Supplementary

This supplementary document provides additional qualitative and quantitative results that provide further insights into the results discussed in the main paper. A more detailed description of the datasets is given in § I and the pose clusters discussed in § 3.1 of the main paper are visualized in § II. Additional experimental results and visualizations that show the effectiveness of different components of our framework are given in § III and § IV, respectively.

I. Datasets

I.1. IMDB

We created the IMDB database to train the actor classifier in the movie scenario. The scenario is different from most of the person recognition in that the test set contains a single movie with lesser variation in appearance between multiple instances of an actor in terms of age, style of clothing, etc. We assume that there are no labeled images within the movie and hence training data is not a part of the movie. To create a training set, the images are collected from the IMDB profile[2] of actors appearing in the movie, which are then manually cropped and annotated. Few images from IMDB database are shown in Figure 12. We relied on text tags associated with photos for annotation whenever the photos contain multiple confusing identities. Apart from illumination, resolution, and pose variations, there is a large age variations among IMDB instances. In addition, there is a large domain contrast between IMDB and Hannah test set in terms of lighting, camera and imaging conditions. This creates a more challenging setting to match identites between IMDB and Hannah instances.

I.2. Soccer

Soccer is another scenario where there are a significant number of frames in which the face is not visible and the subjects are often occluded by other players. We show more examples from our soccer dataset in Figure 14. In many instances, head is largely occluded, and in back-view unlike PIPA and Hannah instances, which contain visible head and torso regions. Also, soccer instances exhibit large body deformations, are of low resolution with significant blur. The soccer dataset therefore offers different kinds of challenges for recognition that are not seen in PIPA and Hannah.

II. Pose clusters

We obtain a set of prominent views to facilitate pose-specific representations as discussed in § 3.1. To achieve this, we annotated 14 body keypoints for 29,223 PIPA train instances which are then used for clustering. More

[2] http://www.imdb.com/title/tt0091167/fullcredits

Figure 12: **IMDB:** Each row shows few images of an actor from the dataset. We used IMDB dataset to train classifiers for actor recognition in the Hannah movie.

Figure 13: **Pose clusters:** Each row from top to bottom shows people from PIPA with particular body orientation clustered using orientation and keypoint visibility features.

Figure 14: Images from soccer dataset. It offers a challenging person recognition scenario due to low resolution, high occlusion, deformation and motion blur exhibited by soccer instances.

examples of our pose clusters are shown in Figure 13. Each row from top to bottom contain images from `right`, `semi-right`, `frontal`, `semi-left`, `left`, `back` and `partial body` views. The orientation and keypoint visibility features produced tight clusters containing images with particular body orientation. The last cluster captures the instances with partial upper body such as head or shoulder, etc, in the images that are commonly seen in social media photos and movies. While we considered seven prominent views in this work, we note that generating a large number of views can be helpful, provided there are enough training samples in each cluster to train the convnets.

III. Quantitative Results and Analysis

We provide more insightful results that help to understand merits and challenges of different recognition settings that are considered.

Recognition per subject: Figure 15 shows the number of images for each actor in IMDB and Hannah test sets along with their individual recognition performances. We observe that, for those subjects with sufficiently large number of training instances (*Michael Caine, Barbara Harshey, Woody Allen, Julia Louis-Dreyfus, and Mia Farrow*), the performance is high as expected. For subjects with less than 20 training instances, the performance is very low.

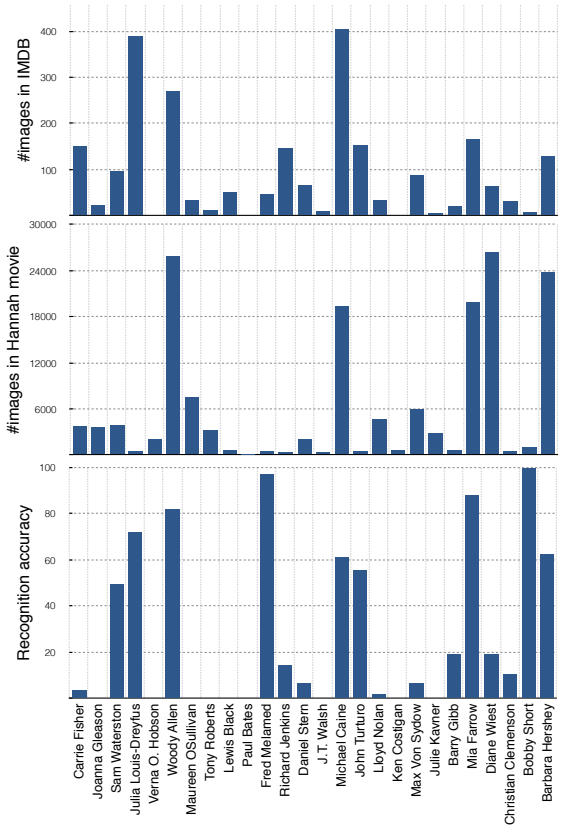

Figure 15: Number of images for each actor in (top) IMDB and (middle) Hannah movie test set. We show the (bottom) recognition performance of each actor on the test set.

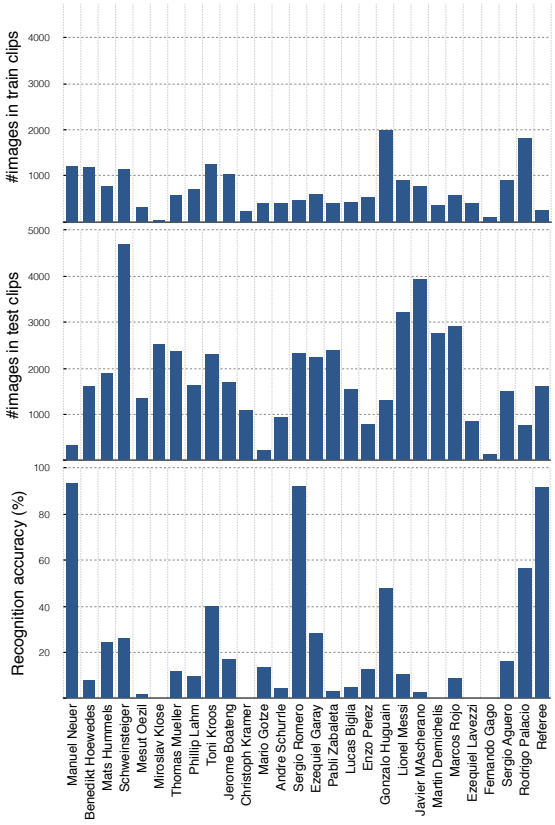

Figure 16: Number of images for each player in the training (top) and test (middle) split of Soccer dataset. We also show the (bottom) recognition performance of each player.

However, whenever there is a large difference in age between train and test instances (*Carrie Fisher, Dianne West, Richard Jenkins*), the performance is poor despite having enough training examples.

Similarly, we show the statistics of soccer players along with their individual performances in Figure 16. We see a similar trend of high performance for subjects (*Gonzalo Huguain* and *Rodrigo Palacio*) with sufficient training instances. We also observe a near 100% accuracy for goal keepers (*Manuel Neuer* and *Sergio Romero*) and the referee due to clothing cues, which are discussed next.

Recognition performance of top subjects: We compare the recognition performance of various approaches on 5 most occurring movie and soccer subjects in Figure 17 and Figure 18, respectively. Our approach reaches an accuracy of 61.17% on top actors, which is significantly better than `naeil`. Note that the overall performance of `naeil` with 17 models is comparable to head and upper body. Unlike photo-albums, clues such as scene and human attributes like age, glasses, and hair color are less useful in the movie setting. For actors with less change in appearance over time (*Michael Caine* and *Woody Allen*: See row three and five in Figure 12), face is found to be extremely informative and robust compared to head.

On the soccer dataset, the overall performance is poor for all the approaches. This suggest to develop better representations that are able to recognize people at a distance.

How informative is clothing? Though it is intuitively obvious that clothing helps in recognition, a qualitative evaluation is not done previously. We perform such a study using the soccer dataset. We show the performance of different approaches on three subjects (*Manuel Neuer*, *Sergio Romero* and *Referee*) with unique clothing in Figure 19. The first two subjects are the goal keepers of the Germany and Argentina, respectively.

As seen in Figure 19, upper body region, which is often less informative compared to head, outperforms head by a large margin due to clothing. The concatenation of head and upper body obtained through separate training is worse than upper body feature alone. On the other hand, the concatenation of features using jointly trained model is more robust and performs much better as it provide more flexibility to focus on selective regions. Finally, the overall performance of pose aware models and `naeil` are identical.

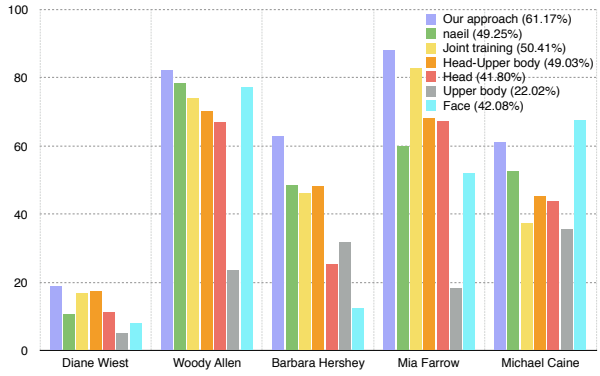

Figure 17: Recognition performance of five lead actors in Hannah dataset.

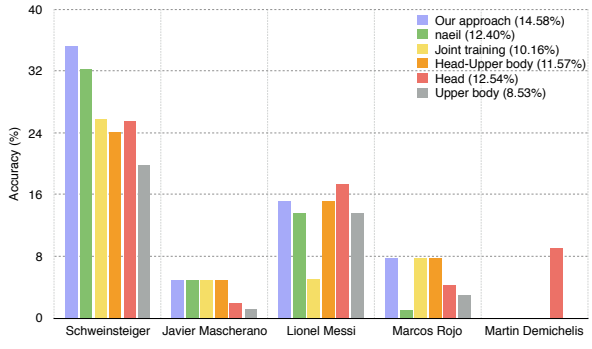

Figure 18: Recognition performance of five most occurring players in Soccer dataset.

It is interesting to note that, convnets that are trained for identity recognition can distinguish clothing without any explicit modeling or hand-crafted features [29].

Confusion between identities: We show the recognition confusion matrix for Hannah and Soccer datasets in Figure 21 and Figure 22 respectively, with and without tracking. We notice two important points related to gender and clothing. As seen from Figure 21, female subjects are mostly getting confused with female subjects, and similarly the male subjects are confused with male subjects. In Figure 22, we notice that players from each team are mislabeled with the members from the same team. These studies show the effectiveness of convnets in capturing human attributes without any explicit training. Finally, majority voting over a track helps to produce consistent predictions.

Domain gap: To understand the effect of domain contrast between train and test instances, we conduct an experiment adding different number of Hannah instances per subject to the IMDB training gallery. The results are shown in Figure 20. As seen from the graph, the addition of even a few instances from the test domain results in a very large improvement in the recognition performance.

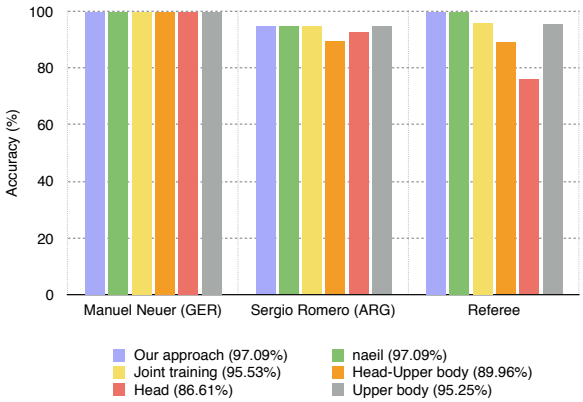

Figure 19: Effect of clothing on recognition.

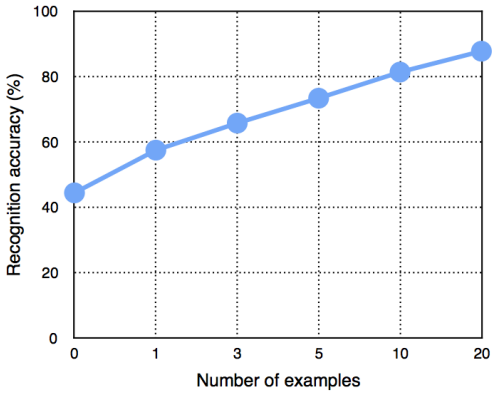

Figure 20: Recognition performance of Hannah movie set using IMDB plus samples from the Hannah test set.

IV. Qualitative Results

We show some qualitative results in Figures 23 to 27. Figure 23 shows the success and failure cases of joint training and separate training of body regions. We notice an over-influence of clothing while using separately trained and concatenated regional features, compared to the jointly training features. In Figure 24, we show the effectiveness of using multiple classifiers from each *PSM*. As seen in the figure, the concatenated head and upper body features ($\mathcal{F}$) may predict incorrect labels even when one (or two) of these features predict correctly, due to the over influence of less informative body region. Combining these three features is found to be more robust.

We show the top scoring predictions obtained from each pose-specific *PSM* in Figure 25. It clearly shows how each *PSM* helps in the prediction of instances in that particular pose when the base model is unable to predict correctly. Finally, we show the success and failure cases of our approach on Hannah and Soccer datasets in Figure 26 and Figure 27 respectively, and compare with the `naeil`.

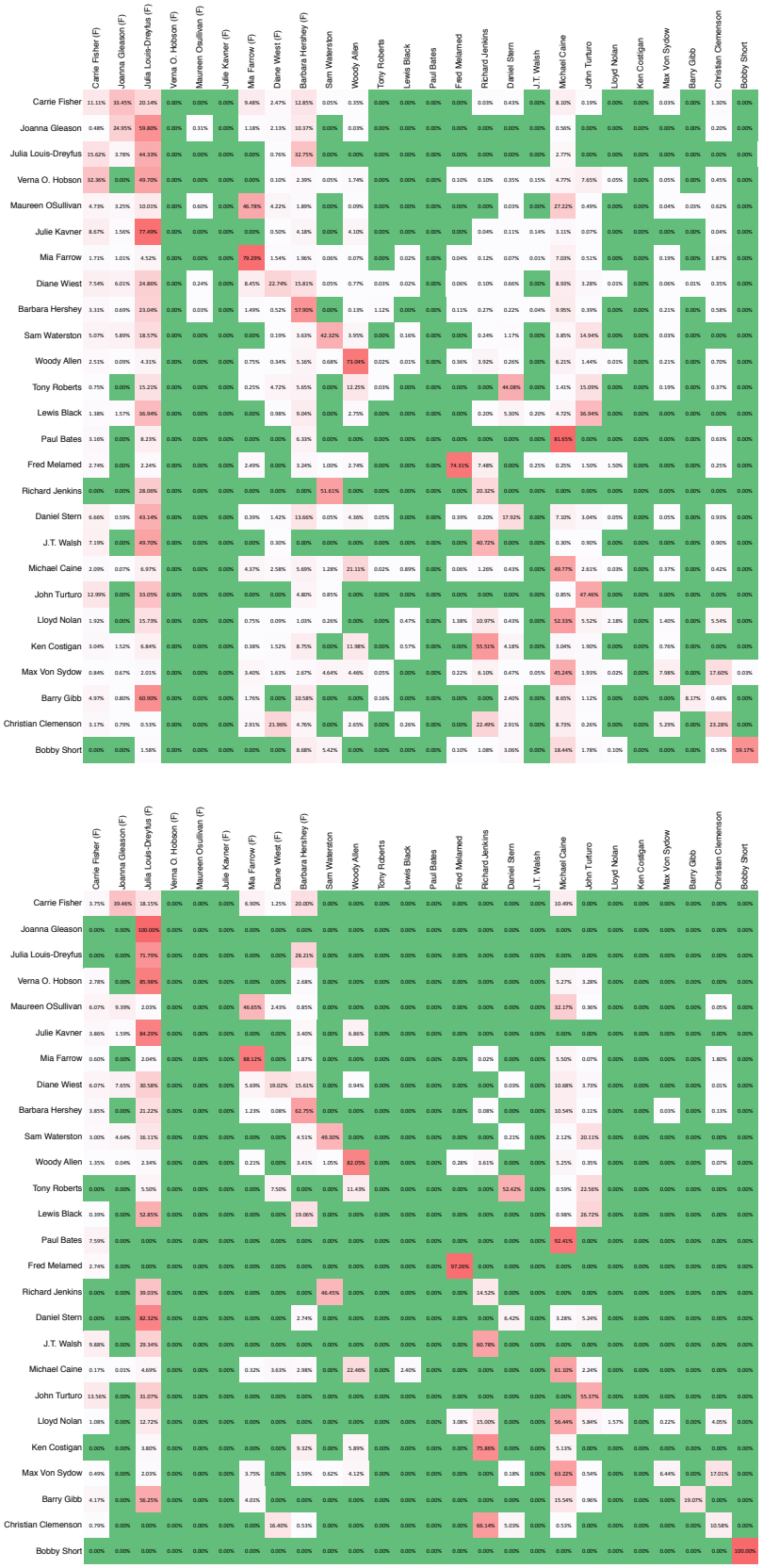

Figure 21: **Confusion matrix** on Hannah dataset (top) with and (bottom) without tracking.

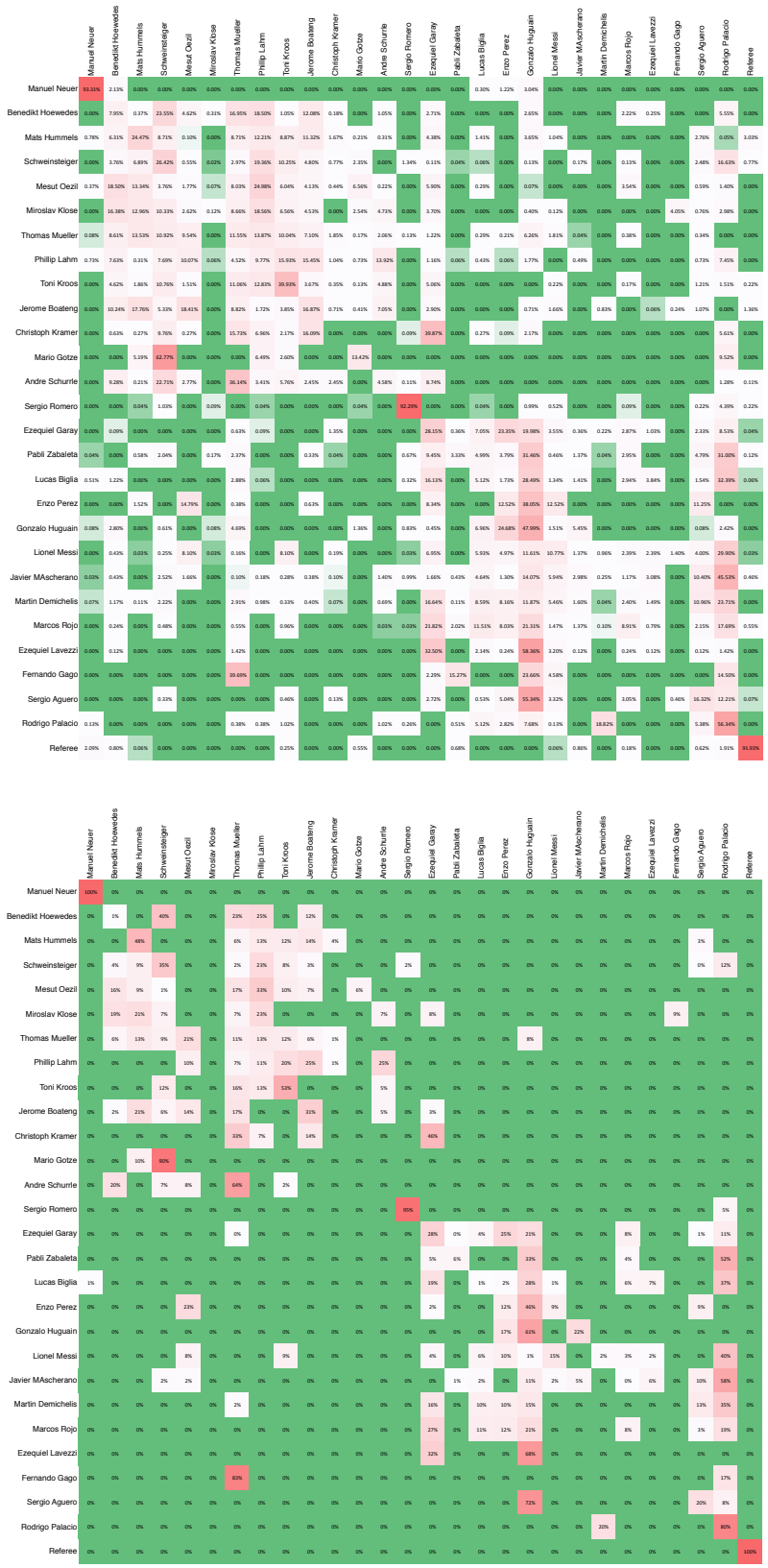

Figure 22: **Confusion matrix** on Soccer dataset (top) with and (bottom) without tracking.

Figure 23: Success and failure cases of separate and joint training of body regions on PIPA dataset. Column one shows the test images and the column two shows the training images belonging to the predicted subject. (Left) shows the success and failure case of joint training (JT) and separate training (ST), respectively and the reverse is shown in (right).

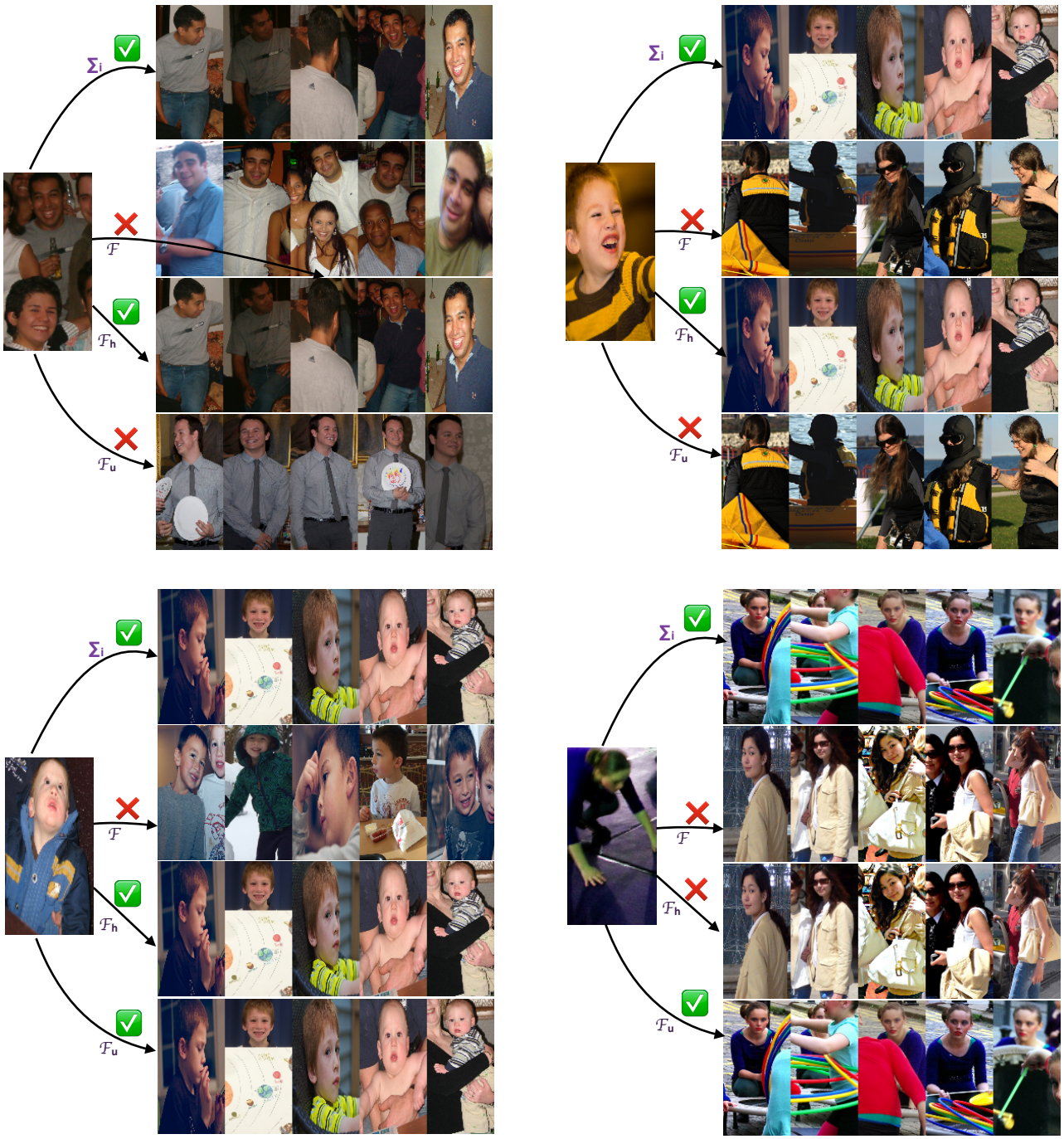

Figure 24: **Effectiveness of multiple classifiers from each *PSM*:** Column one shows the PIPA test images and the column two shows the training images belonging to the predicted subject using different approaches. The four approaches considered are the classifiers trained on head ($\mathcal{F}_h$) and upper body ($\mathcal{F}_u$) features, a classifier trained on concatenated head and upper body ($\mathcal{F}$) feature, and linear combination of three classifiers ($\sum_i$) trained on these features. It clearly shows that it is advantageous to consider individual classifiers trained on regional features and their combination for improved performance.

Figure 25: **Success cases of pose-specific models (*PSM*s) on PIPA dataset.** Each row shows the success predictions of our approach where the improvement is obtained primarily due to the specific-pose model *i.e.*, `base` model wrongly predicts but `base` + correct `PSM` predicts correctly. Green and yellow boxes indicate the success and failure result of `naeil` respectively.

Figure 26: Comparison of our approach with `naeil` on Hannah dataset. Column one shows the test images and the column two shows the training images belonging to the predicted subject. (Left) in green shows the success case of our approach and the failure case of `naeil`. (Right) in red shows the failure case of our approach and the success case of `naeil`.

Figure 27: Comparison of our approach with `naeil` on Soccer dataset. Column one shows the test images and the column two shows the training images belonging to the predicted subject. (Left) in green shows the success case of our approach and the failure case of `naeil`. (Right) in red shows the failure case of our approach and the success case of `naeil`.