Age Estimation Using Expectation of Label Distribution Learning

Bin-Bin Gao[1], Hong-Yu Zhou[1], Jianxin Wu[1], Xin Geng[2]

[1]LAMDA Group, Nanjing University, China

[2]PALM Group, Southeast University, China

Jul. 19, 2018 Stockholm

Background

Face information

- Identity
- Emotion
- Ethnicity
- Gender
- Attractiveness
- **Age**
- ……

This information plays a significant role during face-to-face communication between humans.

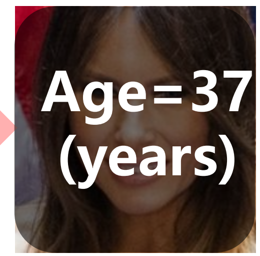
Background

What is facial age estimation?

It attempts to automatically predict age based on an individual face.

Testing face

Age Model



Training images:

Background

Potential applications

Law enforcement	Security control	Recommendations

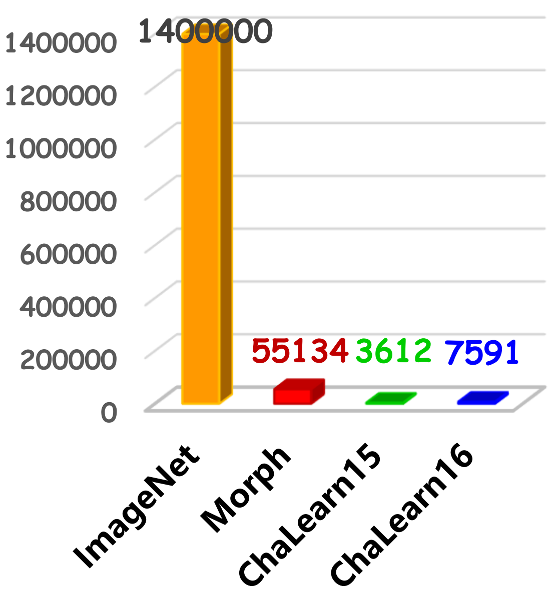
……

Automatic age estimation from face images is an attractive yet challenging topic.

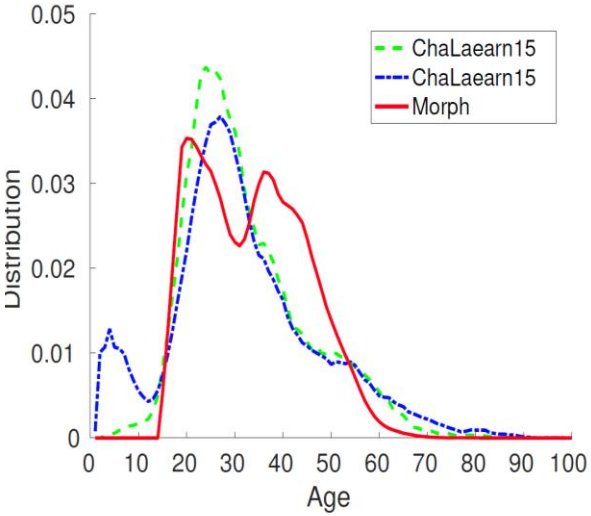
Background

Challenges

Insufficiency



Imbalance



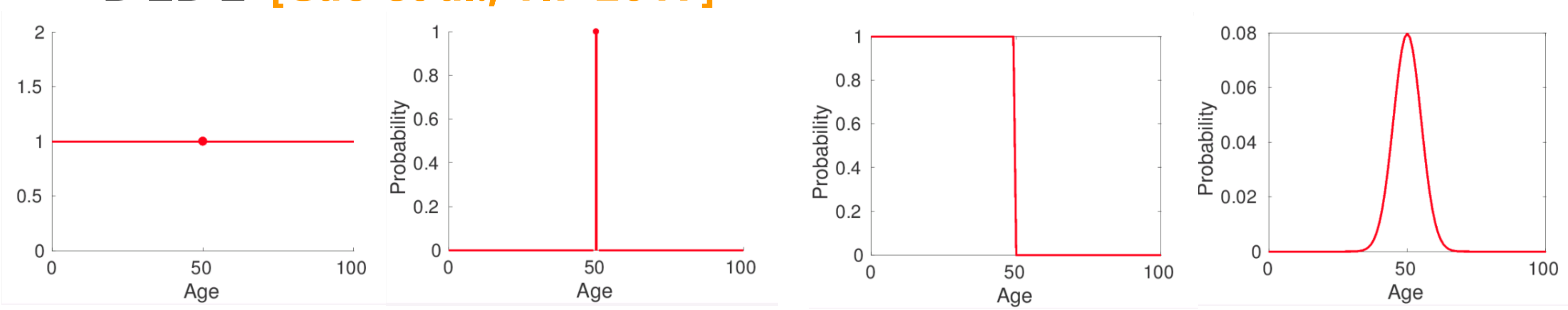
Fine-grained Recognition

36 37

Related Works

Plenty of deep methods are proposed,

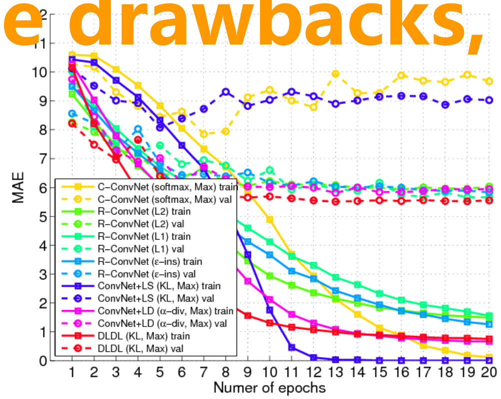
- MR: Metric Regression **[Ranjan et al., FG 2017]**
- DEX: Classification **[Rothe et al., IJCV 2016]**
- Ranking **[Chen et al., CVPR 2017]**
- DLDL **[Gao et al., TIP 2017]**



Regression | Classification | Ranking | DLDL

Motivation

Pervious works have some notable drawbacks,



- Classification and regression may lead to an **unstable training** procedure.

- There is an **inconsistency** between the training objectives and evaluation metric in DLDL and Ranking.

Objective: Fit Distribution

Evaluation: MAE

- Almost all state-of-the-arts have **huge computational cost and storage overhead**.

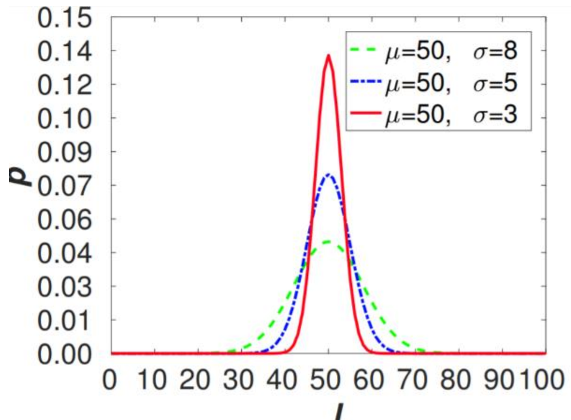
VGG16 iPhone6 iPhone7 iPhonex

500M 1G 2G 3G

Proposed Method

LAMDA
Learning And Mining from DatA
http://lamda.nju.edu.cn

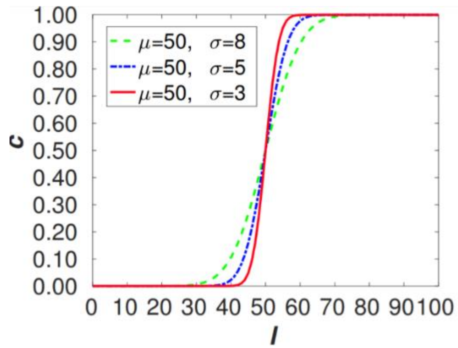
Ranking is learning label distribution



Label Distribution

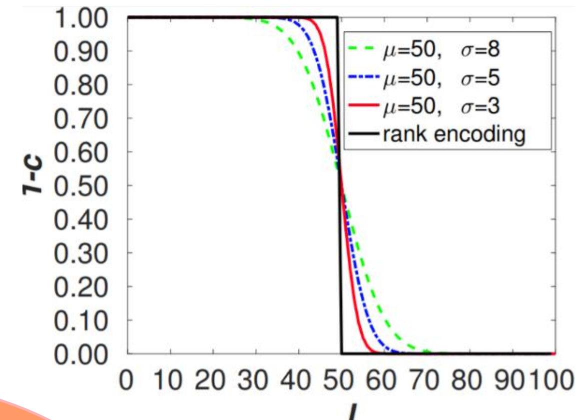
$$\boxed{p^{ld} = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(l-y)^2}{2\sigma^2}\right)}$$

Normal Distribution



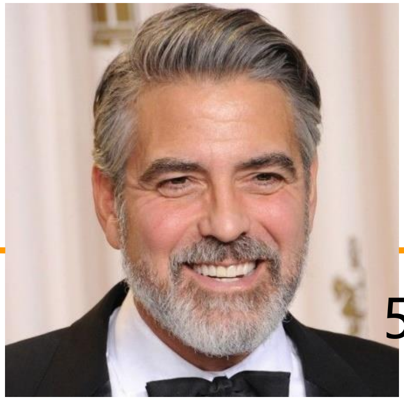
c.d.f

$$c = \frac{1}{2}\left[1+\operatorname{erf}\left(\frac{l-y}{\sigma\sqrt{2}}\right)\right]$$



Ranking Encoding

$$\boxed{p^{rank} = [\underbrace{1,\ldots,1}_{|y|},\underbrace{0,\ldots,0}_{|K-1-y|}]}$$



50-year-old

Proposed Method

Learning And Mining from DatA
http://lamda.nju.edu.cn

Ranking is learning label distribution

Label Distribution

$$\boldsymbol{p}^{ld} = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(\boldsymbol{l}-\mathrm{y})^2}{2\sigma^2}\right)$$

CDF

$$\boldsymbol{c} = \frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{\boldsymbol{l}-\mathrm{y}}{\sigma\sqrt{2}}\right)\right]$$

Ranking Encoding

$$\boldsymbol{p}^{rank} = [\underbrace{1,\ldots,1}_{\lfloor \mathrm{y} \rfloor}, \underbrace{0,\ldots,0}_{\lfloor K-1-\mathrm{y} \rfloor}]$$

$$\boldsymbol{c} = \mathbf{T}\boldsymbol{p}^{ld}$$

$$\boldsymbol{p}^{rank} \approx \mathbf{1} - \boldsymbol{c}$$

$$\boldsymbol{p}^{rank} \approx \mathbf{1} - \mathbf{T}\boldsymbol{p}^{ld}$$

There is a linear relationship.

- Label distribution can represent more meaningful age information.
- Label distribution learning is more efficient.

Proposed Method

DLDL-v2

- Label Distribution Module
  - Linear transformation

$$\boldsymbol{x} = \mathbf{W}^T \boxed{\boldsymbol{f}} + \boldsymbol{b}$$

CNN feature

  - Label distribution

$$\hat{p}_k = \boxed{\frac{\exp(x_k)}{\sum_t \exp(x_t)}}$$

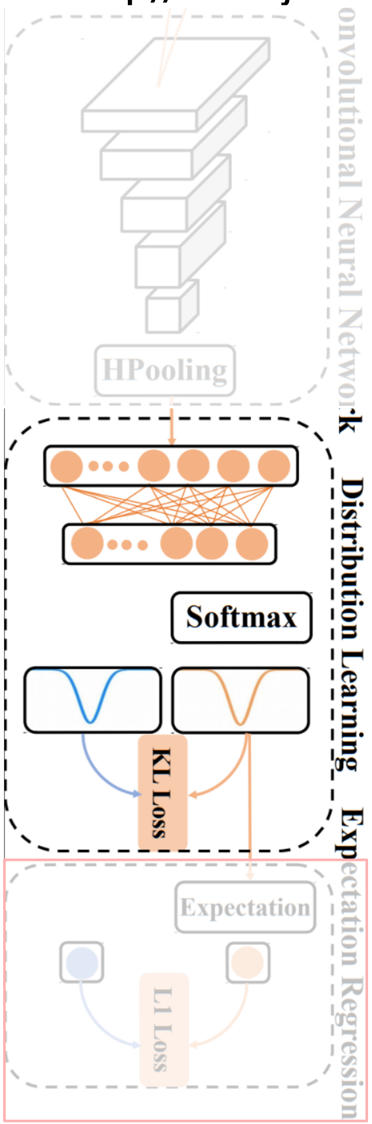
Softmax

  - Loss: KL-Div

$$L_{ld} = \sum_k p_k \ln \frac{\boxed{p_k}}{\boxed{\hat{p}_k}}$$

Label Dis

Pred Dis



Proposed Method

DLDL-v2

- Expectation Regression Module
  - Expectation layer

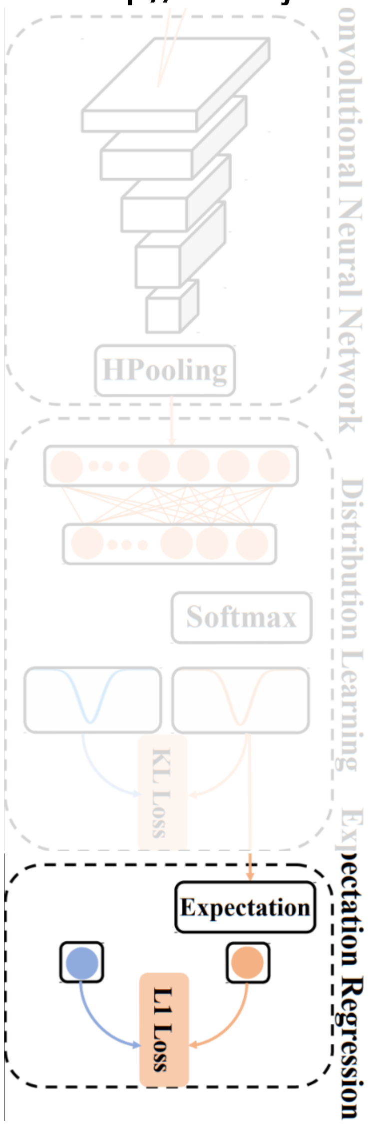
$$\hat{y} = \sum\nolimits_k \hat{p}_k \boxed{l_k}$$

Label Set

  - Loss: l_1

$$L_{er} = |\hat{y} - y|$$

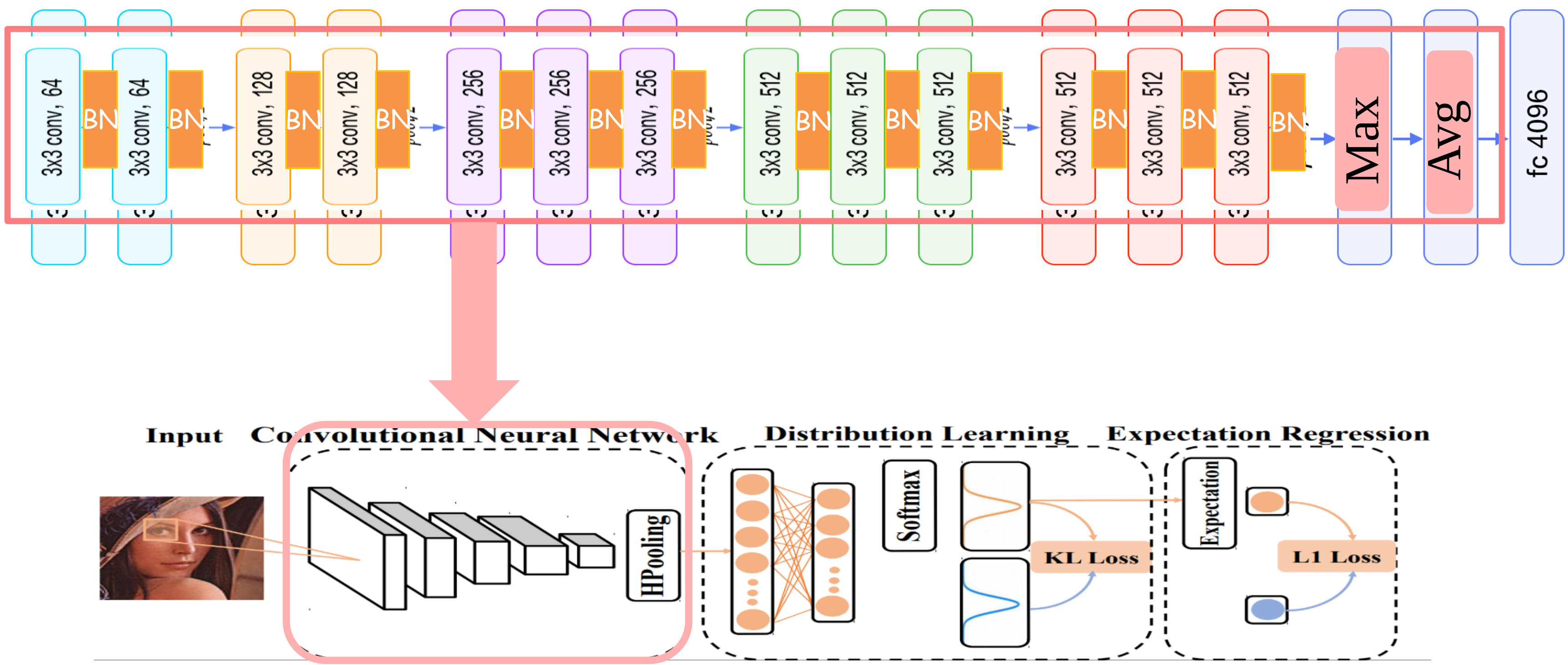
This module does not introduce any new parameter.



Proposed Method

DLDL-v2

- Network Architecture



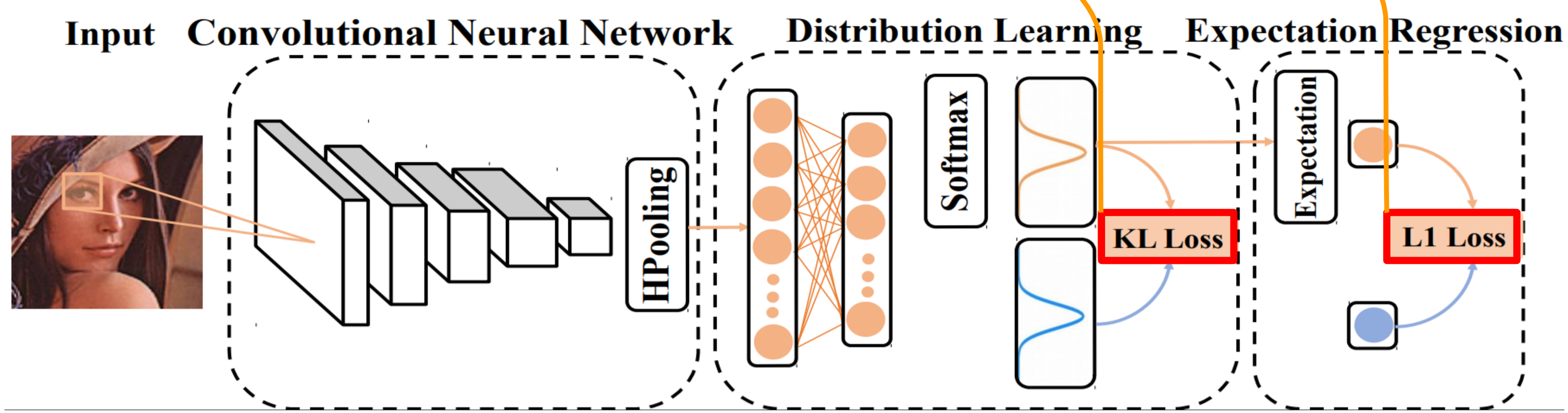
Proposed Method

DLDL-v2

- Jointly Learning (SGD algorithm)

Weight

$$L = \boxed{L_{ld}} + \boxed{\lambda}\boxed{L_{er}}$$



Experiments

Learning And Mining from DatA
http://lamda.nju.edu.cn

Datasets

- Apparent age
  - ChaLearn15 (2476+1136)
  - ChaLearn16 (5613+1978)

- Real age
  - Morph (55134: 80%+20%)

Evaluation metric

- MAE：mean average error
- e-error: It is defined by the ChaLearn.

Experiments

Comparisons with state-of-the-arts

Table 1: Comparisons with state-of-the-art methods for apparent and real age estimation.

Methods	External Data	ChaLearn15 MAE	ChaLearn15 ϵ-error	ChaLearn16 MAE	ChaLearn16 ϵ-error	Morph MAE
Human [Han *et al.*, 2015]	×	-	0.34	-	-	6.30
OR-CNN [Niu *et al.*, 2016]	×	-	-	-	-	3.34
DEX [Rothe *et al.*, 2018]	×	5.369	0.456	-	-	3.25
DEX [Rothe *et al.*, 2018]	√	3.252	0.282	-	-	2.68
DLDL [Gao *et al.*, 2017]	×	3.51	0.31	-	-	2.42[1]
Ranking [Chen *et al.*, 2017]	×	-	-	-	-	2.96
LDAE [Antipov *et al.*, 2017]	√	-	-	-	0.241[2]	2.35
DLDL-v2 (TinyAgeNet)	×	3.427	0.301	3.765	0.291	2.291
DLDL-v2 (ThinAgeNet)	×	**3.135**	**0.272**	**3.452**	**0.267**	**1.969**

[1]Used 90% of Morph images for training and 10% for evaluation;
[2]Used multi-model ensemble;

Table 2: Comparisons of model parameters and forward times with state-of-the-arts.

Methods	#Param(M)	#Time(ms)
DEX [Rothe *et al.*, 2018]	134.6	133.30
DLDL [Gao *et al.*, 2017]	134.6	133.30
LDAE [Antipov *et al.*, 2017]	1480.6	1446.30
DLDL-v2 (TinyAgeNet)	0.9	24.26
DLDL-v2 (ThinAgeNet)	3.7	51.05

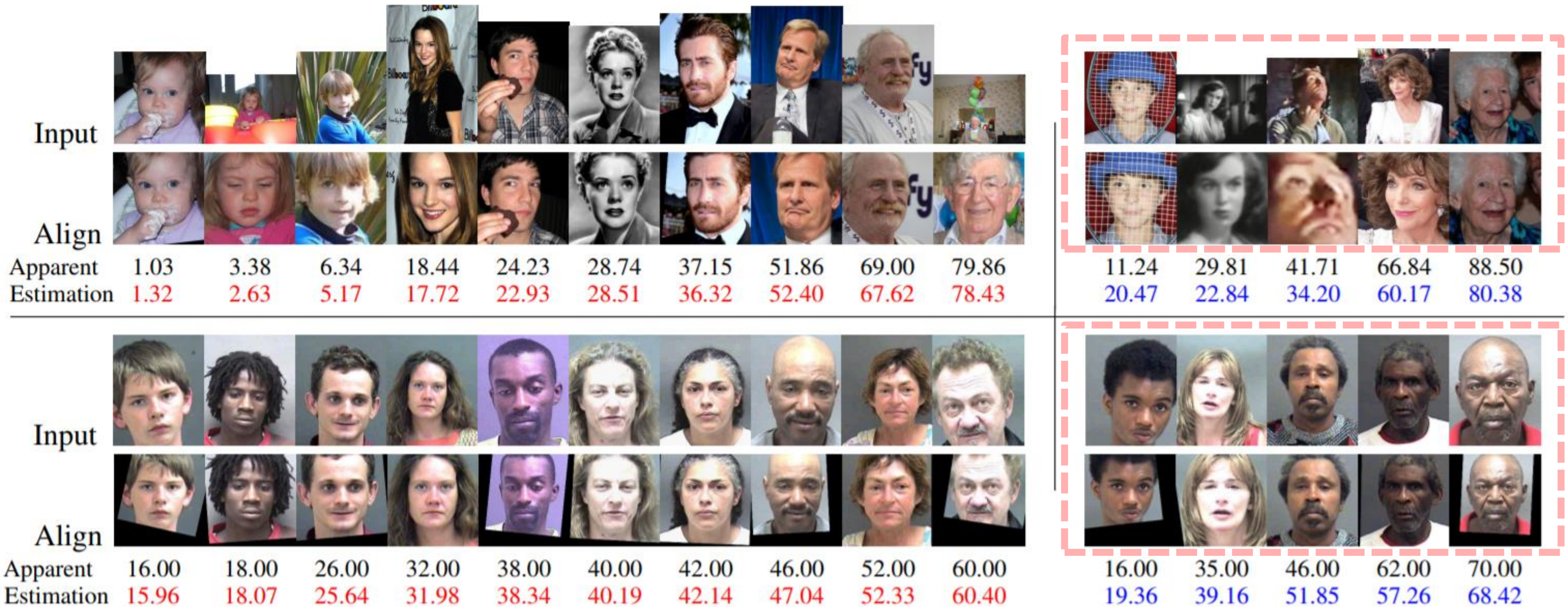
32 images in ms on one M40 GPU

150×
36×

5.5×
2.6×

Experiments

Visual assessment



Good examples

Poor examples

Experiments

Ablation study

- Comparisons

Table 3: Comparison of different methods.

Methods	Factors		ChaLearn15		ChaLearn16		Morph
	Aug	Pool	MAE	ϵ-error	MAE	ϵ-error	MAE
DLDL-v2	×	HP	3.399	0.303	3.717	0.290	2.346
	√	GAP	3.210	0.282	3.539	0.274	2.039
	√	HP	**3.135**	**0.272**	**3.452**	**0.267**	**1.969**
MR (ℓ_2)	√	HP	3.665	0.337	3.696	0.294	2.282
MR (ℓ_1)	√	HP	3.655	0.334	3.722	0.301	2.347
DEX	√	HP	3.558	0.306	4.163	0.332	2.311
Ranking	√	HP	3.365	0.298	3.645	0.290	2.164
ER (ℓ_1)	√	HP	3.287	0.291	3.641	0.282	2.214
DLDL	√	HP	3.228	0.285	3.509	0.272	2.132

It means that erasing the inconsistency between training and evaluation stages can help us make a better prediction.

Experiments

Ablation study

- Sensitivity of hyper-parameters

λ: Loss weight

$$0.01 \leq \lambda \leq 10$$

$\Delta l\ (K)$ The number of discrete labels

$$0.25 \leq \Delta l \leq 4.$$

Table 4: The influences of hyper-parameters.

Hyper-param		ChaLearn15		ChaLearn16		Morph
λ	$\Delta l\ (K)$	MAE	ϵ-error	MAE	ϵ-error	MAE
0.01	1 (101)	3.223	0.282	3.493	0.270	**1.960**
0.10	1 (101)	3.188	0.278	3.455	0.268	1.972
1.00	1 (101)	**3.135**	**0.272**	**3.452**	0.267	1.969
10.00	1 (101)	3.144	0.273	3.487	0.270	1.977
1.00	4 (26)	3.182	0.276	3.473	0.270	1.963
1.00	2 (51)	3.184	0.274	3.484	0.271	1.963
1.00	0.50 (201)	3.184	0.278	3.484	0.269	1.992
1.00	0.25 (401)	3.167	0.274	3.459	**0.265**	2.028

Our method is not sensitive to these hyper-parameters.

Understanding DLDL-v2

How does DLDL-v2 estimate facial age?

The network uses different patterns to estimate different age.

Conclusion

- We provide the first analysis and show that *the ranking method is in fact learning label distribution implicitly*. This result thus unifies existing state-of-the-art facial age estimation methods into the DLDL framework.

- We propose an end-to-end learning framework which *jointly learns age distribution and regresses single-value age in both feature learning and classifier learning*.

- We *create new state-of-the-art results on facial age estimation tasks* using single and small model without external age labeled data or multi-model ensemble.

Thanks !

Projects