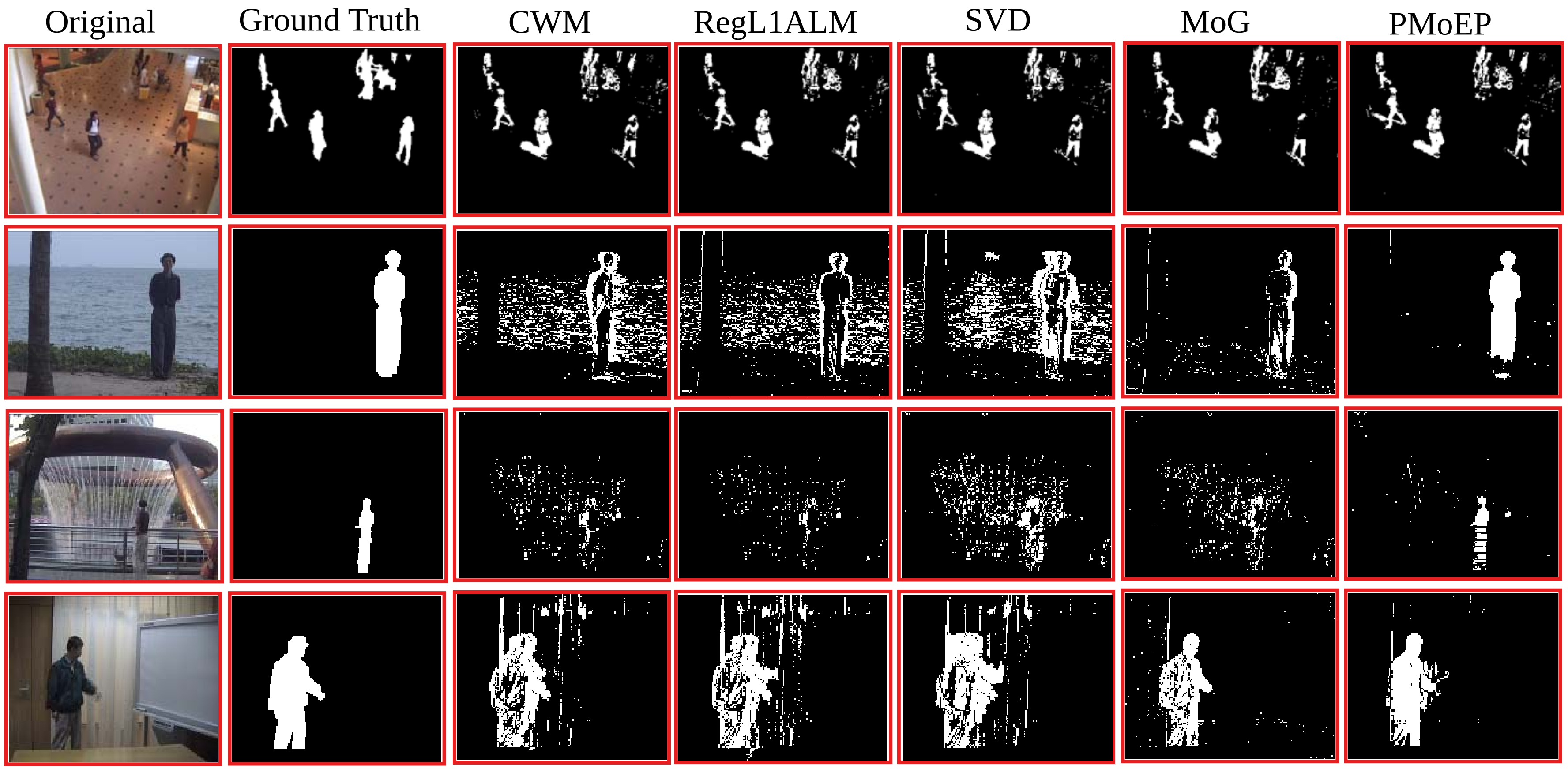
Structure from motion



Background subtraction



Application: Background Subtraction



Original video

Background video

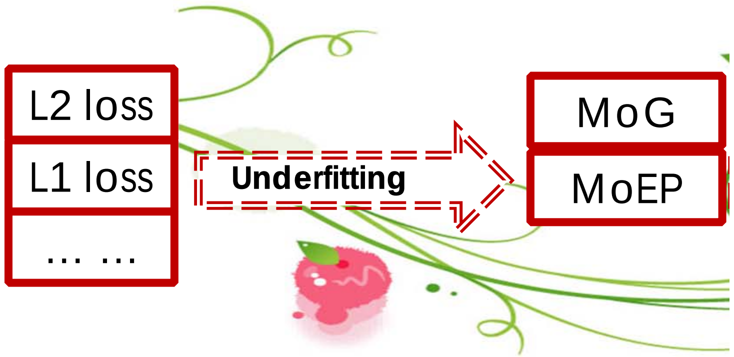
Foreground objects

Shadows of objects

Camera noise

DY Meng, D Fernando, ICCV 2013; Q, Zhao, DY Meng, et al., ICML, 2014

Extend MoG to MoEP



$$e \sim \sum_k \pi_k EP_{p_k}(e|0, \eta_k)$$

XY Cao, Q Zhao, DY Meng, et al., ICCV 2015

$$e \sim \sum_k \pi_k EP_{p_k}(e|0,\eta_k)$$

Cao Xiangyong

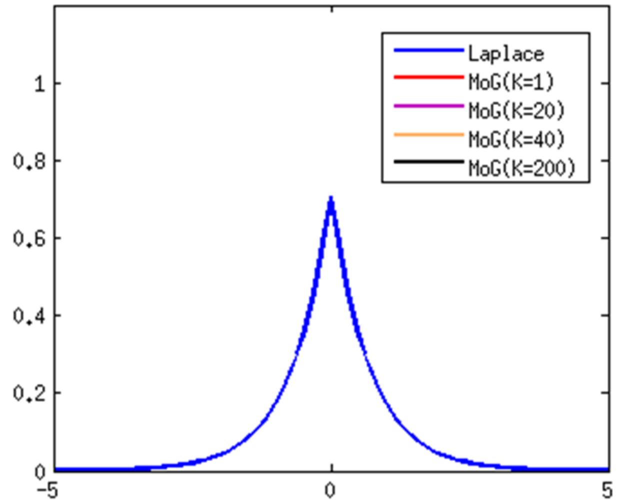
Candidate Components

EP_{p_1} EP_{p_2} ... EP_{p_k}

- Heuristic strategies (Meng et.al. ICCV 2013, Zhao et.al. ICML 2014)
- Information criteria: AIC (Akaike ISIT 1973), BIC(Schwarz et.al. AOS 1978) and etc.
- Bayesian methods: variational inference (Bishop et.al. PRML 2006), Dirichlet prior based method (Ormoneit et.al. TNN 1998, Zivkovic et.al. TPAMI 2004).
- Penalty methods: penalized likelihood (Huang et.al. arxiv 2013).

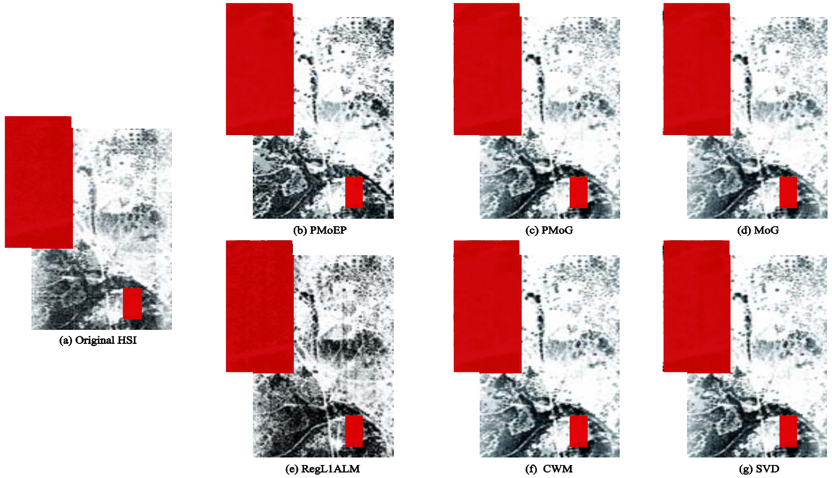
Mixture of Exponential Power Distribution

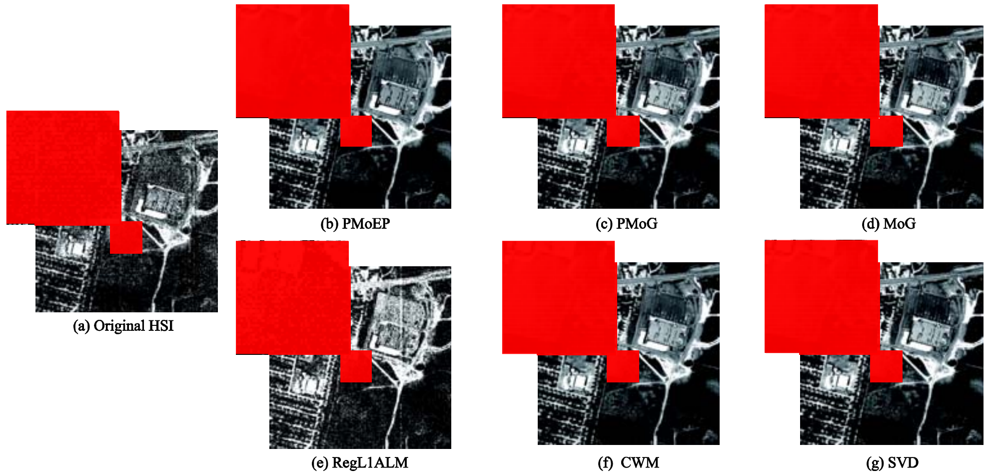
- The candidates p_k can be set the same or different
- Representation capacity of MoEP significantly expand that of MoG!
- All previous models can be considered as special cases of this MoEP framework:
  - L2 norm loss model: EP_2
  - L1 norm loss model: EP_1
  - L2+L1 norm loss model: $EP_2 + EP_1$
  - L_∞ norm loss model: EP_∞
  - MoG: $EP_2 + EP_2 + \cdots$
  - Mixture of Laplacian: $EP_1 + EP_1 + \cdots$

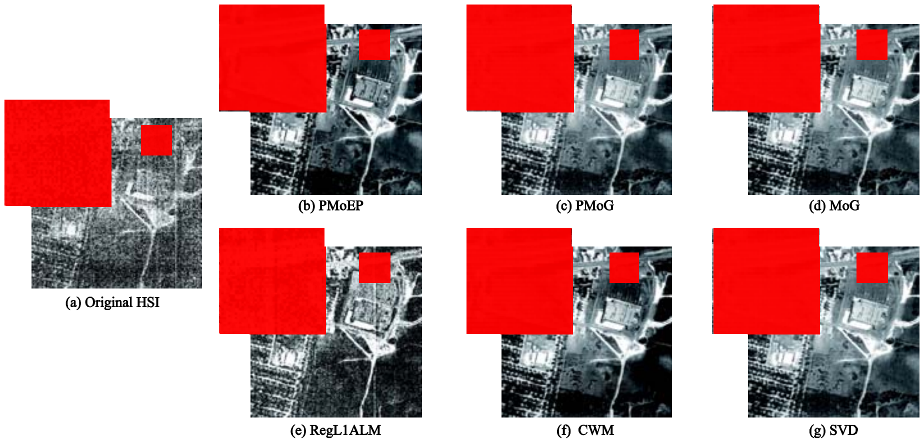


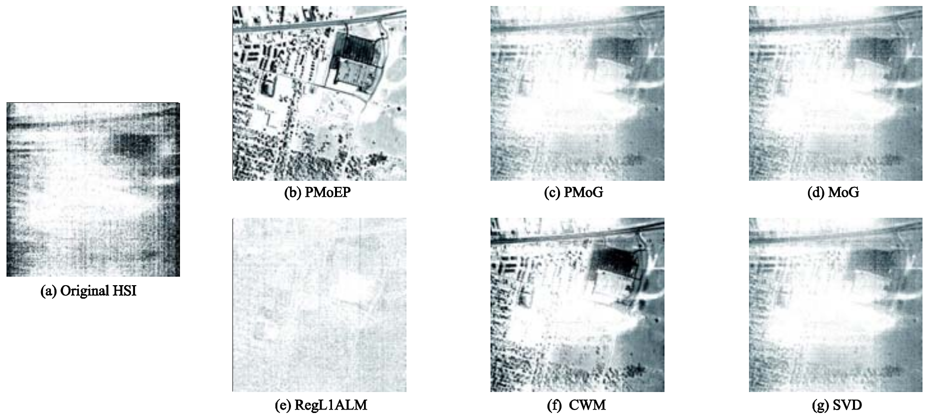
XY Cao, Q Zhao, DY Meng, et al., ICCV 2015

Hyperspectral Image Denoising





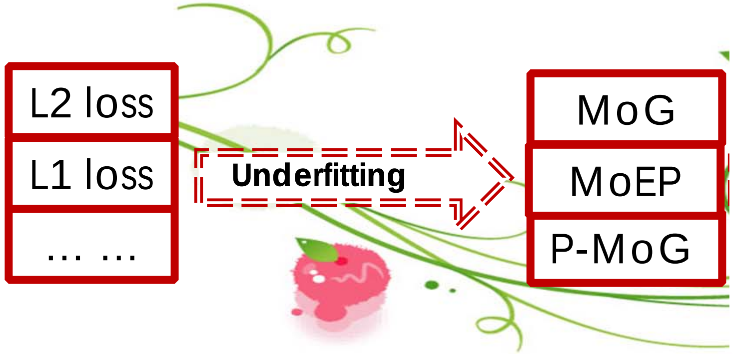




XY Cao, Q Zhao, DY Meng, et al., ICCV 2015

Sometimes noise has intrinsic structures!

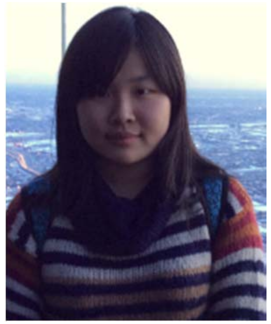
Extend pixel-wise MoG to Patch-wise MoG



$$\boldsymbol{e} \sim \sum_{k} \pi_k N(\boldsymbol{e} | 0, \Sigma_k)$$

W Wei, LX Yi, DY Meng, et al., ICCV 2017

Wei Wei



Yi Lixuan

Previous methods encode rain streaks by their:

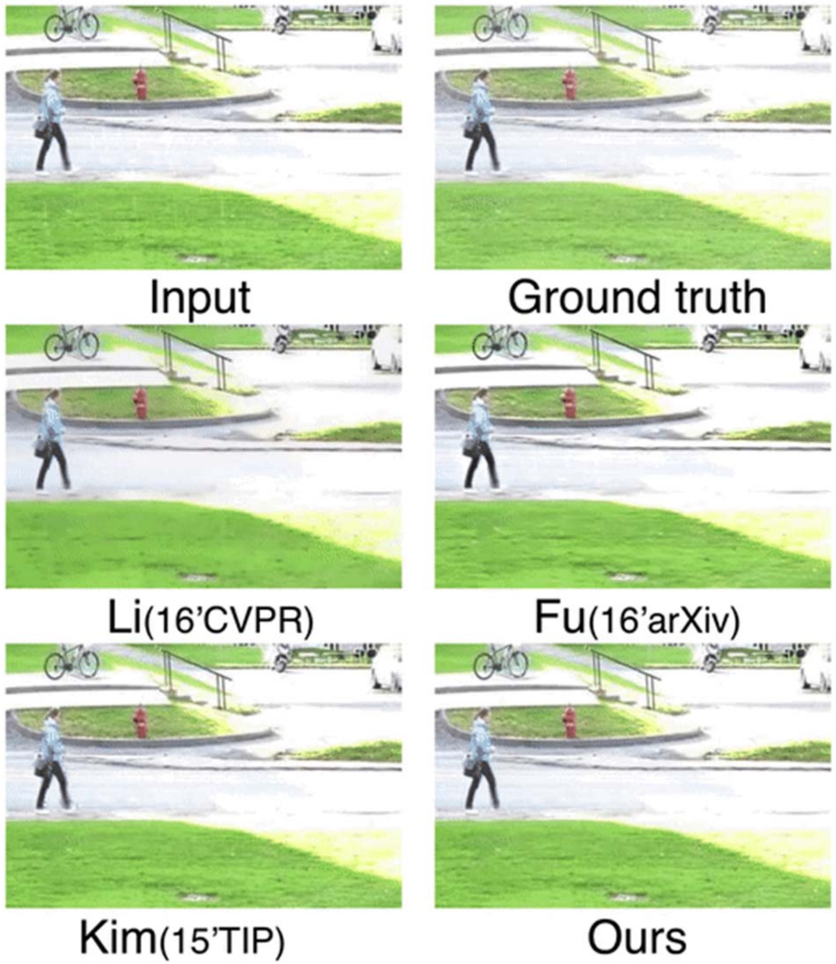
- **Photometric appearance** (K. Garg and S. K. Nayar, ICCV 2005)
- **Chromatic consistency** (P. Liu, J. Xu, J. Liu, X. Tang, CIS 2009)
- **Spatiotemporal configurations** (A. Tripathi, S. Mukhopadhyay, LIP 2012)
- **Local structure correlations** (Y.-L. Chen and C.-T. Hsu, ICCV 2013)
- **Discriminative structures** (J. H. Kim, J. Y. Sim, and C. S. Kim, TIP 2015)

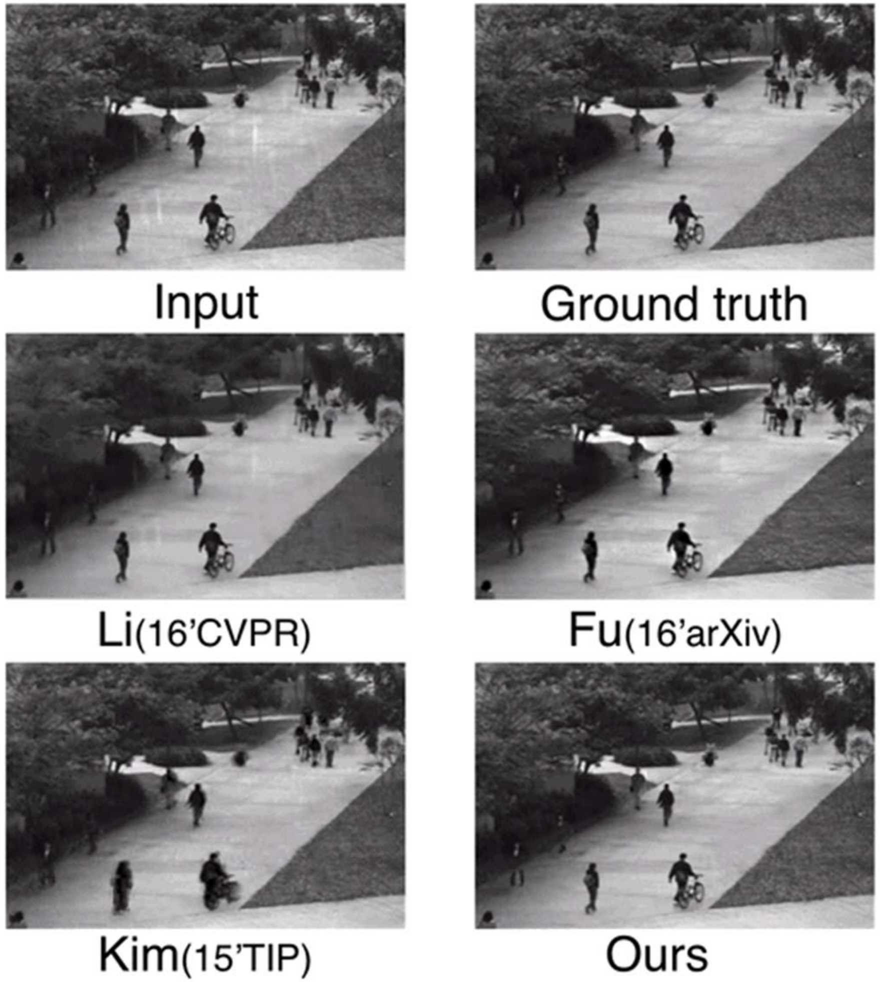
They take rain streaks as **Deterministic Knowledge!**

We might better encode rain streaks as stochastic!

- Background: Low-rank
- Moving objects: Spatial smoothness
- Rain steaks: Patch-based MoG

$$\min_{\Theta} -\sum_{n=1}^{n_p} \log \sum_{k=1}^{K} \pi_k \mathcal{N}(f(\mathcal{H}^{\perp} \circ \mathcal{R})_n | 0, \Sigma_k)$$
$$+ \alpha ||\mathcal{H}||_{3DTV} + \beta ||\mathcal{H}||_1$$
$$\text{s.t. } \mathcal{H}^{\perp} \circ \mathcal{R} = \mathcal{H}^{\perp} \circ (\mathcal{D} - \mathcal{B}) \quad \mathbf{B} = \mathbf{U}\mathbf{V}^T$$



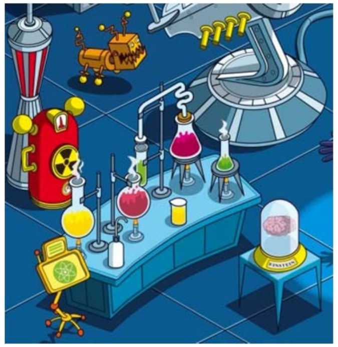


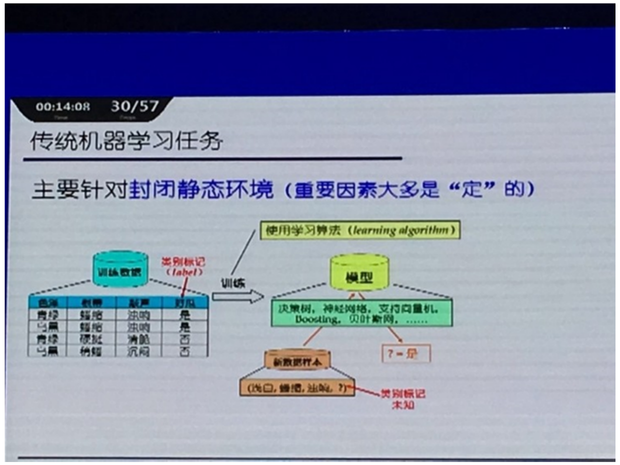
Dataset	Dataset 1				Dataset 2				Dataset 3				Dataset 4			
Metrics	VIF	SSIM	FSIM	UQI	VIF	SSIM	FSIM	UQI	VIF	SSIM	FSIM	UQI	VIF	SSIM	FSIM	UQI
Input	0.846	0.981	0.991	0.934	0.731	0.950	0.975	0.927	0.591	0.877	0.935	0.816	0.717	0.917	0.970	0.763
Fu [10]	0.696	0.956	0.968	0.847	0.673	0.948	0.971	0.923	0.530	0.887	0.933	0.812	0.670	0.935	0.967	0.808
Garg [14]	0.862	0.984	0.990	0.949	0.745	0.961	0.979	0.944	0.712	0.935	0.969	0.887	0.707	0.920	0.972	0.772
Kim [17]	0.810	0.981	0.987	0.941	0.642	0.949	0.968	0.933	0.666	0.943	0.967	0.907	0.589	0.912	0.960	0.758
Ours	**0.904**	**0.993**	**0.993**	**0.969**	**0.786**	**0.977**	**0.985**	**0.968**	**0.757**	**0.960**	**0.980**	**0.952**	**0.768**	**0.949**	**0.981**	**0.891**

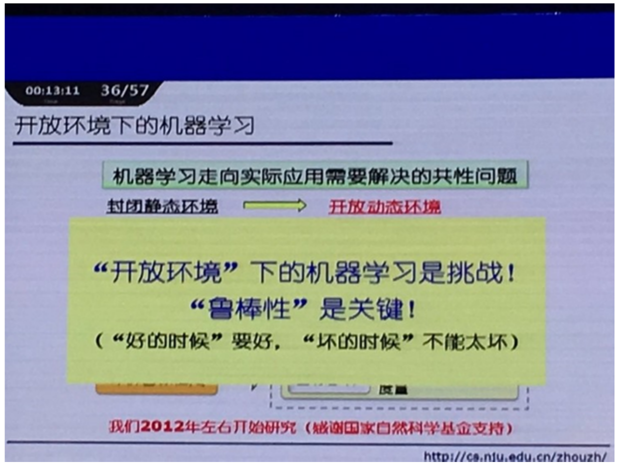
W Wei, LX Yi, DY Meng, et al., ICCV 2017

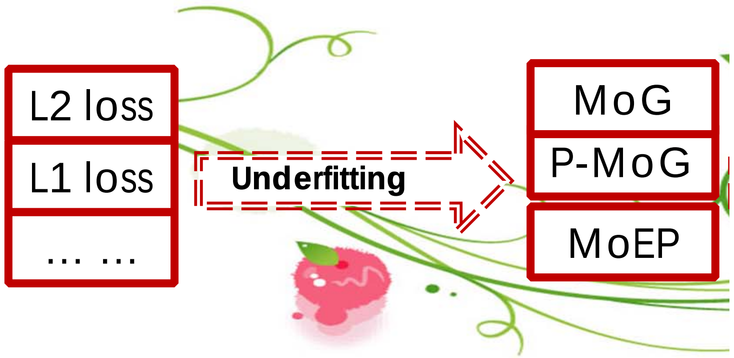
W Wei, LX Yi, DY Meng, et al., ICCV 2017

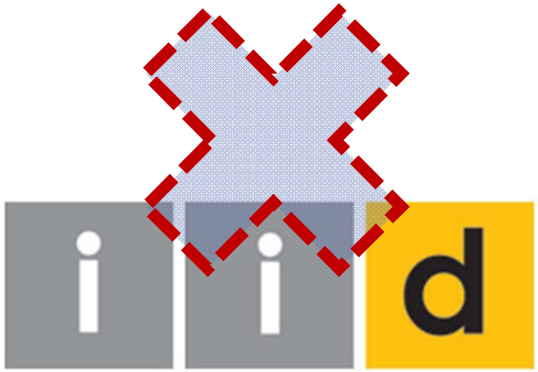
Next Generation of ML:
From Laboratory to the Wild







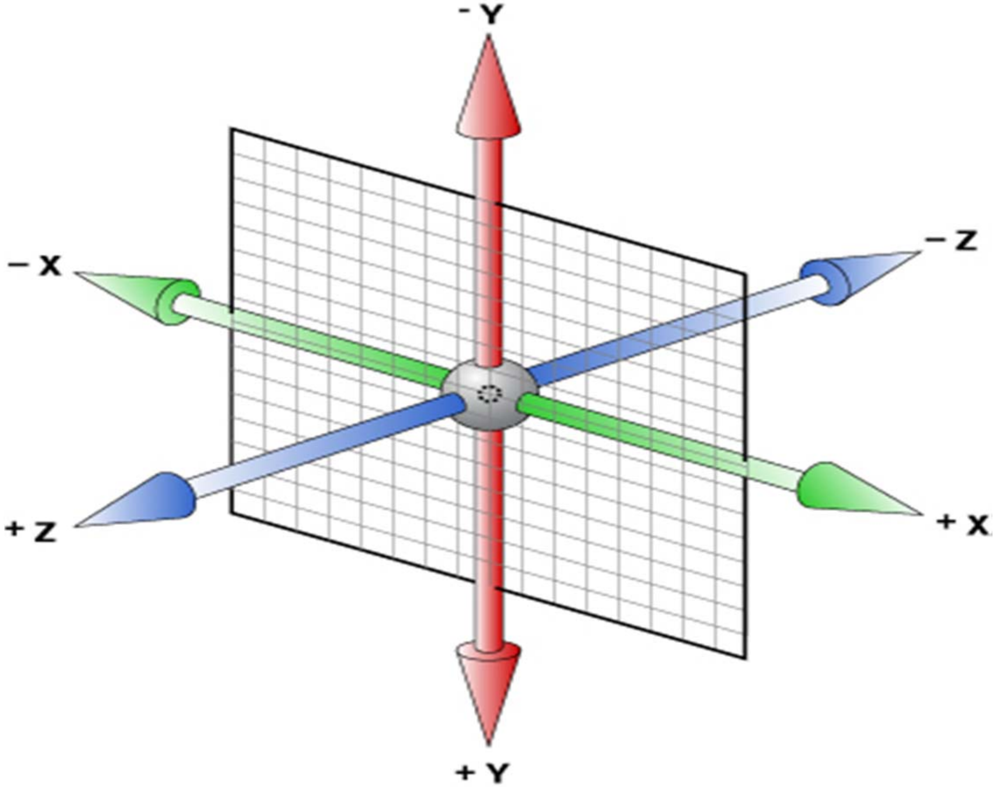

L2 loss
L1 loss
… …
Underfitting
MoG
P-MoG
MoEP



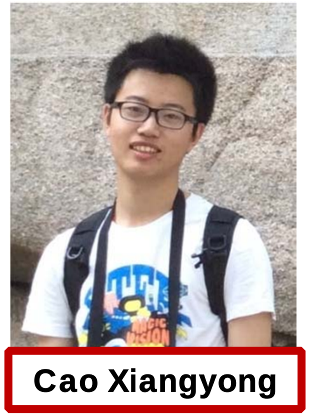
Not identically distributed
AND
Not independently distributed

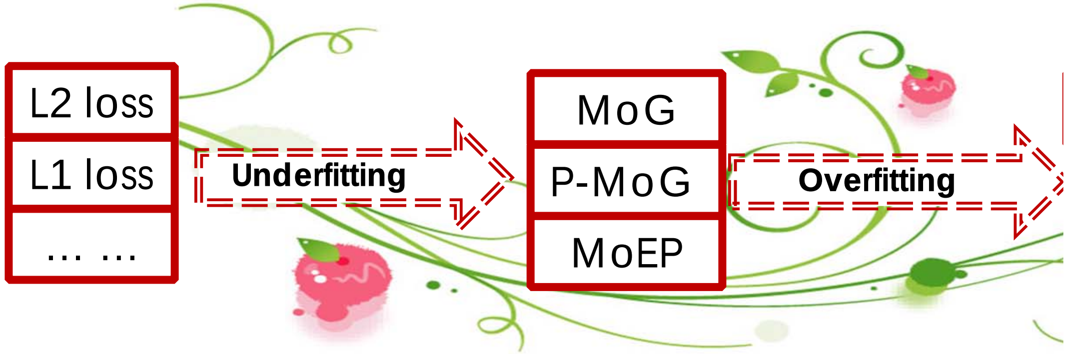
Dimensional/Spectral prior

Spatial prior

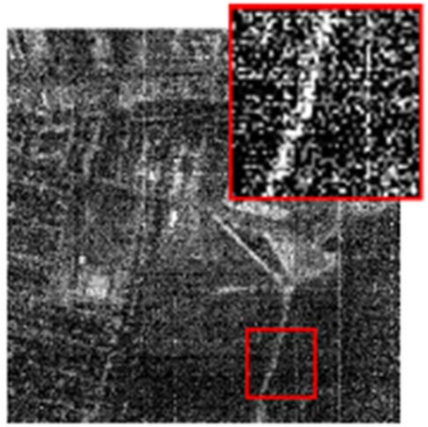


Temporal prior

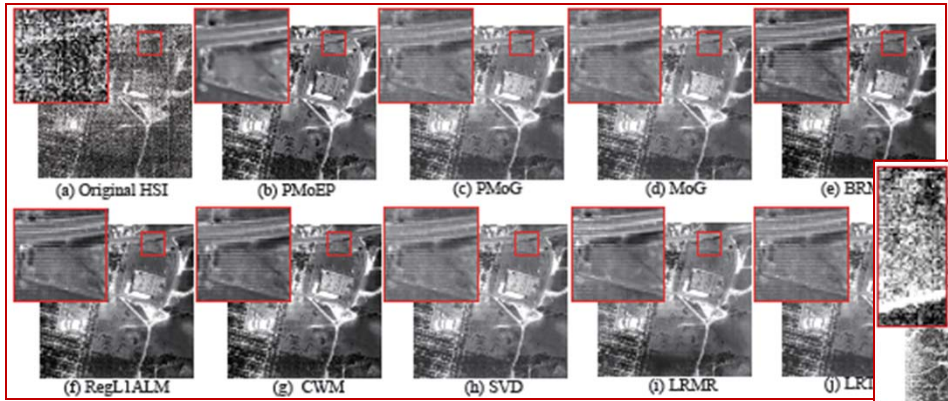


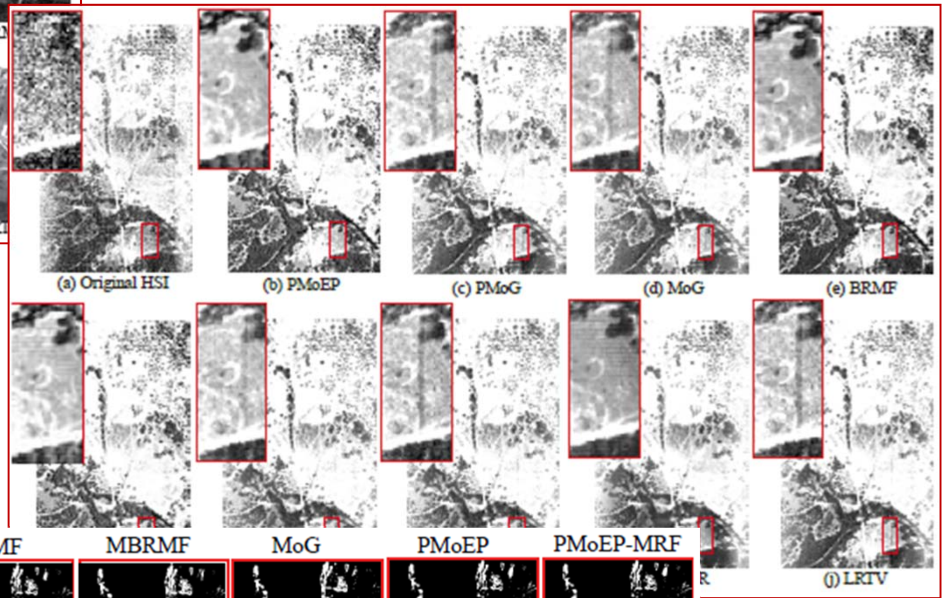


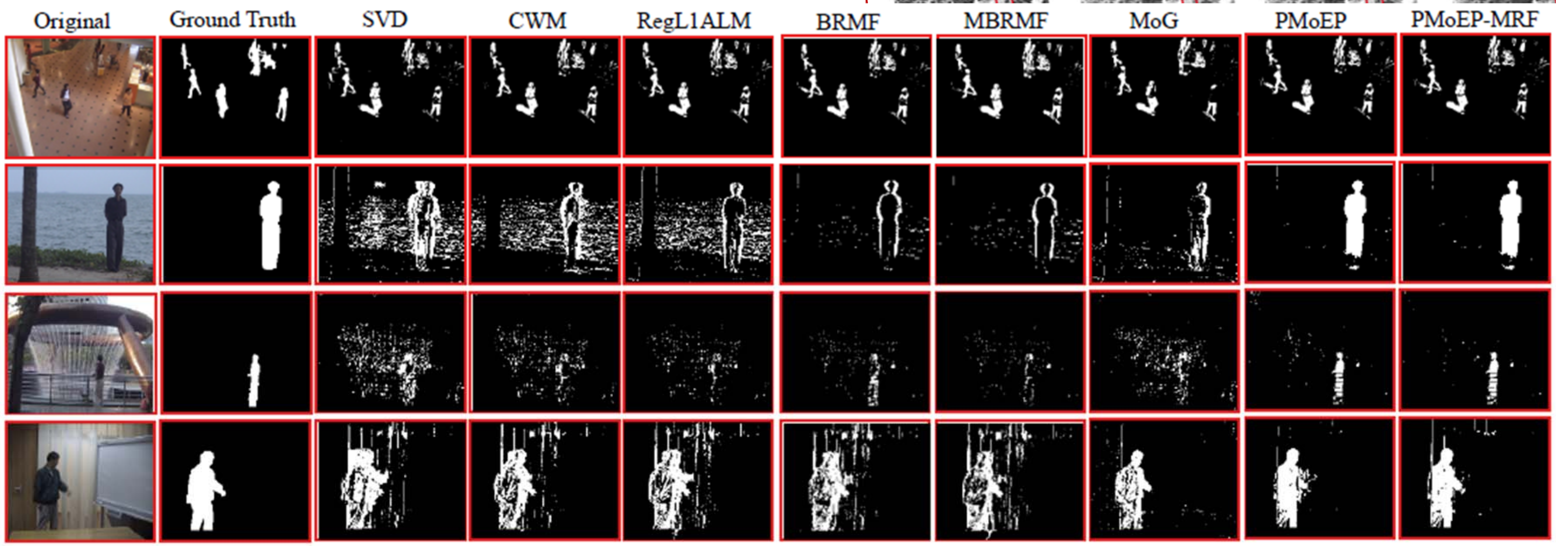
From Spatial Perspective



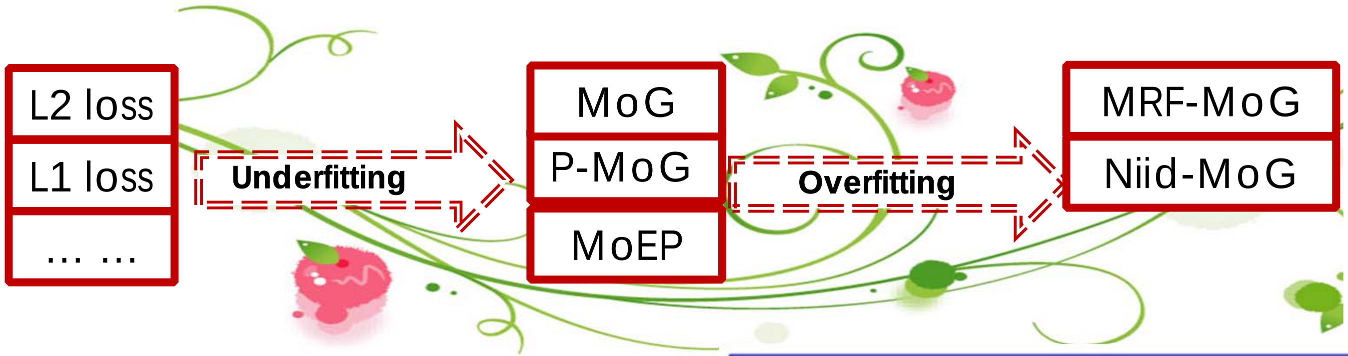
XY Cao, Q Zhao, DY Meng, et al., TIP 2016



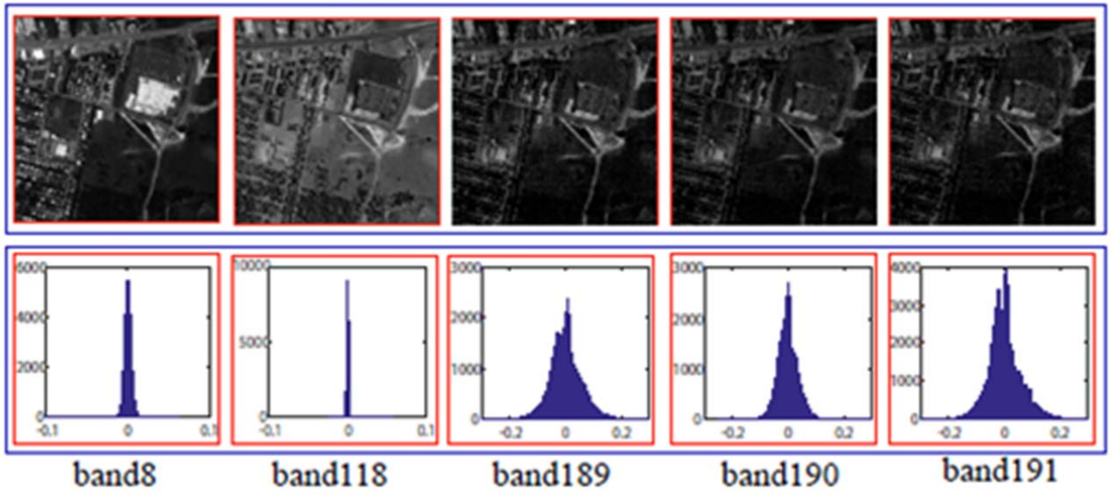




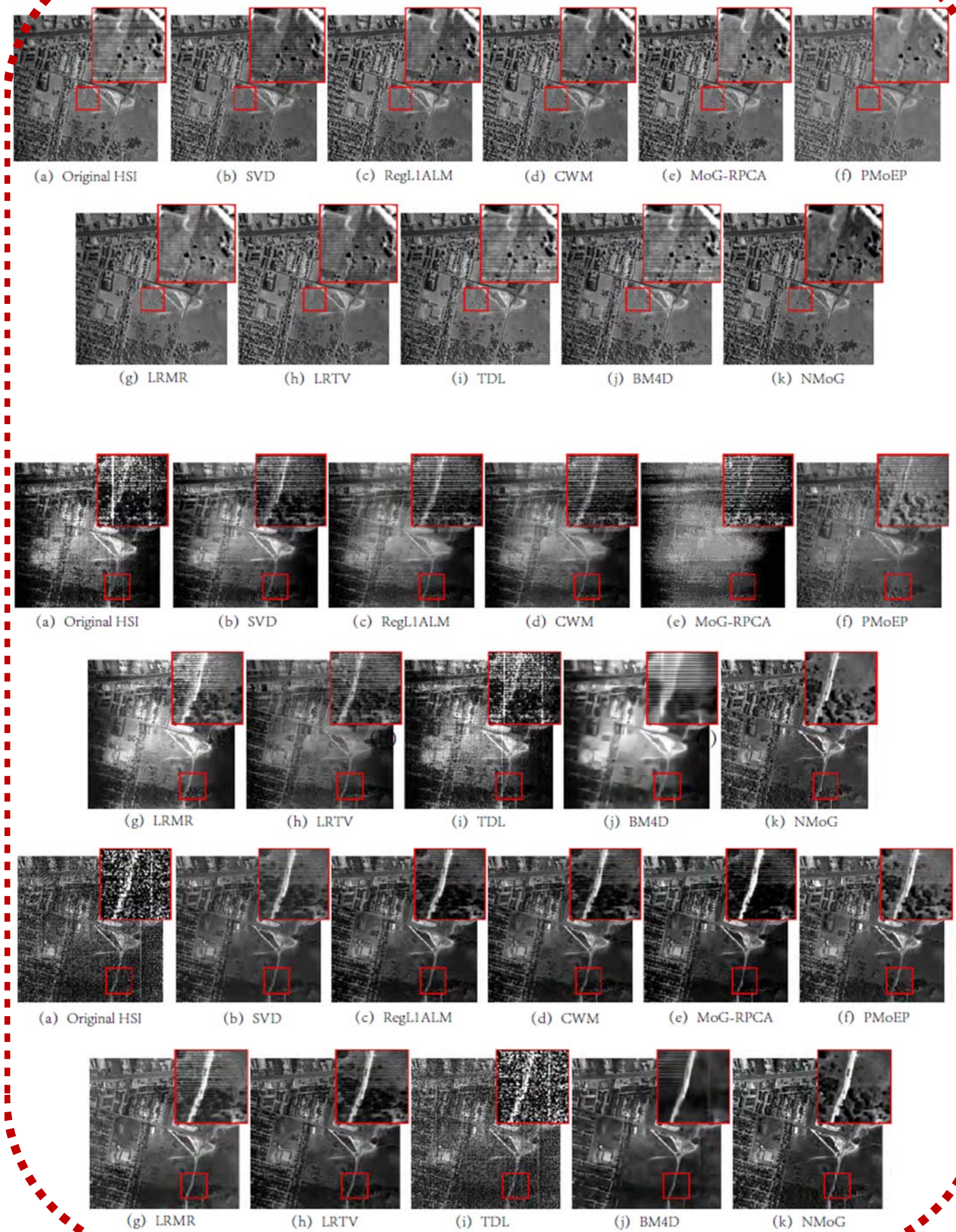
XY Cao, Q Zhao, DY Meng, et al., TIP 2016

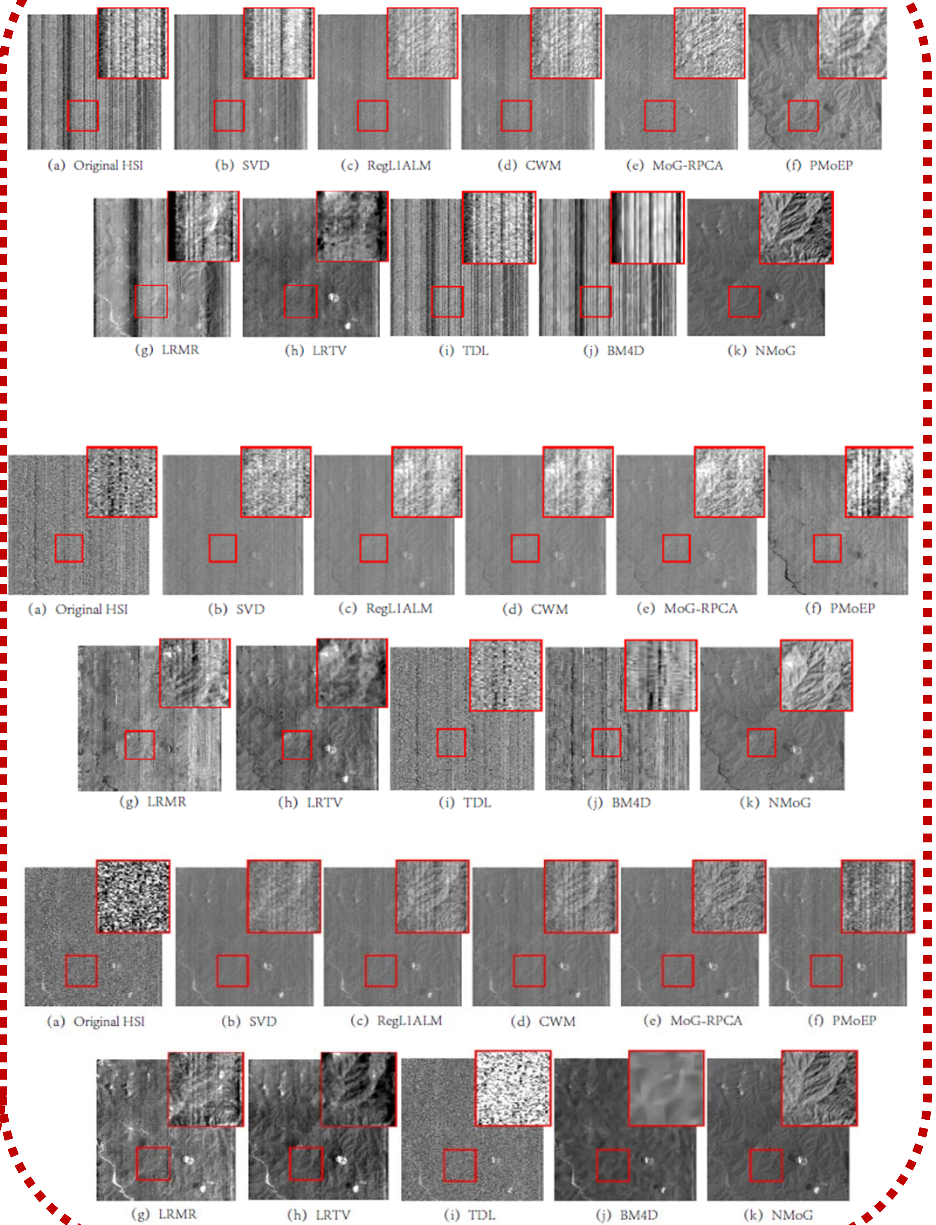


From Spectial Perspective



Y Chen, XY Cao, Q Zhao, et al., TC 2017





Y Chen, XY Cao, Q Zhao, et al., TC 2017

Yong Hongwei

L2 loss
L1 loss
… …

Underfitting →

MoG
P-MoG
MoEP

Overfitting →

MRF-MoG
Niid-MoG
Online-MoG

From Temporal Perspective

Online method should be

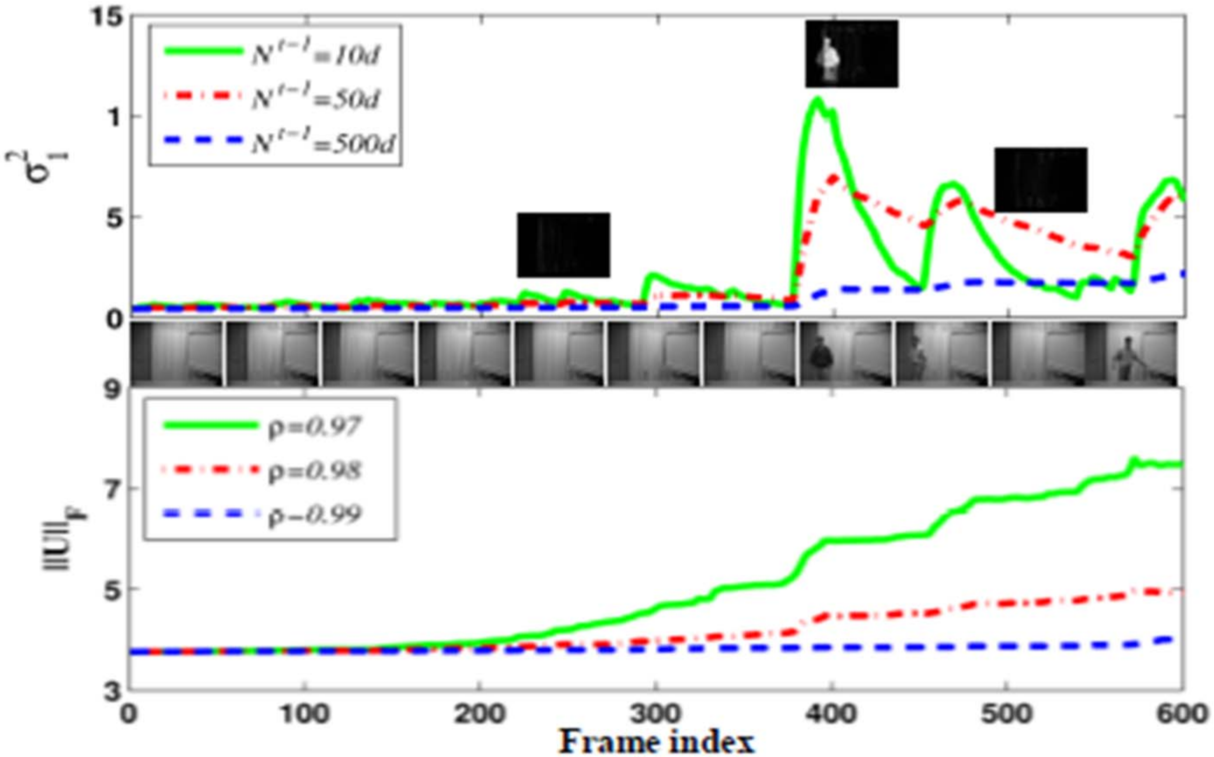
- *More efficient!*
- *More accurate!*

HW Yong, DY Meng, et al., TPAMI, 2017

$$\mathcal{L}^t(\mathbf{\Pi},\mathbf{\Sigma},\mathbf{v},\mathbf{U}) = -\ln p(\mathbf{x}^t|\mathbf{\Pi},\mathbf{\Sigma},\mathbf{v},\mathbf{U}) + \mathcal{R}_F^t(\mathbf{\Pi},\mathbf{\Sigma}) + \mathcal{R}_B^t(\mathbf{U})$$

$$\begin{aligned}\mathcal{R}_F^t(\mathbf{\Pi},\mathbf{\Sigma}) = & \sum_{k=1}^{K} N_k^{t-1} D_{KL}(\mathcal{N}(x|0,{\sigma_k^{t-1}}^2)||\mathcal{N}(x|0,{\sigma_k}^2)) \\ & + N^{t-1} D_{KL}(\mathrm{Multi}(\mathbf{z}|\mathbf{\Pi}^{t-1})||\mathrm{Multi}(\mathbf{z}|\mathbf{\Pi})) + C \\ = & N^{t-1} D_{KL}(p(x,\mathbf{z}|\mathbf{\Pi}^{t-1},\mathbf{\Sigma}^{t-1})||p(x,\mathbf{z}|\mathbf{\Pi},\mathbf{\Sigma})) + C,\end{aligned}$$

$$\mathcal{R}_B^t(\mathbf{U}) = \rho \sum^{d} (\mathbf{u}_i - {\mathbf{u}_i}^{t-1})^T (\mathbf{A}_i^{t-1})^{-1} (\mathbf{u}_i - {\mathbf{u}_i}^{t-1})$$

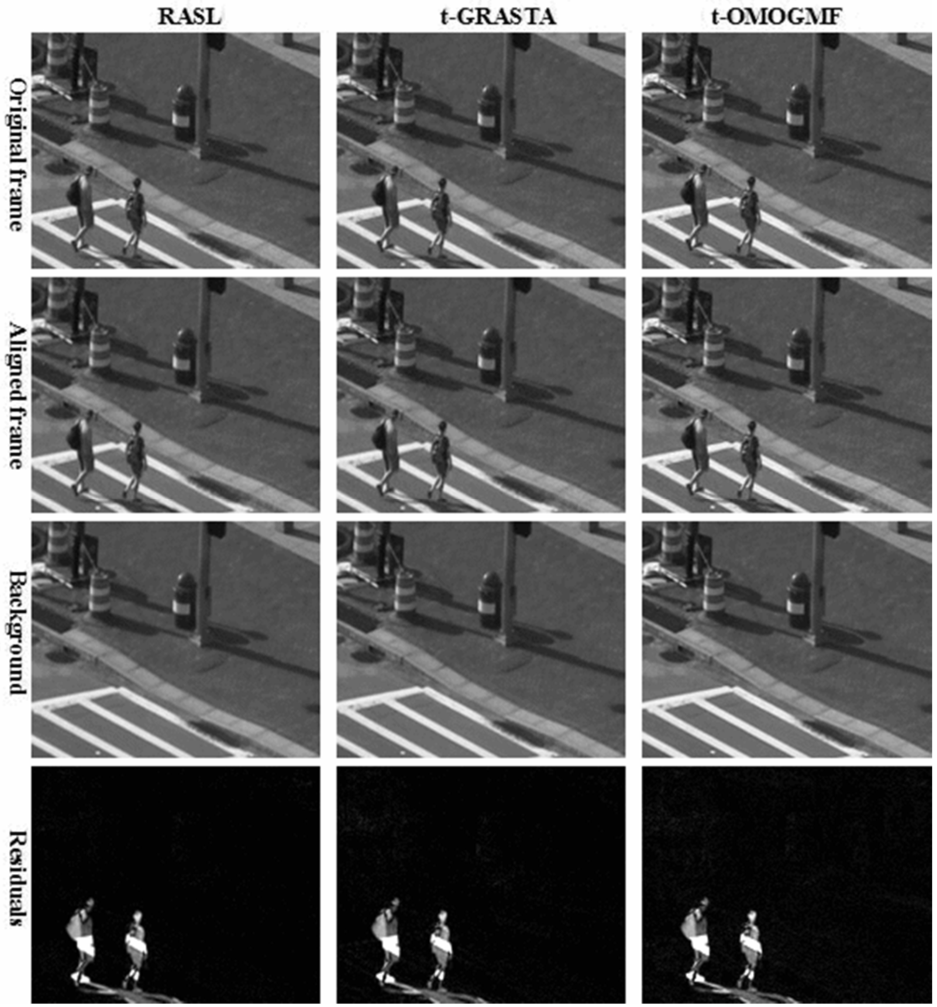


Methods	data									
	air.	*boo.*	*sho.*	*lob.*	*esc.*	*cur.*	*cam.*	*wat.*	*fou.*	Average
RPCA [16]	71.11	**67.67**	**72.79**	78.12	64.09	81.65	44.56	65.56	72.39	68.66
GODEC [19]	62.69	58.39	70.71	73.29	57.42	59.84	43.71	48.79	66.01	60.09
RegL1 [29]	65.63	62.46	71.97	75.27	60.95	62.69	44.42	57.86	73.17	63.82
PRMF [17]	65.87	62.29	71.99	75.32	60.20	65.17	44.04	61.95	72.98	64.42
OPRMF [17]	66.17	61.82	71.95	73.99	60.12	70.86	42.89	61.89	71.80	64.61
GRASTA [21]	61.87	58.07	71.47	60.98	57.26	68.20	44.53	75.88	69.23	63.05
incPCP [38]	59.84	62.47	71.28	75.83	45.59	61.10	44.55	74.94	70.49	62.90
PracReProCS [37]	70.01	*63.71*	71.61	61.89	56.08	77.74	42.28	*87.53*	62.76	65.96
OMoGMF	*74.08*	59.87	71.80	78.01	61.42	86.08	44.48	87.34	71.78	70.54
DECOLOR [18]	63.98	59.97	65.37	68.93	*75.93*	*89.56*	**77.14**	64.03	**86.76**	*72.41*
GOSUS [22]	65.80	61.95	72.12	*80.97*	**86.27**	68.26	51.30	84.37	73.15	71.35
OMoGMF+TV	**77.20**	61.17	*72.43*	**83.47**	66.37	**92.54**	*65.88*	**93.14**	*82.53*	**77.19**

- ***Better foreground object detection***

Video	*esc.*	*air.*	*sho.*
Frame Size	130× 160	144× 176	256× 320
OPRMF [17]	0.5	0.4	0.1
PracReProCS [37]	1.5	1.2	0.2
GOSUS [22]	3.8	2.7	0.6
OMoGMF+TV	18.5	14.8	3.5
OMoGMF	99.6	63.0	5.2
GRASTA [21]	*166.9*	*123.9*	*28.7*
incPCP [38]	**274.5**	**220.8**	**85.2**
GRASTA&1%SS	303.2	246.7	65.5
OMoGMF&1%SS	**332.0**	**263.6**	**104.7**

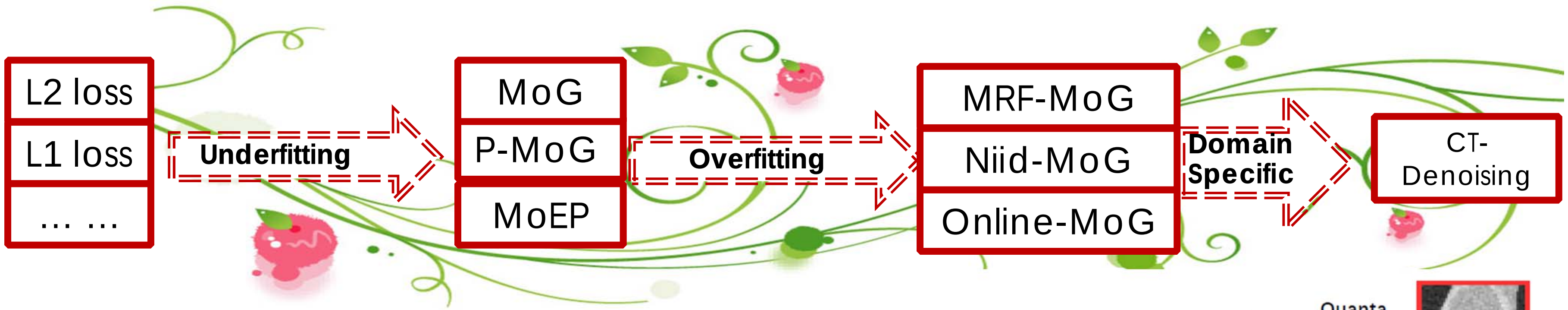
- ***Faster computational speed***



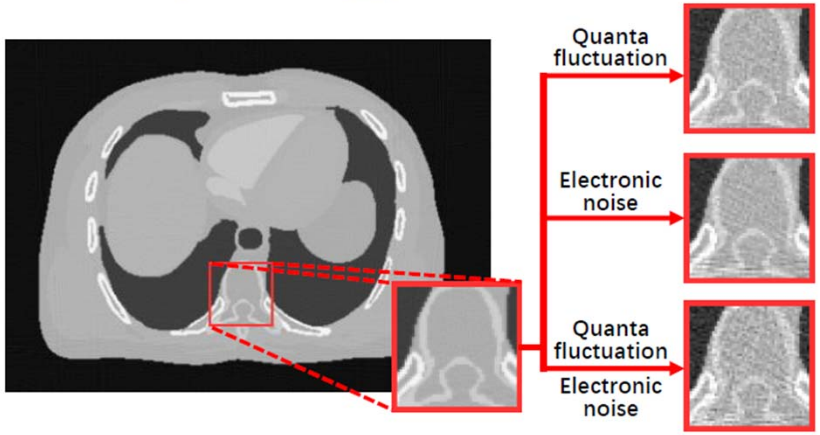
- ***Better background scene subtraction***

Please see more demos in http://gr.xjtu.edu.cn/web/dymeng/7.

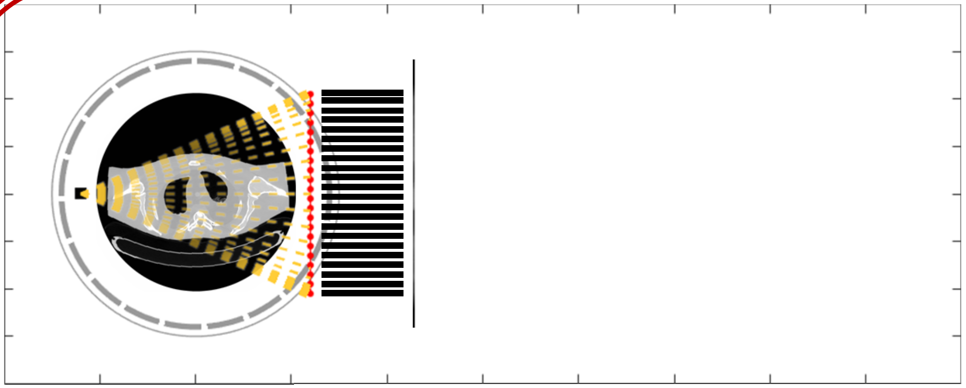
HW Yong, DY Meng, et al., TPAMI 2017

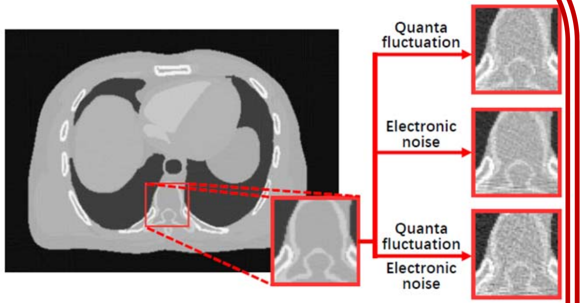


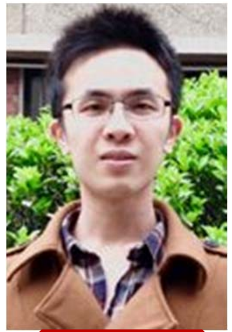
Design noise structure based on domain knowledge



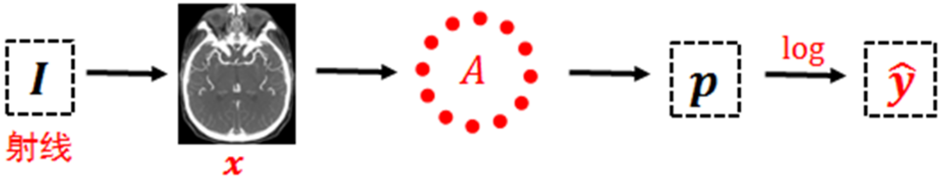
Q Xie, D Zeng, Q Zhao et al., TMI, 2017







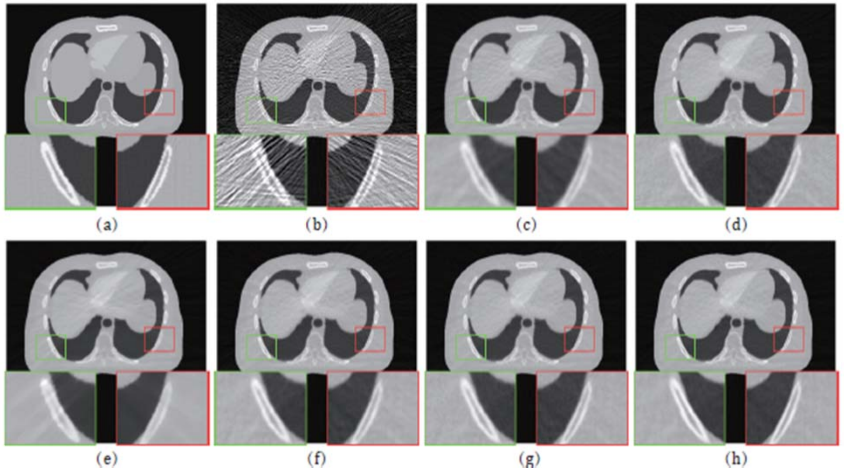
Xie Qi

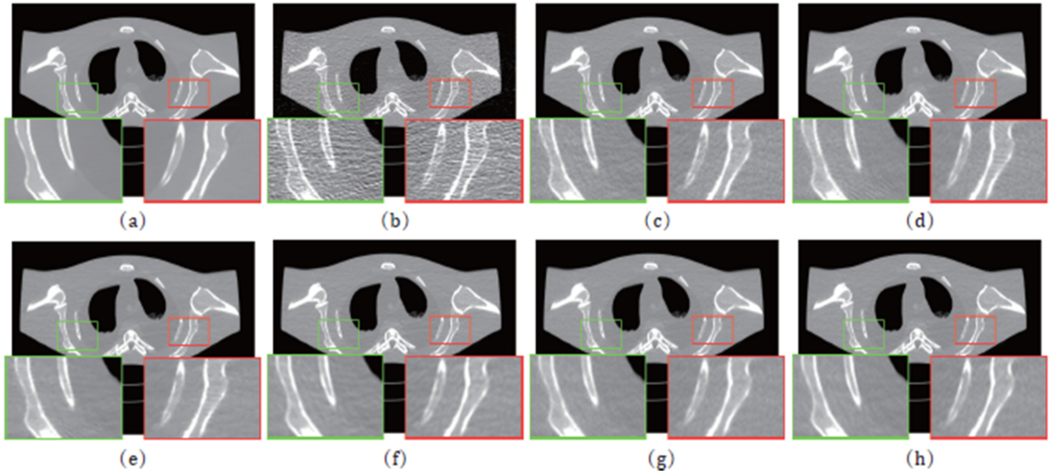


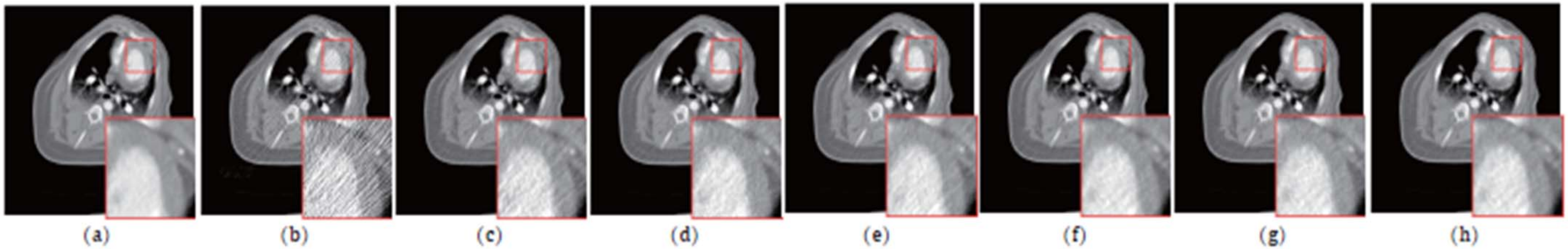
Ma Jianhua

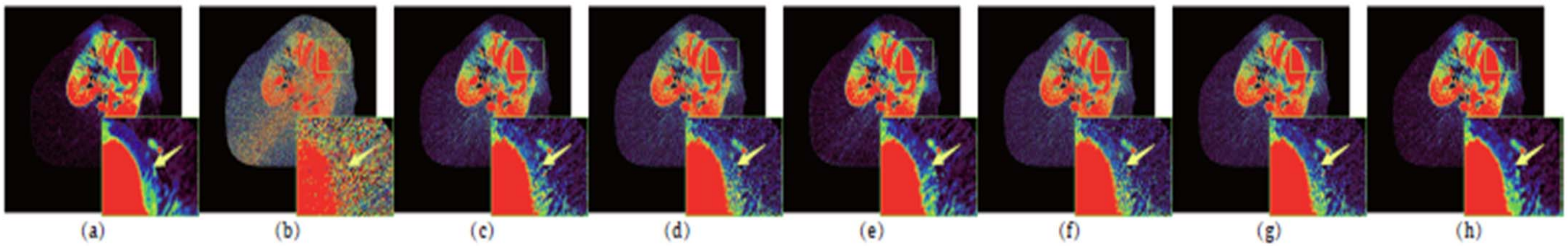
$$p(I,Y,b|P)=\frac{p(P,I|Y)p(Y,b)}{p(P)}$$
$$\propto\frac{1}{b^M}\exp\left(-\frac{\|P-I\|_2^2}{2\sigma^2}-\frac{\|f(Y)\|_1}{b}\right)\prod_{i=1}^{N}\left(\frac{(I_{0_i}e^{-Y_i})^{I_i}}{I_i!}\exp\left(-I_{0_i}e^{-Y_i}\right)\right)$$

$$\max_{I,Y,Q,b}\sum_{i=1}^{N}\left(\frac{(P_i-I_i)^2}{2\sigma^2}+I_i\ln(I_{0_i})-I_iY_i-\ln(I_i!)-I_{0_i}e^{-Y_i}\right)$$
$$-\frac{1}{b}\|\ln(Z+\epsilon)-\ln(\epsilon)\|_1-M\ln(b)$$
$$\text{s.t. } D_2Y=Z$$

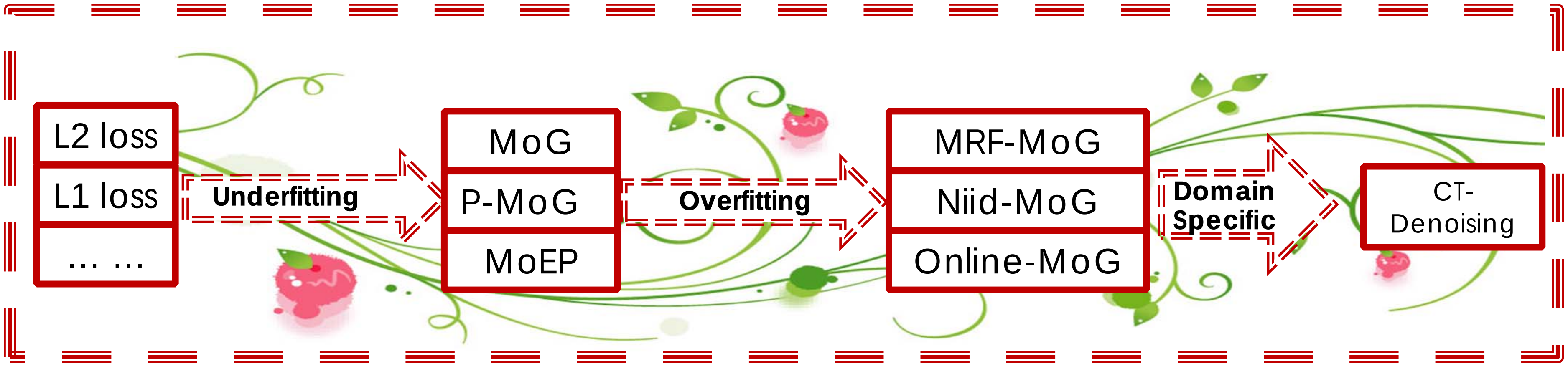








Q Xie, D Zeng, Q Zhao et al., TMI, 2017


L2 loss
L1 loss
… …
Underfitting
MoG
P-MoG
MoEP
Overfitting
MRF-MoG
Niid-MoG
Online-MoG
Domain
Specific
CT-
Denoising

- Non-L2 robust loss:
  - **L1-SPCA** <u>*Pattern Recognition*</u> ***2012***
  - **L1-Small Target detection** <u>*TIP*</u> ***2013***
  - **L1-LRMF** <u>*AAAI*</u> ***2013***
  - **L1-Bayes-LRMF** <u>*TNNLS*</u> ***2015***

- General iid Noise Modeling:
  - **MoG-LRMF** <u>*ICCV*</u> ***2013***
  - **MoG-RPCA** <u>*ICML*</u> ***2014***
  - **MoEP** <u>*ICCV*</u> ***2015***
  - **MoG-TF** <u>*CVPR*</u> ***2016***
  - **P-MoG** <u>*ICCV*</u> ***2017***

- Noise Modeling with T/S/D Prior:
  - **MRF-MoEP** <u>*TIP*</u> ***2016***
  - **Non-iid MoG** <u>*TC*</u> ***2017***
  - **OMoGMF** <u>*TPAMI*</u> ***2017***

- Domain Specific Noise Modeling:
  - **CT-Denoising** : <u>*TMI*</u> ***2017***
  - **Lesion detection**: *In progress*


中国人工智能学会通讯

/科技前沿/

误 差 建 模 原 理

孟德宇 / 西安交通大学

机器学习的优化问题

著名机器学习专家，卡内基梅隆大学的 Tom Mitchell 教授曾经用三个要素定义了机器学习的基本概念[1]——基于经验 E，针对学习目标 T，提升表现度量 P。经验是学习的信息来源，代表了预先获取并输入机器学习算法的训练数据或信号观察。如对于有监督学习问题，经验体现为数据及其标号构成的有标注数据集合；而对于无监督学习问题，经验就仅体现为数据本身构成的无标注数据集。学习目标是学习的最终任务，体现为机器学习方法预期获得的最终输出结果。如对于判别问题，学习目标为对未知数据能够执行判别的决策函数；对于信号复原问题，学习目标为对输入信号进行恢复的信号等。表现度量是学习的实现手段，对应于一个衡量学习效果满意与否的量化标准。通过将表现度量作为优化问题的目标函数并设计优化算法对其进行求解，机器学习的任务得以实现。

如何构造机器学习的表现度量，或者说如何对机器学习的优化问题进行建模，是对给定数据与特定学习目标进行有效机器学习的核心问题。通常采用如下模型来实现这一目的：

$$\min_{f(W)\in\Omega} L(f(W),D)+P(W) \quad (1)$$

该模型中，D 代表了训练数据 $f(W)$ 代表了学习目标 W 代表了模型参数。模型中主要包含三个因素，其物理意义分别如下：$L(f(W),D)$ 为误差函数，代表了学习目标对训练数据的拟合精度，最常见的形式包括最小二乘误差（即 L_2 范数误差）与稳健误差（如 L_1 范数误差等）等；$P(W)$ 为正则项，编码了模型参数的先验信息，常采用的形式包括 Ridge 正则（即 L_2 范数正则）或稀疏正则（如 L_1 范数正则等）；Ω 称为学习机，代表了一个预先设置的学习目标可行集，其功能为对学习目标的学习范围进行约束。

机器学习首先要关注的问题，当然是学习目标对训练数据的有效拟合。这一目标又可称为经验风险极小化原理 (ERM)[2]。具体来说，就是如何针对输入数据 D，使在经验数据上的误差，$L(f(W),D)$ 尽可能小的问题。神经网络是基于这一原理实现的典型机器学习方法，其跌宕起伏的发展历程，几乎可以概括为一部"成也拟合，败也拟合"的历史。最早期由 Rosenblatt 提出的感知机概念[3-4]，事实上对应于一个包含简单加减法的两层神经网络。尽管在字母识别等应用中体现了一定的应用效果，但其拟合能力有限的问题很快受到其他科学家的质疑。特别是，在 1969 年著名科学家 Minsky 与 Papert 关于感知机的著作中[5]，评价感知机甚至无法对经典异或问题进行有效拟合。此问题导致之后十多年的时间，神经网络的研究几乎停滞不前。其再一次复兴的标志成果为 1975 年出现的反向传播算法（BP 算法）[6]。该算法解决了神经网络对异或回路的拟合学习问题，并实现了多层神经网络的有效训练策略。由于理论可以证明，多层网络具有对广泛函数的万有逼近（universal approximation）性能[7]，机器学习对于数据的拟合问题似乎得以很好的解决，这也带来了神经网络在上世纪八九十年代的又一次研究热潮。

然而，学者们很快发现，片面地追求在训练数据上的拟合精度往往也是有问题的。当采用形态较为复杂的学习目标对学习任务进行训练时，在训练数据上拟合精度可能很高；然而当对未包含在训练数据中的测试数据进行预测时，精度却可能较差。

01