RULe: Relocalization-Uniformization-Landmark Estimation Network for Real-Time Face Alignment in Degraded Conditions

Arnaud Dapogny, Gauthier Tallec, Jules Bonnard, Edouard Yvinec, Kevin Bailly

RULe: Relocalization-Uniformization-Landmark Estimation Network for Real-Time Face Alignment in Degraded Conditions

Arnaud Dapogny[1], Gauthier Tallec[2], Jules Bonnard[2], Edouard Yvinec[1,2] and Kévin Bailly[1,2]
[1] Datakalab, 114 Boulevard Malesherbes, 75017 Paris, France
[2] Sorbonne Université, 4 Place Jussieu, 75005 Paris, France

***Abstract*— Face alignment refers to the process of estimating the position of a number of salient landmarks on face images or videos, such as mouth and eye corners, nose tip, etc. With the availability of large annotated databases and the rise of deep learning-based methods, face alignment as a domain has matured to a point where it can be applied in more or less unconstrained conditions, e.g. non-frontal head poses, presence of heavy make-up or partial occlusions. However, when considering real-case alignment on videos with possibly low frame rates, we need to make sure that the algorithms are robust to jittering of the face bounding box localization, low-resolution of the face crops, possible bad environmental lighting, brightness, and presence of noise.**

To tackle these issues, we propose RULe, a three-staged Relocalization-Uniformization-Landmark Estimation network. In the first stage, an initial loosely localized bounding box gets refined to output a well centered face crop, thus reducing the variability of the images prior to passing them to the subsequent stage. Then, in the second stage, the face style is uniformized (using adversarial learning as well as perceptual losses) to correct low resolution or variations of brightness/contrast. Finally, the third stage outputs a precise landmark estimation given such enhanced face crop using a cascaded compact model trained using hint-based knowledge distillation. We show through a variety of experiments that RULe achieves real-time face alignment with state-of-the-art precision in heavily degraded conditions.

I. Introduction

Face alignment denotes the process of localizing a number of landmarks on a face image, such as lips or eye corners, pupils or nose tips. It is an essential preprocessing for many computer vision applications, such as facial expression recognition [6], face synthesis [15], or facial performance reenactment [25]. However, most state-of-the-art approaches have focused on estimating face landmarks in nominal conditions, *i.e.* starting from a perfect bounding box and with moderate to high quality images. This does not always translate well to real-world scenarios, e.g. when tracking faces on videos and the bounding box has to be re-estimated from one frame to another and in the case where a number of frames may be skipped. Such temporal video downsampling can cause imprecision in bounding box localization prior to landmark estimation. Similarly, spatial downsampling as well as poor lighting conditions can lead to severe out-of-distribution problems, leading to major inconsistencies in landmark estimation.

This work has been supported by Datakalab as well as the French National Agency (ANR) in the frame of its Technological Research JCJC program (FacIL, project ANR-17-CE33-0002).

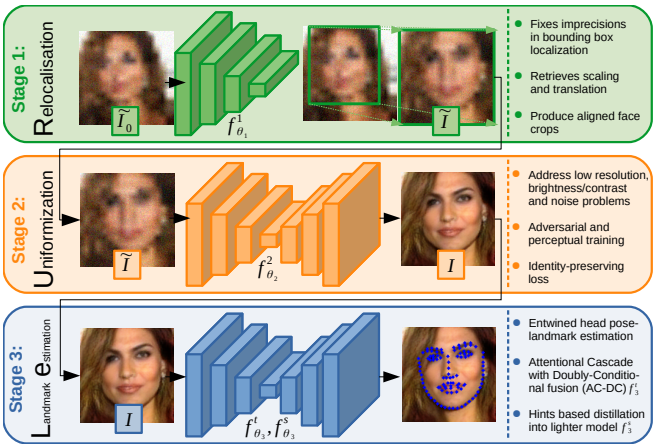

Fig. 1. Overview of our **RULe** method and summary of its main features. First, given a roughly estimated crop of a face image, a first pass through a **R**elocalization network refines the bounding box, providing aligned crops for the subsequent stage. Second, the face crop pass to a style **U**niformization network that fixes low-resolution, bad environmental lighting as well as noise problems. Third, a **L**andmark **e**stimation network localizes face landmarks on this enhanced face crop.

In this paper, we propose a three-staged RULe network for face alignment in degraded conditions. The generic flowchart of RULe is illustrated on Figure 1. First, the bounding box is corrected using a Relocalization **(R)** network: this first step produces well registered face crops, a *sine qua non* condition for reducing variability for the subsequent step. Second, a style Uniformization **(U)** network (trained using adversarial learning as well as perceptual losses) is applied to the corrected crop to enhance its quality, and address low resolution as well as poor lighting conditions. Third, a Landmark Estimation **(Le)** network is applied. The latter consists of a cascaded compact model trained using hint-based knowledge distillation, that allows precise landmark analysis in degraded conditions with reasonable speed and can be integrated into the proposed three-staged approach. To sum it up, the contributions of this paper are the following:

- We propose a three-staged Relocalization-Uniformization-Landmark estimation (RULe) method to address poor bounding box initialization as well as degraded image quality (low-resolution, as well as poor environmental lighting conditions and noise) for face landmark estimation.
- We provide practical solutions using hint-based knowledge distillation to design compact cascaded models for

real-time, yet precise face landmark estimation methods to be integrated into the proposed three-staged approach.
- We demonstrate the effectiveness of our three-staged RULe framework for face alignment in nominal and degraded conditions for still images and video settings.

II. Related work

Landmark estimation can BE formulated as either 2D or 3D alignment. The former consists in predicting 2D landmark localization, usually for face images with low to medium yaw angles. On the one hand, popular methods for 2D face alignment either belong to cascaded regression or deep learning-based, end-to-end approaches. Popular exemples of cascaded regression approaches include SDM [29], LBF [20], CSP-GNdF [5] and more recently DAN [17] and tree-gated MoE [2]. A natural pitfall of such approaches is that the regressors are not learned jointly in a end-to-end fashion, thus there is no guarantee that the whole cascade might be optimal. Tackling this issue, MDM [26] improves the feature extraction process by sharing CNN layers among cascade stages, which are formulated as a recurrent neural network. This results in a more optimized landmark trajectory throughout the cascade. On the other hand, exemples of the deep end-end-end trainable methods include TCDCN [33], which involves pretraining on a wide facial attributes database [19]. More recently, SAN [10] uses generative adversarial networks to convert images from different styles to an aggregated style before performing landmark localization. Authors of [28] propose to use edge map estimation as an intermediate representation to drive the landmark prediction task. Authors in [11] use a surrogate loss to enhance training of deep networks. AAN [30] proposes to use intermediate feature maps as attentional masks to select relevant spatial regions. In previous work, [7], [8] we proposed a deep convolutional cascade that lies in-between the deep and cascaded approaches, allowing to iteratively learning a coarse-to-fine landmark estimation, while benefitting from end-to-end training. Huang *et al.* [14] propose to model the per-landmark error statistics to better handle the annotation variability. Lastly, authors in [18] propose to measure and address the quantization error caused by the downsampling process inherent to deep neural network processing.

Addressing degraded conditions: most of the above-mentionned approaches focus on designing efficient face landmark estimation in nominal conditions, that is to say assuming that a perfect bounding box and face crop are provided as the input of the proposed system. However, when applying face landmark estimation in a real-case scenario, this bounding box is usually provided by either an off-the-shelf face detector or, when performing landmark estimation on video, as the result of the processing of previous frames in the sequence. As a consequence, in such scenario, the accuracy of landmark estimation might be impacted heavily by degradation of these initial conditions. Most approaches [29], [20], [3] addresses this problem by augmenting the training with perturbation of the original bounding box within the previously estimated standard deviation of a face detector. Furthermore, fully-convolutional alignment methods [30] have a built-in invariance to small translation of the input, hence to small aspect-equivariant variations of the original bounding box. This may not, however, be sufficient to ensure precise landmark estimation when the bounding box is too far from the target face. In addition, traditional face alignment methods are not robust to heavy downsampling of the images: for instance, the performance of FAN [3] starts to drop down significantly when images are downsampled by 2. More recently, authors of [4] addresses the low-resolution problem by learning a face super-resolution model using generative adversarial network (GAN) prior to performing landmark estimation on top of the generated image. In contrast, in this paper, we address the imprecision in the bounding box localization (translation and changes in aspect ratio), in addition to tackling the issue of low-quality images as a whole: this includes low-resolution images, poor environmental lighting, brightness or contrast conditions, as well as the presence of noise.

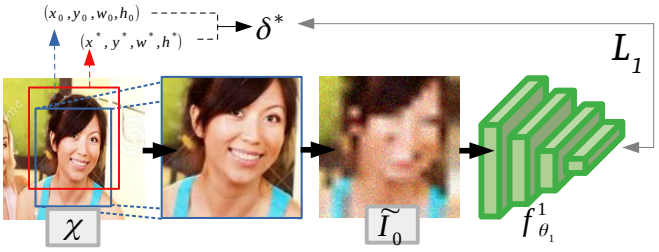

Fig. 2. Flowchart of the relocalization network for the first **(R)** stage. At train time, a crop $\tilde{I}_0$ of an image χ is produced, by applying bounding box augmentation as well as downsampling, random Hue/Saturation/Brightness as well as additive noise. The relocalization network $f^1_{\theta_1}$ aims at retrieving the translation-scale bounding box transformation.

III. RULe: a three-stages landmark localization model

Our RULe model is summarized on Figure 1, and is composed of three stages: in the first stage, starting from an initial guess, the face bounding box is corrected (Section III-A) using a Relocalization (R) network. Then, the face crop pass through a Uniformization (U - Section III-B) network that standardizes its quality. Finally, Landmark Estimation (LE) is performed *via* a compact cascaded model trained using hint-based knowledge distillation (see Section III-C).

A. Bounding box relocalization

Figure 2 Illustrates the relocalization **(R)** stage. In this first stage, we consider a face image $\mathcal{X}$ and assume that we have an initial bounding box guess (x_0, y_0, w_0, h_0), where x_0 and y_0 denote the (x, y) coordinates of the bounding box center, and w_0 and h_0 its width and height, respectively. Typically, when performing landmark estimation on video, we periodically apply an off-the-shelf face detection algorithm [27] to locate the face region of interest before trying to locate the face landmarks. In other cases, the aligned landmarks for the previous frames provide a raw initial guess for aligning landmarks on the current frame. In either case,

directly performing landmark alignment by cropping the face image $\mathcal{X}$ using such initial guess generally leads to imprecise alignment. To address this problem, we propose a bounding box correction scheme using a first relocalization stage.

1) Bounding box correction with relocalization network: Let's consider $\tilde{I}_0$ an (imprecise) $n \times n$ face crop obtained from the initial bounding box, *i.e.* $\tilde{I}_0$ is obtained by reshaping $\mathcal{X}[y_0 - h_0/2 : y_0 + h_0/2, x_0 - w_0/2 : x_0 + w_0/2]$ to $n \times n$, and $f^1_{\theta_1} : \tilde{I}_0 \rightarrow \{\delta_x, \delta_y, \delta_w, \delta_h\}$ denote a function with parameters θ_1 that takes as input this initial face crop guess and provides a transformation vector that "corrects" the corresponding bounding box in terms of translation and scaling. Formally, if we have:

$$\begin{cases} x_1 = x_0 + \delta_x \\ y_1 = y_0 + \delta_y \\ w_1 = w_0 \times \delta_w \\ h_1 = h_0 \times \delta_h \end{cases} \tag{1}$$

then the face crop $\mathcal{X}[y_1 - h_1/2 : y_1 + h_1/2, x_1 - w_1/2 : x_1 + w_1/2]$ obtained from this new region of interest is a more suitable face crop $\tilde{I}_1$ for the subsequent stages. In other words, the function $f^1_{\theta_1}$ maps an imprecise face crop to the parameters of a rigid transformation (in translation δ_x-δ_y and scale δ_w-δ_h) that refines it. We use a deep neural network to model $f^1_{\theta_1}$. Also, in what follows, we explain how we can optimize its parameters θ_1 using existing datasets.

2) Learning to correct bounding boxes from still images: The function $f^1_{\theta_1}$ can be trained on any landmark-annotated still image dataset. For a face image $\mathcal{X}$ we derive a ground truth bounding box (x^*, y^*, w^*, h^*) from the ground truth landmark annotation by considering the min-max localization of the landmarks and adding margins in both width and height (e.g. 25% of the min-max bounding box dimensions). From this ground truth bounding box we generate random augmentations in translation and scale $\delta^* = (\delta^*_x, \delta^*_y, \delta^*_w, \delta^*_h)$ which leads to a new initial guess $\tilde{I}_0$. This, in turn, gives rise to the following bounding box relocalization loss function:

$$\mathcal{L}_1(\theta_1) = |f^1_{\theta_1}(\tilde{I}_0) - \delta^*| \tag{2}$$

The first stage relocalization network is trained to optimize the objective function $\mathcal{L}_1$. Note at this point that, in order to mimic real-world conditions where the face crops can be arbitrarily small, or may present large variations of illumination, we augment the images with random resolution as well as brightness/contrast for training the relocalization network for it to be able to correct the face bounding box in degraded conditions. In what follows, we show how we can explicitly uniformize the face crop style in a second stage, prior to landmark estimation.

B. Style uniformization

The Uniformization (**U**) stage takes as input a correctly localized face crop and corrects style variations such as low resolution as well as brightness/contrast. Figure 3 provides an overview of the training of the uniformization network

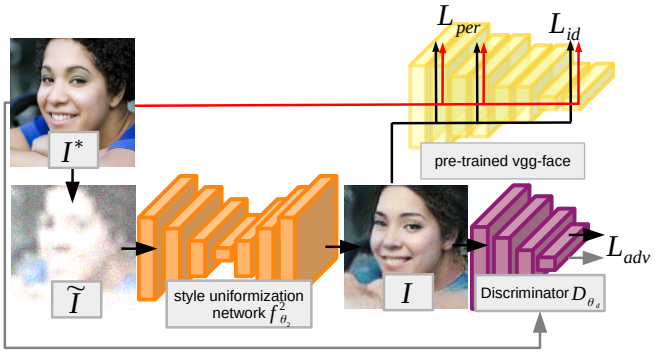

Fig. 3. Flowchart of the style uniformization network for the second (**U**) stage. At train time, a corrupted version $\tilde{I}$ of an image I^* is enhanced. Adversarial training as well as perceptual and identity-preserving losses help produce quality reconstructions.

designed to tackle this problem. At train time, it uses adversarial learning and perceptual/identity-preserving loss.

1) Adversarial training: In this section, we denote $\tilde{I}$ a corrupted version of this face crop I^*, obtained by applying random downsampling and variations in brightness and contrast. We denote $f^2_{\theta_2} : \tilde{I} \rightarrow I$ the style uniformization network that aims at restoring I^* knowing $\tilde{I}$ only. As it is classical in the image edition litterature [4], we use generative adversarial networks (GAN) for that purpose. We thus define a discriminator network D_{θ_d} with parameters θ_d, aiming at distinguishing the real images I^* from the generated ones I, by minimizing the following loss:

$$\mathcal{L}_{Disc}(\theta_d) = -\log D_{\theta_d}(I^*) - \log[1 - D_{\theta_d}(f^2_{\theta_2}(\tilde{I}))]. \tag{3}$$

While the generator minimizes the following:

$$\mathcal{L}_{Gen}(\theta_2) = -\log D_{\theta_d}(f^2_{\theta_2}(\tilde{I})) \tag{4}$$

The generator and discriminator networks thus play a min-max game where the later tries to correctly distinguish the real images from the fake, generated ones, whereas the generator has to fool the discriminator, thus generating perceptually real images. By doing so, the generator captures the underlying distribution of the real, high quality images. Note at this point that while GANs encountered great successes in image distribution when the target distribution is relatively easy to model (e.g. on CelebA database [19]), they may struggle in more difficult cases, e.g. on unaligned images or datasets with broader semantic content such as ImageNet or Pascal VOC 2012 datasets, as also reported in many papers on image generation [1], [12], [9]. Thus, the relocalization and uniformization stages complement each other well to a certain extent, as the relocalization network ensures that the face crops that pass through the second (uniformization) stage are correctly centered. Last but not least, in our case, we seek to not only produce realistic images but to reconstruct the original image as accurately as possible, hence we need to add a reconstruction loss such as perceptual loss [16].

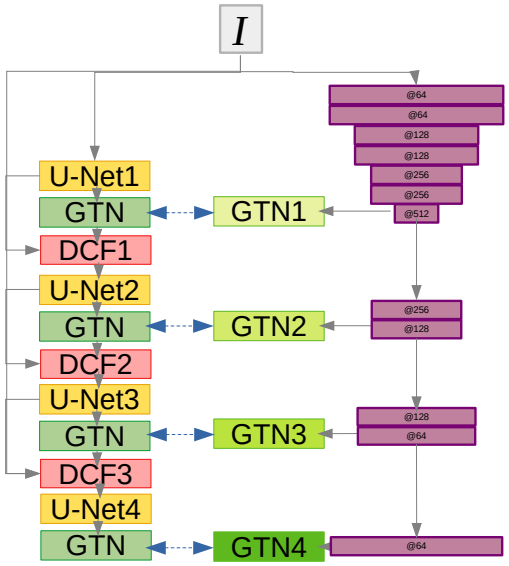

Fig. 4. Illustration of teacher (left)-student (right) learning. In particular, the GTN output at each stage is used to provide hints for the student model at each upsampling step, *via* a specific GTN.

2) Perceptual and Identity-preserving loss: Perceptual loss [16] is an upgrade to the standard (e.g. $\mathcal{L}_1$ or $\mathcal{L}_2$) reconstruction losses. It consists in matching the generated and target images in the embedding space of the L layers $\{\psi_l\}_{l=1,...,L}$ of a pre-trained (e.g. VGG-16) network. This loss thus can be written as:

$$\mathcal{L}_{per}(\theta_2) = \sum_{l=1}^{L} \lambda_l^{per} ||\psi_l(f_{\theta_2}^2(\tilde{I})) - \psi_l(I^*)|| \quad (5)$$

Where λ_l^{per} weights the importance of the l-th layer in the loss term. Traditionally, the consensus is to put more weights on the first layers (e.g. using $\lambda_1^{per} = 1$, $\lambda_2^{per} = 0.5$, and so on) to match the low-level features (analog as gradients) of the generated and target image. In this work, we use a different setting: First, instead of a traditional VGG-16 network, we use VGG-face, i.e. the same architecture trained for face recognition as in [23]. The rationale behind this is that the basis of learned features are more adapted to detect subtle discriminative face patterns. Furthermore, we introduce a identity-preserving term by considering the last layer $x \rightarrow \psi_L(x)$ of VGG-face network:

$$\mathcal{L}_{id}(\theta_2) = ||\psi_L(f_{\theta_2}^2(\tilde{I})) - \psi_L(I^*)|| \quad (6)$$

The total loss optimized to train the second **(U)** stage is:

$$\mathcal{L}_2(\theta_2, \theta_d) = \mathcal{L}_{Gen}(\theta_2) + \mathcal{L}_{Disc}(\theta_d) + \lambda^{per}\mathcal{L}_{per}(\theta_2) + \lambda^{id}\mathcal{L}_{id}(\theta_2) \quad (7)$$

With λ^{per} and λ^{id} controlling the effect of perceptual and identity preserving loss, respectively. The values of these hyperparameters are specified in Section IV-B.

C. Landmark localization with a distilled cascaded network

Similarly to [21] we apply knowledge distillation [13] with a lighter student architecture to enhance the runtime capabilities of an existing architecture [8]. This architecture [8] is composed of 4 cascade stages that use geometry transfer networks (GTN) composed of 1×1 convolutions that convert landmark geometry format to integrate together heterogeneous landmark and head pose annotation markups. Let's now suppose that we have such a model ouputting landmark estimates $\{s_i^{L_k}\}_{i=1,...,4,k=1,...,K}$ and head pose estimates $\{\Omega_i\}_{i=1,...,4}$ for all cascade stages indexed by i. If we also consider a fully-convolutional encoder-decoder student model f_3^{stu} (with parameters θ_3^{stu}), we can write it as $f_3^{stu}(I) = d_4 \circ d_3 \circ d_2 \circ d_1 \circ e(I)$, where $e, d_1, ..., d_4$ indicate the successive encoding and decoding (upsampling) layers. As illustrated on Figure 4, we plug a GTN after each decoding layers d_i, that outputs landmark estimates $\{\tilde{s}_i^{L_k}\}_{i=1,...,4,k=1,...,K}$ and head pose estimates $\{\tilde{\Omega}_i\}_{i=1,...,4}$ at this upsampling layer. We can thus match the outputs of the teacher and student models, both in terms of landmark-wise attention maps, and head pose estimates. Formally, the loss of the student model can be written as:

$$\mathcal{L}_s(\theta_3^{stu}) = \sum_{i=1}^{4} \lambda_i \sum_{k=1}^{K} (1-\lambda_{kd})|\tilde{s}_i^{L_k} - s^{L_k*}| + \lambda_{kd}|\tilde{s}_i^{L_k} - s_i^{L_k}| \quad (8)$$

for the landmark estimation objective, and

$$\mathcal{L}_\Omega(\theta_3^{stu}) = \sum_{i=1}^{4} \lambda_i \sum_{k=1}^{K} (1-\lambda_{kd})|\tilde{\Omega}_i - \Omega^*| + \lambda_{kd}|\tilde{\Omega}_i - \Omega_i| \quad (9)$$

for the head pose estimation term. Again, the total loss is:

$$\mathcal{L}_3^{stu}(\theta_3) = \mathcal{L}_s(\theta_3^{stu}) + \mathcal{L}_\Omega(\theta_3^{stu}) \quad (10)$$

Where the first term is a supervised term and the second corresponds to the distillation term, weighted by hyperparameter λ_{kd} and λ_i for each stage. Note that, in such a case, θ_3^{stu}, the collection of parameters for the student model, encompasses both the parameters of the encoder-decoder architecture $e, d_1, ..., d_4$ as well as the GTN. Also, because each GTN has to deal with different input sizes, we apply different networks with no parameter sharing for the student model. However, as the GTN are composed of only 1×1 convolutions and soft-argmax blocks, they encompasses very few parameters. Hence, most of the gradient flow through the parameters of the student network, which facilitates training. With this in mind, we can train high-precision models for entwined Landmark estimation (**Le**) and use them to train lighter student models to enable real-time processing on lighter devices such as CPU. This, in turn, allows real-time processing when integrated within the three-staged model.

IV. Experiments

First, we provide a brief summary of the datasets involved as well as a summary for the whole three-staged **R-U-Le**

architecture and hyperparameters. Then, we evaluate the distilled models on clean data. Finally, we validate **R-U-Le** in degraded conditions both for still images and videos.

A. Databases

The **300W** database, introduced in [22], contains moderate variations in pose and expressions. It also embraces a few occluded images. It consists in four databases: **LFPW** (811 images for train / 224 images for test), **HELEN** (2000 images for train / 330 images for test), **AFW** (337 images for train) and **IBUG** (135 images for test), for a total of 3148 images annotated with 68 landmarks for training the models. As state-of-the-art approaches already outputs very high accuracy on this dataset, authors of [34] introduced the **300W-LP** database, which is a large-pose dataset synthethized from 300W. It contains 100842 train images and 21608 images following the same partitions as in 300W, but with yaw angles covering the $[-90, 90]$ degrees range. Authors of [34] also proposed AFLW2000-3D database, which contains example synthethized from the 2000 first images of AFLW database using the same protocol as in 300W-LP.

The **CelebA** database [19] is a large-scale face attribute database which contains 202599 218×178 celebrity images coming from 10177 identities, each annotated with 40 binary attributes (such as *gender*, *eyeglasses*, *smile*), and the localization of 5 landmarks (nose, left and right pupils, mouth corners). In our experiments, we use the train partition that contains 162770 images from $8k$ identities to train our models. The test partition contains 19962 instances from $1k$ identities that are different from the training set identities.

The **Wider Facial Landmarks in the Wild or WFLW** database [28] contains 10000 faces (7500 for training and 2500 for testing) with 98 annotated landmarks. This database also features rich attribute annotations in terms of occlusion, head pose, make-up, illumination, blur and expressions.

The **300VW** database [24] is a video alignment database that contains 114 videos making a total of 218,595 frames, which are divided into three subsets of various difficulty (categories A, B and C, C being the most challenging).

B. Implementation details

The relocalisation network, style uniformization network, and landmark estimation network are trained separately using Tensorflow, by using ADAM optimizer with a $5e^{-4}$ learning rate with $\beta_1 = 0.9$ and learning rate annealing with power 0.9 for the three stages. We apply $200k$ updates with batch size 32 for the first two stages. For the third stage, we apply $400k$ updates with batch size 8 for each database, with alternating updates between the databases.

The relocalization network (R-1^{st} stage) is trained by optimizing Equation (2) on WFLW database. It takes 128×128 grayscale face crops and consists in subsequents applications of 3×3 conv with stride 2, batch norm, ReLU with $64 \rightarrow 64 \rightarrow 128 \rightarrow 256 \rightarrow 512 \rightarrow 1024$ channels followed by 2 dense layers with 1024 and 4 outputs, respectively.

The style uniformization network (U-2^{nd} stage) is trained by optimizing Equation (7) on WFLW and takes 128×128 RGB face crops. The generator is composed of an encoder and decoder part. Its encoder part performs blocks of 3×3 conv, batch norm, ReLU, followed 3×3 conv with stride 2, batch norm, ReLU. The number of channels is $64 \rightarrow 64 \rightarrow 128 \rightarrow 128 \rightarrow 256 \rightarrow 256$. We then flatten the 4×4 feature maps and add a large 4096-units dense layer. Then, the feature maps are reshaped to 4×4 and are passed to the decoder, that mirrors the encoder with upsampling operators. The discriminator mirrors the encoder part, except it does not have dense layers and has leaky ReLU activation everywhere. We set the hyperparameters to $\lambda^{per} = \lambda^{id} = 1$ which allows to generate high quality images in practice.

The landmark estimation network (LE-3^{rd} stage): is trained by optimizing Equation (10) on the train partitions of WFLW, 300W-LP, 300W and CelebA databases. We benchmarked several architectures, the first of which (AC-DC-s) is a 13-layers U-net like architecture with $64 \rightarrow 64 \rightarrow 128 \rightarrow 128 \rightarrow 256 \rightarrow 256 \rightarrow 512$ encoder part. The decoder mirrors the encoder part with a hint-based distillation using a particular GTN after each upsampling. Following [13], we set the distillation hyperparameter $\lambda_{kd} = 0.75$. We use AC-DC network [8] as our teacher architecture, taking as inputs 128×128 grayscale images. Also, as in [8], we set $\lambda_1 = 0.125, \lambda_2 = 0.25, \lambda_3 = 0.5, \lambda_4 = 1$.

We evaluate on the test partitions of WFLW as well as AFLW2000-3D. We also report results on degraded versions of WFLW and 300VW databases, that we refer to as **WFLW-degraded** and **300VW**, respectively. We report three evaluation metrics, the normalized mean error (NME), the failure rate or FR@0.1 and the AUC@0.1. For 2d alignment, the NME denotes the average landmark-wise distance normalized by the inter-ocular distance (distance between the outer eye corners). For 3d face alignment, as it is traditionnally done in the literature, we normalize the distances using the square root of the bounding box $height \times width$, as proposed in [35]. The FR@0.1 corresponds to the proportion of examples for which the NME is larger than 0.1, and AUC@0.1 is the integral or the cumulative error distribution (CED) curve for examples for which the NME is below 0.1. For head pose estimation, we report the average mean absolute difference (MAE) across the 3 Euler angles.

C. Evaluation of distilled models

Table I summarizes results obtained with distillation of AC-DC models. The teacher AC-DC model is very accurate, however, like LAB [28] it is very slow (500ms on an I7 CPU). As such, applying knowledge distillation to learn a lighter student network allows to dramatically speed-up the landmark estimation pipeline. AC-DC-s is the base student network described in Section IV-B. AC-DC-s-small denotes the same network as AC-DC-s that takes as input downsampled 64×64 face crops. Ac-DC-s-thin and AC-DC-s-sep are similar to AC-DC-s-small but with 2 times less channels per layer, or separable convolution everywhere, respectively. AC-DC-s-small appears as a suitable compromise, allowing for real-time

TABLE I

COMPARISON OF DISTILLED MODELS WITH BASELINES ON WFLW (LANDMARKS NME) AND AFLW (LANDMARKS NME AND HEAD POSE MAE) AS WELL AS INFERENCE TIME FOR A SINGLE IMAGE ON CPU.

architecture	WFLW	AFLW2000-3D	Pose	CPU Time(ms)
LAB [28]	5.27	-	-	800
3DDFA [35]	-	3.79	7.39	63
AC-DC	4.49	3.40	5.29	500
AC-DC-s	4.79	3.72	5.55	105
AC-DC-s-small	**4.94**	**3.80**	**5.59**	**60**
AC-DC-s-thin	5.37	4.07	5.69	48
AC-DC-s-sep	5.51	4.13	5.7	50

alignment on CPU at ≈ 16 fps with reasonable accuracy (e.g. compared with state-of-the-art methods such as LAB [28] and 3DDFA [35]). Furthermore, using smaller models result in downgraded accuracy and do not significantly decrease the computational burden. For the model involving separable convolution, however, this may be due to known issues in tensorflow implementation. As such, these distilled models for the **LE-3^{rd} stage)** are compact enough to be integrated into the whole three staged network.

D. Face alignment on degraded still images

Figure 5 shows comparisons between several models. We investigate multiple scenarios on WFLW database, including image downsampling (up to a factor 8), the addition of random hue/saturation/brightness/noise, as well as random scaling ($Sc \in [2/3; 4/3]$) and translation ($tr \in [-0.4; 0.4] \times$ iod). Figure 5 shows that while adding a first relocalization step allows a steady increase in accuracy, the most important gap is due to the addition of the uniformization stage in most configurations. Furthermore, relocalization and uniformization benefit from each other, thus the accuracy of the three-staged R-U-Le model is the best overall by a significant margin: for instance, with all the degradations at once the LE model has a NME of 32.98 whereas R-U-Le has 10.85, which makes an $\approx 67\%$ error decrease. Qualitative comparison between the models on Figure 6 illustrates that using both the relocalization (R) and uniformization (U) steps before estimating the landmark localization results in substantial improvement in the degraded case. Figure 6 shows a qualitative comparison between Le, R-Le, U-Le and R-U-Le models on WFLW-degraded. While on certain images (e.g. on top row), simply adding the relocalization step (second column) already provides a decent landmark estimation, in a number of cases (e.g. third row), the combination of both relocalization and uniformization stages is required to output a precise landmark estimation.

Figure 7 shows examples of successful landmark estimation with RULe on WFLW-degraded. First, we see that, generally speaking, given an approximate original crop, the relocalization network successfully outputs a transformation to retrieve a consistent, well-positionned face crop. From this face crop, the uniformization network can generate an upsampled, enhanced image, upon which a precise landmark estimation can be performed. We draw attention on the fact that, to be learned correctly, the uniformization networks requires (at train time) well-aligned images, as it is classical with GANs. Consequently, at test time, the capacity of the uniformization network to generate visually appealing images is closely related to the quality of the output of the relocalization network. Thus, the R, U, and Le networks complement each other to a certain extent. Notice, as a result, the quality of the uniformized images and subsequent landmark estimation, even with bad initial bounding box and $\times 8$ downsampling.

E. Face alignment on degraded video

Next, we measure the landmark estimation accuracy of R-U-Le on 300VW-degraded with both spatial ($\times 8$) and temporal downsampling ($\times 10$, which means we only keep one frame out of 10, causing localization errors as we initialize the bounding box with the last aligned frame). Results are displayed on Figure 8. In terms of NME (left graph), RULe is better in all cases: most notably, it allows to reduce the error from 15.81 to 6.6 in case of spatial downsampling, and from 19.84 to 11.08 $\%$ NME when we apply both spatial and temporal video downsampling. Moreover, it allows to substantially improve the AUC as well as to reduce the failure rate (FR, visible on the right graph) in case of temporal, spatial, and spatial+temporal downsampling. Thus, the 3-staged RULe architecture allows to mitigate the accuracy loss that results from skipped frames or fast head movement during tracking thanks to its relocalization stage. Furthermore, to a certain extent, it allows to precisely align face on low-resolution or badly illuminated/noisy images, as well as to precisely track small faces that are far from the camera.

V. DISCUSSION AND CONCLUSION

In this paper, we proposed a three-staged Relocalization-Uniformization-Landmark estimation (RULe) method for face alignment in degraded conditions. In the first stage, the relocalization network corrects imprecision in the face bounding box localization. In the second stage, from a refined face crop, the uniformization network enhances this crop quality, correcting illumination and low-resolution issues. On this enhanced image, face landmarks can be precisely localized. For that matter, we propose a compact cascaded design that is trained using hint-based knowledge distillation, allowing real-time processing on a single CPU. To further decrease the inference runtime, recent pruning [31] or quantization [32] can also be used in addition.

We showed experimentally that we can obtain state-of-the-art accuracy on both landmark estimation and head pose estimation by applying knowledge distillation on lighter models, while enabling real-time alignment on CPU. Moreover, when trying to estimate facial landmarks in degraded conditions (e.g. with low-resolutions images with noise, poor environmental lighting and brightness/contrast or, in the case of landmark estimation on video, poor bounding box initialization), the proposed three-staged RULe architecture dramatically increases the landmark estimation accuracy as compared to a single-staged one. This is due to the complementarity between its two first stages (relocalization

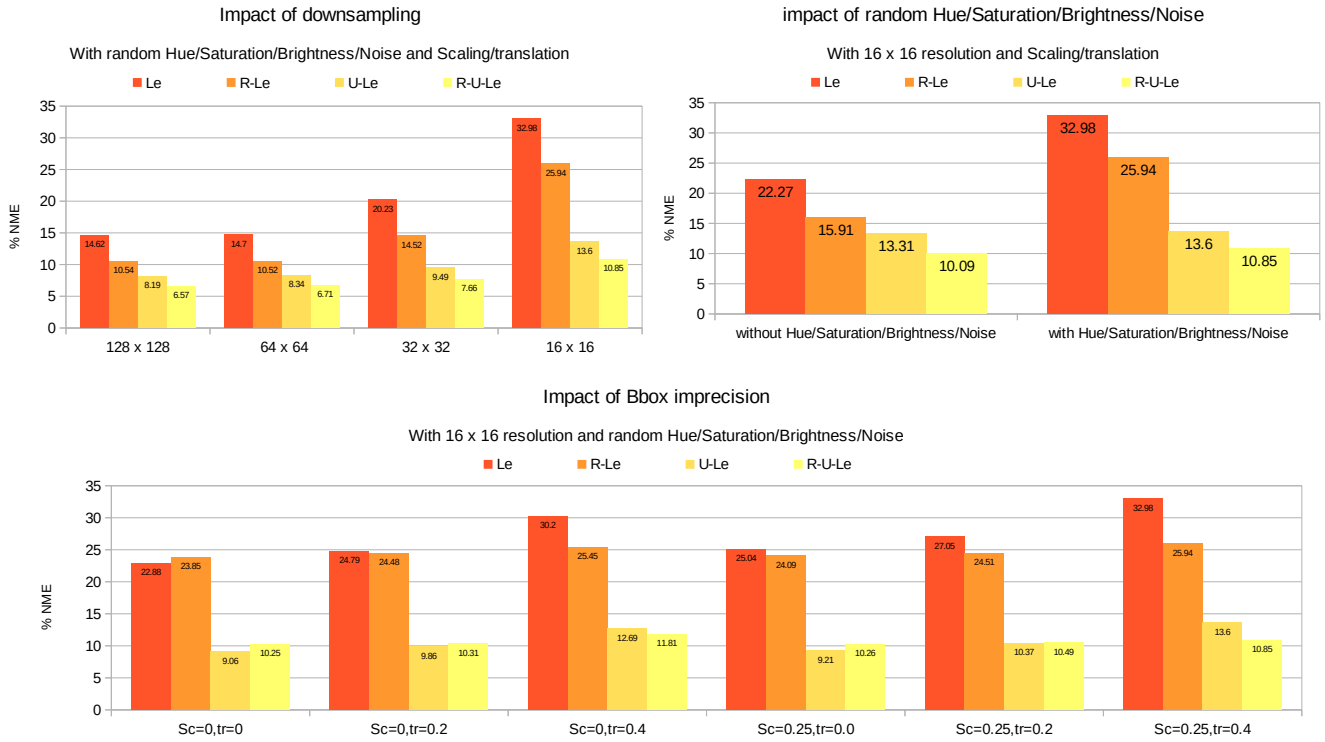

Fig. 5. Comparison on on WFLW-degraded in terms of robustness to downsampling, hue/saturation/brightness/noise and bounding box imprecision.

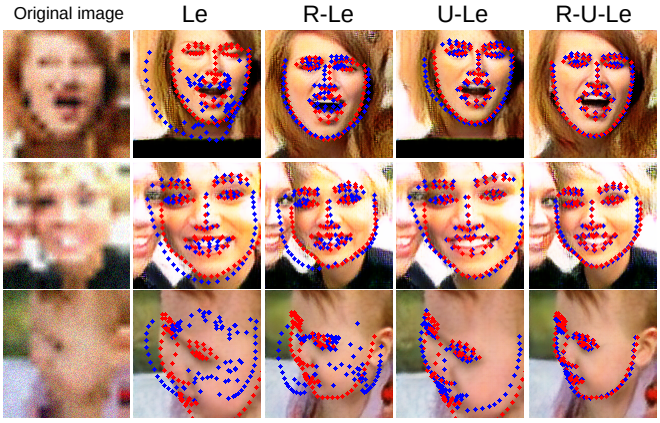

Fig. 6. Qualitative comparison on WFLW-degraded. From left to right: original images, and alignment (red: ground truth, blue: estimated landmarks) overlayed on clean image with Le, R-Le, U-Le and R-U-Le. The tree-staged R-U-Le architecture significantly improves the landmark estimation accuracy.

and uniformization), that allows to first accurately estimate a suitable face bounding box. Seeing such pre-registered crops that follows a similar distribution than that of the training phase, the uniformization stage can enhance the image quality in order to successfully estimate the landmark localization.

The proposed three-staged alignment framework is quite general and could be applied in a variety of cases. For example, the uniformization stage could theoretically encompass domain adaptation objectives, to e.g. perform face alignment on infra-red images. Similarly, it could be applied to closely related domains such as body pose estimation.

ACKNOWLEDGMENTS

This work has been supported by the French National Agency (ANR) (FacIL, project ANR-17-CE33-0002).

REFERENCES

[1] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein gan. arXiv preprint arXiv:1701.07875, 2017.
[2] E. Arnaud, A. Dapogny, and K. Bailly. Tree-gated deep mixture-of-experts for pose-robust face alignment. IEEE Transactions on Biometrics, Behavior, and Identity Science, 2019.
[3] A. Bulat and G. Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In CVPR, 2017.
[4] A. Bulat and G. Tzimiropoulos. Super-fan: Integrated facial landmark localization and super-resolution of real-world low resolution faces in arbitrary poses with gans. In CVPR, 2018.
[5] A. Dapogny and K. Bailly. Face alignment with cascaded semi-parametric deep greedy neural forests. Pattern recognition letters, 2018.
[6] A. Dapogny and K. Bailly. Investigating deep neural forests for facial expression recognition. In FG 2018 workshop on human behavior understanding), 2018.
[7] A. Dapogny, K. Bailly, and M. Cord. Decafa: Deep convolutional cascade for face alignment in the wild. ICCV, 2019.
[8] A. Dapogny, K. Bailly, and M. Cord. Deep entwined learning head pose and face alignment inside an attentional cascade with doubly-conditional fusion. In FG 2020, 2020.
[9] A. Dapogny, M. Cord, and P. Perez. The missing data encoder: Cross-channel image completion with hide-and-seek adversarial network. AAAI, 2020.
[10] X. Dong, Y. Yan, W. Ouyang, and Y. Yang. Style aggregated network for facial landmark detection. In CVPR, 2018.
[11] Z.-H. Feng, J. Kittler, M. Awais, P. Huber, and X.-J. Wu. Wing loss for robust facial landmark localisation with convolutional neural networks. In CVPR, 2018.
[12] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville. Improved training of wasserstein gans. In NeurIPS, 2017.
[13] G. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531, 2015.

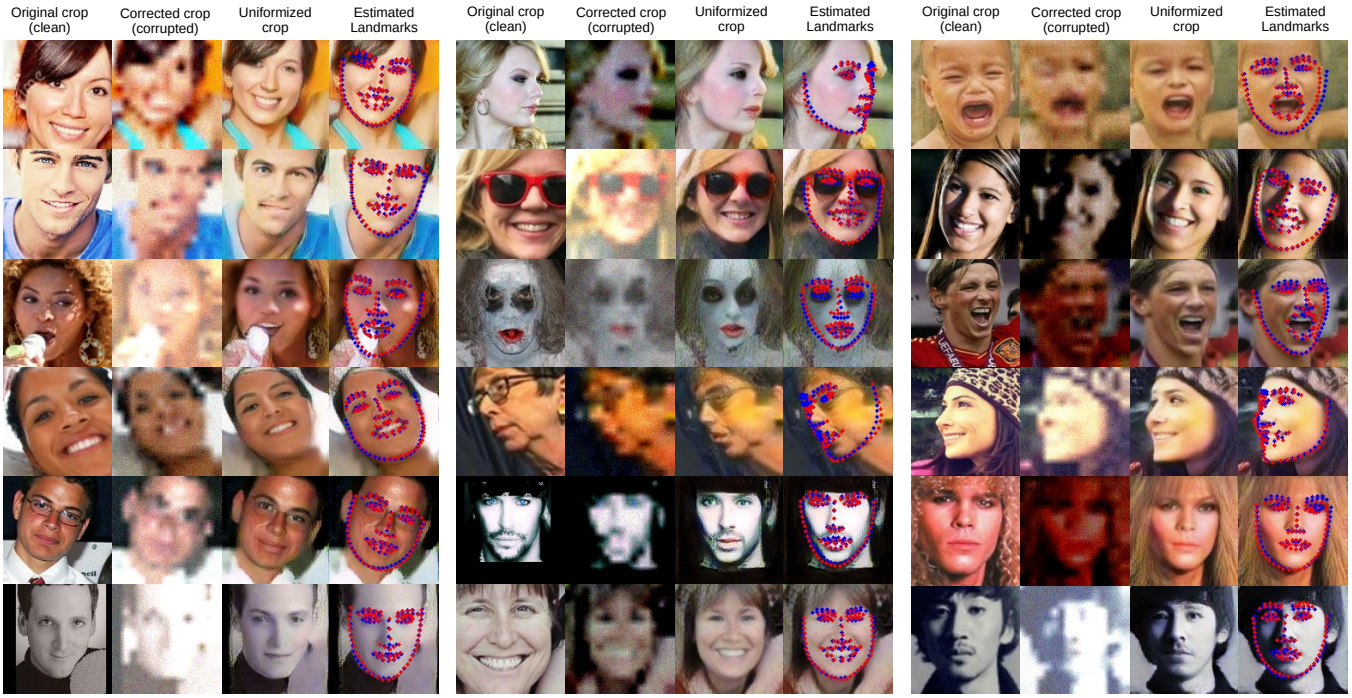

Fig. 7. Visualizations of bounding box relocalization, style uniformization and landmark estimation under scale/translation, low resolution and degraded lighting/noise conditions. Blue: Estimated landmarks, Red: ground truth localization.

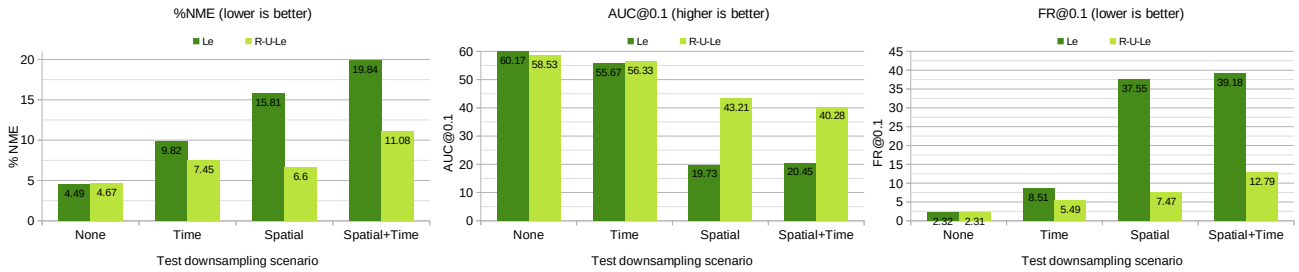

Fig. 8. Benchmark on 300VW-degraded. Comparisons between models (NME, AUC and FR@0.1) under time, spatial downsampling or both.

[14] Y. Huang, H. Yang, C. Li, J. Kim, and F. Wei. Adnet: Leveraging error-bias towards normal direction in face alignment. In ICCV, 2021.

[15] G.-S. Jison Hsu, C.-H. Tang, and M. Hoon Yap. Face synthesis and recognition using disentangled representation-learning wasserstein gan. In CVPR Workshops, 2019.

[16] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In ECCV, 2016.

[17] M. Kowalski, J. Naruniec, and T. Trzcinski. Deep alignment network: A convolutional neural network for robust face alignment. In CVPR workshops, 2017.

[18] X. Lan, Q. Hu, and J. Cheng. Revisting quantization error in face alignment. In ICCV workshops, 2021.

[19] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In ICCV, 2015.

[20] S. Ren, X. Cao, Y. Wei, and J. Sun. Face alignment at 3000 FPS via regressing local binary features. CVPR, 2014.

[21] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio. Fitnets: Hints for thin deep nets. arXiv preprint arXiv:1412.6550, 2014.

[22] C. Sagonas, E. Antonakos, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. 300 Faces In-The-Wild Challenge: database and results. IVC, 2015.

[23] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In CVPR, 2015.

[24] J. Shen, S. Zafeiriou, G. G. Chrysos, J. Kossaifi, G. Tzimiropoulos, and M. Pantic. The first facial landmark tracking in-the-wild challenge: Benchmark and results. In ICCV Workshops, 2015.

[25] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In CVPR, 2016.

[26] G. Trigeorgis, P. Snape, M. A. Nicolaou, E. Antonakos, and S. Zafeiriou. Mnemonic Descent Method: A Recurrent Process Applied for End-to-End Face Alignment. CVPR, 2016.

[27] P. Viola, M. Jones, et al. Rapid object detection using a boosted cascade of simple features. CVPR, 2001.

[28] W. Wu, C. Qian, S. Yang, Q. Wang, Y. Cai, and Q. Zhou. Look at boundary: A boundary-aware face alignment algorithm. In CVPR, 2018.

[29] X. Xiong and F. De La Torre. Supervised descent method and its applications to face alignment. CVPR, 2013.

[30] L. Yue, X. Miao, P. Wang, B. Zhang, X. Zhen, and X. Cao. Attentional alignment network. BMVC, 2018.

[31] E. Yvinec, A. Dapogny, M. Cord, and K. Bailly. Red: Looking for redundancies for data-freestructured compression of deep neural networks. Advances in Neural Information Processing Systems, 2021.

[32] E. Yvinec, A. Dapogny, M. Cord, and K. Bailly. Spiq: Data-free per-channel static input quantization. WACV, 2022.

[33] Z. Zhang, P. Luo, C. C. Loy, and X. Tang. Learning deep representation for face alignment with auxiliary attributes. PAMI, 2016.

[34] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li. Face alignment across large poses: A 3d solution. In CVPR, pages 146–155, 2016.

[35] X. Zhu, X. Liu, Z. Lei, and S. Z. Li. Face alignment in full pose range: A 3d total solution. PAMI, 2019.