ASD: TOWARDS ATTRIBUTE SPATIAL DECOMPOSITION FOR PRIOR-FREE FACIAL ATTRIBUTE RECOGNITION

Chuanfei Hu\(^1\), Hang Shao\(^2\), Bo Dong\(^3\), Zhe Wang\(^3\), Yongxiong Wang\(^3\)

\(^1\)School of Automation, Southeast University
\(^2\)School of Computer Science and Engineering, Nanjing University of Science and Technology
\(^3\)School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology

cfhu@seu.edu.cn, shaohang@njust.edu.cn, bodong.cv@gmail.com, wang94627@gmail.com, wyxiong@usst.edu.cn

ABSTRACT

Representing the spatial properties of facial attributes is a vital challenge for facial attribute recognition (FAR). Recent advances have achieved the reliable performances for FAR, benefiting from the description of spatial properties via extra prior information. However, the extra prior information might not be always available, resulting in the restricted application scenario of the prior-based methods. Meanwhile, the spatial ambiguity of facial attributes caused by inherent spatial diversities of facial parts is ignored. To address these issues, we propose a prior-free method for attribute spatial decomposition (ASD), mitigating the spatial ambiguity of facial attributes. The attribute components could be formally described in terms of the spatial locations without any extra prior information. Experimental results demonstrate the superiority of ASD compared with state-of-the-art prior-based methods on both CelebA and LFW A.

Index Terms—Facial attribute recognition, deep learning, attribute decomposition, prior-free method

1. INTRODUCTION

Facial attribute recognition (FAR) aims to predict multiple biometrics from a given facial image, such as gender, mustache, lip thickness, and hair color, which has been widely contributed to numerous image/video analysis applications, e.g., facial identification [1], face generation [2], and face retrieval [3]. However, FAR is still an open issue in the wild, since facial appearance is variable significantly caused by the complicated factors of capturing, such as illumination, pose, and occlusion.

Recent efforts have been made towards improving deep learning-based FAR methods [5–13] via additional annotations, which can be seen as an insight of modeling the spatial properties of facial attributes via extra prior information. These prior-based methods can be categorized as the explicit and implicit prior information-based methods. The explicit prior information-based methods introduce the additional annotations, such as facial landmarks [8], identifications [6,10], and face parsing masks [7,10], to construct a multi-task learning or self-supervised learning, in which the additional annotations can be regarded as the prior information to enhance the discriminability between the features of facial attributes. On the other hand, the implicit prior information-based methods design a multi-task learning network via attribute groups [5,6,9,11], which can group the attributes manually according to locations or semantics. Although these methods achieve the promising results in FAR, there are three major drawbacks. First, the explicit prior information might not be always available, resulting in the restricted application scenario of these methods. Meanwhile, the manual grouping of facial attributes is not suitable and optimal, since the different individuals might give different partitions according to locations or semantics. It means that the relationships among facial attributes could not be defined sufficiently by prior experiences. Furthermore, the inherent spatial diversities of facial parts, caused by different individuals and various poses, are considered insufficiently. For instance, we illustrate the distributions of facial landmarks, including nose, eyes, and mouth, from all individuals in CelebA [4], as shown in Figure 1. The conspicuous overlapping regions of the convex hulls reveal that the spatial ambiguity of facial attributes is inherent. In the above mentioned methods, flattening and global pooling are utilized to vectorize the convolutional features, where the spatial properties are diluted, resulting in the spatial ambiguity of facial attributes, as shown in Figure 2. Therefore, a challenging issue remains:

This work was supported in part by the Natural Science Foundation of Shanghai under Grant 22ZR1443700. (Corresponding author: Yongxiong Wang.)
Specifically, the spatial ambiguity of facial attributes means that multiple attribute features might be represented on the same spatial location of feature map. Intuitively, to alleviate the spatial ambiguity, conducting a classifier to recognize the attributes from each spatial location of feature map might be an available approach. However, the traversal of all positions would produce redundant predictions. Since the spatial-wise labels of attributes are unavailable, we have to build a full label-space with all attributes for each spatial location, resulting in many trivial prediction results. Furthermore, a strategy of integrating all predictions would raise the issue of increase in complexity.

In this paper, we focus on FAR method without any extra prior information, denoted as prior-free FAR method. A novel module whose capability of attribute spatial decomposition (ASD) is proposed, termed as assignment-embedding module (AEM). Motivated by hidden factor analysis (HFA) [14], a facial image can be seen as a combination of facial attribute components. Therefore, ASD can be intuitively cast as decomposing the attribute components in the spatial dimension via latent factors, incorporated into a paradigm of supervised learning-based FAR. To further enlarge the discriminability of attribute embeddings, correlation matrix minimization (CMM) is introduced to constrain the correlations among latent factors, muting the impact caused by relationships among latent factors on the procedure of assignment. To summarize, the main contributions are as follows:

- A prior-free FAR method is proposed to decompose the attribute components in the spatial dimension, thereby mitigating the spatial ambiguity of facial attributes.
- A novel module, termed as assignment-embedding module (AEM), is proposed to enable the procedure of ASD via attribute-to-location assignment and location-to-attribute embedding.
- Correlation matrix minimization (CMM) is introduced to decorrelate the latent factors, strengthening the discriminability of the decomposed attribute embeddings.

2. RELATED WORKS

2.1. Facial attribute recognition

With the significant success of deep learning in computer vision community [15–20], deep learning-based methods [5–13] have emerged in the field of FAR. The above mentioned methods often achieve the competitive performances relying on extra prior information. However, the extra prior information, such as identifications, landmarks, and face parsing masks, might be not available, resulting the limitation of these methods. In this paper, we focus on modeling FAR method without any extra prior information.

2.2. Decomposition of facial components

Face is composed of different characteristics physiologically, such as eyes, nose, and lips. Therefore, the facial representation in facial analysis tasks can be intuitively decomposed into various task-related facial components. Recently, inspired by hidden factor analysis (HFA) [14], the insight of decomposition is introduced to the deep learning-based facial analysis methods [21–23]. The above mentioned methods of facial analysis decompose the face into global-level facial components, ignoring the ambiguity of facial components on the spatial locations. Here, we propose ASD which focuses on the decomposition of facial attribute components in the spatial dimension based on the spirit of HFA. The attribute components are described via latent factors on each spatial location to alleviate the spatial ambiguity.

3. METHODOLOGY

3.1. Overview

ASD aims to represent facial attribute components for each spatial location of a facial image via latent factors, alleviating the spatial ambiguity without any extra prior information. ASD is a key idea in the procedure of ASD, which is adapted for a deep learning-based paradigm, as shown in Figure 3.

Specifically, given a facial image \(X\) whose feature map \(f \in \mathbb{R}^{H \times W \times C}\) is first obtained via a CNN-based feature extractor \(F\) as \(f = F(X; \theta_F)\), where \(\theta_F\) denotes the learnable parameters of feature extractor, \(C\) is the number of channels and \(H \times W\) is the resolution.

Then, AEM is modeled to decompose the various attribute components of \(f\) in the spatial dimension via attribute-to-location assignment, while the attribute components are integrated via location-to-attribute embedding. The two operations of AEM can be formulated as \(g = \psi(\phi(f; z), f)\), where \(z\) represents the latent factors, \(\phi\) and \(\psi\) denote attribute-to-location assignment and location-to-attribute embedding, respectively.

Finally, a classifier \(P\) is conducted to project the integrated attribute embeddings \(g\) to the label-space \(asp = \)
where \( \alpha \) is defined as noise. \( D \) magnitudes of assignment matrix formally. linearly with assigning magnitudes. Equation 1 can explicitly of attributes, noise and mean of facial feature are combined \( \hat{f} \) can be defined as follows:
\[
\hat{f}_i = \sum_{j=1}^{D} \alpha_{i,j} f_i + \alpha_{i,D+1} f_i + \bar{f},
\]
where \( \bar{f} \in \mathbb{R}^C \) is the mean of feature map, \( f_i \in \mathbb{R}^C \) denotes the feature on the \( i \)-th spatial location of \( f' \). The components of attributes, noise and mean of facial feature are combined linearly with assigning magnitudes. Equation 1 can explicitly describe the magnitude of each component in the spatial dimension, alleviating the spatial ambiguity of facial attributes formally. \( \alpha_{i,j} \) is the assigning magnitude in the \((i,j)\) position of assignment matrix \( A \) which can be generated as follows:
\[
A = \begin{bmatrix}
\alpha_{1,1} & \cdots & \alpha_{1,D+1} \\
\vdots & \ddots & \vdots \\
\alpha_{M,1} & \cdots & \alpha_{M,D+1}
\end{bmatrix},
\]
where \( A \in \mathbb{R}^{M \times (D+1)} \) denotes the assignment matrix. \( \alpha_{M,D+1} = \frac{e^{s(z_{D+1},f_M)}}{\sum_{j=1}^{D+1} e^{s(z_j,f_M)}} \). The assigning magnitudes \( \alpha \) are defined to describe the magnitude of components for each spatial location, including \( D \) attributes and noise. \( s \) denotes a cosine similarity between \( f' \) and \( z \), formulated as \( s(a,b) = \frac{|a \cdot b|}{|a| |b|} \). \( | \cdot | \) is the modulo operation of vector. \( z_j \in \mathbb{R}^C \) is the \( j \)-th learnable latent factor of \( z \), representing the \( j \)-th attribute in the latent space. \( z_{D+1} \) denotes the latent noise factor which is an additional term in the formulation. We argue that \( z_{D+1} \) can guide an explicit assignment of the attribute-unrelated components, enhancing the discriminability of the attribute-related.

**Location-to-attribute Embedding.** After attribute-to-location assignment, attribute components can be extracted from each spatial location explicitly via assigning magnitudes. Then, the attribute components from all locations are integrated to attribute embeddings via location-to-attribute embedding as \( g_j = \sum_{i=1}^{M} \alpha_{i,j} f_i + \bar{f}, \) where \( g_j \in \mathbb{R}^C \) is the \( j \)-th attribute embedding. Furthermore, the entire attribute embeddings \( g \) can be simplified with matrix operations as \( g = A^T f' + \tilde{f}_{(D+1) \times C} \), where \( \tilde{f}_{(D+1) \times C} \in \mathbb{R}^{(D+1) \times C} \) denotes a matrix whose entities are \( \bar{f} \). To formulate the attribute embeddings generally, we preserve the noise embedding as the \((D+1)\)-th embedding. The two operations of AEM are summarized as the pseudo code in Supplementary Material Algorithm S1.

### 3.3. Correlation Matrix Minimization

Since the high correlation among latent factors would weaken the distinctions of assignment magnitudes in \( A \), we introduce a regularization, termed as correlation matrix minimization (CMM), to reduce the correlation among \( z \). The correlation matrix of \( z \) can be obtained as follows:
\[
H = \begin{bmatrix}
s(z_1,z_1) & \cdots & s(z_1,z_{D+1}) \\
\vdots & \ddots & \vdots \\
s(z_{D+1},z_{D+1}) & \cdots & s(z_{D+1},z_{D+1})
\end{bmatrix},
\]
where \( s \) is a cosine similarity function. Then, CMM can be formulated as follows:
\[
\mathcal{L}_{cmm} = \sum_{i=1}^{D+1} (1 - s(z_i,z_i))^2 + \sum_{i=1}^{D+1} \sum_{j=1,i \neq j}^{D+1} s^2(z_i,z_j),
\]
where the first term can be omitted obviously, since \( s(z_i,z_i) \triangleq 1 \), and the second term enforces the off-diagonal elements of correlation matrix to 0, decorrelating the latent factors explicitly.

### 3.4. Classification and Loss Function

For each attribute embedding, we cast the attribute predictions as multiple binary classification tasks. The \( j \)-th representation of attribute embedding is projected to a logit value linearly, and then a sigmoid function \( \sigma \) is utilized to convert the logit value as a probability \( p_j \), which can be formulated as
p_j = \sigma(w^T_j g_j + b_p)$. \(w_p\) and \(b_p\) refer to the weight vector and bias, respectively. The final prediction \(p\) is generated via concatenating the probabilities of attribute embeddings as \(p = [p_1, p_2, \ldots, p_D, p_{D+1}]\), where \(p \in \mathbb{H}^{D+1}\) is the prediction of \(D\) attributes and noise. \(\mathbb{H}\) represents Hamming space. The loss function of the classification is cross entropy loss which can be formulated as follows:

\[
\mathcal{L}_{cls} = -\frac{1}{D+1} \sum_{j=1}^{D+1} y_j \log p_j, \tag{5}
\]

where \(y_j\) is the label of the \(j\)-th attribute. \(y_{D+1}\) is set to 0, since the label of noise is not available. The final loss function can be defined as \(\mathcal{L} = \mathcal{L}_{cls} + \gamma \mathcal{L}_{cnn}\), where \(\gamma\) is a hyper-parameter to balance the importance between the classification and the constraint of latent factors.

### 4. EXPERIMENT

#### 4.1. Datasets

We conduct experiments on two public facial attribute datasets including CelebA [4] and LFWA [4], which are widely used to evaluate the method of FAR.

CelebA is a large-scale facial attribute dataset containing 202,599 facial images divided into 3 subsets in terms of training, validation, and testing. The numbers of facial images for 3 subsets are 162,770, 19,867, and 19,962, respectively. The facial images are annotated with 40 attribute labels. LFWA is another popular facial attribute dataset composed of 13,143 facial images with 6,263 for training, 2,800 for validation, and 4,080 for testing. The facial images are also annotated with the same attribute labels as CelebA. In the experiments, we follow the protocol of CelebA and LFWA that the default training set is used to train our method, while the performance of our method is evaluated on the default testing set.

#### 4.2. Implementation

The experiments are conducted on a workstation with NVIDIA RTX 2080Ti GPUs. The proposed method is implemented based on PyTorch deep learning framework. The backbone of feature extractor is ResNet50 [24] pre-trained on ImageNet [25]. Following [14, 21], the latent factors \(z\) are initialized with the distribution of \(\mathcal{N}(0, I)\). In the training phase, Adam is used to optimize the learnable parameters with a weight decay of 5e-4. The total epochs are 80 and the initial learning rate is 3e-4, while the learning rate is reduced with the decay ratio of 0.1 after every 20 epochs. The value of \(\gamma\) is set to 2e-2 empirically. The input images are scaled as 224 \(\times\) 224 and random flip is used as data augmentation. Particularly, the sizes of batches for training on CelebA and LFWA are 64 and 32, respectively.

#### 4.3. Ablation studies

**Effect of ASD for different feature extractors.** To clarify the effect of ASD, we conduct the ablation experiments based on the different feature extractors. Three popular CNNs for FAR are introduced, including ResNet18 [24], ResNet50 [24], and ResNeXt50 [26]. Specifically, two versions of FAR method for the corresponding feature extractors are constructed. The first version of the method is implemented without ASD, in which the global pooled feature is directly used as the input of the fully connected layer-based classifier. In contrast, the second version of method is designed with ASD. The results of the different feature extractors-based methods in terms of average accuracy are reported in Table 1. The experimental results demonstrate the effect of ASD whose superiority is unaffected by the different feature extractors. Moreover, to interpret ASD intuitively, we visualize the assignment matrix \(A\) which is the key idea of ASD, as shown in Supplementary Material Figure S2 and Figure S3. The heatmaps highlight the corresponding attribute-related regions evidently. It argues that the attribute components can be described via ASD in the spatial dimension. The spatial ambiguity of facial attributes is alleviated, thereby improving the performance of FAR effectively. To consider the trade-off between efficiency and performance, we implement ASD based on ResNet50 in the subsequent experiments.

**Latent noise factor and mean of feature in AEM.** To investigate the impact of latent noise factor \(z_{D+1}\) and mean of feature \(\bar{f}\), we design the methods based on ResNet50 without \(z_{D+1}\) and \(\bar{f}\), separately. Meanwhile, the method without both \(z_{D+1}\) and \(\bar{f}\) is constructed, termed as vanilla AEM. As shown in Table 2, the performance of the method including \(z_{D+1}\) is higher than the vanilla AEM, while the formulation of \(\bar{f}\) in

| Backbone  | w/ ASD | CelebA (%) | LFWA (%) |
|-----------|--------|------------|----------|
| ResNet18  | ✓      | 90.93      | 85.75    |
| ResNet18  | ✓      | 91.67      | 86.68    |
| ResNet50  | ✓      | 91.39      | 86.47    |
| ResNet50  | ✓      | 92.22      | 87.43    |
| ResNeXt50 | ✓      | 91.63      | 86.78    |
| ResNeXt50 | ✓      | 92.21      | 87.44    |

| w/o \(z_{D+1}\) | w/o \(\bar{f}\) | CelebA (%) | LFWA (%) |
|-----------------|-----------------|------------|----------|
| ✓               | ✓               | 91.76      | 86.98    |
| ✓               | ✓               | 91.91      | 87.23    |
| ✓               | ✓               | 91.93      | 87.22    |

**Table 2.** Ablation analysis of \(z_{D+1}\) and \(\bar{f}\). “w/ \(z_{D+1}\)” and “w/o \(\bar{f}\)” denote AEM without latent noise factor and mean of feature, respectively.

| \(\gamma\) | CelebA (%) | LFWA (%) |
|------------|------------|----------|
| 0.0        | 92.03      | 87.27    |
| 2e-1       | 92.05      | 87.24    |
| 2e-2       | 92.22      | 87.43    |
| 2e-3       | 92.12      | 87.33    |

**Table 3.** Ablation analysis of CMM with various \(\gamma\).
the AEM also improves the overall performance.

Effect of CMM. To validate the effect of CMM, we further train the method based on ResNet50 with various $\gamma$. The quantitative results are reported in Table 3. It can be seen that a proper value of $\gamma$ improves the performance of our method obviously. CMM with $\gamma = 2e-2$ achieves a higher accuracy about 0.2% increases on both CelebA and LFWA. It reveals that the proper decorrelation among latent factors can enhance the discriminability of $g$ to improve the overall performance of ASD. However, the high value of $\gamma$ would degenerate the performance of our method. We argue that the excessive minimization of correlation among the latent factors might suppress the useful relations between the attributes implicitly.

4.4. Comparison with State-of-the-art Methods

We compare our method with 12 state-of-the-art methods lately reported on CelebA and LFWA, which can be categorized into as prior-based and prior-free methods. The prior-based methods include AIRRAIR [5], PS-MCNN [6], HSA [7], DMM [8], SlimCNN [9], SSPL [10], MGG-Net [11], HFE [27], APS [12], and ICKD [13], where prior information, such as landmarks, face parsing masks, and attribute prior embeddings, are utilized in the training phase. CSN [28] and TResNetM [29] are general prior-free methods for multi-label classification task. Here, we implement TResNetM on CelebA and LFWA with the training scheme provided in [29], such as optimizer, learning rate, and training epochs. Besides [29], we list the experimental results of other methods reported from corresponding papers, as shown in Table 4.

Comparison with prior-based methods. ASD achieves the competitive performances than state-of-the-art methods on both CelebA and LFWA. It can be observed that there is still a performance gap between ASD and PS-MCNN. However, we argue the superiority of ASD is obvious. The advantage of ASD is that the performance of ASD is not relied on prior extra information. In [6], the performance of PS-MCNN without identified information and prior attribute groups would be deteriorated, whereas the average accuracy of PS-MCNN on CelebA is reduced from 92.98% to 91.15%.

Comparison with prior-free methods. The performance of ASD is superior to CSN and TResNetM. The reason for the low performance of TResNetM might be that the larger size of batches should be conducted, like the setting in [29]. To come up with a fair comparison, we gradually adopt the various sizes of batches for TResNetM trained on CelebA. The experimental results are listed in Table 5. It can be seen that the improvements of TResNetM caused by increasing the size of batches are slight, while there is a significant growth of GPU memory costs in the training phase. Meanwhile, ASD illustrates the superiority in terms of disk memory and GFLOPs, while the comparison is visualized in Supplementary Material Figure S1. Thus, our method achieves a good trade-off between accuracy and efficiency, implying the architecture of ASD without bells and whistles.

Limitation. The major limitation of ASD is the footprint of GPU memory during the training stage. In future works, the lightweight techniques would be further introduced to reduce the training consumption of our method.

5. CONCLUSION

In this paper, we propose a novel prior-free method for facial attribute recognition (FAR), termed as attribute spatial decomposition (ASD), mitigating the spatial ambiguity of facial attributes formally without any extra prior information. Assignment-embedding module (AEM) is modeled to enable ASD via latent factors, while correlation matrix minimization (CMM) is introduced to improve the discriminability of decomposed attribute embeddings. Experimental results demonstrate that ASD achieves the competitive performances without any extra prior information compared with state-of-the-art prior-based methods.

6. REFERENCES

[1] Hugo Proença, “The uu-net: Reversible face de-identification for visual surveillance video footage,” IEEE Trans. Circuits Syst. Video Technol., vol. 32, no. 2, pp. 496–509, 2022.
[2] Hao Wu, Jianyang Gu, Xiaojin Fan, He Li, Lidong Xie, and Jian Zhao, “3d-guided frontal face generation for pose-invariant recognition,” ACM Trans. Intell. Syst. Technol., vol. 14, no. 2, 2023.

[3] Bor-Chun Chen, Yan-Ying Chen, Yin-Hsi Kuo, and Winston H. Hsu, “Scalable face image retrieval using attribute-enhanced sparse codewords,” IEEE Trans. Multimedia, vol. 15, no. 5, pp. 1163–1173, 2013.

[4] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang, “Deep learning face attributes in the wild,” in 2015 ICCV, 2015, pp. 3730–3738.

[5] Jianshu Li, Fang Zhao, Jiashi Feng, Sujoy Roy, Shuicheng Yan, and Terence Sim, “Landmark free face attribute prediction,” IEEE Trans. Image Process., vol. 27, no. 9, pp. 4651–4662, 2018.

[6] Jiajiong Cao, Yingming Li, and Zhongfei Zhang, “Partially shared multi-task convolutional neural network with local constraint for face attribute learning,” in CVPR, 2018, pp. 4290–4299.

[7] Keke He, Yanwei Fu, Wuhao Zhang, Chengjie Wang, Yu-Gang Jiang, Feiye Huang, and Xiangyang Xue, “Harnessing synthesized abstraction images to improve facial attribute recognition,” in IJCAI, 2018, p. 733–740.

[8] Longbiao Mao, Yan Yan, Jing-Hao Xue, and Hanzi Wang, “Deep multi-task multi-label cnn for effective facial attribute classification,” IEEE Trans. Affect. Comput., 2020.

[9] Ankit Kumar Sharma and Hassan Foroosh, “Slim-cnn: A light-weight cnn for face attribute prediction,” in FG, 2020, pp. 329–335.

[10] Ying Shu, Yan Yan, Si Chen, Jing-Hao Xue, Chunhua Shen, and Hanzi Wang, “Learning spatial-semantic relationship for facial attribute recognition with limited labeled data,” in CVPR, 2021, pp. 11911–11920.

[11] Zhenghao Chen, Shuhang Gu, Feng Zhu, Jing Xu, and Rui Zhao, “Improving facial attribute recognition by group and graph learning,” in ICME, 2021, pp. 1–6.

[12] Si Chen, Xinyu Lai, Yan Yan, Da-Han Wang, and Shunzhi Zhu, “Learning an attention-aware parallel sharing network for facial attribute recognition,” J. Visual Commun. Image Represent., p. 103745, 2023.

[13] Si Chen, Xueyan Zhu, Yan Yan, Shunzhi Zhu, Shaozi Li, and Da-Han Wang, “Identity-aware contrastive knowledge distillation for facial attribute recognition,” IEEE Trans. Circ. Syst. Vid. Technol., pp. 1–1, 2023.

[14] Dihong Gong, Zhifeng Li, Dahua Lin, Jianzhuan Liu, and Xiaou Tang, “Hidden factor analysis for age-invariant face recognition,” in ICCV, 2013, pp. 2872–2879.

[15] C. Hu and Y. Wang, “An efficient convolutional neural network model based on object-level attention mechanism for casting defect detection on radiography images,” IEEE Trans. Ind. Electron., vol. 67, no. 12, pp. 10922–10930, 2020.

[16] Yongxiong Wang, Chuanfei Hu, Kai Chen, and Zhong Yin, “Self-attention guided model for defect detection of aluminum alloy casting on x-ray image,” Computers & Electrical Engineering, vol. 88, pp. 106821, 2020.

[17] Chuanfei Hu, Weijie Sheng, Bo Dong, and Xinde Li, “Tntc: two-stream network with transformer-based complementarity for gait-based emotion recognition,” in ICASSP. IEEE, 2022, pp. 3229–3233.

[18] Chenyang Zhao, Chuanfei Hu, Hang Shao, Zhe Wang, and Yongxiong Wang, “Towards trustworthy multi-label sewer defect classification via evidential deep learning,” in ICASSP, IEEE, 2023, pp. 1–5.

[19] Chuanfei Hu, Bo Dong, Hang Shao, Jiapeng Zhang, and Yongxiong Wang, “Toward purifying defect feature for multilabel sewer defect classification,” IEEE Trans. Instrum. and Meas., vol. 72, pp. 1–11, 2023.

[20] Chuanfei Hu, Tianyi Xia, Ying Cui, Quchen Zou, Yuancheng Wang, Wenbo Xiao, Shenghong Ju, and Xinde Li, “Trustworthy multi-phase liver tumor segmentation via evidence-based uncertainty,” Eng. Appl. of Artif. Intel., vol. 133, pp. 108289, 2024.

[21] Yandong Wen, Zhifeng Li, and Yu Qiao, “Latent factor guided convolutional neural networks for age-invariant face recognition,” in CVPR, 2016, pp. 4893–4901.

[22] Yihua Cheng, Yiwei Bao, and Feng Lu, “Puregaze: Purifying gaze feature for generalizable gaze estimation,” in AAAI, 2022, vol. 36, pp. 436–443.

[23] Xiaojin Fan, Mengmeng Liao, Jingfeng Xue, Hao Wu, Lei Jin, Jian Zhao, and Liehuang Zhu, “Joint coupled representation and homogeneous reconstruction for multi-resolution small sample face recognition,” Neurocomputing, vol. 522, pp. 89–104, 2023.

[24] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in CVPR, June 2016, pp. 770–778.

[25] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in CVPR, June 2009, pp. 248–255.

[26] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He, “Aggregated residual transformations for deep neural networks,” in CVPR, 2017, pp. 1492–1500.

[27] Jie Yang, Jiarou Fan, Yiru Wang, Yige Wang, Weihao Gan, Lin Liu, and Wei Wu, “Hierarchical feature embedding for attribute recognition,” in CVPR, 2020, pp. 13052–13061.

[28] Xiangyun Zhao, Yi Yang, Feng Zhou, Xiaoran Yuan, Yingze Bao, and Ying Wu, “Recognizing part attributes with insufficient data,” in 2019 ICCV, 2019, pp. 350–360.

[29] Tal Ridnik, Hussam Lawen, Asaf Noy, Emanuel Ben, Baruch Gilad Shariat, and Itamar Friedman, “Tresnet: High performance gpu-dedicated architecture,” in WACV, 2021, pp. 1399–1408.
Algorithm S1 Pseudo code of assignment-embedding module (AEM) in a PyTorch-like style.

```python
# In: Feature map f. Out: Attribute embeddings g.

class AEM(nn.Module):
    def __init__(self, D, C):
        self.z = nn.Parameter(torch.rand((D+1, C)))

    def forward(self, f):
        W, H, C = f.size()
        # f is flattened as [M, C].
        f = f.view(W*H, C)
        # Attribute-to-location assignment.
        norm_z, norm_f = F.normalize(self.z), F.normalize(f)
        A = torch.exp(norm_f @ norm_z.t())
        A /= torch.sum(A)
        # Location-to-attribute embedding.
        g = A.t() @ f + torch.mean(f)
        return g
```

1. METHODOLOGY

1.1. Assignment-Embedding Module

The two operations (Attribute-to-location Assignment and Location-to-attribute Embedding) of AEM are summarized as the pseudo code in Algorithm S1.

2. EXPERIMENT

2.1. Ablation studies

Effect of ASD for different feature extractors. The experimental results demonstrate the effectiveness of ASD whose superiority is unaffected by the different feature extractors. Moreover, to interpret ASD intuitively, we visualize the assignment matrix $A$ which is the key idea of ASD, as shown in Figure S2 and Figure S3. The heatmaps highlight the corresponding attribute-related regions evidently. It argues that the attribute components can be described via ASD in the spatial dimension. The spatial ambiguity of facial attributes is alleviated, thereby improving the performance of FAR effectively.

2.2. Comparison with State-of-the-art Methods

Comparison with prior-free methods. ASD illustrates the superiority in terms of disk memory and GFLOPs, while the comparison is visualized in Figure S1. It argues that our method achieves a good trade-off between accuracy and efficiency.

Fig. S1. Comparison of our method with TResNetM for the various sizes of batches on CelebA visually. TResNetM denotes the size of batches is 64. The size of circles presents the footprint of disk memory. GFLOPs of TResNetM and ASD are 5.73 and 4.22, respectively, where ASD achieves a good trade-off between accuracy and efficiency.
Fig. S2. The visualization of assignment matrix $A$ for some traditional attributes, such as Black Hair, Arched Eyebrows, Mouth Slightly Open, Big Nose, Wearing Necktie, Wearing Hat, and Sideburns. Specifically, $A$ is rearranged with the size of $H \times W \times (D + 1)$, and the assigning magnitudes are visualized via heatmaps to represent the corresponding attribute components in the spatial dimension. It can be observed that the spatial locations belong to the corresponding facial attributes are highlighted obviously. The capability of ASD to decompose the attribute components is verified visually. Furthermore, we illustrate the assigned noise components in the last column. The attribute-irrelated regions are highlighted via heatmaps, which is argued the insight of introducing the noise embedding $z_{D+1}$. 
**Fig. S3.** The additional visualization of assignment matrix \( A \) for some traditional attributes, such as *Arched Eyebrows*, *Mouth Slightly Open*, and *Big Nose*.