Neural Decision Forests for Semantic Image Labelling

Samuel Rota Bulò
Fondazione Bruno Kessler
Trento, Italy
rotabulo@fbk.eu

Peter Kontschieder
Microsoft Research
Cambridge, UK
pekontsc@microsoft.com

Supplementary Material

In this document we provide the following supplementary contributions:

- In Section 1 we show that the split function quality measures $Q_{\text{Reg}}(\Theta)$ and $Q(\Theta)$ proposed in Eq. (6) and (4) of our CVPR contribution are optimal in terms of ℓ_1-regularized and non-regularized empirical risk under log-loss, respectively;
- In Section 2 we prove that $Q(\Theta)$ is substantially equivalent to the *information gain* criterion if the routing function $f(\cdot;\Theta)$ is binary;
- In Section 3 we prove that the update rule for the child posteriors $\boldsymbol{\pi}$ in Eq. (8) of our CVPR contribution monotonically increases the likelihood and can thus be used to compute $Q(\Theta)$ and $Q_{\text{Reg}}(\Theta)$.
- In Section 4 we show result images for the Labelled Faces in the Wild (LFW) dataset, corresponding to the evaluation results of the main document. In addition, we show the development of the Jaccard score as a function of the forest ensemble size.

1. Split Function Quality Measures and Empirical Risk Minimization

Let $t \in \mathcal{F}$ be a tree of a decision forest $\mathcal{F}$. The *empirical risk* under log-loss for tree $t \in \mathcal{F}$ relative to the training samples $\mathcal{T} \subseteq \mathcal{X} \times \mathcal{Y}$ is given by

$$\mathbb{R}(t;\mathcal{T}) = -\frac{1}{|\mathcal{T}|} \sum_{(\boldsymbol{x},y)\in\mathcal{T}}^{|\mathcal{T}|} \log \mathbb{P}[y|\boldsymbol{x},t]\,.$$

Let $\mathcal{S} \subset \mathcal{T}$ be a subset of the training set that reached a node to be split. Then, by the decomposability of the log-loss, we can rewrite the empirical risk as

$$\mathbb{R}(t;\mathcal{T}) = \frac{|\mathcal{S}|}{|\mathcal{T}|}\mathbb{R}(t;\mathcal{S}) + \frac{|\mathcal{T}\setminus\mathcal{S}|}{|\mathcal{T}|}\mathbb{R}(t;\mathcal{T}\setminus\mathcal{S})\,.$$

Since the class posterior for the samples in $\mathcal{S}$ depends exclusively on the node where they ended up, the splitting process will impact only the empirical risk confined to $\mathcal{S}$ and not the one comprising the rest of the samples. Hence, a decision that is globally optimal in terms of log-loss can be taken at the node to be split by minimizing a local quantity, *i.e.* $\mathbb{R}(t;\mathcal{S})$. Assuming the routing function $f(\cdot;\Theta)$ to express the probability that a sample will be routed left, we can write $\mathbb{R}(t;\mathcal{S})$ in terms of our split function quality measure $Q(\Theta)$ (see, (4) of our CVPR contribution) as follows:

$$\mathbb{R}(t;\mathcal{S},\Theta) = -\frac{1}{|\mathcal{S}|}\log Q(\Theta)\,.$$

Finding a split function yielding the lowest empirical risk is thus equivalent to finding a maximizer of $Q(\Theta)$ with respect to Θ. Similarly, if we consider a ℓ_1-regularized empirical risk, then finding the best split function is equivalent to finding a maximizer of $Q_{\text{Reg}}(\Theta)$. In substance, the split function quality measure that we propose in our contribution is optimal in terms of risk minimization under log-loss.

2. Split Function Quality Measures and Information Gain

The split function quality measure $Q(\Theta)$ is optimal in terms of a log-loss-based risk minimization, as discussed in the previous section. We show in this section that, in the case of standard Random Forests (RF), maximizing $Q(\Theta)$ is equivalent to finding the split function in the random pool yielding the highest information gain. More generally, maximizing the information gain quality measure is substantially equivalent to maximizing a lower bound of $Q(\Theta)$, which is tight in the presence of binary routing functions, but *sub-optimal* in terms of log-loss when the routing functions are not binary. In the Neural Decision Forest (NDF) case the routing functions obtained from the randomized Multi-Layer Perceptron (rMLP) are not necessarily binary. For this reason, maximizing $Q(\Theta)$ (or the regularized counterpart) involves an optimization problem over the neural network's weights as well as the child posterior probabilities in $\boldsymbol{\pi}$, which in general does not follow an information gain maximization criterion, in the usual sense.

Let $f(\cdot;\Theta)$ be a routing function, $\mathtt{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{\mathsf{n}})$ be a matrix with n node samples and $\boldsymbol{y} = (y_1, \ldots, y_{\mathsf{n}})$ be a vector with the sample's class labels. Moreover, let $\hat{\boldsymbol{\pi}}$ denote the overall sample's empirical class distribution given by

$$\hat{\pi}_c = \frac{1}{\mathsf{n}} \sum_{s=1}^{\mathsf{n}} \mathbb{1}_{y_s = c} \,,$$

and let $\hat{\boldsymbol{\pi}}^{(\mathtt{L})}$ and $\hat{\boldsymbol{\pi}}^{(\mathtt{R})}$ be the empirical class distributions for the left and right child, respectively, induced by the routing function, *i.e*.

$$\hat{\pi}_c^{(d)} = \frac{\sum_{s=1}^{\mathsf{n}} \mathbb{1}_{y_s=c} f_d(\boldsymbol{x}_s;\Theta)}{\sum_{s=1}^{\mathsf{n}} f_d(\boldsymbol{x}_s;\Theta)} \,,$$

where $d \in \{\mathtt{L}, \mathtt{R}\}$, $f_{\mathtt{L}} = f$ and $f_{\mathtt{R}} = 1 - f$.

The *information gain* due to the split induced by the routing function is given by

$$\mathbb{IG}(\Theta) = \mathbb{H}(\hat{\boldsymbol{\pi}}) - \mathbb{H}(\hat{\boldsymbol{\pi}} | f(\cdot;\Theta)) \,,$$

where $\mathbb{H}$ is the entropy and

$$\mathbb{H}(\hat{\boldsymbol{\pi}} | f(\cdot;\Theta)) = \sum_{d \in \{\mathtt{L},\mathtt{R}\}} \frac{\sum_{s=1}^{\mathsf{n}} f_d(\boldsymbol{x}_s;\Theta)}{\mathsf{n}} \mathbb{H}(\hat{\boldsymbol{\pi}}^{(d)}) \,.$$

Theorem 1. *For any Θ the following relation holds:*

$$\log Q(\Theta) \geq \mathsf{n} \left[\mathbb{IG}(\Theta) - \mathbb{H}(\hat{\boldsymbol{\pi}}) \right] \,, \tag{1}$$

with equality in the presence of binary routing functions.

Proof. By application of the Jensen's inequality and simple algebraic manipulations, we have

$$\begin{aligned}
\log(Q(\Theta)) &= \max_{\boldsymbol{\pi}} \sum_{s=1}^{\mathsf{n}} \log \left(\sum_{d \in \{\mathtt{L},\mathtt{R}\}} \pi_{y_s}^{(d)} f_d(\boldsymbol{x}_s;\Theta) \right) \\
&\geq \max_{\boldsymbol{\pi}} \sum_{s=1}^{\mathsf{n}} \sum_{d \in \{\mathtt{L},\mathtt{R}\}} f_d(\boldsymbol{x}_s;\Theta) \log \left(\pi_{y_s}^{(d)} \right) && \{\text{by Jensen's inequality}\} \qquad (2) \\
&= \max_{\boldsymbol{\pi}} \sum_{d \in \{\mathtt{L},\mathtt{R}\}} \sum_{c \in \mathcal{Y}} \sum_{s=1}^{\mathsf{n}} \mathbb{1}_{y_s=c} f_d(\boldsymbol{x}_s;\Theta) \log \left(\pi_c^{(d)} \right) \\
&= \max_{\boldsymbol{\pi}} \sum_{d \in \{\mathtt{L},\mathtt{R}\}} \sum_{s=1}^{\mathsf{n}} f_d(\boldsymbol{x}_s;\Theta) \sum_{c \in \mathcal{Y}} \hat{\pi}_c^{(d)} \left[\log \frac{\pi_c^{(d)}}{\hat{\pi}_c^{(d)}} + \log \hat{\pi}_c^{(d)} \right] \\
&= \max_{\boldsymbol{\pi}} - \sum_{d \in \{\mathtt{L},\mathtt{R}\}} \sum_{s=1}^{\mathsf{n}} f_d(\boldsymbol{x}_s;\Theta) \left[\mathbb{H}\left(\hat{\boldsymbol{\pi}}^{(d)} \right) + \mathbb{D}_{\mathrm{KL}} \left(\hat{\boldsymbol{\pi}}^{(d)}, \boldsymbol{\pi}^{(d)} \right) \right] \\
&= - \sum_{d \in \{\mathtt{L},\mathtt{R}\}} \sum_{s=1}^{\mathsf{n}} f_d(\boldsymbol{x}_s;\Theta) \left[\mathbb{H}\left(\hat{\boldsymbol{\pi}}^{(d)} \right) + \underbrace{\min_{\boldsymbol{\pi}^{(d)}} \mathbb{D}_{\mathrm{KL}} \left(\hat{\boldsymbol{\pi}}^{(d)}, \boldsymbol{\pi}^{(d)} \right)}_{=0} \right]
\end{aligned}$$

$$= - \sum_{d \in \{\mathtt{L},\mathtt{R}\}} \sum_{s=1}^{\mathsf{n}} f_d(\boldsymbol{x}_s;\Theta) \mathbb{H}\left(\hat{\boldsymbol{\pi}}^{(d)}\right)$$
$$= \mathsf{n}\left[\mathbb{IG}(\Theta) - \mathbb{H}(\hat{\boldsymbol{\pi}})\right] \tag{3}$$

Therefrom, inequality (1) derives. Moreover, if the routing function f is binary, inequality (2) yields equality and, hence, so does (1). □

Corollary 1. *In the presence of binary routing functions, the following holds:*

$$\arg\max_{\Theta} Q(\Theta) = \arg\max_{\Theta} \mathbb{IG}(\Theta)\,.$$

Proof. The relation follows directly from Theorem 1:

$$\arg\max_{\Theta} Q(\Theta) = \arg\max_{\Theta} \log Q(\Theta) = \arg\max_{\Theta} \mathsf{n}[\mathbb{IG}(\Theta) - \mathbb{H}(\hat{\boldsymbol{\pi}})] = \arg\max_{\Theta} \mathbb{IG}(\Theta)\,.$$

□

3. Update Rule for the Child Posteriors $\boldsymbol{\pi}$

In this section we prove that the multiplicative update rule for $\boldsymbol{\pi}$ that we have given in our CVPR contribution monotonically increases the likelihood. This result justifies the use of the rule for the computation of $Q(\Theta)$ and $Q_{\text{Reg}}(\Theta)$, which involve indeed the maximization of the likelihood with respect to the child posteriors.

Let $\boldsymbol{\pi}(t) = (\boldsymbol{\pi}^{(\mathtt{L})}(t), \boldsymbol{\pi}^{(\mathtt{R})}(t))$ denote a pair of class-posteriors of the left and right child at time $t \geq 0$ and let $L(\boldsymbol{\pi})$ denote the log-likelihood function restricted to $\boldsymbol{\pi}$:

$$L(\boldsymbol{\pi}) = \log \mathbb{P}[\boldsymbol{y}|\mathtt{X}, \boldsymbol{\pi}, \Theta] = \sum_{s=1}^{\mathsf{n}} \log \left(\sum_{d \in \{\mathtt{L},\mathtt{R}\}} \pi_{y_s}^{(d)} f_d(\boldsymbol{x}_s|\Theta) \right)\,,$$

and consider the following multiplicative update rule for the child posteriors, used in our CVPR contribution:

$$\pi_c^{(d)} \leftarrow \frac{1}{Z^{(d)}} \sum_{s=1}^{\mathsf{n}} \frac{\mathbb{1}_{y_s=c}\, \pi_c^{(d)} f_d(\boldsymbol{x}_s;\Theta)}{\pi_c^{(\mathtt{L})} f_{\mathtt{L}}(\boldsymbol{x}_s;\Theta) + \pi_c^{(\mathtt{R})} f_{\mathtt{R}}(\boldsymbol{x}_s;\Theta)}\,. \tag{4}$$

Theorem 2. *Let $\{\boldsymbol{\pi}(t)\}_{t\geq 0}$ be a trajectory of (4). Then for all $t \geq 0$,*

$$L(\boldsymbol{\pi}(t+1)) > L(\boldsymbol{\pi}(t))$$

with equality if and only if $\boldsymbol{\pi}(t+1) = \boldsymbol{\pi}(t)$.

Proof. Consider the following auxiliary function:

$$\phi(\boldsymbol{\pi}, \overline{\boldsymbol{\pi}}) = L(\overline{\boldsymbol{\pi}}) + \sum_{s=1}^{\mathsf{n}} \sum_{d \in \{\mathtt{L},\mathtt{R}\}} \xi_s^{(d)}(\overline{\boldsymbol{\pi}}) \log \left(\frac{\pi_{y_s}^{(d)}}{\overline{\pi}_{y_s}^{(d)}} \right)\,, \tag{5}$$

where

$$\xi_s^{(d)}(\boldsymbol{\pi}) = \frac{\pi_{y_s}^{(d)} f_d(\boldsymbol{x}_s|\Theta)}{\pi_{y_s}^{(\mathtt{L})} f_{\mathtt{L}}(\boldsymbol{x}_s|\Theta) + \pi_{y_s}^{(\mathtt{R})} f_{\mathtt{R}}(\boldsymbol{x}_s|\Theta)}\,.$$

Trivially, $\phi(\boldsymbol{\pi}, \boldsymbol{\pi}) = L(\boldsymbol{\pi})$ holds for all $\boldsymbol{\pi}$ as the logarithms in (5) yield zero for all values of s, d and c. We show now that $\phi(\boldsymbol{\pi}, \overline{\boldsymbol{\pi}}) \leq L(\boldsymbol{\pi})$ holds for all $\boldsymbol{\pi}$. To this end, we use Jensen's inequality to derive:

$$L(\boldsymbol{\pi}) = \sum_{s=1}^{\mathsf{n}} \log \left(\sum_{d \in \{\mathtt{L},\mathtt{R}\}} \pi_{y_s}^{(d)} f_d(\boldsymbol{x}_s|\Theta) \right) \geq \sum_{s=1}^{\mathsf{n}} \sum_{d \in \{\mathtt{L},\mathtt{R}\}} \xi_s^{(d)}(\overline{\boldsymbol{\pi}}) \log \left(\frac{\pi_{y_s}^{(d)} f_d(\boldsymbol{x}_s|\Theta)}{\xi_s^{(d)}(\overline{\boldsymbol{\pi}})} \right) = \phi(\boldsymbol{\pi}, \overline{\boldsymbol{\pi}})\,. \tag{6}$$

The last equality follows by simple algebraic manipulations.

Next, we show that $\boldsymbol{\pi}(t+1)$, which is obtained by applying the update rule (4) to $\boldsymbol{\pi}(t)$, is a global maximizer of $\phi(\cdot, \boldsymbol{\pi}(t))$. The distribution $\boldsymbol{\pi}^{(d)}(t+1)$ can be written in terms of $\xi_s^{(d)}$ as

$$\pi_c^{(d)}(t+1) = \frac{1}{Z^{(d)}} \sum_{s=1}^{\mathsf{n}} \mathbb{1}_{y_s=c}\, \xi_s^{(d)}(\boldsymbol{\pi}(t))\,,$$

where $Z^{(d)}$ is the normalizing factor. By exploiting this equality we have that

$$\begin{aligned}
&\phi(\boldsymbol{\pi}(t+1), \boldsymbol{\pi}(t)) - \phi(\boldsymbol{\pi}, \boldsymbol{\pi}(t)) \\
&= \sum_{s=1}^{\mathsf{n}} \sum_{d\in\{\mathtt{L},\mathtt{R}\}} \xi_s^{(d)}(\boldsymbol{\pi}(t)) \log\left(\frac{\pi_{y_s}^{(d)}(t+1)}{\pi_{y_s}^{(d)}}\right) \\
&= \sum_{d\in\{\mathtt{L},\mathtt{R}\}} \sum_{c\in\mathcal{Y}} \sum_{s=1}^{\mathsf{n}} \mathbb{1}_{y_s=c}\, \xi_s^{(d)}(\boldsymbol{\pi}(t)) \log\left(\frac{\pi_c^{(d)}(t+1)}{\pi_c^{(d)}}\right) \\
&= \sum_{d\in\{\mathtt{L},\mathtt{R}\}} Z^{(d)} \sum_{c\in\mathcal{Y}} \pi_c^{(d)}(t+1) \log\left(\frac{\pi_c^{(d)}(t+1)}{\pi_c^{(d)}}\right) = \sum_{d\in\{\mathtt{L},\mathtt{R}\}} Z^{(d)} \mathbb{D}_{\mathrm{KL}}\left(\boldsymbol{\pi}^{(d)}(t+1) || \boldsymbol{\pi}^{(d)}\right) \geq 0\,.
\end{aligned}$$

holds for all possible $\boldsymbol{\pi}$. This inequality yields equality if and only if $\boldsymbol{\pi} = \boldsymbol{\pi}(t+1)$ as it is well-known that the Kullback-Leibler divergence yields zero if and only if the arguments coincide. This proves that $\boldsymbol{\pi}(t+1)$ is a *strict* global maximizer of $\phi(\cdot, \boldsymbol{\pi}(t))$.

By using this last result and (6), we derive this chain of inequalities:

$$L(\boldsymbol{\pi}(t)) = \phi(\boldsymbol{\pi}(t), \boldsymbol{\pi}(t)) < \phi(\boldsymbol{\pi}(t+1), \boldsymbol{\pi}(t)) \leq L(\boldsymbol{\pi}(t+1))\,,$$

where equality holds if and only if $\boldsymbol{\pi}(t) = \boldsymbol{\pi}(t+1)$, *i.e.* in the presence of a fixed-point of the update rule (4). □

4. Example result images for LFW dataset

In this section we show qualitative results obtained by standard classification forests and our proposed neural decision forests, corresponding to the quantitative evaluations provided in the main document in Table (2). In Figure 1 we show the input RGB-image with ground truth labelling and the respective results of our baseline random forest and the proposed $\text{NDF}_{\mathtt{MLPC}-\ell 1}$. Remember that the input to our classifiers is restricted to RGB raw intensities so with conventional decision forests it is very likely to confuse predictions due to similarities in the colours. This is *e.g.* shown in rows 2-4 of Figure 1 where many background pixels get confused with the facial skin category. Even though our algorithm is not explicitly invariant to such confusions, the images show that we can handle potential confusions much better by learning ad-hoc image representations with the proposed, randomized Multi-Layer Perceptrons installed as split nodes. Consequently, the images show significant reduction in the number of confusions. We also see that the individual categories are segmented in a more coherent way (especially the mouth and eye categories).

In Figure 2 we show the development for the Global and Jaccard scores with respect to the number of trees in the forests. Again, we compare the conventional classification forest with our best-performing approach $\text{NDF}_{\mathtt{MLPC}-\ell 1}$, demonstrating significant improvements also on a per-tree basis. For instance, a single $\text{NDF}_{\mathtt{MLPC}-\ell 1}$-tree has an improvement of 4.7/9% (Global/Jaccard) compared to a single standard classification tree. We can maintain this improvement almost over the entire experiment. Additionally, using only 3 $\text{NDF}_{\mathtt{MLPC}-\ell 1}$trees already outperform an 8-tree ensemble of conventional classification trees by 2/4% for Global/Jaccard score, respectively. These findings support our claims in the main paper that the compression of trees leads to better generalization and therefore higher test accuracy.

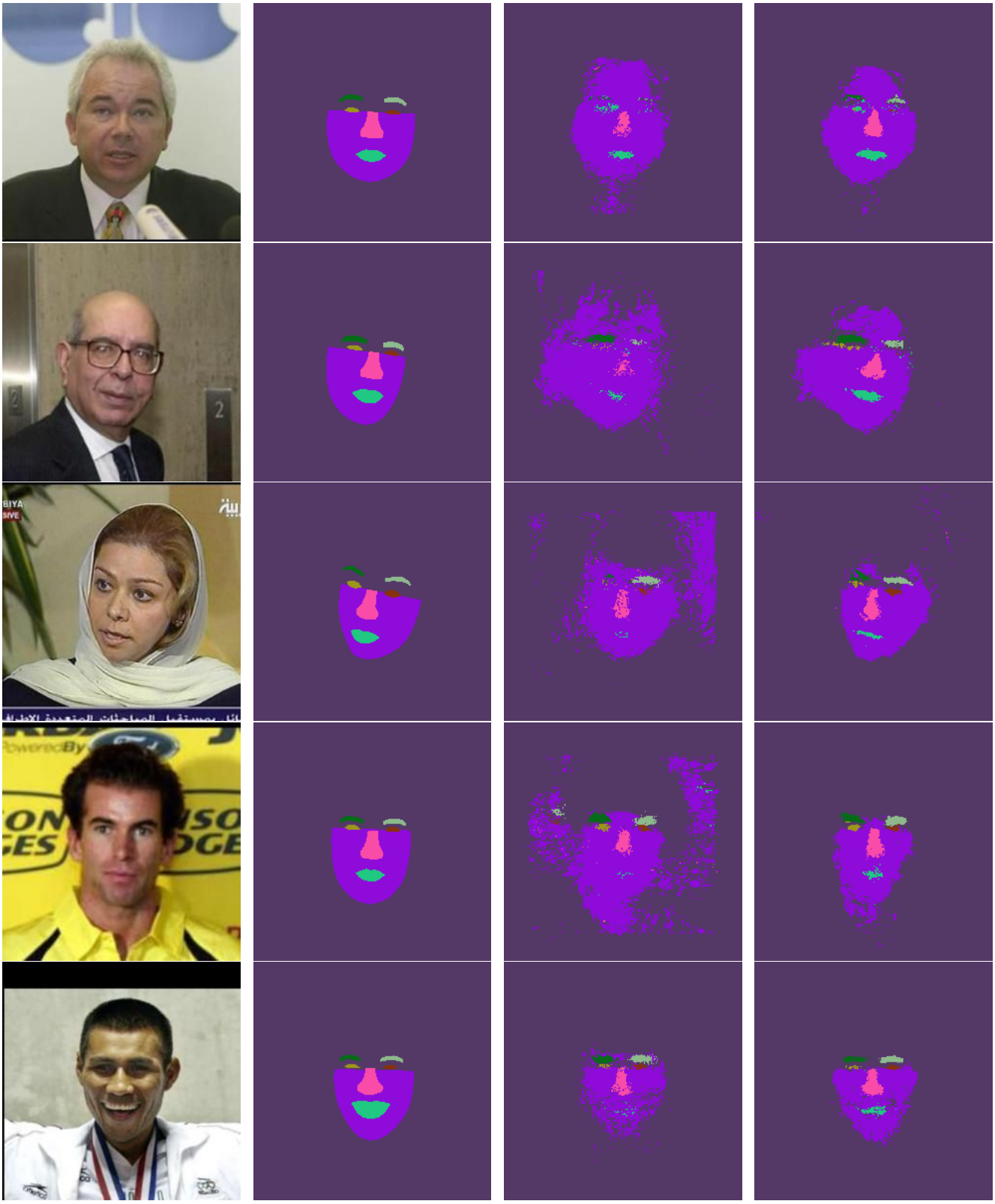

Figure 1. Qualitative results on LFW dataset. From left to right: Input RGB image, ground truth labels, result of baseline random forest, result of our proposed $\text{NDF}_{\text{MLPC}-\ell 1}$. See text for description.

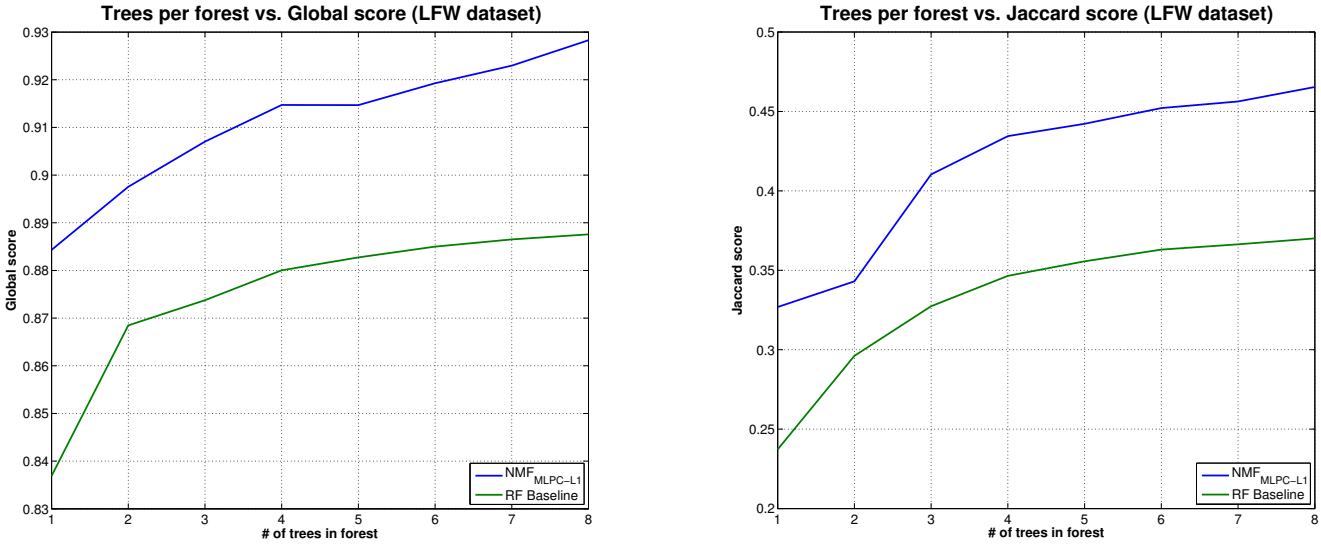

Figure 2. Development of Global (left) and Jaccard (right) scores as a function of trees in the forest ensemble for conventional random forests (green) and our proposed Neural decision forests (blue). See text for description.