FNFORMER: A TRANSFORMER-BASED FACE NORMAL ESTIMATOR

Meng Wang[1], *Xiaojie Guo*[2], *and Jiawan Zhang*[2]*

[1]School of Computer Science and Technology, Tiangong University, Tianjin, China
[2]School of Computer Software, Tianjin University, Tianjin, China
{autohdr, xj.max.guo}@gmail.com, jwzhang@tju.edu.cn

ABSTRACT

Face normal estimation is a crucial step in the development of 3D facial applications, particularly for face modeling and relighting. U-shaped networks are widely used for the task and have witnessed remarkable success. However, CNN-based methods often suffer from unsatisfied generalization ability to out-of-distribution/unseen data, because they do not adequately model long-range dependencies. To address this limitation, Transformer-based approaches have been developed, which benefit from the global self-attention mechanism. Nevertheless, merely using them to learn face normal may lead to limited localization abilities due to insufficient low-level details. In this work, we customize a hybrid model called *FNFormer* that combines Transformer and CNN to achieve accurate face normal estimation. The proposed model encodes tokenized image patches from CNN feature maps as input to extract global context features using Transformer blocks. Additionally, it extracts detailed local spatial information from a U-shaped CNN. Both the CNN and Transformer features are then integrated for further learning, enabling the network to take both the local and global information into account effectively. Extensive experimental results demonstrate that our proposed *FNFormer* achieves state-of-the-art performance on various datasets. Our code is available at `https://github.com/AutoHDR/FNFormer`.

Index Terms— Face normal estimation, hybrid model, CNN, Transformer

1. INTRODUCTION

Face normal estimation aims to reconstruct 3D face from the given face images. It has attracted more attention in recent years as it has shown promising potential in downstream tasks such as face relighting [1, 2, 3] and face editing [4]. Since the groundbreaking work of [5], CNNs have dominated face normal estimation. A standard CNN model for this task follows an encoder-decoder architecture, where the encoder learns feature representations and the decoder predicts these features at a pixel level. Among these two components, feature representation learning (i.e., the encoder) is arguably the most important, which enhances the representation of features by designing different network structures [5, 3, 6, 7].

Fig. 1: Comparisons with 'HFFNE' [7] on normals and shadings generated with different lights on the FFHQ [8].

Although face normal estimation has made considerable improvements, current CNN-based models continue to grapple with certain limitations: 1). *Elaborate network architectures*: These models often resort to intricate and meticulously crafted network structures in order to enhance the capacity of model features and overall performance. 2). *Data requirements for generalization*: To bolster the model's generalization capabilities, a substantial volume of training data is required, underscoring the significance of extensive dataset availability. 3). *Local Region focus*: The convolutional nature of CNNs confines their ability to address only local regions within an image, thereby limiting their efficiency in modeling long-range dependencies. These challenges point towards the need for further research and innovation in the field of face normal estimation to address these limitations and propel the progress in this domain.

In response to these challenges, we embarked on a further exploration of the potential utility of transformers within the context of face normal estimation. However, our investigation revealed that employing transformers to encode tokenized image patches and subsequently upsampling hidden feature representations directly into dense outputs at full res-

*Corresponding author.

olution yielded unsatisfactory outcomes. This was primarily attributed to the inherent nature of transformers, which treat inputs as 1D sequences and focus exclusively on capturing global context. Consequently, the resultant low-resolution features lacked the nuanced localization information essential for accurate face normal estimation.

To harness the benefits of both the long-range dependency capability inherent in transformers and the localized spatial representation prowess of CNNs, we introduce *FNFormer* as a pioneering transformer-based framework tailored for face normal estimation. Our framework pioneers the incorporation of attention mechanisms rooted in the transformer architecture, facilitating the fusion of these two powerful paradigms. Recognizing the challenges posed by transformers in preserving local relationships, we devise a novel hybrid CNN-Transformer architecture. This innovative design effectively amalgamates the intricate and fine-grained spatial information encapsulated within CNN features with the expansive global context adeptly captured by transformers.

To mitigate the challenge of lost local relations in transformer-based methods, we implement a U-shaped architecture to recover the self-attentive features derived from transformers. This strategic fusion culminates in the precise localization of spatial attributes, enriching the quality of face normal estimation(as shown in Fig 1).

The main contributions of this paper are as follows:

- We propose a hybrid framework that combines CNN and Transformer for face normal estimation. This approach involves extracting and fusing both global and local features and establishing long-range dependencies, resulting in improving model performance and bringing strong generalization capabilities.
- Experimental results on different datasets demonstrate that our model achieves state-of-the-art results with better performance compared to previous methods.

2. RELATED WORK

Face shape estimation. Recent researchers have underscored the heightened efficacy of CNNs in addressing and resolving this intricate problem [6]. Some researchers [5, 9] employ the generation of synthetic paired data utilizing the 3DMM to facilitate the training of their networks for predicting real-world facial shapes. Nevertheless, neglecting to account for the disparity between synthetic and real data can lead to a decline in the performance of models.

Recently, to bridge the gap between synthetic and real data, Sengupta *et al.*[3] combined both types of data - real and synthetic to train their model. Abrevaya *et al.*[6] developed a network that utilizes deactivatable skip connections, allowing for the use of both paired and unpaired data. To avoid the gap, Wang *et al.*[7] proposed a two-stage framework with the exemplar-based learning to produce high-quality face normal.

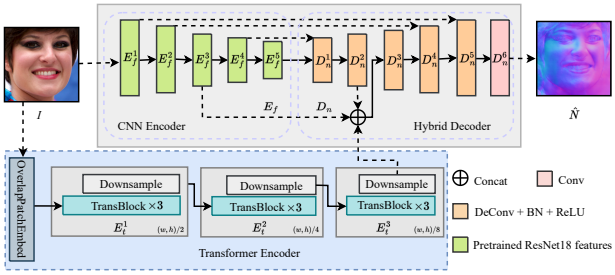

Fig. 2: The whole framework of *FNFormer*, which contains three parts: a CNN encoder $\mathbf{E}_f$, a Transformer encoder $\mathbf{E}_t$ and a hybird decoder $\mathbf{D}_n$. These modules are consecutively applied to the input image, thereby facilitating the extraction of pertinent features to produce high-fidelity face normal.

While these methods can enhance face normal to some extent, they are not effective on face details.

Vision transformers. Recently, ViT, introduced by Dosovitskiy *et al.*[10], allowed for the application of transformer models to image recognition. This approach motivated researchers to explore the use of transformers in low-level vision tasks [11, 10]. Combining CNNs with different forms of transformers has shown promising results in various tasks, such as image restoration [12] and semantic segmentation [13]. For instance, Wang *et al.*[13] proposed Uformer, a general U-shaped transformer for image restoration. The success of the transformer motivated us to explore a pure transformer-based model for face normal estimation.

Combining CNNs with attention mechanisms. Combining CNNs with attention mechanisms is to enhance the performance of CNNs in vision tasks. Attention mechanisms enable the network to learn more distinctive features, thereby improving its accuracy. Several research works [14, 15] have proposed different techniques for combining CNNs with attention mechanisms. For example, Chen *et al.*[16] proposed a Transformer-based model that uses masked attention for universal image segmentation. Deng *et al.*[17] developed a transformer-based style transfer framework that generates stylization results with well-preserved structures and details of the input content image. Inspired by these methods, we design a hybrid CNN-Transformer architecture to learn the characteristics of paired data for face normal estimation.

3. METHOD

3.1. Network Structure

CNN-based face features encoding. In Figure 2, we initiate the process by extracting local facial features using a CNN encoder denoted as $\mathbf{E}_f$, which leverages a pretrained ResNet18 [18]. This strategy enables us to harness the benefits of transfer learning, saving valuable time and computational resources that would have been otherwise required for training from scratch. Given an input image $\mathbf{I} \in \mathcal{R}^{H \times W \times C}$,

we direct it through the CNN encoder module to generate multi-scale feature maps across five distinct layers, denoted as $\{\mathbf{E}_f^i\}_{i=1}^5$. As a result, the feature map $\mathbf{E}_f^1$ has a spatial dimension of $\frac{H}{2} \times \frac{W}{2}$, while $\mathbf{E}_f^5$ spans $\frac{H}{32} \times \frac{W}{32}$.

Transformer-based encoder. As previously discussed, the CNN is primarily tailored for local feature extraction, with a focus on capturing fine-grained details within limited regions of the input image. However, this localized approach may inadvertently overlook essential aspects crucial for high-quality face normal restoration, such as the global facial structure and the overall distribution of normals across the face. In response to this inherent limitation, we introduce a transformer block specifically to collectively capture long-range image relations, thereby aiming to enhance the model's comprehension of the broader global context. This inspiration for incorporating transformer-based mechanisms arises from the remarkable success of transformers in various image analysis tasks. To achieve this objective, we employ the Transformer block(TransBlock), inspired by the work in [12], to construct a transformer encoder, effectively capturing the critical long-range dependency features within the face images.

To achieve this goal, we commence by conducting an overlapped image patch embedding operation(OverlapPatchEmbed) on the input image, employing a 3x3 convolution operation. Subsequently, we integrate three consecutive transformer-based modules to extract long-range dependent features. Each transformer module consists of three TransBlocks, positioned immediately after the downsampling process applied to the extracted features. This operation results in the final size of the extracted feature $\mathbf{E}_t^3$ being reduced to one eighth of the original image size, ensuring a suitable balance between computational efficiency and the preservation of crucial global facial context.

Hybrid normal decoder. The FNFormer model synergistically harnesses the robust feature extraction capabilities of both a CNN encoder and a transformer encoder to capture a comprehensive range of features from the input data. This dual-source feature extraction process serves as the foundation for a series of meticulously orchestrated operations, culminating in the precise prediction of facial normals.

At the outset, the local features extracted by the CNN encoder undergo a dedicated decoding process. As the network advances, a critical point is reached, where the size of the feature map becomes one-eighth of the original dimensions. It is at this juncture that a pivotal operation takes place: the fusion of the global features obtained from the transformer encoder represented as $\mathbf{F} = \mathbf{Con}[\mathbf{E}_t^3, \mathbf{E}_f^3, \mathbf{D}_n^2]$. $\mathbf{Con}$ is meticulously executed through the concatenation operation, strategically positioned to complement and enrich the local features. By integrating global context alongside local details, this fusion step provides the model with a broader and more holistic perspective. Subsequent to the fusion of these enriched features, the combined feature set $\mathbf{F}$ undergoes a pivotal final decoding stage, leading to the generation of the ultimate high-quality face normal estimation denoted as $\hat{N}$.

3.2. Loss Function

Reconstruction Loss. Following a methodology akin to that of [5], our approach aims to enhance the precision of the generated face normal distribution. To achieve this, we employ the cosine loss to evaluate the discrepancy as follows:

$$\mathcal{L}_{recon} = \text{CosineLoss}\left(\hat{N}, N_{gt}\right), \tag{1}$$

where $\hat{N}$ and N_{gt} are the predicted normal and the ground truth normal, respectively.

Adversarial Loss. To capture fine details, using only the cosine loss results in low-frequency normal. Therefore, we incorporate an adversarial loss that can effectively capture high-frequency details of normal. The adversarial loss is given by:

$$\mathcal{L}_{adv} = D_{adv}(\hat{N}), \tag{2}$$

where $D_{adv}(\cdot)$ refers to the normal discriminator used to determine the authenticity of the predicted normal.

Total Variation Loss. To preserve normal structures and produce sharper geometry, we employ a total variation(TV) loss to impose spatial smoothness and reduce the presence of noise or artifacts in the generated normal while preserving face geometry details. The TV loss is formulated as:

$$\begin{aligned}\mathcal{L}_{TV} = \sum_{p,q} (||\bar{N}(p, q+1) - \bar{N}(p, q)||_2 \\ +||\bar{N}(p+1, q) - \bar{N}(p, q)||_2)/(W \times H),\end{aligned} \tag{3}$$

where (p, q) denotes the horizontal and vertical coordinates of the normal's values, W and H denote the resolution sizes of the predicted normal maps, $\bar{N}(p, q) = (\hat{N}/2{+}0.5)$ represents the pixel value of the predicted normal in image space, and $\|\cdot\|_2$ means the L2 norm, respectively.

Therefore, *FNFormer* is optimized by minimizing the following overall objective function:

$$\mathcal{L}_{otal} = \mathcal{L}_{recon} + \lambda_{adv}\mathcal{L}_{adv} + \lambda_{tv}\mathcal{L}_{TV}, \tag{4}$$

where λ_{tv} and λ_{adv} indicate the trade-off parameters for the reconstruction loss and the adversarial loss, respectively.

4. EXPERIMENTS AND RESULTS

4.1. Experiments

Datasets. In our experiments, we use Photoface [19] dataset for our training and evaluate the model validity on 300-W [20], FFHQ [8], Florence [21] and ICT-3DRFE [22] datasets. Following [6, 7], we randomly choose 80% of the images from Photoface for training purposes, while the remaining 20% are reserved for quantitative evaluation.

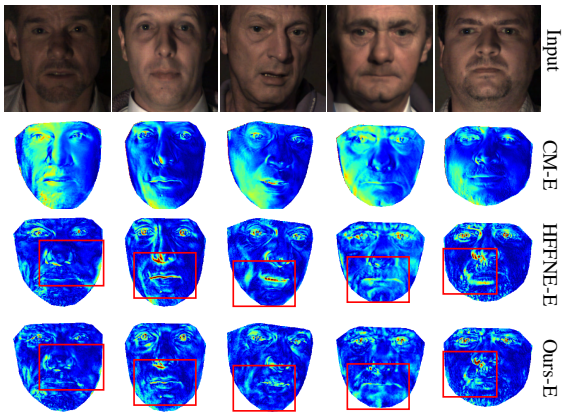

Fig. 3: Normal error comparisons on the Photoface [19]. 'CM-E', 'HFFNE-E', and 'Ours-E' are the 'CM' [6], 'HFFNE' [7] and our error maps, respectively. The color dark blue indicates a smaller estimation error, while the color red indicates a larger estimation error.

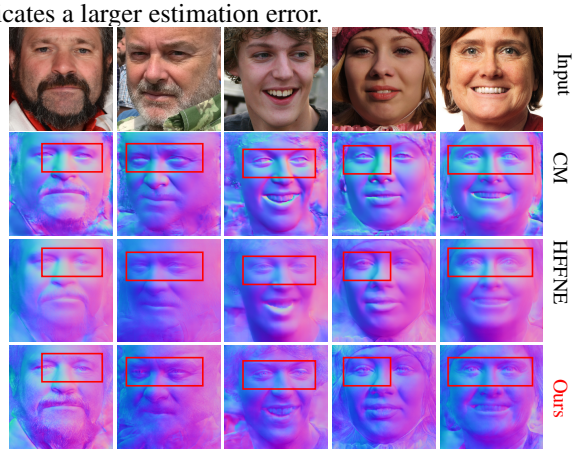

Fig. 4: Normals comparisons on the FFHQ [8]. Compared to 'HFFNE' [7], ours can recover more precise face normals.

Implementation Details. We implement our *FNFormer* using the PyTorch framework. Meanwhile, we optimize our model by Adam and set $\beta_1 = 0.9$ and $\beta_2 = 0.99$. The initial learning rate is set to 1×10^{-4}. For *FNFormer* , we empirically set $\lambda_{tv} = 0.001$ and $\lambda_{adv} = 0.0001$. We also use Adam to optimize our *FNFormer*. Our *FNFormer* is trained on an NVIDIA GTX 3060 GPU with a batch size of 8 and 200 epochs. We adopt the transformer block from Restormer [12] as the fundamental unit for constructing our Transformer-based encoder. Each TransBlock comprises two transformer blocks and two attention heads.

Evaluation Metrics. To evaluate the accuracy of our normal estimation results, we follow previous methods [5, 3, 6, 7] and use two objective metrics: mean and standard deviation angular error between the estimated normal and the ground truth (MSDAE), as well as the percentage of pixels within facial regions with an angular error (PPAE) less than 20°, 25°, and 30°. Furthermore, we utilize geometric shading and a normal error map for qualitative comparisons.

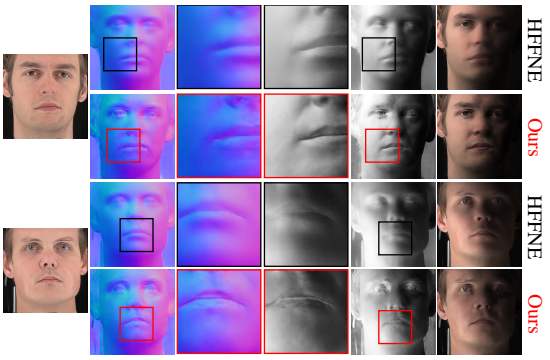

Fig. 5: Comparisons between normals, shadings, and relit faces on the ICT-3DRFE [22]. We focused on magnifying the normal and shading maps of their respective regions to show ours recover more precise normals than 'HFFNE' [7].

Table 1: Normal reconstruction error on the Photoface [19]. The data comes from 'HFFNE' [7], The methods listed in the upper half of the table are tested without training on the data.

Method	Mean ± std	$< 20°$	$< 25°$	$< 30°$
Pix2V	33.9±5.6	24.8%	36.1%	47.6%
Extreme	27.0±6.4	37.8%	51.9%	64.5%
3DMM	26.3±10.2	4.3%	56.1%	89.4%
3DDFA	26.0±7.2	40.6%	54.6%	66.4%
SfSNet	25.5±9.3	43.6%	57.5%	68.7%
PRN	24.8±6.8	43.1%	57.4%	69.4%
Cross-modal	22.8±6.5	49.0%	62.9%	74.1%
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%
Cross-modal-ft	12.0±5.3	85.2%	92.0%	95.6%
LAP	12.3±4.5	84.9%	92.4%	96.3%
HFFNE	11.3±7.7	88.6 %	94.4 %	97.2 %
Ours	**7.7±5.7**	**95.6** %	**97.9** %	**99.1** %

4.2. Comparison with SOTA methods

In Table 1, we compare the performance of the proposed *FNFormer* with several SOTA face normal estimation methods on the Photoface dataset [19]. The table demonstrates that our method yields significantly smaller MSDAE and PPAE values compared to previous methods. Additionally, we show the normal error maps comparison with 'HFFNE' [7], in Fig. 3. It is evident that our normal maps exhibit superior performance in the majority of areas with wrinkles and significant geometric changes on the face. This is consistent with the notion that our *FNFomer* has both local and global learning abilities for normal prediction. 'HFFNE' are capable of accurately estimating normals in areas where the face geometry with slower rates of change. However, in regions where sudden changes, such as around the eyes, corners of the mouth, and wrinkles, 'HFFNE' may provide inaccurate estimates. Nonetheless, our

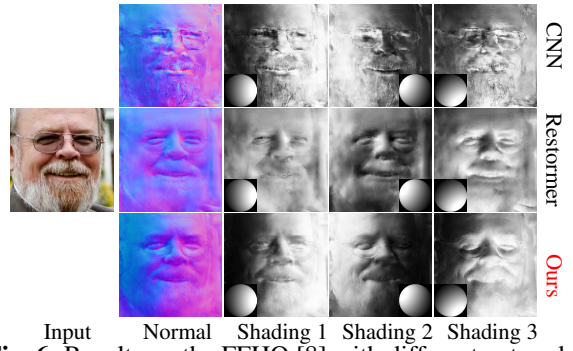

Fig. 6: Results on the FFHQ [8] with different networks.

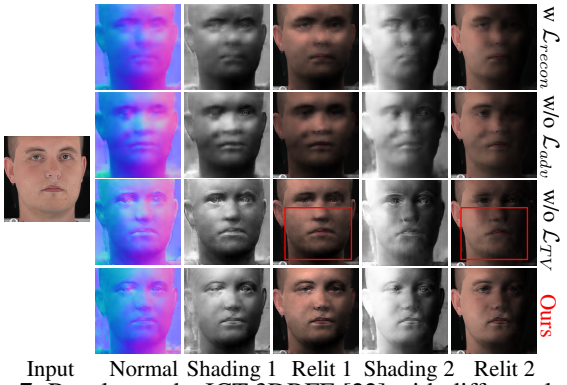

Fig. 7: Results on the ICT-3DRFE [22] with different loss.

approach can effectively handle these challenging situations.

In Fig. 4, our method produces much higher quality face normals than previous methods. When it comes to recovering detail in face normals, 'CM' [6] can generate some level of detail but struggles with certain face images, while 'HFFNE' [7] surpass the previous method 'CM' in terms of normal detail recovery. The figure shows that the 'HFFNE' and 'CM' methods do not perform well in the eye crease region, resulting in a loss of high-fidelity geometric detail, referring to the area highlighted by the red rectangle in Fig. 4. In contrast, both of them sacrifice high-frequency details, while ours can recover finer-grained normals with better accuracy.

In Fig. 5, we present the comparison by rendering new light faces under the Lambertian reflectance model [23]. Compared to the previous method 'HFFNE' [7], our face normal estimation approach is superior in preserving more facial details while accurately estimating normal directions. Moreover, our method can produce high-quality face normals to render realistic relit faces under new lighting conditions. The shading maps demonstrate that our method preserves sharper details in areas with high levels of detail, such as the corners of the mouth. Furthermore, the relit faces exhibit a more authentic appearance in contrast to 'HFFNE', which generates artifacts. Please zoom in to see the details more clearly.

To ensure that our model performs well beyond the training data, we generate paired face with normal from the Florence [21]. We then use our model to predict normal maps and

Table 2: Normal reconstruction error on the unseen faces generated from the Florence [21].

Method	Mean ± std	$< 20°$	$< 25°$	$< 30°$
Extreme	19.2±2.2	64.7%	51.9%	64.5%
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%
Cross-modal	11.3±1.5	89.3%	94.6%	96.9%
HFFNE	10.1±3.4	92.3%	95.6%	97.8%
Ours	**7.2±3.0**	**96.2%**	**98.0%**	**99.1%**

Table 3: Normal reconstruction error on different experiment settings from the Photoface [19].

Method	Mean ± std	$< 20°$	$< 25°$	$< 30°$
CNN	7.8±5.6	95.1%	97.5%	98.6%
Restormer	8.7±6.6	93.9%	96.9%	98.3%
w $\mathcal{L}_{recon}$	7.9±5.9	96.0%	98.4%	99.0%
w/o $\mathcal{L}_{TV}$	8.6±7.4	93.2%	96.5%	98.1%
w/o $\mathcal{L}_{adv}$	9.1±7.1	92.5%	96.3%	98.1%

compared them with those produced through the 3D model. The results are presented in Table 2, which clearly demonstrates that our model has a strong generalization capability and performs exceptionally well on previously unseen data.

In Fig. 1, we generate shadings with varying lighting conditions by rendering the estimated normal maps from 4 different angles of light. The results between ours and 'HFFNE' [7] reveals that our approach preserves fine-grained geometric of the face, especially in regions where facial geometry varies significantly such as beards and wrinkles. In contrast, 'HFFNE' produces less accurate normal maps with a loss of high-frequency details. Furthermore, while the normal maps estimated by 'HFFNE' retain their detail under certain angular light, the shading maps produced under different angles of lighting contain artifacts and exhibit less realistic, as shown in the second row of the figure where 'HFFNE' generates shading maps with artifacts that are not present in ours. Thus, highlighting the precision and accuracy of *FNFormer*.

5. ABLATION STUDIES

Ablation of network architectures. To demonstrate the efficacy of *FNFormer*, we trained two models: a U-shaped CNN [24] and a vision transformer network using Restormer [12]. The results are in the first two rows of Table 3 and Fig. 6. From the table and the figure, CNN modle exhibits certain advantages in quantitative metrics on the test data, but its generalization ability is weak when inferenced on unseen data. The Restormer successfully eliminates artifacts and retains details. Nevertheless, when rendered with different lighting, the shading reveals inaccuracies in the predicted normals.

Ablation of loss functions. Rows 3 to 5 of Table 3 demonstrate that the quantitative evaluation is good, regardless of the loss function. However, without $\mathcal{L}_{adv}$, the predicted normals become blurred and lose high-frequency details (as shown in rows 1 and 2 of Fig. 7). On the other hand, 'w/o $\mathcal{L}_{TV}$' can produce normals with high-frequency detail, but contain artifacts when rendering new relit faces (as shown in row 3 of Fig. 7) due to inaccurate normal orientation. The performance of our full model ('Ours') has a generalization ability.

6. CONCLUSION

In this paper, we have proposed a novel hybrid CNN-Transformer model for face normal estimation, which leverages the strengths of both networks. Our model is capable of generalizing well to new, unseen face images. Furthermore, detailed qualitative and quantitative evaluations show that our *FNFormer* significantly outperforms existing state-of-the-art methods in terms of accuracy, robustness, and generalization. The superior performance of our model is due to its ability to preserve fine-grained details of the face while accurately estimating the normal direction. We believe that our work provides a promising direction for future research in the field of face normal estimation.

7. REFERENCES

[1] Meng Wang, Xiaojie Guo, Wenjing Dai, and Jiawan Zhang, "Face inverse rendering via hierarchical decoupling," *TIP*, vol. 31, pp. 5748–5761, 2022.

[2] Meng Wang, Wenjing Dai, Xiaojie Guo, and Jiawan Zhang, "Face inverse rendering from single images in the wild," in *ICME*. IEEE, 2022, pp. 1–6.

[3] Soumyadip Sengupta, Angjoo Kanazawa, Carlos D Castillo, and David W Jacobs, "Sfsnet: Learning shape, reflectance and illuminance of facesin the wild'," in *CVPR*, 2018, pp. 6296–6305.

[4] Zhixin Shu, Ersin Yumer, Sunil Hadap, Kalyan Sunkavalli, Eli Shechtman, and Dimitris Samaras, "Neural face editing with intrinsic image disentangling," in *CVPR*, 2017, pp. 5541–5550.

[5] George Trigeorgis, Patrick Snape, Iasonas Kokkinos, and Stefanos Zafeiriou, "Face normals" in-the-wild" using fully convolutional networks," in *CVPR*, 2017, pp. 38–47.

[6] Victoria Fernández Abrevaya, Adnane Boukhayma, Philip HS Torr, and Edmond Boyer, "Cross-modal deep face normals with deactivable skip connections," in *CVPR*, 2020, pp. 4979–4989.

[7] Meng Wang, Chaoyue Wang, Xiaojie Guo, and Jiawan Zhang, "Towards high-fidelity face normal estimation," in *ACM MM*, 2022, pp. 5172–5180.

[8] Tero Karras, Samuli Laine, and Timo Aila, "A style-based generator architecture for generative adversarial networks," in *CVPR*, 2019, pp. 4401–4410.

[9] Ayush Tewari, Michael Zollhofer, Hyeongwoo Kim, Pablo Garrido, Florian Bernard, Patrick Perez, and Christian Theobalt, "Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction," in *ICCVW*, 2017, pp. 1274–1283.

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *ICLR*, 2021.

[11] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li, "Uformer: A general u-shaped transformer for image restoration," in *CVPR*, 2022, pp. 17683–17693.

[12] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang, "Restormer: Efficient transformer for high-resolution image restoration," in *CVPR*, 2022, pp. 5728–5739.

[13] Huiyu Wang, Yukun Zhu, Bradley Green, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen, "Axial-deeplab: Stand-alone axial-attention for panoptic segmentation," in *ECCV*. Springer, 2020, pp. 108–126.

[14] Prarthana Bhattacharyya, Chengjie Huang, and Krzysztof Czarnecki, "Sa-det3d: Self-attention based context-aware 3d object detection," in *ICCV*, 2021, pp. 3022–3031.

[15] Shengheng Deng, Zhihao Liang, Lin Sun, and Kui Jia, "Vista: Boosting 3d object detection via dual cross-view spatial attention," in *CVPR*, 2022, pp. 8448–8457.

[16] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar, "Masked-attention mask transformer for universal image segmentation," in *CVPR*, 2022, pp. 1290–1299.

[17] Yingying Deng, Fan Tang, Weiming Dong, Chongyang Ma, Xingjia Pan, Lei Wang, and Changsheng Xu, "Stytr2: Image style transfer with transformers," in *CVPR*, 2022, pp. 11326–11336.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[19] Stefanos Zafeiriou, Mark Hansen, Gary Atkinson, Vasileios Argyriou, Maria Petrou, Melvyn Smith, and Lyndon Smith, "The photoface database," in *CVPRW*. IEEE, 2011, pp. 132–139.

[20] Christos Sagonas, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic, "300 faces in-the-wild challenge: The first facial landmark localization challenge," in *ICCVW*, 2013, pp. 397–403.

[21] Andrew D Bagdanov, Alberto Del Bimbo, and Iacopo Masi, "The florence 2d/3d hybrid face dataset," in *J-HGBU*, 2011, pp. 79–80.

[22] Wan-Chun Ma, Tim Hawkins, Pieter Peers, Charles-Felix Chabert, Malte Weiss, Paul E Debevec, et al., "Rapid acquisition of specular and diffuse normal maps from polarized spherical gradient illumination.," *Rendering Techniques*, vol. 2007, no. 9, pp. 10, 2007.

[23] David W Jacobs and Ronen Basri, "Lambertian reflectance and linear subspaces," Feb. 8 2005, US Patent 6,853,745.

[24] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros, "Image-to-image translation with conditional adversarial networks," in *CVPR*, 2017, pp. 1125–1134.