Faceptor: A Generalist Model for Face Perception
- Supplemental Document -

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

A Supplementary Explanations of the Method

A.1 Categorization for Face Analysis Tasks

Face analysis tasks can be classified into the following three categories based on the differences in shape and granularity of their expected outputs:

1. **Dense prediction** involves tasks like facial landmark localization, face parsing, and depth estimation that require predictions for each pixel in an image.
2. **Attribute prediction** includes tasks such as age estimation, expression recognition, binary attribute classification (e.g., gender classification), race classification, face forgery detection, and face anti-spoofing. The prediction outcome in these tasks is a continuous or discrete label.
3. **Identity prediction**, commonly referred to as face recognition, is a basic face perception task that represents a face identity with a vector.

A.2 Naive Faceptor

As shown in Fig. 1, the Naive Faceptor employs standardized face analysis and face recognition subnets from SwinFace [36] as attribute prediction head and identity prediction head, respectively. The Multi-Level Channel Attention (MLCA) module is integrated into the attribute prediction head, which consists of a Multi-Level Feature Fusion (MLFF) module and a Channel Attention (CA) module. MLFF is used to combine feature maps at different levels enabling the task-specific subnet to rely on both local and global information of the faces and CA emphasizes the contributions of different levels for the specific group of tasks. In addition, we follow the implementation in the FaRL experiment, utilizing UperNet [52] as the dense prediction head to produce dense output.

For dense prediction tasks, encoded features from the 4th, 6th, 8th, and 12th layers of the transformer encoder [8] are passed into dense prediction heads, while for attribute prediction tasks, features from the 6th, 8th, 10th, and 12th

* Corresponding author.

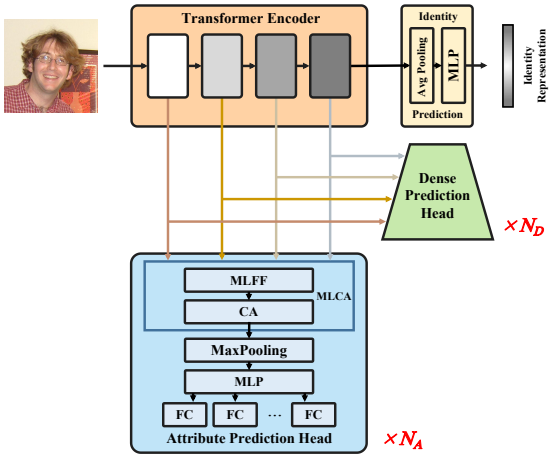

Fig. 1: Overall architecture for the proposed Naive Faceptor

layers are utilized. In our experiments, given the involvement of two dense prediction tasks (facial landmark localization and face parsing) and three attribute prediction tasks (age estimation, expression recognition, and binary attribute classification), N_D and N_A are set to 2 and 3 respectively. In the dense prediction heads, the channel number of the output map is also an adjustable hyperparameter, configured as the number of landmarks and semantic parsing classes for the two tasks respectively. Regarding the attribute prediction heads, the number and the output dimension of FC units are adjustable hyperparameters, set to 1 and 101 for age estimation, 1 and 7 for expression recognition, and 40 and 1 for binary attribute classification.

A.3 Objective Functions

In our experiments, we evaluate the effectiveness of the proposed face generalist models across a diverse set of tasks. The objective function for each task used in our framework is as follows.

Facial Landmark Localization This task aims to predict heatmaps of the landmarks, as is practiced by AWing [47], LUVLi [22] and ADNet [16]. We employ a loss function, combining the binary cross-entropy loss and the L1 loss:

$$L_{lan} = \sum_{k=1}^{N_{lan}} \{-\frac{1}{M} \sum_{m=1}^{M} [(1-p_{k,m}) \log(1-\hat{p}_{k,m}) + p_{k,m} \log \hat{p}_{k,m}] + \lambda [|\hat{l}_{k,x} - l_{k,x}| + |\hat{l}_{k,y} - l_{k,y}|]\}, \quad (1)$$

where $p_{k,m} = 1$ if pixel m of the input image is within the circle centered at landmark k with a radius of 5, otherwise 0. The output heatmap for facial landmark localization is $\mathbf{y}_{lan} \in \mathbb{R}^{N_{lan} \times H \times W}$. The predicted probability $\hat{p}_{k,m}$ for

pixel m at the channel k of the output heatmap is calculated with the sigmoid function. M represents the total number of pixels, which is equal to the product of H and W. $(\hat{l}_{k,x}, \hat{l}_{k,y})$ is the predicted result for the coordinate of landmark k, calculated by normalizing the channel k of the output heatmap. $(l_{k,x}, l_{k,y})$ represents the ground-truth for the coordinate of landmark k. λ is a weight that balances the importance between two types of losses and is set to 1.0.

Face Paring This task is trained with cross-entropy loss for each pixel:

$$L_{par} = -\frac{1}{M}\sum_{m=1}^{M}\sum_{i=1}^{N_{par}} p_{i,m}\log\hat{p}_{i,m}, \tag{2}$$

where $p_{i,m} = 1$ if pixel m of the input image belongs to semantic parsing class i, otherwise 0. The predicted probability $\hat{p}_{i,m}$ is calculated with softmax for each pixel from the output for face parsing $\mathbf{y}_{par} \in \mathbb{R}^{N_{par}\times H\times W}$. M represents the total number of pixels, which is equal to the product of H and W.

Age Estimation We train the task of age estimation by jointly learning label distribution and expectation regression, following DLDL-v2 [11]:

$$L_{age} = -\sum_{i=0}^{100} p_i\log\hat{p}_i + \lambda|\hat{a} - a|, \tag{3}$$

where $\mathbf{p}$ is age label distribution which can be estimated with training samples. $\hat{\mathbf{p}}$ is predicted distribution which should be similar to $\mathbf{p}$. We use a softmax function to turn the output for age estimation $\mathbf{y}_i \in \mathbb{R}^{101}$, into a predicted probability, that is, $\hat{p}_i = \frac{\exp(\mathbf{y}_{age,i})}{\sum_{j=0}^{100}\exp(\mathbf{y}_{age,j})}$. a_i is the ground-truth age. The predicted age $\hat{a}$ can be calculated by: $\hat{a} = \sum_{i=0}^{100} i\hat{p}_i$. λ is a weight that balances the importance between two types of losses and is set to 1.0.

Expression Recognition The expressions include surprise, fear, disgust, happiness, sadness, anger, and neutral. The loss function for training is as follows:

$$L_{exp} = -\sum_{i=1}^{7} p_i\log\hat{p}_i, \tag{4}$$

where $p_i = 1$ if the input sample belongs to expression class i, otherwise 0. The predicted probability is given by $\hat{p}_i$ which is calculated from the output for expression recognition $\mathbf{y}_{exp} \in \mathbb{R}^7$ by softmax.

Binary Attribute Classfication This task involves N_{att} binary labels, and the total loss function is the sum of N_{att} binary cross-entropy loss functions:

$$L_{att} = -\sum_{k=1}^{N_{att}}[(1-p_k)\log(1-\hat{p}_k) + p_k\log\hat{p}_k], \tag{5}$$

where $p_k = 1$ for the k-th attribute exists and 0 otherwise. p_k is the predicted probability that the input face contains the k-th attribute. It is calculated from the $\mathbf{y}_{att} \in \mathbb{R}^{N_{att}}$ by sigmoid function.

Face Recognition We train the task of face recognition with CosFace [45]:

$$L_{rec} = -\log \frac{e^{s(\cos\theta_i - m)}}{e^{s(\cos\theta_i - m)} + \sum_{j=1, j\neq i}^{n} e^{s\cos\theta_j}}. \tag{6}$$

Initially, we feed the facial identity representation obtained from Naive Faceptor or Faceptor into a fully connected layer to predict the identity label of the sample. The weight of the fully connected layer can be written as $\mathbf{W} \in \mathbb{R}^{d\times n}$, where n is the number of identities. We use $\mathbf{W}_j \in \mathbb{R}^d$ to denote the j-th column of the weight $\mathbf{W}$ and $\mathbf{y}_{rec} \in \mathbb{R}^d$ to denote the deep feature for the input sample, belonging to the i-th class. θ_j is the angle between the weight $\mathbf{W}_j$ and the feature $\mathbf{y}_{rec}$. The embedding feature $\|\mathbf{y}_{rec}\|$ is fixed by l_2 normalization and re-scaled to s. m is the CosFace margin penalty. In our implementation, s is set to 64, and m is set to 0.4.

B Implementation details

B.1 Datasets

300W [39]: It is the most commonly used dataset for facial landmark localization which includes 3,148 images for training and 689 images for testing. The training set consists of the full set of AFW [58], the training subset of HELEN [24], and LFPW [2]. The test set is further divided into a challenging subset that includes 135 images (IBUG full set [39]) and a common subset that consists of 554 images (test subset of HELEN and LFPW). Each image in 300W is annotated with 68 facial landmarks.

WFLW [49]: It is collected from the WIDER Face dataset, encompassing large variations in pose, expression, and occlusion. It provides 98 manually annotated landmarks for 10,000 images, 7,500 for training, and 2,500 for testing. The test set is further divided into 6 subsets for different scenarios.

COFW [3]: It contains 1,345 training images and 507 testing images with 29 landmarks.

AFLW-19 [57]: The original AFLW [20] provides at most 21 landmarks for each face but excludes coordinates for invisible landmarks. AFLW-19 provides manually annotated coordinates for these invisible landmarks. The new annotation does not include two ear points because it is very difficult to decide the location

of invisible ears. This causes the point number of AFLW-19 to be 19. The original AFLW does not provide a train-test partition. AFLW-19 adopts a partition with 20,000 images for training and 4,386 images for testing (AFLW-Full). In addition, a frontal subset (AFLW-Frontal) is proposed where all landmarks are visible (a total of 1,165 images).

CelebAMask-HQ [25]: CelebAMask-HQ consists of 30,000 high-resolution face images selected from the CelebA dataset. The masks of CelebAMask-HQ are manually annotated with the size of 512×512 and 19 classes.

LaPa [29]: It consists of more than 22,000 facial images with abundant variations in expression, pose, and occlusion. Each image of LaPa is provided with an 11-category pixel-level label map and 106-point landmarks.

MORPH II [18]: It is an age estimation dataset, which contains 55,134 facial images of 13,617 subjects ranging from 16 to 77 years old. The entire dataset is randomly divided into five folds, with four folds allocated for training and one fold reserved for testing.

UTKFace [53]: It provides about 20,000 facial images ranging from 0 to 116 years old. For a fair comparison, we employ the evaluation protocol using a subset of UTKFace covering faces between 21 and 60 years old as in MWR [42] - 13,147 for training, and 3,287 for testing.

AffectNet [33]: It stands as the largest publicly available dataset for facial expression recognition, comprising about 420K images with manually annotated labels. Due to significant label noise in this dataset and highly imbalanced training data distribution, we opt not to conduct testing on it. We preprocess the dataset according to the methods outlined in HSEmotion [40] for joint training. In our experiments, we utilize the version that includes seven classes of facial expressions.

RAF-DB [27]: It is a real-world expression dataset comprising 29,672 real-world facial images collected through Flickr's image search API and independently labeled by approximately 40 trained human annotators. For our experiments, we utilize the single-label subset, which consists of 15,339 expression images with six basic emotions (happiness, surprise, sadness, anger, disgust, and fear), along with the neutral expression. Among these, 12,271 images are used for training purposes, while the remaining images are reserved for testing.

FERPlus [1]: FERPlus [1] is extended from FER2013 [12] which is a large-scale dataset collected by APIs in the Google image search. It contains 28,709 training, 3,589 validation, and 3,589 test images. In our experiments, we utilize the version that includes seven classes of facial expressions.

CelebA [30]: It is a large-scale collection of facial attributes, comprising 162,770 images for training, 19,867 images for validation, and 19,962 images for testing. Each image in CelebA is extensively annotated with 40 binary attributes.

LFW-73 [23]: It is another challenging facial dataset, comprising 13,143 images annotated with 73 binary facial attributes, 40 of which are shared with CelebA. This dataset is divided in half for training (6,263 images) and testing (6,880 images). We utilize this dataset in the cross-datasets transfer experiment.

MS-Celeb-1M [13]: It is one of the most popular training datasets in the field of facial recognition and we utilize the clean version refined by insightface [6], containing 5.3M images of 93,431 celebrities.

Face Verification Datasets: LFW [15] database contains 13,233 face images from 5,749 different identities, which is a classic benchmark for unconstrained face verification. CFP-FP [41] and CPLFW [55] are built to emphasize the cross-pose challenge while AgeDB-30 [34] and CALFW [56] are built for the cross-age challenge.

B.2 Auxiliary Supervised Learning

In the setting of auxiliary supervised learning, we consider age estimation and expression recognition as the main tasks respectively, while facial landmark localization, face parsing, and face recognition serve as auxiliary tasks. The batch size and weight used for each dataset is presented in Tab. 1. Other hyper-parameters are kept consistent with the first stage of training the Faceptor-Base.

Table 1: The batch size and weight used for each dataset in the setting of auxiliary supervised learning

Category	Task	Dataset	Main Task			
			Age		Expression	
			n_t	α_t	n_t	α_t
Main Task	Age Estimation	MORPH II [18]	64	10.0	-	-
	Expression Recognition	AffectNet [33]	-	-	64	80.0
		RAF-DB [27]	-	-	16	20.0
Auxiliary Tasks	Landmark Localization	300W [39]	4	1000.0	2	1000.0
	Face Parsing	CelebAMask-HQ [25]	4	100.0	2	100.0
	Face Recognition	MS1MV3 [13]	256	10.0	64	1.0

B.3 Cross-Datasets Transfer

Starting from Faceptor-Base, cross-dataset transfer experiments are conducted on AFLW-19 [57], LaPa [29], and LFW-73 [23] with batch sizes set to 8, 8, and 32 respectively. 20000 steps are required for tuning, with 2000 steps reserved for linear warm-up. Other hyper-parameters are kept consistent with the first stage of training the Faceptor-Base.

C Additional Results

C.1 Performance of Early All-In-One Models

Early all-in-one models [37,38] employ significantly simpler testing protocols that are now rarely referenced. In this section, we provide a detailed discussion of the performance of these early models on the tasks they can address. Through indirect comparison, we have demonstrated that the proposed Faceptor outperforms these early all-in-one models significantly.

Facial Landmark Localization HyperFace [37] and AIO [38] report performance for facial landmark localization on AFW [58] and the original AFLW [20] datasets. AFW contains only 205 images with 468 faces. The full set of AFW has already been incorporated into the training samples of the 300W [39] protocol. By manually annotating coordinates for invisible landmarks, the original AFLW dataset has been reprocessed into the more commonly used testing protocol known as AFLW-19 [57]. Our Faceptor has achieved performance surpassing the state-of-the-art method on more challenging 300W and AFLW-19. Although results on AFW and the original AFLW dataset are not reported, it is evident that our Faceptor significantly outperforms early all-in-one methods in facial landmark localization.

Age Estimation AIO [38] provides test results on CLAP2015 [9] and FG-NET [14] datasets. CLAP2015 consists of 2,476 training samples and 1,079 testing samples. FG-NET contains a total of 1,002 face samples and is commonly used for leave-one-person-out testing protocol. To ensure an adequate number of training and testing examples, we employ the MORPH II [18] and UTKFace [53] protocols to evaluate the performance of our proposed models in age estimation. By providing results of MWR [42] on MORPH II, UTKFace, and CLAP2015, we indirectly demonstrate the superior age estimation capabilities of our Faceptor compared to the early all-in-one model.

Table 2: Comparison for age estimation

Methods	MORPH II UTKFace MAE ↓		CLAP2015 ϵ-error ↓
AIO [38]	-	-	0.29
MWR [42]	2.00	4.37	**0.26**
Faceptor-Full	**1.96**	**4.10**	-

Binary Attribute Classification CelebA [30] is the most commonly used binary attribute classification dataset. HyperFace [37] supports only gender classification, while AIO [38] supports gender and smile classification. Our Faceptor supports all 40 attribute classification tasks involved in CelebA. Even in gender and smile classification tasks, our method achieves the same accuracy as AIO.

Table 3: Comparison for binary attribute classification

Methods	Gender Smile Acc ↑		All Attributes mAcc ↑
HyperFace [37]	97	-	-
AIO [38]	**99**	**93**	-
Faceptor-Full	**99**	**93**	91.39

Face Recognition AIO [38] evaluates face recognition on IJB-A [19]. The dataset has been extended to IJB-C [32], which is more challenging. By providing the results of VGGFace2 [4] on IJB-A and IJB-C, we indirectly demonstrate the superior face recognition capability of our method compared to the early all-in-one model.

Table 4: Comparison for face recognition

Methods	IJB-C TAR@FAR ↑		IJB-A TAR@FAR ↑		
	0.001	0.01	0.001	0.01	0.1
AIO [38]	-	-	78.7	89.3	96.8
VGGFace2 [4]	92.7	96.7	**92.1**	**96.8**	**99.0**
Faceptor-Full	**95.7**	**98.1**	-	-	-

C.2 Performance Evaluation for Faceptor

Due to space limitations, we do not include the complete test results of the Faceptor-Full on dense prediction task in the main body of the paper. Here, Tabs. 5 to 8 present the complete results on datasets WFLW [49], 300W [39], COFW [3], AFLW-19 [57], CelebAMask- HQ [25] and Lapa [29].

Table 5: Comparison with other specialized facial landmark localization methods on WFLW [49] and 300W [39]

Methods	WFLW									300W		
	$NME_{inter-ocular}$ ↓							FR^{10} ↓	AUC^{10} ↑	$NME_{inter-ocular}$ ↓		
	Full	Pose	Expr.	Illum.	M.U.	Occl.	Blur	Full		Comm.	Chal.	Full
DAN-Menpo [21]	-	-	-	-	-	-	-	-	-	3.44	4.88	3.09
SAN [7]	-	-	-	-	-	-	-	-	-	3.34	6.60	3.98
LAB [50]	5.27	10.24	5.51	5.23	5.15	6.79	6.32	7.56	53.23	2.98	5.19	3.49
Wing [10]	5.11	8.75	5.36	4.93	5.41	6.37	5.81	6.00	55.40	3.27	7.18	4.04
DeCaFA [5]	4.62	-	-	-	-	-	-	4.84	56.30	2.93	5.26	3.39
Awing [47]	4.36	7.38	4.58	4.32	4.27	5.19	4.96	2.84	57.19	2.72	4.52	3.07
AVS [35]+SAN [7]	4.39	8.42	4.68	4.24	4.37	5.60	4.86	4.08	59.13	3.21	6.49	3.86
HRNet [46]	4.60	7.86	4.78	4.57	4.26	5.42	5.36	-	-	2.91	5.11	3.34
DAG [28]	4.21	7.36	4.49	4.12	4.05	4.98	4.82	3.04	58.93	2.62	4.77	3.04
LUVLi [22]	4.37	-	-	-	-	-	-	3.12	57.70	2.76	5.16	3.23
ADNet [16]	4.14	6.96	4.38	4.09	4.05	5.06	4.79	2.72	60.22	2.53	4.58	2.93
PIPNet [17]	4.31	7.51	4.44	4.19	4.02	5.36	5.02	-	-	2.78	4.89	3.19
SLPT [51]	4.14	-	-	-	-	-	-	2.76	59.50	2.75	4.90	3.17
DTLD+ [26]	4.05	-	-	-	-	-	-	2.68	-	2.60	4.48	2.96
Faceptor-Full	**4.03**	**6.81**	**4.28**	**3.93**	**3.91**	**4.71**	**4.56**	**1.92**	**60.24**	**2.52**	**4.25**	**2.86**

Table 6: Comparison with other specialized facial landmark localization methods on COFW [3] and AFLW-19 [57]

Methods	COFW	AFLW-19			
	$NME_{inter-ocular}$ ↓	NME_{diag} ↓		NME_{box} ↓	AUC_{box} ↑
		Full	Frontal	Full	
SAN [7]	-	1.91	1.85	4.04	54.00
LAB [50]	3.92	1.85	1.62	-	-
Wing [10]	-	1.65	-	-	-
HRNet [46]	3.45	1.57	1.46	-	-
LUVLi [22]	-	1.39	1.19	2.28	68.00
PIPNet [17]	3.08	1.42	-	-	-
SLPT [51]	3.32	-	-	-	-
DTLD+ [26]	3.02	1.37	-	-	-
Faceptor-Full	**3.01**	**0.95**	**0.87**	**1.35**	**81.11**

Table 7: Comparison with other specialized face parsing methods on CelebAMask-HQ [25]. Results are reported in F1 scores (%)

Methods	Face	Nose	Classes	L-Eye	R-Eye	L-B	R-B	L-Ear	R-Ear	Mean
	I-M	U-L	L-L	Hair	Hat	Earring	Necklace	Neck	Cloth	
EHANet [31]	96.0	93.7	90.6	86.2	86.5	83.2	83.1	86.5	84.1	84.0
	93.8	88.6	90.3	93.9	85.9	67.8	30.1	88.8	83.5	
Wei et al. [48]	96.4	91.9	89.5	87.1	85.0	80.8	82.5	84.1	83.3	82.1
	90.6	87.9	91.0	91.1	83.9	65.4	17.8	88.1	80.6	
EAGR [44]	96.2	**94.0**	92.3	88.6	88.7	85.7	85.2	88.0	85.7	85.1
	95.0	88.9	**91.2**	94.9	87.6	68.3	27.6	89.4	85.3	
AGRNet [43]	96.5	93.9	91.8	88.7	89.1	85.5	85.6	88.1	88.7	85.5
	92.0	**89.1**	91.1	95.2	87.2	69.6	32.8	89.9	84.9	
DML-CSR [54]	95.7	93.9	92.6	**89.4**	**89.6**	85.5	85.7	88.3	88.2	86.1
	91.8	87.4	91.0	94.5	88.5	71.4	40.6	89.6	85.7	
Faceptor-Full	**96.6**	93.9	**94.0**	**89.4**	89.1	**86.2**	**86.3**	**88.4**	**88.8**	**88.2**
	91.6	89.0	90.6	**96.2**	90.8	**72.5**	**61.6**	92.4	**91.0**	

Table 8: Comparison with other specialized face parsing methods on LaPa [29]. Results are reported in F1 scores (%)

Methods	Skin	Hair	L-E	R-E	U-L	I-M	L-L	Nose	L-B	R-B	Mean
BASS [29]	97.2	96.3	88.1	88.0	84.4	87.6	85.7	95.5	87.7	87.6	89.8
EHANet [31]	95.8	94.3	87.0	89.1	85.3	85.6	88.8	94.3	85.9	86.1	89.2
Wei et al. [48]	96.1	95.1	88.9	87.5	83.1	89.2	83.8	96.1	86.0	87.8	89.4
EAGR [44]	97.3	96.2	89.5	90.0	88.1	90.0	89.0	97.1	86.5	87.0	91.1
AGRNet [43]	**97.7**	**96.5**	91.6	91.1	**88.5**	**90.7**	90.1	97.3	89.9	90.0	92.3
DML-CSR [54]	97.6	96.4	91.8	91.5	88.0	90.5	89.9	97.3	90.4	90.4	92.4
Faceptor-Full	97.5	96.0	91.2	90.8	86.7	88.9	89.3	97.0	89.0	88.9	91.5
Faceptor-Base+Finetuning	**97.7**	96.3	**92.3**	**92.2**	**88.5**	90.6	**90.5**	**97.5**	**90.7**	**90.5**	**92.7**

C.3 Layer-Attention Mechanism Visualization

Averaging the cross-attention matrices in the transformer decoder, we obtained the visualization results for different tasks (in Fig. 2). It can be observed that

different tasks exhibit distinct preferences for features from various layers. For challenging attribute predictions such as age, expression, and gender, features from the 10th layer prove to be the most crucial, whereas for localized attributes like hair color, features from earlier layers are preferred.

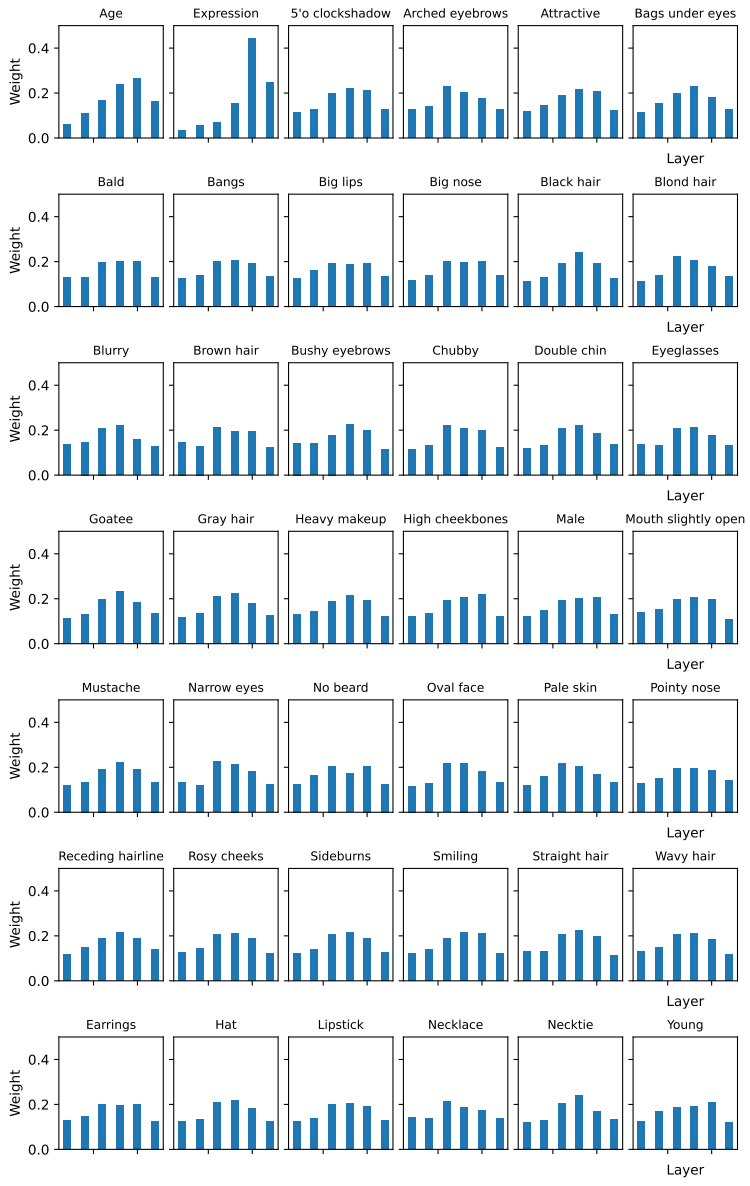

Fig. 2: Layer-Attention Mechanism Visualization. Each subplot's horizontal axis represents layers 2, 4, 6, 8, 10, and 12, respectively.

References

1. 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
2. Belhumeur, P.N., Jacobs, D.W., Kriegman, D.J., Kumar, N.: Localizing parts of faces using a consensus of exemplars. IEEE Trans. Pattern Anal. Mach. Intell. **35**(12), 2930–2940 (2013). https://doi.org/10.1109/TPAMI.2013.23
3. 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
4. Cao, Q., Shen, L., Xie, W., Parkhi, O.M., Zisserman, A.: Vggface2: A dataset for recognising faces across pose and age. In: FG. pp. 67–74. IEEE Computer Society (2018). https://doi.org/10.1109/FG.2018.00020
5. 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
6. 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
7. 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
8. 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)
9. Escalera, S., Fabian, J., Pardo, P., Baró, X., Gonzàlez, J., Escalante, H.J., Misevic, D., Steiner, U., Guyon, I.: Chalearn looking at people 2015: Apparent age and cultural event recognition datasets and results. In: ICCV Workshops. pp. 243–251. IEEE Computer Society (2015). https://doi.org/10.1109/ICCVW.2015.40
10. 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
11. 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)
12. Goodfellow, I.J., Erhan, D., Carrier, P.L., Courville, A.C., Mirza, M., Hamner, B., Cukierski, W., Tang, Y., Thaler, D., Lee, D., Zhou, Y., Ramaiah, C., Feng, F., Li, R., Wang, X., Athanasakis, D., Shawe-Taylor, J., Milakov, M., Park, J., Ionescu, R.T., Popescu, M., Grozea, C., Bergstra, J., Xie, J., Romaszko, L., Xu, B., Zhang, C., Bengio, Y.: Challenges in representation learning: A report on three machine learning contests. In: ICONIP (3). Lecture Notes in Computer Science, vol. 8228, pp. 117–124. Springer (2013). https://doi.org/10.1007/978-3-642-42051-1_16
13. 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
14. Han, H., Otto, C., Jain, A.K.: Age estimation from face images: Human vs. machine performance. In: ICB. pp. 1–8. IEEE (2013). https://doi.org/10.1109/ICB.2013.6613022

15. 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)
16. 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
17. 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
18. 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
19. Klare, B.F., Klein, B., Taborsky, E., Blanton, A., Cheney, J., Allen, K., Grother, P., Mah, A., Burge, M.J., Jain, A.K.: Pushing the frontiers of unconstrained face detection and recognition: IARPA janus benchmark A. In: CVPR. pp. 1931–1939. IEEE Computer Society (2015). https://doi.org/10.1109/CVPR.2015.7298803
20. Köstinger, M., Wohlhart, P., Roth, P.M., Bischof, H.: Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In: ICCV Workshops. pp. 2144–2151. IEEE Computer Society (2011). https://doi.org/10.1109/ICCVW.2011.6130513
21. 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
22. 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
23. 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
24. Le, V., Brandt, J., Lin, Z., Bourdev, L.D., Huang, T.S.: Interactive facial feature localization. In: ECCV (3). Lecture Notes in Computer Science, vol. 7574, pp. 679–692. Springer (2012). https://doi.org/10.1007/978-3-642-33712-3_49
25. 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
26. 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
27. 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
28. 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
29. 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

30. 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
31. Luo, L., Xue, D., Feng, X.: Ehanet: An effective hierarchical aggregation network for face parsing. Applied Sciences **10**(9), 3135 (2020)
32. Maze, B., Adams, J.C., Duncan, J.A., Kalka, N.D., Miller, T., Otto, C., Jain, A.K., Niggel, W.T., Anderson, J., Cheney, J., Grother, P.: IARPA janus benchmark - C: face dataset and protocol. In: ICB. pp. 158–165. IEEE (2018). https://doi.org/10.1109/ICB2018.2018.00033
33. 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
34. 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
35. 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
36. 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
37. 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
38. 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
39. 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
40. Savchenko, A.V.: Facial expression recognition with adaptive frame rate based on multiple testing correction. In: ICML. Proceedings of Machine Learning Research, vol. 202, pp. 30119–30129. PMLR (2023)
41. 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
42. 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
43. 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
44. 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

45. 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
46. 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
47. 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
48. 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
49. 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
50. 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
51. 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
52. 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
53. 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
54. 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
55. 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)
56. 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)
57. 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
58. Zhu, X., Ramanan, D.: Face detection, pose estimation, and landmark localization in the wild. In: CVPR. pp. 2879–2886. IEEE Computer Society (2012). https://doi.org/10.1109/CVPR.2012.6248014