Learning to regulate 3D head shape by removing occluding hair from in-the-wild images

Sohan Anisetty* Varsha Saravanabavan† Cai Yiyu‡
IIT Dharwad NTU Singapore NTU Singapore

ABSTRACT
Recent 3D face reconstruction methods reconstruct the entire head compared to earlier approaches which only model the face. Although these methods accurately reconstruct facial features, they do not explicitly regulate the upper part of the head. Extracting information about this part of the head is challenging due to varying degrees of occlusion by hair. We present a novel approach for modeling the upper head by removing occluding hair and reconstructing the skin, revealing information about the head shape. We introduce three objectives: 1) a dice consistency loss that enforces similarity between the overall head shape of the source and rendered image, 2) a scale consistency loss to ensure that head shape is accurately reproduced even if the upper part of the head is not visible, and 3) a 71 landmark detector trained using a moving average loss function to detect additional landmarks on the head. These objectives are used to train an encoder in an unsupervised manner to regress FLAME parameters from in-the-wild input images. Our unsupervised 3DMM model achieves state-of-the-art results on popular benchmarks and can be used to infer the head shape, facial features, and textures for direct use in animation or avatar creation.

Index Terms: Human-centered computing—Visualization—3D morphable models

1 INTRODUCTION
Blanz and Vetter [3] introduced a general face representation and a principled approach to facial reconstruction. The key idea that any face can be defined as the linear combination of a set of template faces has given rise to 3D Morphable Models (3DMM). Traditionally, optimization-based analysis-by-synthesis methods have been used to infer the 3DMM parameters from a given image. However, this approach requires solving a complex and nonlinear optimization problem to model in-the-wild images. In recent years, morphable models have received increasing attention due to the advent of deep learning. Their nonlinear and deep representations offer a more robust and reliable performance on-in-the-wild images (for a comprehensive overview of 3D morphable models see [41]).

For practical use in 3D avatar creation and animation, we require reconstruction of the full head and not only the frontal face as computed by most of these existing models [8, 11, 12, 32]. Models like DECA [9] and RingNet [28] that are based on the FLAME [22] geometric prior regress complete head shape. However, they train their models on frontal face features and do not account for the top part of the head due to occlusion by hair. Due to this, state-of-the-art models have high accuracy while evaluating with ground truth frontal face scans, but do not have a realistic head shape. As shown in Figure 5, DECA head reconstructions are enlarged compared to the input image. To address these issues we train an encoder that regresses 3DMM parameters from in-the-wild images. We regulate head shape using a dice consistency loss between masks of the input images and output renders. However, the skin mask of the input is not representative of the head shape due to a portion of the head being occluded by hair. To overcome this we develop deep learning-based models that can remove occluding hair and reconstruct the skin, making the person bald. We also quantitatively evaluate our model on the popular NoW benchmark [2], the CoMA [27] dataset, and LYHM [7] dataset.

2 RELATED WORK
2.1 3D reconstruction
The existing monocular reconstruction algorithms can be broadly categorized into optimization-based [17, 30, 31] and learning-based [8, 9, 11, 12, 28]. Optimization-based methods aim to estimate 3DMM parameters by solving a non-linear optimization problem to minimize energy functions between the input and synthesized face image. Such methods are computationally expensive and output only a tightly cropped face mesh, unsuitable for animation.

Learning-based methods utilize deep networks to regress 3DMM parameters. The use of non-linear representation for modeling surpasses classical models and can generalize well to unseen data. Inference speed is magnitudes faster than optimization methods, and training on a large corpus of data makes it robust to occlusions and varying real-world conditions. Learning-based models are usually trained in a supervised [10, 32], weakly supervised [8, 12], or unsupervised [9, 11, 28] fashion. Supervised learning models require 2D-3D paired data in the form of 3D scans or synthetic data. These methods do not generalize well to unseen data due to the low volume of training data that does not capture real-world complexity. Unsupervised approaches like [9, 11, 28] regress 3DMM parameters without the need for explicit supervision. This form of learning eliminates the need for corresponding 3D data by instead minimizing loss functions between the input and rendered image. [9] introduces shape consistency loss which enforces the fact that swapping only the shape parameters between images of the same subject should not affect the rendered image. Unlike our model, none of the existing methods, optimization or deep learning-based, take the upper head into consideration while designing their loss functions.

2.2 Image editing
Generative Adversarial Networks (GAN) [13] learn the underlying patterns in the input data and generate new data that seem to be drawn from the real dataset with the help of an adversarial loss. Image manipulation using GANs gained prominence with the Pix2Pix framework [19] which performed paired translation. However, this model requires paired training data, which is not readily available for hair to bald conversion. CycleGAN [40] uses a novel cycle consistency loss to decipher the underlying style to perform the translation and requires an unpaired collection of photographs in each domain for training.

Another popular method is the latent space exploration of GANs that manipulates images by altering their latent codes. First observed by [29], they succeed in manipulating gender, age, expression, and face pose by learning boundaries between the latent codes of these attributes. GAN inversion [35] is used to encode any image into the
latent code of generative networks, extending latent space editing to in-the-wild images. Although latent space editing is successful on editing attributes such as age and hair color, they struggle with highly correlated attributes.

Image inpainting is the task of filling the missing regions such that the filled pixels are semantically consistent with the rest of the image. Although Inpainting initially supported only regular polygon holes [18], they later grew to support random masks with the help of edge generation [25], and partial convolutions [23]. Recently some works have also used graph neural networks for image manipulation [33, 37] due to their increased overall receptive field leading to better global consistency. The hypergraph-based inpainting approach [33] is highly relevant to our purpose due to its ability to handle long-range contextual information.

2.3 Quantitative evaluation
The NoW benchmark [2] focuses on extreme viewing angles, facial expressions, and partial occlusions. These methods evaluate error metrics only on a frontal crop of the head and not the entire head. However, to properly evaluate the accuracy of head shape reconstruction, we require 360° scans of the subject’s head. For this purpose we propose using the CoMA [27] dataset which provides FLAME registrations of 3D scans of 12 subjects in 12 challenging expressions, and the LYHM [7] dataset which provides FLAME registrations of 1000 subjects with a neutral expression.

3 Approach
The key idea is grounded in the observation that regardless of scale and crop, viewing angles, expression, and occlusion, the shape parameters must remain the same. Even if the upper part of the head is not visible in the input image, the model should accurately estimate the head shape. We enforce this idea by introducing a dice loss between skin masks of the input and rendered output of our model. However, the rendered output is bald, whereas the input has hair occluding the face leading to inconsistent skin masks and inaccurate dice coefficients. Thus, we remove hair from the input image to make a more fair and robust loss computation. In addition to the dice loss, we crop the input at multiple scale factors and shuffle the resulting shape parameters. This forces the encoder to infer complete head shape from facial features.

3.1 Hair removal
We opt for a deep learning approach using generative adversarial models for removing hair on top of the head. We first experiment with popular image translation methods, then with latent space editing, and finally with inpainting methods as described in this section. The results of these methods are consolidated and compared in Section 4.2.1.

3.1.1 Image translation
**Pix2pix:** Pix2Pix [19] requires paired examples to learn the relation between input and output. Due to non-availability of datasets with hair-bald paired examples, we create an artificial dataset in a commercial 3D modeling software [6] and train the Pix2Pix network on this synthetic dataset. This network architecture comprises of a PatchGAN discriminator with a U-Net generator. We also augment training with an additional identity loss to preserve facial characteristics during translation.

**CycleGAN:** The CycleGAN model is trained on an unpaired collection of photographs from each domain (subjects with hair and without) taken from the CelebA dataset [24]. This network consists of a PatchGAN discriminator and a generator with encoder-decoder architecture. We also introduce an identity loss to preserve facial characteristics during translation.

3.1.2 Latent space editing
For each unique d-dimensional latent code \( z \in \mathbb{R}^d \), a GAN generator \( G \) generates a unique image \( G(z) \). [29] observed that this \( z \) vector encodes information about facial characteristics, and modifying the \( z \) vector would result in the modification of those characteristics in the image. Given a separating hyperplane with unit normal vector \( n \in \mathbb{R}^d \) between latent codes corresponding to specific characteristics, we estimate the perturbation magnitude \( \epsilon \) such that \( z_{new} = z + \epsilon n \), where \( z_{new} \) corresponds to latent code with desired attribute. We estimate \( \epsilon \) by trying to increase the predicted score \( \alpha = f(g(z_{new})) \) of desired attribute, where \( f \) is a Resnet [15] based multi-class classifier which predicts probabilities of different attributes present in the image. In our case, we try to maximize the attribute of baldness. We use [39] to obtain latent codes of images of people with hair and without. This set of latent codes was used to train a linear SVM resulting in the desired separating hyperplane. We use [39] to invert the input image into the latent space of StyleGAN [20]. To further improve the quality of the inversion we introduce the M-SSIM loss [34] which additionally gives a quantifiable metric to measure inversion accuracy.

3.1.3 GAN inpainting
We utilize the hypergraph network architecture [33] for inpainting due to the fact that complex relationship among spatial features of an image can be represented effectively by hyperedges that can connect more than two nodes i.e. spatial features, which a traditional CNN struggles to do due to its limited receptive field. Our task is to remove hair using this network, therefore, we denote hair as the missing region or hole which the inpainting network fills with valid pixels(reconstructed skin and background). Given a source image \( I \) and a binary mask \( M \) of the missing region (1 for holes), the inpainting network fills in the missing pixels and gives as output a coarse reconstruction \( I_{coarse} \) further refined using hypergraph layers to give \( I_{fine} \) using a course and refine network respectively. \( I_{coarse} \) is the final output with hair removed and skin reconstructed to be used for training the 3DMM model. This inpainting network is trained on a dataset of bald people overlayed with random hair masks generated from subjects with hair in the CelebA dataset. The training pipeline is shown in Figure 2. Formally, we minimize

\[
L_{total} = L_{hole} + L_{valid} + L_{adv} + L_{ppl} + L_{edge}. \tag{1}
\]

Where, \( L_{hole} \) and \( L_{valid} \) are photometric losses to enforce consistency in the hole region and non-hole region respectively, \( L_{adv} \) is the adversarial loss between the generator(refine network) and discriminator of the GAN model, \( L_{ppl} \) is the perceptual similarity metric [38], and \( L_{edge} \) is the edge consistency loss [25]. These loss functions are defined as,

\[
L_{hole} = ||M \odot (I_{fine} - I_{gt})||_1 + ||M \odot (I_{coarse} - I_{gt})||_1
\]

\[
L_{valid} = ||(1 - M) \odot (I_{fine} - I_{gt})||_1 + ||(1 - M) \odot (I_{coarse} - I_{gt})||_1
\]

\( L_{hole} \) is the loss on the hole pixels values, and \( L_{valid} \) is the loss on the non-pixels values. The adversarial loss \( L_{adv} \) can be formulated as a min-max problem between the inpainting generator network \( G \) which predicts \( I_{fine} \) and the discriminator \( D \).

\[
L_{adv} = \max_{G} \min_{D} \mathbb{E}_{D}[\log(D(I_{gt}, M))] + \mathbb{E}[\log(1 - D(G(I_{in}, M)))] \tag{2}
\]

[38] demonstrated the effectiveness of deep features as a perceptual similarity metric compared to traditional ones like PSNR [16] which fail to account for many nuances of human perception. The perceptual loss is used to encourage a reconstruction to appear more
We evaluate this loss for the final prediction with three landmarks on top of the head in addition to 68 landmarks with valid pixel set to corresponding pixels of 3.2 Landmark detection

We learn to regress FLAME parameters from a single image using an analysis-by-synthesis approach. Given a 2D image I as input we encode the image into FLAME parameters and use it to synthesize a rendered image I_r using the FLAME decoder and a differentiable renderer, and minimize the error between this image and the input image as shown in Figure 1a. The FLAME parameters consists of 100 shape parameters β, 50 expression parameters ψ, 6 pose parameters θ, 50 albedo parameters α, 3 camera parameters c, and 27 lighting I coefficients. The goal of the reconstruction network is to minimize,

\[
L_{full} = L_{link} + L_{pho} + L_{id} + L_{scale} + L_{dice} + L_{enc} + L_{reg},
\]

with landmark loss \(L_{link}\), photometric loss \(L_{pho}\), identity loss \(L_{id}\), scale consistency loss \(L_{scale}\), dice consistency loss \(L_{dice}\), encoder loss \(L_{enc}\), and regularization \(L_{reg}\).

**Landmark loss:** The landmark loss measures the Manhattan distance between ground truth 2D face landmarks \(I\) and corresponding 3D landmarks \(I'\) on the mesh surface projected on the image plane with weights \(ω_i\).

\[
L_{link} = \sum_{i=1}^{71} ω_i ||I_i - I'_i||_1
\]

**Photometric loss:** The photometric loss computes the discrepancy between the input image \(I\) and the rendered image \(I_r\) in regions specified by a binary skin mask \(M_r\) as,

\[
L_{pho} = ||M_r ⊙ (I - I_r)||_1,
\]

where ⊙ denotes the Hadamard product. This skin mask is computed using a segmentation network [1] built upon a Bisenet [36] architecture and trained on the CelebAMask-HQ [21] dataset. Using the skin mask prevents the influence of occlusions on the loss improving the quality of the reconstructed texture.

**Identity loss:** Regardless of expression, pose or illumination, embeddings from the face recognition network [5] map close to each other for the same identity. We quantify the similarity between the
input image and rendered image as the cosine score between their embeddings $f(I)$ and $f(I')$. We define the loss as

$$L_{ad} = 1 - \frac{f(I) f(I')}{||f(I)||_2 \cdot ||f(I')||_2}. \quad (8)$$

Scale consistency loss: Similar to DECA’s shape consistency and RingNet’s ring loss, we use multiple images of the same person to isolate shape parameters from expression and pose. In the scale consistency loss we crop the input image in $N_i$ different ways ranging from a part of the head not being visible, to the entire head being visible as shown in Figure 3b. These crops are used in the scale consistency loss defined below in tandem with multiple views of the input image. We enforce the fact that these shape parameters should be equal by shuffling shape parameters $\beta$ and $\beta_{ib}$ and minimizing the loss between rendered image and input image similar to Equation 9.

$$L_{enc} = \sum_{\beta, \psi, \alpha, \epsilon} ||p_{in} - p_{out}||^2_2. \quad (11)$$

**Regularization:** We define the regularization loss on shape $\beta$, expression $\psi$, and albedo $\alpha$ parameters as

$$L_{reg} = ||\beta||^2 + ||\psi||^2 + ||\alpha||^2. \quad (12)$$

4 EVALUATION
4.1 Qualitative evaluation

Hair removal: Figure 4 compares the outputs of image translation, latent space editing, and GAN inpainting. The Pix2Pix network which is trained on only synthetic data does not replicate the skin tone of the subject. CycleGAN reconstructs the skin with a more suitable skin tone but superimposes estimated skin on top without removing the hair. Light space editing works better than the previous methods for men but does not extend well to women. Further, editing the attribute of baldness causes the attribute of age to increase along with increasing the masculinity of women by introducing facial hair. Inpainting gives the best results both visually and generalizing the best to in-the-wild images. It successfully removes the hair and reconstructs skin with an accurate texture that seamlessly blends into the rest of the image.

Monocular reconstruction: Figure 5 qualitatively compares our method with PRNet [10], 3DDFA-V2 [14], Deep3DRecon [8], RingNet [28], and DECA [9]. We re-optimize the pose and camera parameters of RingNet and DECA to rigidly align the mesh to match the input image. PRNet, Deep3DRecon, and 3DDFA-V2 do not provide a full head representation while DECA and RingNet geometric shapes have an enlarged head compared to the input image. DECA has better facial details due to the generation of a suitable skin tone but superimposes estimated skin on top without removing the hair. Light space editing works better than the previous methods for men but does not extend well to women. Further, editing the attribute of baldness causes the attribute of age to increase along with increasing the masculinity of women by introducing facial hair. Inpainting gives the best results both visually and generalizing the best to in-the-wild images. It successfully removes the hair and reconstructs skin with an accurate texture that seamlessly blends into the rest of the image.

4.2 Quantitative evaluation

Hair removal: We pass $I$ and $I_b$ through the encoder $E_b$ giving shape parameters $\beta_I$ and $\beta_{ib}$. We enforce the fact that these shape parameters should be equal by shuffling $\beta_I$ and $\beta_{ib}$ and minimizing the loss between rendered image and input image similar to Equation 9.

**Encoder loss:** The previous metrics enforced similarity between the rendered and input image. We also minimize the difference between the encoded FLAME parameters of the rendered image $p_{out}$ and the input image $p_{in}$.

$$L_{enc} = \sum_{\beta, \psi, \alpha, \epsilon} ||p_{in} - p_{out}||^2_2. \quad (11)$$

**Regularization:** We define the regularization loss on shape $\beta$, expression $\psi$, and albedo $\alpha$ parameters as

$$L_{reg} = ||\beta||^2 + ||\psi||^2 + ||\alpha||^2. \quad (12)$$

4.2 Hair removal

We develop a binary hair vs bald classifier with a ResNet [15] architecture trained on the CelebA [24] dataset with an accuracy.
of 98%. We then compare the percentage of correctly classified images that were artificially manipulated as bald. These results are summarised in Table 1. The inpainting network achieves the best accuracy with the classifier and is qualitatively better.

4.2.2 Monocular reconstruction

We compare the accuracy of facial reconstruction with publicly available methods, namely DECA [9], 3DDFA-V2 [14], RingNet [28], PRNet [10], 3DMM-CNN [32] and Deep3DRecon [8] using the NoW benchmark [2] in Table 2, and upper head accuracy on the CoMA [27] and LYHM [7] dataset in Table 4.

**NoW benchmark:** Table 2 and the cumulative error plot in Figure 6 compares our results to other popular reconstruction methods on the NoW test set. We have also shown the influence of different cropping levels in Table 3. Although the dice loss offers lower reconstruction error when the entire head is visible, addition of the scale loss results in a more stable reconstruction.

**Full head benchmark:** Our solution aims to solve the problem of accurate reconstruction of the complete head, whereas the NoW benchmark only evaluates metrics on a tightly cropped portion of the face. Thus, we devise an evaluation scheme based on the CoMA [27] and LYHM [7] dataset. We rigidly align the predicted mesh to the refined ground truth similar to the NoW evaluation and calculate the mean square error on a subset of vertices corresponding to the upper head between the predicted mesh and the ground truth. Table 4 summarises the error metrics.

5 CONCLUSION

We address the challenging problem of accurate monocular head reconstruction without 2D-3D supervision. Existing approaches do not explicitly monitor the overall head, causing ambiguity in its shape. To resolve this, we train a stacked hourglass neural network to detect additional landmarks on top of the head for defining its structure. We augment this with a dice consistency loss between the rendered and input images, enabling us to control the fullness of the face and overall shape when detected landmarks do not perfectly align with facial contours. Since the rendered image does not have hair, we require the subject in the input image to be bald for accurate computation of the overlap degree. After experimenting with image translation, latent space editing, and inpainting methods, we utilise a hypergraph-based inpainting network for hair removal and skin reconstruction. This network also learns to reconstruct ears which were previously occluded by hair in the input images. Future work can use the ears in the reconstruction pipeline to improve the 3D pose and shape accuracy. However, this method can produce minor artifacts in the image, especially when long hair obscures the face or in profile images, suggesting the need to train on a more challenging dataset. Further, we introduce a training method to account for different crops of the input image. However, the scale loss improves
stability at the expense of accuracy, suggesting the need for more varied training data. Our model is built upon FLAME [22] and hence, is compatible with DECA’s [9] displacement map generation pipeline. Model weights corresponding to DECA’s detail network are directly transferable to our model. We present a complete framework that regresses accurate head geometry and facial features from monocular in-the-wild images for use in 3D avatar creation.

ACKNOWLEDGMENTS

The first author would like to thank the India Connect program for the opportunity to conduct research at Nanyang Technological University. We would also like to thank S. Sanyal and Yao Feng for their RingNet and DECA pytorch implementation.

REFERENCES

[1] Face parsing using the bisenet architecture. https://github.com/zllrunning/face-parsing.PyTorch, 2018.
[2] NoW challenge, 2019.
[3] V. Blanz and T. Vetter. A morphable model for the synthesis of 3d faces. SIGGRAPH 99 Proceedings of the 26th annual conference on Computer graphics and interactive techniques, 09 2002: doi: 10.1145/311553.311556.
[4] A. Bulat and G. Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230, 000 3d facial landmarks). CoRR, abs/1703.07332, 2017.
[5] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. Vggface2: A dataset for recognising faces across pose and age. CoRR, abs/1710.08092, 2017.
[6] B. O. Community. Blender - a 3D modelling and rendering package. Blender Foundation, 2018.
[7] H. Dai, N. Pears, W. Smith, and C. Duncan. Statistical modeling of craniofacial shape and texture. International Journal of Computer Vision, 128(2):547–571, Nov 2019. doi: 10.1007/s11263-019-01260-7.
[8] Y. Deng, J. Yang, S. Xu, D. Chen, Y. Jia, and X. Tong. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set, 2020.
[9] Y. Feng, H. Feng, M. J. Black, and T. Bolkart. Learning an animatable detailed 3d face model from in-the-wild images, 2021.
[10] Y. Feng, F. Wu, X. Shao, Y. Wang, and X. Zhou. Joint 3d face reconstruction and dense alignment with position map regression network. In ECCV, 2018.
[11] B. Gecer, S. Ploumpis, I. Kotis, and S. Zaferiou. Ganfit: Generative adversarial network fitting for high fidelity 3d face reconstruction. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Jun 2019. doi: 10.1109/cvpr.2019.00125.
[12] K. Genova, F. Cole, A. Maschinot, A. Sarna, D. Vlasic, and W. T. Freeman. Unsupervised training for 3d morphable model reconstruction. CoRR, abs/1806.06099, 2018.
[13] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. Advances in neural information processing systems, 27, 2014.
[14] J. Guo, X. Zha, Y. Yang, F. Yang, Z. Lei, and S. Z. Li. Towards fast, accurate and stable 3d dense face alignment. CoRR, abs/2009.09960, 2020.
[15] K. He, X. Zhang, S. Ren, and J. Sun. Identity mappings in deep residual networks. CoRR, abs/1603.05027, 2016.
[16] A. Horé and D. Ziom. Image quality metrics: Psnr vs. ssim. In 2010 20th International Conference on Pattern Recognition, pp. 2366–2369, 2010. doi: 10.1109/ICPR.2010.579.
[17] A. Ichim, S. Bouaziz, and M. Pauly. Dynamic 3d avatar creation from hand-held video input. ACM Transactions on Graphics, 34:45:1–45:14, 07 2015. doi: 10.1145/2766974.
[18] S. Iizuka, E. Simo-Serra, and H. Ishikawa. Globally and locally consistent image completion. ACM Transactions on Graphics, 36:1–14, 07 2017. doi: 10.1145/3072599.3073659.
[19] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks, 2018.
[20] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks, 2019.
[21] C.-H. Lee, Z. Liu, L. Wu, and P. Luo. Maskgan: Towards diverse and interactive facial image manipulation. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020.
[22] T. Li, T. Bolkart, M. J. Black, H. Li, and J. Romero. Learning a model of facial shape and expression from 4d scans. ACM Transactions on Graphics (TOG), 36:1 – 17, 2017.
[23] G. Liu, F. A. Reda, K. J. Shih, T. Wang, A. Tao, and B. Catanzaro. Image inpainting for irregular holes using partial convolutions. CoRR, abs/1804.07723, 2018.
[24] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In Proceedings of International Conference on Computer Vision (ICCV), December 2015.
[25] K. Nazeri, E. Ng, T. Joseph, F. Z. Qureshi, and M. Ebrahimi. Edgeconnect: Generative image inpainting with adversarial edge learning. CoRR, abs/1901.00212, 2019.
[26] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. CoRR, abs/1603.06937, 2016.
[27] A. Ranjan, T. Bolkart, S. Sanyal, and M. J. Black. Generating 3d faces using convolutional mesh autoencoders. In European Conference on Computer Vision (ECCV), pp. 725–741, 2018.
[28] S. Sanyal, T. Bolkart, H. Feng, and M. J. Black. Learning to regress 3d face shape and expression from an image without 3d supervision, 2019.
[29] Y. Shen, C. Yang, X. Tang, and B. Zhou. Interfacegan: Interpreting the disentangled face representation learned by gans. TPAMI, 2020.
[30] S. Suwajanakorn, I. Kemelmacher-Shlizerman, and S. M. Seitz. Total moving face reconstruction. In ECCV, 2014.
[31] J. Thies, M. Zollhöfer, M. Stammbinger, C. Theobalt, and M. Nießner. Face2face: Real-time face capture and reenactment of rgb videos, 2020.
[32] X. Tran, T. Hassner, I. Masi, and G. G. Medioni. Regressing robust and discriminative 3d morphable models with a very deep neural network. CoRR, abs/1612.04904, 2016.
[33] G. Wadhwa, A. Dhall, S. Murala, and U. Tariq. Hyperrealistic image inpainting with hypergraphs, 2020.
[34] Z. Wang, E. Simoncelli, and A. Bovik. Multiscale structural similarity for image quality assessment. In The Thirty-Seventh Asilomar Conference on Signals, Systems, Computers, 2003. vol.2, 2003. doi: 10.1109/ACSSC.2003.1292216.
[35] W. Xia, Y. Zhang, Y. Yang, J. Xue, B. Zhou, and M. Yang. GAN inversion: A survey. CoRR, abs/2101.05278, 2021.
[36] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation, 2018.
[37] K. Zhang, T. Li, S. Shen, B. Liu, J. Chen, and Q. Liu. Adaptive graph convolutional network with attention graph clustering for co-saliency detection. CoRR, abs/2003.06167, 2020.
[38] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. CoRR, abs/1801.03924, 2018.
[39] J. Zhu, Y. Shen, D. Zhao, and B. Zhou. In-domain gan inversion for real image editing. In Proceedings of European Conference on Computer Vision (ECCV), 2020.
[40] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks, 2020.
[41] M. Zollhöfer, J. Thies, P. Garrido, D. Bradley, T. Beeler, P. Pérez, M. Stammbinger, M. Nießner, and C. Theobalt. State of the art on monocular 3d face reconstruction, tracking, and applications. Computer Graphics Forum, 37:523–550, 05 2018. doi: 10.1111/cgf.13382.