IACT: Intensive Attention in Convolution-Transformer Network for Facial Landmark Localization

Zhanyu Gao*, Kai Chen* and Dahai Yu
TCL Corporate Research (HK) Co., Ltd, China

Keywords: Transformer, Attention, Convolution.

Abstract: Recently, the facial landmarks localization tasks based on deep learning methods have achieved promising results, but they ignore the global context information and long-range relationship among the landmarks. To address this issue, we propose a parallel multi-branch architecture combining convolutional blocks and transformer layer for facial landmarks localization named Intensive Attention in the Convolutional Vision Transformer Network (IACT), which has the advantages of capturing detailed features and gathering global dynamic attention weights. To further improve the performance, the Intensive Attention mechanism is incorporated with the Convolution-Transformer Network, which includes Multi-head Spatial attention, Feature attention, the Channel attention. In addition, we present a novel loss function named *Smooth Wing Loss* that fills the gap in the gradient discontinuity of the Adaptive Wing loss, resulting in better convergence. Our IACT can achieve state-of-the-art performance on WFLW, 300W, and COFW datasets with 4.04, 2.82 and 3.12 in Normalized Mean Error.

1 INTRODUCTION

The facial landmarks localization task is to establish coordinate information around the facial features and the contour of faces, which mainly used for facial expression recognition (Savchenko, 2021), fatigue detection (Parekh et al., 2020).

According to (Xia et al., 2022), the global contextual information and the long-range dependencies between the landmarks are crucial to facial landmarks localization tasks. Although Heatmap-based regression methods (Wan et al., 2020) provide an excellent solution for facial images in extreme conditions, they based on convolutional neural network (CNN) cannot model the global contextual information and long distance relations due to local receptive fields. Heatmap-based regression methods encode the coordinate information of the ground truth heatmap through Gaussian distribution and decode the highest intensity in the heatmaps of the coordinate information.

Facial landmarks localization tasks also have other regression methods - the Coordinate direct regression methods (Dong et al., 2020; Dong et al., 2018; Guo et al., 2019; Feng et al., 2018) that map feature maps to the landmarks via fully-connect layers (FC layers). However, directly projecting feature maps into FC layers leads to losing local feature information and not robustness under challenging conditions.

Transformer used to have achieved impressive success in the natural language processing (NLP) field. Recently, researchers are focusing on introduce transformer to computer vision tasks since the proposed of Vision in Transformer (ViT) [cite] first indicate that transformer can perform comparable result with those convolution-based network [cite]. However, directly apply the ViT architecture to different tasks is not able to achieve considerable result and need to make appropriate modifications.

As shown in Fig.1, we propose parallel multi-branch architecture consisting of convolution blocks and transformer blocks to overcome such barriers and apply the transformer in the facial landmarks localization. The convolution block is designed for extracting pixel-level information and downsampling the feature map size for computational efficiency and parameter-friendly. The transformer block is design for global information interaction through attention module.

To further improve the effectiveness and robustness of the parallel architecture, the Intensive Attention mechanism is introduced into the structure, which contains three parts: the multi-head Spatial attention, the Feature attention, and the Channel attention.The Intensive Attention mechanism improves the detection accuracy of facial landmarks localiza-

*Authors contributed equally

Gao, Z., Chen, K. and Yu, D.
IACT: Intensive Attention in Convolution-Transformer Network for Facial Landmark Localization.
DOI: 10.5220/0011740900003417
In *Proceedings of the 18th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2023) - Volume 4: VISAPP*, pages 401-410
ISBN: 978-989-758-634-7; ISSN: 2184-4321

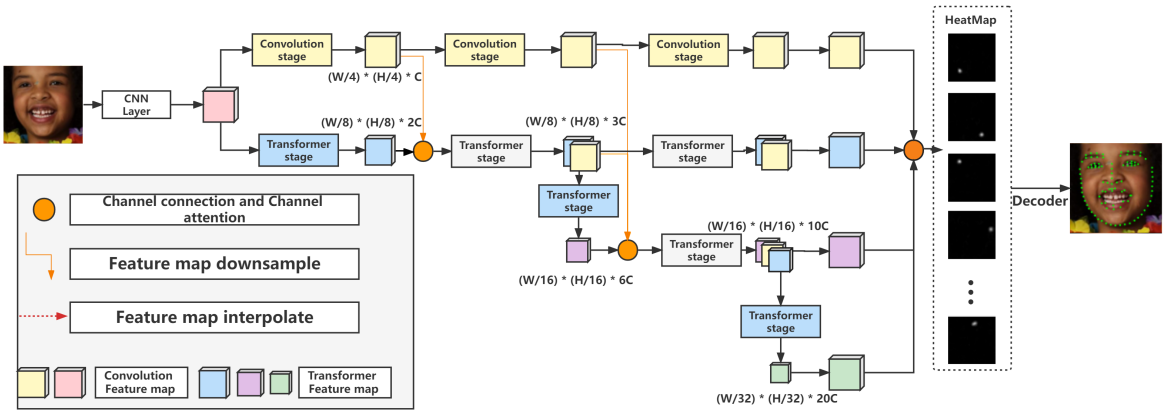


Figure 1: The whole structure of IACT. The parallel multi-branch structure could gain multi-model feature maps with various resolutions. The Transformer stage(the white block) does not downsample the feature maps; it is only used for further integrating multi-modal feature information.

tion and makes IACT comparable to other mainstream methods.

The adaptive wing loss (Wang et al., 2019) performed impressively in the heatmap-based regression tasks. However, the Adaptive Wing loss has an unavoidable problem in that the gradient value of it discontinues around the zero point ($y - \hat{y} = 0$), which will have an impact on training convergence. We proposed a novel loss function called Smooth Wing Loss, it can solve the problem by using a smoother curve. We conduct a series of experiments on the heatmap regression task to compare the training convergence of Adaptive Wing Loss and Smooth Wing Loss and thoroughly verify that Smooth Wing loss converges faster than Adaptive Wing loss. Additionally, Smooth wing loss could further converge at the later training stage.

Our method can achieve state-of-the-art results on three different facial landmarks datasets, 300W (Sagonas et al., 2013a), COFW (Burgos-Artizzu et al., 2013) and WFLW (Wu et al., 2018), which is 2.82, 3.12 and 4.02 compare to the methods (Kumar et al., 2020; Guo et al., 2019; Huang et al., 2021; Xia et al., 2022).

The main contributionscan be summarized as follows:

- We introduce a parallel multi-branch framework, IACT, which combines convolution stages and transformer stages. The proposed IACT has the advantages of capturing detailed features and spatial subsampling from convolution and the abilities of dynamic attention and global attention weights fusion from transformer.
- To further improve the performance, we propose the Intensive Attention mechanism, which can be divided into Multi-head Spatial attention, Feature attention, and Channel attention.
- We present a new loss function, Smooth Wing Loss. Through the detailed experiments, we demonstrate that Smooth Wing loss not only can converge faster than Adaptive Wing loss but also can further converge at the later training stage.

1.1 Related Work

With the development of CNN, face alignment methods have achieved exciting results. The mainstream methods (Wang et al., 2020; Guo et al., 2019; Kumar et al., 2020; Huang et al., 2021; Wang et al., 2019; Huang et al., 2021) can be divided into the Heatmap regression method and Coordinate regression method. Expecting the CNN-based method, the regression method based on vision transformer (Xia et al., 2022) recently also achieved impressive performance.

1.2 Coordinate Regression Method

Coordinate Regression Method (Feng et al., 2018; Zhang et al., 2016) utilize the fully connected layer to project feature maps into the landmarks directly. To acquire more accurate coordinate information, diverse cascaded networks (Trigeorgis et al., 2016) and recurrent networks (Xiao et al., 2016) are utilized to achieve face alignment with multi-stages. In order to improve the robustness of the facial landmarks tasks and solve the Intra-Dataset Variation and Inter-Dataset Variation, Wu et al. (Wu and Yang, 2017) introduce Deep Variation Leveraging Networks (DVLN) using two strong coupling sub-networks. In addition, coordinate regression methods have trouble detecting facial landmarks if they are in extreme con-

ditions. To address this problem, (Guo et al., 2019) proposed a practical face detection network that uses an auxiliary network jointly to predicate the landmarks. (Feng et al., 2018) proposed the Wing loss function that can surpass the performance of L2 loss and demonstrated its effectiveness through experiments.

1.3 Heatmap Regression Method

The Heatmap regression method (Dapogny et al., 2019a; Deng et al., 2019; Kowalski et al., 2017; Jin et al., 2021) maps the input image to probability heatmaps through the network, and the maximum point in each heatmap represents the probability of the coordinate location. (Kowalski et al., 2017) combined heatmap and landmarks for the first time, achieving impressive results; (Kumar et al., 2020) firstly estimates the uncertainty of the predicted locations, and they proposed a novel framework for predicting landmarks by associating uncertainties of these predicated locations and landmark visibilities. Wu et al. (Wu et al., 2018) used the boundary information of the face to locate facial landmarks and proposed the WFLW dataset, which includes facial images affected by different factors. Same as (Wu et al., 2018), (Wang et al., 2019) also used boundary information to enhance the fitting ability of the network and proposed a novel loss function, named the Adaptive wing loss, for heatmap regression tasks which can adapt its shape to different types of ground truth heatmap pixels. But the Adaptive wing loss ignores the issue that the gradient can not smooth and continue at every pixel, and we propose the Smooth Wing loss to fill this gap.

1.4 Vision Transformer

Transformers have achieved excellent results in the NLP field. In the computer vision field dominated by convolution networks, some researchers have noticed that convolutional networks are problematic in modeling the relationship between long-distance pixels, motivating more and more researchers have invested in the work of applying transformers to the CV field. VIT (Dosovitskiy et al., 2020) divides the input images into patches, and maps them into a d-dimension vector as the input of the self-attention layer to model the long-distance relationship between patches and patches, and successfully surpassed other state-of-the-art methods based on CNN by pre-training on large-scale datasets. Swin-Transformer (Liu et al., 2021) uses shift windows to limit self-attention computation to non-overlapping local windows, which can reduce parameters and complexities meanwhile improving performance. SLPT (Xia et al., 2022) proposed a sparse local patch transformer, which can generate the representations of the landmarks from each local patch and learn the inherent relations between the landmarks.

Convolution and transformer have their own merits, so we do not need to argue about which is better or worse, what we need to do is that taking advantage of them to achieve better performance. In this paper, we propose the Intensive Attention in the Convolution-Transformer Network, it has the strengths of both convolution and transformer.

2 METHOD

As shown in Fig.1, IACT has a parallel multi-branch structure containing two stages: Convolution and Transformer stages. IACT can capture detailed features and gather global dynamic attention weights on account of the parallel multi-branch convolution-transformer structure. The Intensive Attention improves the performance of our parallel multi-branch architecture, which includes the Multi-head Spatial Attention, Feature Attention, Channel attention(Woo et al., 2018). We propose the Multi-head Spatial attention in the transformer stages, which uses the ability of sparse sampling to focus attention weights on meaningful locations quickly. We insert the Feature attention in the whole structure; as for it, we introduce the exchanging feature information operation to gain multi-model feature maps for better representations, and we add the Channel attention to model the importance of these individual multi-model features.

2.1 The Multi-head Spatial Attention in Transformer Stages

As shown in Fig.2, the Transformer stage contains three parts: convolution embedding, Multi-head Spatial attention, and Feed-forward Network(FFN).

2.1.1 Convolution Embedding

VIT (Dosovitskiy et al., 2020) divides the input image or 2D feature map $I \in \mathbb{R}^{H\times W\times C}$ into patches of size $P_h \times P_w$ directly, P_h and P_w represents the weigh and the height of each patch. For the same input, different from VIT (Dosovitskiy et al., 2020), we want to take full advantage of CNN to extract more low-level features so that we use convolution to down-sample the input image to gain the feature map, $F \in \mathbb{R}^{\frac{H}{2}\times\frac{W}{2}\times 2C}$, then we flatten the feature map to get the 1D patches, these patches are treated as tokens, whose length is

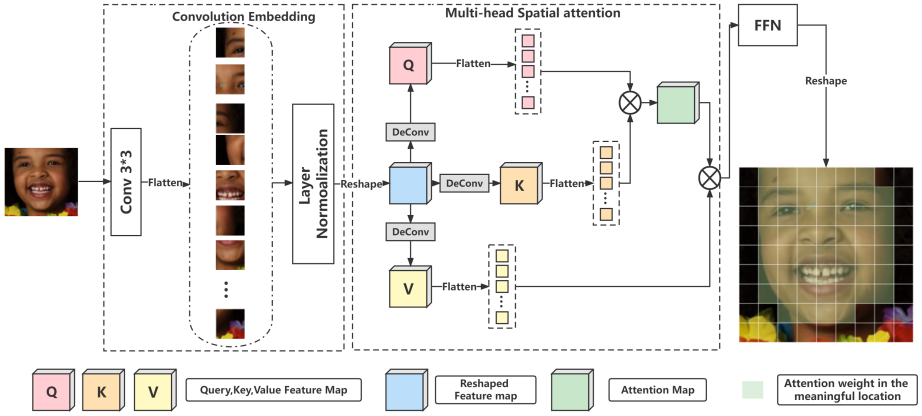


Figure 2: The Multi-head Spatial Attention in the Transformer stage. The flatten operation means that the feature maps are directly flattened into 1×1 patches, the reshape operation indicates that the 1-D vectors are reshaped back into 2-D feature maps. The DeConv represents Deformable Convolution.

$\frac{H}{2} \times \frac{W}{2}$ and dimension is $2C$, then sending them into the Multi-head Spatial attention layer.

We perform the downsample operation in the convolution embedding stages, generating multi-scale feature maps while increasing feature dimension. The convolution embedding process formula is as follows, *LN* represents the Layer Normalization operation.

$$ConvEmbedding(I) = LN(Flatten(Conv(I))) \quad (1)$$

2.1.2 Multi-Head Spatial Attention(MSA)

Different from the Multi-head self-attention structure in the traditional transformer, when calculating the value of Q(query), K(key), and V(value), we implement deformable convolution (Dai et al., 2017) projection instead of the linear projection. In the deformable convolution, it has an offset that can be learned from input data, and then the offset is added to each point of the receptive field, then the receptive field is no longer a regular square shape. The sparse sampling ability and the characteristic of data-dependent from deformable convolution are vital for MSA.

We reshape the 1D vectors gained from convolution embedding layers into 2D feature maps. Suppose a reshaped 2D feature map $I \in \mathbb{R}^{H \times W \times C}$ is fed into the Multi-head Spatial attention. The attention layers have h heads, we get Q (query) feature maps, K (key) feature maps, and V (value) feature maps via deformable convolution projection, and we flatten these 2D feature maps into the 1D vectors $Q_v, K_v, V_v \in \mathbb{R}^{(H \times W) \times C}$, the flatten operation is the same as the embedding phase. C is the number of channels of the feature map, H and W are the height and width of the feature map, respectively, these vectors are equally divided into h sequences $Q_h, K_h, V_h \in \mathbb{R}^{h \times (H \times W) \times \frac{C}{h}}$.

Then we perform matrix computation, soft-max function, and the linear projection on Q_h, K_h, V_h, to obtain vectors containing global attention weights information, eventually giving them into the FFN.

Assuming I is the reshaped 2D feature map, the following equation could summarize the process.

$$Q_h, K_h, V_h = Flatten(I \times (W_h^q, W_h^k, W_h^v)) \quad (2)$$

$$Attention_h(p) = Softmax(\frac{Q_h \times K_h}{\sqrt{C_h}})(V_h) \quad (3)$$

$$MSA(p) = [Attention_1(p); \ldots : Attention_h(p)]W_p \quad (4)$$

Among them, $C_h = \frac{C}{h}$ and $W_h^q, W_h^k, W_h^v \in \mathbb{R}^{C_h \times C_h}$ represents learnable matrices from deformable convolution, $W_p \in \mathbb{R}^{C \times C}$ is the learnable matrix from the linear projection layer.

2.1.3 Feed-Forward Network

The Feed-forward Network(FFN) consists of two linear projection and a non-linear activation:

$$FFN(x) = \sigma(xW_1 + b_1)W_2 + b_2 \quad (5)$$

Where $W_1 \in \mathbb{R}^{C \times K}$ is the weight of the first layer, which projects tokens into a higher dimension K. $W_2 \in \mathbb{R}^{K \times C}$ represents the weight for the second layer, projecting tokens into an original dimension C, b_1 and b_2 represents bias, and $\sigma()$ is the activation of GELU. At the end of FFN, we reshape the 1-D vectors into 2-D feature maps for exchanging information in the following stages.

2.2 The Feature Attention and the Channel Attention in the Parallel Multi-Branch Network Structure

The overall structure of IACT is shown in Fig.1. We construct a parallel multi-branch network structure, modeling the global attention weights information in the Transformer stage, and extracting rich detailed features in the Convolution stages. **The Convolution stages** contain four basic convolutional modules. Each module uses two stride-1 Conv-BN-Relu blocks to extract C-channel features with more information maintained. We also add the Feature Attention mechanism and the Channel attention mechanism into the parallel structure, which can help our IACT gain multi-model feature maps via exchanging feature information and modeling the importance of multi-model features.

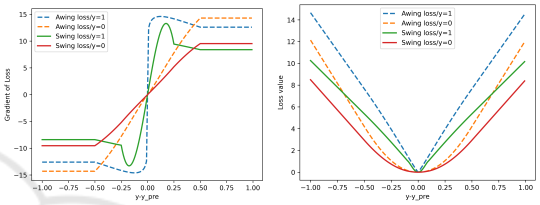
Feature Attention: in the initial stage, we use CNN Layer to extract low-level features from the input image and then feed these feature maps into the parallel structure. To distinguish the feature maps generated by different stages, we name the feature map generated by the Convolution stage as the C-feature map and the feature map generated by the Transformer stage as the T-feature map. Due to the T-feature maps being downsampled in the embedding layer, IACT can gain various-resolution feature maps. The C-feature map benefits from the convolutional network and has rich detailed features, while the T-feature map contains the global attention weights information.

Channel Attention: For better feature representations, we introduce the Feature attention and the Channel attention that exists at the end of each stage, we downsample the higher-resolution feature maps to match the lowest-resolution feature maps, then apply feature fusion operation on these various-resolution feature maps via channel connection, generating multi-model feature maps. Subsequently, the Channel attention is used to model the inherent relations between each multi-model feature. Finally, we feed these multi-model feature maps into the Transformer stage to further strengthen the global contextual information fusion.

Repeating Transformer stages and Convolution stages three times, IACT can obtains four multi-model feature maps with different resolutions, where the C-feature maps are always the highest resolution from the Convolution stage, which contains rich detailed features information, and the other three feature maps are obtained by exchanging feature information between the Convolution stages and the Transformer stages, which contains multi-model feature information.

We interpolate these three feature maps with lower resolutions to restore them to the exact resolution as the C-feature map, performing channel connection operations on them. We use convolution layers to map these feature maps to heatmaps, the number of channels is the same as the number of landmarks in each facial image. We can get the predicted facial landmarks by decoding these heatmaps, which contain coordinate information. We refer to the method (Yu and Tao, 2021), encoding the fractional part of the numerical coordinates via the random rounding method and decoding the maximum activation points in the heatmaps of the numerical coordinates.

2.3 Smooth Wing Loss



(a) Gradient of losses (b) Loss functions

Figure 3: Adaptive Wing Loss and Smooth Wing loss. When y=1, Fig.3(a) shows that the gradient of the Smooth Wing loss is close to zero around the point ($y-\hat{y}=0$), while the Adaptive Wing loss tends to different values.

From Fig.3a, it can be known that when $y=1$, the gradient of the Adaptive Wing loss (Wang et al., 2019) is not continuous at the $y-\hat{y}=0$ (indicting the error), but with large gradient magnitude around this point($y-\hat{y}=0$), there exits a negative impact on training convergence. It makes it difficult for the training network to output zero or slight gradient at $y-\hat{y}=0$ and could cause the training process to be unstable and oscillating.

To solve this problem, we propose a novel loss function for Heatmap regression tasks, Smooth Wing Loss, defined as follows:

$$SWL(y,\hat{y}) = \begin{cases} K\ln(1+(y-\hat{y})^2)+ & \\ M(y-\hat{y})^2 & if \quad 0<|y-\hat{y}|<\theta_1 \\ \omega\ln(1+|\frac{y-\hat{y}}{\varepsilon}|^2) & if \quad \theta_1<|y-\hat{y}|<\theta_2 \\ A|y-\hat{y}|-C & otherwise \end{cases} \tag{6}$$

Where y and $\hat{y}$ are the pixel values on the ground truth heatmap and the predicted heatmap respectively, we set $\omega=14$, $\theta_1=0.05$, $\theta_2=0.5$, $\alpha=2.1$, $\varepsilon=1$. Meanwhile, to make Smooth Wing loss function continuous and smooth at every pixels, we also set $K=(1+\theta_1^2)(B\theta_1-2\omega\ln 1+\theta_1)/(2\theta_1^2-2\ln 1+\theta_1^2)(1+\theta_1^2)$, $B=\omega(a-y)\theta_1^{a-y-1}/(\varepsilon+\theta_1^a-y)$, $M=(A-K\ln 1+\theta_1)/\theta_1^2$. In order to prevent gradient explosion and make Smooth Wing Loss be more robust to

outliers, we set that $A = \omega(1/(1+(\theta_2/\varepsilon)^{a-y}))(a-y)(1/\varepsilon)((\theta_2/\varepsilon)^{a-y-1})$, $C = \theta_2 - \omega \ln 1 + (\theta_2/\varepsilon)^{a-y}$.

In addition, Smooth wing loss could also adaptively adjust the type of loss function according to the value of the ground truth pixels. With significant errors$(y-\hat{y})$, the derivative value of the loss function on the predicted value reaches an upper limit, and the upper limit is the value of C, which does not destroy the network parameters. Smooth Wing loss perfectly solves the defects of the Adaptive Wing loss and ensures that the network tends to be stable in the training stage.

As shown in Fig.3a and 3b, our Smooth Wing loss has a smoother curve, with its small gradient magnitude around the zero-point, and it could decrease rapidly to zero. In the training stage, it could converge more quickly than the Adaptive Wing loss. Meanwhile, the smoother curve contributes to making the training process more stable when the error is close to zero, which means that Smooth wing loss could converge further in the later training process. We carry out a series of experiments on the Smooth Wing loss function and the Adaptive Wing loss function, it shows that our Smooth Wing loss not only converges faster than the Adaptive Wing loss in the beginning training stage, but also contributes to further convergence.

3 EXPERIMENTS

3.1 Datasets

We carry out experiments on three popular benchmarks, including WFLW (Wu et al., 2018), 300W (Sagonas et al., 2013a), COFW (Burgos-Artizzu et al., 2013)

WFLW: dataset contains 10000 face images, of which 7500 face images are training images, and 2500 are test images. Each face image provides 98 manually annotated landmarks and attributed labels, such as make-up and illumination.

300W: dataset has 3837 face images, each image is annotated with 68 facial landmarks. The training and test sets of 300W are composed of AFW (Zhu and Ramanan, 2012), Helen (Le et al., 2012), IBUG (Sagonas et al., 2013b), and LFPW (Belhumeur et al., 2013) together.

COFW: dataset contains 1345 training images with 29 facial landmarks, and the test set provides 507 images. It mainly consists of face images with heavy occlusion and profile faces.

3.2 Evaluation Metrics

Regarding related facial landmarks detection work (Xia et al., 2022; Huang et al., 2021; Wang et al., 2019; Kumar et al., 2020), we use the standard metric: Normalized Mean Error (NME), Failure Rate (FR) and Area Under Curve (AUC) to evaluate the proposed method. The **NME** is defined as follows:

$$NME(S,S_{gt}) = \frac{1}{N}\sum_{i=1}^{N}\frac{||p^i - p^i_{gt}||_2}{d} \times 100\% \quad (7)$$

Where S and S_{gt} represent the predicted and the annotated facial landmarks respectively, p^i and p^i_{gt} represent $i-th$ facial landmarks in S and S_{gt}. N is the total number of landmarks, d could be the distance between outer eye corners (inter-ocular) or the distance between pupil centers (inter-pupils). **FR** indicates the percentage of images in the test set whose NME is higher than a certain threshold. **AUC** is calculated based on the Cumulative Error Distribution (CED) curve. AUC is the area under the CED curve, from zero to the threshold for FR.

3.3 Implementation Details

We crop and resize each image to a resolution of 256×256, and they are downsampled to 64×64 to generate heatmaps. We train the network mentioned above framework using Adam optimizer (Kingma and Ba, 2014), the initial learning rate is set to 0.0005, train 120 epochs, the learning rate decays at the 40th epoch and the 80th epoch, the decay rate is 0.2. We set three parallel multi-branch stages with convolution and transformer, and finally we get four various-resolution feature maps, we interpolate the other three lower resolutions into the highest-resolution feature maps. In addition, we randomly enhance the training data by random flipping (50%), mirroring (40%), masking (30%), color gamut change (30%), rotation (±30°), etc.

3.4 Comparison with State-of-the-Art Method

3.4.1 WFLW

As tabulated in Table.1 and Table.2, our methods demonstrate impressive performance on WFLW. With the help of the Intensive Attention mechanism and the proposed parallel multi-branch structure, the performance of IACT outperforms the state-of-the-art methods (Huang et al., 2021; Xia et al., 2022; Kumar et al., 2020). Specifically, IACT reaches 4.04 NME, 2.62

Table 1: Comparison with state-of-the-art methods on WFLW with inter-ocular NME(%)↓ , Params(M), Flops(G). Key: [BEST,SECOND BEST].

Method	Year	Params	Flops	Full.	Pose	Exp.	Ill.	Mu.	Occ.	Blur
LAB	CVPR_2018	12.26	18.96	5.27	10.24	5.51	5.23	5.15	6.79	6.32
Wing	CVPR_2018	-	25	4.99	8.43	5.21	4.88	5.26	6.21	5.81
MHHN	TIP_2020	-	-	4.77	9.31	4.79	4.72	4.59	6.17	5.82
DeCaFA	ICCV_2019	10	-	4.62	8.11	4.64	4.41	4.63	5.74	5.38
HRNet	TPAMI_2020	9.66	4.75	4.65	7.94	4.85	4.55	5.29	5.44	4.86
LUVLI	CVPR_2020	-	-	4.37	7.56	4.77	4.30	4.33	5.29	4.94
AWing	ICCV_2019	24.15	26.8	4.36	7.38	4.58	4.32	4.27	5.19	4.96
SDFL	TIP_2019	-	5.17	4.35	7.42	4.63	4.29	4.22	5.19	5.08
ADNet	ICCV_2021	13.37	17.04	4.14	6.96	4.38	4.09	4.05	5.06	4.79
SLPT	CVPR_2022	19.45	8.14	4.12	6.99	4.37	4.02	4.03	5.01	4.79
Ours	-	20.29	6.45	4.04	6.54	3.87	3.71	3.42	4.67	4.43

Table 2: Performance comparison of IACT and the state-of-the-art methods on WFLW. The normalization factor is inter-ocular and the threshold for FR is set to 0.1. Key: [Best, Second Best].

Metric	Method	Full	Pose	Exp.	Ill.	Mu.	Occ.	Blur
$FR_{0.1}(\%)$ (↓)	LAB (Wu et al., 2018)	7.56	28.83	6.37	6.73	7.77	13.72	10.74
	SAN (Dong et al., 2018)	6.32	27.91	7.01	4.87	6.31	11.28	6.60
	HRNet (Wang et al., 2020)	4.64	23.01	3.50	4.72	2.43	8.29	6.34
	LUVLi (Kumar et al., 2020)	3.12	15.95	3.18	2.15	3.40	6.39	3.23
	AWing (Wang et al., 2019)	2.84	13.50	2.23	2.58	2.91	5.98	3.75
	SDFL (Lin et al., 2021)	2.72	12.88	1.59	2.58	2.43	5.71	3.36
	SDL (Li et al., 2020)	3.04	15.95	2.86	2.72	1.45	5.29	4.01
	ADNet (Huang et al., 2021)	2.72	12.72	2.15	2.44	1.94	5.79	3.54
	SLPT (Xia et al., 2022)	2.76	12.27	2.23	1.86	3.40	5.98	3.88
	Ours	2.61	8.28	0.62	2.14	2.42	5.02	3.36
$AUC_{0.1}$ (↑)	LAB (Wu et al., 2018)	0.532	0.235	0.495	0.543	0.539	0.449	0.463
	SAN (Dong et al., 2018)	0.536	0.236	0.462	0.555	0.522	0.456	0.493
	HRNet (Wang et al., 2020)	0.524	0.251	0.510	0.533	0.545	0.459	0.452
	LUVLi (Kumar et al., 2020)	0.557	0.310	0.549	0.584	0.588	0.505	0.525
	AWing (Wang et al., 2019)	0.572	0.312	0.515	0.578	0.572	0.502	0.512
	SDFL (Lin et al., 2021)	0.576	0.315	0.550	0.585	0.583	0.504	0.515
	SDL (Li et al., 2020)	0.589	0.315	0.566	0.595	0.604	0.524	0.533
	ADNet (Huang et al., 2021)	0.602	0.344	0.523	0.580	0.601	0.530	0.548
	SLPT (Xia et al., 2022)	0.595	0.348	0.574	0.601	0.605	0.515	0.535
	Ours	0.615	0.409	0.625	0.636	0.619	0.575	0.587

$FR_{0.1}$ and 0.612 $AUC_{0.1}$ on WFLW, which demonstrates that our method could localize the landmarks accurately.

3.4.2 COFW

Following other works, we report results in terms of NME and FR by inter-ocular normalization and inter-pupil normalization. The compare result is shown in Table.3. Compared to other excellent works, our methods still maintain impressive performance and surpass all state-of-the-art methods. Significantly, IACT reaches 3.12 inter-ocular NME and 4.53 inter-pupil NME.

3.4.3 300W

As shown in Table.4, we compared with previous works in inter-ocular NME on the 300W benchmark that contains full set, challenge set, and common set. It is obvious that our methods also gain excellent results. Especially, IACT reaches 2.82 NME on the full set, 2.51 NME on the common set and 4.09 NME on the challenge set.

Table 3: NME and FR comparisons with State-of-the-art methods under inter-ocular normalization and inter-pupil normalization on COFW, the threshold for FR is set to 0.1. Key: [Best, Second Best].

Method	Inter-Ocular		Inter-pupil	
	NME(%)↓	FR(%)↓	NME(%)↓	FR(%)↓
LAB	3.92	0.39	5.58	2.76
SDFL	3.63	0.00	-	-
HRNet	3.45	0.2	-	-
DAC-CSR	-	-	6.03	4.73
Human	-	-	5.60	-
DCFE	-	-	5.27	7.29
MHHN	-	-	4.95	1.78
AWing	-	-	4.94	0.99
ADNet	-	-	4.68	0.59
SLPT	3.32	0.00	4.79	1.18
Ours	3.12	0.00	4.53	0.21

Table 4: Comparison with state-of-the-art methods under inter-ocular NME(%)↓ on 300W. Key: [Best, Second Best].

Method	Full	Com.	Chal.
LAB (Wu et al., 2018)	3.49	2.98	5.19
Wing (Feng et al., 2018)	3.60	3.01	6.01
DCFE (Valle et al., 2018)	3.24	2.76	5.22
DeCaFA (Dapogny et al., 2019b)	3.39	2.93	5.26
Awing (Wang et al., 2019)	3.07	2.72	4.52
HRNet (Wang et al., 2020)	3.32	2.87	5.15
LUVLI (Kumar et al., 2020)	3.23	2.76	5.16
SDFL (Lin et al., 2021)	3.28	2.88	4.93
ADNet (Huang et al., 2021)	2.93	2.53	4.58
SLPT (Xia et al., 2022)	3.17	2.75	4.90
Ours	2.82	2.51	4.09

3.5 Ablation Study

3.5.1 Evaluation on the Multi-Head Spatial Attention

To explore the contribution of the Multi-head Spatial attention module in the Transformer stages, we train our Convolution-Transformer Network with different transformer structures on the WFLW dataset. We introduce Vit structure (Dosovitskiy et al., 2020) and DETR structure (Carion et al., 2020). The results are shown in Table.5. The Vit structure still outperforms other methods because of the parallel multi-branch structure. With the DETR structure, the performance on WFLW is boosted from 4.50% to 4.32% in terms of NME. Our transformer structure with a Spatial multi-head attention module has the best performance, reaching 4.04% in NME.

Table 5: NME,FR,AUC with different transformer structures on WFLW. [Best].

Method	Spa.	VIT	DETR	NME(%)↓	FR(%)↓	AUC(%)↑
Model 1	-	✓	-	4.50	4.28	0.575
Model 2	-	-	✓	4.32	2.92	0.596
Model 3	✓	-	-	4.04	2.63	0.612

Table 6: NME,FR,AUC with/without Feature Attention module on WFLW [Best].

Method	NME(%)↓	FR(%)↓	AUC(%)↑
w/o Feature Attention	4.16	2.83	0.595
w Feature Attention	4.04	2.63	0.612

3.5.2 Evaluation on the Feature Attention

We implement two models with/without the Feature Attention module to explore the influence of multi-model feature information. With the Feature Attention module, the performance of IACT is improved, as shown in Table.6.

3.5.3 Evaluation on Convergence Curves of Smooth Wing Loss and Adaptive Wing Loss

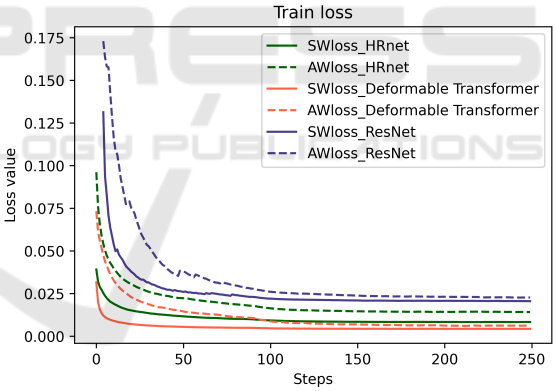


Figure 4: Convergence curves of different backbones using Adaptive Wing Loss or Smooth Wing Loss.

We apply the Smooth Wing Loss and the Adaptive Wing Loss on different backbones, such as HRNet (Wang et al., 2020), ResNet (He et al., 2015), and our IACT. The convergence curves of them are shown in Fig.4. Experiments on different backbones show that Smooth Wing Loss converges faster than Adaptive Wing Loss. Compared with it, Smooth Wing loss converges further in the later training stages.

4 CONCLUSION

In this paper, we propose a parallel multi-branch network, combining the advantages of both the convolution and transformer, which can capture rich detailed features while modelling the long-range relations. Besides, the Intensive attention mechanism is utilized in the network, which enables the network to gain multi-model feature maps with different resolutions for better representations and focus global attention weight rapidly on sparse and meaningful locations. Additionally, we propose a novel and effective loss function, Smooth Wing Loss, which steadily accelerates the convergence speed of the network and can further converge at the later training stage. Extensive experiments show that IACT outperforms the state-of-the-art methods, and the ablation studies prove the effectiveness of the proposed methods.

REFERENCES

Belhumeur, P. N., Jacobs, D. W., Kriegman, D. J., and Kumar, N. (2013). Localizing parts of faces using a consensus of exemplars. *IEEE transactions on pattern analysis and machine intelligence*, 35(12):2930–2940.

Burgos-Artizzu, X. P., Perona, P., and Dollár, P. (2013). Robust face landmark estimation under occlusion. In *Proceedings of the IEEE international conference on computer vision*, pages 1513–1520.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., and Zagoruyko, S. (2020). End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer.

Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., and Wei, Y. (2017). Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 764–773.

Dapogny, A., Bailly, K., and Cord, M. (2019a). Decafa: Deep convolutional cascade for face alignment in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6893–6901.

Dapogny, A., Bailly, K., and Cord, M. (2019b). Decafa: Deep convolutional cascade for face alignment in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6893–6901.

Deng, J., Trigeorgis, G., Zhou, Y., and Zafeiriou, S. (2019). Joint multi-view face alignment in the wild. *IEEE Transactions on Image Processing*, 28(7):3636–3648.

Dong, X., Yan, Y., Ouyang, W., and Yang, Y. (2018). Style aggregated network for facial landmark detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 379–388.

Dong, X., Yang, Y., Wei, S.-E., Weng, X., Sheikh, Y., and Yu, S.-I. (2020). Supervision by registration and triangulation for landmark detection. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3681–3694.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Feng, Z.-H., Kittler, J., Awais, M., Huber, P., and Wu, X.-J. (2018). Wing loss for robust facial landmark localisation with convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2235–2245.

Guo, X., Li, S., Yu, J., Zhang, J., Ma, J., Ma, L., Liu, W., and Ling, H. (2019). Pfld: A practical facial landmark detector. *arXiv preprint arXiv:1902.10859*.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep residual learning for image recognition. cite arxiv:1512.03385Comment: Tech report.

Huang, Y., Yang, H., Li, C., Kim, J., and Wei, F. (2021). Adnet: Leveraging error-bias towards normal direction in face alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3080–3090.

Jin, H., Liao, S., and Shao, L. (2021). Pixel-in-pixel net: Towards efficient facial landmark detection in the wild. *International Journal of Computer Vision*, 129(12):3174–3194.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kowalski, M., Naruniec, J., and Trzcinski, T. (2017). Deep alignment network: A convolutional neural network for robust face alignment. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 88–97.

Kumar, A., Marks, T. K., Mou, W., Wang, Y., Jones, M., Cherian, A., Koike-Akino, T., Liu, X., and Feng, C. (2020). 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.

Le, V., Brandt, J., Lin, Z., Bourdev, L., and Huang, T. S. (2012). Interactive facial feature localization. In *European conference on computer vision*, pages 679–692. Springer.

Li, W., Lu, Y., Zheng, K., Liao, H., Lin, C., Luo, J., Cheng, C.-T., Xiao, J., Lu, L., Kuo, C.-F., et al. (2020). Structured landmark detection via topology-adapting deep graph learning. In *European Conference on Computer Vision*, pages 266–283. Springer.

Lin, C., Zhu, B., Wang, Q., Liao, R., Qian, C., Lu, J., and Zhou, J. (2021). Structure-coherent deep feature learning for robust face alignment. *IEEE Transactions on Image Processing*, 30:5313–5326.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022.

Parekh, V., Shah, D., and Shah, M. (2020). Fatigue detection using artificial intelligence framework. *Augmented Human Research*, 5(1):1–17.

Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., and Pantic, M. (2013a). 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.

Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., and Pantic, M. (2013b). 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.

Savchenko, A. V. (2021). Facial expression and attributes recognition based on multi-task learning of lightweight neural networks. In *2021 IEEE 19th International Symposium on Intelligent Systems and Informatics (SISY)*, pages 119–124. IEEE.

Trigeorgis, G., Snape, P., Nicolaou, M. A., Antonakos, E., and Zafeiriou, S. (2016). Mnemonic descent method: A recurrent process applied for end-to-end face alignment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4177–4187.

Valle, R., Buenaposada, J. M., Valdés, A., and Baumela, L. (2018). A deeply-initialized coarse-to-fine ensemble of regression trees for face alignment. In Ferrari, V., Hebert, M., Sminchisescu, C., and Weiss, Y., editors, *ECCV (14)*, volume 11218 of *Lecture Notes in Computer Science*, pages 609–624. Springer.

Wan, J., Lai, Z., Liu, J., Zhou, J., and Gao, C. (2020). Robust face alignment by multi-order high-precision hourglass network. *IEEE Transactions on Image Processing*, 30:121–133.

Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., et al. (2020). Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3349–3364.

Wang, X., Bo, L., and Fuxin, L. (2019). Adaptive wing loss for robust face alignment via heatmap regression. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6971–6981.

Woo, S., Park, J., Lee, J.-Y., and Kweon, I. S. (2018). Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19.

Wu, W., Qian, C., Yang, S., Wang, Q., Cai, Y., and Zhou, Q. (2018). Look at boundary: A boundary-aware face alignment algorithm. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2129–2138.

Wu, W. and Yang, S. (2017). Leveraging intra and inter-dataset variations for robust face alignment. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 150–159.

Xia, J., Huang, W., Zhang, J., Wang, X., Xu, M., et al. (2022). Sparse local patch transformer for robust face alignment and landmarks inherent relation learning. *arXiv preprint arXiv:2203.06541*.

Xiao, S., Feng, J., Xing, J., Lai, H., Yan, S., and Kassim, A. (2016). Robust facial landmark detection via recurrent attentive-refinement networks. In *European conference on computer vision*, pages 57–72. Springer.

Yu, B. and Tao, D. (2021). Heatmap regression via randomized rounding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhang, K., Zhang, Z., Li, Z., and Qiao, Y. (2016). Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, 23(10):1499–1503.

Zhu, X. and Ramanan, D. (2012). Face detection, pose estimation, and landmark localization in the wild. In *2012 IEEE conference on computer vision and pattern recognition*, pages 2879–2886. IEEE.