FaceXFormer: A Unified Transformer for Facial Analysis

Kartik Narayan* Vibashan VS* Rama Chellappa Vishal M. Patel
{knaraya4, vvishnu2, rchella4, vpatel36}@jhu.edu
https://kartik-3004.github.io/facexformer/

Abstract

In this work, we introduce FaceXFormer, an end-to-end unified transformer model capable of performing ten facial analysis tasks within a single framework. These tasks include face parsing, landmark detection, head pose estimation, attribute prediction, age, gender, and race estimation, facial expression recognition, face recognition, and face visibility. Traditional face analysis approaches rely on task-specific architectures and pre-processing techniques, limiting scalability and integration. In contrast, FaceXFormer employs a transformer-based encoder-decoder architecture, where each task is represented as a learnable token, enabling seamless multi-task processing within a unified model. To enhance efficiency, we introduce FaceX, a lightweight decoder with a novel bi-directional cross-attention mechanism, which jointly processes face and task tokens to learn robust and generalized facial representations. We train FaceXFormer on ten diverse face perception datasets and evaluate it against both specialized and multi-task models across multiple benchmarks, demonstrating state-of-the-art or competitive performance. Additionally, we analyze the impact of various components of FaceXFormer on performance, assess real-world robustness in "in-the-wild" settings, and conduct a computational performance evaluation. To the best of our knowledge, FaceXFormer is the first model capable of handling ten facial analysis tasks while maintaining real-time performance at 33.21 FPS.

1. Introduction

Face analysis is a crucial problem as it has broad range of application such as face verification and identification [92, 93], surveillance [25], face swapping [14], face editing [137], de-occlusion [120], 3D face reconstruction [111], retail [1], image generation [118] and face retrieval [122]. Facial analysis tasks (Figure 1 include face parsing [33, 107], landmarks detection [51, 135], head pose estimation [13, 134], facial attributes recognition [63, 71], age/gender/race estimation [9, 48], facial expression recognition [85], face recognition [39], and face visibility prediction [41, 60]. Therefore, developing a generalized and robust face model for all tasks is a crucial and longstanding problem in the face community.

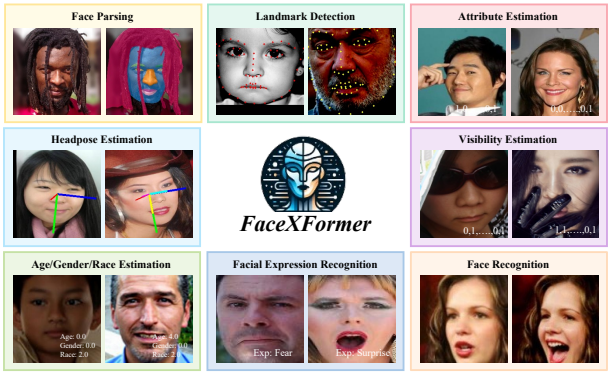

Figure 1. *FaceXFormer* an end-to-end unified transformer model for 10 different facial analysis tasks such as face parsing, landmark detection, head pose estimation, attributes recognition, age, gender, and race estimation, facial expression recognition, face recognition, and face visibility prediction.

cross-attention 2 chiều

Why Unified Model ? In recent years, significant advancements have been made in facial analysis, developing state-of-the-art methods and face libraries for various tasks [13, 14, 48, 120, 134, 135]. Despite these methods achieving promising performance, they cannot be integrated into a single pipeline due to their specialized model designs and task-specific pre-processing techniques. Furthermore, deploying multiple specialized models simultaneously is computationally intensive and impractical for real-time applications, leading to increased system complexity and resource consumption. These challenges emphasis the need for a unified model that can concurrently handle multiple facial analysis tasks efficiently (see Table 1). A single model capable of addressing multiple facial tasks is desirable because it: (1) learns a robust and generalized face representation capable of handling "in-the-wild" images; (2) intra-task modeling helps the models to learn task-invariant representation;

*Equal contribution

Methods	FP	LD	HPE	Attr	Age	Gen	Race	Vis	Exp	FR
Single-Task Models										
DML-CSR [130]	✓									
FP-LIIF [83]	✓									
SegFace [69]	✓									
Wing [23]		✓								
HRNet [101]		✓								
WHENet [134]			✓							
TriNet [10]			✓							
img2pose [3]			✓							
TokenHPE [124]			✓							
SSPL [88]				✓						
VOLO-D1 [42]					✓					
DLDL-v2 [24]					✓					
3DDE [97]								✓		
MNN [98]								✓		
KTN [46]									✓	
DMUE [85]									✓	
CosFace [100]										✓
ArcFace [16]										✓
AdaFace [39]										✓
Multi-Task Models										
SSP+SSG [35]	✓			✓						
Hetero-FAE [28]				✓	✓	✓	✓		✓	
FairFace [36]					✓	✓	✓			
MiVOLO [42]					✓	✓				
MTL-CNN [141]		✓		✓					✓	
ProS [18]	✓	✓		✓						
FaRL [133]	✓	✓		✓	✓	✓				
HyperFace [77]		✓	✓			✓		✓		✓
AllinOne [78]		✓	✓		✓	✓			✓	✓
Swinface [73]				✓	✓				✓	✓
QFace [91]			✓	✓	✓				✓	
Faceptor [74]	✓	✓		✓	✓	✓			✓	✓
FaceXFormer	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Table 1. Comparison with representative methods under different task settings. Our proposed *FaceXFormer* can perform various facial analysis tasks in a single model. FP - Face Parsing, LD - Landmarks Detection, HPE - Head Pose Estimation, Attr - Attributes Recognition, Age - Age, Gen - Gender, Race - Race Estimation, Exp - Facial Expression Recognition, FR - Face Recognition, and Vis - Face Visibility Prediction

and (3) simplifies deployment pipelines by reducing computational overhead and enabling faster inference.

Proposed *FaceXFormer* Architecture: To this end, we introduce *FaceXFormer*, an end-to-end unified model designed for ten different facial analysis tasks, as depicted in Figure 1. These tasks include face parsing, landmark detection, head pose estimation, attributes recognition, age/gender/race estimation, facial expression recognition, face recognition and face visibility prediction. *FaceXFormer* enables task unification by leveraging transformers and learnable tokens as its core components. Specifically, we employ a transformer-based encoder-decoder structure, where the encoder extracts hierarchical face representations and fuses them using a MLP fusion module. The fused features are then processed in the decoder, where each facial analysis task is represented by a unique learnable token, allowing for the simultaneous and effective processing of multiple tasks. In particular, we propose a lightweight decoder, *FaceX*, which processes both face and task tokens together using bi-directional cross-attention mechanism (Section 3.2), enabling the model to learn robust face representations that generalize across various tasks. The bi-directional cross-attention mechanism enables a 2-layer lightweight decoder, allowing the model to operate in real time. After modeling the intra-task and face-token relationships in the *FaceX* decoder, the task tokens are fed into a unified head, which converts these task tokens into corresponding task predictions.

Our extensive experiments demonstrate that *FaceXFormer* achieves state-of-the-art or competitive performance compared to specialized models and existing multi-task models across multiple benchmarks, while supporting more tasks than any previous multi-task model. Moreover, we show that our model effectively handles images "in the wild", demonstrating its robustness and generalizability across ten different tasks. This robustness is critical for real-world applications where uncontrolled conditions and diverse inputs are common. *FaceXFormer* achieves state-of-the-art performance at 33.21 FPS, representing a significant 69.44% speed boost over prior multi-task models, making it highly suitable for real-world applications.

In summary, our paper's contributions are as follows:

1. We introduce *FaceXFormer*, a unified transformer-based framework capable of simultaneously processing ten different facial analysis tasks, achieving real-time performance of 33.21 FPS.
2. We propose *FaceX*, a lightweight decoder that employs the proposed bi-directional cross-attention mechanism, enabling joint processing of face and task tokens.
3. We conduct extensive experiments and analyses to demonstrate that our approach achieves state-of-the-art performance with reduced inference time compared to specialized and multi-task models across multiple tasks.

2. Related Work

Facial analysis tasks: Facial analysis tasks involve face parsing [12, 33, 69, 130], landmarks detection [51, 61, 135], head pose estimation [13, 98, 124, 134], facial attributes recognition [63, 71, 88, 133], age/gender/race estimation [9, 42, 45, 48], facial expression recognition [46, 47], face recognition [16, 100] and face visibility prediction [41, 60]. These tasks hold significance in various applications such as face swapping [14, 67], face editing [137], de-occlusion [120], 3D face reconstruction [111], driver assistance [66], human-robot interaction [89], retail [1], face verification and identification [92, 93], image generation [118], image retrieval [122] and surveillance [25, 68]. Specialized models excel in their respective tasks but cannot be easily integrated with other tasks due to the need for extensive task-specific pre-processing [52, 135]. Generally, these models under-perform when applied to tasks beyond

	QFace [91]	Faceptor [74]	*FaceXFormer*
Tasks	4	7	10
Model Size	-	178.9 M	109.29 M
Face Pretrained	Yes	Yes	No
FPS (fp32)	-	14.30	33.21
Decoder	Query2Label [54]	Pixel Decoder & Transformer Decoder	FaceX
Decoder Layers	9	9	2

Table 2. Comparison of multi-task face analysis methods.

their specialization as their design is specific to their designated tasks. Some works [30, 62, 127, 129] perform multiple tasks simultaneously but utilize the additional tasks for guidance or auxiliary loss calculation to enhance the performance of the primary task.

Multi-task learning for face analysis: HyperFace [77] and AllinOne [78] are early convolution-based models that aim to perform multiple tasks. Recent multi-task frameworks, such as QFace [91] and Faceptor [74], are also inspired from DETR [11] and propose a unified model structure consisting of learnable tokens. However, these previous works differ from the proposed method in several key aspects, as summarized in Table 2. Specifically, QFace [91] employs a feature fusion module that uses stage embeddings to aggregate features from the encoder. Faceptor [74] introduces a layer-attention mechanism to fuse features from different encoder layers and incorporates two separate decoders. Both methods, employ a 9-layer transformer decoder, also Faceptor additionally includes a Pixel Decoder. These architectural components increase computational overhead, resulting in slower inference times. In contrast, *FaceXFormer* proposes a bi-directional cross-attention mechanism, which enables efficient task-specific feature extraction from face tokens resulting in a 2-layer lightweight decoder. This design choice is the primary reason for *FaceXFormer*'s superior speed and performance. Notably, unlike previous methods, *FaceXFormer* does not rely on face-specific pertaining backbone.

Unified transformer models: In recent years, the rise of transformers [20, 99] have paved the way for the unification of multiple tasks within a single architecture. Unified transformer architectures are being explored across various computer vision problems, including segmentation [49, 142], visual question answering (VQA) [102, 121], tracking [105, 136], detection [106]. While these models may not achieve state-of-the-art (SOTA) performance and may under-perform compared to specialized models on some tasks, they demonstrate competitive performance across a variety of tasks. Such unification efforts have led to the development of foundational models like SAM [40], CLIP [75], LLaMA [96], GPT-3 [7], DALL-E [76], etc. However, these models are computationally intensive and not suitable for facial analysis applications that require real-time performance. Motivated by this challenge, we propose *FaceXFormer*: the first lightweight, transformer-based model capable of performing multiple facial analysis tasks. It delivers real-time performance at 33.21 FPS and can be seamlessly integrated into existing systems providing additional annotations for the person of interest.

3. FaceXFormer

kiến trúc enc-dec tiêu chuẩn

In our framework, we follow a standard encoder-decoder structure as illustrated in Fig. 2. For an input face image $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$, we extract coarse to fine-grained multi-scale features $\mathbf{S}_i$, where i belongs to the i-th encoder output. To learn a unified face representation $\mathbf{F}$, these multi-scale features are then fused using a MLP-Fusion $\mathbf{M}$ module. Following fusion, we initialize a series of task-specific tokens $\mathbf{T} = \langle T_1, \ldots, T_n \rangle$, with each t_i representing a face task. Afterward, we initialize task tokens $\mathbf{T} = \langle T_1, \ldots, T_n \rangle$, where T_i denotes each task. Face tokens $\mathbf{F}$ and task tokens $\mathbf{T}$ are then processed by a lightweight Decoder $\mathbf{FXDec}$ where task tokens are attended with face tokens to learn relevant task representation.

$$\langle \hat{\mathbf{T}} \rangle = \mathbf{FXDec}\left(\langle \mathbf{F}, \mathbf{T} \rangle; \mathbf{S}_i\right)$$

Here, $\hat{\mathbf{T}}$ represents the output task tokens. These tokens are then fed into unified heads, where each task token is refined and passed to its respective task head for prediction.

3.1. Multi-scale Encoder

In the encoder, we employ a multi-scale encoding strategy to address the varying feature requirements intrinsic to each face analysis task. For instance, age estimation requires a global representation, while face parsing necessitates a fine-grained representation. Given an input image $\mathbf{I}$, it is processed through a set of encoder layers. For each encoder layer, the output captures information at varying levels of abstraction and detail, generating multi-scale features $\{\mathbf{S}_i\}_{i=1}^{n}$, where i ranges from 1 to 4. This results in a hierarchical structure of features, wherein each feature map $\mathbf{S}_i$ transitions from a coarse to a fine-grained representation suitable for diverse facial analysis tasks.

MLP-Fusion: Assigning each feature-map $\mathbf{S}_i$ to each face task is sub-optimal; rather, learning a unified face representation is more optimal and parameter-efficient. Following [116], we utilize a MLP-Fusion module $\mathbf{M}$ to generate a fused face representation from the multi-scale features $\{\mathbf{S}_i\}_{i=1}^{n}$. In this framework, each feature map $\mathbf{S}_i$ is initially passed through a separate MLP layer, standardizing the channel dimensions across scales to facilitate fusion. The transformed features are then concatenated and passed through a fusion MLP layer to aggregate a fused represen-

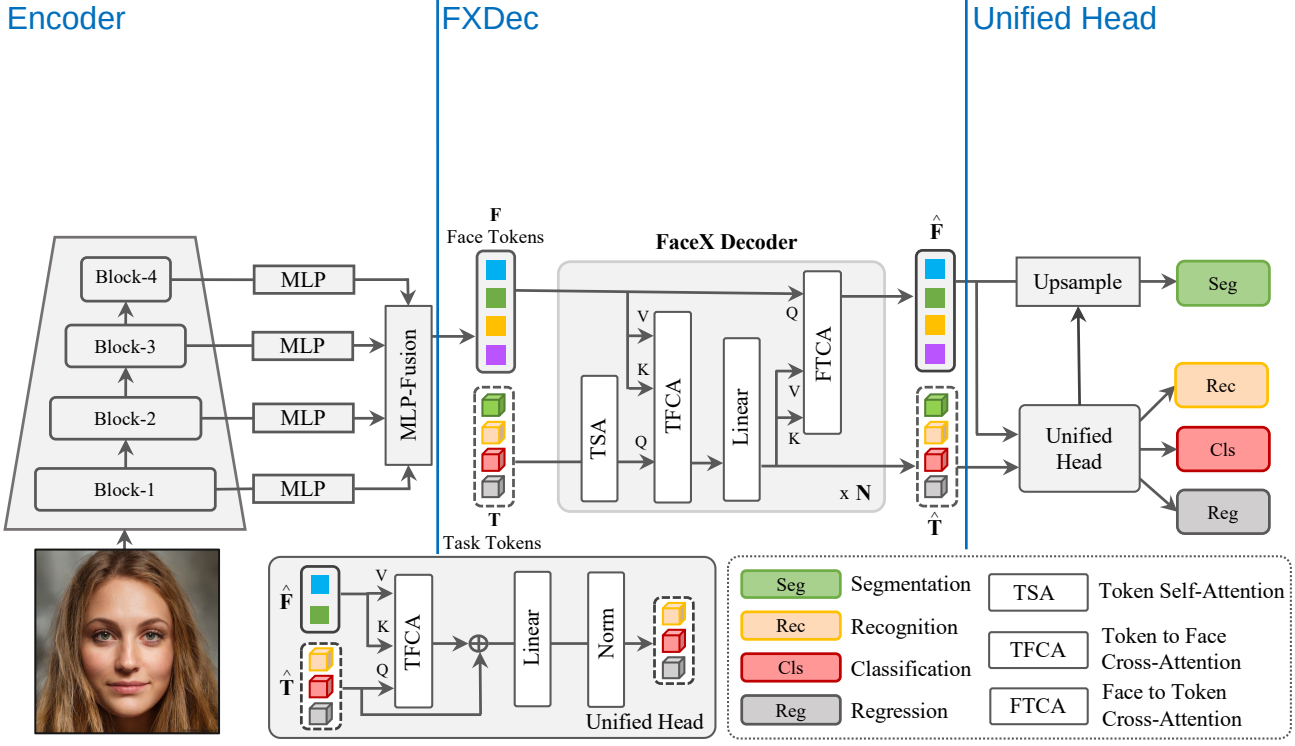

Figure 2. Overview of our proposed framework. The *FaceXFormer* employs an encoder-decoder architecture, extracting multi-scale features from the input face image $\mathbf{I}$, and fusing them into a unified representation $\mathbf{F}$ via MLP-Fusion. Task tokens $\mathbf{T}$ are processed alongside face representation $\mathbf{F}$ in the FaceX Decoder **FXDec**, resulting in refined task-specific tokens $\hat{\mathbf{T}}$. These refined tokens are then used for task-specific predictions by passing through the unified head. *FaceXFormer* performs ten tasks, including face parsing, landmark detection, head pose estimation, attribute prediction, age, gender, and race estimation, facial expression recognition, face recognition, and face visibility prediction, achieving state-of-the-art performance at a real-time FPS of 33.21.

tation $\mathbf{F}$ as follows:

$$\hat{\mathbf{S}}_i = \text{MLP}_{\text{proj}}(D_i, D_t)(\mathbf{S}_i), \forall i \in \{1, \ldots, n\},$$
$$\mathbf{F}_{\text{cat}} = \text{Concat}(\hat{\mathbf{S}}_1, \hat{\mathbf{S}}_2, \ldots, \hat{\mathbf{S}}_n),$$
$$\mathbf{F} = \text{MLP}_{\text{fusion}}(nD_t, D_t)(\mathbf{F}_{\text{cat}}),$$

where D_i and D_t are the multi-scale feature channel dimensions of $\mathbf{S}_i$ and the target channel dimension, respectively. The MLP-fusion design ensures minimal computational overhead (983k parameters) while maintaining the ability to perform efficient feature fusion, which is crucial for real-time application based face analysis tasks.

3.2. FaceX Decoder

Detection transformer (DETR) [11] employs object tokens to learn bounding box predictions for each object. Inspired by this approach, we introduce Task Tokens, whereby each task token is designed to learn specific facial tasks leveraging the fused face representation. However, existing decoders such as DETR [11] and Deformable-DETR [139] are computationally intensive, impacting runtime significantly. To address this, we propose *FaceX* (**FXDec**) a lightweight decoder designed to efficiently model the task tokens with face tokens. Specifically, each task token learns a task-related representation by interacting with other task tokens $\mathbf{T}$ and face tokens $\mathbf{F}$, enhancing the overall representation. The Lightweight Decoder comprises of three main components: 1) Task Self-Attention, 2) Task-to-Face Cross-Attention, and 3) Face-to-Task Cross-Attention as illustrated in Figure 2.

Task Self-Attention (TSA): The Task Self-Attention module is designed to refine the task-specific representations within the set of task tokens $\mathbf{T} = \langle T_1, \ldots, T_n \rangle$. Each task token T_i is an embedded representation that corresponds to a specific facial task. In TSA, each T_i is updated by attending to all other task tokens to capture task-specific interactions. Formally, the updated task token T_i' is computed as:

$$\mathbf{T}_i' = \text{SelfAttn}(\mathbf{Q} = T_i', \mathbf{K} = \mathbf{T}, \mathbf{V} = \mathbf{T}),$$

where Attention denotes the multi-headed self-attention mechanism, and $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ represent the queries, keys, and values, respectively. Therefore, TSA essentially helps the model to learn task-invariant representation.

Task-to-Face Cross-Attention (TFCA): The Task-to-Face Cross-Attention module allows each task token to interact with the fused face representation $\mathbf{F}$. This enables each task token to gather information relevant to its specific facial task from the fused face features. In this module, the fused face representation $\mathbf{F}$ acts as both key and value, while the task tokens serve as queries. The updated task token $\hat{T}_i$ is then computed as follows:

$$\hat{T}_i = \text{CrossAttn}(\mathbf{Q} = T_i', \mathbf{K} = \mathbf{F}, \mathbf{V} = \mathbf{F}),$$

where $\hat{\mathbf{T}} = \langle \hat{T}_1, \ldots, \hat{T}_n \rangle$ is the output task token. Thus, TFCA enables direct interaction between the task-specific

tokens and the compact facial features, facilitating task-focused feature extraction.

Face-to-Task Cross-Attention (FTCA): Conversely, the Face-to-Task Cross-Attention module is designed to refine the fused face representation $\mathbf{F}$ based on the information from the updated task tokens. This process aids in enhancing the face representation with task-specific details, thereby improving the extraction of overall fused representation. In FTCA, the set of updated task tokens $\mathbf{T}' = \{\mathbf{T}'_1, \mathbf{T}'_2, \ldots, \mathbf{T}'_m\}$ acts as both keys and values, while the fused face features $\mathbf{F}$ serve as queries. The refined face representation $\hat{\mathbf{F}}$ is computed as:

$$\hat{\mathbf{F}} = \text{CrossAttn}(\mathbf{Q} = \mathbf{F}, \mathbf{K} = \mathbf{T}', \mathbf{V} = \mathbf{T}').$$

Through this inverse attention mechanism, the face representation is augmented with critical task-specific details, enabling a robust approach towards facial task unification.

3.3. Unified-Head

In Unified-Head, the task tokens are processed to obtain corresponding task predictions. As shown in Figure 2, the output face tokens $\hat{\mathbf{F}}$ and task tokens $\hat{\mathbf{T}}$ are processed through a Task-to-Face Cross-Attention mechanism to obtain final refined features. Then, the output tokens are fed into their corresponding task heads. The task head for landmark detection is a hourglass network, for head pose estimation is a regression MLP, and for face recognition is PartialFC [4], while the tasks of age, gender and race estimation, facial expression recognition, face visibility prediction, and attributes prediction utilize classification MLPs. For face parsing, we leverage the output $\hat{\mathbf{F}}$ and process it through an upsampling layer, then perform a cross-product with the face parsing token to obtain a segmentation map. The number of tokens for segmentation corresponds to the total number of classes. For landmark prediction, it corresponds to the number of landmarks (i.e., 68). For head pose estimation, the number of tokens is 9, representing the 3×3 rotation matrix. For other tasks, one token is used for each.

Số lượng token

3.4. Multi-Task Training

We aim to train *FaceXFormer* for multiple facial analysis tasks simultaneously, however each task requires distinct and sometimes conflicting pre-processing steps. For instance, landmark detection typically requires keypoint alignment of faces, which contradicts the needs for head pose estimation, as it may eliminate the natural variability of headposes. Due to these reasons, integrating all tasks into a single model poses significant challenges. To address this, *FaceXFormer* incorporates task-specific tokens designed to extract task-specific features from the fused representation. These task tokens compel the backbone to learn a unified representation capable of supporting a broad spectrum of facial analysis tasks. We employ different loss functions for each task and combine them in a joint objective for training. The final loss function is given as:

$$L = \lambda_{seg}L_{seg} + \lambda_{lnd}L_{lnd} + \lambda_{hpe}L_{hpe} + \lambda_{attr}L_{attr} + \lambda_a L_a + \lambda_{g/r}L_{g/r} + \lambda_{exp}L_{exp} + \lambda_{fr}L_{fr} + \lambda_{vis}L_{vis}$$

where L_{seg} is the mean of dice loss [90] and Cross-Entropy (CE) loss for face parsing, L_{lnd} is STAR loss [135] for landmarks prediction, L_{hpe} is geodesic loss [124] for head pose estimation, $L_{g/r}$ is CE loss for gender/race estimation, L_a is mean of L1 loss and CE loss for age estimation, L_{exp} is CE loss for facial expression recognition, L_{fr} is ArcFace [16] loss for face recognition, and L_{attr} and L_{vis} are Binary Cross-Entropy with logits loss for attributes prediction and face visibility prediction respectively.

4. Experiments and Results

4.1. Datasets and Metrics

We perform co-training, where the model is simultaneously trained for multiple tasks using a total of 10 datasets with task-specific annotations. We conduct a comprehensive evaluation, comparing our approach with both task-specific and multi-task models. We present our results on the test sets according to the standard protocol for each task using the following datasets:

Train: *Face Farsing*: CelebAMaskHQ [44]; *Landmarks Detection*: 300W [82]; *Head Pose Estimation*: 300W-LP [138]; *Attributes Prediction*: CelebA [56]; *Facial Expression Recognition*: RAF-DB [47], AffectNet [64]; *Age/Gender/Race estimation*: UTKFace [128], FairFace [37]; *Face Recognition*: MS1MV3 [26]; *Visibility Prediction*: COFW [8].

Test: *Face Parsing*: CelebAMaskHQ [44]; *Landmarks Detection*: 300W [138], 300VW [86]; *Head Pose Estimation*: BIWI [21]; *Attributes Prediction*: CelebA [56], LFWA [110]; *Facial Expression Recognition*: RAF-DB [47]; *Age/Gender/Race Estimation*: UTKFace [128], FairFace [37]; *Face Recognition*: LFW [32], CFP-FP [84], AgeDB [65], CALFW [132], CPLFW [131] ; *Visibility Prediction*: COFW [8].

The evaluation metrics used are the F1-score for face parsing, Normalized Mean Error (NME) for landmark prediction, Mean Absolute Error (MAE) for head pose estimation and age estimation, accuracy for facial expression recognition, attributes prediction, gender estimation, race estimation, 1:1 verification accuracy for face recognition, and recall at 80% precision for face visibility prediction.

4.2. Implementation Details

We train our models using a distributed PyTorch setup on eight A6000 GPUs, each equipped with 48GB of memory. The models' backbones are initialized with ImageNet pretrained weights and processes input images at a resolution of 224×224. We employ the AdamW optimizer with a

Method	Input Res.	Skin	Hair	Nose	L-Eye	R-Eye	L-Brow	R-Brow	L-Lip	I-Mouth	U-Lip	Mean F1
Wei et al. [108]	512	96.4	91.1	91.9	87.1	85.0	80.8	82.5	91.0	90.6	87.9	88.43
EHANet [57]	512	96.0	93.9	93.7	86.2	86.5	83.2	83.1	90.3	93.8	88.6	89.53
EAGRNet [94]	473	96.2	94.9	**94.0**	88.6	89.0	85.7	85.2	**91.2**	**95.0**	88.9	90.87
AGRNet [95]	473	96.5	87.6	93.9	88.7	89.1	85.5	85.6	91.1	92.0	89.1	89.91
$\text{FaRL}_{\text{scratch}}$ [133]	512	96.2	94.9	93.8	89.0	89.0	85.3	85.4	90.0	91.7	88.1	90.34
DML-CSR [130]	473	95.7	94.5	93.9	89.4	89.6	85.5	85.7	91.0	91.8	89.1	90.62
FP-LIIF [83]	256	96.4	95.1	93.7	88.5	88.5	84.5	84.3	90.3	92.1	87.5	90.09
SwinFace [73]	×	×	×	×	×	×	×	×	×	×	×	×
QFace [91]	×	×	×	×	×	×	×	×	×	×	×	×
Faceptor [74]	512	**96.6**	**96.2**	93.9	89.4	89.1	**86.2**	**86.3**	90.6	91.6	89.0	90.89
FaceXFormer	224	96.4	95.7	93.8	**90.1**	**90.3**	86.0	85.9	90.6	92.1	**89.2**	**92.01**

Table 3. Performance comparison for face parsing on the CelebAMask-HQ dataset [44]. The symbol × indicates that the model does not perform the corresponding task. **Red** = First Best, Blue = Second Best. × indicates a model that doesn't perform the task.

Methods	Expression (RAF-DB)	Methods	Visibility (COFW)	Methods	Age (MAE) UTKFace
DLP-CNN [47]	80.89	RCPR [8]	40	OR-CNN [70]	5.74
gACNN [50]	85.07	Wu et al. [114]	44.43	Axel Berg et al. [5]	4.55
IPA2LT [123]	86.77	Wu et al. [113]	49.11	CORAL [9]	5.47
RAN [104]	86.90	ECT [125]	63.4	Gustafsson et al. [27]	4.65
CovPool [2]	87.00	3DDE [97]	63.89	R50-SORD [72]	4.36
SCN [103]	87.03	MNN [98]	72.12	VOLO-D1 [42]	4.23
DACL [22]	87.78			DLDL-v2 [24]	4.42
KTN [46]	88.07			MWR [87]	4.37
DMUE [85]	88.76				
SwinFace [73]	86.54	SwinFace [73]	×	SwinFace [73]	×
QFace [91]	**92.86**	QFace [91]	×	QFace [91]	×
Faceptor [74]	87.58	Faceptor [74]	×	Faceptor [74]	**4.10**
FaceXFormer	**88.24**	***FaceXFormer***	**72.56**	***FaceXFormer***	4.17

Table 4. Performance comparison on facial expression recognition, face visibility prediction and age estimation.

weight decay of $1e^{-5}$. All models are trained for 12 epochs with a batch size of 48 on each GPU, and an initial learning rate of $1e^{-4}$, which decays by a factor of 10 at the 6^{th} and 10^{th} epochs. We train the model for three additional epochs for some tasks. For data augmentation, we randomly apply Gaussian blur, grayscale conversion, gamma correction, occlusion, horizontal flipping, and affine transformations, such as rotation, translation and scaling. The number of *FaceX* decoder N is set to two. To ensure stable training across tasks when using multiple datasets of varying sample sizes, we equalize the representation of each task's samples in every batch through upsampling. Additional details on our implementation are provided in the Appendix F.

4.3. Main results

In Table 3, Table 4, Table 6, Table 5, we present a comparative analysis of *FaceXFormer* against recent methods across a variety of tasks. A key highlight of our work is its unique capability to deliver promising results across multiple tasks at real-time inference speed using a single unified model. Specifically, *FaceXFormer* achieves state-of-the-art performance in face parsing, with a mean F1 score of 92.01 on CelebAMaskHQ at a resolution of 224×224, which is half the input size required by other state-of-the-art methods. Furthermore, it demonstrates superior performance in head pose estimation and landmark detection, achieving a mean MAE of 3.52 and a mean NME of 4.67, respectively. Additionally, *FaceXFormer* provides a significant performance boost in attributes prediction and visibility prediction, achieving an accuracy of 91.83% on the CelebA dataset and 72.56% on COFW. It also performs competitively in age estimation, achieving the second-best score of 4.17, and achieves an accuracy of 88.24% in facial expression recognition. In face recognition, *FaceXFormer* outperforms Faceptor, achieving a mean accuracy of 95.94% compared to 95.28%. However, we observe that multi-task models generally underperform compared to specialized ones in this task. This can be attributed to conflicting training objectives, which force the model to learn identity-invariant features rather than identity-specific representations crucial for accurate recognition. The results on gender estimation across different race categories is shown in Table 8. We present additional cross-dataset results in Appendix D.

Recent models such as SwinFace [73], QFace [91], and Faceptor [74] also aim to unify multiple tasks but only address a subset of them. These approaches often exclude complex tasks such as segmentation, head pose estimation, and landmark prediction. Moreover, they rely on multiple decoders and computationally expensive attention mechanisms, adding to the overall computational overhead. In contrast, *FaceXFormer* seamlessly unifies these complex tasks using a lightweight decoder and achieves state-of-the-art performance across them at a real-time FPS of 33.21. It outperforms previous multi-task models in segmentation, head pose estimation, landmark prediction, attribute prediction, and face visibility prediction, while achieving the second-best performance in age estimation. In this work, we simultaneously train for ten heterogeneous tasks, posing a more formidable challenge than previous approaches. Despite this, *FaceXFormer* effectively handles multiple tasks, achieving SOTA or competitive performance in real time. This success can be attributed to the efficiency of the

Methods	Headpose (BIWI) Yaw	Pitch	Roll	MAE	Methods	Landmarks (300W) Full	Com	Chal	Methods	CelebA Acc.
HopeNet [81]	4.81	6.61	3.27	4.89	LAB [112]	3.49	2.98	5.19	PANDA-1 [126]	85.43
QuatNet [31]	5.49	4.01	2.94	4.15	Wing [23]	4.04	3.27	7.18	LNets+ANet [55]	87.33
FSA-Net [119]	4.27	5.49	2.93	4.14	DeCaFa [15]	3.39	2.93	5.26	SSP+SSG [35]	88.24
EVA-GCN [117]	6.01	4.78	2.98	3.98	HRNet [101]	3.32	2.87	5.15	MOON [80]	90.94
TriNet [10]	4.11	4.75	3.04	3.97	PicassoNet [109]	3.58	3.03	5.81	NSA [58]	90.61
img2pose [3]	4.56	**3.54**	3.24	3.78	AVS+SAN [19]	3.86	3.21	6.46	MCNN-AUX [29]	91.29
MNN [98]	3.98	4.61	**2.39**	3.66	LUVLi [41]	3.23	2.76	5.16	MCFA [141]	91.23
MFDNet [53]	**3.40**	4.68	2.77	3.62	HIH [43]	3.09	**2.65**	4.89	DMM-CNN [59]	91.70
TokenHPE [124]	3.95	4.51	2.71	3.72	PIPNet [34]	3.19	2.78	4.89	SSPL [88]	91.77
WHENet [134]	3.99	4.39	3.06	3.81	SLPT [115]	3.17	2.75	4.90	FaRL [133]	91.39
SwinFace [73]	×	×	×	×	SwinFace [73]	×	×	×	SwinFace [73]	91.38
QFace [91]	–	–	–	–	QFace [91]	×	×	×	QFace [91]	91.56
Faceptor [74]	×	×	×	×	Faceptor [74]	3.16	2.75	4.84	Faceptor [74]	91.39
FaceXFormer	3.91	3.97	2.67	**3.52**	***FaceXFormer***	**3.05**	2.66	**4.67**	***FaceXFormer***	**91.83**

Table 5. Performance comparison on headpose, landmark detection, and attribute recognition. The symbol × indicates that the model does not perform the corresponding task, while – denotes that results for this dataset are not provided. **Red** = First Best, Blue = Second Best.

Method	LFW	CFP-FP	AgeDB	CALFW	CPLFW	Mean
CosFace [100]	99.81	98.12	98.11	95.76	92.28	96.81
ArcFace [16]	99.83	98.27	98.28	95.45	92.08	96.78
VPL-ArcFace [17]	99.83	99.11	**98.60**	**96.12**	93.45	**97.42**
AdaFace [39]	99.83	**99.11**	98.17	96.02	**93.93**	97.41
SwinFace [73]	**99.87**	98.60	98.15	96.10	93.42	97.22
QFace [91]	×	×	×	×	×	×
Faceptor [74]	99.40	96.34	93.65	94.75	92.27	95.28
FaceXFormer	99.68	96.75	96.35	95.50	91.41	95.94

Table 6. Performance comparison for face recognition.

MLP Fusion	Self Attn.	Cross Attn.	Bi-dir. CA	HPE (MAE)	Lnd (NME)	Attr. (Acc.)	Age (MAE)
	✓		✓	3.70	4.70	91.21	4.21
✓	✓			17.16	31.49	79.90	16.15
✓	✓	✓		4.63	5.30	89.98	4.84
✓	✓		✓	**3.52**	**4.67**	**91.83**	**4.17**

Table 7. Impact of various components on performance.

proposed lightweight decoder, which employs a novel bi-directional cross-attention mechanism.

4.4. Qualitative "in-the-wild" results

In this section, we present the qualitative results of *FaceXFormer* on randomly selected "in-the-wild" images. We select four random images and showcase the results for face parsing, head pose estimation, landmarks prediction, age estimation, gender and race classification, and attributes prediction in Figure 3. Notably, the model successfully performs complex tasks such as face segmentation, head pose estimation, and landmark prediction, even when input samples exhibit extreme poses, occlusions, or blurring. Furthermore, *FaceXFormer* can be effectively used as a tool to generate multiple annotations for each image, making it valuable for various downstream tasks. These results highlight *FaceXFormer's* robust performance in challenging, real-world scenarios.

5. Ablation Study

In this section, we explore the impact of different components of *FaceXFormer* on performance. Additionally, we demonstrate that the proposed model exhibits minimal bias compared to other models by evaluating age and gender prediction across various demographics. Furthermore, we analyze the computational performance of different components of *FaceXFormer* and compare it with existing multi-task models. Additional ablation studies on the impact of using different backbones of varying sizes are provided in the Appendix C.

5.1. Impact of various components in FaceXFormer

To evaluate the contribution of each component in *FaceXFormer*, we conduct an ablation study focusing on the importance of specific design choices and their impact on performance across various tasks. The results of these experiments are summarized in Table 7. We observe that without MLP fusion (row 1), there is a drop in performance, highlighting the importance of of integrating multi-scale features to capture both global and local information essential for accurate predictions. The model performs extremely poorly (row 2) without cross-attention in the decoder, which is expected, as there is no interaction between face tokens and task tokens in this case. Introducing the proposed bi-directional cross-attention (row 4), which corresponds to *FaceXFormer*, in the decoder provides a significant boost compared to using standard cross-attention (row 3), yielding improvements of 1.11 MAE in head pose estimation, 0.63 NME in landmark detection, 1.85 accuracy points in attribute prediction, and 0.67 MAE in age estimation. These

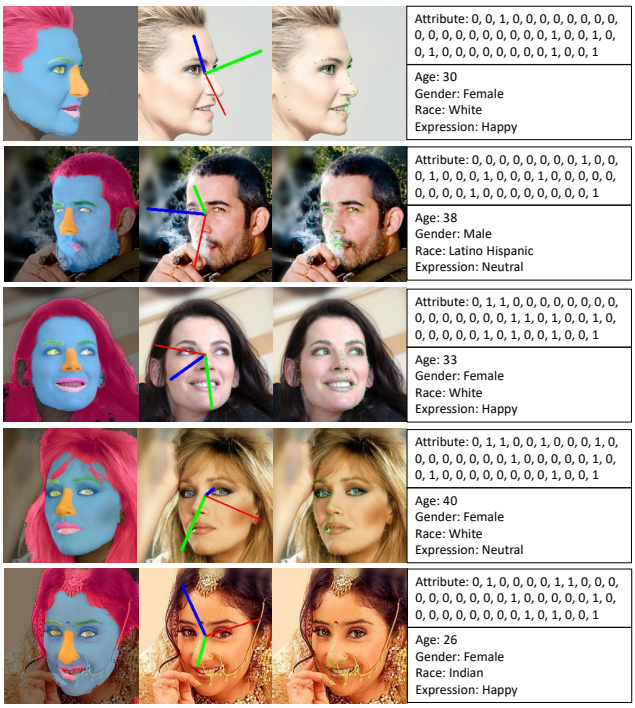

Figure 3. *FaceXFormer* predictions on "in-the-wild" images

results demonstrate the importance of MLP fusion and bi-directional cross-attention in the *FaceXFormer* architecture.

5.2. Bias Analysis and Ethical Considerations

In our work, we utilize 17 unique datasets for training and evaluation. We obtained these datasets following the procedures stated on their respective pages and signed the license agreements if and when necessary. As we train our models on multiple datasets designed for different tasks, the subjects across different age groups, genders, and races is not equal. This imbalance may introduce bias in the model. Therefore, we provide an analysis using the FairFace [36] dataset, which is balanced in terms of age, gender and race. We follow [75] and define the "Non-white" group to include multiple racial categories: "Black", "Indian", "East Asian", "Southeast Asian", "Middle Eastern" and "Latino". As can be seen from Table 8, *FaceXFormer* shows the smallest performance discrepancy across different racial groups and exhibits minimal bias compared to other models despite being trained on fewer data points. This can be attributed to race estimation being the task in co-training.

5.3. Computational Performance Analysis

We present a computational performance analysis of the proposed method compared to previous multi-task models in Table 9 to highlight its efficiency. *FaceXFormer* achieves the fastest inference speed among multi-task face analysis models, with an FPS of 33.2 (FP32) and 100.1 (FP16), outperforming previous multi-task model Faceptor [74]. This improvement is attributed to the proposed *FaceX* decoder, which employs a novel bi-directional cross-attention mechanism, enabling *FaceXFormer* to maintain only two decoder layers while ensuring effective face feature extraction. Moreover, *FaceXFormer* significantly reduces computational cost, requiring only 114 GFLOPs compared to 167 GFLOPs in Faceptor, leading to a substantial reduction in latency from 69.9 ms to 30.1 ms in FP32 and from 23.7 ms to 10.0 ms in FP16. With its reduced computational cost and faster inference, *FaceXFormer* achieves state-of-the-art performance across most tasks, demonstrating the effectiveness of its lightweight yet powerful design.

	Model	Data Points	White	Non-white	Average	Discrepancy
Age	FairFace	100K	60.05	60.63	60.52	0.58
	CLIP	400M	62.25	61.95	62.00	-0.30
	FaRL	20M	61.49	61.84	61.78	0.35
	FaceXFormer	5.9M	58.94	59.44	59.34	0.50
Gender	FairFace	100K	94.15	94.41	94.36	0.26
	CLIP	400M	94.87	95.78	95.61	0.91
	FaRL	20M	95.16	95.77	95.65	0.61
	FaceXFormer	5.9M	95.34	95.19	95.22	-0.09

Table 8. Age and gender accuracy w.r.t race groups on FairFace

Component	Model	FLOPs (GFLOPs)	Latency (ms)		FPS	
			FP32	FP16	FP32	FP16
Backbone	Faceptor [74]	64	19.1	6.5	52.3	153.8
	FaceXFormer	47	12.5	4.0	80.0	250.0
Decoder	Faceptor [74]	85	34.0	11.5	29.4	87.0
	FaceXFormer	55	13.6	4.5	73.5	222.2
Task Heads	Faceptor [74]	18	6.5	2.2	153.8	454.5
	FaceXFormer	12	4.0	1.5	250.0	666.7
Total Inference Time	Faceptor [74]	167	69.9	23.7	14.3	42.2
	FaceXFormer	114	30.1	10.0	**33.2**	**100.1**

Table 9. Computational performance: *FaceXFormer* vs Faceptor.

6. Conclusion

FaceXFormer introduces a novel end-to-end unified model that efficiently handles a wide range of facial analysis tasks in real-time. By adopting a transformer-based encoder-decoder architecture and representing each task as a learnable token, our approach seamlessly integrates multiple tasks within a single framework while maintaining minimal computational cost and fast inference times. The proposed lightweight decoder, *FaceX, incorporates a novel bi-directional cross-attention mechanism, enhancing the model's ability to learn robust and* generalized face representations across diverse tasks. Comprehensive experiments demonstrate that *FaceXFormer* achieves state-of-the-art performance across multiple facial analysis tasks, achieving a real-time FPS of 33.21. In broader applications, *FaceXFormer* can serve as an annotator for large-scale face datasets and can be integrated into existing facial analysis systems to provide extra information, making it a valuable tool for surveillance, subject analysis, and image retrieval.

Acknowledgment

This research is based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via [2022-21102100005]. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The US Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

[1] B Abirami, TS Subashini, and V Mahavaishnavi. Gender and age prediction from real time facial images using cnn. *Materials Today: Proceedings*, 33:4708–4712, 2020. 1, 2

[2] Dinesh Acharya, Zhiwu Huang, Danda Pani Paudel, and Luc Van Gool. Covariance pooling for facial expression recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 367–374, 2018. 6

[3] Vitor Albiero, Xingyu Chen, Xi Yin, Guan Pang, and Tal Hassner. img2pose: Face alignment and detection via 6dof, face pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7617–7627, 2021. 2, 7

[4] Xiang An, Xuhan Zhu, Yuan Gao, Yang Xiao, Yongle Zhao, Ziyong Feng, Lan Wu, Bin Qin, Ming Zhang, Debing Zhang, et al. Partial fc: Training 10 million identities on a single machine. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1445–1449, 2021. 5

[5] Axel Berg, Magnus Oskarsson, and Mark O'Connor. Deep ordinal regression with label diversity. In *2020 25th international conference on pattern recognition (ICPR)*, pages 2740–2747. IEEE, 2021. 6

[6] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators, 2024. 15

[7] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 3, 15

[8] Xavier P Burgos-Artizzu, Pietro Perona, and Piotr Dollár. Robust face landmark estimation under occlusion. In *Proceedings of the IEEE international conference on computer vision*, pages 1513–1520, 2013. 5, 6, 18

[9] Wenzhi Cao, Vahid Mirjalili, and Sebastian Raschka. Rank consistent ordinal regression for neural networks with application to age estimation. *Pattern Recognition Letters*, 140:325–331, 2020. 1, 2, 6

[10] Zhiwen Cao, Zongcheng Chu, Dongfang Liu, and Yingjie Chen. A vector-based representation to enhance head pose estimation. In *Proceedings of the IEEE/CVF Winter Conference on applications of computer vision*, pages 1188–1197, 2021. 2, 7

[11] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 3, 4

[12] Liang-Chieh Chen, Yi Yang, Jiang Wang, Wei Xu, and Alan L Yuille. Attention to scale: Scale-aware semantic image segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3640–3649, 2016. 2

[13] Alejandro Cobo, Roberto Valle, José M Buenaposada, and Luis Baumela. On the representation and methodology for wide and short range head pose estimation. *Pattern Recognition*, 149:110263, 2024. 1, 2

[14] Kaiwen Cui, Rongliang Wu, Fangneng Zhan, and Shijian Lu. Face transformer: Towards high fidelity and accurate face swapping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 668–677, 2023. 1, 2

[15] Arnaud Dapogny, Kevin Bailly, and Matthieu Cord. Decafa: Deep convolutional cascade for face alignment in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6893–6901, 2019. 7

[16] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019. 2, 5, 7

[17] Jiankang Deng, Jia Guo, Jing Yang, Alexandros Lattas, and Stefanos Zafeiriou. Variational prototype learning for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11906–11915, 2021. 7

[18] Xing Di, Yiyu Zheng, Xiaoming Liu, and Yu Cheng. Pros: Facial omni-representation learning via prototype-based self-distillation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6087–6098, 2024. 2

[19] Xuanyi Dong, Yan Yan, Wanli Ouyang, and Yi Yang. Style aggregated network for facial landmark detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 379–388, 2018. 7

[20] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[21] Gabriele Fanelli, Matthias Dantone, Juergen Gall, Andrea Fossati, and Luc Van Gool. Random forests for real time 3d face analysis. *International journal of computer vision*, 101:437–458, 2013. 5, 17

[22] Amir Hossein Farzaneh and Xiaojun Qi. Facial expression recognition in the wild via deep attentive center loss. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2402–2411, 2021. 6

[23] Zhen-Hua Feng, Josef Kittler, Muhammad Awais, Patrik Huber, and Xiao-Jun Wu. Wing loss for robust facial landmark localisation with convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2, 7

[24] Bin-Bin Gao, Xin-Xin Liu, Hong-Yu Zhou, Jianxin Wu, and Xin Geng. Learning expectation of label distribution for facial age and attractiveness estimation. *arXiv preprint arXiv:2007.01771*, 2020. 2, 6

[25] Asma El Kissi Ghalleb, Safa Boumaiza, and Najoua Essoukri Ben Amara. Demographic face profiling based on age, gender and race. In *2020 5th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP)*, pages 1–6. IEEE, 2020. 1, 2

[26] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, pages 87–102. Springer, 2016. 5, 17

[27] Fredrik K Gustafsson, Martin Danelljan, Goutam Bhat, and Thomas B Schön. Energy-based models for deep probabilistic regression. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*, pages 325–343. Springer, 2020. 6

[28] Hu Han, Anil K Jain, Fang Wang, Shiguang Shan, and Xilin Chen. Heterogeneous face attribute estimation: A deep multi-task learning approach. *IEEE transactions on pattern analysis and machine intelligence*, 40(11):2597–2609, 2017. 2

[29] Emily Hand and Rama Chellappa. Attributes for improved attributes: A multi-task network utilizing implicit and explicit relationships for facial attribute classification. In *Proceedings of the AAAI conference on artificial intelligence*, 2017. 7

[30] Hui-Lan Hsieh, Winston Hsu, and Yan-Ying Chen. Multi-task learning for face identification and attribute estimation. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2981–2985, 2017. 3

[31] Heng-Wei Hsu, Tung-Yu Wu, Sheng Wan, Wing Hung Wong, and Chen-Yi Lee. Quatnet: Quaternion-based head pose estimation with multiregression loss. *IEEE Transactions on Multimedia*, 21(4):1035–1046, 2018. 7

[32] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. 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. 5, 17

[33] Aaron S Jackson, Michel Valstar, and Georgios Tzimiropoulos. A cnn cascade for landmark guided semantic part segmentation. In *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14*, pages 143–155. Springer, 2016. 1, 2

[34] Haibo Jin, Shengcai Liao, and Ling Shao. Pixel-in-pixel net: Towards efficient facial landmark detection in the wild. *International Journal of Computer Vision*, 129(12):3174–3194, 2021. 7

[35] Mahdi M Kalayeh, Boqing Gong, and Mubarak Shah. Improving facial attribute prediction using semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6942–6950, 2017. 2, 7

[36] Kimmo Karkkainen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1548–1558, 2021. 2, 8

[37] Kimmo Karkkainen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1548–1558, 2021. 5, 17

[38] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 17

[39] Minchul Kim, Anil K Jain, and Xiaoming Liu. Adaface: Quality adaptive margin for face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18750–18759, 2022. 1, 2, 7

[40] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 3, 15

[41] Abhinav Kumar, Tim K Marks, Wenxuan Mou, Ye Wang, Michael Jones, Anoop Cherian, Toshiaki Koike-Akino, Xiaoming Liu, and Chen Feng. Luvli face alignment: Estimating landmarks' location, uncertainty, and visibility likelihood. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8236–8246, 2020. 1, 2, 7

[42] Maksim Kuprashevich and Irina Tolstykh. Mivolo: Multi-input transformer for age and gender estimation. *arXiv preprint arXiv:2307.04616*, 2023. 2, 6

[43] Xing Lan, Qinghao Hu, Qiang Chen, Jian Xue, and Jian Cheng. Hih: Towards more accurate face alignment via heatmap in heatmap. *arXiv preprint arXiv:2104.03100*, 2021. 7

[44] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 5, 6, 16

[45] Gil Levi and Tal Hassner. Age and gender classification using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 34–42, 2015. 2

[46] Hangyu Li, Nannan Wang, Xinpeng Ding, Xi Yang, and Xinbo Gao. Adaptively learning facial expression representation via cf labels and distillation. *IEEE Transactions on Image Processing*, 30:2016–2028, 2021. 2, 6

[47] Shan Li, Weihong Deng, and JunPing Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2852–2861, 2017. 2, 5, 6, 17

[48] Wanhua Li, Xiaoke Huang, Jiwen Lu, Jianjiang Feng, and Jie Zhou. Learning probabilistic ordinal embeddings for uncertainty-aware regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13896–13905, 2021. 1, 2

[49] Xiangtai Li, Haobo Yuan, Wei Li, Henghui Ding, Size Wu, Wenwei Zhang, Yining Li, Kai Chen, and Chen Change Loy. Omg-seg: Is one model good enough for all segmentation? *arXiv preprint arXiv:2401.10229*, 2024. 3

[50] Yong Li, Jiabei Zeng, Shiguang Shan, and Xilin Chen. Occlusion aware facial expression recognition using cnn with attention mechanism. *IEEE Transactions on Image Processing*, 28(5):2439–2450, 2018. 6

[51] Chunze Lin, Beier Zhu, Quan Wang, Renjie Liao, Chen Qian, Jiwen Lu, and Jie Zhou. Structure-coherent deep feature learning for robust face alignment. *IEEE Transactions on Image Processing*, 30:5313–5326, 2021. 1, 2

[52] Jinpeng Lin, Hao Yang, Dong Chen, Ming Zeng, Fang Wen, and Lu Yuan. Face parsing with roi tanh-warping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5654–5663, 2019. 2

[53] Hai Liu, Shuai Fang, Zhaoli Zhang, Duantengchuan Li, Ke Lin, and Jiazhang Wang. Mfdnet: Collaborative poses perception and matrix fisher distribution for head pose estimation. *IEEE Transactions on Multimedia*, 24:2449–2460, 2021. 7

[54] S Liu, L Zhang, X Yang, H Su, and J Zhu. Query2label: A simple transformer way to multi-label classification. arxiv 2021. *arXiv preprint arXiv:2107.10834*, 2021. 3

[55] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015. 7

[56] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2015. 5, 16, 17

[57] Ling Luo, Dingyu Xue, and Xinglong Feng. Ehanet: An effective hierarchical aggregation network for face parsing. *Applied Sciences*, 10(9):3135, 2020. 6

[58] Upal Mahbub, Sayantan Sarkar, and Rama Chellappa. Segment-based methods for facial attribute detection from partial faces. *IEEE Transactions on Affective Computing*, 11(4):601–613, 2018. 7

[59] Longbiao Mao, Yan Yan, Jing-Hao Xue, and Hanzi Wang. Deep multi-task multi-label cnn for effective facial attribute classification. *IEEE Transactions on Affective Computing*, 13(2):818–828, 2020. 7

[60] Chen Mi, Baoxi Yuan, Peng Ma, Yingxia Guo, Le Qi, Feng Wang, Wenbo Wu, and Lingling Wang. Visibility prediction based on landmark detection in foggy weather. In *2020 International Conference on Robots & Intelligent System (ICRIS)*, pages 134–137, 2020. 1, 2

[61] Paul Micaelli, Arash Vahdat, Hongxu Yin, Jan Kautz, and Pavlo Molchanov. Recurrence without recurrence: Stable video landmark detection with deep equilibrium models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22814–22825, 2023. 2

[62] Zuheng Ming, Junshi Xia, Muhammad Muzzamil Luqman, Jean-Christophe Burie, and Kaixing Zhao. Dynamic multi-task learning for face recognition with facial expression. *arXiv preprint arXiv:1911.03281*, 2019. 3

[63] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1979–1993, 2018. 1, 2

[64] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, 2017. 5, 17

[65] Stylianos Moschoglou, Athanasios Papaioannou, Christos Sagonas, Jiankang Deng, Irene Kotsia, and Stefanos Zafeiriou. Agedb: the first manually collected, in-the-wild age database. In *proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 51–59, 2017. 5, 17

[66] Erik Murphy-Chutorian, Anup Doshi, and Mohan Manubhai Trivedi. Head pose estimation for driver assistance systems: A robust algorithm and experimental evaluation. In *2007 IEEE intelligent transportation systems conference*, pages 709–714. IEEE, 2007. 2

[67] Kartik Narayan, Harsh Agarwal, Kartik Thakral, Surbhi Mittal, Mayank Vatsa, and Richa Singh. Df-platter: Multi-face heterogeneous deepfake dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9739–9748, 2023. 2

[68] Kartik Narayan, Nithin Gopalakrishnan Nair, Jennifer Xu, Rama Chellappa, and Vishal M Patel. Petalface: Parameter efficient transfer learning for low-resolution face recognition. *arXiv preprint arXiv:2412.07771*, 2024. 2

[69] Kartik Narayan, Vibashan VS, and Vishal M Patel. Segface: Face segmentation of long-tail classes. *arXiv preprint arXiv:2412.08647*, 2024. 2

[70] Zhenxing Niu, Mo Zhou, Le Wang, Xinbo Gao, and Gang Hua. Ordinal regression with multiple output cnn for age estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4920–4928, 2016. 6

[71] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European conference on computer vision*, pages 69–84. Springer, 2016. 1, 2

[72] Jakub Paplhám, Vojt Franc, et al. A call to reflect on evaluation practices for age estimation: Comparative analysis of

the state-of-the-art and a unified benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1196–1205, 2024. 6

[73] Lixiong Qin, Mei Wang, Chao Deng, Ke Wang, Xi Chen, Jiani Hu, and Weihong Deng. Swinface: a multi-task transformer for face recognition, expression recognition, age estimation and attribute estimation. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023. 2, 6, 7

[74] Lixiong Qin, Mei Wang, Xuannan Liu, Yuhang Zhang, Wei Deng, Xiaoshuai Song, Weiran Xu, and Weihong Deng. Faceptor: A generalist model for face perception. In *European Conference on Computer Vision*, pages 240–260. Springer, 2025. 2, 3, 6, 7, 8

[75] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3, 8

[76] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 3

[77] Rajeev Ranjan, Vishal M Patel, and Rama Chellappa. Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. *IEEE transactions on pattern analysis and machine intelligence*, 41(1):121–135, 2017. 2, 3

[78] Rajeev Ranjan, Swami Sankaranarayanan, Carlos D Castillo, and Rama Chellappa. An all-in-one convolutional neural network for face analysis. In *2017 12th IEEE international conference on automatic face & gesture recognition (FG 2017)*, pages 17–24. IEEE, 2017. 2, 3

[79] Karl Ricanek and Tamirat Tesafaye. Morph: A longitudinal image database of normal adult age-progression. In *7th international conference on automatic face and gesture recognition (FGR06)*, pages 341–345. IEEE, 2006. 17

[80] Ethan M Rudd, Manuel Günther, and Terrance E Boult. Moon: A mixed objective optimization network for the recognition of facial attributes. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, pages 19–35. Springer, 2016. 7

[81] Nataniel Ruiz, Eunji Chong, and James M Rehg. Fine-grained head pose estimation without keypoints. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 2074–2083, 2018. 7

[82] Christos Sagonas, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 397–403, 2013. 5, 16

[83] Mausoom Sarkar, Mayur Hemani, Rishabh Jain, Balaji Krishnamurthy, et al. Parameter efficient local implicit image function network for face segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20970–20980, 2023. 2, 6

[84] S. Sengupta, J.C. Cheng, C.D. Castillo, V.M. Patel, R. Chellappa, and D.W. Jacobs. Frontal to profile face verification in the wild. In *IEEE Conference on Applications of Computer Vision*, 2016. 5, 17

[85] Jiahui She, Yibo Hu, Hailin Shi, Jun Wang, Qiu Shen, and Tao Mei. Dive into ambiguity: Latent distribution mining and pairwise uncertainty estimation for facial expression recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6248–6257, 2021. 1, 2, 6

[86] Jie Shen, Stefanos Zafeiriou, Grigoris G Chrysos, Jean Kossaifi, Georgios Tzimiropoulos, and Maja Pantic. The first facial landmark tracking in-the-wild challenge: Benchmark and results. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 50–58, 2015. 5, 16

[87] Nyeong-Ho Shin, Seon-Ho Lee, and Chang-Su Kim. Moving window regression: A novel approach to ordinal regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18760–18769, 2022. 6

[88] Ying Shu, Yan Yan, Si Chen, Jing-Hao Xue, Chunhua Shen, and Hanzi Wang. Learning spatial-semantic relationship for facial attribute recognition with limited labeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11916–11925, 2021. 2, 7

[89] Dominykas Strazdas, Jan Hintz, and Ayoub Al-Hamadi. Robo-hud: Interaction concept for contactless operation of industrial cobotic systems. *Applied Sciences*, 11(12):5366, 2021. 2

[90] Carole H Sudre, Wenqi Li, Tom Vercauteren, Sebastien Ourselin, and M Jorge Cardoso. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 14, Proceedings 3*, pages 240–248. Springer, 2017. 5

[91] Haomiao Sun, Mingjie He, Shiguang Shan, Hu Han, and Xilin Chen. Task-adaptive q-face. *arXiv preprint arXiv:2405.09059*, 2024. 2, 3, 6, 7

[92] Yi Sun, Yuheng Chen, Xiaogang Wang, and Xiaoou Tang. Deep learning face representation by joint identification-verification. *Advances in neural information processing systems*, 27, 2014. 1, 2

[93] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1701–1708, 2014. 1, 2

[94] Gusi Te, Yinglu Liu, Wei Hu, Hailin Shi, and Tao Mei. Edge-aware graph representation learning and reasoning for face parsing. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*, pages 258–274. Springer, 2020. 6

[95] Gusi Te, Wei Hu, Yinglu Liu, Hailin Shi, and Tao Mei. Agrnet: Adaptive graph representation learning and reasoning for face parsing. *IEEE Transactions on Image Processing*, 30:8236–8250, 2021. 6

[96] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 3

[97] Roberto Valle, José M Buenaposada, Antonio Valdés, and Luis Baumela. Face alignment using a 3d deeply-initialized ensemble of regression trees. *Computer Vision and Image Understanding*, 189:102846, 2019. 2, 6

[98] Roberto Valle, José M Buenaposada, and Luis Baumela. Multi-task head pose estimation in-the-wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43 (8):2874–2881, 2020. 2, 6, 7

[99] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3

[100] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5265–5274, 2018. 2, 7

[101] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3349–3364, 2020. 2, 7

[102] Junke Wang, Dongdong Chen, Zuxuan Wu, Chong Luo, Luowei Zhou, Yucheng Zhao, Yujia Xie, Ce Liu, Yu-Gang Jiang, and Lu Yuan. Omnivl: One foundation model for image-language and video-language tasks. *Advances in neural information processing systems*, 35:5696–5710, 2022. 3

[103] Kai Wang, Xiaojiang Peng, Jianfei Yang, Shijian Lu, and Yu Qiao. Suppressing uncertainties for large-scale facial expression recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6897–6906, 2020. 6

[104] Kai Wang, Xiaojiang Peng, Jianfei Yang, Debin Meng, and Yu Qiao. Region attention networks for pose and occlusion robust facial expression recognition. *IEEE Transactions on Image Processing*, 29:4057–4069, 2020. 6

[105] Qianqian Wang, Yen-Yu Chang, Ruojin Cai, Zhengqi Li, Bharath Hariharan, Aleksander Holynski, and Noah Snavely. Tracking everything everywhere all at once. *arXiv preprint arXiv:2306.05422*, 2023. 3

[106] Zhenyu Wang, Yali Li, Xi Chen, Ser-Nam Lim, Antonio Torralba, Hengshuang Zhao, and Shengjin Wang. Detecting everything in the open world: Towards universal object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11433–11443, 2023. 3

[107] Zhen Wei, Yao Sun, Jinqiao Wang, Hanjiang Lai, and Si Liu. Learning adaptive receptive fields for deep image parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2434–2442, 2017. 1

[108] Zhen Wei, Si Liu, Yao Sun, and Hefei Ling. Accurate facial image parsing at real-time speed. *IEEE Transactions on Image Processing*, 28(9):4659–4670, 2019. 6

[109] Tiancheng Wen, Zhonggan Ding, Yongqiang Yao, Yaxiong Wang, and Xueming Qian. Picassonet: Searching adaptive architecture for efficient facial landmark localization. *IEEE Transactions on Neural Networks and Learning Systems*, 34(12):10516–10527, 2023. 7

[110] Lior Wolf, Tal Hassner, and Yaniv Taigman. Effective unconstrained face recognition by combining multiple descriptors and learned background statistics. *IEEE transactions on pattern analysis and machine intelligence*, 33(10): 1978–1990, 2010. 5, 17

[111] Erroll Wood, Tadas Baltrušaitis, Charlie Hewitt, Matthew Johnson, Jingjing Shen, Nikola Milosavljević, Daniel Wilde, Stephan Garbin, Toby Sharp, Ivan Stojiljković, et al. 3d face reconstruction with dense landmarks. In *European Conference on Computer Vision*, pages 160–177. Springer, 2022. 1, 2

[112] Wayne Wu, Chen Qian, Shuo Yang, Quan Wang, Yici Cai, and Qiang Zhou. Look at boundary: A boundary-aware face alignment algorithm. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2129–2138, 2018. 7

[113] Yue Wu and Qiang Ji. Robust facial landmark detection under significant head poses and occlusion. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3658–3666, 2015. 6

[114] Yue Wu, Chao Gou, and Qiang Ji. Simultaneous facial landmark detection, pose and deformation estimation under facial occlusion. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3471–3480, 2017. 6

[115] Jiahao Xia, Weiwei Qu, Wenjian Huang, Jianguo Zhang, Xi Wang, and Min Xu. Sparse local patch transformer for robust face alignment and landmarks inherent relation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4052–4061, 2022. 7

[116] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021. 3

[117] Miao Xin, Shentong Mo, and Yuanze Lin. Eva-gcn: Head pose estimation based on graph convolutional networks. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 1462–1471, 2021. 7

[118] Xinchen Yan, Jimei Yang, Kihyuk Sohn, and Honglak Lee. Attribute2image: Conditional image generation from visual attributes. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14,*

2016, Proceedings, Part IV 14, pages 776–791. Springer, 2016. 1, 2

[119] Tsun-Yi Yang, Yi-Ting Chen, Yen-Yu Lin, and Yung-Yu Chuang. Fsa-net: Learning fine-grained structure aggregation for head pose estimation from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1087–1096, 2019. 7

[120] Xiangnan Yin, Di Huang, Zehua Fu, Yunhong Wang, and Liming Chen. Segmentation-reconstruction-guided facial image de-occlusion. In *2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–8. IEEE, 2023. 1, 2

[121] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021. 3

[122] Alireza Zaeemzadeh, Shabnam Ghadar, Baldo Faieta, Zhe Lin, Nazanin Rahnavard, Mubarak Shah, and Ratheesh Kalarot. Face image retrieval with attribute manipulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12116–12125, 2021. 1, 2

[123] Jiabei Zeng, Shiguang Shan, and Xilin Chen. Facial expression recognition with inconsistently annotated datasets. In *Proceedings of the European conference on computer vision (ECCV)*, pages 222–237, 2018. 6

[124] Cheng Zhang, Hai Liu, Yongjian Deng, Bochen Xie, and Youfu Li. Tokenhpe: Learning orientation tokens for efficient head pose estimation via transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8897–8906, 2023. 2, 5, 7

[125] Hongwen Zhang, Qi Li, Zhenan Sun, and Yunfan Liu. Combining data-driven and model-driven methods for robust facial landmark detection. *IEEE Transactions on Information Forensics and Security*, 13(10):2409–2422, 2018. 6

[126] Ning Zhang, Manohar Paluri, Marc'Aurelio Ranzato, Trevor Darrell, and Lubomir Bourdev. Panda: Pose aligned networks for deep attribute modeling. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1637–1644, 2014. 7, 16

[127] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Facial landmark detection by deep multi-task learning. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI 13*, pages 94–108. Springer, 2014. 3

[128] Zhifei Zhang, Yang Song, and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5810–5818, 2017. 5, 17

[129] Rui Zhao, Tianshan Liu, Jun Xiao, Daniel PK Lun, and Kin-Man Lam. Deep multi-task learning for facial expression recognition and synthesis based on selective feature sharing. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 4412–4419. IEEE, 2021. 3

[130] Qingping Zheng, Jiankang Deng, Zheng Zhu, Ying Li, and Stefanos Zafeiriou. Decoupled multi-task learning with cyclical self-regulation for face parsing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4156–4165, 2022. 2, 6

[131] Tianyue Zheng and Weihong Deng. Cross-pose lfw: A database for studying cross-pose face recognition in unconstrained environments. *Beijing University of Posts and Telecommunications, Tech. Rep*, 5(7):5, 2018. 5, 17

[132] Tianyue Zheng, Weihong Deng, and Jiani Hu. Cross-age lfw: A database for studying cross-age face recognition in unconstrained environments. *arXiv preprint arXiv:1708.08197*, 2017. 5, 17

[133] Yinglin Zheng, Hao Yang, Ting Zhang, Jianmin Bao, Dongdong Chen, Yangyu Huang, Lu Yuan, Dong Chen, Ming Zeng, and Fang Wen. General facial representation learning in a visual-linguistic manner. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18697–18709, 2022. 2, 6, 7

[134] Yijun Zhou and James Gregson. Whenet: Real-time fine-grained estimation for wide range head pose. *arXiv preprint arXiv:2005.10353*, 2020. 1, 2, 7

[135] Zhenglin Zhou, Huaxia Li, Hong Liu, Nanyang Wang, Gang Yu, and Rongrong Ji. Star loss: Reducing semantic ambiguity in facial landmark detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15475–15484, 2023. 1, 2, 5, 15, 16

[136] Jiawen Zhu, Simiao Lai, Xin Chen, Dong Wang, and Huchuan Lu. Visual prompt multi-modal tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9516–9526, 2023. 3

[137] Peihao Zhu, Rameen Abdal, Yipeng Qin, and Peter Wonka. Sean: Image synthesis with semantic region-adaptive normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5104–5113, 2020. 1, 2

[138] Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan Z Li. Face alignment across large poses: A 3d solution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 146–155, 2016. 5, 17

[139] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 4

[140] Zheng Zhu, Guan Huang, Jiankang Deng, Yun Ye, Junjie Huang, Xinze Chen, Jiagang Zhu, Tian Yang, Jiwen Lu, Dalong Du, et al. Webface260m: A benchmark unveiling the power of million-scale deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10492–10502, 2021. 15

[141] Ni Zhuang, Yan Yan, Si Chen, and Hanzi Wang. Multi-task learning of cascaded cnn for facial attribute classification. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 2069–2074. IEEE, 2018. 2, 7

[142] Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Wang, Lijuan Wang, Jianfeng Gao, and Yong Jae Lee. Segment everything everywhere all at once. *Advances in Neural Information Processing Systems*, 36, 2024. 3

Appendix

A. Overview

As part of the Appendix, we present the following as an extension to the ones shown in the paper:

- Broader Impact (Section B)
- Ablation study (Section C)
- Cross-dataset Evaluation (Section D)
- In-the-wild Visualization (Section E)
- Dataset details (Section F)

B. Discussion

The world is moving towards transformers because of its potential to model large amounts of data [6, 7, 40]. Presently, the face community lacks large-scale annotated datasets to train foundational models capable of performing a wide spectrum of facial tasks. The largest clean dataset, WebFace42M [140], lacks annotations for face parsing, landmarks detection, headpose, expression, race and facial attributes. *FaceXFormer* can be used as an annotator for large-scale data, and can be continually improved through successive rounds of annotation and fine-tuning. We aim to propel the face community towards developing foundation models that cater to a variety of facial tasks. Additionally, *FaceXFormer* is a lightweight model that provides real-time output based on task-specific queries and can be appended with existing facial systems to provide additional information. It can also serve as a valuable tool in surveillance, and provide auxiliary information for subject analysis and image retrieval.

C. Ablation Study

To evaluate the impact of different backbones on performance and FPS, we conduct an ablation study comparing various backbone architectures in *FaceXFormer*. We categorize head pose estimation, landmark prediction, and age estimation as regression (Reg) tasks, while attribute prediction and facial expression recognition fall under classification (Cls). Additionally, face parsing is denoted as segmentation (Seg). The results of these experiments are summarized in Table C.1.

Backbone	Seg	Reg	Cls	FPS	Params
MobileNet	91.21	4.64	88.22	39.76	25.32
ResNet101	91.49	4.37	88.91	34.98	65.54
ConvNext-B	92.08	4.35	89.09	36.61	110.19
Swin-B	92.01	4.12	90.03	33.21	109.29

Table C.1. Effect of different backbones on performance and FPS.

From the results, we observe that ConvNeXt achieves the best performance in segmentation with an F1 score of 92.08%. The Swin Transformer backbone excels in both regression and classification tasks, with a mean error of 4.12 and a mean accuracy of 90.03%, respectively. In contrast, MobileNet demonstrates the lowest performance metrics, including an F1 score of 91.21% and a mean error of 4.64, highlighting its limitations in handling larger, more complex datasets due to its smaller receptive field compared to the Swin Transformer. The selection of the Swin Transformer as the backbone for *FaceXFormer* is driven by its superior scalability and global contextual understanding, both of which are essential for facial analysis tasks.

D. Cross-Dataset Evaluation

We conduct additional cross-dataset experiments to demonstrate the effectiveness of *FaceXFormer* in scenarios that closely resemble real-life conditions. These scenarios involve previously unseen, unconstrained face images characterized by significant variability in background, lighting, pose, and other factors. As shown in Table D.1, *FaceXFormer* outperforms the existing state-of-the-art model, STARLoss [135], on the 300VW dataset. This highlights *FaceXFormer*'s effectiveness in landmark detection under in-the-wild scenarios. The cross-dataset results support the rationale presented in this paper: the necessity of a unified facial analysis model capable of performing multiple tasks on unconstrained, in-the-wild faces, particularly for real-time applications. *FaceXFormer* addresses this gap and achieves state-of-the-art performance.

Method	300VW (Cat.A)	300VW (Cat.B)	300VW (Cat.C)	LFWA (Gender)
	NME	NME	NME	Acc.
PANDA [126]	-	-	-	92.00
STARLoss [135]	3.97	**3.39**	8.42	-
FaceXFormer	**3.90**	3.58	**6.75**	**92.74**

Table D.1. Cross Dataset evaluation of *FaceXFormer*.

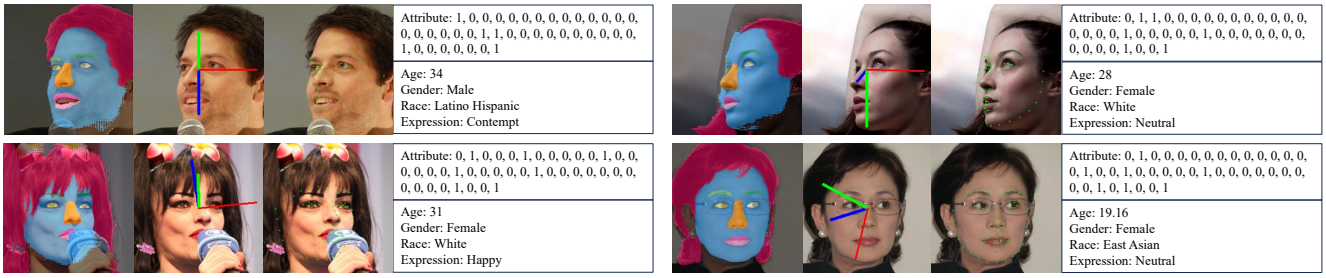

Figure E.1. Visualization of "in-the-wild" images for multiple tasks. Attributes represent the 40 binary attributes defined in the CelebA [56] dataset, indicating the presence (1) or absence (0) of specific facial attributes.

E. In-the-wild Visualization

We randomly selected images from the web and treated them as "in-the-wild" images. The qualitative results for all tasks are presented in Figure E.1. Our observations indicate that *FaceXFormer* produces promising results even in the presence of occlusions, extreme angles, and accessories.

F. Datasets and Implementation Details

In this section, we detail the dataset characteristics and the augmentations applied to each dataset during training. *FaceXFormer* is trained using multiple datasets, which have varying sample sizes. Datasets with a larger number of images may dominate the training process and create bias. To mitigate this, we employ upsampling to ensure that each batch is represented by samples from every dataset. This is achieved by repeating the samples of smaller datasets through upsampling and then randomly sampling images from the upsampled set. The model is trained for 12 epochs with a total batch size of 384 and an initial learning rate of $1e^{-4}$, which decays by a factor of 10 at the 6^{th} and 10^{th} epochs. We use the AdamW optimizer with a weight decay of $1e^{-5}$ for gradient updates.

F.1. Face Parsing

We use CelebAMask-HQ [44] for training and evaluation of *FaceXFormer*. CelebAMask-HQ contains 30,000 high-resolution face images annotated with 19 classes. The classes used for training and evaluation include: skin, face, nose, left eye, right eye, left eyebrow, right eyebrow, upper lip, mouth, and lower lip. During training, we resize the images to 224×224, before feeding them into the model.

F.2. Landmarks Detection

We utilize the 300W dataset [82] for the training and evaluation of *FaceXFormer*. The 300W dataset contains 3,148 images in its training set and 689 test images, which are categorized into three overlapping test sets: common (554 images), challenge (135 images), and full (689 images). It encompasses a wide variety of identities, expressions, illumination conditions, poses, occlusions, and face sizes. All images are annotated with 68 landmark points. For cross-dataset testing of multi-task methods, we employ the 300VW dataset [86]. This dataset provides three test categories: Category-A (well-lit conditions, comprising 31 videos with 62,135 frames), Category-B (mildly unconstrained conditions, consisting of 19 videos with 32,805 frames), and Category-C (challenging conditions, including 14 videos with 26,338 frames). We report the results for all three categories. During training, we apply various data augmentations such as random rotation ($\pm 18°$), random scaling ($\pm 10\%$),

random translation ($5\% \times 224$), random horizontal flip (50%), random gray (20%), random Gaussian blur (30%), random occlusion (40%) and random gamma adjustment(20%). Additionally, we align the images using five landmarks points.

F.3. Head Pose Estimation

We utilize the 300W-LP dataset [138], which contains approximately 122,000 samples. For performance evaluation, we use the BIWI dataset [21], comprising 15,678 images of 20 individuals (6 females and 14 males, with 4 individuals recorded twice). The head pose range spans approximately $\pm 75^{\circ}$ yaw and $\pm 60^{\circ}$ pitch. During training, we loosely crop the face images based on the landmarks and apply several augmentations, including random gray (10%), random Gaussian blur (10%), random resized crop ($80\% to 100\%$)and random gamma adjustment(10%).

F.4. Attributes Prediction

We utilize the CelebA [56] dataset for training and the LFWA [110] dataset for cross-dataset evaluation of multi-task methods. CelebA comprises 202,599 facial images, each annotated with 40 binary labels that indicate various facial attributes such as hair color, attractive, bangs, big lips, and more. LFWA consists of 13,143 facial images, annotated with the same set of facial attributes. During training, we apply several augmentations, including random rotation ($\pm 18^{\circ}$), random scaling ($\pm 10\%$), random translation ($1\% \times 224$), random horizontal flip (50%), random gray (10%), random Gaussian blur (10%), and random gamma adjustment(20%).

F.5. Age/Gender/Race Estimation

We utilize the FairFace [37] and UTKFace [128] datasets for training, and the FFHQ [38] dataset for cross-dataset testing. FairFace comprises 108,501 images, balanced across seven racial groups: White, Black, Indian, East Asian, Southeast Asian, Middle Eastern, and Latino. The UTKFace dataset contains 20,000 facial images annotated with age, gender, and race. In our work, we follow the 'race-4' annotation scheme, categorizing individuals into five racial labels: White, Black, Indian, Asian, and Others. Age annotations are categorized into decade bins: 0–9, 10–19, 20–29, 30–39, 40–49, 50–59, 60–69, and over 70. Gender is annotated with two labels: male and female. Additionally, we incorporate the MORPH-II dataset [79], which contains 55,134 facial images of 13,617 subjects aged between 16 and 77 years. This dataset provides annotations for age, gender, and race, with a predominance of male subjects and a significant representation of Black and White individuals. For age estimation tasks, we train on both UTKFace and MORPH-II datasets and evaluate our models on the MORPH-II dataset to assess performance. During training, we apply augmentations such as random rotation ($\pm 18^{\circ}$), random scaling ($\pm 10\%$), random translation ($1\% \times 224$), random horizontal flip (50%), random grayscale conversion (10%), random Gaussian blur (10%), and random gamma adjustment (10%).

F.6. Facial Expression Recognition

We utilize the RAF-DB [47] and AffectNet [64] datasets for training and RAF-DB [47] dataset for intra-dataset evaluation. RAF-DB is a facial expression dataset with approximately 30,000 images. The dataset includes variability in subjects' age, gender, ethnicity, head poses, lighting conditions, and occlusions (e.g., glasses, facial hair, or self-occlusion). RAF-DB provides annotations for seven basic emotions that are surprise, fear, disgust, happiness, sadness, anger, and neutral. AffectNet is one of the largest facial expression datasets with approximately 440,000 images that are manually annotated for the presence of eight discrete facial expressions: neutral, happy, angry, sad, fear, surprise, disgust, contempt. During training, we apply augmentations such as random rotation ($\pm 18^{\circ}$), random scaling ($\pm 10\%$), random translation ($1\% \times 224$), random horizontal flip (50%), random grayscale conversion (10%), random Gaussian blur (10%), random color jitter (10%), and random gamma adjustment (10%).

F.7. Face Recognition

We utilize the MS1MV3 [26] dataset for training our face recognition models and evaluate their performance using LFW [32], CFP-FP [84], AgeDB [65], CALFW [132], and CPLFW [131]. MS1M-V3 is a cleaned version of the MS-Celeb-1M dataset, containing approximately 5.1 million images of 93,000 identities, making it suitable for large-scale face recognition training. For evaluation, LFW (Labeled Faces in the Wild) consists of 13,233 images of 5,749 individuals and is designed for face verification in unconstrained environments. CFP-FP (Celebrities in Frontal-Profile) contains 7,000 images of 500 subjects and focuses on frontal-to-profile face verification. AgeDB provides 12,240 images of 440 subjects, spanning ages from 3 to 101 years, to evaluate age-invariant face verification. CALFW (Cross-Age LFW) introduces age variations by selecting positive pairs with large age gaps and negative pairs with similar age, race, and gender attributes. CPLFW (Cross-Pose LFW) is derived from LFW and emphasizes pose variation by selecting positive pairs with different poses and negative pairs with

similar pose, race, and gender. These datasets collectively cover diverse challenges, including pose, age, and other variations, enabling a comprehensive evaluation of face recognition models. We do not apply any augmentations during training but preprocess images by aligning them based on five facial keypoints before feeding them into the model.

F.8. Visibility Prediction

We utilize the COFW [8] dataset, which is annotated with 29 landmarks for landmarks visibility prediction. Each landmark is associated with 29 binary labels that indicate its visibility. We loosely crop the faces and apply augmentations, including random horizontal flip (50%), random gray (10%), random Gaussian blur (10%), and random gamma adjustment(10%).