Faceptor: A Generalist Model for Face Perception

Lixiong Qin[1], Mei Wang[2], Xuannan Liu[1], Yuhang Zhang[1], Wei Deng[1], Xiaoshuai Song[1], Weiran Xu[1]⋆, and Weihong Deng[1]

[1] Beijing University of Posts and Telecommunications, Beijing, China
{lxqin,xuweiran,whdeng}@bupt.edu.cn
[2] Beijing Normal University, Beijing, China
wangmei1@bnu.edu.cn

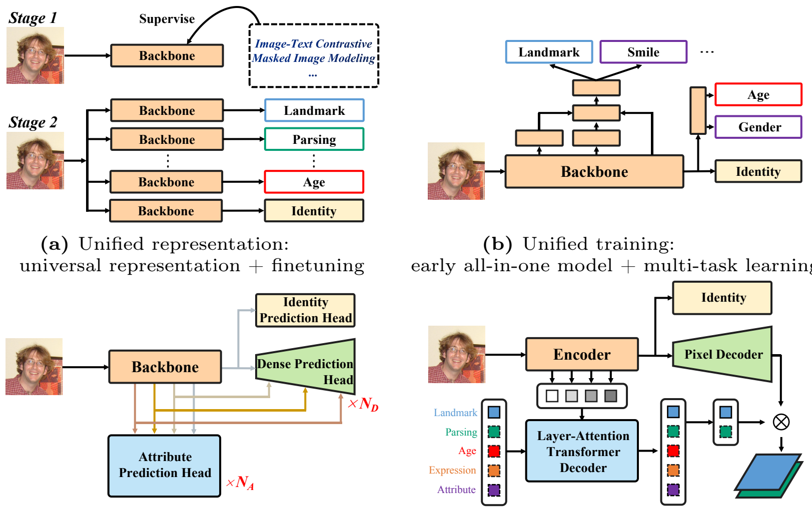
Abstract. With the comprehensive research conducted on various face analysis tasks, there is a growing interest among researchers to develop a unified approach to face perception. Existing methods mainly discuss unified representation and training, which lack task extensibility and application efficiency. To tackle this issue, we focus on the unified model structure, exploring a face generalist model. As an intuitive design, **Naive Faceptor** enables tasks with the same output shape and granularity to share the structural design of the standardized output head, achieving improved task extensibility. Furthermore, **Faceptor** is proposed to adopt a well-designed single-encoder dual-decoder architecture, allowing task-specific queries to represent new-coming semantics. This design enhances the unification of model structure while improving application efficiency in terms of storage overhead. Additionally, we introduce Layer-Attention into Faceptor, enabling the model to adaptively select features from optimal layers to perform the desired tasks. Through joint training on 13 face perception datasets, Faceptor achieves exceptional performance in facial landmark localization, face parsing, age estimation, expression recognition, binary attribute classification, and face recognition, achieving or surpassing specialized methods in most tasks. Our training framework can also be applied to auxiliary supervised learning, significantly improving performance in data-sparse tasks such as age estimation and expression recognition. The code and models will be made publicly available at `https://github.com/lxq1000/Faceptor`.

Keywords: Face perception · Unified model · Transformer

1 Introduction

In recent years, substantial strides have been made in face perception research. Numerous methods have been developed to enhance performance in face analysis tasks such as facial landmark localization [36, 81], face parsing [69, 90], age estimation [20, 68], expression recognition [35, 87], binary attribute classification [24, 49] and face recognition [14, 43, 72]. There are several concerns related to these methods which necessitate a distinct deep model for each task. Firstly,

⋆ Corresponding author.



(a) Unified representation: universal representation + finetuning

(b) Unified training: early all-in-one model + multi-task learning

(c) Unified model structure (ours, shared structural designs): Naive Faceptor

(d) Unified model structure (ours, shared parameters): Faceptor

Fig. 1: Existing efforts for unified face perception mainly concentrate on representation and training. Our work focuses on unified model structure, achieving improved task extensibility and increased application efficiency.

from a methodological perspective, it is not cost-effective to conduct large-scale data collection and model training for each task due to the fact that there is only one object of interest - the human face. Secondly, from a practical perspective, real-world applications often simultaneously require a set of face analysis tasks to cater to specific businesses. It is inefficient to deploy numerous models.

In light of this, researchers have naturally turned their attention toward achieving a unified approach for face perception. Existing efforts mainly concentrate on the following two aspects: (1) Unified representation. As shown in Fig. 1a, FRL [5] and FaRL [93] initially obtain a task-agnostic backbone through universal facial representation learning (unsupervised learning [9], self-supervised learning [2,53], and natural language supervised learning [27,57,58]). By avoiding the need to collect large-scale datasets specifically for supervised pre-training of each task, these approaches improve data efficiency. However, they still require separate finetuning for each downstream task, resulting in low application efficiency in terms of the training process, inference speed, and storage overhead. (2) Unified training. As shown in Fig. 1b, HyperFace [59] and AIO [60] employ a multi-task learning framework to simultaneously handle a predefined set of face analysis tasks, eliminating the repetitiveness in model training. However, due to the empirically determined output structures for each task, these early all-in-one models are unable to address new-coming tasks, resulting in a lack of task extensibility. Furthermore, these early models lack robust pre-training and are now considered to have performed inadequately.

Our work aims to explore a face generalist model, which is initialized with a task-agnostic backbone (unified representation) and can handle any user-chosen set of face analysis tasks with a multi-task learning framework (unified training). To achieve improved task extensibility and increased application efficiency, we focus on the unified model structure. Two ideas are presented as follows:

(1) Shared structural designs: dealing with new-coming tasks using standardized output heads. We have observed significant variations in the expected outputs of different face analysis tasks in terms of shape and granularity. Based on these observations, we categorize all face analysis tasks into three distinct categories: dense prediction, attribute prediction, and identity prediction. An intuitive model design can consist of a backbone and three types of standardized output heads, each dedicated to a specific task category, as illustrated in Fig. 1c, referred to as **Naive Faceptor**. All tasks share a common backbone, enabling the proposed model to achieve higher application efficiency than the unified representation approaches. Tasks within the same category will share structural designs, thus avoiding the need to design new output structures based on experience for new-coming tasks, and ensuring the extensibility of the model. However, a notable limitation of this design is the lack of parameter sharing among heads across tasks. This results in a linear growth of the number of heads as the tasks increase, leading to significant storage overhead.

(2) Shared parameters: dealing with new-coming semantics using task-specific queries. To further enhance the unification of model structure while maintaining the model's performance on individual tasks, we propose **Faceptor**, which adopts a single-encoder dual-decoder architecture, as shown in Fig. 1d. The transformer encoder extracts shared features while the transformer decoder attends to particular semantic information. Additionally, the pixel decoder is used for restoring the image spatial scale for dense prediction tasks. Inspired by previous works [8, 10, 11, 76, 81], we introduce task-specific queries from single-task methods into our unified structure to model the semantics of different tasks, minimizing the use of non-shared parameters and achieving a significantly higher storage efficiency. We also introduce the Layer-Attention mechanism in the transformer decoder to model the preferences of different tasks towards features from different layers. With layer-aware embeddings in the transformer decoder, Faceptor can adaptively assign weights for the features from different layers.

Multi-task learning aims to achieve optimal performance across all tasks, while auxiliary supervised learning leverages some tasks to enhance the performance of others. In our training framework, auxiliary supervised learning can be performed by adjusting the weights and batch sizes of involved tasks. Experiments indicate that harnessing landmark localization, face parsing and face recognition tasks can significantly enhance the performance of tasks such as age estimation and expression recognition, which suffer from limited available data.

Our contributions can be summarized as follows:

1. To the best of our knowledge, our work is the first to explore a face generalist model, with unified representation, training, and model structure. Our main focus is on the development of unified model structures.

2. With one shared backbone and three types of standardized output heads, **Naive Faceptor** achieves improved task extensibility and increased application efficiency.
3. With task-specific queries to deal with new-coming semantics, **Faceptor** further enhances the unification of model structure and employs significantly fewer parameters than Naive Faceptor.
4. The proposed Faceptor demonstrates outstanding performance under both multi-task learning and auxiliary supervised learning settings.

2 Related Works

Universal Facial Representation: FRL [5] and FaRL [93] address face analysis tasks by following a pipeline that involves (1) collecting a large-scale facial dataset, (2) pre-training a task-agnostic network to achieve universal facial representation learning, and (3) fine-tuning the network for specific facial tasks in the user-chosen set. FaRL [93] combines natural language supervised and self-supervised learning, extracting high-level semantic meaning from image-text pairs using contrastive loss [27, 57, 58], while also exploring low-level information through masked image modeling [2, 53]. Robust pre-training is crucial for face generalist models. In our experiments, we utilize the ViT [17] model pre-trained with the FaRL framework as the initialization for the transformer encoder.

Multi-task Learning for Face Perception: HyperFace [59] and AIO [60] are early classic works of multi-task learning, employing CNN as the backbone and leveraging experiential knowledge to determine the appropriate layer of features for different tasks. However, since these models are designed for predefined task sets, they are not able to deal with new-coming tasks. In contrast, SwinFace [56] adopts standardized subnets for task extensibility, with face analysis and recognition subnets handling attribute and identity prediction tasks respectively. In our experiments, the Naive Faceptor is primarily inspired by SwinFace but includes an additional subnet [82] to handle dense prediction tasks.

Transformer Encoder-Decoder Architecture for Computer Vision: The success of DETR [8] in object detection has motivated researchers to investigate the utilization of transformer encoder-decoder architecture in computer vision tasks. MaskFormer [11] presents a unified approach to tackle semantic and instance-level segmentation tasks through the introduction of a single-encoder dual-decoder structure. In MaskFormer, each segment is represented by a query. In SLPT [81] and RLPFER [76], individual facial landmarks or expressions are considered distinct semantic information and are represented as task-specific queries. To the best of our knowledge, there is no existing work in the field of face perception that comprehensively unifies all face analysis tasks and employs task-specific queries to represent diverse semantic information.

3 Method

In this section, we first offer a brief introduction to the structure of Naive Faceptor. Next, we provide the details of the Faceptor design, highlighting the Layer-

Attention mechanism. Then, we present the training framework and discuss the objective functions. Lastly, we provide a comprehensive comparison between our proposed face generalist models and previous efforts for face perception.

3.1 Naive Faceptor

We briefly describe the structure of Naive Faceptor. For a fair comparison, the backbone of Naive Faceptor and the encoder of Faceptor utilize the same transformer encoder architecture, initialized by the FaRL [93] framework. Details regarding the transformer encoder will be provided in Sec. 3.2. We employ standardized face analysis and face recognition subnets from SwinFace [56] as attribute prediction head and identity prediction head, respectively. In addition, we follow the implementation in the FaRL experiment, utilizing UperNet [82] as the dense prediction head to produce dense output. We provide an illustration of Naive Faceptor in the appendix, offering more details.

3.2 Faceptor

Faceptor adopts a single-encoder dual-decoder architecture, as shown in Fig. 2.

Transformer Encoder: We utilize a 12-layer ViT-B [17] as the transformer encoder, which is pre-trained with FaRL [93] framework. When an image $\mathbf{X}$ of size $H \times W$ is given as input, the encoder produces a feature $\mathbf{F}^l \in \mathbb{R}^{C_{en} \times \frac{H}{S} \times \frac{W}{S}}$ at the l-th layer. Here, C_{en} represents the number of channels, and S represents the stride of patch projection, which are 768 and 16 respectively. To handle input images of varying resolutions (512×512 for dense prediction tasks, and 112×112 for attribute and identity prediction tasks), we employ a shared learnable positional embedding $\mathbf{E}_{en-pos}$ with a size of 32×32, and interpolate it based on the spatial size of the input image after patch projection. We retain the features obtained from all 12 layers of the encoder for future use. Therefore, the encoded feature F can be formulated as:

$$\mathbf{F} = \mathrm{TransformerEncoder}(\mathbf{X}, \mathbf{E}_{en\_pos}) \in \mathbb{R}^{12 \times C_{en} \times \frac{H}{S} \times \frac{W}{S}}, \tag{1}$$

where $\mathbf{F} = [\mathbf{F}^1; \mathbf{F}^2; \cdots; \mathbf{F}^{12}]$.

Transformer Decoder: We employ a 9-layer standard transformer decoder [71] to compute the task-specific tokens based on the encoded features and task-specific queries. To begin, we define task-specific queries, which are applicable to dense prediction and attribute prediction tasks. The task queries for task t are denoted as:

$$\mathbf{Q}_t = [\mathbf{q}_{t,1}, \mathbf{q}_{t,2}, \mathbf{q}_{t,3}, ..., \mathbf{q}_{t,N_t}], \tag{2}$$

where N_t represents the number of queries that convey different semantic meanings in task t. A landmark, a semantic parsing class, and a binary attribute are each represented by one query for facial landmark localization, face parsing, and binary attribute classification respectively. 101 queries represent ages 0-100 for

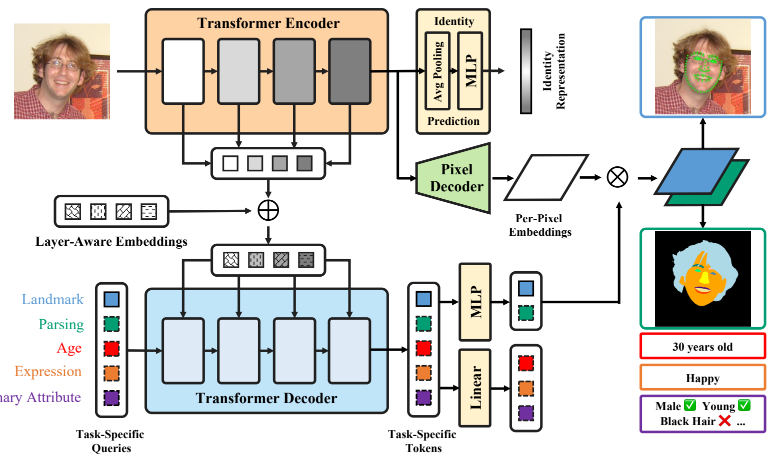


Fig. 2: Overall architecture for the proposed Faceptor

age estimation. 7 queries represent expressions (surprise, fear, disgust, happiness, sadness, anger, neutral) for expression recognition. Following established conventions [10,71], all task-specific queries $\mathbf{Q}_t$ are accompanied by a positional embedding $\mathbf{E}_{de\_pos,t}$, which has the same dimension as $\mathbf{Q}_t$ and is not shared across tasks.

Typically, when using the transformer decoder in visual tasks, only the encoded feature from the top layer, denoted as $\mathbf{F}^{top}$, is utilized for computation. However, the features obtained from the encoder contain decreasing geometric information and increasing semantic information from the bottom to the top layers. Different tasks have varying preferences for features from different layers. To enable the transformer decoder to leverage features from multiple layers, we uniformly extract six layers of features from $\mathbf{F}$ and project them into the dimension of the decoder tokens, denoted as C_{de} and set to 256, resulting in:

$$\hat{\mathbf{F}} = \text{Projection}([\mathbf{F}^2; \mathbf{F}^4; \mathbf{F}^6; \mathbf{F}^8; \mathbf{F}^{10}; \mathbf{F}^{12}]) \in \mathbb{R}^{6 \times C_{de} \times \frac{H}{S} \times \frac{W}{S}}. \tag{3}$$

After processing with the transformer decoder, task-specific tokens for dense prediction or attribute prediction task t are obtained:

$$\mathbf{T}_t = \text{TransformerDecoder}(\hat{\mathbf{F}}, \mathbf{Q}_t, \mathbf{L}_t, \mathbf{P}, \mathbf{E}_{de\_pos,t}) \in \mathbb{R}^{N_t \times C_{de}}, \tag{4}$$

where $\mathbf{L}_t$ and $\mathbf{P}$ are the layer-aware embedding and positional embedding associated with $\hat{\mathbf{F}}$, respectively. Further details are provided in Sec. 3.3.

Pixel Decoder: The pixel decoder is used to gradually upsample the features in order to produce per-pixel embeddings:

$$\mathbf{E}_{pixel} = \text{PixelDecoder}(\mathbf{F}) \in \mathbb{R}^{C_{de} \times \frac{H}{s} \times \frac{W}{s}}, \tag{5}$$

where s is set to 4 in our implementation. It should be noted that any per-pixel classification-based segmentation model can be employed as a pixel decoder. In our implementation, we extract the feature $\mathbf{F}^{12}$ from the top layer of the encoder, and then pass it through two consecutive 2×2 deconvolutional layers to obtain the per-pixel embedding $\mathbf{E}_{pixel}$. Experimental results have demonstrated that this simple pixel decoder has been capable of achieving excellent performance in facial landmark localization and face parsing.

Outputs: Similar to Naive Faceptor, Faceptor also includes specifically designed output modules for three categories of tasks. For the dense prediction tasks, the task-specific tokens need to be passed through a shared MLP to align with the per-pixel embeddings outputted by the pixel decoder. The dot product of these two is then linearly interpolated to obtain the final dense prediction output $\mathbf{y}_{map} \in \mathbb{R}^{N_t \times H \times W}$. For the attribute prediction tasks, the task-specific tokens produced by the decoder can directly go through a shared linear layer to obtain the final prediction result $\mathbf{y}_{value} \in \mathbb{R}^{N_t}$. For the identity prediction task, the features from the top layer of the transformer, denoted as $\mathbf{F}^{12}$, are first passed through an average pooling layer to obtain a vector. Then, following the implementation of SwinFace [56], the vector is processed by an FC-BN-FC-BN structure to obtain the final identity representation $\mathbf{y}_{vector} \in \mathbb{R}^{d}$, where d is set to 512. It is important to note that in Faceptor, all parameters of output modules are shared among multiple tasks of the same category, whereas in Naive Faceptor, tasks of the same category share only the structural design of output modules without sharing parameters.

3.3 Layer-Attention Mechanism

In the transformer decoder, cross-attention can be represented as:

$$\mathrm{CrossAttention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathrm{Softmax}(\mathbf{Q}\mathbf{K}^T/\sqrt{d})\mathbf{V}. \tag{6}$$

For the l-th layer, the query is $\mathbf{Q} = \mathbf{H}_t^{l-1} + \mathbf{E}_{de\_pos,t}$, where $\mathbf{H}_t^{l-1}$ is the output of the previous layer of the decoder and $\mathbf{H}_t^0 = \mathbf{Q}_t$. The value is $\mathbf{V} = \hat{\mathbf{F}}$. We implement Layer-Attention by introducing layer-aware embeddings $\mathbf{L}_t \in \mathbb{R}^{6 \times C_{de}}$ for task t into the key, obtaining:

$$\mathbf{K} = \hat{\mathbf{F}} + \mathrm{Repeat}(\mathbf{L}_t) + \mathrm{Repeat}(\mathbf{P}), \tag{7}$$

where $\mathbf{P} \in \mathbb{R}^{C_{de} \times \frac{H}{S} \times \frac{W}{S}}$ is the learnable positional embeddings randomly initialized, and the Repeat function extends the input features in a repeated manner to a scale of $\mathbb{R}^{6 \times C_{de} \times \frac{H}{S} \times \frac{W}{S}}$.

For simplification, we use $\hat{\mathbf{L}}_t$ and $\hat{\mathbf{P}}$ to represent $\mathrm{Repeat}(\mathbf{L}_t)$ and $\mathrm{Repeat}(\mathbf{P})$ respectively. In Eq. (6), $\mathbf{Q}\mathbf{K}^T$ can be expanded as $\mathbf{Q}\hat{\mathbf{F}}^T + \mathbf{Q}\hat{\mathbf{L}}_t^T + \mathbf{Q}\hat{\mathbf{P}}^T$. The term $\mathbf{Q}\hat{\mathbf{P}}^T$ reflects the model's preference for features at different positions, typically taken into account by existing models. In contrast, $\mathbf{Q}\hat{\mathbf{L}}_t^T$ represents the model's preference for features from different layers, which has often been neglected in previous research.

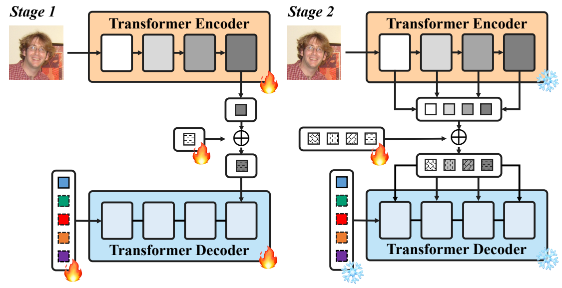


Fig. 3: Two-stage training process to ensure the effectiveness of Layer-Attention mechanism.

In practice, we found that directly introducing Layer-Attention can not improve the model's performance on various tasks, and even result in significant deterioration in the age estimation task. We believe that this is because both $\mathbf{Q}_t$ and $\mathbf{E}_{de\_pos}$ are randomly initialized, which causes, at the beginning of training, $\mathbf{Q}_t$ to be unable to represent semantic information and $\mathbf{Q}\hat{\mathbf{L}}_t^T$ to be inadequate in reflecting task t's preference for features from different layers. To address this issue, we introduce a two-stage training process, as shown in Fig. 3. In the first stage, only the features from the top layer, namely, Projection($\mathbf{F}^{12}$), are used for training to enable $\mathbf{Q}_t$ to learn the semantic representation of task t. In the second stage, the transformer decoder is allowed to access $\hat{\mathbf{F}}$, and most of the model parameters are frozen except for $\mathbf{L}_t$, which is allowed to be learned. It should be noted that since $\mathbf{L}_t$ is not shared across tasks, if there is no performance improvement on task t after the second stage of training, the Layer-Attention mechanism can be excluded during inference for task t. Experimental results show that attribute prediction tasks such as age estimation, expression recognition, and binary attribute classification can benefit from the introduction of the Layer-Attention mechanism.

3.4 Objective Functions

We employ a multi-task learning framework to enable the model to simultaneously tackle a variety of face analysis tasks. The overall objective function is:

$$L_{all} = \frac{\sum_{t \in T} \alpha_t \frac{1}{n_t} \sum_{i=1}^{n_t} L(\mathbf{y}_{t,i})}{\sum_{t \in T} \alpha_t}, \tag{8}$$

where T represents the user-chosen task set, α_t is the weight of task t, n_i is the number of samples for task t in each training batch, $\mathbf{y}_{t,i}$ is the output of Faceptor for the i-th sample in task t, and $L(\mathbf{y}_{t,i})$ is the loss function for single sample. Auxiliary supervised learning can be performed by adjusting the α_t and n_i. Please refer to the appendix for the specific loss function used for each individual task.

3.5 Comparison of Task Extensibility and Application Efficiency

Table 1 presents a semi-quantitative comprehensive comparison between our proposed models and previous unified approaches in task extensibility and application efficiency. Assuming there are N tasks in the user-chosen set. It is

Table 1: Semi-quantitative comparison of task extensibility and application efficiency. $\mathcal{B}$ represents backbones, $\mathcal{O}$ represents output modules, and $\mathcal{Q}$ represents queries in the transformer decoder.

Paradigms or Models	Focus for Unified Face Perception	Extensible?	Application Efficiency		
			Training Cycles	Inference Calculation	Storage Parameter
Universal Representation + Finetuning	Representation	Yes	N	$N\mathcal{B}+N\mathcal{O}$	$N\mathcal{B}+N\mathcal{O}$
Early All-In-One Model	Training	No	1	$1\mathcal{B}+N\mathcal{O}$	$1\mathcal{B}+N\mathcal{O}$
Our Naive Faceptor	Model Structure	Yes	1	$1\mathcal{B}+N\mathcal{O}$	$1\mathcal{B}+N\mathcal{O}$
Our Faceptor	Model Structure	Yes	1	$1\mathcal{B}+N\mathcal{O}$	$1\mathcal{B}+1\mathcal{O}+N\mathcal{Q}$, $\mathcal{Q}\ll\mathcal{O}$

noticed that the number of parameters in the queries is much less than that in the output modules. As N increases, the number of parameters in Faceptor will be significantly less than that in Naive Faceptor. To sum up, our Faceptor can achieve improved task extensibility and the highest application efficiency.

4 Experiments

4.1 Implementation Details

Datasets: To validate the effectiveness of our proposed generalist models, we have collected 13 training datasets covering 6 tasks within 3 categories. In our experiments, Naive Faceptor and the base version of Faceptor (referred to as Faceptor-Base) are trained with only the 7 datasets highlighted in bold in Tab. 2. To explore the performance ceiling of Faceptor, we further train Faceptor-Full using all 13 datasets. Table 2 presents the number of training samples in each dataset after preprocessing. For dense prediction, we apply the data augmentation methods used in the FaRL [93]'s downstream experiment. For attribute prediction, we employ horizontal flip, Randaugment [12], and Random Erasing [84]. For identity prediction, we use only horizontal flip for data augmentation. It is worth noting that we do not perform uniform alignment for training samples used but still achieve excellent performance. Please refer to the appendix for more details of the datasets.

Training for Faceptor: For the first stage, we employ an AdamW [46] optimizer for 50,000 steps, using a cosine decay learning rate scheduler and 2000 steps of linear warm-up. The base learning rate for the Transformer Encoder is 5.0×10^{-5}, and the learning rate for the remaining parts is 10 times that of the Transformer Encoder. A weight decay of 0.05 is used. For the second stage, only 20000 steps are required, with 2000 steps reserved for linear warm-up. All parameters except for layer-aware embeddings are frozen. The other hyper-parameters remain consistent with the first stage. Due to the small number of parameters being trained, the second stage can be completed quickly. Table 2 presents the batch size and weight used for each dataset during the training of Faceptor-Base and Faceptor-Full. All training is conducted on 4 NVIDIA Tesla V100 GPUs.

Table 2: The face analysis tasks included in our experiment and the corresponding datasets used

Task Category	Task	Datasets for Training	Number of Samples	Faceptor-Base n_t	Faceptor-Base α_t	Faceptor-Full n_t	Faceptor-Full α_t
Dense Prediction	Landmark Localization	**300W** [63]	3, 148	4	1000.00	4	250.00
		WFLW [79]	7, 500	-	-	4	250.00
		COFW [6]	1, 345	-	-	4	250.00
		AFLW-19 [94]	20, 000	-	-	4	250.00
	Face Parsing	**CelebAMask-HQ** [33]	27, 176	4	100.00	4	100.00
		LaPa [44]	20, 168	-	-	4	100.00
Attribute Prediction	Age Estimation	**MORPH II** [29]	44, 194	64	6.00	64	4.00
		UTKFace [88]	13, 144	-	-	16	1.00
	Expression Recognition	**AffectNet** [50]	282, 829	64	4.00	64	6.66
		RAF-DB [37]	12, 271	16	1.00	16	1.67
		FERPlus [3]	28, 127	-	-	16	1.67
	Binary Attribute Classification	**CelebA** [45]	182, 637	64	2.00	64	2.00
Identity Prediction	Face Recognition	**MS1MV3** [22]	5, 179, 510	256	5.00	256	5.00

Training for Naive Faceptor: During the training of the Naive Faceptor, we have observed that this structure is not sensitive to the weight changes of the tasks. Therefore, the weights for all tasks are set to 1.0. Other settings are kept consistent with the first stage of training the Faceptor-Base.

4.2 Comparison Between Naive Faceptor and Faceptor

Table 3 presents a comparison between Naive Faceptor and Faceptor-Base in terms of parameters and performance. Overall, Faceptor-Base demonstrates similar performance to Naive Faceptor while utilizing significantly fewer parameters. Specifically, Faceptor exhibits slight enhancements in facial landmark localization, face parsing, age estimation, and binary attribute estimation tasks, along with a notable improvement in expression recognition by 2.80%. Only for face recognition, Faceptor indicates a slight decrease. Faceptor consists of a total of 103.2M parameters, distributed as follows: 86.8M for the transformer encoder, 14.7M for the transformer decoder, 0.5M for the pixel decoder, and 1.2M for the remaining components. In Naive Faceptor, the standardized output heads for dense, attribute, and identity prediction tasks respectively contain approximately 39.3M, 3.4M, and 1.0M parameters. Consequently, Naive Faceptor encompasses a total of 178.9M parameters for the six tasks, which is 73% more than Faceptor. As the number of tasks increases, this parameter difference between the two models will become even more pronounced. The experimental results indicate that Faceptor, with higher storage efficiency and comparable performance with the naive counterpart, should be favored as a unified model structure. For this reason, we conduct larger-scale experiments in Sec. 4.4 to compare the performance of our Faceptor with specialized models.

It is worth noting that we have omitted the performance comparison of our proposed models with early all-in-one models [59, 60], as those early models utilized significantly simpler testing protocols that are now rarely referenced, and

Table 3: Comparison between Naive Faceptor and Faceptor-Base

Methods	Landmark 300W Comm.	Chal.	Full	Parsing CelebAMask-HQ	Age MORPH II	Expression RAF-DB	Attribute CelebA
	$\text{NME}_{\text{inter-ocular}} \downarrow$			F1-mean ↑	MAE ↓	Acc ↑	mAcc ↑
Naive	2.75	4.84	3.16	88.04	1.873	87.58	91.40
Faceptor	**2.60**	**4.60**	**3.00**	**88.22**	**1.869**	**90.38**	**91.43**

Methods	Face Recognition 1:1 Verification Accuracy ↑ LFW	CFP-FP	AgeDB-30	CALFW	CPLFW	Mean	Params (M)
Naive	99.50	**96.17**	**94.35**	**95.13**	**92.68**	**95.57**	178.9
Faceptor	**99.52**	95.86	93.33	94.70	92.12	95.10	**103.2**

Table 4: Comparison under three settings. LA stands for Layer-Attention.

Settings	Age MORPH II MAE ↓	Expression RAF-DB Acc ↑	Attribute CelebA mAcc ↑
w/o LA	1.882	89.80	91.40
LA (Directly)	1.970	90.03	91.40
LA (Two-stage)	**1.869**	**90.38**	**91.43**

their task sets are also smaller. Given that our generalist models perform well on more challenging and diverse testing protocols, it is evident that our models surpass the early all-in-one models. The appendix provides further discussion on the performance of early models.

4.3 Layer-Attention Mechanism

Table 4 presents the performance of Faceptor-Base on age estimation, expression recognition, and binary attribute classification tasks under three settings: without using the Layer-Attention mechanism, using the Layer-Attention mechanism directly, and using the Layer-Attention mechanism with two-stage training process. It can be observed that when using the Layer-Attention mechanism directly, Faceptor does not always achieve improved performance and even exhibits significant degradation in age estimation. However, employing two-stage training generally leads to performance improvement, especially in expression recognition, where a 0.58% improvement is achieved on RAF-DB [37].

4.4 Comprehensive Performance Evaluation for Faceptor

To explore the upper limit of Faceptor's performance, we have trained Faceptor-Full using 13 training datasets. Tables 5 to 7 present the performance of Faceptor-Full in various tasks. In most tasks, Faceptor-Full achieves comparable or superior performance to state-of-the-art specialized models, except face recognition where it slightly lags behind the state-of-the-art method. A detailed analysis of the performance is presented below.

Dense Prediction. Thanks to the masked image modeling [2, 53] incorporated into the FaRL framework [93], our model achieves outstanding performance in dense prediction tasks. Faceptor-Full outperforms existing methods on all facial landmark localization and face parsing datasets except for LaPa,

Table 5: Comparison with other specialized models for dense prediction tasks

Methods	WFLW Full	300W Comm.	300W Chal.	300W Full	COFW -	AFLW-19 Full	Methods	CelebA Mask-HQ	LaPa
	NME$_{\text{inter-ocular}}$ ↓					NME$_{\text{diag}}$ ↓		F1-mean ↑	
DAN-Menpo [30]	-	3.44	4.88	3.09	-	-	Lee et al. [89]	80.3	-
SAN [16]	-	3.34	6.60	3.98	-	1.91	BASS [44]	-	89.8
LAB [80]	5.27	2.98	5.19	3.49	3.92	1.85	EHANet [47]	84.0	89.2
Wing [19]	5.11	3.27	7.18	4.04	-	1.65	Wei et al. [78]	82.1	89.4
DeCaFA [13]	4.62	2.93	5.26	3.39	-	-	EAGR [70]	85.1	91.1
AWing [77]	4.36	2.72	4.52	3.07	-	-	AGRNet [69]	85.5	92.3
AVS [55]+SAN [16]	4.39	3.21	6.49	3.86	-	-	DML-CSR [90]	86.1	**92.4**
HRNet [73]	4.60	2.91	5.11	3.34	3.45	1.57			
DAG [39]	4.21	2.62	4.77	3.04	-	-			
LUVLi [31]	4.37	2.76	5.16	3.23	-	1.39			
ADNet [26]	4.14	2.53	4.58	2.93	-	-			
PIPNet [28]	4.31	2.78	4.89	3.19	3.08	1.42			
SLPT [81]	4.14	2.75	4.90	3.17	3.32	-			
DTLD+ [36]	4.05	2.60	4.48	2.96	3.02	1.37			
Faceptor	**4.03**	**2.52**	**4.25**	**2.86**	**3.01**	**0.95**	**Faceptor**	**88.2**	91.5

Table 6: Comparison with other specialized models for attribute prediction tasks

Methods	Age MORPH II	Age UTKFace	Methods	Expression RAF-DB	Expression FERPlus	Methods	Attribute CelebA
	MAE ↓			Acc ↑			mAcc ↑
OR-CNN [52]	3.27	5.74	DLP-CNN [37]	80.89	-	PANDA-1 [85]	85.43
DEX [61]	2.68	-	gACNN [40]	85.07	-	LNets+ANet [45]	87.33
DLDL [21]	2.42	-	IPA2LT [83]	86.77	-	MOON [62]	90.94
DLDLF [67]	2.24	-	RAN [75]	86.90	88.55	NSA [48]	90.61
DRFs [66]	2.17	-	CovPool [1]	87.00	-	MCNN-AUX [24]	91.29
MV [54]	2.16	-	SCN [74]	87.03	89.35	MCFA [95]	91.23
Axel Berg et al. [4]	-	4.55	DACL [18]	87.78	-	DMM-CNN [49]	**91.70**
CORAL [7]	-	5.47	KTN [35]	88.07	**90.49**	SwinFace [56]	91.32
Gustafsson et al. [23]	-	4.65	DMUE [65]	88.76	88.64		
BridgeNet [38]	2.38	-	RUL [86]	88.98	88.75		
OL [41]	2.22	-	EAC [87]	88.99	89.64		
DRC-ORID [34]	2.16	-	SwinFace [56]	90.97	-		
PML [15]	2.15	-					
DLDL-v2 [20]	1.97	4.42					
MWR [68]	2.00	4.37					
Faceptor	**1.96**	**4.10**	**Faceptor**	**91.26**	90.40	**Faceptor**	91.39

as shown in Tab. 5. However, for LaPa, our model's performance declines due to the introduction of Tanh-warping [42] to balance segmentation performance between the inner facial components and hair region. We conduct experiments using Faceptor-Base for transfer learning on the LaPa dataset, achieving a mean F1 score of 92.7, as shown in Tab. 9. This score is higher than that of the state-of-the-art specialized methods, demonstrating our model's strong understanding of dense prediction tasks.

Attribute Prediction. Faceptor-Full achieves state-of-the-art results in age estimation and expression recognition with 1.96 and 4.10 MAE on MORPH II [29] and UTKFace [88] respectively, and 91.26% accuracy on RAF-DB [37], while it performs on par with the state-of-the-art on binary attribute classification. The training samples used for age estimation and expression recognition are insufficient relative to the complexity of these tasks. During joint training, these tasks can benefit from the initialization of universal representation and multi-task learning, obtaining improved performances. In contrast, for the binary attribute classification task, the availability of ample data from CelebA [45] with around 183K training samples has led to saturated performance across existing methods.

Table 7: Comparison for face recognition. The 1:1 verification accuracies on the LFW [25], CFP-FP [64], AgeDB-30 [51], CALFW [92] and CPLFW [91] are provided.

Methods	Face Verification Accuracy					
	LFW	CFP-FP	AgeDB-30	CALFW	CPLFW	Mean
ViT [17]+CosFace [72]	99.83	96.19	97.82	95.92	92.55	96.46
FaRL [93]+CosFace [72]	99.60	96.70	95.55	95.43	92.38	95.93
Faceptor	99.40	96.34	93.65	94.75	92.27	95.28

Identity Prediction. The performances of specialized models trained using the MS-Celeb-1M [22] dataset and the CosFace [72] loss function starting from randomly initialized ViT-B [17] and FaRL pretraining are presented in Tab. 7, allowing a fair comparison to Faceptor-Full. Evaluation results on several face verification test datasets indicate that Faceptor-Full performs lower than ViT trained from scratch. This performance decline can be attributed to two main reasons. Firstly, Faceptor-Full is initialized from FaRL, which provides facial representations combining high-level and low-level information not specifically tailored for the face recognition task. The inferior performances of specialized models starting from FaRL pre-training compared to those trained from scratch validate this point. Secondly, Faceptor-Full involves tasks that inherently have conflicting objectives. While face recognition requires the model to learn to extract identity representations ignoring variations in facial texture and movements, face dense and attribute prediction tasks demand the opposite. Despite the slight decline in face recognition, Faceptor-Full achieves or surpasses state-of-the-art results in all other tasks, underscoring the significant potential of the proposed face generalist model with a highly unified model structure.

4.5 Auxiliary Supervised Learning

The performance improvement of certain attribute prediction tasks is limited due to insufficient data, with age estimation and expression recognition being two typical tasks. In our experiment, we consider these two tasks as the main tasks and introduce auxiliary tasks such as facial landmark localization, face parsing, and face recognition to provide additional supervised signals. Our results (as shown in Tab. 8) show that Faceptor with auxiliary supervised learning outperforms the same model which is under single-task or multi-task learning settings. Moreover, our model achieves significant improvements over the state-of-the-art in age and expression tasks, with an MAE of 1.787 on MORPH II [29], reducing by 0.183, and an accuracy of 91.92% on RAF-DB [37], increasing by 0.95%. This indicates that our proposed method can effectively enhance data efficiency by leveraging rich supervised signals from auxiliary tasks, thus enabling better performance for main tasks with insufficient data. For more experimental details on auxiliary supervised learning, please refer to the appendix.

Table 8: Comparison for auxiliary supervised learning. STL is short for Single-Task Learning. MTL is short for Multi-Task Learning. ASL is short for Auxiliary Supervised Learning.

Methods	Age MORPH II MAE ↓	Expression RAF-DB Acc ↑
SOTA (STL)	1.970 [20]	90.97 [56]
Naive Faceptor (STL)	2.070	91.33
Faceptor (STL)	2.238	91.10
Faceptor (MTL)	1.869	90.38
Faceptor (ASL)	**1.787**	**91.92**

Table 9: Cross-datasets transfer performances under different settings. EM is short for Early Methods. PT is short for Prompt Tuning. DFT is short for Decoder Finetuning. FPFT is short for Full-Parameter Finetuning.

Settings	Landmark AFLW-19 [94] $\mathrm{NME_{diag}}$ ↓	Parsing LaPa [44] F1-mean ↑	Attribute LFW-73 [32] mAcc ↑
EM	1.91 [16]	89.8 [44]	-
PT	1.89	84.0	85.56
DFT	1.06	89.9	87.81
FPFT	0.89	92.7	87.95

4.6 Cross-Datasets Transfer

We aim to explore the performance of Faceptor in cross-dataset transfer scenarios where subtle semantic variations exist in certain tasks, as shown in Tab. 9. We have observed that facial landmark localization datasets encompass different landmarks, face parsing datasets involve varying semantic parsing classes, and binary attribute classification datasets have different attribute labels. Starting from Faceptor-Base, we try to transfer its capabilities to unseen datasets with novel semantics. By considering the diverse trainable parameters, we investigate three settings: training only task-specific queries (prompt tuning), training only the decoders and other output structures (output module fine-tuning), and training all parameters (full-parameter fine-tuning). The experiments reveal that in facial landmark localization, prompt tuning results even outperform the early method [16]. In face parsing, the results of prompt tuning can approach the performance of the early method [44]. In binary attribute classification, prompt tuning can achieve performance close to that of full-parameter fine-tuning. These experimental findings demonstrate the potential of prompt tuning for Faceptor. For more experimental details, please refer to the appendix.

5 Conclusion

To the best of our knowledge, this is the first work that explores face generalist models. Naive Faceptor consists of one shared backbone and 3 types of standardized output heads, obtaining improved task extensibility and increased application efficiency. Compared to Naive Faceptor, Faceptor is more unified in structure and offers higher storage efficiency with a single-encoder dual-decoder architecture and task-specific queries for semantics. We demonstrate the effectiveness of the proposed models on a task set including 6 tasks, achieving excellent performance. In particular, we introduce a Layer-Attention mechanism that models the preferences of different tasks towards features from different layers, thereby enhancing performance further. The two-stage training process ensures the effectiveness of the Layer-Attention mechanism. Additionally, our training framework can also perform auxiliary supervised learning to improve performance on attribute prediction tasks with insufficient data.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant No.62076036, No.62076031, No.62236003, and No.62306043.

References

1. Acharya, D., Huang, Z., Paudel, D.P., Gool, L.V.: Covariance pooling for facial expression recognition. In: CVPR Workshops. pp. 367–374. Computer Vision Foundation / IEEE Computer Society (2018). https://doi.org/10.1109/CVPRW.2018.00077
2. Bao, H., Dong, L., Piao, S., Wei, F.: Beit: BERT pre-training of image transformers. In: ICLR. OpenReview.net (2022)
3. Barsoum, E., Zhang, C., Canton-Ferrer, C., Zhang, Z.: Training deep networks for facial expression recognition with crowd-sourced label distribution. In: ICMI. pp. 279–283. ACM (2016). https://doi.org/10.1145/2993148.2993165
4. Berg, A., Oskarsson, M., O'Connor, M.: Deep ordinal regression with label diversity. In: ICPR. pp. 2740–2747. IEEE (2020). https://doi.org/10.1109/ICPR48806.2021.9412608
5. Bulat, A., Cheng, S., Yang, J., Garbett, A., Sánchez-Lozano, E., Tzimiropoulos, G.: Pre-training strategies and datasets for facial representation learning. In: ECCV (13). Lecture Notes in Computer Science, vol. 13673, pp. 107–125. Springer (2022). https://doi.org/10.1007/978-3-031-19778-9_7
6. Burgos-Artizzu, X.P., Perona, P., Dollár, P.: Robust face landmark estimation under occlusion. In: ICCV. pp. 1513–1520. IEEE Computer Society (2013). https://doi.org/10.1109/ICCV.2013.191
7. Cao, W., Mirjalili, V., Raschka, S.: Rank consistent ordinal regression for neural networks with application to age estimation. Pattern Recognit. Lett. **140**, 325–331 (2020). https://doi.org/10.1016/J.PATREC.2020.11.008
8. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: ECCV (1). Lecture Notes in Computer Science, vol. 12346, pp. 213–229. Springer (2020). https://doi.org/10.1007/978-3-030-58452-8_13
9. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. In: NeurIPS (2020)
10. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: CVPR. pp. 1280–1289. IEEE (2022). https://doi.org/10.1109/CVPR52688.2022.00135
11. Cheng, B., Schwing, A.G., Kirillov, A.: Per-pixel classification is not all you need for semantic segmentation. In: NeurIPS. pp. 17864–17875 (2021)
12. Cubuk, E.D., Zoph, B., Shlens, J., Le, Q.V.: Randaugment: Practical automated data augmentation with a reduced search space. In: CVPR Workshops. pp. 3008–3017. Computer Vision Foundation / IEEE (2020). https://doi.org/10.1109/CVPRW50498.2020.00359
13. Dapogny, A., Cord, M., Bailly, K.: Decafa: Deep convolutional cascade for face alignment in the wild. In: ICCV. pp. 6892–6900. IEEE (2019). https://doi.org/10.1109/ICCV.2019.00699
14. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: CVPR. pp. 4690–4699. Computer Vision Foundation / IEEE (2019). https://doi.org/10.1109/CVPR.2019.00482

15. Deng, Z., Liu, H., Wang, Y., Wang, C., Yu, Z., Sun, X.: PML: progressive margin loss for long-tailed age classification. In: CVPR. pp. 10503–10512. Computer Vision Foundation / IEEE (2021). https://doi.org/10.1109/CVPR46437.2021.01036
16. Dong, X., Yan, Y., Ouyang, W., Yang, Y.: Style aggregated network for facial landmark detection. In: CVPR. pp. 379–388. Computer Vision Foundation / IEEE Computer Society (2018). https://doi.org/10.1109/CVPR.2018.00047
17. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR. OpenReview.net (2021)
18. Farzaneh, A.H., Qi, X.: Facial expression recognition in the wild via deep attentive center loss. In: WACV. pp. 2401–2410. IEEE (2021). https://doi.org/10.1109/WACV48630.2021.00245
19. Feng, Z., Kittler, J., Awais, M., Huber, P., Wu, X.: Wing loss for robust facial landmark localisation with convolutional neural networks. In: CVPR. pp. 2235–2245. Computer Vision Foundation / IEEE Computer Society (2018). https://doi.org/10.1109/CVPR.2018.00238
20. Gao, B., Liu, X., Zhou, H., Wu, J., Geng, X.: Learning expectation of label distribution for facial age and attractiveness estimation. CoRR **abs/2007.01771** (2020)
21. Gao, B., Xing, C., Xie, C., Wu, J., Geng, X.: Deep label distribution learning with label ambiguity. IEEE Trans. Image Process. **26**(6), 2825–2838 (2017). https://doi.org/10.1109/TIP.2017.2689998
22. Guo, Y., Zhang, L., Hu, Y., He, X., Gao, J.: Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In: ECCV (3). Lecture Notes in Computer Science, vol. 9907, pp. 87–102. Springer (2016). https://doi.org/10.1007/978-3-319-46487-9_6
23. Gustafsson, F.K., Danelljan, M., Bhat, G., Schön, T.B.: Energy-based models for deep probabilistic regression. In: ECCV (20). Lecture Notes in Computer Science, vol. 12365, pp. 325–343. Springer (2020). https://doi.org/10.1007/978-3-030-58565-5_20
24. Hand, E.M., Chellappa, R.: Attributes for improved attributes: A multi-task network utilizing implicit and explicit relationships for facial attribute classification. In: AAAI. pp. 4068–4074. AAAI Press (2017). https://doi.org/10.1609/AAAI.V31I1.11229
25. Huang, G.B., Mattar, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In: Workshop on faces in'Real-Life'Images: detection, alignment, and recognition (2008)
26. Huang, Y., Yang, H., Li, C., Kim, J., Wei, F.: Adnet: Leveraging error-bias towards normal direction in face alignment. In: ICCV. pp. 3060–3070. IEEE (2021). https://doi.org/10.1109/ICCV48922.2021.00307
27. Jia, C., Yang, Y., Xia, Y., Chen, Y., Parekh, Z., Pham, H., Le, Q.V., Sung, Y., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: ICML. Proceedings of Machine Learning Research, vol. 139, pp. 4904–4916. PMLR (2021)
28. Jin, H., Liao, S., Shao, L.: Pixel-in-pixel net: Towards efficient facial landmark detection in the wild. Int. J. Comput. Vis. **129**(12), 3174–3194 (2021). https://doi.org/10.1007/S11263-021-01521-4
29. Jr., K.R., Tesafaye, T.: MORPH: A longitudinal image database of normal adult age-progression. In: FGR. pp. 341–345. IEEE Computer Society (2006). https://doi.org/10.1109/FGR.2006.78

30. Kowalski, M., Naruniec, J., Trzcinski, T.: Deep alignment network: A convolutional neural network for robust face alignment. In: CVPR Workshops. pp. 2034–2043. IEEE Computer Society (2017). https://doi.org/10.1109/CVPRW.2017.254
31. Kumar, A., Marks, T.K., Mou, W., Wang, Y., Jones, M., Cherian, A., Koike-Akino, T., Liu, X., Feng, C.: Luvli face alignment: Estimating landmarks' location, uncertainty, and visibility likelihood. In: CVPR. pp. 8233–8243. Computer Vision Foundation / IEEE (2020). https://doi.org/10.1109/CVPR42600.2020.00826
32. Kumar, N., Berg, A.C., Belhumeur, P.N., Nayar, S.K.: Describable visual attributes for face verification and image search. IEEE Trans. Pattern Anal. Mach. Intell. **33**(10), 1962–1977 (2011). https://doi.org/10.1109/TPAMI.2011.48
33. Lee, C., Liu, Z., Wu, L., Luo, P.: Maskgan: Towards diverse and interactive facial image manipulation. In: CVPR. pp. 5548–5557. Computer Vision Foundation / IEEE (2020). https://doi.org/10.1109/CVPR42600.2020.00559
34. Lee, S., Kim, C.: Deep repulsive clustering of ordered data based on order-identity decomposition. In: ICLR. OpenReview.net (2021)
35. Li, H., Wang, N., Ding, X., Yang, X., Gao, X.: Adaptively learning facial expression representation via C-F labels and distillation. IEEE Trans. Image Process. **30**, 2016–2028 (2021). https://doi.org/10.1109/TIP.2021.3049955
36. Li, H., Guo, Z., Rhee, S., Han, S., Han, J.: Towards accurate facial landmark detection via cascaded transformers. In: CVPR. pp. 4166–4175. IEEE (2022). https://doi.org/10.1109/CVPR52688.2022.00414
37. Li, S., Deng, W., Du, J.: Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In: CVPR. pp. 2584–2593. IEEE Computer Society (2017). https://doi.org/10.1109/CVPR.2017.277
38. Li, W., Lu, J., Feng, J., Xu, C., Zhou, J., Tian, Q.: Bridgenet: A continuity-aware probabilistic network for age estimation. In: CVPR. pp. 1145–1154. Computer Vision Foundation / IEEE (2019). https://doi.org/10.1109/CVPR.2019.00124
39. Li, W., Lu, Y., Zheng, K., Liao, H., Lin, C., Luo, J., Cheng, C., Xiao, J., Lu, L., Kuo, C., Miao, S.: Structured landmark detection via topology-adapting deep graph learning. In: ECCV (9). Lecture Notes in Computer Science, vol. 12354, pp. 266–283. Springer (2020). https://doi.org/10.1007/978-3-030-58545-7_16
40. Li, Y., Zeng, J., Shan, S., Chen, X.: Occlusion aware facial expression recognition using CNN with attention mechanism. IEEE Trans. Image Process. **28**(5), 2439–2450 (2019). https://doi.org/10.1109/TIP.2018.2886767
41. Lim, K., Shin, N., Lee, Y., Kim, C.: Order learning and its application to age estimation. In: ICLR. OpenReview.net (2020)
42. Lin, J., Yang, H., Chen, D., Zeng, M., Wen, F., Yuan, L.: Face parsing with roi tanh-warping. In: CVPR. pp. 5654–5663. Computer Vision Foundation / IEEE (2019). https://doi.org/10.1109/CVPR.2019.00580
43. Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., Song, L.: Sphereface: Deep hypersphere embedding for face recognition. In: CVPR. pp. 6738–6746. IEEE Computer Society (2017). https://doi.org/10.1109/CVPR.2017.713
44. Liu, Y., Shi, H., Shen, H., Si, Y., Wang, X., Mei, T.: A new dataset and boundary-attention semantic segmentation for face parsing. In: AAAI. pp. 11637–11644. AAAI Press (2020). https://doi.org/10.1609/AAAI.V34I07.6832
45. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: ICCV. pp. 3730–3738. IEEE Computer Society (2015), https://doi.org/10.1109/ICCV.2015.425
46. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: ICLR (Poster). OpenReview.net (2019)

47. Luo, L., Xue, D., Feng, X.: Ehanet: An effective hierarchical aggregation network for face parsing. Applied Sciences **10**(9), 3135 (2020)
48. Mahbub, U., Sarkar, S., Chellappa, R.: Segment-based methods for facial attribute detection from partial faces. IEEE Trans. Affect. Comput. **11**(4), 601–613 (2020). https://doi.org/10.1109/TAFFC.2018.2820048
49. Mao, L., Yan, Y., Xue, J., Wang, H.: Deep multi-task multi-label CNN for effective facial attribute classification. IEEE Trans. Affect. Comput. **13**(2), 818–828 (2022). https://doi.org/10.1109/TAFFC.2020.2969189
50. Mollahosseini, A., Hassani, B., Mahoor, M.H.: Affectnet: A database for facial expression, valence, and arousal computing in the wild. IEEE Trans. Affect. Comput. **10**(1), 18–31 (2019). https://doi.org/10.1109/TAFFC.2017.2740923
51. Moschoglou, S., Papaioannou, A., Sagonas, C., Deng, J., Kotsia, I., Zafeiriou, S.: Agedb: The first manually collected, in-the-wild age database. In: CVPR Workshops. pp. 1997–2005. IEEE Computer Society (2017). https://doi.org/10.1109/CVPRW.2017.250
52. Niu, Z., Zhou, M., Wang, L., Gao, X., Hua, G.: Ordinal regression with multiple output CNN for age estimation. In: CVPR. pp. 4920–4928. IEEE Computer Society (2016). https://doi.org/10.1109/CVPR.2016.532
53. van den Oord, A., Vinyals, O., Kavukcuoglu, K.: Neural discrete representation learning. In: NIPS. pp. 6306–6315 (2017)
54. Pan, H., Han, H., Shan, S., Chen, X.: Mean-variance loss for deep age estimation from a face. In: CVPR. pp. 5285–5294. Computer Vision Foundation / IEEE Computer Society (2018). https://doi.org/10.1109/CVPR.2018.00554
55. Qian, S., Sun, K., Wu, W., Qian, C., Jia, J.: Aggregation via separation: Boosting facial landmark detector with semi-supervised style translation. In: ICCV. pp. 10152–10162. IEEE (2019). https://doi.org/10.1109/ICCV.2019.01025
56. Qin, L., Wang, M., Deng, C., Wang, K., Chen, X., Hu, J., Deng, W.: Swinface: A multi-task transformer for face recognition, expression recognition, age estimation and attribute estimation. IEEE Trans. Circuit Syst. Video Technol. (2023). https://doi.org/10.1109/TCSVT.2023.3304724
57. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: ICML. Proceedings of Machine Learning Research, vol. 139, pp. 8748–8763. PMLR (2021)
58. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: ICML. Proceedings of Machine Learning Research, vol. 139, pp. 8748–8763. PMLR (2021)
59. Ranjan, R., Patel, V.M., Chellappa, R.: Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. IEEE Trans. Pattern Anal. Mach. Intell. **41**(1), 121–135 (2019). https://doi.org/10.1109/TPAMI.2017.2781233
60. Ranjan, R., Sankaranarayanan, S., Castillo, C.D., Chellappa, R.: An all-in-one convolutional neural network for face analysis. In: FG. pp. 17–24. IEEE Computer Society (2017). https://doi.org/10.1109/FG.2017.137
61. Rothe, R., Timofte, R., Gool, L.V.: Deep expectation of real and apparent age from a single image without facial landmarks. Int. J. Comput. Vis. **126**(2-4), 144–157 (2018). https://doi.org/10.1007/S11263-016-0940-3
62. Rudd, E.M., Günther, M., Boult, T.E.: MOON: A mixed objective optimization network for the recognition of facial attributes. In: ECCV (5). Lecture Notes in

Computer Science, vol. 9909, pp. 19–35. Springer (2016). https://doi.org/10.1007/978-3-319-46454-1_2
63. Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., Pantic, M.: 300 faces in-the-wild challenge: The first facial landmark localization challenge. In: ICCV Workshops. pp. 397–403. IEEE Computer Society (2013). https://doi.org/10.1109/ICCVW.2013.59
64. Sengupta, S., Chen, J., Castillo, C.D., Patel, V.M., Chellappa, R., Jacobs, D.W.: Frontal to profile face verification in the wild. In: WACV. pp. 1–9. IEEE Computer Society (2016). https://doi.org/10.1109/WACV.2016.7477558
65. She, J., Hu, Y., Shi, H., Wang, J., Shen, Q., Mei, T.: Dive into ambiguity: Latent distribution mining and pairwise uncertainty estimation for facial expression recognition. In: CVPR. pp. 6248–6257. Computer Vision Foundation / IEEE (2021). https://doi.org/10.1109/CVPR46437.2021.00618
66. Shen, W., Guo, Y., Wang, Y., Zhao, K., Wang, B., Yuille, A.L.: Deep regression forests for age estimation. In: CVPR. pp. 2304–2313. Computer Vision Foundation / IEEE Computer Society (2018). https://doi.org/10.1109/CVPR.2018.00245
67. Shen, W., Guo, Y., Wang, Y., Zhao, K., Wang, B., Yuille, A.L.: Deep differentiable random forests for age estimation. IEEE Trans. Pattern Anal. Mach. Intell. **43**(2), 404–419 (2021). https://doi.org/10.1109/TPAMI.2019.2937294
68. Shin, N., Lee, S., Kim, C.: Moving window regression: A novel approach to ordinal regression. In: CVPR. pp. 18739–18748. IEEE (2022). https://doi.org/10.1109/CVPR52688.2022.01820
69. Te, G., Hu, W., Liu, Y., Shi, H., Mei, T.: Agrnet: Adaptive graph representation learning and reasoning for face parsing. IEEE Trans. Image Process. **30**, 8236–8250 (2021). https://doi.org/10.1109/TIP.2021.3113780
70. Te, G., Liu, Y., Hu, W., Shi, H., Mei, T.: Edge-aware graph representation learning and reasoning for face parsing. In: ECCV (12). Lecture Notes in Computer Science, vol. 12357, pp. 258–274. Springer (2020). https://doi.org/10.1007/978-3-030-58610-2_16
71. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: NIPS. pp. 5998–6008 (2017)
72. Wang, H., Wang, Y., Zhou, Z., Ji, X., Gong, D., Zhou, J., Li, Z., Liu, W.: Cosface: Large margin cosine loss for deep face recognition. In: CVPR. pp. 5265–5274. Computer Vision Foundation / IEEE Computer Society (2018). https://doi.org/10.1109/CVPR.2018.00552
73. Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., Liu, W., Xiao, B.: Deep high-resolution representation learning for visual recognition. IEEE Trans. Pattern Anal. Mach. Intell. **43**(10), 3349–3364 (2021). https://doi.org/10.1109/TPAMI.2020.2983686
74. Wang, K., Peng, X., Yang, J., Lu, S., Qiao, Y.: Suppressing uncertainties for large-scale facial expression recognition. In: CVPR. pp. 6896–6905. Computer Vision Foundation / IEEE (2020). https://doi.org/10.1109/CVPR42600.2020.00693
75. Wang, K., Peng, X., Yang, J., Meng, D., Qiao, Y.: Region attention networks for pose and occlusion robust facial expression recognition. IEEE Trans. Image Process. **29**, 4057–4069 (2020). https://doi.org/10.1109/TIP.2019.2956143
76. Wang, W., Sebe, N., Lepri, B.: Rethinking the learning paradigm for facial expression recognition. CoRR **abs/2209.15402** (2022). https://doi.org/10.48550/ARXIV.2209.15402
77. Wang, X., Bo, L., Li, F.: Adaptive wing loss for robust face alignment via heatmap regression. In: ICCV. pp. 6970–6980. IEEE (2019). https://doi.org/10.1109/ICCV.2019.00707

78. Wei, Z., Liu, S., Sun, Y., Ling, H.: Accurate facial image parsing at real-time speed. IEEE Trans. Image Process. **28**(9), 4659–4670 (2019). https://doi.org/10.1109/TIP.2019.2909652
79. Wu, W., Qian, C., Yang, S., Wang, Q., Cai, Y., Zhou, Q.: Look at boundary: A boundary-aware face alignment algorithm. In: CVPR. pp. 2129–2138. Computer Vision Foundation / IEEE Computer Society (2018). https://doi.org/10.1109/CVPR.2018.00227
80. Wu, W., Qian, C., Yang, S., Wang, Q., Cai, Y., Zhou, Q.: Look at boundary: A boundary-aware face alignment algorithm. In: CVPR. pp. 2129–2138. Computer Vision Foundation / IEEE Computer Society (2018). https://doi.org/10.1109/CVPR.2018.00227
81. Xia, J., Qu, W., Huang, W., Zhang, J., Wang, X., Xu, M.: Sparse local patch transformer for robust face alignment and landmarks inherent relation learning. In: CVPR. pp. 4042–4051. IEEE (2022). https://doi.org/10.1109/CVPR52688.2022.00402
82. Xiao, T., Liu, Y., Zhou, B., Jiang, Y., Sun, J.: Unified perceptual parsing for scene understanding. In: ECCV (5). Lecture Notes in Computer Science, vol. 11209, pp. 432–448. Springer (2018). https://doi.org/10.1007/978-3-030-01228-1_26
83. Zeng, J., Shan, S., Chen, X.: Facial expression recognition with inconsistently annotated datasets. In: ECCV (13). Lecture Notes in Computer Science, vol. 11217, pp. 227–243. Springer (2018). https://doi.org/10.1007/978-3-030-01261-8_14
84. Zhang, C., Zhong, W., Li, C., Deng, H.: Random walk-based erasing data augmentation for deep learning. Signal Image Video Process. **17**(5), 2447–2454 (2023). https://doi.org/10.1007/S11760-022-02461-3
85. Zhang, N., Paluri, M., Ranzato, M., Darrell, T., Bourdev, L.D.: PANDA: pose aligned networks for deep attribute modeling. In: CVPR. pp. 1637–1644. IEEE Computer Society (2014). https://doi.org/10.1109/CVPR.2014.212
86. Zhang, Y., Wang, C., Deng, W.: Relative uncertainty learning for facial expression recognition. In: NeurIPS. pp. 17616–17627 (2021)
87. Zhang, Y., Wang, C., Ling, X., Deng, W.: Learn from all: Erasing attention consistency for noisy label facial expression recognition. In: ECCV (26). Lecture Notes in Computer Science, vol. 13686, pp. 418–434. Springer (2022). https://doi.org/10.1007/978-3-031-19809-0_24
88. Zhang, Z., Song, Y., Qi, H.: Age progression/regression by conditional adversarial autoencoder. In: CVPR. pp. 4352–4360. IEEE Computer Society (2017). https://doi.org/10.1109/CVPR.2017.463
89. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: CVPR. pp. 6230–6239. IEEE Computer Society (2017). https://doi.org/10.1109/CVPR.2017.660
90. Zheng, Q., Deng, J., Zhu, Z., Li, Y., Zafeiriou, S.: Decoupled multi-task learning with cyclical self-regulation for face parsing. In: CVPR. pp. 4146–4155. IEEE (2022). https://doi.org/10.1109/CVPR52688.2022.00412
91. Zheng, T., Deng, W.: Cross-pose lfw: A database for studying cross-pose face recognition in unconstrained environments. Beijing University of Posts and Telecommunications, Tech. Rep **5**, 7 (2018)
92. Zheng, T., Deng, W., Hu, J.: Cross-age LFW: A database for studying cross-age face recognition in unconstrained environments. CoRR **abs/1708.08197** (2017)
93. Zheng, Y., Yang, H., Zhang, T., Bao, J., Chen, D., Huang, Y., Yuan, L., Chen, D., Zeng, M., Wen, F.: General facial representation learning in a visual-linguistic manner. In: CVPR. pp. 18676–18688. IEEE (2022). https://doi.org/10.1109/CVPR52688.2022.01814

94. Zhu, S., Li, C., Loy, C.C., Tang, X.: Unconstrained face alignment via cascaded compositional learning. In: CVPR. pp. 3409–3417. IEEE Computer Society (2016). https://doi.org/10.1109/CVPR.2016.371
95. Zhuang, N., Yan, Y., Chen, S., Wang, H.: Multi-task learning of cascaded CNN for facial attribute classification. In: ICPR. pp. 2069–2074. IEEE Computer Society (2018). https://doi.org/10.1109/ICPR.2018.8545271