MCFL: Multi-label Contrastive Focal Loss for Pedestrian Attribute Recognition

Xiaoqiang Zheng[1,2], Zhenxia Yu[1], Lin Chen[2*], Fan Zhu[2] , Shilog Wang[1]

[1]Chengdu University of Information Technology, China

[2]Chongqing Institute of Green and Intelligent Technology, Chinese Academy of Science

Corresponding author: Lin Chen, chenlin@cigit.ac.cn

Outline

1. Background and motivation

2. The proposed model

3. Experiments

4. Conclusion

Outline

1. Background and motivation

2. The proposed model

3. Experiments

4. Conclusion

1. Background and motivation: Introduction

- **Pedestrian Attribute Recognition** is a key element in video surveillance, aiming at predict a group of attributes to describe the characteristic of human
- **Proving useful clues for smart video analysis**
  - Person Retrieval
  - Human Identification
  - Customer Analysis

Male
Age31-45
BodyNormal
Customer
hs-BlackHair
hs-Glasses
ub-ShortSleeve
lb-LongTrousers
shoes-Leather

Male
Age17-30
BodyNormal
Customer
hs-BlackHair
hs-Glasses
ub-Cotton
lb-LongTrousers
shoes-Casual
attach-Backpack

Female
Age17-30
BodyNormal
Customer
hs-LongHair
hs-BlackHair
hs-Glasses
ub-Cotton
lb-TightTrousers
shoes-Boots

Male
Age17-30
BodyNormal
Customer
hs-BlackHair
hs-Glasses
ub-Shirt
ub-Tshirt
lb-LongTrousers
shoes-Leather

Male
Age31-45
BodyNormal
Customer
hs-BlackHair
hs-Glasses
ub-Jacket
lb-LongTrousers
shoes-Sport
attach-SingleShoulderBag

Female
Age17-30
BodyFat
Customer
hs-BlackHair
ub-TShirt
lb-TightTrousers
shoes-Boots
attach-Other

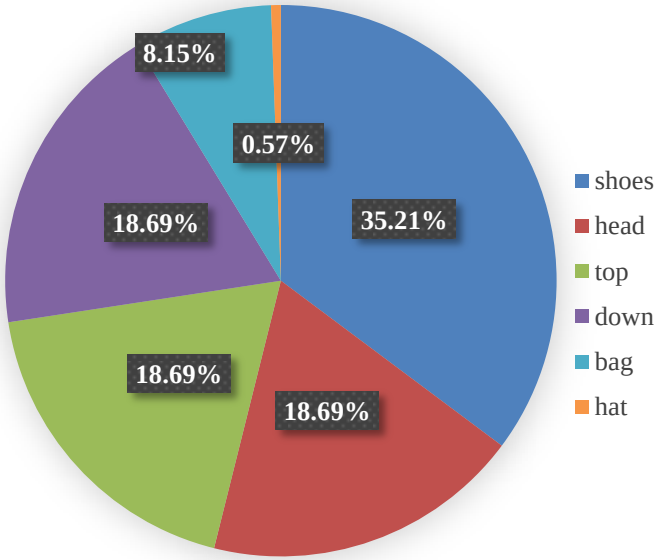
Female
Age31-45
BodyNormal
Clerk
hs-LongHair
hs-BlackHair
ub-Sweater
ub-Vest
lb-TightTrousers

1. Background and motivation: Introduction

- Deep learning based algorithms have been proposed and achieved remarkable results for pedestrian attributes recognition.
- However, it is still a challenging work.

Pose Variation

Occlusion



Data Imbalance

Outline

1. Background and motivation

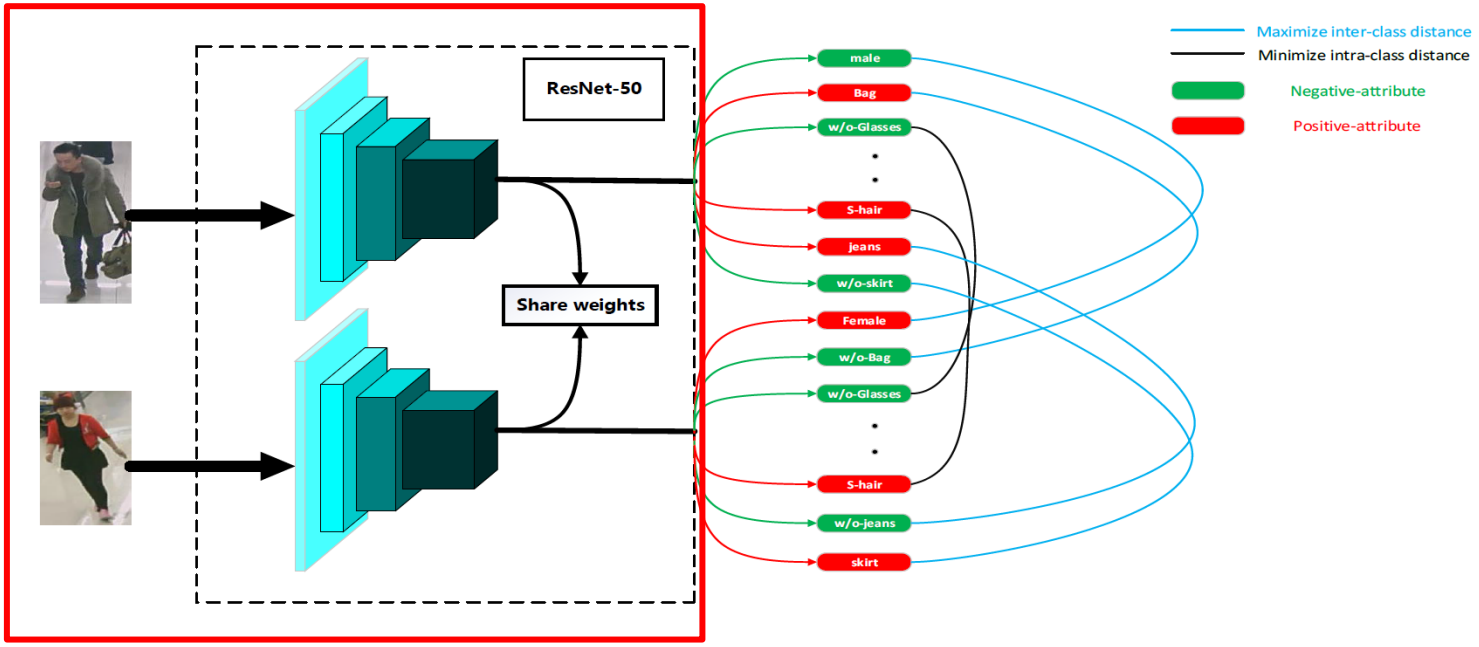
2. The proposed model

3. Experiments

4. Conclusion

2. The proposed loss function: MCFL

- **Multi-label Contrastive Focal Loss(MCFL), integrating multi-label focal loss and contrastive loss simultaneously.**
- **Focusing on the difficult and error-prone positive attributes.**
- **Enlarging the discriminate gap between positive and negative samples in multi-label attributes learning.**



Siamese neural network

2. The proposed loss function: MCFL

➢ Multi-label Contrastive Focal Loss

- **Multi-label Focal Loss：**

$$L_i^p = -\mathbf{sw}_i(1 - Pt_i)\mathbf{y_i}log(Pt_i)$$

$$L_i^n = -\mathbf{sw}_i Pt_i(1 - \mathbf{y_i})log(1 - Pt_i)$$

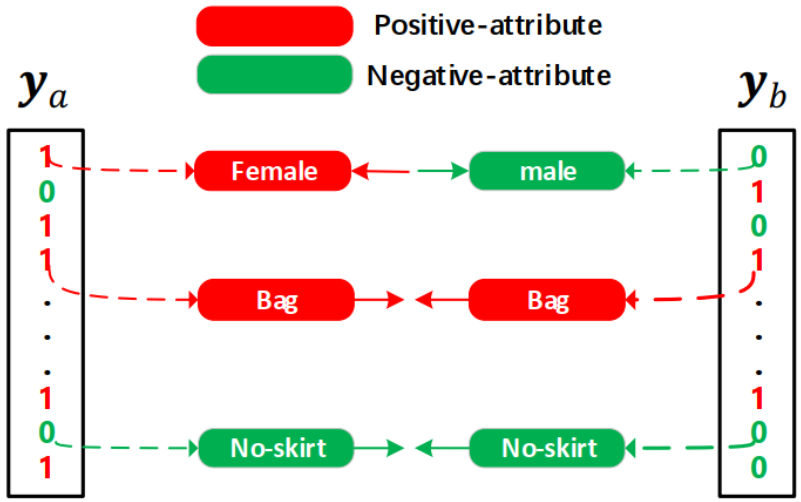
$$\mathbf{sw}_i = exp(\mathbf{y}_i(1 - \frac{\sum_{i=1}^{N}\mathbf{y}_i}{N}) + (1 - \mathbf{y}_i)\frac{\sum_{i=1}^{N}\mathbf{y}_i}{N})$$

- **Multi-label Contrastive Loss**

$$S_p = \frac{\sum_{j=1}^{M} Pt_{aj}Pt_{bj}w_j^p}{\sqrt{\sum_{j=1}^{M}(Pt_{aj}w_j^p)^2\sum_{j=1}^{M}(Pt_{bj}w_j^p)^2}}$$

$$S_n = \frac{\sum_{j=1}^{M} Pt_{aj}Pt_{bj}w_j^n}{\sqrt{\sum_{j=1}^{M}(Pt_{aj}w_j^n)^2\sum_{j=1}^{M}(Pt_{bj}w_j^n)^2}}$$

$$S_d = \frac{\sum_{j=1}^{M} Pt_{aj}Pt_{bj}w_j^d}{\sqrt{\sum_{j=1}^{M}(Pt_{aj}w_j^d)^2\sum_{j=1}^{M}(Pt_{bj}w_j^d)^2}}$$



2. The proposed loss function: MCFL

- **Multi-label Contrastive Focal Loss**

$$L_{intra\_p} = \mathbf{w}^p L_a^p \left(\frac{1}{S_p + eps} + \alpha\right)$$

$$L_{intra\_n} = \mathbf{w}^n L_b^n \left(\frac{1}{S_n + eps} + \beta\right)$$

$$L_{inter\_p} = \mathbf{w}^d \left(\sum_{i=a}^{b} L_i^p\right)(S_d + \alpha)$$

$$L_{inter\_n} = \mathbf{w}^d \left(\sum_{i=a}^{b} L_i^n\right)(S_d + \beta)$$

$$L_{total} = L_{intra\_p} + L_{intra\_n} + L_{inter\_p} + L_{inter\_n}$$

Outline

1. Background and motivation

2. The proposed model

3. Experiments

4. Conclusion

3. Experiments: Evaluation metrics

Pedestrian multi-attribute recognition: Following the evaluation metrics defined

$$mA = \frac{1}{2N}\sum_{i=1}^{L}\left(\frac{TP_i}{P_i} + \frac{TN_i}{N_i}\right)$$

$$Accuracy = \frac{1}{N}\sum_{i=1}^{L}\frac{|Y_i \cap f(x_i)|}{|Y_i \cup f(x_i)|} \qquad Recall = \frac{1}{N}\sum_{i=1}^{L}\frac{|Y_i \cap f(x_i)|}{|Y_i|}$$

$$Precision = \frac{1}{N}\sum_{i=1}^{L}\frac{|Y_i \cap f(x_i)|}{|f(x_i)|} \qquad F1 = \frac{2\times(Precision \times Recall)}{Precision + Recall}$$

3. Experiments: Comparison with STOA

- **Richly Annotated Pedestrian (RAP)** dataset: is one of the few large-scale datasets containing 41,585 samples annotated with 51attributes.
- Totally 33,268 images are used for training and the rest 8,317 images are used for testing.

Methods	BackBone	mA	Acc	Prec	Rec	F1
HPNet(ICCV17) [19]	InceptionNet	76.12	65.39	77.33	78.79	78.05
LGNet(BMVC18) [18]	InceptionV2	78.68	68.00	80.36	79.82	80.09
PGDM(ICME18) [11]	CaffeNet	74.31	64.57	78.86	75.90	77.35
JLPLS-PAA(TIP19) [29]	–	81.25	67.91	78.56	81.45	79.98
RA(AAAI19) [35]	InceptionV3	81.16	–	79.45	79.23	79.34
AAP(AAAI19) [5]	ResNet50	81.42	68.37	**81.04**	80.27	80.65
ALM(ICCV19) [30]	BNInception	81.87	68.17	74.71	**86.48**	80.16
StrongBaseline(2020) [8]	ResNet50	80.52	68.44	79.91	80.64	79.89
MCFL(ours)	ResNet50	**82.06**	**69.01**	77.47	84.91	**81.02**

3. Experiments: Comparison with STOA

PETA [2]: The PETA dataset is composed of 10 small pedestrian re-identification datasets. It consists of 19,000 images including 8,705 people. Following the default setting, the dataset is randomly partitioned into 9500 for training, 1900 for validation, and 7600 for testing, meanwhile, 35 binary attributes are evaluated in experiments.

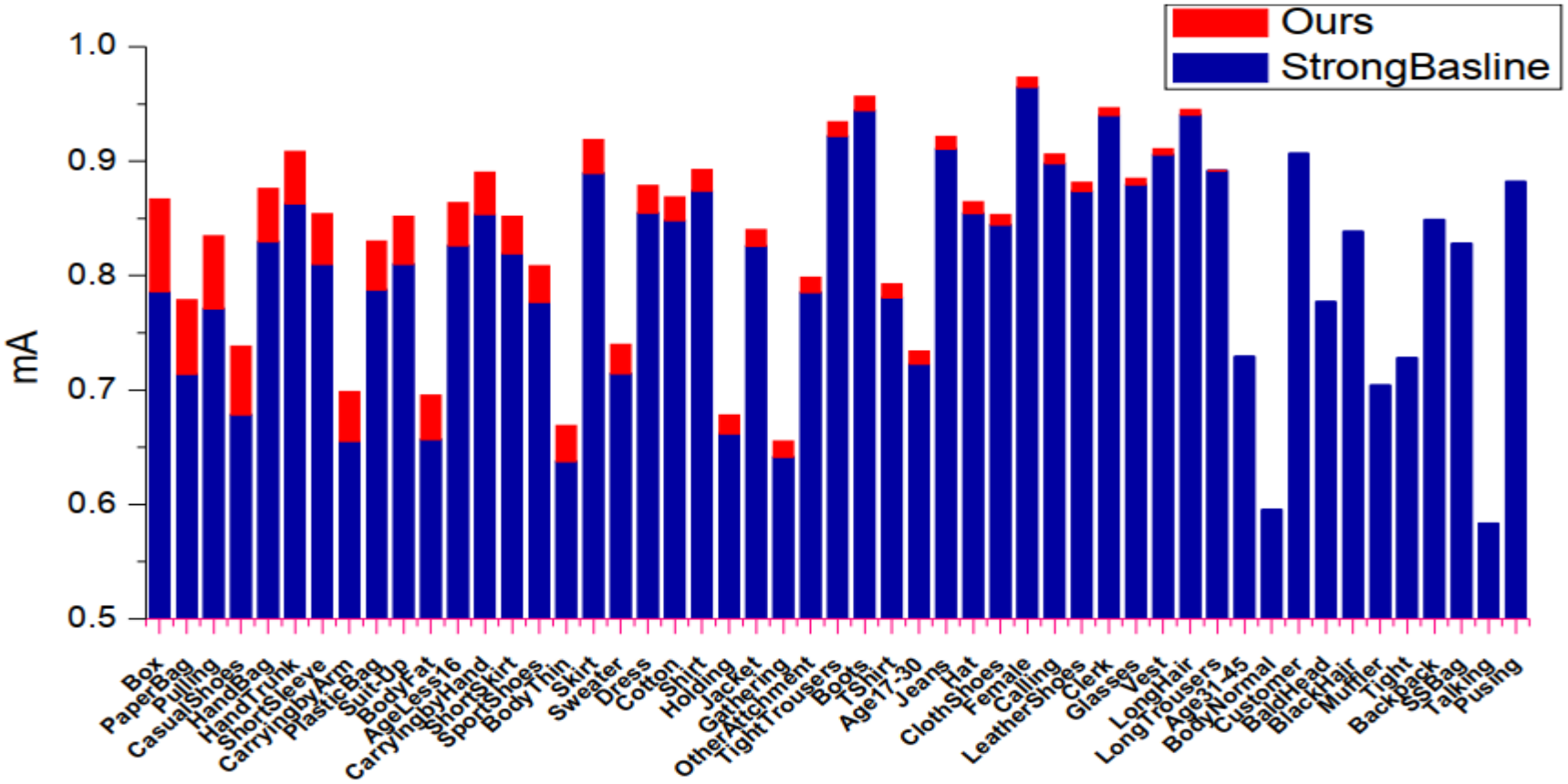
Methods	BackBone	mA	Acc	Prec	Rec	F1
HPNet(ICCV17) [19]	InceptionNet	81.77	76.13	84.92	83.24	84.07
PGDM(ICME18) [11]	CaffeNet	82.97	78.08	86.86	84.68	85.76
RA(AAAI19) [35]	InceptionV3	86.11	–	84.69	88.51	86.56
MsVAA(ECCV18) [22]	ResNet101	84.59	78.56	86.79	86.12	86.46
JLPLS-PAA(TIP19) [29]	–	84.88	**79.46**	87.42	86.33	**86.87**
MT-CAS(ICME20) [33]	ResNet50	83.17	78.78	**87.49**	85.35	86.41
StrongBaseline(2020) [8]	ResNet50	85.19	79.14	87.11	86.18	86.36
MCFL(ours)	ResNet50	**86.84**	78.78	83.68	**89.97**	86.71

3. Experiments: Comparison with STOA

PA100K : The PA-100K dataset is collected real outdoor surveillance scenarios from 598 cameras, it is the largest pedestrian attribute dataset so far containing 100,000 images. The whole dataset is randomly divided into training and testing with the ratio of 9:1. 26 binary attributes are evaluated for each image.

Methods	BackBone	mA	Acc	Prec	Rec	F1
HPNet(ICCV17) [19]	InceptionNet	74.21	72.19	82.97	82.09	82.53
LGNet(BMVC18) [18]	InceptionV2	76.96	75.55	86.99	83.17	85.04
PGDM(ICME18) [11]	CaffeNet	74.95	73.08	84.36	82.24	83.29
AAP(AAAI19) [5]	ResNet50	80.56	78.30	**89.49**	84.36	86.85
ALM(ICCV19) [30]	BNInception	80.68	77.08	84.21	**88.84**	86.46
MT-CAS(ICME20) [33]	ResNet50	77.20	78.09	88.46	84.86	86.62
StrongBaseline(2020) [8]	ResNet50	80.50	78.84	87.24	87.12	86.78
MCFL(ours)	ResNet50	**81.11**	**79.01**	86.67	88.15	**87.41**

3. Experiments: Comparison with STOA



Attributes	Male	Age17-30	Body-Normal	Box	CasualShoes	Short-hair
Ground-truth	✓	✓	✓	✓	✓	✓
StrongBaseline	✓	✗	✓	✗	✗	✓
Ours	✓	✗	✓	✓	✓	✓

Attributes	Female	Age31-45	Body-Normal	HandBag	CasualShoes	LongHair
Ground-truth	✓	✓	✓	✓	✓	✓
StrongBaseline	✓	✓	✗	✗	✓	✓
Ours	✓	✓	✓	✓	✓	✓

3. Experiments: Ablation Study

- Compared with the conventional BCE loss function, the proposed MCFL further improves mA and F1 values by 1.24% and 0.48% over MFL, achieving the highest mA, Rec, and F1 among the tested models

TABLE IV: Ablation study on RAP dataset.

Methods	mA	Acc	Prec	Rec	F1
ResNet50+BCE	80.11	67.92	79.43	80.42	79.55
ResNet50+MFL	80.82	**69.33**	**80.30**	80.79	80.54
ResNet50+MCFL(ours)	**82.06**	69.01	77.47	**84.91**	**81.02**

- The best performance can be achieved when setting α=1.5 and β=1, these values are used in our testing experiments

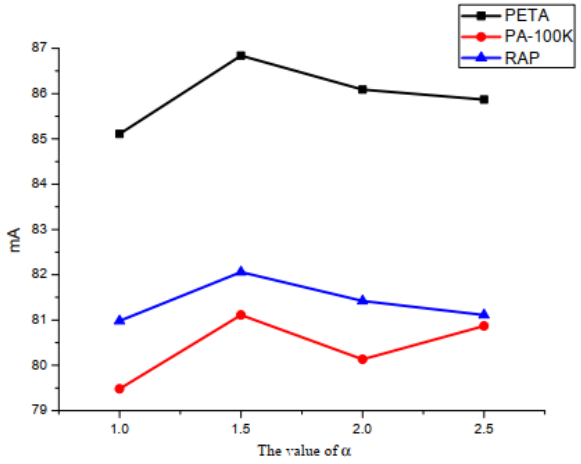


Fig. 6: The impact of α is visualized in terms of mA

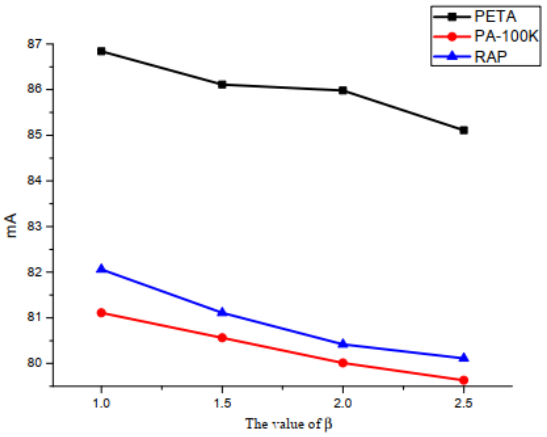


Fig. 7: The impact of β is visualized in terms of mA

4. Conclusion

- **We introduced a method with novel multi-label contrastive focal loss(MCFL) function to improve PAR performance.**
- **This proposed MCFL separates the losses of positive and negative attributes in order to emphasize the hard samples from minority class. Meanwhile, the multi-label contrastive loss is proposed to force CNNs to extract more discriminative features.**
- **Experiments results demonstrate the proposed method outperforms other state-of-the-art methods and can achieve better prediction results on test datasets.**


中国科学院重庆绿色智能技术研究院
Chongqing Institutes of Green and Intelligent Technology, Chinese Academy of Sciences

THANKS!