Cas-FNE: Cascaded Face Normal Estimation

Meng Wang, Jiawan Zhang, *Senior Member, IEEE*, Jiayi Ma, *Senior Member, IEEE*, and Xiaojie Guo, *Senior Member, IEEE*

***Abstract*—Capturing high-fidelity normals from single face images plays a core role in numerous computer vision and graphics applications. Though significant progress has been made in recent years, how to effectively and efficiently explore normal priors remains challenging. Most existing approaches depend on the development of intricate network architectures and complex calculations for in-the-wild face images. To overcome the above issue, we propose a simple yet effective cascaded neural network, called Cas-FNE, which progressively boosts the quality of predicted normals with marginal model parameters and computational cost. Meanwhile, it can mitigate the imbalance issue between training data and real-world face images due to the progressive refinement mechanism, and thus boost the generalization ability of the model. Specifically, in the training phase, our model relies solely on a small amount of labeled data. The earlier prediction serves as guidance for following refinement. In addition, our shared-parameter cascaded block employs a recurrent mechanism, allowing it to be applied multiple times for optimization without increasing network parameters. Quantitative and qualitative evaluations on benchmark datasets are conducted to show that our Cas-FNE can faithfully maintain facial details and reveal its superiority over state-of-the-art methods. The code is available at https://github.com/AutoHDR/CasFNE.git.**

***Index Terms*—Cascaded learning, face normal, progressive refinement, shared-parameter.**

I. Introduction

RECONSTRUCTING 3D surfaces is one of the most fundamental and important tasks in computer vision and graphics. Recently, deep learning-based methods [1]−[4] have significantly enhanced the quality of reconstructing 3D facial geometries from images captured under diverse conditions, achieving an immersive experience in VR/AR demands. However, most of these methods have trouble in capturing fine details such as beards and wrinkles on the face. Compared to 3D geometry information, face normals [5]−[10] also encode 3D surface details, therefore being more informative. In this paper, we propose a method to recover high-fidelity face normal from single images captured in unconstrained environments.

Unlike general normal estimation, which requires a full interpretation of all items within a scene [11]−[13], face normal estimation focuses on high-quality reconstruction of facial geometric details, such as subtle wrinkles and contours around the eyes, nose and mouth. As a fundamental problem in computer vision and graphics, face normal estimation has been extensively studied with numerous applications, like face editing [14], [15], face relighting [8], [16], [17], and face animation [18], [19], to name just a few. However, predicting face normals from single (in-the-wild) images poses a significant challenge, primarily due to the scarcity of a substantial amount of labeled data. Collecting such data is expensive in practice, while on the other hand, the models often suffer from unsatisfactory generalization ability to in-the-wild cases because to the uneven distribution of data between training and testing. Therefore, it is essential to develop effective techniques that improve the performance of face normal estimation and improve their generalization capability using limited labeled data.

To address the need for labeled data, Trigeorgis and Snape [20] trained a model on a synthetic dataset. However, this approach does not consider the distribution gap between the training data and real-world data. To depress the issue of distribution inconsistency, Sengupta *et al.* [21] trained a model using synthetic data to generate pseudo-labels for real face images. Then, the synthetic and pseudo-labeled pairs are fed into the network to address the problem of inconsistent data distribution. But, in this manner, the ability to capture high-frequency details may be limited in order to encompass the full range of variations in real-world scenarios. For instance, facial expressions, poses, and complex backgrounds are often not fully represented. As a result, the performance of the model trained on synthetic data is highly likely to degrade when applied to images in real-world scenarios. Recently, in order to enhance performance on natural/in-the-wild scenes, Abrevaya *et al.* [10] adopted a combination of labeled and unlabeled in-the-wild images and approached the task as an image translation problem. However, accurately extracting facial structural features and normal features from both labeled and in-the-wild images necessitates the careful design of a sophisticated network architecture. In addition, their method produces normal maps that lack high-frequency details.

Despite the progress, existing methods still face model degradation due to the lack of a large number of high-preci-

Manuscript received June 13, 2024; revised August 7, 2024; accepted August 31, 2024. This work was supported by the National Natural Science Foundation of China (62072327). Recommended by Associate Editor Wenguan Wang. *(Corresponding authors: Jiayi Ma and Xiaojie Guo.)*

Citation: M. Wang, J. Zhang, J. Ma, and X. Guo, "Cas-FNE: Cascaded face normal estimation," *IEEE/CAA J. Autom. Sinica*, vol. 11, no. 12, pp. 2423–2434, Dec. 2024.

M. Wang is with the Tianjin Key Laboratory of Autonomous Intelligence Technology and Systems, Tianjin 300350, and also with the School of Computer Science and Technology, Tiangong University, Tianjin 300387, China (e-mail: mengwang@tiangong.edu.cn).

J. Zhang and X. Guo are with the College of Intelligence and Computing, Tianjin University, Tianjin 300365, China (e-mail: jwzhang@tju.edu.cn; xiaojie.guo@tju.edu.cn).

J. Ma is with the Electronic Information School, Wuhan University, Wuhan 430072, China (e-mail: jiayima@whu.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JAS.2024.124899

sion face normal datasets for recovering high-quality, fine-grained face normals. This predicament arises from the lack of comprehensive, high-precision face normal datasets. Although Wang *et al.* [9] introduced a coarse-to-fine approach that employs exemplar-based learning for generating high-quality normals with a small amount of labeled data, it is important to note that the normal priors used during training are derived from secondary sources. These prior beliefs may introduce noise originating from the initial stage, resulting in potential inaccuracies in the second learning stage. Such discrepancies may lead to the accumulation of errors and, as a result, hinder the overall performance of the model. This could potentially compromise the precision and reliability of the final predictions.

In this paper, we present a cascaded method to recover face normal with "high-fidelity", which indicates preserving facial detail and accuracy with little artifact. The incorporation of chunking-based human learning theory [22], which has inspired a cascaded decoupled encoding and decoding strategy for face normal estimation, aims to repeatedly refine the estimating process. This approach allows for iterative refinement, effectively reducing the potential for error. Simultaneously, our cascaded model offers the opportunity to reduce the heavy reliance on large training data through iterative optimization techniques. The proposed cascaded face normal estimation network (Cas-FNE) systematically enhances the accuracy of normal prediction stage-by-stage without increasing the number of model parameters. Please see Fig. 1 for examples.

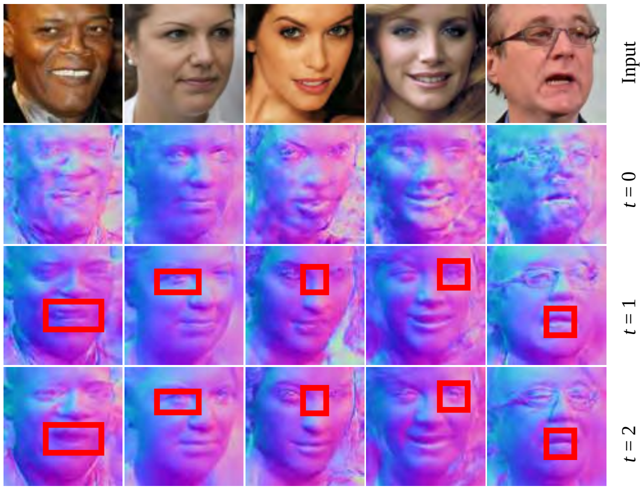

Fig. 1. Predicted normals at different cascaded stages (t) for samples from the CelebA dataset [23].

Our approach differs from previous methods in several ways. Firstly, we use a shared-parameter cascaded block that can be iteratively refined on specific features, leading to more rational training than single-stage architectures used in prior works. Secondly, our method employs a progressive refinement pipeline that introduces face structure features into normal branches at each stage. The learned normal features from previous stages are embedded and fed to subsequent stages for further refinement. Incorporating previous knowledge into model training makes it easier to extract robust features, which improves the model's refining capacity. The robust features improve the model's generalization capabilities while reducing model degradation caused by disparities in the distribution of training and testing data. Third, unlike previous methods, "CM" [10] with elaborate network designs and "HF" [9] with substantial computational overhead, our approach is distinguished by simplicity, low computational costs, and can be trained in an end-to-end manner. The major contributions of this work can be summarized as follows.

1) Inspired by chunking-based human learning theory, we propose a face normal estimation approach with cascaded decoupled encoding and decoding manner that can iteratively refine face normals, thereby enhancing the overall representation and generalization ability on in-the-wild face images.

2) Analogous to human step-wise cognitive learning, we design a shared-parameter cascaded block to efficiently and effectively learn features/patterns from different priors, thereby mitigating the reliance of large-scale data.

3) Our model demonstrates superiority compared to other state-of-the-art methods in different metrics, including accuracy, generalization ability, and computational efficiency, particularly when dealing with images containing intricate facial details.

II. Related Work

1) Data-Driven Based Normal Estimation: Benefiting from the success of CNNs, deep learning-based methods [1], [2], [9]−[11], [24] have revolutionized the field of normal estimation by providing more flexible and adaptable solutions that are less dependent on handcrafted priors. For instance, Shu *et al.* [14] develop an end-to-end network that can deduce a face-specific disentangled representation of intrinsic face components, such as normal. However, the smooth constraint of 3DMM normal utilized in Shu *et al.* [14] lead to results that are deficient in high-frequency details. To recover high-frequency normals, Trigeorgis *et al.* [20], Sengupta *et al.* [21] have started generating a large amount of synthetic data for model training. The trained models can capture the complex variations in real-world images, leading to improved performance in normal recovering from a single image. However, the performance of these models may be suboptimal when applied to real-world scenarios because of the distribution discrepancy between the synthetic and real data.

Recently, Abrevaya *et al.* [10] designed a network that transfer facial features between the image and normal domains to produce face normals. However, their architecture requires an elaborate design to achieve better results. Wang *et al.* [9], inspired by exemplar-based learning, considered normal estimation as a two-stage problem to generate fine-grained normals. Nevertheless, this two-stage training approach may not efficiently capture the underlying data distribution or extract relevant features for prediction. In contrast, our proposed cascaded model incorporates a strategic utilization of knowledge accumulated from previous stages in an end-to-end learning framework, which can facilitate the gradual refinement of face normal prediction.

2) Shape-From-Shading: Shape-from-shading (SFS) [25] is a technique that aims to recover the 3D shape from a gradual variation of shading in the image. To solve this ill-posed problem, many methods [26]–[30] assume images captured under a Lambertian model to solve 3D shape. For example, Wang *et al.* [26] propose a 3D spherical harmonic morphable model that can incorporate an integration of spherical harmonics into the morphable model framework to estimate lighting, shape and albedo. Xiong *et al.* [30] recover the local shape of a surface from its shading pattern based on the idea that the local shape of a surface can be represented as a combination of basis functions. To solve ill-posed problem, all these methods make assumptions to simplify. However, it is important to note that these assumptions may not always hold in the unconstrained case, where the real-world conditions may vary widely.

Recent works [31]–[34] have made significant advancements in recovering facial shape, which effectively enhance the realism of synthesized faces. However, these methods, although focusing on improving the overall fidelity of face synthesis, predominantly overlook the subtle yet critical aspects of facial geometry. Moreover, [1], [3], [4] concentrates on restoring 3D facial avatars, wherein facial expression is more desired. In contrast, our approach simultaneously takes care of recovering facial shape and high-frequency details.

3) Cascaded Networks: Cascaded networks are a prominent strategy that has been widely applied to improve performance in various vision tasks, such as face recognition [35]–[37], face detection [38]–[40], face expression editing [41], face generation [42] and image semantic segmentation [43]–[47]. For example, Xue *et al.* [37] proposed a cascaded network for video facial expression recognition based on prior classified face emotions. Wu *et al.* [41] successfully combined the advantages of GANs and cascaded models to enhance the accuracy of facial expression editing. Chen *et al.* [42] developed a cascade talking face video generating strategy, which used facial landmarks as intermediary high-level representations to bridge the gap between two different modalities. Ding *et al.* [47] introduced a cascaded network for instance segmentation, which leveraged previous shape features to iteratively enhance the bounding box detection in the current stage. However, their architectures are not weight-sharing, leading to an expansion in the model's parameter space and a redundancy in feature representation. In contrast, we propose a shared-parameters cascaded face normal estimation network, learning normal prior from previous prediction to enhance subsequent refinement. Furthermore, the cascaded manner contributes to improve performance while also achieving a significant reduction in model parameters, without compromising performance.

III. Method

A. Problem Analysis

Face normal estimation aims to predict the face normal N from the input face image I. In a deep-learning manner, the problem aims to construct a network $\mathcal{N}_\theta$ that can effectively generate $N = \mathcal{N}_\Theta(I)$. From the perspective of maximizing a posterior (MAP), the training objective can be generally written as follows:

$$\min_{\Theta} \Psi(\mathcal{N}_\Theta(I), N_{gt}) + \Phi(\mathcal{N}_\Theta(I)) \tag{1}$$

where N_{gt} represents the ground-truth normal, $\Phi(N)$ stands for the regularizer on the estimation, and $\Psi(\mathcal{N}_\Theta(I), N_{gt})$ designates the fidelity measurement.

Typically, optimizing (1) is a complex and non-convex task. An intuitive approach to address this issue is to construct a relatively deep network. Inspired by the framework of majorization minimization (MM) [48], [49], an effective technique for solving complex optimization problems, we can cope with the target problem in a similar fashion. The fundamental idea is to successively minimize an easy-to-tackle surrogate associated with the current estimate. Consequently, given an estimate at the t-th stage $\hat{N}_t$, the problem of (1) can be modified as

$$\min_{\theta} \sum_{t=0}^{T} \Psi(\mathcal{N}_\theta(I, \hat{N}_t), N_{gt}) + \Phi(\mathcal{N}_\theta(I, \hat{N}_t)) \tag{2}$$

where T is a pre-defined total stage/step number. Moreover, $\mathcal{N}_\theta$ shall be a (much) smaller network compared to $\mathcal{N}_\Theta$, performing like the surrogate in MM. Please notice that $\mathcal{N}_\theta$ takes the original image together with the estimation from the previous stage as input, and is expected to dynamically adjust according to different intermediate estimates $\hat{N}_t$. The above analysis motivates us to construct a cascaded network to accomplish the task. Each step is in nature conceptualized as a sub-module of the network. The subsequent section will detail our proposed network.

B. Network Architecture

Fig. 2 illustrates the overall framework of our cascaded network structure for face normal estimation. The model consists of multiple cascaded blocks, each refining the estimated normals based on the early predictions. Specifically, the early prediction is encoded into normal feature, and the feature is combined with the structure features in a cascaded manner to achieve progressive refinement of the normal estimations. This iterative process allows the model to enhance the accuracy and capture fine-grained details of the face normals at each stage. As a result, our model produces more accurate and realistic estimations of the face normals. The cascaded network structure enables the model to effectively learn from limited data and generalize well to unseen face images, making it a powerful and effective approach for face normal estimation.

1) Cascaded Block: Our *CasNet* consists of cascaded blocks that can be used to the refinement at different stages. To validate the primary claim, we choose a U-shaped network as the base due to its simplicity and efficiency. In Fig. 2, a shared-parameter cascaded block houses a pair of interconnected subnetworks, denoted as E_{cn} and D_{cn}, respectively. These subnetworks share parameters, but perform distinct yet synergistic functions in the process of estimating normals. Specifically, E_{cn} assumes the role of encoding normal features, while D_{cn} is responsible for the important task of refining the normal estimation. The normal features z extracted from the previously predicted normal by E_{cn} are seamlessly integrated

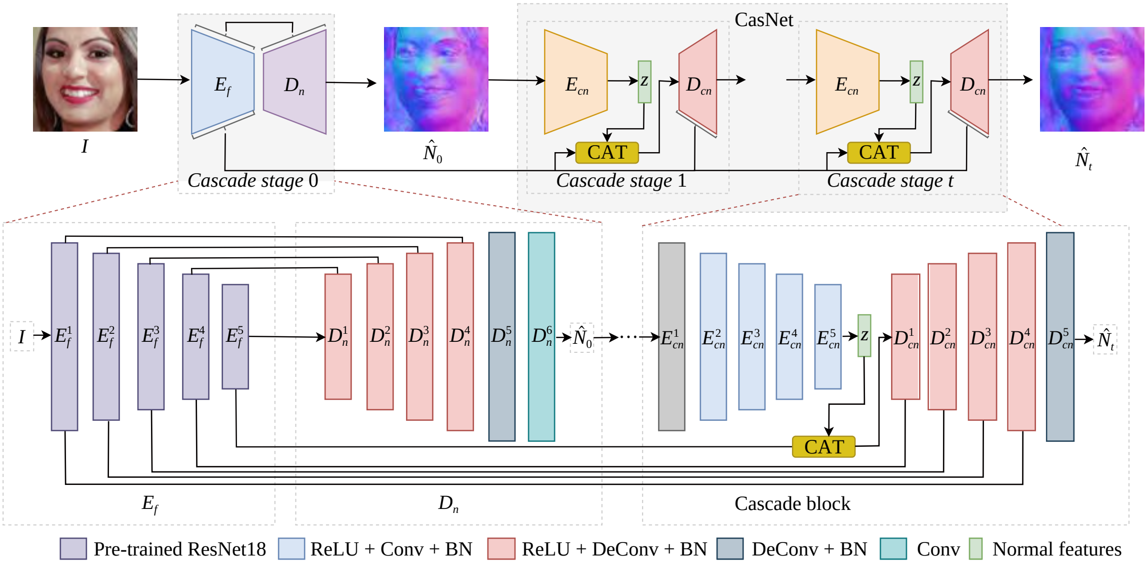

Fig. 2. The framework of our Cas-FNE can produce face normals through the refinement of multiple cascaded stages. The initial predictions are encoded using normal features and then combined with the structure features to further normal refinement.

with the structural features acquired through encoding by the network component E_f. The fusion features are then sent to D_{cn} for further refinement. The entire process can be represented as follows:

$$\hat{N}_t = D_{cn}^{j+1}(\mathbf{CAT}[D_{cn}^{j}(\mathbf{CAT}(z, E_f^5)), E_f^{5-j}]) \tag{3}$$

where t represents the number of cascaded refinement, z signifies the normal features encoded by E_{cn}, E_f^{5-j} denotes the facial structure extracted at the $(5-j)$-th layer, with j ranging from 1 to 4. Meanwhile, D_{cn}^j corresponds to the j-th layer of the cascaded normal decoder. The normal features extracted by E_{cn} in each stage provide valuable prior knowledge for predicting D_{cn} in the subsequent stage. This iterative refinement not only guides the model towards improved results but also streamlines the model training, leading to enhanced model performance.

2) Initial Block: Instead of optimizing the problem from a random starting point, we aim to establish a robust initialization for the cascaded network. For this purpose, a suitable foundation for subsequent refinement is obtained via an initial block (cascaded stage $t=0$). In this work, we utilize a pre-trained ResNet18 [50] on the ImageNet [51] as the initialization. This operation ensures a favorable warm start for the procedure and accelerates model convergence. Specifically, in our approach, the initial block consists of two sub-networks, as shown in Fig. 2: E_f aims to learn the face structure features, while D_n aims to learn normal priors based on these structure features. E_f utilizes the encoder from the pre-trained ResNet18 and divides it into five multi-scale feature blocks E_f^i (i ranges from 1 to 5). Additionally, D_n is created using five deconvolution layers followed by a convolution layer. The multi-scale structure features E_f^i are propagated to D_n through a skip-connection. The initial normal $\hat{N}_0$ for further refinement as follows:

$$\hat{N}_0 = \mathbf{Con}(D_n^{i+1}(\mathbf{CAT}[D_n^i(E_f^5), E_f^{5-i}])) \tag{4}$$

where E_f^5 and E_f^{5-i} are the features extracted by E_f and D_n at layer 5 and $(5-i)$. D_n^{i+1} and D_n^i are the layers of the normal decoder D_n at $i+1$ and i, i starts from 1 to 5. **CAT** represents the concatenation operation. **Con** represents as a 1×1 convolutional layer to combine these recovered features using a nonlinear mapping, which helps capture higher-level features and improves the generalization.

C. Loss Function

1) Reconstruction Loss (Ψ): Similarly to other face normal estimation tasks, we propose the utilization of a reconstruction loss function, which is defined as follows:

$$\mathcal{L}_{\text{rec}} = \sum_{0}^{t} CosLoss\left(\hat{N}_t, N_{gt}\right) \tag{5}$$

where N_{gt} is a unit vector, representing the ground truth normal, $\hat{N}_t$ is a unit vector regarded as predicted normal. t represents the number of cascaded refinement utilized in our model, and t $(0 \le t)$ is an integer that indicates the current stage of normal refinement. The *CosLoss* function calculates the cosine similarity loss between the predicted $\hat{N}_t$ and the ground truth N_{gt}.

2) Total-Variation (TV) Loss (Φ): To enhance the predicted normal quality, we utilize TV loss [52] to eliminate noise and artifacts, while retaining crucial details and structures. The TV loss is defined as

$$\mathcal{L}_{\text{tv}} = \frac{1}{\mathcal{M}} \sum_{0}^{t} \sum_{x,y} \left\|\nabla(\hat{N}_t(x,y)/2+0.5)\right\| \tag{6}$$

where $\mathcal{M}$ denotes the pixels on the I and $(\hat{N}_t(x,y)/2+0.5)$ is used to convert the normal range from −1 to 1 to the range of 0 to 1, ∇ represents the gradient of the predicted normal at each pixel (x,y), and $\|\cdot\|_1$ denotes the gradient magnitude.

3) Adversarial Loss (Φ): To improve the accuracy of predicted normals and align the distribution with the ground truth, we incorporate an adversarial loss [53] into our model.

The purpose of incorporating the adversarial loss is to improve the quality and realism of the predicted normal maps, resulting in more precise and visually appealing outcomes. The adversarial loss is represented as follows:

$$\mathcal{L}_{\text{adv}} = \sum_{0}^{t} \mathcal{D}_{\text{adv}}(\hat{N}_t) \tag{7}$$

where $\mathcal{D}_{\text{adv}}$ is the discriminator, $\hat{N}_t$ represents the predicted normal by the cascaded network, where $0 \leq t$.

In summary, the overall loss function is defined as

$$\mathcal{L}_{\text{total}} = \lambda_{\text{rec}}\mathcal{L}_{\text{rec}} + \lambda_{\text{adv}}\mathcal{L}_{\text{adv}} + \lambda_{\text{tv}}\mathcal{L}_{\text{tv}} \tag{8}$$

where λ_{rec}, λ_{tv} and λ_{adv} are the weight of the normal reconstruction term, perceptual term and adversarial term.

IV. Experiments

A. Datasets

The Photoface [54] is a widely utilized collection of facial images that captures four faces under different lighting conditions. This dataset can utilize the photometric stereo methodology to produce face normals, which are a valuable resource for face analysis, including face normal estimation [9], [10]. According to the setting [9], [10], [20], [21], we randomly split approximately 80% of the data for training and used the remaining data for testing.

In order to show that the proposed *Cas-FNE* can produce generic face normals, we show the results of *Cas-FNE* performed on five face datasets. There are 300-W [55], CelebA [23], FFHQ [56], Florence [57] and ICT-3DRFE [58]. 300-W is a face image dataset that contains 300 indoor images and 300 outdoor images captured in the wild. CelebA is a large-scale real-face dataset with diverse images covering various poses and background variations. FFHQ is a face image dataset that contains a wide range of ages and ethnicity. Florence is a face 3D-models dataset that produces face normals that can be used to evaluate the generalizability of *Cas-FNE* on unseen face images. Following [9], we also relight faces under new light conditions after predicting face normals on ICT-3DRFE to demonstrate the accuracy of our *Cas-FNE*.

B. Evaluations

Following [9], [10], we also employ the mean angular error between the predicted normals and the ground truth normals. Additionally, we consider the percentage of pixels within the facial region that exhibit angular errors less than 20°, 25°, and 30° as additional evaluation metrics for face normal estimation. For qualitative comparisons, we introduce geometric shading and normal error maps. These visualizations help to understand the quality of the estimated normals and provide insights into the performance of the model. Furthermore, to assess the realism of normal estimation on the ICT-3DRFE [58] dataset, we re-render the faces with predicted normals under new illuminations, enabling a comprehensive comparison of the results.

C. Implementation Details

We implemented our framework in PyTorch [59] with a learning rate of 10^{-4}, and use Adam [60] as our optimizer with default parameters. The *Cas-FNE* is trained about 500 K on a single 2080Ti GPU, using a batch size of 8. Following previous method [10], cropped face with a size of 256×256 are used for our training and testing. To balance training time and model computational complexity, we use two cascaded steps ($t = 2$) in our paper.

The hyperparameters $\lambda_{\text{rec}} = 1$, $\lambda_{\text{adv}} = 0.0001$ and $\lambda_{\text{tv}} = 0.002$ were determined through a systematic grid search approach over the parameter space. This method allows us to explore a range of parameter values to identify the optimal set that yields the possible best performance. The process evaluates different combinations of λ values to balance the reconstruction loss, adversarial loss, and total variation loss.

D. Comparison

We compare the predicted normals from different cascaded stages, as shown in Figs. 3 and 4. The results on out-of-training data distribution in Fig. 3 and the error map in Fig. 4 highlight the gradual refinement of the estimated normals of our method. Furthermore, the results on out-of-training data provides evidence of the strong performance of our method on data beyond the training set. The one-stage trained model's capacity for generalization is constrained by the distribution gap existing between the training data and real-world data. This limitation can result in artifacts or loss of high-frequency details at $t = 0$ (as shown in Fig. 3). Following several iterative refinements facilitated by our cascaded block, the *Cas-FNE* model demonstrates its capability to consistently generate high-fidelity face normals. More specifically, $t = 1$ can remove a significant number of artifacts with one cascaded

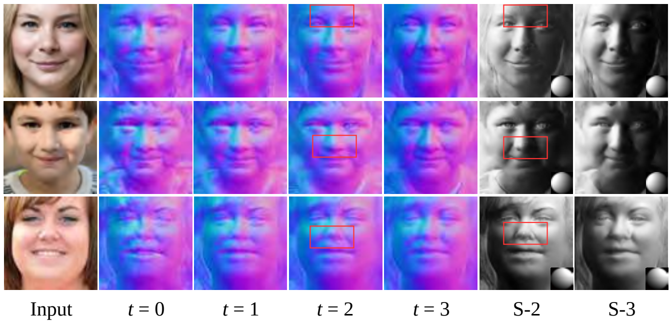

Fig. 3. Progressively refinement normal results and rendered shadings on previously unseen data from the FFHQ [56] at stage $t = 3$. From left to right, it is evident that the estimated outgoing normals show a progressive improvement. "S-t" represents the rendered shading at the current stage t.

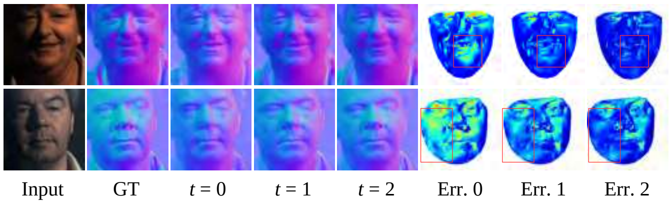

Fig. 4. Progressively refinement normal results and their error maps on the Photoface [54]. "GT" and "Err. t" represent the ground truth and error maps at the current cascaded stage t.

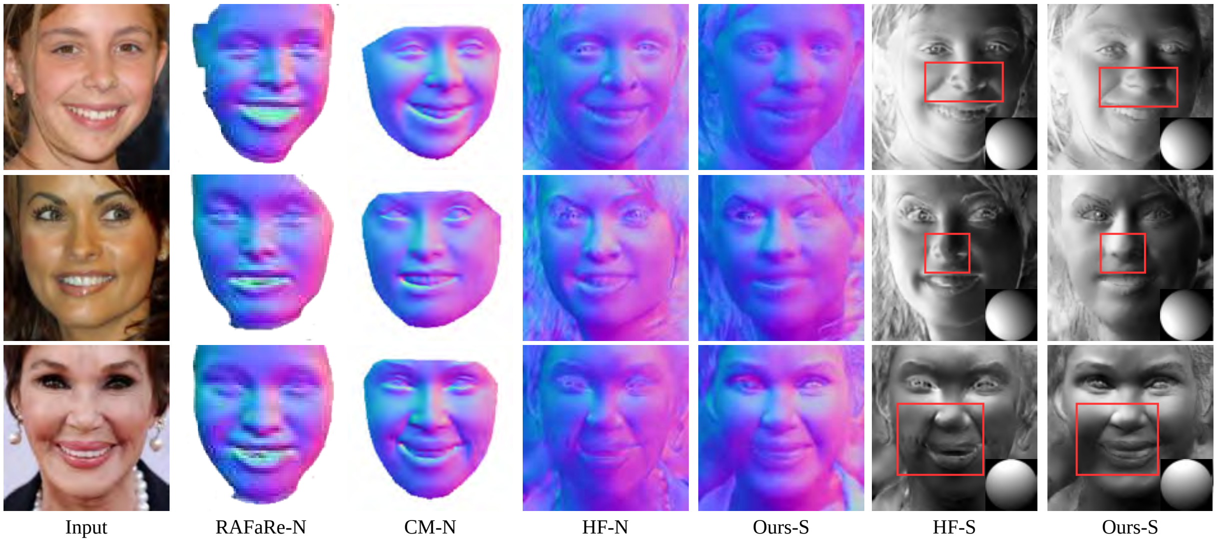

Fig. 5. Normals comparison with the state-of-the-art method “RAFaRe” [61], “CM” [10] and “HF” [9] on the CelebA dataset [23]. “-N” and “-S” are the predicted normal and rendered shading, respectively.

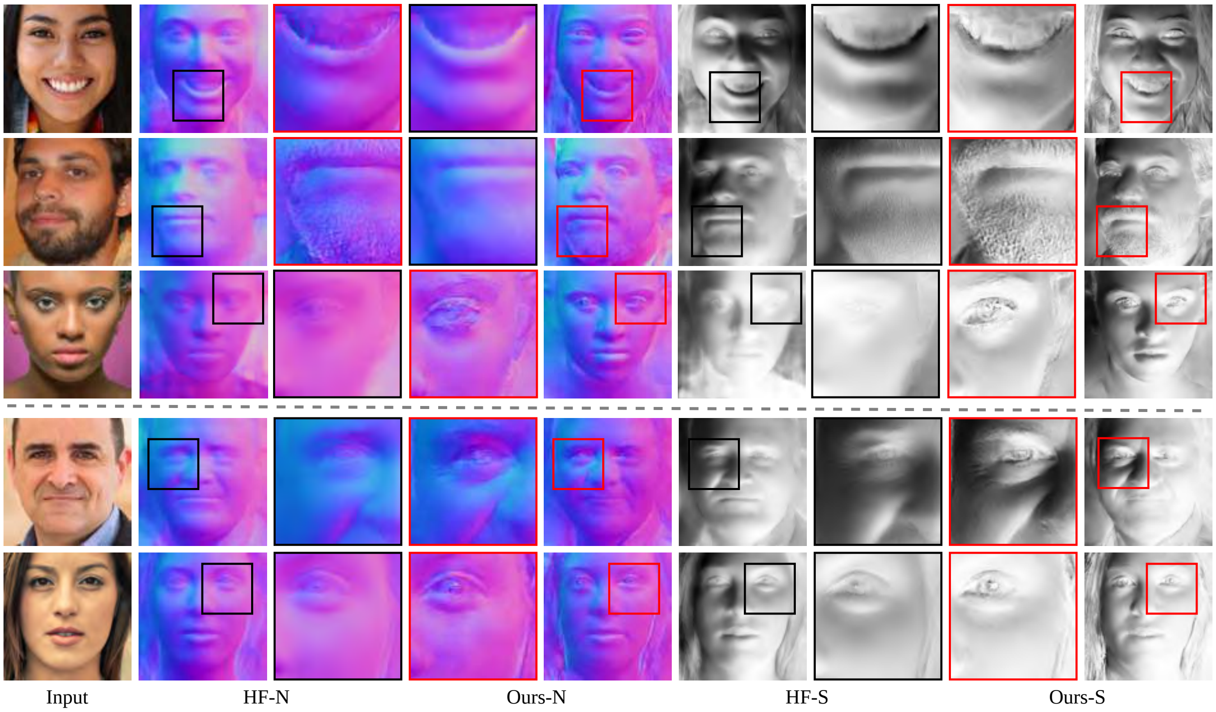

Fig. 6. Visual comparison with the state-of-the-art method “HF” on the FFHQ dataset. “-N” and “-S” are the predicted normal and rendered shading, respectively.

block; however, the results are not entirely satisfactory. By adding another cascaded block, it is possible to refine the estimation of normals and enhance their level of detail, ultimately leading to a desirable outcome, as demonstrated at $t = 2$. In Fig. 3, we present our performance on out-of-distribution data at $t = 3$. The artifacts are gradually eliminated as the number of cascaded block increases. This improvement is also evident from the last 5 rows of Table I.

Figs. 5 and 6 provide compelling evidence of the exceptional performance of our method on in-the-wild face images. As shown in Fig. 5, “RAFaRe-N” [61] demonstrates proficiency in face normal recovery; however, it suffers from a limitation in preserving details. “CM-N” [10] preserves certain facial details, but it has a generally smooth appearance and cannot generate accurate shading maps (“CM-S”) under specific lighting. While “HF-S” [9] and “Ours-S” can accurately depict the light direction. “HF-N” is capable of recovering certain geometric details, but it faces difficulties when dealing with significant changes in gradients, such as areas like beards and corners of the eyes. They still have room for improvement when it comes to recovering details. Our proposed method achieves superior performance by effectively capturing subtle facial details, as demonstrated by “Our-N” and “Ours-S”. This comparison aims to verify the generaliza-

TABLE I
COMPARISON IN NORMAL RECONSTRUCTION ERROR ON PHOTOFACE DATASET WITH DIFFERENT SETTINGS

Exp.	Mean ± std	< 20°	< 25°	< 30°
CARN	15.89±9.54	73.48%	84.98%	91.55%
CBE	14.63±8.56	79.89%	89.53%	94.48%
BNN	11.79±8.91	86.08%	93.81%	96.04%
Res	10.49±7.46	90.15%	95.32%	97.51%
Rec	9.23±7.67	92.68%	95.78%	97.25%
Rec+adv	8.91±7.33	92.99%	96.14%	97.87%
Rec+tv	8.70±7.12	93.48%	96.58%	98.02%
AdaIN	8.89±7.69	92.39%	95.66%	97.26%
FMM	8.63±6.82	92.12%	95.64%	97.22%
t = 0	10.21 ±7.75	90.03%	94.92 %	97.21 %
t = 1	8.02 ±6.17	95.31%	97.80 %	98.41 %
t = 2/Ours	**7.31±5.92**	**96.26%**	**98.17%**	**99.01%**
t = 3	7.28 ±5.87	96.29%	98.21 %	99.04 %
t = 4	7.25 ±5.76	96.31%	98.22%	99.05 %
t = 5	7.24 ±5.78	96.32%	98.25%	99.07 %

tion ability of our model when dealing with data that is different from the training dataset. From the figure, it is evident that both "HF" and our method outperform "CM" in recovering geometric details of the human face. The normal maps generated by "CM" exhibit incorrect lighting effects when rendered with specific lighting conditions, while the normal maps produced by "HF" and our method display more accurate lighting effects. Moreover, when compared to "HF", our method is able to effectively reduce artifacts, particularly around the nose area (as indicated by the red box in the figure). "HF" often exhibits artifacts in this region, while our proposed method successfully mitigates these artifacts, resulting in a more visually pleasing and accurate representation of the facial structure.

In Fig. 6, different lighting angles are provided to observe the detailed geometry of the face. Both our method and "HF" [9] are capable of recovering some details, as shown in the image below the dashed line. However, above the dashed line, the performance of "HF" appears to be degraded and does not perform well. On the other hand, our method can still recover fine-grained normals. In comparison to "HF", our method can recover normals and capture finer facial details, which significantly improves the quality on geometry shading.

In Fig. 7, we present a comparative analysis between our method and "HF" [9] regarding normals, shadings, and relit faces. The relit faces are presented under different lighting to better observe the details of the geometry. The shadings on the left of the dashed line indicate that both ours and "HF" can successfully restore fine details. However, when it comes to re-rendering faces under different light, "HF" tends to smooth out due to the loss of high-frequency details resulting in a less realistic. In contrast, our approach preserves high-frequency details more effectively and generates more realistic relit faces. On the right side of the dashed line, the presence of "HF" results in artifacts that ultimately impact the authenticity of relit faces under specific lighting conditions.

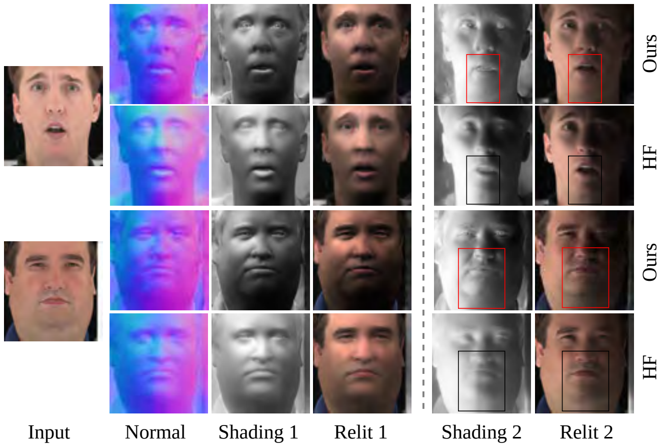

Fig. 7. Comparisons in normals, shadings and relit faces on the ICT-3DRFE (best viewed by zooming in).

In addition, Fig. 8 provides a qualitative comparison between our proposed method and "HF" [9], conducted on the Photoface dataset [54]. Upon a thorough examination of the recovered normals and error maps produced by both methods, a discernible contrast emerges. "HF" exhibits a propensity for generating smoother normals, primarily attributable to inaccuracies in its estimation process. Although "HF" can capture certain details, it falls short of achieving the desired level of accuracy in normal recovery. In contrast, our method demonstrates a significant advantage over "HF" in terms of accuracy, as corroborated by the normal error maps. These error maps clearly indicate that our approach excels in recovering highly detailed normals with superior accuracy, thereby yielding results that are both more realistic and visually appealing. This constitutes compelling evidence of the discernible advancement achieved by our methods beyond those of "HF".

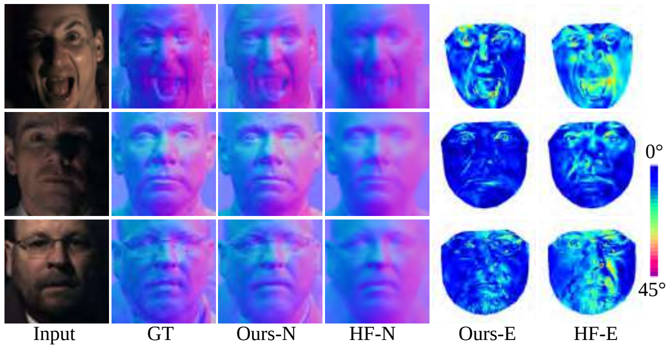

Fig. 8. Normal and error maps comparisons on the Photoface dataset [54]. "GT", "-N", and "-E" are ground truths, predicted normals, error maps, respectively.

In Table II, we compared the quantitative results of our refinement normals with those obtained by other methods, including "RAFaRe" [61], "HF" [9], "LAP" [62], "CM" [10], "PRN" [63], "SfSNet" [21], "Marr Rev" [64], "NiW" [20] and "UberNet" [65], "3DDFA" [66]. We evaluate these methods based on mean angular errors (in degrees) and percentage of errors below $< 20^\circ$, $< 25^\circ$ and $< 30^\circ$. The methods located

TABLE II
NORMAL RECONSTRUCTION ERROR COMPARISONS ON THE PHOTOFACE DATASET [54]

Method	Mean ± std	< 20°	< 25°	< 30°
SfSNet	25.5±9.3	43.6%	57.5%	68.7%
PRN	24.8±6.8	43.1%	57.4%	69.4%
CM	22.8±6.5	49.0%	62.9%	74.1%
RAFaRe	24.3±8.4	45.6%	58.2%	71.2%
UberNet	29.1±11.5	30.8%	36.5%	55.2%
NiW	22.0±6.3	36.6%	59.8%	79.6%
Marr Rev	28.3±10.1	31.8%	36.5%	44.4%
SfSNet-ft	12.8±5.4	83.7%	90.8%	94.5%
CM-ft	12.0±5.3	85.2%	92.0%	95.6%
LAP	12.3±4.5	84.9%	92.4%	96.3%
HF	11.3±7.7	88.6 %	94.4 %	97.2 %
Ours	**7.31±5.92**	**96.26%**	**98.17%**	**99.01%**

above the horizontal line indicate that the model was not trained using the data situated in that region. The others were trained on the Photoface [54]. "SfSNet-ft" and "CM-ft" were fine-tuned on the Photoface. The lower Mean ± std error is better, while a higher is better for the percentage of correct pixels at various thresholds. In contrast to the state-of-the-art methods, our proposed method exhibits superior performance. "RAFaRe" incorporates pseudo-training data into the training; however, the weak generalization ability results in a decline in performance when faced with novel data. This comparison underscores that our model has strong generalization ability.

To evaluate the performance in the context of downstream task called normal enhance geometric, which ensures that the recovered 3D face has high fidelity details. We compare the mesh results presented in Fig. 9. Specifically, we examined the normal mapping over the same base mesh, which is obtained using "PRN" [63] for all of our method "Ours+PRN", "HF" [9] "HF+PRN", and "CM" [10] "CM+PRN". Our approach demonstrates the ability to recover significantly more refined details and enhance the base mesh effectively, as highlighted in the red rectangular box in the figure. In comparison to normal enhancement meshes, our method offers a substantial improvement in detail recovery. Furthermore, we compare our method against geometric approaches such as "HRN" [2], "SMIRK" [1], "EMOCA" [3], and "DECA" [4]. Unlike these methods, our approach does not introduce unnecessary additional noise. As observed by other authors, "HRN" excels at recovering intricate details in regions with significant gradient variations but tends to introduce noise in smoother regions of the face, compromising the overall visual quality. Conversely, "HF+PRN" maintains noise-free smooth regions, thus preserving the visual integrity of these areas. However, it introduces noise in regions with large gradient variations, negatively impacting the accuracy and quality of detail recovery in these more complex areas. The downstream task of normal improvement geometry demonstrates the capacity of our approach to recover details while reducing noise, resulting in high-quality mesh enhancement.

Following previous works [9], [10], we also generate face normals from Florence [57] to evaluate the performance on the out-of-distribution face images. The results are presented in Table III, indicating that our proposed method achieves superior values for both mean angular error and percentage under $< 20^\circ$, $< 25^\circ$, and $< 30^\circ$ degrees. These findings underscore the effectiveness of our approach in handling out-of-distribution data, which is critical for real-world applications where input images vary widely from the training data.

We conducted a comparative analysis of computational efficiency between our method and two existing approaches, namely, "CM" [10] and "HF" [9]. The results, as summarized in Table IV, unequivocally highlight the advantages of our approach. Specifically, "Ours" exhibits superior performance in terms of model parameters and FLOPs, signifying its efficiency. Furthermore, "Ours" outperforms both "HF" and "CM" in terms of accuracy, as evident from Tables II−IV. It is noteworthy that "Ours" not only reduces model parameters but also significantly simplifies the model while concurrently enhancing the accuracy of normal estimation, thus affirming its efficiency and effectiveness.

V. ABLATION STUDY

1) Cascaded Block Number: We conducted performance testing Cas-FNE with varying values of t ranging from 1 to 5, respectively. In Figs. 1 and 4, Tables I and IV, the performance of our model consistently improves with the addition of cascade blocks. In addition, we show the results with $t = 2$, $t = 3$, $t = 4$, and $t = 5$ cascaded blocks in Figs. 10−13 and Table I, respectively. These figures and the table illustrate the effectiveness of our cascaded network structure in handling out-of-distribution data and show how the performance improves as the number of cascaded refinement increases. However, the improvement becomes less significant when more than $t = 3$ cascade blocks are added, possibly due to limitations in the parameters. To strike a balance between training time and performance, we employ $t = 2$ in our paper. This configuration offers a good balance between model performance and computational efficiency on unseen data.

2) Architecture: We conducted comparative evaluations with several baseline methods on the CelebA [23], including the cascaded model "CARN" [67], "CBE" [68], "BNN" [69], and transformer-based "Restormer" [70], as shown in Fig. 14 and Table I. The model utilizing these networks demonstrates satisfactory performance on the test dataset, as shown in the table. However, model degradation occurs when applied to data that is not included in the training set, as shown in Fig. 14. While "Restormer" can recover normals that appear to be true, their normal directions are not entirely accurate, as shown in the rendered shadings ("Res-S" and "Ours-S"). In contrast, our results accurately depict the lighting effect under directional light. To ensure efficient feature utilization and avoid redundancy in the cascaded architecture, we directly propagate the features extracted in the previous stage to the subsequent stages. This strategy not only reduces the number of model parameters but also preserves model accuracy.

3) Loss Term: In Fig. 15 and the first two rows of Table I, we show the quantitative and qualitative results of our network trained only with reconstruction loss ("Rec"), with

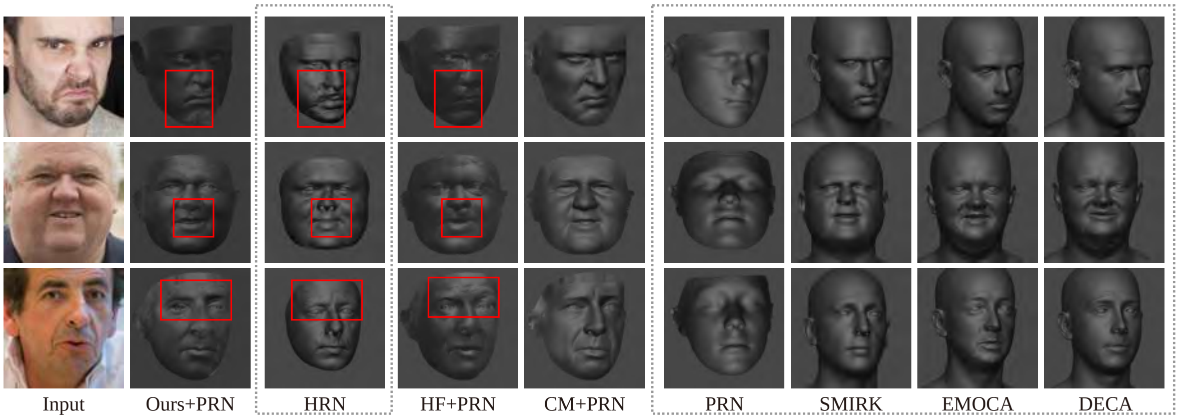

Fig. 9. Geometric comparisons as a downstream task of normal estimation with the state-of-the-art methods in the FFHQ dataset [56]. Where "Ours+PRN", "HF+PRN", and "CM+PRN" are the results of normal enhanced geometry based on "PRN", and the results of the dashed box produced by facial geometry reconstruction methods, "HRN" [2], "PRN" [63], "SMIRK" [1], "EMOCA" [3] and "DECA" [4]. Given the scarcity of face normal estimation, we compare our normal enhanced geometric to 3D face methods. The enhanced geometric shows that our normal can preserve high-fidelity details.

TABLE III
NORMAL RECONSTRUCTION ERROR COMPARISONS ON THE FLORENCE DATASET [57]

Method	Mean ± std	< 20°	< 25°	< 30°
SfSNet	18.7±3.2	63.1%	77.2%	86.7%
3DDFA	14.3±2.3	79.7%	87.3%	91.8%
PRN	14.1±2.2	79.9%	88.2%	92.9%
CM	11.3±1.5	89.3%	94.6%	96.9%
HF	10.1±3.4	92.3%	95.6%	97.8%
RAFaRe	14.0±2.1	80.1%	88.5%	93.1%
Ours	**7.24±2.1**	**96.54%**	**98.23%**	**99.1%**

TABLE IV
COMPARISON IN FLOPS, PARAMETERS AND RUNTIME ON DIFFERENT CASCADED STAGES TESTED ON 256×256 IMAGE SIZE

Exp.	FLOPs	Parameters	Runtime
CM	49 G	35.2 M	33 ms
HF	234 G	126.4 M	79 ms
AdaIN	25.8 G	29.3 M	12 ms
FMM	35.8 G	36.8 M	14 ms
$t = 1$	14.2 G	22.9 M	11 ms
$t = 2$/Ours/CAT	20.2 G	29.3 M	12 ms
$t = 3$	26.2 G	29.3 M	12 ms
$t = 4$	32.2 G	29.3 M	13 ms
$t = 5$	38.2 G	29.3 M	13 ms

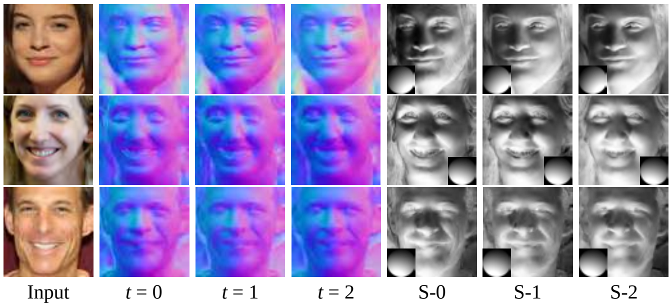

Fig. 10. Progressively refinement normal results and rendered shadings on previously unseen data from the FFHQ [56] at stage $t = 2$. The "S-t" denotes the rendered shading at the current stage t, providing a more favorable perspective for discerning intricate details by analyzing changes in shading (please zoom in for details).

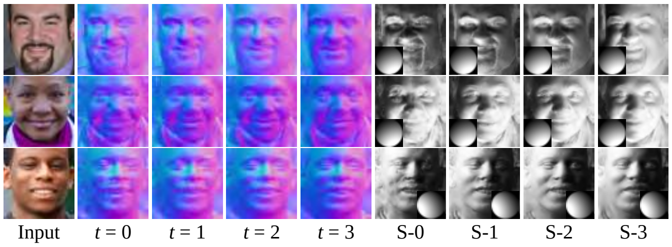

Fig. 11. Progressively refinement normals and rendered shadings on unseen data from the FFHQ [56] at stage $t = 3$. The shadings proves advantageous for discerning intricate details.

reconstruction loss and adversarial loss ("Rec+adv"), with reconstruction loss and TV loss ("Rec+tv") and our full model. Table I shows that "Rec" performs well on the test dataset but performs poorly on the out-of-distribution data. On the other hand, "Rec+adv", the quantitative performance on the test dataset is not as good as "Rec"; nevertheless, the performance is better on the out-of-distinct data. However, when applied to out-of-distribution data, the "Rec+adv" shows certain artifacts, as shown in Fig. 15. While "Rec+tv" effectively mitigates the presence of artifacts, it results in a concomitant reduction of high-frequency details within the reconstructed normals. In contrast, "Ours", while scoring lower quantitatively than the others, successfully reduces artifacts and improves the visual quality of normal on out-of-distribution data, resulting in better model generalization. Our shadings show that our model effectively recovers high-frequency and retains details.

4) Feature Fusion Strategy: We compare the feature fusion strategy in Fig. 16, Tables I and IV. In Fig. 16, the feature modulation module [71] ("FFM") produces artifacts that may negatively affect both the accuracy and visual quality of out-

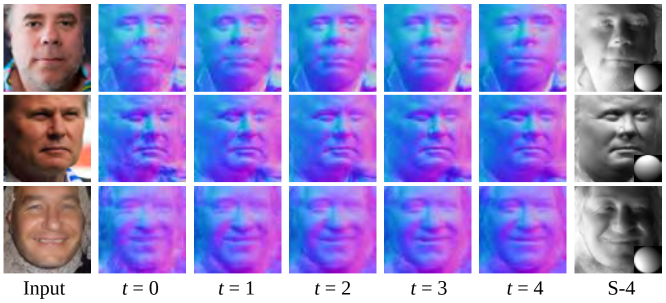

Fig. 12. Progressively refinement normal results and rendered shadings on previously unseen data from the FFHQ [56] at stage $t = 4$. It is evident that the details in the face normal maps improve progressively.

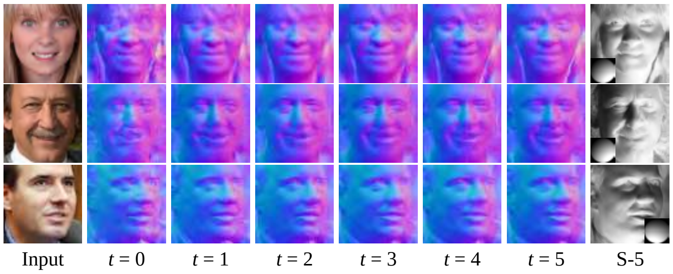

Fig. 13. Progressively refinement normal results and rendered shadings on previously unseen data from the FFHQ [56] at stage $t = 5$.

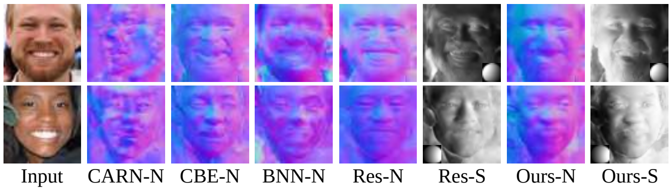

Fig. 14. Comparative evaluation of model performance with different network structures on the out-distribution data.

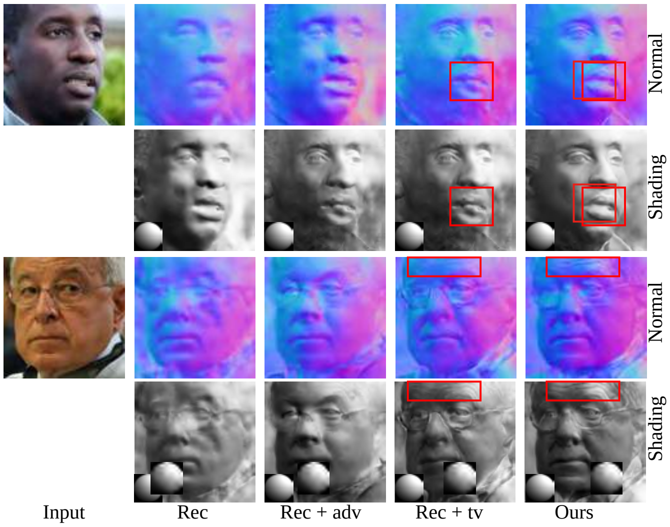

Fig. 15. Normal and shading comparative evaluation of model performance with different loss terms.

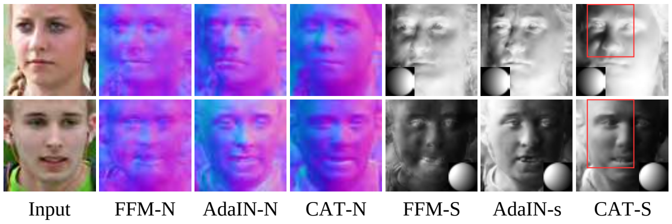

Fig. 16. Normal and shading comparative evaluation of model performance with different feature fusion strategies.

put. The AdaIN [72] appears to generate a more realistic normal. When the shading is rendered with different lighting, artifacts become visible in the shading ("AdaIN-S"). This may emerge as a consequence of the imperative to modify features simultaneously, thus impacting their individual performance.

VI. Concluding Remarks

In this paper, we have introduced a novel framework for face normal estimation called Cas-FNE. Our approach is based on an encoder-decoder structure with a shared-parameter cascaded block that can progressively refine the predicted face normal with fewer parameters. Specifically, Cas-FNE converts the previous prediction into normal prior, and then the prior is leveraged to generate outcomes of heightened accuracy by means of an iterative refinement process facilitated by the cascade block. Importantly, our method does not require increasing the network depth or parameter amount to improve the performance compared to existing methods. Extensive experiments have demonstrated that our approach outperforms state-of-the-art methods. However, like most existing methods, Cas-FNE has limitations when dealing with faces under extreme lighting conditions (a), occlusion (b), and low-quality images (c) as shown in Fig. 17. Therefore, future work should focus on developing advanced versions of our framework that can effectively address these challenging cases.

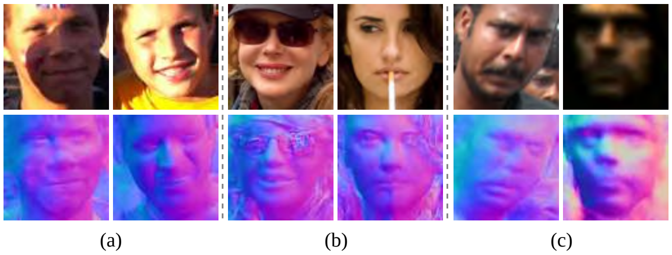

Fig. 17. Failure cases with extreme lighting: (a) Occluded; (b) Low-quality; (c) Faces on the 300-W [55].

References

[1] G. Retsinas, P. P. Filntisis, R. Danecek, V. F. Abrevaya, A. Roussos, T. Bolkart, and P. Maragos, "3D facial expressions through analysis-by-neural-synthesis," arXiv preprint arXiv: 2404.04104, 2024.

[2] B. Lei, J. Ren, M. Feng, M. Cui, and X. Xie, "A hierarchical representation network for accurate and detailed face reconstruction from in-the-wild images," arXiv preprint arXiv: 2302.14434, 2023.

[3] R. Daněček, M. J. Black, and T. Bolkart, "Emoca: Emotion driven monocular face capture and animation," arXiv preprint arXiv: 2204.11312, 2022.

[4] Y. Feng, H. Feng, M. J. Black, and T. Bolkart, "Learning an animatable detailed 3d face model from in-the-wild images," *ACM Trans.*

Graphics, vol. 40, no. 4, p. 88, Aug. 2021.

[5] Z. Li, Z. Lu, H. Yan, B. Shi, G. Pan, Q. Zheng, and X. Jiang, "Spin-up: Spin light for natural light uncalibrated photometric stereo," arXiv preprint arXiv: 2404.01612, 2024.

[6] B. Yu, J. Ren, J. Han, F. Wang, J. Liang, and B. Shi, "EventPS: Real-time photometric stereo using an event camera," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, 2024, pp. 9602–9611.

[7] S. Ikehata, "Scalable, detailed and mask-free universal photometric stereo," arXiv preprint arXiv: 2303.00308, 2023.

[8] M. Wang, X. Guo, W. Dai, and J. Zhang, "Face inverse rendering via hierarchical decoupling," *IEEE Trans. Image Processing*, vol. 31, pp. 5748–5761, Aug. 2022.

[9] M. Wang, C. Wang, X. Guo, and J. Zhang, "Towards high-fidelity face normal estimation," in *Proc. 30th ACM Int. Conf. Multimedia*, Lisboa, Portugal, 2022, pp. 5172–5180.

[10] V. F. Abrevaya, A. Boukhayma, P. H. S. Torr, and E. Boyer, "Cross-modal deep face normals with deactivable skip connections," arXiv preprint arXiv: 2003.09691, 2020.

[11] X. Chen, Y. Zheng, Y. Zheng, Q. Zhou, H. Zhao, G. Zhou, and Y.-Q. Zhang, "DPF: Learning dense prediction fields with weak supervision," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Vancouver, Canada, 2023, pp. 15347–15357.

[12] M. Rossi, M. El Gheche, A. Kuhn, and P. Frossard, "Joint graph-based depth refinement and normal estimation," arXiv preprint arXiv: 1912.01306, 2020.

[13] Y. Zhang, S. Song, E. Yumer, M. Savva, J.-Y. Lee, H. Jin, and T. Funkhouser, "Physically-based rendering for indoor scene understanding using convolutional neural networks," arXiv preprint arXiv: 1612.07429, 2017.

[14] Z. Shu, E. Yumer, S. Hadap, K. Sunkavalli, E. Shechtman, and D. Samaras, "Neural face editing with intrinsic image disentangling," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Honolulu, USA, 2017, pp. 5444–5453.

[15] K. Zhang, Y. Su, X. Guo, L. Qi, and Z. Zhao, "MU-GAN: Facial attribute editing based on multi-attention mechanism," *IEEE/CAA J. Autom. Sinica*, vol. 8, no. 9, pp. 1614–1626, Sep. 2021.

[16] M. Wang, W. Dai, X. Guo, and J. Zhang, "Face inverse rendering from single images in the wild," in *Proc. IEEE Int. Conf. Multimedia and Expo*, Taipei, China, 2022, pp. 1–6.

[17] A. Lattas, S. Moschoglou, S. Ploumpis, B. Gecer, J. Deng, and S. Zafeiriou, "FitMe: Deep photorealistic 3D morphable model avatars," arXiv preprint arXiv: 2305.09641, 2023.

[18] N. Yang, B. Xia, Z. Han, and T. Wang, "A domain-guided model for facial cartoonlization," *IEEE/CAA J. Autom. Sinica*, vol. 9, no. 10, pp. 1886–1888, Oct. 2022.

[19] Y. Zheng, Y. Wang, G. Wetzstein, M. J. Black, and O. Hilliges, "PointAvatar: Deformable point-based head avatars from videos," arXiv preprint arXiv: 2212.08377, 2023.

[20] G. Trigeorgis, P. Snape, I. Kokkinos, and S. Zafeiriou, "Face normals "in-the-wild" using fully convolutional networks," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Honolulu, USA, 2017, pp. 340–349.

[21] S. Sengupta, A. Kanazawa, C. D. Castillo, and D. W. Jacobs, "SfSNet: Learning shape, reflectance and illuminance of faces 'in the wild'," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Salt Lake City, USA, 2018, pp. 6296–6305.

[22] J. E. Laird, P. S. Rosenbloom, and A. Newell, "Towards chunking as a general learning mechanism," in *Proc. 4th AAAI Conf. Artificial Intelligence*, Austin, USA, 1984, pp. 188–192.

[23] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proc. IEEE Int. Conf. Computer Vision*, Santiago, Chile, 2015, pp. 3730–3738.

[24] X. Long, Y. Zheng, Y. Zheng, B. Tian, C. Lin, L. Liu, H. Zhao, G. Zhou, and W. Wang, "Adaptive surface normal constraint for geometric estimation from monocular images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 9, pp. 6263–6279, Sep. 2024.

[25] R. Zhang, P.-S. Tsai, J. E. Cryer, and M. Shah, "Shape-from-shading: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 21, no. 8, pp. 690–706, Aug. 1999.

[26] Y. Wang, L. Zhang, Z. Liu, G. Hua, Z. Wen, Z. Zhang, and D. Samaras, "Face relighting from a single image under arbitrary unknown lighting conditions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 11, pp. 1968–1984, Nov. 2009.

[27] S. Biswas, G. Aggarwal, and R. Chellappa, "Robust estimation of albedo for illumination-invariant matching and shape recovery," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 5, pp. 884–899, May 2009.

[28] M. K. Johnson and E. H. Adelson, "Shape estimation in natural illumination," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Colorado Springs, USA, 2011, pp. 2553–2560.

[29] J. T. Barron and J. Malik, "Shape, albedo, and illumination from a single image of an unknown object," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Providence, USA, 2012, pp. 334–341.

[30] Y. Xiong, A. Chakrabarti, R. Basri, S. J. Gortler, D. W. Jacobs, and T. Zickler, "From shading to local shape," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 1, pp. 67–79, Jan. 2015.

[31] Y. Mei, Y. Zeng, H. Zhang, Z. Shu, X. Zhang, S. Bi, J. Zhang, H. J. Jung, and V. M. Patel, "Holo-relighting: Controllable volumetric portrait relighting from a single image," arXiv preprint arXiv: 2403.09632, 2024.

[32] Y. Cheng, Z. Chen, X. Ren, W. Zhu, Z. Xu, D. Xu, C. Yang, and Y. Yan, "3D-aware face editing via warping-guided latent direction learning," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, 2024, pp. 916–926.

[33] Z. Cai, K. Jiang, S.-Y. Chen, Y.-K. Lai, H. Fu, B. Shi, and L. Gao, "Real-time 3d-aware portrait video relighting," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, 2024, pp. 6221–6231.

[34] A. Ranjan, K. M. Yi, J.-H. R. Chang, and O. Tuzel, "FaceLit: Neural 3D relightable faces," arXiv preprint arXiv: 2303.15437, 2023.

[35] L. Zhang, J. Liu, B. Zhang, D. Zhang, and C. Zhu, "Deep cascade model-based face recognition: When deep-layered learning meets small data," *IEEE Trans. Image Process.*, vol. 29, pp. 1016–1029, Sep. 2019.

[36] Q. Wang, T. Wu, H. Zheng, and G. Guo, "Hierarchical pyramid diverse attention networks for face recognition," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Seattle, USA, 2020, pp. 8323–8332.

[37] F. Xue, Z. Tan, Y. Zhu, Z. Ma, and G. Guo, "Coarse-to-fine cascaded networks with smooth predicting for video facial expression recognition," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, New Orleans, USA, 2022, pp. 2411–2417.

[38] B. Yu and D. Tao, "Anchor cascade for efficient face detection," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2490–2501, May 2019.

[39] J. Lou, X. Cai, J. Dong, and H. Yu, "Real-time 3D facial tracking via cascaded compositional learning," *IEEE Trans. Image Process.*, vol. 30, pp. 3844–3857, Mar. 2021.

[40] S. Ma, Y. Wang, Y. Wei, J. Fan, T. H. Li, H. Liu, and F. Lv, "CAT: Localization and identification cascade detection transformer for open-world object detection," arXiv preprint arXiv: 2301.01970, 2023.

[41] R. Wu, G. Zhang, S. Lu, and T. Chen, "Cascade EF-GAN: Progressive facial expression editing with local focuses," arXiv preprint arXiv: 2003.05905, 2020.

[42] L. Chen, R. K. Maddox, Z. Duan, and C. Xu, "Hierarchical cross-modal talking face generation with dynamic pixel-wise loss," arXiv preprint arXiv: 1905.03820, 2019.

[43] X. Li, Z. Liu, P. Luo, C. C. Loy, and X. Tang, "Not all pixels are equal: Difficulty-aware semantic segmentation via deep layer cascade," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Honolulu, USA, 2017, pp. 6459–6468.

[44] D. Lin, G. Chen, D. Cohen-Or, P.-A. Heng, and H. Huang, "Cascaded feature network for semantic segmentation of RGB-D images," in *Proc. IEEE Int. Conf. Computer Vision*, Venice, Italy, 2017, pp. 1320–1328.

[45] P. Hu, G. Wang, X. Kong, J. Kuen, and Y.-P. Tan, "Motion-guided cascaded refinement network for video object segmentation," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Salt Lake City, USA, 2018, pp. 1400–1409.

[46] K. Chen, J. Pang, J. Wang, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Shi, W. Ouyang, C. C. Loy, and D. Lin, "Hybrid task cascade for instance segmentation," arXiv preprint arXiv: 1901.07518, 2019.

[47] H. Ding, S. Qiao, A. Yuille, and W. Shen, "Deeply shape-guided cas-

cade for instance segmentation," arXiv preprint arXiv: 1911.11263, 2021.

[48] E. Chouzenoux, J. Idier, and S. Moussaoui, "A majorize–minimize strategy for subspace optimization applied to image restoration," *IEEE Trans. Image Process.*, vol. 20, no. 6, pp. 1517–1528, Jun. 2011.

[49] J. Xie, J. Yang, J. J. Qian, Y. Tai, and H. M. Zhang, "Robust nuclear norm-based matrix regression with applications to robust face recognition," *IEEE Trans. Image Process.*, vol. 26, no. 5, pp. 2286–2295, May 2017.

[50] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," arXiv preprint arXiv: 1512.03385, 2016.

[51] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vision*, vol. 115, no. 3, pp. 211–252, Apr. 2015.

[52] L. A. Gatys, A. S. Ecker, and M. Bethge, "A neural algorithm of artistic style," arXiv preprint arXiv: 1508.06576, 2015.

[53] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," arXiv preprint arXiv: 1611.07004, 2017.

[54] S. Zafeiriou, M. Hansen, G. Atkinson, V. Argyriou, M. Petrou, M. Smith, and L. Smith, "The photoface database," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Colorado Springs, USA, 2011, pp. 132–139.

[55] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "300 faces in-the-wild challenge: The first facial landmark localization challenge," in *Proc. IEEE Int. Conf. Computer Vision Workshops*, Sydney, Australia, 2013, pp. 397–403.

[56] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," arXiv preprint arXiv: 1812.04948, 2019.

[57] A. D. Bagdanov, A. Del Bimbo, and I. Masi, "The Florence 2D/3D hybrid face dataset," in *Proc. Joint ACM Workshop on Human Gesture and Behavior Understanding*, Scottsdale, USA, 2011, pp. 79–80.

[58] W.-C. Ma, T. Hawkins, P. Peers, C.-F. Chabert, M. Weiss, and P. Debevec, "Rapid acquisition of specular and diffuse normal maps from polarized spherical gradient illumination," in *Proc. 18th Eurographics Conf. Rendering Techniques*, Grenoble, France, 2007, pp. 183–194.

[59] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deeplearning library," in *Proc. 33rd Int. Conf. Neural Information Processing Systems*, Vancouver Canada, 2019, pp. 721.

[60] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv: 1412.6980, 2014.

[61] L. Guo, H. Zhu, Y. Lu, M. Wu, and X. Cao, "RAFaRe: Learning robust and accurate non-parametric 3D face reconstruction from pseudo 2D&3D pairs," in *Proc. 37th AAAI Conf. Artificial Intelligence*, Washington, USA, 2023, pp. 719–727.

[62] Z. Zhang, Y. Ge, R. Chen, Y. Tai, Y. Yan, J. Yang, C. Wang, J. Li, and F. Huang, "Learning to aggregate and personalize 3D face from in-the-wild photo collection," arXiv preprint arXiv: 2106.07852, 2021.

[63] Y. Feng, F. Wu, X. Shao, Y. Wang, and X. Zhou, "Joint 3D face reconstruction and dense alignment with position map regression network," in *Proc. 15th European Conf. Computer Vision*, Munich, Germany, 2018, pp. 557–574.

[64] A. Bansal, B. Russell, and A. Gupta, "Marr revisited: 2D-3D alignment via surface normal prediction," arXiv preprint arXiv: 1604.01347, 2016.

[65] I. Kokkinos, "UberNet: Training a universal convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory," arXiv preprint arXiv: 1609.02132, 2016.

[66] X. Zhu, X. Liu, Z. Lei, and S. Z. Li, "Face alignment in full pose range: A 3D total solution," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 1, pp. 78–92, Jan. 2019.

[67] N. Ahn, B. Kang, and K.-A. Sohn, "Fast, accurate, and lightweight super-resolution with cascading residual network," in *Proc. 15th European Conf. Computer Vision*, Munich, Germany, 2018, pp. 256–272.

[68] L. Zhang, Y. He, Q. Zhang, Z. Liu, X. Zhang, and C. Xiao, "Document image shadow removal guided by color-awarebackground," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Vancouver, Canada, 2023, pp. 1818–1827.

[69] J. Li, Z. Zhang, X. Liu, C. Feng, X. Wang, L. Lei, and W. Zuo, "Spatially adaptive self-supervised learning for real-world image denoising," arXiv preprint arXiv: 2303.14934, 2023.

[70] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M.-H. Yang, "Restormer: Efficient transformer for high-resolution image restoration," arXiv preprint arXiv: 2111.09881, 2022.

[71] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of styleGAN," arXiv preprint arXiv: 1912.04958, 2020.

[72] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proc. IEEE Int. Conf. Computer Vision*, Venice, Italy, 2017, pp. 1510–1519.

Meng Wang received the Ph.D. degree in software engineering from the College of Intelligence and Computing, Tianjin University. He is currently a Lecturer at Tiangong University. His research interests include computer vision, machine learning, and pattern recognition.

Jiawan Zhang (Senior Member, IEEE) received the M.Sc. and Ph.D. degrees in computer Science from Tianjin University in 2001 and 2004, respectively. Currently, he is a Full Professor at the College of Intelligence and Computing, Tianjin University. His research interests include computer vision visualization and visual analysis. He serve(d) for academic events including the General Co-Chair of ChinaVis (2015, 2016), PacificVis (2019, 2020). He also serve(d) as the Program Committee Member or Reviewer for many conferences and journals including CVPR, ICCV, AAAI, VIS, PacificVis, EuroVis, IEEE TVCG, IEEE TIP.

Jiayi Ma (Senior Member, IEEE) received the B.S. degree in information and computing science and the Ph.D. degree in control science and engineering from the Huazhong University of Science and Technology, in 2008 and 2014, respectively. He is currently a Professor with the Electronic Information School, Wuhan University. His research interests include computer vision, machine learning, and robotics. He has authored or co-authored more than 300 refereed journal and conference papers, including IEEE TPAMI/TIP, IJCV, CVPR, ICCV, ECCV, etc. He has been identified in the 2019–2022 Highly Cited Researcher lists from the Web of Science Group. He is an Area Editor of *Information Fusion*, and an Associate Editor of *IEEE/CAA Journal of Automatica Sinica*.

Xiaojie Guo (Senior Member, IEEE) received the Ph.D. degree in computer science from the School of Computer Science and Technology, Tianjin University (TJU) in 2013. He is currently an Associate Professor with tenure (Peiyang Young Scientist) with Tianjin University. Prior to joining TJU, he spent about four years with the Institute of Information Engineering, Chinese Academy of Sciences. His research interests include computer vision and machine learning. He was a recipient of the Piero Zamperoni Best Student Paper Award in ICPR 2010, and the Best Student Paper Runner-up in ICME 2018.