Cross-modal Deep Face Normals with Deactivable Skip Connections

Victoria Fernández Abrevaya*[1], Adnane Boukhayma*†[2], Philip H. S. Torr[3], Edmond Boyer[1]
[1] Inria, Univ. Grenoble Alpes, CNRS, Grenoble INP, LJK, France
[2] Inria, Univ. Rennes, CNRS, IRISA, M2S, France
[3] University of Oxford, UK
{victoria.fernandez-abrevaya, adnane.boukhayma, edmond.boyer}@inria.fr
philip.torr@eng.ox.ac.uk

Abstract

We present an approach for estimating surface normals from in-the-wild color images of faces. While data-driven strategies have been proposed for single face images, limited available ground truth data makes this problem difficult. To alleviate this issue, we propose a method that can leverage all available image and normal data, whether paired or not, thanks to a novel cross-modal learning architecture. In particular, we enable additional training with single modality data, either color or normal, by using two encoder-decoder networks with a shared latent space. The proposed architecture also enables face details to be transferred between the image and normal domains, given paired data, through skip connections between the image encoder and normal decoder. Core to our approach is a novel module that we call deactivable skip connections, which allows integrating both the auto-encoded and image-to-normal branches within the same architecture that can be trained end-to-end. This allows learning of a rich latent space that can accurately capture the normal information. We compare against state-of-the-art methods and show that our approach can achieve significant improvements, both quantitative and qualitative, with natural face images.

1. Introduction

3D reconstruction of the human face is a long-standing problem in computer vision, with a wide range of applications including biometrics, forensics, animation, gaming, and human digitalization. In many of these applications monocular inputs are considered in order to limit the acquisition constraints, hence enabling uncontrolled environments as well as efficient information usage for *e.g.* facial telecommunication and entertainment. Although significant progress has been recently made by the scientific community, recovering detailed 3D face models given only single images is still an open problem.

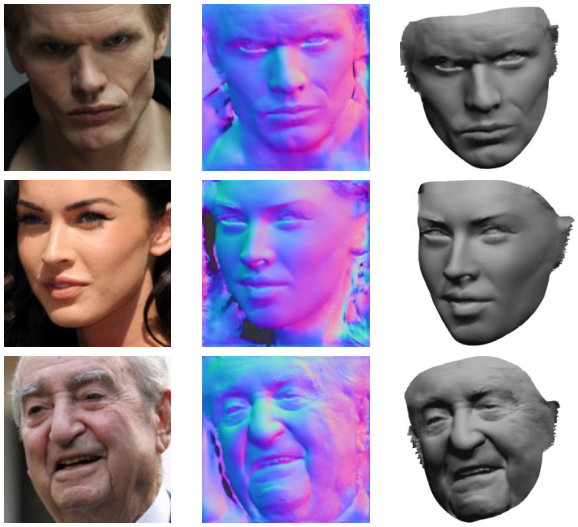


Figure 1: Our model predicts accurate normals from a single input image that can be used to enhance a coarse geometry (e.g. PRN [15]).

Monocular face reconstruction is in essence an ill-posed problem which requires strong prior knowledge. Assuming a simple shading model, seminal shape-from-shading (SfS) approaches [21, 65] were estimating shape normals by considering local pixel intensity variations. Fine scale surface details can be recovered using this strategy, however the applicability to in-the-wild images is limited by the simplified image formation model that is assumed. Later on, a more global strategy was proposed with parametric face models [7, 60]. They allow fitting a template face controlled by only a few coefficients, resulting hence in improved robustness. While being widely adopted, parametric models are inherently restricted in expressiveness and have difficulties

* Authors contributed equally
† This work was done while the author was at University of Oxford

in recovering small surface details, as a consequence of their low dimensional representation. Recently, deep learning methods that exploit large-scale face image datasets have been investigated with the aim of a better generalization. While most work in this category are trained to estimate the coefficients of a parametric model [54, 53, 19, 28, 44], a few other approaches infer directly per-pixel depth [45], UV position maps [15] or surface normals [58, 46].

As observed in previous work [48, 66], regressing depth information alone can lead to suboptimal results, especially detail-wise, as the inherent scale ambiguity with single images can make convergence difficult for neural networks. On the other hand, the estimation of normals appears to be an easier task for such networks, given that normals are strongly correlated to pixel intensities and depend mostly on local information, a fact already exploited by SfS techniques. Still, only a few approaches have been proposed in this line for facial images [47, 46], mostly due to the limited available ground-truth data. We propose here a method that overcomes this limitation and can leverage all data available through the use of cross-modal learning. Our experiments demonstrate that this strategy can estimate more accurate and sharper facial surface normals from single images.

The proposed approach recovers accurate normals corresponding to the facial region within an RGB image, with the goal of enhancing an existing coarse reconstruction, [15] in our experiments. We cast the problem as a color-to-normal image translation, which can be in principle solved by combining an image encoder E_I with a normal decoder D_N as in [58], and including skip connections between E_I and D_N [42] in order to transfer details from the image domain to the normals domain. However, training such a network can prove difficult unless a large dataset of image/normal pairs, that ideally contains images in-the-wild, is available. In practice few such datasets are currently publicly available, *e.g.* [64], which were moreover captured under controlled conditions. To improve generalization, we propose to augment the architecture with a normal encoder E_N and an image decoder D_I, where all encoders/decoders share the same latent space. This augmented architecture provides additional constraints on the latent space with the auto-encoded image-to-image and normal-to-normal branches, allowing therefore for a much wider range of training datasets. In order to keep advantage of the skip connections between E_I and D_N, while avoiding the resulting bonded connections between E_N with D_N that hamper the architecture, we introduce the *deactivable skip connections*. This allows skip connections to be turned on and off during training according to the type of data.

In summary, this work contributes (1) a framework that leverages cross-modal learning for the estimation of normals from a single face image in-the-wild; (2) the introduction of the *deactivable skip connection*; and (3) an extensive evaluation that shows that our approach outperforms state-of-the-art methods on the Photoface [64] and Florence [3] datasets, with up to nearly 10% improvements in angular error on the Florence dataset, as well as visually compelling reconstructions.

2. Related Work

We focus the discussion below on methods that consider 3D face reconstruction, or normal estimation, given single RGB images.

Reconstruction with Parametric Models 3D reconstruction from a single image is ill-posed and many methods resort therefore to strong priors with parametric face models such as blendshape [16, 9, 55] or statistical models, typically the 3D Morphable Model (3DMM) [7]. These models are commonly used within an analysis-by-synthesis optimization [41, 22, 13, 8, 18] or, more recently, using deep learning to regress model parameters [39, 40, 54, 59, 19, 15, 28, 52, 44], or alternatively to regress other face information using 3DMM training data, for instance volumetric information [25], UV position map [15], normal map [58], depth map [45], or the full image decomposition [47, 46, 29]. This strategy has proven robustness, however it is constrained by the parametric representation that offers limited expressiveness and fails in recovering fine scale details.

In order to improve the quality of the reconstructions several works have proposed to add medium-scale correctives on top of the parametric model [33, 17, 53], to train a local wrinkle regressor [9], or to learn deep non-linear 3DMM [57, 67] that can capture higher-frequency details. Our method also enables to enhance a face prediction through the estimation of more accurate normals.

Normal Estimation with Shape from Shading Shape-from-shading (SfS) [21, 65] is a well-studied technique that aims at recovering detailed 3D surfaces from a single image based on shading cues. It estimates surface normals using the image irradiance equation, as well as illumination model parameters when these are unknown. SfS is inherently limited by the simplified image formation model assumed but has inspired numerous works that build on the correlation between pixel intensity and normals, either explicitly or implicitly. For instance, a few works on faces combined SfS with a data-driven model, *e.g.* [49, 27, 50], which helps to avoid some of the limitations such as ill-posedeness and ambiguities *e.g.* [6]. The recent works of Shu *et al.* [47] and Sengupta *et al.* [46] use deep neural networks to decompose in-the-wild facial images into surface normals, albedo and shading, assuming Lambertian reflectance and using a semi-supervised learning approach inspired by SfS. Our work follows a similar direction and estimates the normal information from a single image, but unlike [47, 46] we do

not rely on an image formation model and let instead the network learn such a transformation from real data.

Normal Estimation with Deep Networks Closely related to our work are methods that recover surface normals from an image using deep neural networks, *e.g*. [61, 14, 63, 32, 5, 37, 66, 38, 12, 48, 2]. Yoon *et al*. [63] and Bansal *et al*. [4] focus on the normal prediction task in order to recover detailed surfaces. Eigen and Fergus [14] simultaneously regress depth, normal and semantic segmentation using a multi-scale approach. Zhang and Funkhouser [66] predict surface normal and occlusion boundaries to later optimize for depth completion; a similar direction was followed by [38] for outdoor scenes. Trigeorgis *et al*. [58] estimate facial normals with a supervised approach trained on synthetic data. Our approach differs from the aforementioned methods with a new architecture that enables cross-modal learning, hence improving performances in monocular 3D face normal estimation.

Geometry Enhancement using Deep Networks Methods have been proposed that directly enhance face models using deep neural networks. Richardson *et al*. [40] use two networks where the first estimates a coarse shape and the second one refines the depth map from the previous branch, using an SfS-inspired unsupervised loss function. Sela *et al*. [45] recover the depth and correspondence maps coupled with an off-line refinement step. The works of [62, 23] estimate high frequency details by training with very accurate ground-truth data, which requires a careful acquisition process and high-quality inputs. Tran *et al*. [56] estimate a per-pixel bump map, where the ground-truth data is obtained by applying an SfS method offline. The work of [10] learns to estimate a geometric proxy and a displacement map for details primarily for high resolution images (2048×2048). While they mention limitations with low resolution images, we show results with resolutions as low as 256×256.

3. Method

We propose to predict face normals from a single color image using a deep convolutional encoder-decoder network. A natural solution to this purpose is to combine an image encoder E_I with a normal decoder D_N, as in *e.g*. [58]. However training such an architecture requires pairs of normal and color images in correspondence. Although a few public datasets are available that contain high-quality 3D or normal ground-truth information for faces, for instance ICT-3DRFE [51] or Photoface [64], they were obtained under controlled conditions and do not, therefore, really cover the distribution of images in-the-wild. On the other hand, numerous large datasets of natural images are publicly available, for example CelebA [34] and AffectNet [35], yet without the associated accurate and detailed ground-truth normal values. Whereas other works have approached this by augmenting the training corpus with synthetic ground-truth [58, 46], we propose instead a method based on cross-modal learning that can leverage all available data, even unpaired.

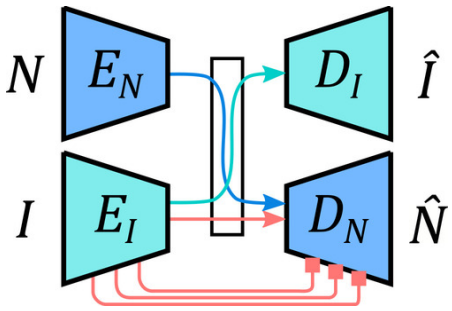
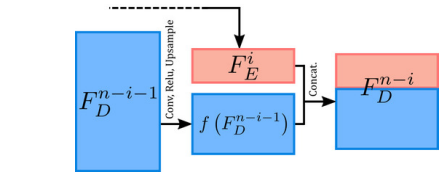


Figure 2: Overview of the proposed approach. Our cross-modal architecture allows exploitation of paired and unpaired image/normal data for image-to-normal translation (Red), by means of further image-to-image (Green) and normal-to-normal (Blue) regularizations during training. The *deactivable skip connections* allow to transfer details from the image encoder E_I to the normal decoder D_N without having to link the normal encoder E_N to the normal decoder D_N.

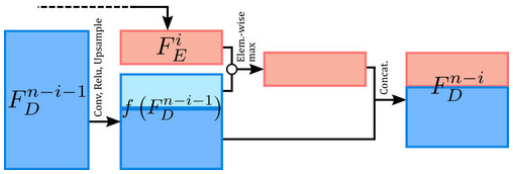
3.1. Cross-modal Architecture

As depicted in Fig. 2, we use two encoder/decoder networks, one for images E_I/D_I and one for normals E_N/D_N, sharing the same latent space. This architecture is trained with image-to-image, normal-to-normal and image-to-normal supervision simultaneously in order to obtain a rich and robust latent representation. To this purpose, we exploit paired images of normal and color information on faces, as available from [51, 64], in addition to individual images of either color or normal information, from *e.g*. CelebA-HQ [26] and BJUT-3D [1]. To improve the overall performance we augment this architecture with long skip connections between E_I and D_N, as it favors the transfer of details between the image and normal domains, and since it has been shown to significantly increase performance in several image translation tasks *e.g*. [24]. In practice we use a U-Net+ResNet [42, 20] architecture that combines the benefits of both short and long skip connections.

Training such an architecture end-to-end raises an obstacle: the skip connections from E_I to D_N ($E_I \rightarrow D_N$), which are based on concatenating feature maps, impose, by construction, to also have skip connections between the encoder and decoder of the normal modality, i.e. $E_N \rightarrow D_N$. This is counterproductive in practice: by setting skip connections within the same modality, it is in fact easier for the normal autoencoder to transfer features from the earliest layers of its encoder to the last layers of its decoder



(a) Standard skip connection



(b) Deactivable skip connection

Figure 3: Instead of concatenating the encoder features (red) and decoder features (blue), as with standard skip connections, we fuse the encoder features with part of the decoder features (light blue), to be able to deactivate this operation when needed.

through the skip connection, thus depriving the deeper layers of any meaningful gradients during training. Not only will this fail to improve the latent face representation, but it will also alter the coefficients of the normal decoder for the image-to-normal inference task.

For this reason, we introduce the *deactivable skip connections* as shown in Fig. 3 and detailed in Sec. 3.2. This allows us to train the framework end-to-end by setting long connections solely between E_I and D_N, thus learning a rich latent space that encodes facial features from both color and normal images while profiting from all available data.

3.2. Deactivable Skip Connections

As mentioned earlier, skip connections are well suited to our problem as they allow sharing of low-level information at multiple scales while still preserving the general structure. In the implementation of standard skip connections, as in [42, 24], the decoder features at the $(n-i)^{th}$ layer F_D^{n-i} are the concatenation of the processed previous layer features $f(F_D^{n-i-1})$ and the encoder features at layer i, F_E^i, where n is the total number of layers (see Fig.3a).

Let $m\left(F_{E_I}^i\right)$ be the number of feature maps at the i^{th} layer of E_I. The proposed architecture (Fig. 2) requires to set connections from the image encoder E_I to the normal decoder D_N, and as a consequence, each layer features $F_{D_N}^{n-i}$ of D_N are expected to always have an additional $m\left(F_{E_I}^i\right)$ channels. In order to gain generalization over each domain, both the color and the normal images can be auto-encoded during training. However, since the concatenation is expected during training on the decoder D_N side, features of the normal encoder E_N must be concatenated as well, which is as discussed detrimental to our model.

The *Deactivable Skip Connections* are designed such that, during training, the transfer of feature maps from encoders to decoders can be selectively activated or deactivated. Compared to a decoder equipped with standard skip connections, the processed features $f\left(F_D^{n-i-1}\right)$ of our decoder include $m\left(F_E^i\right)$ extra channels (light blue in Fig.3b). During a normal-to-normal pass, the skip connections are deactivated and the $(n-i)^{th}$ layer features of the normal decoder correspond to the processed previous layer features *e.g.* $F_D^{n-i} = f\left(F_D^{n-i-1}\right)$. During an image-to-normal pass, the skip connection is activated: we first perform an element-wise max-pooling between the i^{th} layer features of the encoder F_E^i and the last $m\left(F_E^i\right)$ channels of the processed $(n-i-1)^{th}$ layer features of the decoder $f\left(F_D^{n-i-1}\right)$, as illustrated in Fig. 3b. The result is stacked back with the remaining of the processed previous layer features thus forming the final $(n-i)^{th}$ decoder layer features F_D^{n-i}. Doing so allows to transfer the information from encoder to decoder without degrading performances when the transfer operation does not occur, as when auto-encoding normals.

3.3. Training

We train the framework end-to-end using both supervised and unsupervised data, where the latter includes individual image and normal datasets. During training, the skip connections are deactivated when doing a normal-to-normal pass. For the supervised case, and for unsupervised normals, the loss function is the cosine distance between the output and the ground-truth, which in our experiments gave better results than the $L1/L2$ norm:

$$\mathcal{L}_{nrm}(N,\hat{N}) = 1 - \frac{1}{|N|}\sum_{(i,j)}\frac{N(i,j)^\top \cdot \hat{N}(i,j)}{||N(i,j)||_2||\hat{N}(i,j)||_2}, \quad (1)$$

where $N(i,j)$ and $\hat{N}(i,j)$ are the normal vectors at pixel (i,j) in the ground-truth and output normal images N and $\hat{N}$ respectively, and $|N|$ is the number of pixels in N. For unsupervised image data we use the $L2$ loss:

$$\mathcal{L}_{img}(I,\hat{I}) = ||I-\hat{I}||_2^2, \quad (2)$$

where $\hat{I}$ is the output color image and I the ground-truth. In both cases, the loss is applied only on facial regions segmented using masks obtained as explained in Sec.4.1.

In practice, as we can only perform a training iteration for one input modality at a time, either an input batch of images or normals, we train our model as follows: when loading a batch of images with image/normal ground-truth pairs, we perform a normal-to-normal iteration first, followed by an image-to-normal plus image-to-image iteration, where both losses in the latter iteration are summed with equal

weights. When loading a batch of images only, we perform an image-to-image iteration. Finally, with a batch of normals only, we naturally proceed with a normal-to-normal iteration alone.

4. Evaluation

We report below on the accuracy of the normals estimated with our approach on standard datasets [64, 3]. We compare against state-of-the-art methods on normal estimation and 3D reconstruction, and show significant improvements in terms of normal prediction accuracy. This is supported by compelling reconstructions of images in-the-wild from 300-W [43], as can be seen in Figs. 4 and 5.

Following previous works [46, 58], we evaluate with the mean angular error between the output and the ground-truth normals, as well as percentage of pixels within the facial region with an angular error of less than 20°, 25° and 30°. For qualitative comparisons we show both the output normal map, as well as the mesh results obtained by enhancing the output of PRN [15] using normal mapping [11]: we append the predicted normals to the the PRN meshes pixel-wise thus rendering enhanced geometric shading.

4.1. Implementation Details

The framework was implemented in PyTorch [36], and all experiments were run on a GTX TITAN Black. The networks were trained for 40 epochs using ADAM solver [31] with a learning rate of 10^{-4}. We use a ResNet-18 [20] architecture and set five skip connections, one at the output of the initial layer and the rest at the output of each of the four residual blocks. Each mini-batch during training consists of data of the same type, *i.e.* images only, normals only or image-normal pairs, as this worked best for us empirically.

Similar to prior work, input images are crops of fixed size around the face. We extract 2D keypoints with a face detector [30] and create masks on the facial region by finding the tightest square of edge size l around the convex hull of the points. The images are then cropped with a square patch of size $1.2 \times l$ centered at the same 2D location as the previously detected box, and subsequently resized to 256×256. The code will be made publicly available.

4.2. Datasets

Our training set comprises multiple datasets: ICT-3DRFE [51] and Photoface [64] which provide image/normal pairs, CelebA-HQ [26] which only contains 2D images, and BJUT-3D [1], which consists of high-quality 3D scans.

We generated 8625 image/normal pairs from ICT-3DRFE by randomly rotating the 345 3D models and relighting them using the provided albedos. We sampled random rotation axes and angles in $[-\pi/4, \pi/4]$, random lighting directions with positive z, and random intensities in $[0, 2]$. For Photoface, following the setting in [58, 46], we randomly selected a training subset of 353 people resulting in 9478 image/normal pairs. We also generated 5000 high resolution facial images from CelebA-HQ, which is used to train the image-to-image branch exclusively. In addition, we render 3000 normal images from the 500 scans of BJUT-3D, rotated with random axes and angles in $[-\pi/4, \pi/4]$. We only render normal images from this dataset as the original scan color images are not provided.

For evaluation purposes we use the remaining testing subset of Photoface, which consists of 100 subjects not seen during training and 1489 image/normal pairs. This subset challenges the reconstruction with very severe lighting conditions. Following the work of [15], we create an additional evaluation set by rendering 530 color and normal facial images from the 53 3D models of the Florence dataset [3], rotated with random axes and angles in $[-\pi/4, \pi/4]$. This allows to evaluate on a completely unseen dataset. Finally, we use the 300-W dataset [43] of 2D face images to assess qualitative performances in-the-wild. Note that for both training and testing, we limited ourselves to 3D face datasets of high quality and details.

	Mean±std	$< 20^o$	$< 25^o$	$< 30^o$
Pix2V [45]	33.9±5.6	24.8%	36.1%	47.6%
Extreme [56]	27.0±6.4	37.8%	51.9%	64.5%
3DMM [58]	26.3±10.2	4.3%	56.1%	**89.4%**
3DDFA [68]	26.0±7.2	40.6%	54.6%	66.4%
SfSNet [46]	25.5±9.3	43.6%	57.5%	68.7%
PRN [15]	24.8±6.8	43.1%	57.4%	69.4%
Ours	**22.8±6.5**	**49.0%**	**62.9%**	74.1%
UberNet [32]	29.1±11.5	30.8%	36.5%	55.2%
NiW [58]	22.0±6.3	36.6%	59.8%	79.6%
Marr Rev [4]	28.3±10.1	31.8%	36.5%	44.4%
SfSNet-ft [46]	12.8±5.4	83.7%	90.8%	94.5%
Ours-ft	**12.0±5.3**	**85.2%**	**92.0%**	**95.6%**

Table 1: Quantitative comparisons on the Photoface dataset [64] with mean angular errors (degrees) and percentage of errors below 20°, 25° and 30°. `-ft` means that the method was fine-tuned on Photoface.

4.3. Comparisons

We compare our results to methods that explicitly recover surface normals, either for facial images (SfSNet [46], NiW [58]) or for general scenes (Marr Rev [4], UberNet [32]). We also compare against state-of-the-art approaches for 3D face reconstruction, namely the classic 3DMM fitting method used in [58], 3DDFA [68], the bump map regression based approach of [56] and the combined regression+shape-from-shading approach of [45].

Quantitative results can be found in Table 1 for Photo-

(a) input (b) Ours (c) SfSnet (d) PRN (e) Extreme (f) Pix2V (g) 3DDFA

Figure 4: Qualitative comparisons on normals in the 300-W dataset [43].

	Mean±std	$< 20^o$	$< 25^o$	$< 30^o$
Extreme [56]	19.2±2.2	64.7%	75.9%	83.3%
SfSNet [46]	18.7±3.2	63.1%	77.2%	86.7%
3DDFA [68]	14.3±2.3	79.7%	87.3%	91.8%
PRN [15]	14.1±2.16	79.9%	88.2%	92.9%
Ours	**11.3±1.5**	**89.3%**	**94.6%**	**96.9%**

Table 2: Quantitative comparisons on the Florence dataset [3] with mean angular errors (degrees) and percentage of errors below 20°, 25° and 30°.

face and Table 2 for Florence datasets. We show results of our method both with (`Ours-ft`) and without (`Ours`) fine-tuning of the training split of Photoface in the upper and lower parts of Table 1 respectively. The same is done with SfSNet. The error values on Photoface for the methods of [46, 58, 45, 4, 32] are as reported in [46], and we use the publicly available implementations of [56, 68, 15] for the others. For the Florence dataset we use the publicly available implementations. Note that, to be able to evaluate the per-pixel normal accuracy, we can only compare to 3D reconstruction methods whose output is aligned with the image. For a fair comparison, all methods were given facial images of size 256×256 as input, resized if necessary.

The proposed approach shows the best values both in mean angular error and percentage under 20°, 25° and 30° degrees, only outperformed by 3DMM on errors under 30°. As noted by the authors in [58], 3DMM fitting performs well under 30° because of the coarseness of the model and the keypoint supervision, but its performance on tighter angles drops drastically as it lacks precision. We found that, although [45, 56] usually provide seemingly detailed reconstructions, the actual normals of these methods lack accuracy as witnessed by their numbers.

Our good performance is also confirmed by qualitative comparisons over images in-the-wild in various head poses and under arbitrary lighting conditions as can be seen in

(a) Input (b) Ours+PRN (c) SfSNet+PRN (d) PRN (e) Extreme (f) Pix2Vertex (g) 3DDFA

Figure 5: Qualitative comparisons on geometries in the 300-W dataset [43].

Figs. 4 and 5. For comparisons with mesh results (Fig. 5), we show for both our approach and SfSNet [46] the normal mapping over the same base mesh, obtained using PRN [15], and we refer to these as Ours+PRN and SfSNet+PRN respectively. We show our meshes from two views to illustrate that the output is not optimized for a particular viewpoint, a known limitation with SfS. Compared to SfSNet we recover much more refined details that significantly enhance the base mesh. Compared to Extreme [56] our approach does not include unnecessary additional noise. As observed by other authors, Pix2Vertex [45] cannot handle difficult poses or illuminations, and sometimes simply fails to converge. Both PRN and 3DDFA can correctly recover the general structure of the face, although their goal was not to recover surface details as we do.

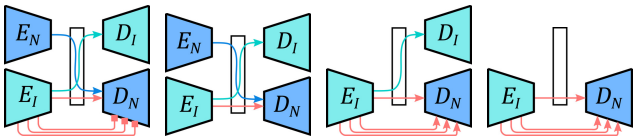
We believe our improved results are due to the fact that we do not rely on a parametric model for training data generation, as was done in *e.g.* [46], as well as the strongly regularized latent space that is learned through the two encoder/decoder networks, in addition to the skip connections that can transfer the necessary details.

4.4. Ablation

We evaluate here the influence of the proposed architectural components. In particular, we compare against the alternatives shown in Fig. 6: our model without skip connections (Fig. 6b), without the normal encoder E_N (Fig. 6c), and without both the normal encoder E_N and image decoder D_I (Fig. 6d), *i.e.* a basic encoder-decoder architecture. Since there is no need in the last two cases for deactivable skip connections we use standard ones. We show quantitative results in Table 7, and qualitative examples in Fig. 8.

Our final model outperforms the alternatives both quantitatively and qualitatively which validates the proposed cross-modal architecture design, and the benefit of the introduced deactivable skip connections.

For example, we can see in the geometric shape of the



(a) Ours (b) w/o skip co. (c) w/o E_N (d) w/o E_N, D_I

Figure 6: Architectures for the ablation test: (a) our proposed architecture, (b) without skip connections, (c) without the normal encoder and (d) without the normal encoder and the image decoder.

	Mean±std	$< 20^o$	$< 25^o$	$< 30^o$
w/o skip co.(Fig.6b)	24.4 ±6.7	46.6%	60.6%	72.0%
w/o E_N, D_I(Fig.6d)	23.3 ±6.3	47.7%	61.9%	73.3%
w/o E_N(Fig.6c)	23.0 ±6.8	47.6%	61.5%	73.1%
Ours (Fig.6a)	**22.8±6.5**	**49.0%**	**62.9%**	**74.1%**

(a) On Photoface [64]

	Mean±std	$< 20^o$	$< 25^o$	$< 30^o$
w/o skip co.(Fig.6b)	12.6±1.4	85.8%	92.6%	95.8%
w/o E_N(Fig.6c)	12.4±1.6	86.0%	92.6%	95.9%
w/o E_N, D_I(Fig.6d)	12.0±1.2	87.8%	94.1%	96.7%
Ours (Fig.6a)	**11.3±1.5**	**89.3%**	**94.6%**	**96.9%**

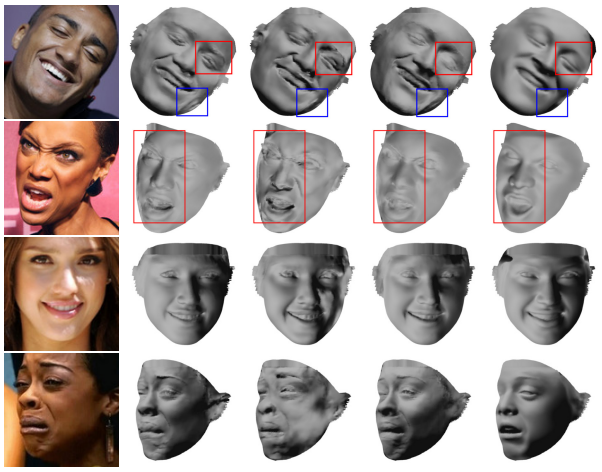
(b) On Florence [3]

Figure 7: Quantitative comparisons between architectures: the proposed architecture (*Ours*), without skip connections (*w/o skip co.*), without the normal encoder (*w/o* E_N) and without the normal encoder and the image decoder (*w/o* E_N, D_I).

eyelids in the first row of Fig. 8 and the shading in the second row that our final model gets the best from each of the alternatives. Our correct global shape estimate is comparable to that of the cross-modal model without skip connections, although the latter is smoother and clearly lacks details. Additionally we can see that removing the image decoder D_I and normal encoder E_N (*i.e.* a standard encoder-decoder with skip connections) gives poor results for images-in-the-wild, due to the domain gap between training and evaluation. This can be visualized particularly in the artifacts appearing on the third and fourth examples, or the inaccurate shadings of the second example. Finally, our fine details are comparable to those of the model with skip connections but without the normal encoder E_N, which in turn has a reduced ability to represent the shape accurately, since it has not learned an additional prior on the geometric aspects of the face.

4.5. Limitations

The proposed method still has limitations, some of which are shown in Fig. 9. These belong to extreme situations that represent outliers to the training data, including faces in very severe lighting/shades (Fig.9a,9b), occlusion (Fig.9c,9d), very low quality images (Fig.9e) and unusual facial textures (Fig.9f).



(a) Input (b) Ours (c) w/o E_N D_I (d) w/o E_N (e) w/o skip co.

Figure 8: Qualitative comparisons between architectures: (b) our proposed architecture, (c) without the normal encoder and the image decoder, (d) without the normal encoder, and (e) without skip connections.

(a) (b) (c) (d) (e) (f)

Figure 9: Failure cases.

5. Conclusion

We presented a novel deep-learning based approach for the estimation of facial normals in-the-wild. Our method is centered on a new architecture that combines the robustness of cross-modal learning and the detail transfer ability of skip connections, enabled thanks to the proposed *deactivable skip connections*. By leveraging both paired and unpaired data of image and normal modalities during training, we achieve state-of-the-art results on angular estimation errors and obtain visually compelling enhanced 3D reconstructions on challenging images in-the-wild. Among the limitations of our work are the inability to properly handle occlusions (as it is mostly a local method) and to recover finer-details, *e.g.* pore-level details, which are directions that will be tackled in future work.

References

[1] The bjut-3d large-scale chinese face database. 3, 5

[2] Thiemo Alldieck, Gerard Pons-Moll, Christian Theobalt, and Marcus Magnor. Tex2shape: Detailed full human body geometry from a single image. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019. 3

[3] Andrew D. Bagdanov, Alberto Del Bimbo, and Iacopo Masi. The florence 2d/3d hybrid face dataset. In *ACM Workshop on Human Gesture and Behavior Understanding*, 2011. 2, 5, 6, 8

[4] Aayush Bansal, Bryan Russell, and Abhinav Gupta. Marr revisited: 2d-3d alignment via surface normal prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 3, 5, 6

[5] Jan Bednarik, Pascal Fua, and Mathieu Salzmann. Learning to reconstruct texture-less deformable surfaces from a single view. In *Proceedings of International Conference on 3D Vision*, 2018. 3

[6] Peter N Belhumeur, David J Kriegman, and Alan L Yuille. The bas-relief ambiguity. *International Journal of Computer Vision*, 1999. 2

[7] Volker Blanz, Thomas Vetter, et al. A morphable model for the synthesis of 3d faces. In *Proceedings of ACM Siggraph*, 1999. 1, 2

[8] James Booth, Anastasios Roussos, Evangelos Ververas, Epameinondas Antonakos, Stylianos Ploumpis, Yannis Panagakis, and Stefanos Zafeiriou. 3d reconstruction of in-the-wild faces in images and videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018. 2

[9] Chen Cao, Derek Bradley, Kun Zhou, and Thabo Beeler. Real-time high-fidelity facial performance capture. *ACM Transactions on Graphics*, 2015. 2

[10] Zhang Chen, Guli Zhang, Ziheng Zhang, Kenny Mitchell, Jingyi Yu, et al. Photo-realistic facial details synthesis from single image. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019. 3

[11] Jonathan Cohen, Marc Olano, and Dinesh Manocha. Appearance-preserving simplification. In *Proceedings of the Annual Conference on Computer Graphics and Interactive Techniques*, 1998. 5

[12] Dapeng Du, Limin Wang, Huiling Wang, Kai Zhao, and Gangshan Wu. Translate-to-recognize networks for rgb-d scene recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 3

[13] Bernhard Egger, Sandro Schönborn, Andreas Schneider, Adam Kortylewski, Andreas Morel-Forster, Clemens Blumer, and Thomas Vetter. Occlusion-aware 3d morphable models and an illumination prior for face image analysis. *International Journal of Computer Vision*, 2018. 2

[14] David Eigen and Rob Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *Proceedings of the IEEE international conference on computer vision*, 2015. 3

[15] Yao Feng, Fan Wu, Xiaohu Shao, Yanfeng Wang, and Xi Zhou. Joint 3d face reconstruction and dense alignment with position map regression network. In *Proceedings of the European Conference on Computer Vision*, 2018. 1, 2, 5, 6, 7

[16] Pablo Garrido, Levi Valgaerts, Chenglei Wu, and Christian Theobalt. Reconstructing detailed dynamic face geometry from monocular video. *ACM Transactions on Graphics*, 2013. 2

[17] Pablo Garrido, Michael Zollhöfer, Dan Casas, Levi Valgaerts, Kiran Varanasi, Patrick Pérez, and Christian Theobalt. Reconstruction of personalized 3d face rigs from monocular video. *ACM Transactions on Graphics*, 2016. 2

[18] Baris Gecer, Stylianos Ploumpis, Irene Kotsia, and Stefanos Zafeiriou. Ganfit: Generative adversarial network fitting for high fidelity 3d face reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2

[19] Kyle Genova, Forrester Cole, Aaron Maschinot, Aaron Sarna, Daniel Vlasic, and William T Freeman. Unsupervised training for 3d morphable model regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 3, 5

[21] Berthold KP Horn and Michael J Brooks. *Shape from shading*. 1989. 1, 2

[22] Patrik Huber, Guosheng Hu, Rafael Tena, Pouria Mortazavian, P Koppen, William J Christmas, Matthias Ratsch, and Josef Kittler. A multiresolution 3d morphable face model and fitting framework. In *Proceedings of the International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, 2016. 2

[23] Loc Huynh, Weikai Chen, Shunsuke Saito, Jun Xing, Koki Nagano, Andrew Jones, Paul Debevec, and Hao Li. Mesoscopic facial geometry inference using deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 3

[24] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 3, 4

[25] Aaron S Jackson, Adrian Bulat, Vasileios Argyriou, and Georgios Tzimiropoulos. Large pose 3d face reconstruction from a single image via direct volumetric cnn regression. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017. 2

[26] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017. 3, 5

[27] Ira Kemelmacher-Shlizerman and Ronen Basri. 3d face reconstruction from a single image using a single reference face shape. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010. 2

[28] Hyeongwoo Kim, Pablo Garrido, Ayush Tewari, Weipeng Xu, Justus Thies, Matthias Niessner, Patrick Pérez, Christian Richardt, Michael Zollhöfer, and Christian Theobalt. Deep video portraits. *ACM Transactions on Graphics*, 2018. 2

[29] Hyeongwoo Kim, Michael Zollhöfer, Ayush Tewari, Justus Thies, Christian Richardt, and Christian Theobalt. Inverse-facenet: Deep monocular inverse face rendering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2

[30] Davis E. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 2009. 5

[31] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[32] Iasonas Kokkinos. Ubernet: Training a universal convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 3, 5, 6

[33] Hao Li, Jihun Yu, Yuting Ye, and Chris Bregler. Realtime facial animation with on-the-fly correctives. *ACM Transactions on Graphics*, 2013. 2

[34] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, 2015. 3

[35] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 2017. 3

[36] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*. 2019. 5

[37] Xiaojuan Qi, Renjie Liao, Zhengzhe Liu, Raquel Urtasun, and Jiaya Jia. Geonet: Geometric neural network for joint depth and surface normal estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 3

[38] Jiaxiong Qiu, Zhaopeng Cui, Yinda Zhang, Xingdi Zhang, Shuaicheng Liu, Bing Zeng, and Marc Pollefeys. Deeplidar: Deep surface normal guided depth prediction for outdoor scene from sparse lidar data and single color image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 3

[39] Elad Richardson, Matan Sela, and Ron Kimmel. 3d face reconstruction by learning from synthetic data. In *Proceedings of the intenational conference on 3D vision*, 2016. 2

[40] Elad Richardson, Matan Sela, Roy Or-El, and Ron Kimmel. Learning detailed face reconstruction from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2, 3

[41] Sami Romdhani and Thomas Vetter. Estimating 3d shape and texture using pixel intensity, edges, specular highlights, texture constraints and a prior. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2005. 2

[42] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2015. 2, 3, 4

[43] Christos Sagonas, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2013. 5, 6, 7

[44] Soubhik Sanyal, Timo Bolkart, Haiwen Feng, and Michael J Black. Learning to regress 3d face shape and expression from an image without 3d supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2

[45] Matan Sela, Elad Richardson, and Ron Kimmel. Unrestricted facial geometry reconstruction using image-to-image translation. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017. 2, 3, 5, 6, 7

[46] Soumyadip Sengupta, Angjoo Kanazawa, Carlos D Castillo, and David W Jacobs. Sfsnet: Learning shape, reflectance and illuminance of facesin the wild'. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2, 3, 5, 6, 7

[47] Zhixin Shu, Ersin Yumer, Sunil Hadap, Kalyan Sunkavalli, Eli Shechtman, and Dimitris Samaras. Neural face editing with intrinsic image disentangling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2

[48] David Smith, Matthew Loper, Xiaochen Hu, Paris Mavroidis, and Javier Romero. Facsimile: Fast and accurate scans from an image in less than a second. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019. 2, 3

[49] William AP Smith and Edwin R Hancock. Recovering facial shape using a statistical model of surface normal direction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2006. 2

[50] Patrick Snape and Stefanos Zafeiriou. Kernel-pca analysis of surface normals for shape-from-shading. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014. 2

[51] Giota Stratou, Abhijeet Ghosh, Paul Debevec, and Louis-Philippe Morency. Effect of illumination on automatic expression recognition: a novel 3d relightable facial database. In *The IEEE International Conference on Automatic Face and Gesture Recognition*, 2011. 3, 5

[52] Ayush Tewari, Florian Bernard, Pablo Garrido, Gaurav Bharaj, Mohamed Elgharib, Hans-Peter Seidel, Patrick Pérez, Michael Zollhofer, and Christian Theobalt. Fml: face model learning from videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2

[53] Ayush Tewari, Michael Zollhöfer, Pablo Garrido, Florian Bernard, Hyeongwoo Kim, Patrick Pérez, and Christian Theobalt. Self-supervised multi-level face model learning for monocular reconstruction at over 250 hz. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2

[54] Ayush Tewari, Michael Zollhofer, Hyeongwoo Kim, Pablo Garrido, Florian Bernard, Patrick Perez, and Christian Theobalt. Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017. 2

[55] Diego Thomas and Rin-Ichiro Taniguchi. Augmented blendshapes for real-time simultaneous 3d head modeling and facial motion capture. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 2

[56] Anh Tuan Tran, Tal Hassner, Iacopo Masi, Eran Paz, Yuval Nirkin, and Gérard G Medioni. Extreme 3d face reconstruction: Seeing through occlusions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 3, 5, 6, 7

[57] Luan Tran, Feng Liu, and Xiaoming Liu. Towards high-fidelity nonlinear 3d face morphable model. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2

[58] George Trigeorgis, Patrick Snape, Iasonas Kokkinos, and Stefanos Zafeiriou. Face normals" in-the-wild" using fully convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2, 3, 5, 6

[59] Anh Tuan Tran, Tal Hassner, Iacopo Masi, and Gérard Medioni. Regressing robust and discriminative 3d morphable models with a very deep neural network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2

[60] Daniel Vlasic, Matthew Brand, Hanspeter Pfister, and Jovan Popović. Face transfer with multilinear models. In *ACM Transactions on Graphics*, 2005. 1

[61] Xiaolong Wang, David Fouhey, and Abhinav Gupta. Designing deep networks for surface normal estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 3

[62] Shuco Yamaguchi, Shunsuke Saito, Koki Nagano, Yajie Zhao, Weikai Chen, Kyle Olszewski, Shigeo Morishima, and Hao Li. High-fidelity facial reflectance and geometry inference from an unconstrained image. *ACM Transactions on Graphics*, 2018. 3

[63] Youngjin Yoon, Gyeongmin Choe, Namil Kim, Joon-Young Lee, and In So Kweon. Fine-scale surface normal estimation using a single nir image. In *Proceedings of the European Conference on Computer Vision*, 2016. 3

[64] Stefanos Zafeiriou, Mark Hansen, Gary Atkinson, Vasileios Argyriou, Maria Petrou, Melvyn Smith, and Lyndon Smith. The photoface database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2011. 2, 3, 5, 8

[65] Ruo Zhang, Ping-Sing Tsai, James Edwin Cryer, and Mubarak Shah. Shape-from-shading: a survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1999. 1, 2

[66] Yinda Zhang and Thomas Funkhouser. Deep depth completion of a single rgb-d image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2, 3

[67] Yuxiang Zhou, Jiankang Deng, Irene Kotsia, and Stefanos Zafeiriou. Dense 3d face decoding over 2500fps: Joint texture & shape convolutional mesh decoders. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2

[68] Xiangyu Zhu, Xiaoming Liu, Zhen Lei, and Stan Z Li. Face alignment in full pose range: A 3d total solution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. 5, 6