Gender classification and landmark detection in celebrity face images Applied Machine Learning for Computer Vision

João Carlos Ramos Gonçalves de Matos[1]
[1]*Faculty of Engineering of University of Porto (Portugal)*

***Abstract*—After comparing a single-task with a multi-task setting for both gender classification and face landmark localization, we propose and implement a self-attention-based architectural improvement to the multi-task baseline setting. The goal is to encode non-local relationships to gather task-oriented global information for each classification head, which convolutional layers may not be able to attain. For landmarks detection, we also propose a novel assessment metric, that we refer to as operating curve. A VGG-16 is used as a backbone in all tested architectures. Four sets of experiments are conducted at 10 epochs with the goal of tuning the hyperparameters of the four final models we propose; whenever necessary, longer training is conducted before testing, at 50 epochs. For both tasks, single-task approaches yield better performances. For gender classification, the single-task architecture achieves an AUC ROC of 0.9984, accuracy of 0.9600 and F1-Score of 0.9600; for landmarks detection, the single-task approach produces a NME of (6.2618 ± 3.5099)%, a $FR_{0.1}$ of 17.6% and an area under the operating curve of 0.9275. Multi-task setting shows very reasonable results, as well, that, though do not surpass the baseline single-tasks, are quite close to them. Finally, the self-attention layers as an improvement to the baselines also show viability and to be suitable for the problem, though there is room for further experiments to be conducted. While our improvement proposal outperforms the multi-task baseline setting in gender classification results, it stays a bit behind in the landmarks task.**

***Index Terms*—Computer Vision, Deep Learning, Gender Classification, Face Landmarks Regression, VGG-16, Multi-Task Setting, Self-Attention**

I. Introduction

Computer Vision seeks to understand and automate tasks that the human visual system can do [1]. Common tasks in Computer Vision include object detection; identity recognition; semantic, instance and panoptic segmentation; action and behavior recognition; 2D and 3D human pose estimation.

Over the last years, Deep Learning methods have been shown to outperform previous state-of-the-art Machine Learning techniques in several fields, with computer vision being one of the most prominent cases [2].

In this work, we will approach two different analysis tasks over face images: gender classification and face landmark localization. For that, we compare a single-task setting with a multi-task one. After that, we propose and implement a self-attention-based architectural improvement to the multi-task baseline setting.

Gender classification entails recognising a subject's gender through an image of his/her face. Automatic gender classification is currently receiving increasing attention. A gender classification system can be used for many purposes, such as biometrics-related applications, or access-control in sensitive locations [3].

Fig. 1. CelebA-mini example faces.

On the other hand, facial landmark detection is one of the key elements of face processing pipeline. It is used in virtual face reenactment, emotion recognition, or driver status tracking. Neural networks have shown an astonishing qualitative improvement for in-the-wild face landmark detection problem, and are now being studied by many researchers in the field [4].

II. Methodology

A. Dataset

The dataset this project was developed upon was a subset of the Large-scale CelebFaces Attributes (CelebA) Dataset [5]. The dataset, to which we will refer to as CelebA-mini onwards, contains 500 face images of 500 different celebrities, with data balance gender-wise. Each image is associated with a gender label (0 for women, 1 for men), and a set of 5 landmark (x,y) coordinates corresponding to the two eyes, the nose, and the two sides of the mouth. Some examples from the dataset can be found in figure 1.

B. Data Split

The data was split into train, validation and test sets. For test, 10% of the data (50 images) was kept apart. From the remaining 450 images, we did another 90-10% split and used 405 images for training and 45 for validation. Since data was balanced in the original CelebA-mini, that was taken into account when performing data split and each set presents equal number of men and women subjects. The validation set was used to fine-tune hyperparameters during training, as well as to select as final model the one that resulted from the epoch with best performance in validation.

C. Image Pre-processing

To assure that all images presented the same dimensions and were normalized, a set of operations are performed when loading the data. We first rescale images to have the lower

dimension with 256 and, then, apply a squared, centered crop to get 3 x 224 x 224 images. The exact same transformations are applied to the landmarks' coordinates, to assure these will match the transformed image. Since the models to be used are loaded with pre-trained with ImageNet's weights, a normalization is applied to each channel: ImageNet's mean is subtracted and the result is divided by ImageNet's standard deviation. Finally, images are converted into a Tensor.

D. Single-Task Gender Classification

1) Architecture: The chosen backbone to use across all experiments in this project was VGG-16, the convolution neural network (CNN) architecture that won ILSVR (ImageNet database) competition in 2014 [6]. It is considered to be one of the best Computer Vision architectures till date. Its architecture is depicted in figure 2. Since only 2 classes are present (woman or man) in the task of gender classification, the last convolution layer was modified to accommodate this restriction. The number of channels being output was set to 2, instead of 1000. Finally, to compute each class' probability, the softmax function (equation 1) is used, and the final prediction will be the class yielding the highest probability.

$$\sigma(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \quad for\ i = 1, 2, \ldots, K \tag{1}$$

In a recent work [7] with gender classification from faces, compared with ResNet-50 and MobileNet, VGG-16 showed delivered the highest recognition accuracy on a database consisting of 1000 images.

The VGG-16 was loaded with weights that have been pre-trained on ImageNet dataset and made available online. Though the train domain is different from CelebA-mini, as ImageNet was designed for visual object recognition [8], networks that have been pre-trained with this dataset can already do meaningful representations of images in the feature space. Therefore, it is usually simple to do Transfer Learning to another domain involving visual classification.

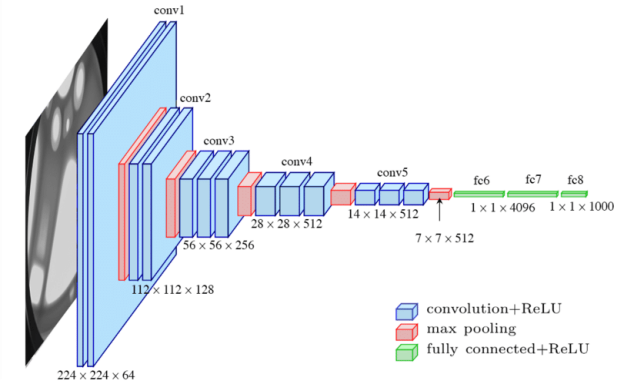

Fig. 2. VGG-16 architecture. Image borrowed from [6].

2) Loss: As we are dealing with a binary classification, the Cross-Entropy Loss was considered to be the most suitable to address the task, which is a calculated with equation 2,

$$CE = -(y \log(p) + (1 - y) \log(1 - p)) \tag{2}$$

where y is binary indicator if class label is correct and p the predicted probability that the observation is that class.

3) Evaluation Metrics: Following several past classification tasks in literature, the performance of our classifier during training and validation was assessed with Accuracy (equation 3), Precision (equation 4), Recall (equation 5), and F1-Score (equation 6).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} = \frac{2 * TP}{2 * TP + FP + FN} \tag{6}$$

where TP denotes true positive, FP false positive, TN true negatives, and FN false negatives.

In addition to these, during testing, Confusion Matrices, which plot the number of TP, FP, TN, and FN, were assessed. Finally, the Receiver Operating Characteristic (ROC) curve and the area under it were assessed. The ROC curve illustrates the diagnostic ability of a binary classifier system as the discrimination threshold (or cut-off value) is varied, plotting the true positive rate (TPR) against the false positive rate (FPR), defined in equation 7 and 8. This is the primary metric looked upon for the gender classification task, in this project, and following work done in literature.

$$TPR = Sensitivity = Recall = \frac{TP}{TP + FN} \tag{7}$$

$$FPR = 1 - Specificity = \frac{TN}{FP + TN} \tag{8}$$

4) Experiments: This project was developed using PyTorch package [9], using Kaggle's Python environment, with a Nvidia K80 GPU (rather than the widely used Google Colab, since Kaggle has a fixed quota of 30h of GPU per week, that the user can manage as prefered, without worrisome timeouts).

For the training setting of this first task, we fixed as hyperparameters a (fixed) learning rate of 0.001, a batch size of 15 and 10 epochs. From the 10 trained epochs, the best model is selected out of the epoch with lowest validation loss.

To choose the most suitable optimizer, we ran training experiments to compare Adam, Adamax [10], Adagrad [11] and RMS Prop [12]. The best performing one will be taken to test.

E. Single-Task Face Landmarks Regression

1) Architecture: The decisions, proposals and experiments for this task were based on a literature review from April 2022 [4] on Facial Landmark Detection. It compares 22 architectures that can be of direct regression (D), Heatmap Regression (H), or combined Heatmap and Direct methods (H + D). Figure 3 depicts the inference time for each algorithm and type of inference device, and figure 4 the performance in 3 different benchmarks.

A conclusion from these figures is that direct regression methods yield the most interesting inference times, while being able to perform equally well 4 across different benchmarks, when compared to more complex techniques.

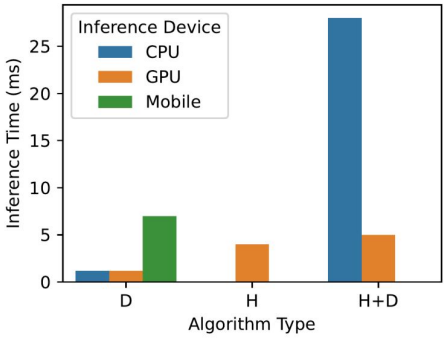

Fig. 3. According to the executed study, the best inference time for each algorithm type: direct (D), heatmap-based (H), combined heatmap and direct (H+D). Results are shown for different inference devices: CPU, GPU and Mobile. Direct regression algorithms have the best speed. Mixed H+D algorithm time is unexpectedly high. Image borrowed from [4]

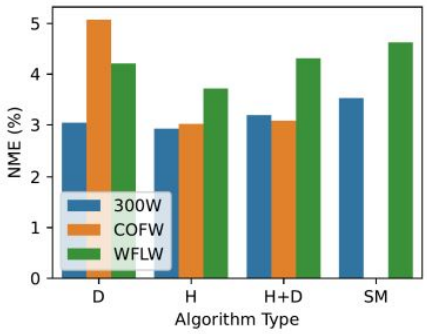

Fig. 4. The best NME for each algorithm type: direct (D), heatmap-based (H), combined heatmap and direct (H+D), shape-model-based (SM). Heatmap-based approaches offer the best quality. Note, similar performance of direct and heatmap-based approaches on 300W. Image borrowed from [4]

As one of the main conclusions of this review is that most architectures being developed nowadays do not take into consideration the applicability in real-world resource-constrained environments, we decided to go for a simpler approach and solve this task with a direct regression. Though this kind of algorithm may not be as much robust as a heatmap regression, for example, it is more likely to be used in a real-world application, where computational power is limited.

In landmarks' direct regression, a model detects the landmark coordinates represented by a vector from a facial image [13]. The dimension of the vector is the twice number of landmarks (coordinates x and y) and, thus, for this task, 10 classes are considered.

The chosen backbone was VGG-16; the last convolutional layer is modified to accommodate 10 classes in the output. Since regression is being performed, this time there is no need to apply the softmax function to the output of the network.

2) Loss: Even though several interesting losses have been proposed for Landmark Regression, such as Adaloss, [14], Adaptive Wing Loss [15] and 2D Wasserstein Loss [16], these are typically used in Heatmap Regression Models (HRMs).

Taking that into account, suitable and simple losses for a direct landmark regression include L2 Loss - or mean squared error - (equation 9), L1 Loss - or mean absolute error - (equation 10) and L1 smoothed loss (equation 11).

$$L2 = \sum_{i=1}^{D} (x_i - y_i)^2 \tag{9}$$

$$L1 = \sum_{i=1}^{D} |x_i - y_i| \tag{10}$$

$$L1_{smooth} = \begin{cases} 0.5(x_n - y_n)^2/\beta, & \text{if } |x_n - y_n| < \beta \\ |x_n - y_n| - 0.5 * \beta, & \text{otherwise} \end{cases} \tag{11}$$

L2, L1 and smooth L1 losses yield very small values for small landmark location differences, which hinders network training in cases where the error is very small. In contrast, Huber (equation 12) and Wing Loss (equation 13) [17] can be less sensitive to outliers and much more sensitive to medium to-small errors, which can improve training overall [4].

$$L_{Huber} = \begin{cases} \frac{1}{2}(y - \hat{y})^2, & \text{if } |(y - \hat{y})| < \delta \\ \delta((y - \hat{y}) - \frac{1}{2}\delta), & \text{otherwise} \end{cases} \tag{12}$$

$$L_{Wing} = \begin{cases} \omega \ln(1 + |x|/\epsilon), & \text{if } |x| < \omega \\ |x| - C, & \text{otherwise} \end{cases} \tag{13}$$

3) Evaluation Metrics: Each landmark is assessed individually and as part of a global assessment per subject. For that, two main metrics are used: the Normalized Mean Error (NME), defined in equations 14 and 15, and the Failure Rate (FR), in equation 16 [4].

$$NME = \frac{1}{K} \sum_{k=1}^{K} NME_K \tag{14}$$

$$NME_k = \frac{1}{N_L} \sum_{i=1}^{N_L} \frac{Y_i - \hat{Y}_i}{d} * 100 \tag{15}$$

where Y is the matrix of true landmark locations, $\hat{Y}$ is the matrix of predicted landmark locations, d is the normalization coefficient (different for each subject), N_L is the number of

facial landmarks per face in the dataset, K is the number of images in the test set. Lower metric values are better.

$$FR_{0.1} = \frac{1}{K}\sum_{k=1}^{K}[NME_K > 100\%] * 100 \tag{16}$$

denoting the number of images with NME above the 10 % threshold. The lower the metric values, the better.

Finally, we propose an operating curve to assess this task, which is a variation of Cumulative Error Distribution (and Area Under Curve, CED-AUC) from literature, which originally plots the fraction of images for which NME is less than or equal to the NME value on X axis. Instead, we propose plotting Success Rate vs. the NME Threshold, thus computing the success rate (inverse of FR) at thresholds that vary from 0 to 100. This curve behaves similarly to the ROC curve and, as such, its area under the curve can be an interesting evaluation metric.

4) Experiments: The hyperparameters were set similarly to the previous task. The learning rate was of 0.001, batch size 15 and the number of epochs 10. The optimizer was the one with best performance in the validation phase of the previous task. As experiments, we try the 5 different losses presented before and compare the validation results. The best model is trained with 50 epochs and taken to test.

F. Multi-Task Setting

1) Architecture: The approach to the multi-task setting was quite straightforward and in line with literature [18]. Using the previous VGG-16 networks as a baseline, we transformed the single classification head into a double classification head. One fully connected network will be responsible for performing gender classification and the other one for direct landmark regression. The proposed architecture is depicted in figure 5.

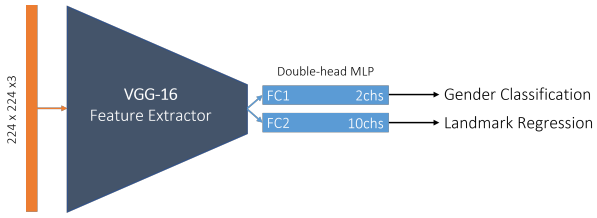

Fig. 5. Double-head architecture for multi-task, with VGG-16 backbone.

2) Loss: The proposed loss for this task is composed of a linear combination between the loss for gender classification and landmarks regression, as depicted in equation 17. Each loss is weighted by λ_1 and λ_2. While the loss for the gender classification is the Cross-Entropy Loss (equation 2), the loss for Landmark regression will be the one with highest performance in the second task.

$$L_{multi_task} = \lambda_1 L_{gender} + \lambda_2 L_{landmarks} \tag{17}$$

3) Evaluation Metrics: The same as in both gender classification and landmarks regression.

4) Experiments: We will study the influence of the weights λ_1 and λ_2, making them vary to give more importance to gender classification or landmark regression. The best model is then trained with 50 epochs and is taken to test.

Past works in literature [18] intuitively tell us that landmarks task should be assigned a higher loss weight, since its a more complex task, from which gender classification is hypothesised to be highly dependent.

G. Proposed Improvements

1) Architecture: Transformers and attention-based architectures have attracted much attention lately, outperforming several existing models in different benchmarks in literature. Applying self-attention (depicted in figure 6 and computed with equation 18) to images in a naive way would require that each pixel attends to every other pixel, which cannot be scaled to actual images, since there is a quadratic cost in the number of pixels [19].

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{18}$$

where Q denotes a query, K a key, V a value, and d_k the key size. The original paper has shown that reducing the attention key size hurts model quality.

To apply Transformers in the context of image processing, several approximations have been tried before Vision Transformers (ViT) [20], such as combining CNNs with forms of self-attention. A first example is augmenting feature maps for image classification [21] or further processing the output of a CNN using self-attention [22].

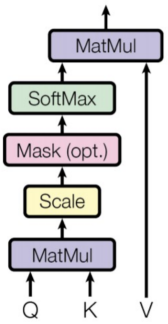

Fig. 6. Scaled dot-product attention, as proposed by [19]. In our application of it, the optional mask is not used.

In this part of the project we explore the effect of self-attention layers placed after the VGG-16's feature extraction. This simple improvement is built upon the multi-task architecture explored before, as one can see in figure 7. The rationale behind this proposal is that different tasks (in this case, our 2 tasks) may require considering different long-ranging interactions within the same input feature map. The self-attention layers would be able to encode non-local relationships to gather task-oriented global information for each classification head, which convolutional layers may not able to attain.

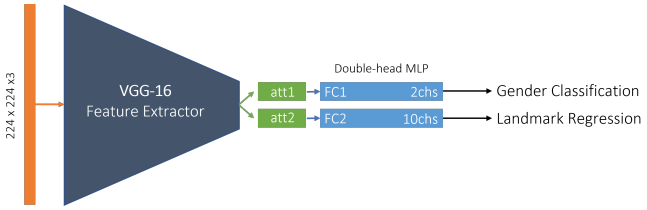

Fig. 7. Proposed improvement for our multi-task setting, based on the idea of self-attention [19].

2) Loss and Evaluation Metrics: To enable comparisons and drive a fair analysis, the hyperparameters, loss, and evaluation metrics will be the same as before, for the multi-task setting.

3) Experiments: The experiments for this section will be focused on the value of d_k. The original self-attention paper [19] proposes a value of 64, but that is in a setting with 8 parallel attention layers, to optimize performance. In our proposal, we only have 2 attention layers. The value of d_k, which represents the size of the keys, will be kept as common for the size of queries, q, and values, v, for simplicity, though it could also be interesting to study the effect of different values. The best model that results from this experiment is then trained with 50 epochs and taken to test.

III. Results & Discussion

In this section, we will go through the executed experiments, draw some conclusions and provide a discussion, in line with previous works in literature. In figure 10, one can find a summary of the conducted experiments, as well as the rationale to obtain the 4 finals models of this project, which will be compared in the Final Testing section.

A. Single-Task Gender Classification

Table I depicts the obtained results for this experiment. The Adamax optimizer shows, over 10 epochs, the best validation loss of 0.0763, significantly below all the other tested ones. It also results in a model that outperforms the other models in terms of the other considered metrics, with an AUC ROC of 1.000 in the validation set for the best epoch. For all experiments onwards, Adamax will be the used optimizer.

In figure 8, one can see the learning curves for each studied optimizer. The lowest validation loss is often achieved early in the process and typically the validation loss starts increasing after the 4th epoch (with exception of RMSprop). 10 epochs may be too much for this single-task gender classification models. The early stopping strategy we implement is fundamental to guarantee that best model is saved and used.

B. Single-Task Face Landmarks Regression

Table II depicts the results for the comparison of different losses in the single-task 2 for landmarks regression. L1 Loss yields the lowest NME and FR, of (11.230 ± 5.669)% and 44.9%, respectively, showing to be the most adequate type of loss for landmarks detection.

L1 Smooth loss also shows an interesting performance, which makes sense since it is a variation from the L1 Loss. On the other hand, the Wing Loss, which was expected (according to literature) to outperform all other losses, presents the poorest result in terms of NME: (16.469 ± 5.286)%.

C. Multi-Task Setting

Table III depicts the results, both in terms of gender classification and landmarks detection, for each λ_1 and λ_2. The combination that yields the most interesting result is for $\lambda_1 = 1$ and $\lambda_2 = 20$. While having a validation AUC ROC of 1.000 in the gender classification task, the achieved NME is of (12.863 ± 5.404)%, which is considerably lower.

On the other hand, models with big weights attributed to gender classification produce catastrophic results for the landmarks regression, which gets a NME as low as (108.861 ± 44.017)% and a FR of 100% for $\lambda_1 = 20$ and $\lambda_2 = 1$.

Though it could be expected that the higher the λ that concerns the task, the higher the metrics related to that task, these experiments show gender classification highly benefits from the contribution from the landmarks labels. In fact, the best AUC ROC (of 1.000) are only obtained when the losses are balanced, with the same weights, or landmarks have a higher weight. On the other hand, when gender classification-related loss is favoured, the performance of such task gets lower. An hypothesis is that, in those scenarios, the landmark detection with such a low loss weight may be just "confusing" the network and hindering the learning for both tasks. Conversely, when the landmarks detection loss is favoured, since it is a more complex task and that can actually support the gender classification decision, all metrics get better, overall.

Further tuning of these weights would, most likely, produce even better results, but we consider that these reduced experiments already provide an interesting insight on the trend regarding the effect of λ_1 and λ_2 in the model's performance.

D. Proposed Improvements

Table IV shows the results obtained on the multi-task setting when self-attention layers are added as an improvement. Varying d_k from 32 to 512, it is not very clear which kind of value is better for the architecture here. Though the lowest NME, (17.310 ± 6.138)%, is obtained by $d_k = 32$, the best FR, 64.4% is obtained by $d_k = 512$. Yet, the $d_k = 64$ architecture produces the best gender classification results, but performs poorly for Landmarks Regression.

One could argue that the architecture with $d_k = 32$ would be the best choice, since it yields good overall results, while considerably lowering the computational cost of the architecture (the FC networks for classification get smaller if the q,k,v have a dimension lower than 512). However, we opted for the architecture with $d_k = 512$, which produced a model with good metrics overall and the lowest total validation loss - which has been our primary criterion for the choice of the best models, in this project.

For these experiments, it could be interesting to test more values of d_k, make them different for each self-attention layer

TABLE I

EXPERIMENTS FOR SINGLE-TASK 1 GENDER CLASSIFICATION. ADAM, ADAMAX, ADAGRAD AND RMSPROP OPTIMIZERS WERE TESTED. UNDERLINED ARE THE BEST VALUES FOR EACH METRIC.

Optimizer	Val. Loss	Accuracy	Recall	Precision	F1-Score	AUC ROC
Adam	0.2400	0.956	0.958	0.958	0.958	0.970
Adamax	0.0763	0.978	1.000	0.952	0.974	1.000
Adagrad	0.1126	0.933	0.958	0.933	0.941	1.000
RMSprop	0.2623	0.889	0.812	0.944	0.869	0.969

TABLE II

EXPERIMENTS FOR SINGLE-TASK 2 LANDMARKS REGRESSION. L1, L2, L1 SMOOTH, HUBER AND WING LOSSES WERE COMPARED. UNDERLINED ARE THE BEST VALUES FOR EACH METRIC.

Loss	Val. Loss	NME (%)	$FR_{0.1}$ (%)
L1	0.0176	11.230 ± 5.669	44.9
L2	0.0011	15.717 ± 5.652	64.0
L1 Smooth	0.0003	12.452 ± 5.730	52.9
Huber	0.0004	14.044 ± 5.623	57.8
Wing	0.1939	16.469 ± 5.286	58.7

TABLE III

EXPERIMENTS FOR MULTI-TASK SETTING. COMBINATIONS OF $lambda_1$ AND $lambda_2$ RANGING FROM 1 TO 50 ARE COMPARED. UNDERLINED ARE THE BEST VALUES FOR EACH METRIC.

			Gender Classification						Landmarks Regression		
λ_1	λ_2	Val. Loss	CE Loss	Accuracy	Recall	Precision	F1-Score	AUC ROC	L1 Loss	NME (%)	$FR_{0.1}$ (%)
1	**1**	0.2108	0.1202	0.933	0.963	0.917	0.933	1.000	0.0906	56.726 ± 17.344	97.8
1	**5**	0.2145	0.0641	0.956	1.000	0.930	0.963	1.000	0.0301	18.739 ± 7.829	78.2
5	**1**	0.9504	0.1672	0.956	1.000	0.939	0.967	0.969	0.1144	71.032 ± 21.243	99.6
1	**20**	0.4960	0.0939	0.956	1.000	0.915	0.955	1.000	0.0201	12.863 ± 5.404	56.9
20	**1**	2.7217	0.1274	0.933	0.880	0.963	0.915	0.986	0.1729	108.861 ± 44.017	100

TABLE IV

EXPERIMENTS FOR SELF-ATTENTION IMPROVEMENT TO THE MULTI-TASK SETTING. THE VALUE OF d_k, WHICH REPRESENTS THE SIZE OF Q, K, AND V IN THE ARCHITECTURE WAS OF 32, 64, 128, 256, AND 512. UNDERLINED ARE THE BEST VALUES FOR EACH METRIC.

		Gender Classification						Landmarks Regression		
d_k	Val. Loss	Val CE Loss	Accuracy	Recall	Precision	F1-Score	AUC ROC	Val L1 Loss	NME (%)	$FR_{0.1}$ (%)
32	0.7061	0.1574	0.956	1.000	0.921	0.958	0.994	0.0274	17.310 ± 6.138	77.3
64	0.6853	0.1018	0.956	1.000	0.933	0.963	0.994	0.0292	18.540 ± 7.789	72.0
128	0.6864	0.0601	0.978	0.967	1.000	0.982	1.000	0.0313	19.809 ± 6.348	74.2
256	0.8931	0.1671	0.933	0.958	0.917	0.930	0.994	0.0363	23.346 ± 7.838	77.3
512	0.6386	0.0862	0.933	0.967	0.903	0.930	0.993	0.0276	17.472 ± 5.719	64.4

TABLE V

FINAL VALIDATION RESULTS FOR MODEL 2 (SINGLE TASK LANDMARKS REGRESSION), 3 (MULTI TASK SETTING) AND 4 (SELF-ATTENTION IMPROVEMENT), WHEN SUBJECT TO LONGER TRAINING SETTING OF 50 EPOCHS.

		Gender Classification						Landmarks Regression		
Model	Val. Loss	Val CE Loss	Accuracy	Recall	Precision	F1-Score	AUC ROC	Val L1 Loss	NME (%)	$FR_{0.1}$ (%)
2	0.0098	-	-	-	-	-	-	0.0098	6.417 ± 3.795	16.0
3	0.3532	0.1399	0.911	0.841	0.933	0.883	0.982	0.0111	7.239 ± 3.620	20.4
4	0.3238	0.1064	0.956	1.000	0.897	0.944	0.993	0.0109	7.129 ± 3.552	18.7

TABLE VI

TEST PERFORMANCE OF EACH TRAINED MODEL (1: SINGLE-TASK GENDER CLASSIFICATION, 2: SINGLE-TASK LANDMARKS REGRESSION, 3: MULTI-TASK SETTING, 4: SELF-ATTENTION IMPROVEMENT.)

	Gender Classification					Landmarks Regression			
Model	Accuracy	Recall	Precision	F1-Score	AUC ROC	NME (%)	$FR_{0.1}$ (%)	AUC	Inference Time (ms)
1	0.9600	0.9600	0.9600	0.9600	0.9984	-	-	-	4.7134
2	-	-	-	-	-	6.2618 ± 3.5099	17.6	0.9275	4.7523
3	0.9400	0.9200	0.9583	0.9388	0.9920	6.3802 ± 3.2614	17.6	0.9260	5.6764
4	0.9600	1.000	0.9259	0.9615	0.9952	7.2158 ± 3.8830	20.0	0.9167	5.8463

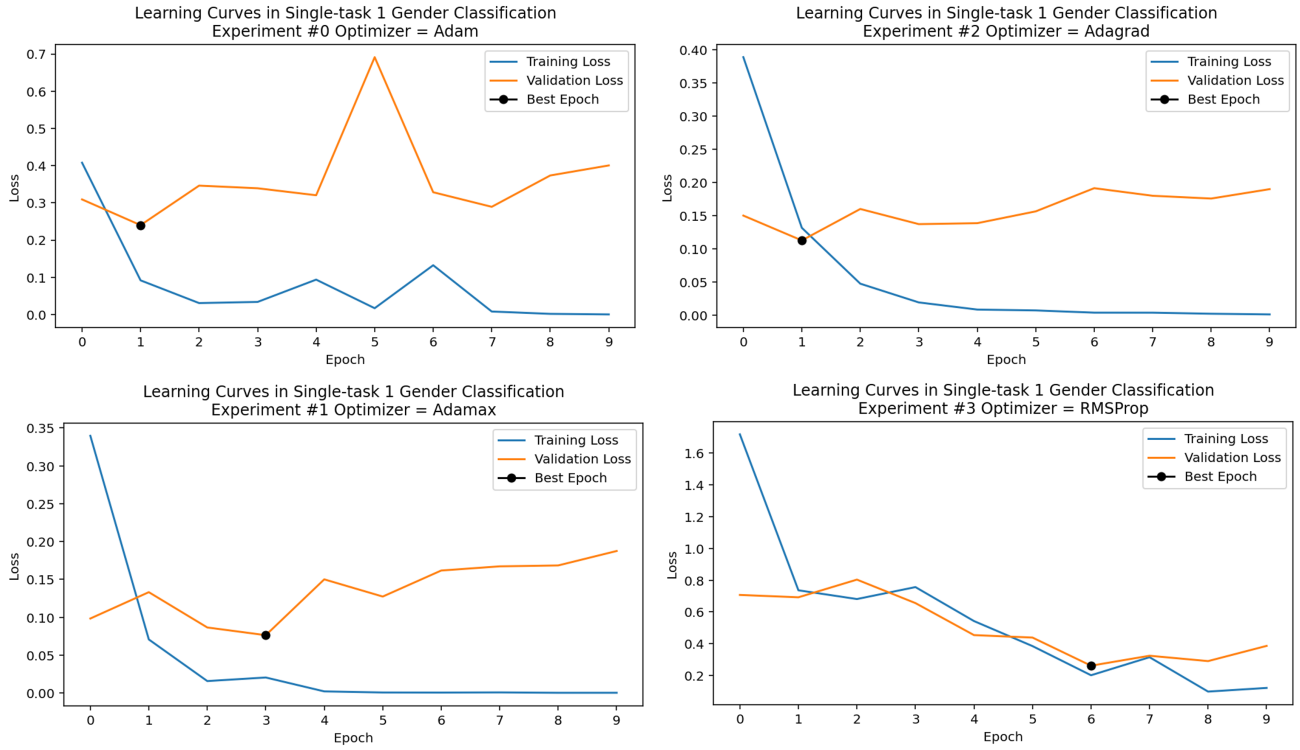

Fig. 8. Learning curves for the first experiment: single-task gender classifcation. The compared optimizers were Adam, Adamax, Adagrad and RMSprop.

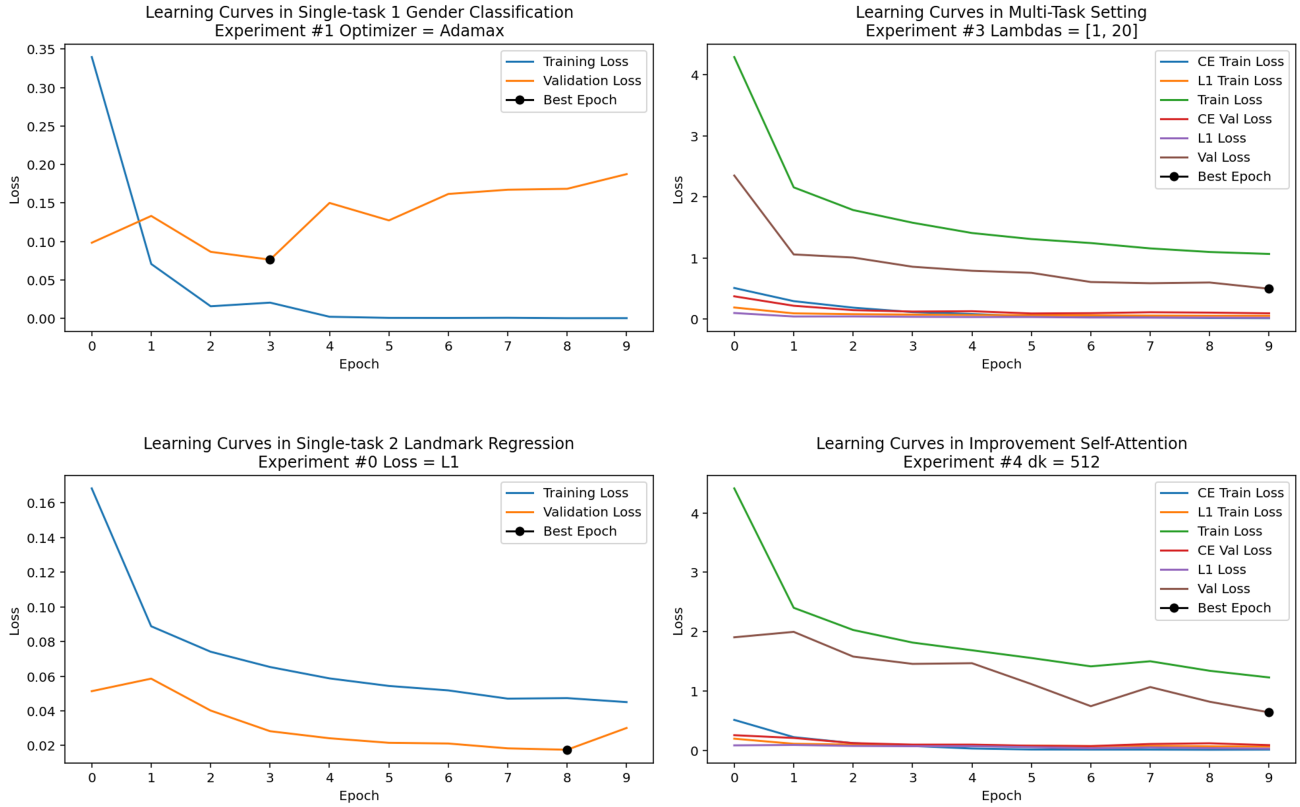

Fig. 9. Learning curves for the best models resulting from the conducted experiments.

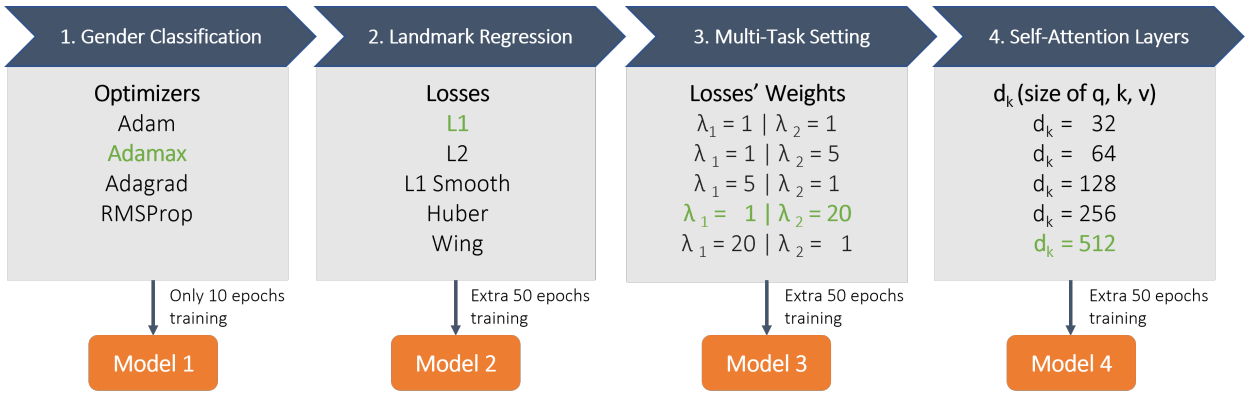

Fig. 10. Conducted experiments, compared variables and flow followed to obtain the 4 final models.

(each is linked to one task) or even make the size of q, k, and v vary among themselves. Moreover, including more attention heads, as often is seen in literature, could produce even more interesting results.

E. 50 Epochs Setting

Figure 9 shows the learning curves for the best model from each of the 4 ran experiments, explored in the previous subsections. It is noticeable that, apart from the single-task gender classification, all models seem to have not converged yet, with the best epochs being achieved in the last epochs. We can conclude that, even though 10 epochs may be enough to conduct the experiments as we did, optimal performances could need longer training settings.

Therefore, for model 2, 3 and 4, we decided to pick the best models from each set of experiments and further train them, before subjecting the models to final test. We considered that better results would probably be achieved, since the landmarks regression task seems to require longer training.

Table V shows the results after training these 3 models with 50 epochs. Indeed, as expected, metrics related to landmarks regression got considerably better and the errors decreased. For example, taking the example of the baseline multi-task model, validation loss reduced from 0.4960 to 0.3535; AUC ROC decreased slightly from 1.000 to 0.982; and, the most significant improvement: NME went from (12.863 ± 5.404)% down to (7.239 ± 3.620)%, and FR from 56.9% to 20.4%. As expected, models seem to benefit from these longer training settings.

As a final remark about the conducted experiments, before getting into testing, it is important to refer that these were quite limited and it would be very interesting to assess the effect of longer training setting right from the beginning, as well as performing the variation of the studied parameters all at the same time. Optimal values could be hidden in combinations that were not tested, since we fixed all parameters but the one that was being analysed individually.

F. Final Testing

Fig. 12. Prediction examples with model 3.

Finally, the best models from the previous experiments were subject to testing, after being subject to 50 epochs training if that was the case. This flow is represented in figure 10.

The final test results can be found in the plots from figure 11, as well in table VI. For both tasks, single-task approaches yield better performances. For gender classification, the single-task architecture achieves an AUC ROC of 0.9984, accuracy of 0.9600 and F1-Score of 0.9600; for landmarks detection, the single-task approach produces a NME of (6.2618 ± 3.5099)%, a $FR_0.1$ of 17.6% and an area under the operating curve of 0.9275.

The multi-task setting is not able to outperform each single-task, as it should if the tasks are highly related, which we hypothesised to be, but maybe are not that much. In fact, looking at a face, the position of the eyes, nose and mouth might not as important to decide someone's gender as we intuitively would say it is. On the other hand, this kind of conclusion would be better supported if we conducted these studies with a bigger dataset (500 images is not much), with longer training settings across all experiments and further tune of the studied hyperparameters.

In any case, yielding reasonable results for both tasks, the multi-task setting shows to be promising in the sense it takes advantages of a common feature extractor to simultaneously solve 2 tasks. It achieves an AUC ROC of 0.9920, accuracy of 0.9400, F1-Score, of 0.9388 (for gender), NME of (6.3802

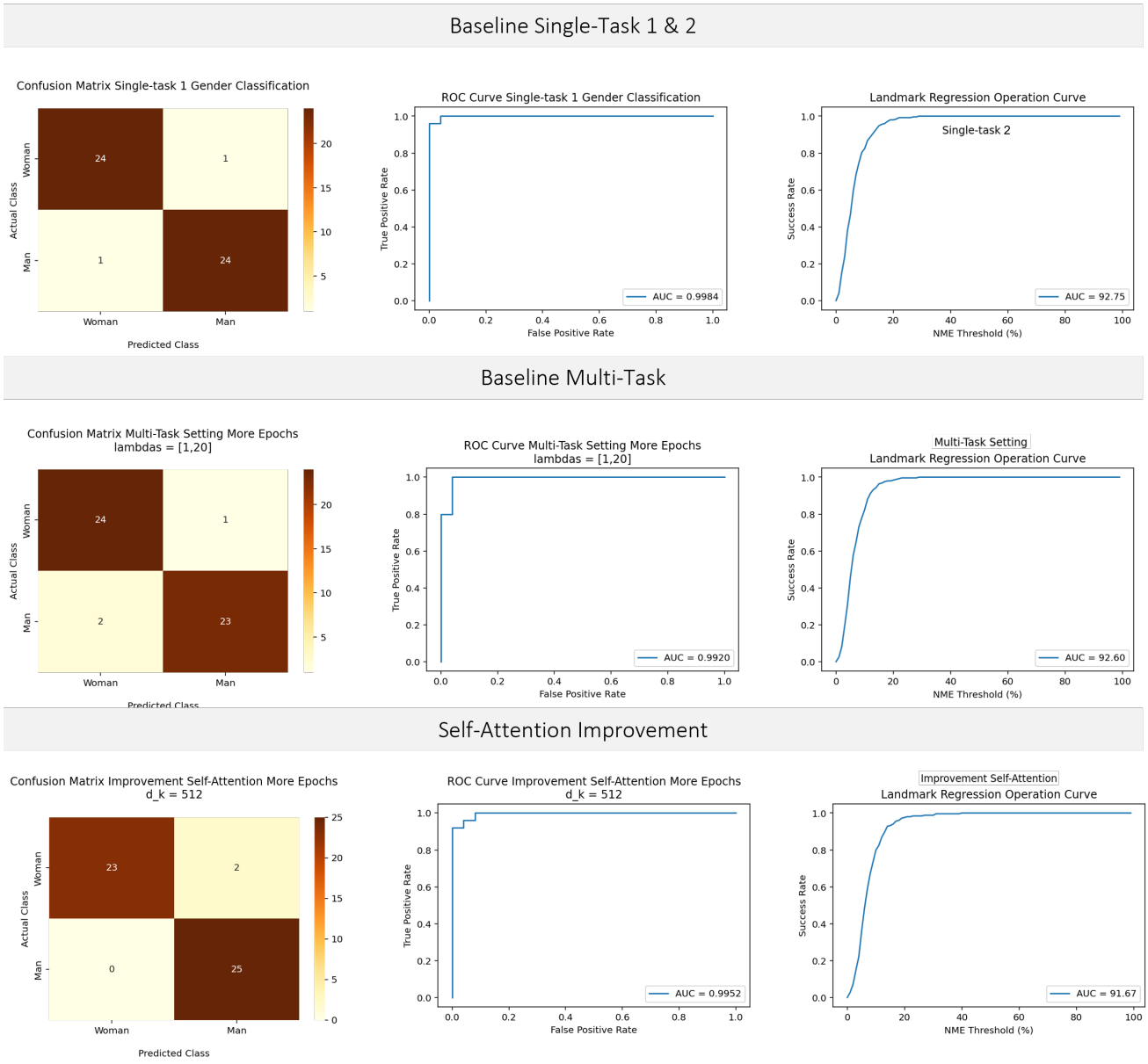

Fig. 11. Test Results in all final models: single task 1 & 2 for gender classification and landmarks regression in the top row; multi-task setting in the mid row; self-attention improvement in the bottom row.

± 3.2614)%, $FR_0.1$ of 17.6% and area under the operating curve of 0.9260 (for landmarks) - all quite close to baseline single-task values. When looking at inference times, single tasks take 9.4657ms in total, while the multi-task setting only takes 5.7664ms.

When it comes to the proposed self-attention layers as an improvement to the baselines, it does not show to outperform the previous models, overall. As seen in table V, our improvement proposal outperforms the multi-task baseline in validation, in both gender and landmarks tasks. However, as we get to the test our improvement proposal shows to have an effect improving the gender classification task, but underperforms in the landmarks detection. It achieves an AUC ROC of 0.9952, accuracy of 0.9600, F1-Score, of 0.9615 (for gender), NME of (7.2158 ± 3.8830)%, $FR_0.1$ of 20.0% and area under the operating curve of 0.9167 (for landmarks). Though we consider it can be a promising improvement, yielding results that we consider good enough, these are not groundbreaking.

In fact, when looking into literature, one can verify that there may be types of attention layer that suit CNNs better [23], [24]. Furthermore, as referred before, since our experiments were not exhaustive, there is the possibility that a better tuning of hyperparameters and longer training can produce results that outperform the baseline (though this is also true in the other way round).

While the gender classification results are close to perfection, with good-looking ROC curves, AUCs close to 1.000 and an accuracy of about 0.9600, the face landmarks detection still has room for improvement, with a NME of (6.2618 ± 3.5099)%. In fact, when looking at the results for the 300W benchmark, the AnchorFace model can achieve a mean NME of 3.12% [25], which is approximately twice as better as our best result (model 2). AnchorFace proposes a novel split-and-aggregate strategy, which is neither a direct regression, nor a heatmap regression (nor a combination of both). In a future work, it could be interesting to test a landmarks detection algorithm that is not a direct regression, which have shown better results (together with higher computational cost, though).

To conclude, as an example of the predictions that we were able to achieve, figure 12 shows some gender and face landmarks prediction. The results seem to be good enough.

IV. Conclusion

In this work, we used image faces to perform gender classification and face landmark localization. For that, we compare a single-task setting with a multi-task one. After that, we proposed and implemented a self-attention-based architectural improvement to the multi-task baseline setting.

For both tasks, single-task approaches yield better performances. For gender classification, the single-task architecture achieves an AUC ROC of 0.9984, accuracy of 0.9600 and F1-Score of 0.9600; for landmarks detection, the single-task approach produces a NME of (6.2618 ± 3.5099)%, a $FR_{0.1}$ of 17.6% and an area under the operating curve of 0.9275.

Nevertheless, the multi-task setting shows very reasonable results that, though do not surpass the baseline single-tasks, are quite close to them. As for the self-attention layers as an improvement to the baselines, these also show viability and to be suitable for the problem, though there is much work to be done and further experiments to conduct. While our improvement proposal outperforms the multi-task baseline setting in gender classification results, it stays a bit behind in the landmarks task.

References

[1] D. H. Ballard and C. M. Brown, *Computer vision*. Englewood Cliffs, N.J: Prentice-Hall, 1982.

[2] A. Voulodimos, N. Doulamis, A. Doulamis, and E. Protopapadakis, "Deep Learning for Computer Vision: A Brief Review," *Computational Intelligence and Neuroscience*, vol. 2018, p. e7068349, Feb. 2018. [Online]. Available: https://www.hindawi.com/journals/cin/2018/7068349/

[3] M. Alghaili, Z. Li, and H. A. Ali, "Deep feature learning for gender classification with covered/camouflaged faces," *IET Image Processing*, vol. 14, no. 15, pp. 3957–3964, 2020. [Online]. Available: https://ietresearch.onlinelibrary.wiley.com/doi/abs/10.1049/iet-ipr.2020.0199

[4] K. Khabarlak and L. Koriashkina, "Fast facial landmark detection and applications: A survey," *Journal of Computer Science and Technology*, vol. 22, no. 1, p. e02, apr 2022.

[5] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[6] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014. [Online]. Available: https://arxiv.org/abs/1409.1556

[7] T. V. Janahiraman and P. Subramaniam, "Gender classification based on asian faces using deep learning," in *2019 IEEE 9th International Conference on System Engineering and Technology (ICSET)*, 2019, pp. 84–89.

[8] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. IEEE, 2009, pp. 248–255.

[9] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds. Curran Associates, Inc., 2019, pp. 8024–8035. [Online]. Available: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

[10] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014. [Online]. Available: https://arxiv.org/abs/1412.6980

[11] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *Journal of Machine Learning Research*, vol. 12, no. 61, pp. 2121–2159, 2011. [Online]. Available: http://jmlr.org/papers/v12/duchi11a.html

[12] A. Graves, "Generating sequences with recurrent neural networks," 2013. [Online]. Available: https://arxiv.org/abs/1308.0850

[13] C.-F. Hsu, C.-C. Lin, T.-Y. Hung, C.-L. Lei, and K.-T. Chen, "A detailed look at cnn-based approaches in facial landmark detection," 2020. [Online]. Available: https://arxiv.org/abs/2005.08649

[14] B. Teixeira, B. Tamersoy, V. Singh, and A. Kapoor, "Adaloss: Adaptive loss function for landmark localization," 2019. [Online]. Available: https://arxiv.org/abs/1908.01070

[15] X. Wang, L. Bo, and F. Li, "Adaptive wing loss for robust face alignment via heatmap regression," *CoRR*, vol. abs/1904.07399, 2019. [Online]. Available: http://arxiv.org/abs/1904.07399

[16] Y. Yan, S. Duffner, P. Phutane, A. Berthelier, C. Blanc, C. Garcia, and T. Chateau, "2d wasserstein loss for robust facial landmark detection," *CoRR*, vol. abs/1911.10572, 2019. [Online]. Available: http://arxiv.org/abs/1911.10572

[17] Z.-H. Feng, J. Kittler, M. Awais, P. Huber, and X.-J. Wu, "Wing loss for robust facial landmark localisation with convolutional neural networks," 2017. [Online]. Available: https://arxiv.org/abs/1711.06753

[18] R. Ranjan, V. M. Patel, and R. Chellappa, "Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition," 2016. [Online]. Available: https://arxiv.org/abs/1603.01249

[19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2017. [Online]. Available: https://arxiv.org/abs/1706.03762

[20] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *CoRR*, vol. abs/2010.11929, 2020. [Online]. Available: https://arxiv.org/abs/2010.11929

[21] I. Bello, B. Zoph, A. Vaswani, J. Shlens, and Q. V. Le, "Attention augmented convolutional networks," 2019. [Online]. Available: https://arxiv.org/abs/1904.09925

[22] B. Wu, C. Xu, X. Dai, A. Wan, P. Zhang, Z. Yan, M. Tomizuka, J. Gonzalez, K. Keutzer, and P. Vajda, "Visual transformers: Token-based image representation and processing for computer vision," 2020. [Online]. Available: https://arxiv.org/abs/2006.03677

[23] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," 2018. [Online]. Available: https://arxiv.org/abs/1807.06521

[24] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," 2017. [Online]. Available: https://arxiv.org/abs/1709.01507

[25] Z. Xu, B. Li, M. Geng, and Y. Yuan, "Anchorface: An anchor-based facial landmark detector across large poses," 2020. [Online]. Available: https://arxiv.org/abs/2007.03221