S2F2: Self-Supervised High Fidelity Face Reconstruction from Monocular Image

Abdallah Dib  Junghyun Ahn  Cédric Thébault  Philippe-Henri Gosselin  Louis Chevallier

InterDigital R&I

Fig. 1. Given a single image, our method achieves appealing 3D face reconstruction and estimates a dense detailed face geometry, spatially varying face reflectance (diffuse and specular albedos) and high frequency scene illumination.

Abstract—We present a novel face reconstruction method capable of reconstructing detailed face geometry, spatially varying face reflectance from a single monocular image. We build our work upon the recent advances of DNN-based auto-encoders with differentiable ray tracing image formation, trained in self-supervised manner. While providing the advantage of learning-based approaches and real-time reconstruction, the latter methods lacked fidelity. In this work, we achieve, for the first time, high fidelity face reconstruction using self-supervised learning only. Our novel coarse-to-fine deep architecture allows us to solve the challenging problem of decoupling face reflectance from geometry using a single image, at high computational speed. Compared to state-of-the-art methods, our method achieves more visually appealing reconstruction.

I. INTRODUCTION

Fast, robust and high fidelity 3D face reconstruction has a wide range of applications in many domains such as interactive face editing, video-conferencing, XR, Metaverse applications, and visual effects for movies post-production. Several approaches such as [41], [19], [20], [30] achieve high fidelity face reconstruction, but require complex hardware setups (multi-view, lightstage). They are therefore not easily usable in most of the aforementioned applications. Significant progress was made to achieve high quality reconstruction from monocular image/video using optimization-based frameworks. Such methods [17], [12] are generally slow, of limited robustness and not suitable for interactive scenarios. Also, their performances in challenging conditions (non-uniform lighting, extreme poses) are limited.

Deep-based analysis-by-synthesis approaches have been investigated to leverage the generalization capabilities of machine learning. However, these methods [37], [39] generally sacrifice reconstruction quality. Methods combining CNNs with differentiable rendering trained in a self-supervised manner have been introduced by Tewari et al. [38], [37], [36]. These methods directly regress the parameters of a statistical morphable model and achieve real-time performance but fall short of the quality and fidelity because their estimated geometry and reflectance are bound by the statistical prior space which limits its generalization regarding the real diversity of face geometry and reflectance.

More recently, many works aim to improve the realism and fidelity of deep-based methods by capturing either detailed geometry or reflectance but not both, which we discuss next. First, to capture detailed geometry, and because of the complexity of the problem, several methods rely on ground truth dataset obtained either from multi-view reconstruction setup (and/or lightstage) [44], [8], [25], [45], from synthetic data [32], [33] or from a mixture of both [46], [1]. Feng et al. [16] is the only self-supervised method that captures detailed geometry. However, this method only captures medium-scale geometry details and misses high-frequency geometry variations. Additionally, its estimated reflectance is restricted to the statistical prior space which limits its generalization regarding the real diversity of face geometry and reflectance. Second, and to improve the reflectance, Dib et al. [13] combined ray tracing and self-supervised learning to capture medium-scale reflectance details. However, this method
restricts the estimated geometry to a parametric face model preventing high-frequency facial details (such as wrinkles, folds...) to be captured. To our knowledge, there is no existing self-supervised methods that can jointly estimate detailed geometry and reflectance.

The first contribution of this work, is the introduction of the first self-supervised method that jointly estimate detailed geometry and reflectance. This is accomplished via our novel coarse-to-fine architecture, with an adapted training strategy which allows our method to efficiently solve the ambiguous problem of separating detailed geometry from reflectance from a single image taken under uncontrolled lighting conditions.

The second contribution, is the combination, for the first time, of differentiable ray tracing with vertex-based renderer at training time to overcome the problem of edge discontinuities of the ray tracing. This allow our method to benefit from both renderers. On one hand, ray tracing accurately models self-shadows and on the second hand, the vertex-based renderer evaluates correctly the whole geometry including boundaries. This leads to a significant improvement in the estimated geometry compared to Dib et al. [13] that uses only ray tracing.

Finally, the aforementioned contributions enable to take a big leap forward in reconstruction quality for self-supervised methods and lead to superior face reconstruction when compared to recent state-of-the art methods. To our knowledge, this is the first time a self-supervised method reaches this level of fidelity and realism. Our robust face attributes estimation (diffuse, specular and normal) leads to practical applications such as face attribute editing and relighting.

II. RELATED WORKS

Methods such as In this work, we are interested in face reconstruction/tracking methods that only use image or video as input and do not require any external hardware setup beyond the camera. These methods can be split into two categories: optimization-based and learning-based approaches.

Geometry and reflectance modeling Statistical 3D Morphable Models - 3DMM - is the main building block for a wide range of optimization-based and learning-based methods [5], [14]. This statistical model adds a lot of structure to the face reconstruction problem from monocular image or video and makes it tractable. However, due to the low-dimensional space of 3DMM, subject specific medium and high frequency geometry and albedo details cannot be modeled. [34] proposes a reflectance model for 3DMM incorporating a diffuse and specular priors. In this work, we base our reconstruction on the 3DMM geometry with the statistical diffuse and specular prior of [34] and we train a novel multi-stage deep network to capture fine diffuse and geometry details.

Most of optimization-based methods like [17], [2], [12] rely on the same 3DMM parametrization, they provide generally precise reconstruction at the expense of a high computation cost and are sensitive to difficult lighting conditions.

Among the learning-based methods, deep convolution neural networks (CNN) are effective at direct face reconstruction [39], [36], [24], [10], [40], [21]. The advantage of these self-supervised methods is that they can be trained on large corpus of unlabeled images. However they generally fall short of reconstruction precision because of their simplified underlying scene parameterization (pure-Lambertian BRDF to model skin reflectance and low-order spherical harmonics to model light). Their inability to model self-shadows is also a possible reason for their instability under challenging lighting conditions. Dib et al. [13] proposes a self-supervised method that uses ray tracing for image formation which significantly improves over these methods and solves many of these limitations. However the reconstructed geometry of their method is still limited by 3DMM space.

Detailed geometry reconstruction Capturing fine detailed geometry on top of global face structure is a pre-condition to achieve high fidelity face reconstruction. Because of the complexity of the problem, methods such Cao et al. [7] relies on ground truth dense data [3], or data captured using a lightstage or multi-view ([45], [44], [8], [25]) to solve the problem. However, acquiring such ground truth data is not always possible.

Optimization based methods like [22], [17] use shape-from-shading [47] to capture fine geometry details. However these methods may not generalize well and are computationally expensive.

Some deep-based methods rely partially on synthetic data ([32], [29], [46]) or a mix of labeled and unlabeled data [1] to capture fine detailed geometry. The bias introduced by these methods may impede the resulting precision due to the mismatch with real data distribution. Recently, Feng et al. [16] learns an ‘expression-dependent’ displacement model in-the-wild and is the only method that relies only on unlabeled images for end-to-end training. However this method only captures medium-frequency displacement map and their estimated reflectance is restricted to the statistical albedo prior space which limits their reconstruction quality.

To our knowledge, our method is the first self-supervised method that jointly estimates: geometry at high frequency, spatially varying personalized skin reflectance with diffuse and specular albedos and high frequency illumination from a single monocular image.

Differentiable rendering Differentiable rendering is a key block in the context of analysis-by-rendering and several implementations exist. Tewari et al. [38] proposed an efficient vertex-based differentiable rendering that can only handle pure Lambertian BRDF and cannot capture self-shadows. Dib et al. [11], [12] introduced a method which uses differentiable ray tracing for face reconstruction within a classic optimization framework. The key advantage of ray tracing over vertex-based renderer is the capacity of ray tracing to handle self-shadows where a visibility mask is calculated for each surface point with respect to each light during direct illumination pass. However, differentiable ray tracing is computationally expensive and memory consum-
ing. Recently, Dib et al. [13] uses differentiable ray tracing in conjunction with a deep neural architecture for direct face reconstruction. In this scheme, inference does not incur the speed penalty of ray tracing and delivers near real-time performance with robust reconstruction in challenging lighting conditions. Regarding the optimization process, a limitation of ray tracing is the noise on gradients originating at the objects boundaries as they are sampled by very few points. In this work, we combine a vertex-based renderer with a ray tracing renderer together with a deep neural architecture that takes advantage of both renderers.

III. METHOD

Our goal is to obtain a high fidelity face reconstruction with faithful separation between reflectance and geometry attributes. To solve this ill-posed problem, we propose a novel multi-stage deep architecture, wherein different stages progressively refine the reconstruction.

Our network architecture, depicted in Figure 2, is composed of 3 stages denoted: ‘Coarse’, ‘Medium’ and ‘Fine’. The ‘Coarse’ reconstruction relies on the statistical geometry and albedo priors space. This base reconstruction lacks important geometry and albedo (diffuse and specular) details and albedo priors space. This base reconstruction lacks important geometry and albedo (diffuse and specular) details and loss function minimization during training:

\[ E_d(\chi) + E_p(\alpha, \beta) + E_b(\delta), \]

where \( E_d \) is the data term equal to:

\[ E_d(\chi) = E^{S}_{ph}(\chi) + w_{dr}E^{R}_{ph}(\chi) + w_{lm}E_{land}(\chi), \]

with \( E^{S}_{ph} \) is the pixel-wise photo-consistency loss between input and ray traced pixels, \( p_i^S \in \mathbb{R}^3 \):

\[ E^{S}_{ph}(\chi) = \sum_{i} |p_i^S(\chi) - p_i|, \]

where \( p_i^S = F(\chi) \), with \( F \) the Monte Carlo estimator of the rendering equation [23]. \( E^{R}_{ph} \) is the vertex-based photo-consistency loss between the projected mesh and the corresponding pixels in the input image, defined as follows:

\[ E^{R}_{ph}(\chi) = \sum_{i=1}^{N} |B(n_i, c_i, R_i) - I(\Pi \circ C(v_i))|, \]

where \( N \) is the number of vertices, \( C(v_i) \) is the projection of vertex \( v_i \) in the real image, equal to: \( R^{-1}(v_i - T) \). \( II \) is the perspective camera matrix that projects a 3D vertex to a 2D pixel. \( B \) is the final irradiance equal to the sum of the diffuse and specular terms weighted by the specular intensity \( s_i \) (details on \( B \) in supplementary material section II). \( E_{land} \) is the landmark loss, which measures the distance between the \( L = 68 \) predicted facial landmarks and the projection of their corresponding vertex on the input image. These landmarks are obtained using the landmarks detector of [6]. \( E_{p} \) is the statistical prior that regularizes against implausible face geometry and reflectance [12]. \( E_b(\delta) \) is a soft-box constraint that maintains \( \delta \) in the range \([0, 1]\).

Edge discontinuities Ray tracing can naturally models self-shadows by building a visibility mask for each surface point with respect to each light. However, the major drawback of differentiable ray tracing is the discontinuities along geometry edges ([27], [26]). In fact, when solving for the rendering equation via Monte Carlo ray tracing [23], very

\[ |B(n_i, c_i, R_i) - I(\Pi \circ C(v_i))|, \]

where \( N \) is the number of vertices, \( C(v_i) \) is the projection of vertex \( v_i \) in the real image, equal to: \( R^{-1}(v_i - T) \). \( II \) is the perspective camera matrix that projects a 3D vertex to a 2D pixel. \( B \) is the final irradiance equal to the sum of the diffuse and specular terms weighted by the specular intensity \( s_i \) (details on \( B \) in supplementary material section II). \( E_{land} \) is the landmark loss, which measures the distance between the \( L = 68 \) predicted facial landmarks and the projection of their corresponding vertex on the input image. These landmarks are obtained using the landmarks detector of [6]. \( E_{p} \) is the statistical prior that regularizes against implausible face geometry and reflectance [12]. \( E_b(\delta) \) is a soft-box constraint that maintains \( \delta \) in the range \([0, 1]\).

Edge discontinuities Ray tracing can naturally models self-shadows by building a visibility mask for each surface point with respect to each light. However, the major drawback of differentiable ray tracing is the discontinuities along geometry edges ([27], [26]). In fact, when solving for the rendering equation via Monte Carlo ray tracing [23], very...
few points are sampled on these areas. As a result, backpropagation fails to handle geometry edges accurately during the optimization. Several solutions have been proposed to overcome this limitation but they are generally very computationally expensive ([28], [26]). For instance, [26] explicitly samples the geometry edges, which extremely penalizes the training time as it needs to calculate the silhouette edges of the geometry. While landmarks are mainly used to guide the training, in the particular case of ray tracing they can help mitigating the aforementioned limitation. However the geometry is not as precise as it could be. As an efficient solution, we introduce in this work a new loss term $E_{ph}^d$ (eq. 4) which relies on a vertex-based differentiable renderer. This leads to a more accurate reconstruction, by taking advantage of ray tracing (which can model self-shadows) and vertex-based rendering (for better gradients on geometry edges) without significant cost. For instance, it only takes 370 ms to process (forward-backward) an image using our method compared to 42 seconds for the method of [26].

C. Medium stage

The albedos (diffuse and specular) and geometry estimated by the previous stage are bound by the statistical prior space and can only capture low frequency components of the skin reflectance and geometry. Our goal is to obtain personalized albedos (outside this space) with detailed geometry. Estimating these parameters jointly in a self-supervised manner is challenging. For this, we proceed with a coarse-to-fine strategy and we start by capturing personalized medium diffuse and specular albedos outside the statistical prior space. The challenge here is to avoid mixing diffuse and specular albedos (outside this space) with detailed geometry. Estimating reflectance and geometry. Our goal is to obtain personalized albedos (outside this space) with detailed geometry. Estimating reflectance and geometry. Our goal is to obtain personalized albedos (outside this space) with detailed geometry.

Fig. 2. Our network architecture, trained end-to-end in a self-supervised manner, estimates face attributes (reflectance and detailed geometry) in a coarse-to-fine fashion (refer to section III).

$\Delta_t$ increments to be added on top of the previously estimated textures, $D$ and $S$, respectively. The resultant textures, $\hat{D}$ and $\hat{S}$, are used to generate a new image $I_{cur}^2$. We note that the second stage has only access to latent space of $E$ allows this stage to focus on separating medium diffuse from specular albedo without worrying about high frequency geometry details that are discarded naturally by design. We define $\chi = \{\alpha, \delta, \phi, \gamma, \hat{D}, \hat{S}\}$ and we minimize the following loss function:

$$E_d(\chi) + E_{sym}(\hat{A}, \hat{A}) + w_m E_m(\hat{A}) + w_b E_b(\hat{A}),$$

where $E_{sym}(\hat{A}, A) = w_s E_s(\hat{A}) + w_s E_s(\hat{A}, A)$, and $A$ is either $D$ or $S$. $E_m$ and $E_b$ are the symmetry and consistency regularizers (similar to [13]) used to avoid baking residual shadows in the personalized albedos. $E_m$ is a constraint term that ensures local smoothness at each vertex, with respect to its first ring neighbors.

D. Fine stage

While the previous stage allows to obtain more personalized diffuse and specular albedos, these albedos remain generally blurry and still miss details. Also the geometry is restricted to the low-dimensional space of 3DMM. For this, we leverage the U-net based architecture (with skip connections) which are very efficient at capturing these fine details. First, using the pose and the geometry produced by the first stage, we project the input image $I$ in the uv-space. This projection, denoted as $I_{uv}$, is passed to two U-net networks, $U_G$ and $U_D$. $U_G$ predicts a normal map $\Delta_n$, and $U_D$ predicts an increment $\Delta_d$ that
is added to the estimated diffuse albedo $\bar{D}$ of the previous stage. We denote $\bar{D}$ as the resulting texture.

We define $\bar{\chi} = \{a, \delta, \phi, \gamma, M, D\}$ and we train $U_G$ and $U_D$ in a self-supervised manner by minimizing the following loss function:

$$E_\alpha(\bar{\chi}) + E_{ac}(\bar{D}, \bar{D}) + w_s^{\|}E_m(\bar{A}) + w_s^{\|}E_b(\bar{A}), \tag{6}$$

where $E_{ac}(\bar{D}, \bar{D}) = w_s^{\|}E_s(\bar{D}) + w_s^{\|}E_c(\bar{D}, \bar{D})$, and $\bar{A}$ is either $D$ or $M$. The regularization terms $E_s$ and $E_c$ play an important role in avoiding baking unexplained shadows in diffuse texture $\bar{D}$ in case our lighting model did not recover the light correctly. Also these regularizers contribute in producing a good separation between diffuse and geometry details. However, they sacrifice some albedo details (please refer to the ablation section V). Finally, we note that we experimented using an additional U-net to capture a specular increment $\Delta_s$ (similar to the diffuse) but we did not obtain substantial improvements in the reconstruction quality.

### E. Training strategy

We proceed with the following training strategy. We first train $E$ (with the fully-connected layer) for 30 epochs, in a non-supervised manner, to directly regress $\gamma$ by minimizing eq. 1. Next, we train $D_A$, $D_s$, and $E$ for 10 epochs while minimizing the loss in eq. 5. We follow the same training strategy proposed by [13] to separate diffuse and specular albedos, which consists in starting with a high regularization value of $w_c$ for diffuse texture (in eq. 5), and then in progressively relaxing its value during training to allow for the diffuse albedo to capture more details. Next, we fix $D_A$ and $D_s$, then we train $U_G$, $U_D$ and $E$ for 5 epochs, to estimate a normal map and an enhanced diffuse map respectively, by minimizing the photo-consistency loss in eq. 6. To avoid over-fitting one component, and to obtain a plausible separation between these attributes, we start with a high weight for $w_f^s$ and we progressively relax this constraint to allow $D$ to capture more albedo details.

Finally, we note that the vertex-based loss in eq. 4 is used in the whole training process with the goal to assist and guide the pixel-wise photo-consistency loss of ray tracing (eq. 3) at different stages, so to alleviate the problem of noisy edge gradients that ray tracking exhibits.

### IV. Results

Figure 1 and supp. material (section I) show successful face reconstruction of more than 100 subjects with challenging face details, extreme lighting conditions, challenging head pose/expression and different skin type (makeup, beards)... For all these subjects, our method successfully estimates personalized albedos and captures fine detailed geometry, which leads to appealing reconstruction at high fidelity. Even in challenging lighting conditions, our method successfully estimates shadow-free maps (diffuse, specular and normal). All this, at very high computational speed, where at inference, our method takes 131 ms to process an image on a Nvidia RTX 2080 Ti. We note that while ray tracing penalizes our training time, it is not needed at inference time and our estimated attributes are compatible with existing rendering engines. Finally, our robust estimation of scene light, face reflectance and geometry provides explicit controls over the face attributes which leads to practical applications such as face attribute editing (aging) and relighting as shown in supp. material (section I). Finally, in supp. video, we also show successful reconstruction on video sequence (Implementation details are in supp. material, section III).

### V. Ablation

#### Importance of vertex-based renderer

In this experiment, we study the importance of the vertex-based renderer to overcome the problem of noisy edge gradients of the ray tracer. For this, we trained $E$ by dropping the vertex-based renderer based loss term (eq 4) from equation 2. We compare the estimated mesh to the one that use the full energy term (ray tracing + vertex-wise). The results in Figure 3 a) show that the estimated meshes using both the vertex-based renderer and ray tracer are more accurate than the ones obtained using ray tracing only (especially around the mouth edges). Quantitatively, we evaluate both methods on 100 subjects from Facescape dataset [45] with various type of facial expressions. We compute the vertex position error with respect to ground truth mesh, and we obtain 2.288/1.671 mm (mean error/std deviation) for the ’vertex-based + ray tracing’ compared to 2.831/1.782 mm for ’ray tracing only’.

#### Regularization

In this experiment, we study the importance of the symmetry and consistency regularizers ($E_{sc}$) used in equation 6 to separate the diffuse and geometric details faithfully. For this, we trained $U_G$ and $U_D$ by dropping these two regularizers. For both subjects in Figure 3 b), without these regularizers, some geometric details get baked in the albedo and leads to sub-optimal separation. Adding these regularizers produce more convincing separation. While these regularizers play an important role in obtaining a correct separation between diffuse and geometry details, they sacrifice some albedo details.

#### Multi-Stage reconstruction

In this experiment we show the
importance of the 'Medium' and 'Fine' stages to improve the realism of the 'Coarse' reconstruction. Figure 3 c) shows the reconstruction obtained from the 'Coarse' stage with the estimated statistical albedo priors. For the 'Medium' stage, we show the final reconstruction with the final diffuse and normal maps. We note that the diffuse albedo in 'Medium' stage is generally blurry and lacks some details and is significantly enhanced in the 'Fine' stage. Also, the detailed geometry captured in the 'Fine' stage significantly improves the realism of the final reconstruction. Quantitatively, we calculate the SSIM between the reconstruction of each stage and the original input image. On 1000 images, we obtain an average of 0.89/0.91/0.97 for the coarse/medium/fine stages respectively (higher is better).

VI. COMPARISON

In this section, we compare, qualitatively and quantitatively, our method to the state-of-the-art methods.

A. Qualitative comparison

Figure 4 shows comparison against Chen et al. [8], Yang et al. [45] and Feng et al. [16]. For [8] and [16], results are from authors’ open implementation. For [45] results are generated by the authors. The methods of [8] and [45] use ground truth (GT) data to train their generative network while our method and [16] are self-supervised methods and do not require any GT data. For all subjects, our method successfully capture most of geometry details especially for the top subjects that present challenging details. These details are barely captured or missed by other methods. Also our method shows better results on the wrinkles formed by the zygomaticus muscles (smile wrinkles around nose and mouth). Our method has a significantly better shape and expression recovery than all other methods, especially around the mouth.

Figure 5 show comparison against the method of Abrevaya et al. [1] on subjects with challenging facial details. The method of Abrevaya et al. [1] uses a combination of labeled and unlabeled data to train the network that predicts normal map of the face. It also produces a complete normal map for the entire head (including eyes and moth interior) while our method restricts the reconstruction to the frontal region of the face (without eyes and mouth interior). However, our method captures more facial details (especially around eyes) than [1].

Compared to Dib et al. [13] (Figure 6), our method achieves robustness against challenging lighting conditions similar to [13] and produces shadow free albedos (first two subjects). In addition, our method estimates better diffuse map and captures detailed geometry, while [13] restricts the geometry reconstruction to the low-dimensional space of 3DMM. This leads to a superior and high fidelity reconstruction of our method compared to [13].

A. Quantitative comparison

Geometric evaluation We first compare our estimated geometry to the state-of-the-art methods of Chen et al. [8], Feng et al. [16], and Dib et al. [13] on the Superfaces dataset [4] composed of 20 high resolution 3D ground truth (GT) face meshes (Table I and Figure 7). Table I reports, for each method, on all subjects, the average error $\mu$ and standard deviation $\sigma$ for vertex position error. As shown in Figure 7, the same mask is used for all methods. Feng et al.[16] achieves slightly better results than our method on average error while ours has a smaller standard deviation. Our method achieves better score than Chen et al. [8] and Dib et al. [13]. As depicted in Figure 7, our approach measures lowest error around the mouth.

We also evaluate our method on the NoW dataset [31], which is composed of 80 subjects with a total of 1702
images. Results are reported in Table II. We note that this dataset only evaluates mesh in neutral pose, so expression accuracy is not evaluated. Nevertheless, our method achieves competitive results, better than Dib et al. [13] and on par with Feng et al. [16].

We note that for both datasets, for our method, the mesh used for comparison is the base 3DMM geometry estimated by the "coarse" stage as our method only estimates a normal map to model the finer details.

Table II

| Name          | Mean±Std | < 20° | < 25° | < 30° |
|---------------|----------|------|------|------|
| Ours          | 22.9±15.3| 51.6%| 67.0%| 77.1%|
| Feng et al. [16]| 24.2±14.5| 43.8%| 59.7%| 72.4%|
| Chen et al. [8]| 26.2±15.6| 39.4%| 55.8%| 68.5%|
| Ours          | 16.8±9.9 | 71.2%| 84.7%| 92.8%|
| Feng et al. [16]| 15.2±12.2| 77.2%| 84.5%| 89.4%|
| Chen et al. [8]| 17.0±13.9| 72.4%| 82.7%| 88.2%|

Conclusion

In this work, we push to a new level the principle of analysis-by-rendering within self-supervised learning framework by refining the modelling, the training as well as the rendering stages. First, by combining, for the first time, ray tracing with the vertex-based renderer at training time, to solve the problem of edge-discontinuities of ray tracing which significantly improves the overall geometry and suggests an improvement on the original ray tracing algorithm. Second, by introducing a coarse-to-fine deep architecture, with adapted training strategy, that solve, for the first time, the highly challenging problem of separating detailed face attributes (reflectance and geometry) from a single image, within a self-supervised setup, and under uncontrolled lighting conditions. Compared to recent state-of-the-art, our method achieves superior reconstruction quality and produces more visually appealing results and define a new baseline for self-supervised monocular face reconstruction methods "in-the-wild".

REFERENCES

[1] V. F. Abrevaya, A. Boukhayma, P. H. Torr, and E. Boyer. Cross-modal deep face normals with deactivable skip connections. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 4979–4989, 2020.
