Response to Reviewer 3 Comments

1. Summary

Thank you very much for taking the time to review this manuscript. Please find the detailed responses below and the corresponding revisions/corrections highlighted/in track changes in the re-submitted files. *[This is only a recommended summary. Please feel free to adjust it. We do suggest maintaining a neutral tone and thanking the reviewers for their contribution although the comments may be negative or off-target. If you disagree with the reviewer's comments please include any concerns you may have in the letter to the Academic Editor.]*

2. Point-by-point response to Comments and Suggestions for Authors

Comments 1: In the review part, we should mention the works devoted to optimizing not only the quality but also the processing speed (10.3103/S87566990210101012X, 10.1007/s11633-023-1423-y), as well as the application of simple neural networks and the study of their quality under different parameters (10.1134/S1054661821030032)

Response 1: **We agree that not only quality but also processing speed should be given important weight in facial analysis task, and thank you for introducing several related works. We could not add the 10.3103/S87566990210101012X paper because we could not confirm it, but for the other two papers, we used them as references in the network architecture section (Section 3.2.1).**

#page 11

"To show that our proposed learning method is feasible, we chose ResNet[58], a network structure commonly used in face recognition-based networks, as a baseline and used the ResNet50 structure for our pretrained model to make the comparison as fair as possible. Furthermore, we used the modified IResNet50 [66], which is used in the recent face recognition-based methods such as ArcFace [4] and Partial FC [5], as the backbone of the multitask learning network. IResNet has no critical differences from ResNet, but we chose and modified the IResNet structure because it allows us to appropriately modify the details of the normalization and activation layers required by each task in face recognition, facial landmark detection, and head pose estimation. While tuning the network architecture for each task, we also considered maximizing shared parameters to minimize inference time loss. To do this, we referenced several face analysis tasks [7, 44, 67, 68] and finally split shared and independent parameters in a 3:1 ratio for each task in the baseline network. The modified IResNet50 consists of four layers, where the first three layers are shared layers, and the last layer is divided into sublayers customized for each task, as shown in Fig. 7."

Comments 2: Line 189: Paragraph (3.1.1) cannot start with a figure. Preliminary text needs to be added.

Response 2: Thank you for your keen observation and response. We moved the sentence to start first at the beginning of the section.

#page 5

3.1.1. Dataset selection strategy Table 1 lists the commonly used datasets for each task. We focus on the ratios of the datasets, number of images per ID, and total amount of data. We selected three datasets as the pseudo-labeled training dataset: two FR datasets, VGGFace2 [28] and MS1MV3 [29], and 300W-LP [22], which is the labeled dataset for HPE and FLD.

Comments 3: Figure 3: Text extends beyond the borders (green box) on the bottom left. Needs to be fixed.

Response 3: Thank you for your comments on this aspect. The text at the bottom of Figure 3 was intended to be outside of the figure, but as you mentioned, it looks more like it's sticking out, so we fixed it.

#page 7

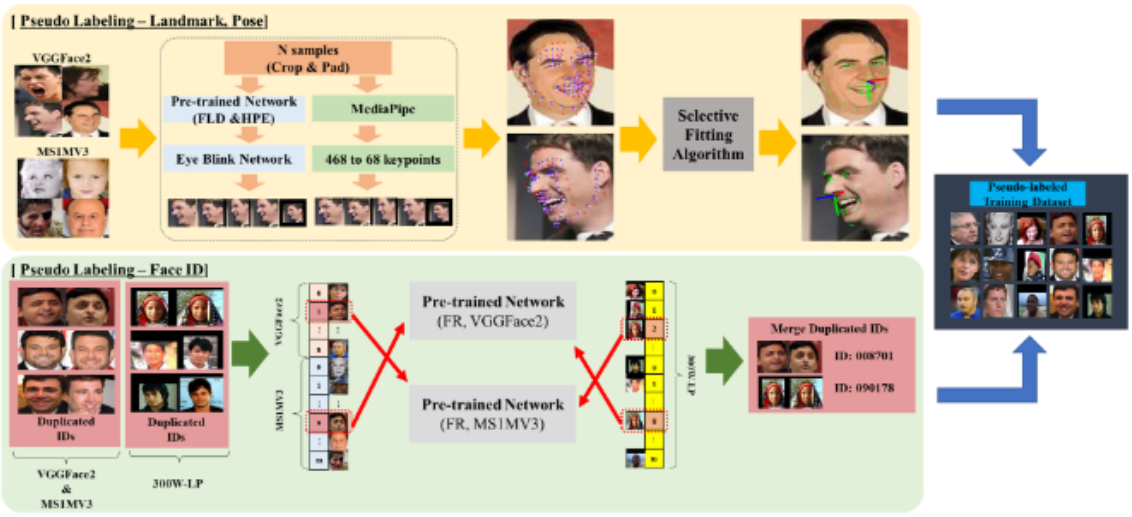

Figure 3. Overview of pseudo-labeled training dataset generation for multitask learning. The upper part illustrates the application of the pseudo-labeling framework of landmark and pose to the face recognition datasets. The bottom part describes the method for pseudo-labeling face IDs for our training datasets, which performs duplicate ID verification and assigns new face ID labels to the VGGFace2 [28], MS1MV3 [29], and 300W-LP [22] datasets.

Comments 4: It would be interesting if the authors added training plots for the loss function they introduced.

Response 4: Thank you for your suggestion about adding a training plot. We thought it was significant to show the training convergence curve of each loss function, but due to the experiments with coefficient variation for each task analyzed in Table 10, we felt that there were too many labels to plot, so we added the loss convergence ratio of each task during the final training process to Table 10.

#page 11

where L_{FR}, L_{FLD}, and L_{HPE} denote the loss functions of face recognition, facial landmark detection, and head pose estimation, respectively. We set the base criterion at 1 for L_{FR} and applied the learning weights λ_1 and λ_2 to L_{FLD} and L_{HPE}, respectively. These weights were adjusted based on the sum of the values represented by the final output of each task. We fine-tuned each task's contribution to the overall multitask framework by choosing λ_1 to be 100 and λ_2 to be 10 to provide an appropriate learning ratio based on the final convergence of each task's loss function. We created our multitask loss function in Equation 1 by computing the total loss value L_{total} by summing all weighted losses from each task.

#page 20

In multitask learning, adjusting the learning weights for the loss function designed for each task is essential for the optimization process. In this paper, we take the loss function for face recognition as the baseline in the unified loss function designed in Equation 1 and set the hyperparameter coefficients of λ_1 and λ_2 to the other tasks, facial landmark detection and head pose estimation, respectively. Table 10 shows the results of experiments conducted to find the optimal hyperparameters. In detail, the loss convergence ratio is the ratio of L_{FR}, $\lambda_1 * L_{FLD}$, and $\lambda_2 * L_{FLD}$ in equation 1 based on the loss value of each task when learning converges to the maximum. The results show that the convergence ratio should be distributed to a certain extent to converge to more optimal multitask learning. Learning by increasing the hyperparameter coefficients for the remaining two tasks is more effective in our proposed multitask learning framework than learning by expanding the hyperparameter coefficients for face recognition, which constitutes most of the pseudo-labeled training dataset.

Table 10. Experimental results showing the impact of multitask learning hyperparameters λ_1 and λ_2 on test accuracy with various benchmarks. These hyperparameters are used in equation 3 to balance the contribution of each task in the overall multitask learning framework. To check the change caused by λ_1 and λ_2, we obtained the loss convergence ratio based on the convergence value of the loss function of each task.

λ_1	λ_2	IJB-C(1e-4)	AFLW2000-3D(NMSE)	AFLW2000-3D(MAE)	Loss Convergence Ratio
0.1	0.1	94.22	5.12	4.14	0.98:0.00:0.02
0.1	1	94.21	4.77	3.42	0.86:0.00:0.14
0.1	10	94.22	4.56	3.35	0.37:0.00:0.63
0.1	100	94.23	4.98	3.35	0.08:0.00:0.92
1	0.1	94.20	4.61	3.98	0.96:0.01:0.03
1	1	94.19	4.21	3.41	0.85:0.01:0.14
1	10	94.12	4.37	3.39	0.37:0.01:0.62
1	100	94.18	4.27	3.38	0.08:0.01:0.91
10	0.1	94.88	3.99	3.94	0.94:0.05:0.02
10	1	94.19	3.65	3.54	0.82:0.04:0.14
10	10	94.17	3.61	3.33	0.35:0.02:0.63
10	100	94.17	3.60	3.33	0.06:0.00:0.93
100	0.1	94.18	3.74	3.99	0.67:0.32:0.01
100	1	94.19	3.55	3.59	0.61:0.29:0.10
100	**10**	**94.17**	**3.53**	**3.32**	0.31:0.15:0.54
100	100	94.17	3.54	3.32	0.04:0.03:0.94

*[*1] Zhu, Jun-Yan, et al. "Unpaired image-to-image translation using cycle-consistent adversarial networks." Proceedings of the IEEE international conference on computer vision. 2017.*

Comments 5: Performance (FPS) data for tables 3, 4 and 5(6) should be added.
Response 5: Thank you for your thoughtful comment adding performance for FPS. We wanted to calculate and add FPS for all of the comparative papers, but we were not able to add all of them, and for those references that had executable code, we measured the FPS performance and reflected it in each table.

However, we excluded table 4 because we thought it would be better to cover it in table 8, and we added its FPS to table 8.

We also added a brief analysis of it, and we think that this comparison of FPS performance helps the reader to understand a little more the research direction that we want to present in this manuscript.

#page 16

Table 3. Evaluation results (%) for face recognition datasets LFW, CFP-FP, AgeDB, and IJB-C. Using TAR@FAR=1e-4 for IJB-C. **Bold** denotes the best performance, and higher is better.

Method	Dataset	LFW	CFP-FP	AgeDB	IJB-C	FPS
SphereFace [37]	CASIA	99.42	-	-	-	-
CosFace [38]	CASIA	99.73	-	-	-	-
SphereFace (Re-Imp)	MS1MV2	99.67	98.46	98.17	91.77	-
CosFace (Re-Imp)	MS1MV2	99.78	98.26	98.17	95.56	-
CircleLoss [70]	MS1M	99.73	96.02	-	93.95	-
CurricularFace [71]	MS1MV2	99.80	98.37	98.32	96.10	329.8
ArcFace [4]	MS1MV2	99.82	98.49	98.05	96.03	331.5
ArcFace [4]	MS1MV3	99.83	**99.03**	**98.17**	96.5	331.5
MagFace [72]	MS1MV2	99.83	98.46	98.17	95.97	254.3
Partial FC [5]	MS1MV3	99.85	98.7	98.11	96.08	255.1
SwinFace [21]	MS1MV2	**99.87**	98.60	98.15	**96.73**	70.3
Ours (pretrained) - ResNet50	VGGFace2	99.33	92.38	94.20	88.05	321.7
Ours (pretrained) - ResNet50	MS1MV3	99.85	98.70	98.11	96.08	321.7
Ours (STL) - IResNet50	PLTD	99.72	97.73	96.90	94.05	253.2
Ours (MTL) - IResNet50-Multi	PLTD	99.70	97.26	96.53	**94.17**	221.2
Ours (MTL) - IResNet100-Multi	PLTD	99.68	97.22	96.29	94.09	112.0

These results indicate that filtering and learning from them is feasible even if the pseudo-labeled dataset incorporates some outliers. It also exhibits the potential to effortlessly integrate various facial analysis tasks beyond the proposed methods, highlighting the versatility and adaptability of our approach despite anomalies and discrepancies within the data. In addition, we compared the inference time with similar papers under identical conditions and conducted a frame-per-second (fps) comparison for the ResNet50 and IResNet50 models used in the comparison papers. Unlike the recent trend of using arrays of multiple GPUs for training face recognition methods, our fps comparison was conducted using just a single GPU. Despite performing multitask computations, our proposed method performed well in terms of fps.

#page 17

Table 5. Mean absolute error (MAE) comparison on AFLW2000-3D [22] for Head Pose Estimation. **Bold** denotes the best performance and lower is better.

Methods	yaw	pitch	roll	MAE	FPS
FAN (12 points) [48]	6.36	12.3	8.71	9.12	-
HopeNet [73]	6.47	6.56	5.44	6.16	323.2
FSANet [44]	4.5	6.08	4.64	5.07	389.6
3DDFA-TPAMI [54]	4.33	5.98	4.30	4.87	-
3DDFA-V2 [53]	4.06	5.26	3.48	4.27	-
QuatNet [74]	3.97	5.62	3.92	4.15	-
TriNet [45]	4.2	5.77	4.04	3.97	311.3
RankPose [46]	2.99	4.75	3.25	3.66	291.7
SynergyNet [7]	3.42	**4.09**	**2.55**	3.35	128.1
Ours (Pretrained)	2.89	4.77	3.35	3.67	320.8
Ours (MTL)	**2.80**	4.27	2.9	**3.32**	221.2

#page 18

Table 6. Mean absolute error (MAE) comparison on BIWI [42] for Head Pose Estimation. **Bold** denotes best performance.

Methods	yaw	pitch	roll	MAE	FPS
3D-FAN [48]	8.53	7.48	7.63	7.89	-
HopeNet [73]	4.81	6.61	3.27	4.90	323.2
FSANet [44]	4.27	4.96	2.76	4.00	389.6
QuatNet [74]	4.01	5.49	2.94	4.15	-
TriNet [45]	4.11	**4.76**	3.05	3.97	311.3
RankPose [46]	3.59	4.77	2.76	3.71	291.7
Ours (pretrained)	4.34	5.18	2.61	4.04	320.8
Ours (MTL)	**3.23**	5.03	**2.36**	**3.54**	221.2

#page 19

Table 7. Normalized Mean Square Error (NMSE) percentage comparison on AFLW2000-3D [22] for 3D facial landmark detection across three yaw angle ranges: 0° to 30°, 30° to 60°, 60° to 90°, and overall mean error. **Bold** denotes the best performance, and lower is better.

Methods	$0° \sim 30°$	$30° \sim 60°$	$60° \sim 90°$	mean	FPS
3DDFA [22]	3.43	4.24	7.17	4.94	-
3DSTN [75]	3.15	4.33	5.98	4.49	-
3D-FAN [48]	3.16	3.53	4.6	3.76	84.3
PRNet [76]	2.75	3.51	4.61	3.62	261.3
3DDFA-PAMI [54]	2.84	3.57	4.96	3.79	-
3DDFA-v2 [53]	2.63	3.42	4.48	3.51	-
SynergyNet [7]	2.65	**3.30**	**4.27**	**3.41**	128.1
Ours (pretrained)	3.08	3.74	4.53	3.78	322.5
Ours (MTL)	**2.62**	3.45	4.51	3.53	221.2

#page 19

Table 8. Comparison of the performance of the baseline model in single-task learning (STL) and multitask learning (MTL) configurations across a variety of tasks, including face recognition (FR), head pose estimation (HPE), and facial landmark detection (FLD).

Methods	FR	HPE	FLD	IJB-C	AFLW2000-3D (MSE)	AFLW2000-3D (MAE)	Params	FPS
STL	✓	-	-	94.05	-	-	24.6M	321.7
STL	-	✓	-	-	3.67	-	23.5M	320.9
STL	-	-	✓	-	-	3.78	23.8M	322.5
MTL	✓	✓	-	94.16	3.54	-	56.8M	243.8
MTL	✓	-	✓	**94.17**	-	3.34	60.1M	237.9
MTL	-	✓	✓	-	**3.52**	3.33	47.3M	252.7
MTL	✓	✓	✓	**94.17**	3.53	**3.32**	73.3M	221.2