Red Carpet to Fight Club: Partially-supervised Domain Transfer for Face Recognition in Violent Videos

Supplementary Material

Yunus Can Bilge*[1], Mehmet Kerim Yucel*[1], Ramazan Gokberk Cinbis[2], Nazli Ikizler-Cinbis[1], and Pinar Duygulu[1]

[1]*Department of Computer Engineering, Hacettepe University*
[2]*Department of Computer Engineering, Middle East Technical University*
{*yunuscanbilge,nazli,pinar*}*@cs.hacettepe.edu.tr, mkerimyucel@hacettepe.edu.tr, gcinbis@ceng.metu.edu.tr*

1. Outline

This supplementary material document presents additional qualitative results for the proposed Attentive Temporal Pooling layer and detailed statistics for the dataset splits.

2. Qualitative results for Attentive Temporal Pooling

In Section 4.2 of the main manuscript, **attentive temporal pooling** (ATP) is introduced for exploiting the hidden pose information in a trainable fashion to extract useful information in the noisy sequences of video frames. In the discussion following Eq. 8, we note that ATP scheme effectively assigns attention weights to the frames in a facial image sequence, where all unnormalized per-frame weights are given by $\Gamma(v)\mathbf{1}_K$. Here, $\Gamma(v)$ is the attention function computing a $|v| \times K$ attention matrix, as defined in Eq. 8 in the manuscript.

In Figure 2, we give qualitative examples for the attention distributions generated by ATP. On top of the each image, we show the relative attention scores obtained according to $\Gamma(v)\mathbf{1}_K$. Overall, we observe that ATP gives higher weights to cleaner frames compared to the blurry ones.

In addition to temporal adaptation, we also note that while we have investigated several other variations for the self-attention, such as using affine transforms for the key, query and/or value embeddings and stacking more than one self-attention layers, we have not observed any significant improvements.

*equal contribution

Figure 1. Images from train,validation and test sets (left to right) for four individuals. Age discrepancy and other changes in apperance can be seen.

3. Dataset splits

Training, test and validation sets for WildestFaces are split person-wise. Intra-class variance is inherently amplified with this split procedure; an actor can be represented with their early career videos in the training set whereas

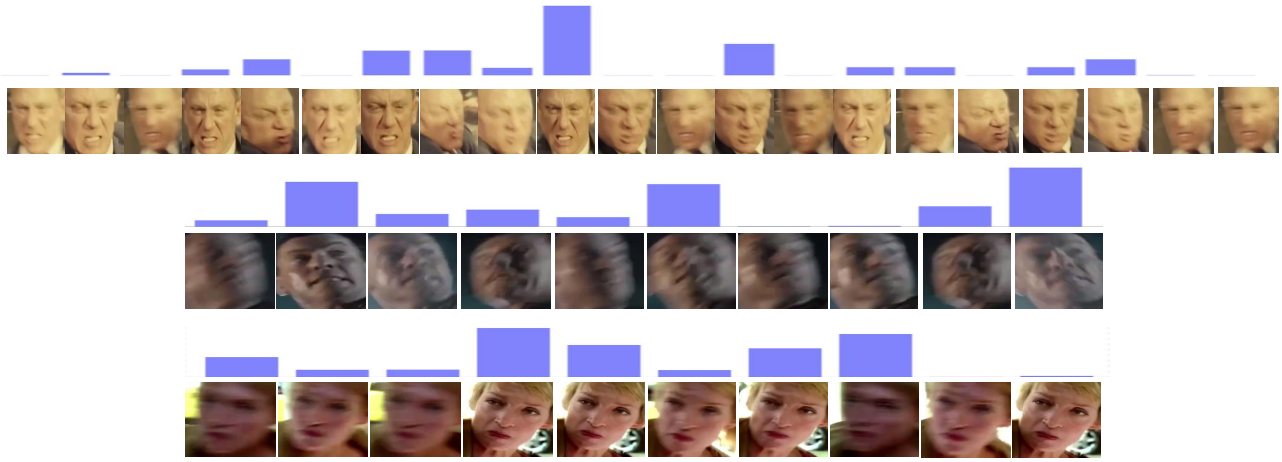

Figure 2. Per-frame attention scores obtained using ATP.

validation and test sets can be represented with their late career videos, introducing a potentially significant age variance (see Figure 1 for qualitative examples).

In Figure 3, Figure 4, and Figure 5, we present the person (*i.e*. face class) counts for the train, validation and test splits of the proposed WildestFaces dataset. We note that the list of people appearing in the training set (40 classes) is a strict subset of those in the validation set (40+10 classes). Similarly, the list of people appearing in the validation set is a strict subset of those in the test set (40+10+14 classes).

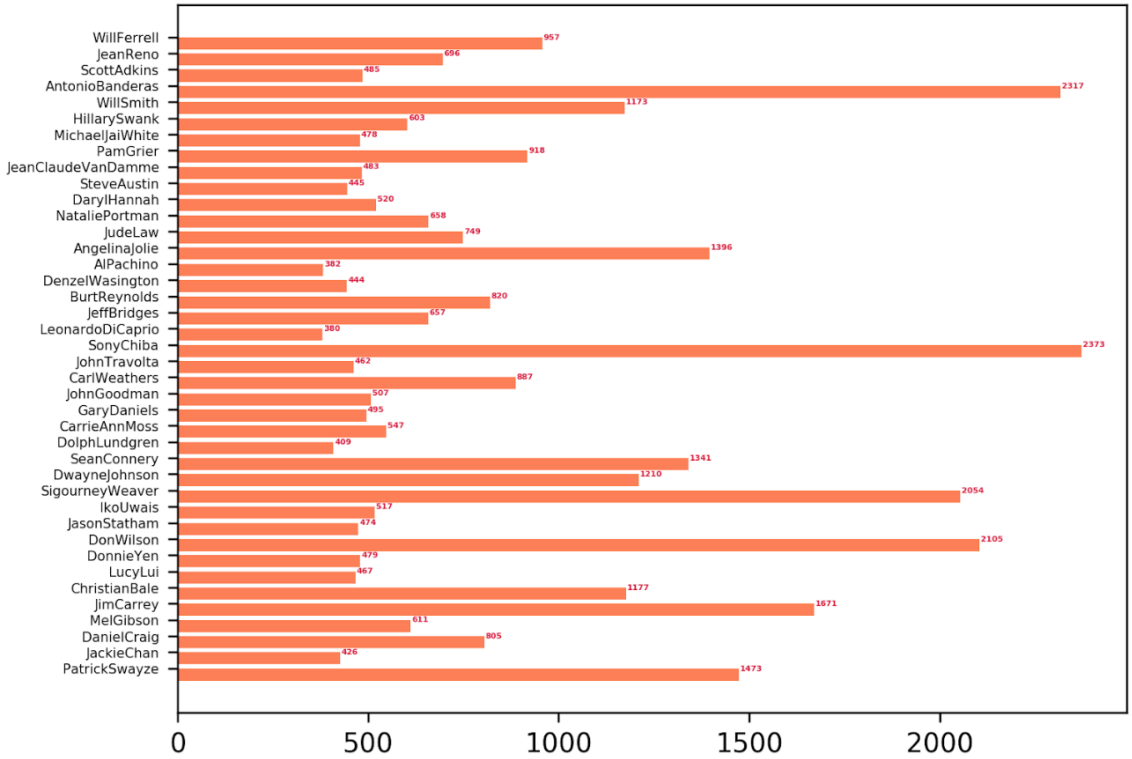

Figure 3. Number of instances of each class in the training split of the WildestFaces dataset.

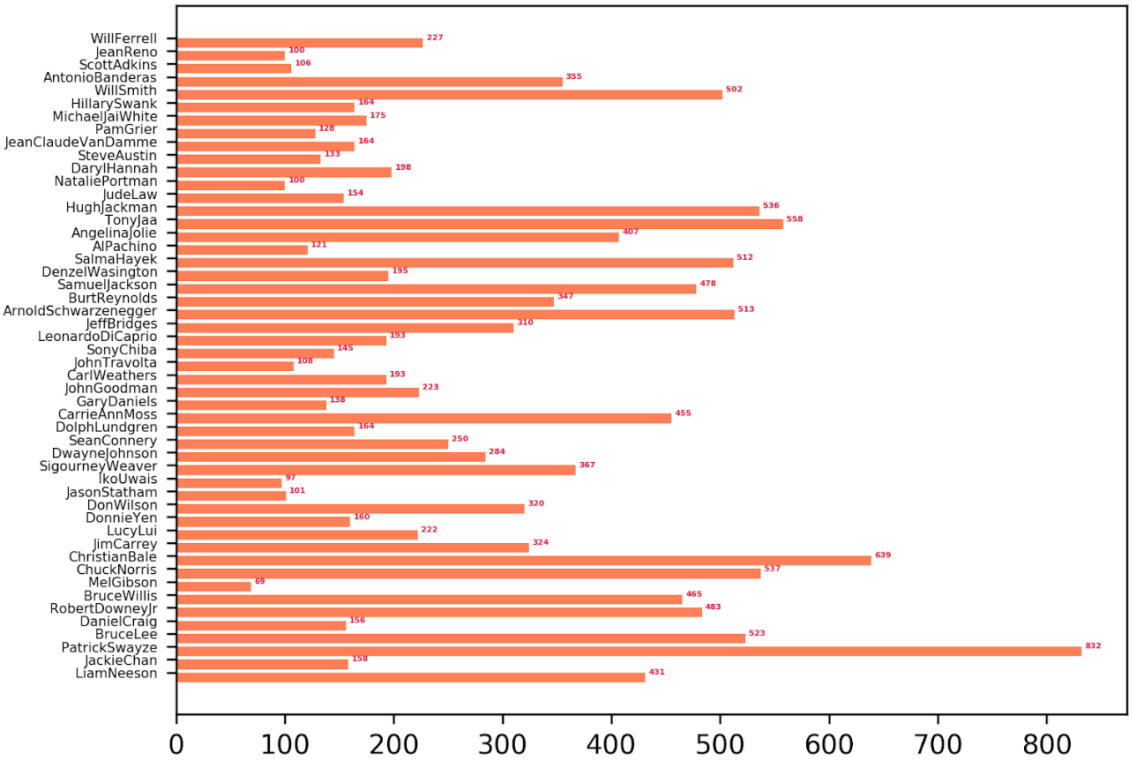

Figure 4. Number of instances of each class in the validation split of the WildestFaces dataset.

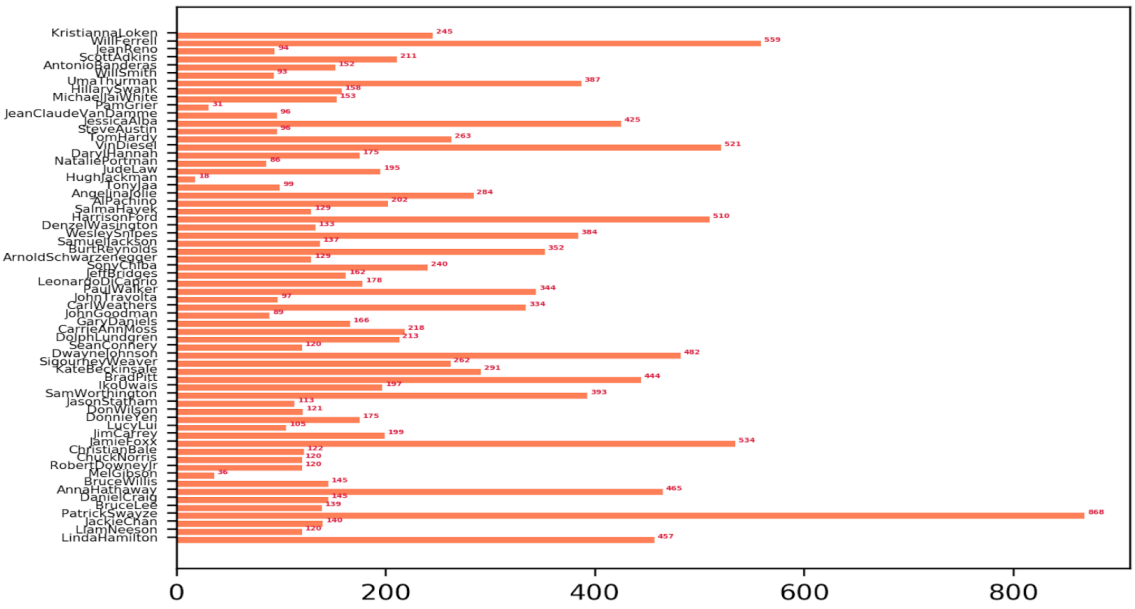

Figure 5. Number of instances of each class in the test split of the WildestFaces dataset.