3D-Guided Face Manipulation of 2D Images for the Prediction of Post-Operative Outcome After Cranio-Maxillofacial Surgery

Robin Andlauer, Andreas Wachter, Matthias Schaufelberger, Frederic Weichel, Reinald Kühle, Christian Freudlsperger, and Werner Nahm

***Abstract*— Cranio-maxillofacial surgery often alters the aesthetics of the face which can be a heavy burden for patients to decide whether or not to undergo surgery. Today, physicians can predict the post-operative face using surgery planning tools to support the patient's decision-making. While these planning tools allow a simulation of the post-operative face, the facial texture must usually be captured by another 3D texture scan and subsequently mapped on the simulated face. This approach often results in face predictions that do not appear realistic or lively looking and are therefore ill-suited to guide the patient's decision-making. Instead, we propose a method using a generative adversarial network to modify a facial image according to a 3D soft-tissue estimation of the post-operative face. To circumvent the lack of available data pairs between pre- and post-operative measurements we propose a semi-supervised training strategy using cycle losses that only requires paired open-source data of images and 3D surfaces of the face's shape. After training on "in-the-wild" images we show that our model can realistically manipulate local regions of a face in a 2D image based on a modified 3D shape. We then test our model on four clinical examples where we predict the post-operative face according to a 3D soft-tissue prediction of surgery outcome, which was simulated by a surgery planning tool. As a result, we aim to demonstrate the potential of our approach to predict realistic post-operative images of faces without the need of paired clinical data, physical models, or 3D texture scans.**

***Index Terms*— Cranio-maxillofacial surgery, post-operative face, surgery planning, face manipulation, face editing, generative adversarial network, CycleGAN, cycle loss, GAN, unsupervised learning, 3D morphable model.**

I. Introduction

Cranio-maxillofacial surgery is a common treatment of temporomandibular disorders or skeletal malocclusion. Besides improvement of function, this surgical intervention often changes the aesthetics or identity of the face which can be a heavy burden for the patient. To support the patient's decision-making in favor of or against surgery, having a prediction of the patient's face after surgery is highly desirable. At present, physicians can predict the virtual post-operative face using surgery planning tools like IPS CaseDesigner® [1] or Dolphin 3D® [2]. These surgery planning tools typically require a tomography scan of the patients face which includes both segmented soft-tissue and segmented bone structure. The surgery planning tool allows the physician to virtually manipulate the bone structure e.g. to cut and move the jaw and subsequently predict the deformation of the soft-tissue using e.g. finite element methods [3]–[5] or mass tensor models [6]. In a next step, the texture of the face has to be predicted to allow a rendering of the post-operative face. For this, a 3D scan of the facial texture must be captured by a 3D camera system, wrapped on the virtual pre-operative face, and subsequently interpolated according to the predicted deformation of the soft-tissue [7]. This procedure to predict the post-operative texture has multiple disadvantages:

1) The procedure requires a 3D texture scanner which might not be available at every clinical site. In such a case, patients can be only provided with a single-color prediction of the post-operative face.
2) The quality of the mapped texture of the face is limited by the registration accuracy and the resolution of both the texture and the tomography scans.
3) Existing methods to translate the pre-operative texture to the predicted post-operative soft-tissue e.g. interpolation might be unsuitable to predict realistic textures since they do not consider illumination or skin properties.

In practice, these disadvantages often result in predictions of the post-operative face that do not look realistic or lively looking and are therefore ill-suited to support the patient's decision-making.

In this study, we propose a novel deep learning-based idea to directly predict a realistic 2D image of the post-operative face given only a 2D image of the patient before surgery and a 3D simulation of the post-operative soft-tissue. In other words, we hypothesize that the current method to capture, wrap and

Manuscript received September 2, 2020; revised March 17, 2021 and May 5, 2021; accepted June 26, 2021. Date of publication July 15, 2021; date of current version August 24, 2021. This work was supported by HeiKA funding. The associate editor coordinating the review of this manuscript and approving it for publication was Mr. Chuang Gan. *(Robin Andlauer and Andreas Wachter are co-first authors.) (Corresponding author: Andreas Wachter.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Ethical Committee of the Medical Faculty of the University of Heidelberg under Application No. S-039/2016.

Robin Andlauer, Andreas Wachter, Matthias Schaufelberger, and Werner Nahm are with the Institute of Biomedical Engineering, Karlsruhe Institute of Technology (KIT), 76131 Karlsruhe, Germany (e-mail: publications@ibt.kit.edu).

Frederic Weichel, Reinald Kühle, and Christian Freudlsperger are with the Department of Oral and Cranio-Maxillofacial Surgery, Heidelberg University Hospital, 69120 Heidelberg, Germany.

Data is available on-line at https://github.com/KIT-IBT/3D-Guided-Face-Mani.

Digital Object Identifier 10.1109/TIP.2021.3096081

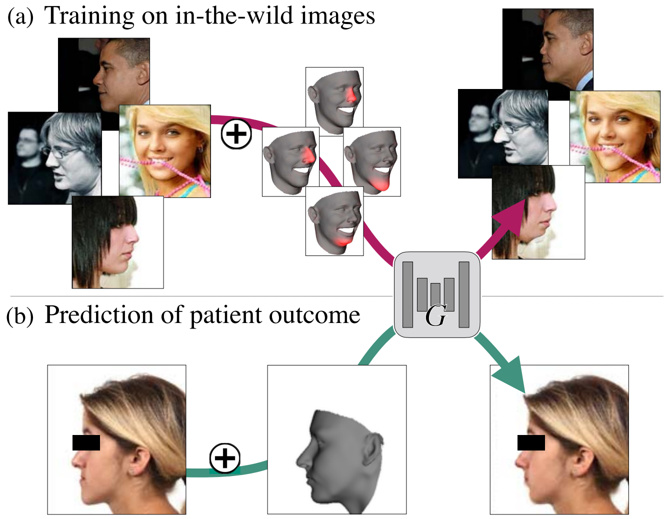

(a) Training on in-the-wild images
(b) Prediction of patient outcome
G

Fig. 1. Overview of our approach to predict the post-operative face. In (a) we first train our model to predict various face modifications (marked in red) at the chin and the nose using a CycleGAN strategy. After training, we are able to apply local modifications to a 2D image of a face as seen on the right. Afterwards, we transfer our model to (b) where we use our trained model to predict the face of a patient after cranio-maxillofacial surgery. More precisely, we use a 2D image of the pre-operative face as shown on the left and a 3D simulation of the post-operative surface as shown in the middle to generate a prediction of the post-operative face as shown on the right. Hereby, our approach requires neither clinical data for training nor 3D texture scans for inference.

interpolate the 3D texture can be replaced by a neural network to make a realistic prediction of the post-operative face and thus, does not require a 3D texture scanner. Our main contribution is a conditional generative adversarial network (cGAN) for post-operative face prediction, which translates a 2D image of the pre-operative face of a patient to a 2D image of the post-operative face. Compared to previous approaches to predict the post-operative face [1]–[7], we propose a deep learning-based solution i.e. we aim to train a suitable model directly from data. However, acquiring large numbers of corresponding image pairs between pre- and post-operative faces of cranio-maxillofacial surgery is difficult and often includes data with large time gaps of several months between images of a pair due to the long healing phase of the swelling. To bypass this lack of feasible training data we propose a semi-supervised CycleGAN [8] strategy to train our model on non-clinical data and subsequently transfer our model to predict the post-operative face as shown in Fig. 1.

More precisely, we first train a modified CycleGAN on “in-the-wild” images of the 3DDFA dataset [9] where we aim to manipulate distinct local face properties of 2D images such as changing the size of the chin or the nose. In contrast to recent state-of-the-art models used to manipulate facial properties, we cannot describe the desired manipulation or a surgery plan by discrete attributes (e.g. brown/blond/black hair color in StarGAN [10], [11]), domain transfer between two images [11], [12], or concealed representations in latent space [13], [14]. Instead, we define the geometric shape of the manipulation using a 3D surface template of the face which enables a precise manipulation according to the 3D shape of the face. Using this 3D template, we use a statistical model to generate distinct local face modifications and pass these locally modified 3D faces together with an unmodified 2D image to our model to predict a face that comprises the desired local manipulation. We then transfer our model to the second stage of our study where we predict the post-operative face for two different views of four clinical subjects that underwent cranio-maxillofacial surgery. To create such a prediction, we simulate a 3D face template of the post-operative face without texture using a surgery planning tool and pass it together with an image of the pre-operative face to our model. As a result, we demonstrate the reasonability of our approach to train on non-clinical data and subsequently predicting realistic 2D images of the post-operative face. Based on these promising first results, we believe that our approach has a high potential as a future tool for post-operative face prediction. Compared to previous approaches [1]–[7], our approach does not require 3D texture scans or registration procedures for inference, nor do we need sparsely available clinical data, physical models or detailed surgery expertise for training.

II. Related Work

Our study aims to mainly contribute to the state-of-the-art in image processing for predicting the post-operative face. In the following, we describe the state-of-the-art for the prediction of post-operative outcome as well as the state-of-the-art for manipulating 2D images of faces. Moreover, we highlight the differences of previous work compared to our study.

A. Post-Operative Face Prediction

Previous research studies [5], [7], [15]–[17], on the prediction of the post-operative face as well as commercially available surgery planning tools [1], [2] are mainly focused on the needs of orthodontics i.e. the planning of bone structures and the prediction of the facial soft-tissue. As a result, the prediction of the post-operative texture is often neglected or replaced by a texture of constant skin color [5], [15] which is poorly suited to guide the patient’s decision whether or not to undergo surgery. On the other hand, commercial planning software such as IPS CaseDesigner® [1] or Dolphin 3D® [2] as well as Harris *et al.* in [16] and Premjani *et al.* [18] offer the prediction of the post-operative texture based on a 3D picture of the pre-operative face. Hereby, the soft-tissue and the bone structure are typically extracted from a cone beam computed tomography (CBCT) scan while the facial texture is captured using a 3D stereo camera system [1], [2], [7], [16], [18]. The 3D texture is then registrated and wrapped on the segmented soft-tissue. However, this procedure is both time-consuming and potentially inaccurate since the registration of the texture must typically be accomplished by either surface matching algorithms using manually annotated landmarks [7], [18] or manual alignment [16]. To overcome this registration problem, other studies have proposed a simultaneous data acquisition of the stereo camera scan and the CBCT scan [17], [19]. However, such stereo photogrammetry systems are expensive to acquire and seldomly available since they offer hardly any additional benefit for clinical diagnostics. Once the texture is wrapped on the soft-tissue of the face, available surgery planning tools allow the physician to virtually cut and move the bone structure of the face and subsequently simulate the deformation of the corresponding soft-tissue. Afterwards, the texture of the pre-operative face must be interpolated according to the simulated soft-tissue deformation to enable a rendering of the predicted post-operative face. This procedure to simulate the post-operative texture typically does not result in lively looking and realistic rendering of faces (compare e.g. [16], [17]) since the texture quality is limited by the resolution of the stereo camera system, the resolution of the soft-tissue scan, and the registration accuracy. Additionally, the interpolation method to manipulate the pre-operative texture according to the soft-tissue deformation does not account for illumination properties of the skin or the preservation of high-frequency details which might further reduce image quality. In contrast, we propose a GAN-based neural network to directly manipulate a 2D image according to a 3D plan of the simulated post-operative soft-tissue. To the best of our knowledge, using neural networks to predict the post-operative face has never been proposed before. As its most important advantage, our approach neither requires acquisition nor registration of 3D texture scans. With regard to the impressive results of recent GANs to generate and manipulate fine-detailed and realistic images of faces in high-resolution, we further hypothesize that a GAN-based approach is able to generate more realistically-looking images of the post-operative face compared to traditional approaches and therefore, might be better suited to guide the patient's decision-making before surgery.

B. Face Manipulation of 2D Images

In recent years, GANs have shown remarkable success in generating and manipulating 2D images of faces. To manipulate a face according to a desirable attribute, numerous studies have been proposed for both purely generative models and cGANs. To enable a controlled manipulation of the face, the desired manipulation has to be represented as an interpretable input to the model. To achieve this, Guan [20] and Liu *et al.* [13] both found representations in the input feature vector of GANs to manipulate desired properties of the face image. In contrast, He *et al.* (AttGAN) [14] defined the manipulation by using both discrete attributes as well as a feature vector in latent space to manipulate facial properties of an image. Alternatively, Bao *et al.* [12] and Shen and Liu [21] trained a cGAN to swap key facial properties between two images which enabled manipulation of e.g. expression, illumination, pose, wearing sunglasses, or having beards. Also, Bansal *et al.* (RecycleGAN) [22] proposed a CycleGAN [8] to transfer facial expressions of video data from one person to another person. Closely related to this work, Choi *et al.* (StarGAN) [10] adapted a CycleGAN strategy and defined the manipulation information by a vector of discrete attributes like hair color, gender or age to manipulate faces in 2D. Most recently, Choi *et al.* [11] released StarGAN v2 which receives both discrete attributes and style information from another image to manipulate faces in a 2D image. As seen above, all the described studies to modify faces in 2D images defined the modification information by either abstract features in latent space, information extraction by transferring properties of other images, or discrete attributes. However, none of the above studies manipulated 2D images according to a 3D plan of the face. Consequently, their approaches would be unsuitable to manipulate a face according to a precise and individual surgery plan. In contrast, we propose a model which receives a representation of a 3D surface mesh of the face to define the manipulation of the 2D face. Hereby, our representation of the face modification is continuous, easily interpretable and enables a precise definition of the targeted geometrical shape of the face. To the best of our knowledge, such a representation to manipulate distinct properties of a face according to a 3D plan of facial shape has never been proposed before.

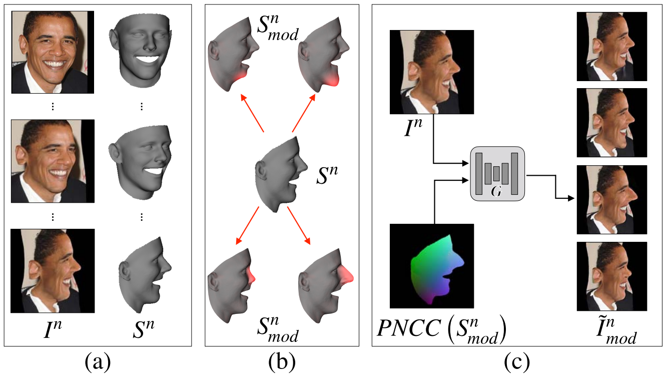


Fig. 2. Pre-processing of the training data. (a) shows an examples of the 300W-LP dataset with in-plane face rotations around the vertical axis with the angle θ by Zhu *et al.* [9]. Hereby, the upper image shows the original image and the lower two images show the augmentated images. Additionally, the estimated 3D shape S^n of the face was also given by the dataset. (b) shows all four face modifications S^{mod} that we applied to S^n to create the modified faces S^n_{mod} (c) shows the inputs and different outputs of the neural network G. Hereby, G received an image I^n and a PNCC projection of the modified 3D shape S^n_{mod} as input. The model then predicted the desired modification in the image $\tilde{I}^n_{mod}$.

III. Methods

The goal of this study was to train a single generator G which receives a 2D image of a face and a modified 3D shape of a face as inputs. Then, the model generates a 2D image of a face that yields the desired modification as an output. More precisely, let I^n be an image of a person's face n and S^n be the corresponding estimation of the 3D shape of the same person's face as shown in Fig. 2 (a). Subsequently, we applied a local modification S_{mod} to every unmodified 3D shape S^n in our dataset to create a locally modified 3D shape $S^n_{mod} = S^n + S_{mod}$.

As a proof-of-concept, we applied four distinct modifications S_{mod} to each face in this study: an increased size of the chin, an increased size of the nose, a decreased size of the chin, and a decreased size of the nose as seen in Fig. 2 (b). Technically, this approach can be extended to other deformations of the face (e.g. mouth or head modifications) as well. Next, we trained a neural network G to apply the modification described by S^n_{mod} on the original image I^n which was supposed to result in a modified image $\tilde{I}^n_{mod}$:

$$\tilde{I}^n_{mod} = G(I^n, S^n_{mod}) \tag{1}$$

For example, this might be an image of a face with an enlarged nose as seen in Fig.2 (c). To train such a model, we utilized the corresponding image I^n and 3D shape S^n pairs from the open-source "in-the-wild" 300W-LP dataset [9] and propose a semi-supervised training strategy inspired by CycleGANs [8]. Since the ground-truth of the modified faces I^n_{mod} are unknown for these "in-the-wild" images, we instead propose a training strategy that leverages four sources of a-priori knowledge to formulate our training objective:

1) a reconstruction loss as introduced by Zhu *et al.* [8]
2) knowledge of the statistics of real world images of faces via an adversarial discriminator,
3) a learned mapping to translate a 2D image to a 3D shape of a face, and
4) information of the approximate location of the local modification in the image.

In the following, the applied local modification S_{mod} and the objectives for training are described in more detail.

A. Local Face Modifications

The ultimate incentive of training G was to create a model which ultimately can predict a 2D image of a patient's face after cranio-maxillofacial surgery. Generally, cranio-maxillofacial surgery does not only affect the appearance of the jaw but other regions of the face as well e.g. the nose and the mouth for the treatment of cleft palates. Therefore, we aimed to demonstrate that our model can be trained on arbitrary regions of the face. In this preliminary study we chose to train our model on size variations of the nose and the chin. These local regions of the face have the advantage that they are easily recognizable in almost all images of faces and in most head positions. For training, the applied modification was required to be automatically applicable and physically plausible i.e. that the existence of S^n_{mod} in the real world was theoretically possible. To achieve this, we expressed each 3D face S^n and each local modification S_{mod} as a parameter vector of the BFM2009 [23] statistical point distribution model. To find such local modifications S_{mod} we annotated different local regions of the 3D face template of the BFM2009. We then implemented an optimization algorithm to find parameter vectors that result in a maximal deformation of the annotated region while minimally deflecting all other regions of the face (see Appendix A for more details). Next, we scaled the computed parameter vectors in negative and positive direction until the deformation of the desired region was maximally deflected without being unrealistic as judged by subjective inspection. The resulting four local deformations S_{mod} can be seen in Fig. 2 (c). During training, we randomly drew one of these four modifications for each sample and applied it to the estimated 3D shape of each unmodified shape S^n of the dataset to create a modified 3D face S^n_{mod}:

$$S^n_{mod} = S^n + S_{mod} \tag{2}$$

B. Objectives

1) Image Reconstruction Loss: To enforce the preservation of the identity of the face in $\tilde{I}^n_{mod}$, we minimized the "identity reconstruction" loss $\mathcal{L}_{I-rec}$ where we aimed to reconstruct the original image I^n from the predicted modified image $\tilde{I}^n_{mod}$:

$$\mathcal{L}_{I-rec} = \mathbb{E}_{I^n,\, \tilde{I}^n_{mod},\, S^n}\left[\mathcal{L}_{Perceptual}\left(I^n, G\left(\tilde{I}^n_{mod}, S^n\right)\right)\right] \tag{3}$$

As seen in the equation, we calculated the image distance $\mathcal{L}_{Perceptual}$ to compare the original image I^n with the reconstructed image $\tilde{I}^n = G\left(\tilde{I}^n_{mod}, S^n\right)$ as illustrated in Fig. 3. The incentive of L_{I-rec} was to ensure that G only changes the geometric shape in $\tilde{I}^n_{mod}$ without modifying properties like skin color, facial hair or other facial attributes independent from facial shape that are required to translate back to the original image I^n. Consequently, these shape independent properties would have to be present in $\tilde{I}^n_{mod}$ to achieve a perfect reconstruction score. As stated in the original CycleGAN paper [8], Zhu *et al.* struggled to translate between images which required geometric changes (e.g. translate dogs to cats). As a possible solution, Gokaslan *et al.* [24] proposed the use of a perceptual loss instead of a pixel-wise loss and achieved convincing results to translate between geometrically changing images using CycleGANs. Motivated by these results, we also used a perceptual loss $\mathcal{L}_{Perceptual}$ [25] to compare between I^n and $\tilde{I}^n$. During training we had to prevent G from learning an "arranged" encoding of these properties in $\tilde{I}^n_{mod}$ and learning a specialized decoder to reconstruct $\tilde{I}^n$. To impede the learning of such an arranged encoding, we predicted $\tilde{I}^n_{mod}$ with G and then froze the weights of G for the reconstruction of the original image $\tilde{I}^n$ i.e. all gradients induced by the second forward pass were not considered for updating G as indicated in Fig. 3.

2) Shape Reconstruction Loss: To enforce a face manipulation in $\tilde{I}^n_{mod}$ we aimed to reconstruct the modified input shape S^n_{mod} from $\tilde{I}^n_{mod}$. For this, we first trained another neural network G_S to predict the 3D shape of an image: $\tilde{S}^n = G_S(I^n)$. After training G_S, we froze the model weights of G_S and optimized the weights of G to minimize the distance $\mathcal{L}_{S-rec}$ between the input modification S^n_{mod} and the reconstructed shape prediction $\tilde{S}^n_{mod} = G_S(\tilde{I}^n_{mod})$:

$$\mathcal{L}_{S-rec} = \mathbb{E}_{S^n_{mod},\, \tilde{I}^n_{mod}}\left[\mathcal{L}_{Shape}\left(S^n_{mod}, G_S\left(\tilde{I}^n_{mod}\right)\right)\right] \tag{4}$$

with $\mathcal{L}_{Shape}$ being a distance metric (described below) between S^n_{mod} and $G_S(\tilde{I}^n_{mod})$. To calculate $\mathcal{L}_{S-rec}$, we first trained G_S to reach convergence using the image-shape pairs (I^n, S^n) of the 300W-LP dataset [9] and minimized the prediction error $\mathcal{L}_{G_S}$ in the shape domain:

$$\mathcal{L}_{G_S} = \mathbb{E}_{I^n, S^n}\left[\mathcal{L}_{Shape}\left(S^n, G_S\left(I^n\right)\right)\right] \tag{5}$$

Hereby, we assumed that the estimated 3D shape S^n of the 300W-LP dataset of each image was the ground-truth. Estimating a 3D shape of a face from a single 2D image is a highly ill-posed problem that is yet to be resolved. To solve this estimation task current state-of-the-art studies propose either the use of iterative template fitting approaches [23], [26], [27] or regression approaches using neural networks [9], [28]–[31]. In initial experiments we considered openly available algorithms or neural networks to serve as G_S and calculate $\mathcal{L}_{S-rec}$. However, we concluded that iterative algorithms [23] are unfeasible for backpropagation and

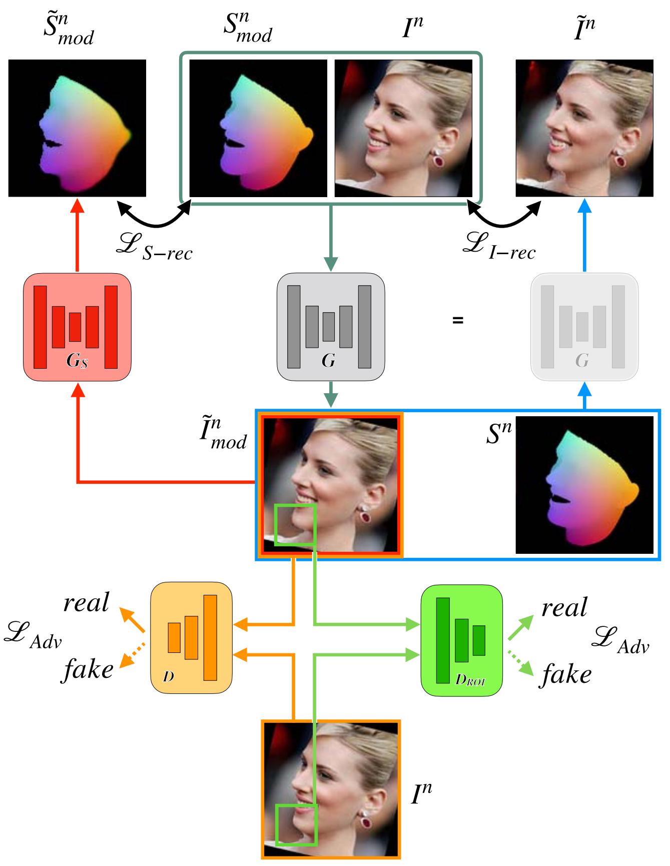


Fig. 3. Schematic overview of our training strategy. As input, G receives an image I^n and a modified shape S^n_{mod} represented as a PNCC and predicts a modified image $\tilde{I}^n_{mod}$. To the left, G_S estimates the projected shape $\tilde{S}^n_{mod}$ from $\tilde{I}^n_{mod}$ which is then compared with the input shape S^n_{mod} to enforce the visibility of the modified shape in $\tilde{I}^n_{mod}$. On the right, G is supposed to reconstruct the original image I^n from the modified prediction $\tilde{I}^n_{mod}$ and the unmodified shape S^n. Notably, the gradients of the right-side pass were not considered for updating G during training. On the bottom, the two adversarials D and D_{Roi} aim to distinguish between generated predictions by G (fake data) and images from our dataset (real data). Hereby, D received all images at full resolution (128 × 128 pixels) while D_{Roi} received all images at various resolutions centered around the local modification (here: 32 × 32 pixels).

publicly available neural networks required either too much GPU RAM [28] or were locally too inaccurate [9], [30] to be used in our CycleGAN setup to calculate $\mathcal{L}_{S-rec}$. Compared to these previous studies, we aimed to facilitate the estimation task for G_S by predicting only a projection of the 3D shape. For the 3D shape we used the projected normalized coordinate code (PNCC) proposed by Zhu *et al.* [9]. To calculate the PNCC we first converted the XYZ coordinates of the mean face of the BFM2009 to the RGB color range by normalizing the coordinate range to [0, 1]. We mapped these colors on the 3D shape S^n and subsequently calculated the projection to the image plane using the projection parameters of S^n as seen in Fig. 2 (c). By using a PNCC to represent the 3D shape of a face we attempted to facilitate the estimation task for G_S and additionally, we changed the translation task of our CycleGAN to a 2D problem which saved GPU RAM and enabled the use of established 2D CNN architectures. Notably, we also split the prediction task of G_S by separately predicting a color map of the PNCC and a mask of the PNCC which in practise, appeared to strongly increase convergence speed when training G_S. To calculate the distance metric between the PNCCs $\mathcal{L}_{Shape}$, we calculated the binary cross entropy $\mathcal{L}_{CE}$ between the masks and the L1 norm $\mathcal{L}_1$ between the color maps:

$$\mathcal{L}_{Shape}(S,\tilde{S})=\mathcal{L}_{CE}(S_{Mask},\tilde{S}_{Mask})+\lambda\mathcal{L}_1(S_{Color},\tilde{S}_{Color}) \quad (6)$$

Hereby, with $\lambda = 10$ and we masked $\tilde{S}_{Color}$ with the predicted PNCC mask $\tilde{S}_{Mask}$. In the following, we name S^n or S^n_{mod} and mean the PNCC representation of the 3D face.

3) Adversarial Loss: To restrict G to only predict realistic images $\tilde{I}^n_{mod}$ we used an adversarial loss. To calculate the adversarial loss $\mathcal{L}_{Adv}$ we randomly drew images I^n from the real data distribution $\mathcal{P}_R$ and modified images $\tilde{I}^n_{mod}$ from the fake data distribution $\mathcal{P}_G$ generated by G and random modifications S^n_{mod}. Thereafter, we alternately approximated the Wasserstein-distance by training a evaluator D and minimizing the estimated Wasserstein-distance by optimizing G according to WGAN theory [32]. To enforce local 1-Lipschitz continuity in D we adopted the gradient penalty loss (WGAN-GP) by Gulrajani *et al.* [33]:

$$\begin{aligned}\mathcal{L}_{Adv} &= \mathbb{E}_{I^n}\left[D\left(I^n\right)\right]-\mathbb{E}_{\tilde{I}^n_{mod}}\left[D\left(\tilde{I}^n_{mod}\right)\right]\\ &\quad -\lambda_{GP}\mathbb{E}_{\dot{I}}\left[\left(\left\|\nabla D\left(\dot{I}\right)\right\|_2-1\right)^2\right] \quad (7)\end{aligned}$$

with $\dot{I}$ being a linear interpolation between an image pair $(I^n, \tilde{I}^n_{mod})$ and $\lambda_{GP} = 10$. Hereby, D aimed to maximize $\mathcal{L}_{Adv}$ while G aimed to minimize $\mathcal{L}_{Adv}$. To increase convergence speed and image quality, we implemented a multi-scale discriminator setup by training a second evaluator D_{Roi} on a cropped region of the local modification as seen in Fig. 3. To automatically compute this region of interest, images I^n and $\tilde{I}^n_{mod}$ were cropped around the location of the modification i.e. an annotated center point of the nose or the chin. To find the approximate center in the prediction of the modified face $\tilde{I}^n_{mod}$ we projected the center-point of the modified 3D shape S^n_{mod} on the image plane. During training of G and D_{Roi}, we varied the size of these regions of interest between 16 × 16 pixels and 48 × 48 pixels before presenting them to our second evaluator D_{Roi} (see Section III-D for further details). As a result, the use of this second discriminator appeared to significantly increase convergence speed and the image quality of the predictions $\tilde{I}^n_{mod}$.

C. Datasets

For training we used the 300W-LP dataset by Zhu *et al.* [9] which comprises corresponding pairs of 2D images I^n of faces, estimated parameters of the BFM2009 [23] statistical point distribution model, and projection parameters of each face. For augmentation, Zhu *et al.* rotated and flipped all faces in-plane around the vertical axis to comprise more training samples with high degrees of face rotations θ as seen in Fig. 2 (a). These in-plane rotations resulted in 300 575 image and 3D shape pairs with a baseline of 7690 independent pairs. For validation we excluded 8 baseline pairs before augmentation which resulted in 112 pairs after augmentation.

Additionally, we rotated each image between −90° and 90° for further augmentation during training. Lastly, we cropped the original image resolution of 450 × 450 pixels around the center of the face to a resolution of 315 × 315 pixels and subsequently rescaled each image to a resolution of 128 × 128 pixels due to memory constraints of our GPU during training. For testing we used the AFLW2000 dataset by Zhu *et al.* [9] which include 2000 images, fitted 3D shapes, and projection parameters derived using the same semi-automatic template fitting approach by Paysan *et al.* [23] as the 300W-LP dataset.

D. Implementation Details

1) Training Strategy: In a first step, we trained the shape estimator G_S on the 300W-LP dataset by minimizing $\mathcal{L}_{G_S}$:

$$\min_{G_S} \mathcal{L}_{G_S} \tag{8}$$

We used Adam [34] for optimization with $\beta_1 = 0.5$, $\beta_2 = 0.999$, a batch-size of 32, and a constant learning rate of $lr = 10^{-4}$ over the first 150 000 iterations. Then we linearly decreased lr to zero over another 150 000 iterations which took approximately 2 days on a Nvidia RTX2080ti. After training G_S, the mean absolute pixel-wise error was $L_1 = 0.015$ and the cross entropy loss was $L_{CE} = 0.160$ on the 300W-LP dataset and $L_1 = 0.072$ and $L_{CE} = 0.338$ on the AFLW2000 dataset. Next, we trained our CycleGAN by updating the evaluators D and D_{Roi} alternatingly on every iteration using the objective function

$$\max_{D} \mathcal{L}_{Adv} \tag{9}$$

while updating the generator G every fifth iteration using the objective function:

$$\min_{G} \mathcal{L}_G = \lambda_1 \mathcal{L}_{I-rec} + \lambda_2 W \mathcal{L}_{S-rec} + \lambda_3 \mathcal{L}_{Adv,D} + \lambda_4 \mathcal{L}_{Adv,D_{Roi}} \tag{10}$$

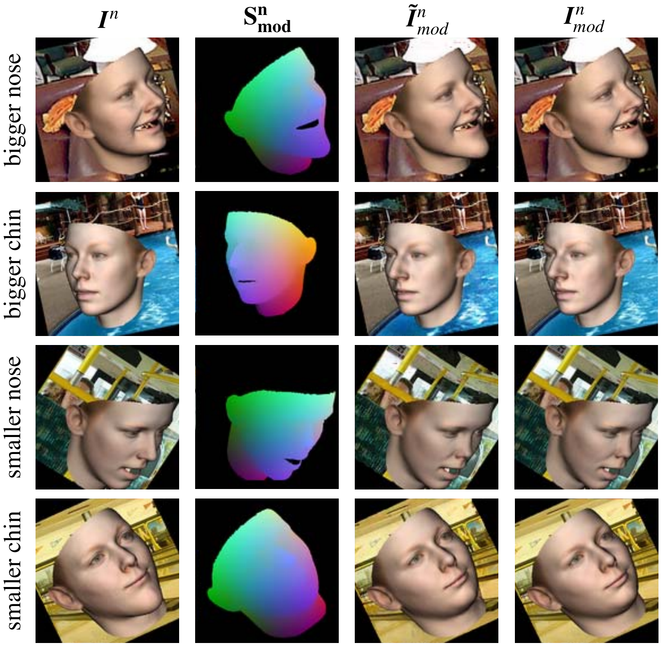


Fig. 4. Four randomly chosen images of the synthetic dataset used for pre-training G. Each row shows one of the four image modifications that were considered in this study. The figure shows synthetically generated samples of the training set after training G over 1 million iterations. The columns show the input image I^n, the modified input shape S^n_{mod}, the prediction $\tilde{I}^n_{mod}$ of G and the synthetic ground truth I^n_{mod}.

with $\lambda_1 = 10$, $\lambda_2 = 75$, $\lambda_3 = 1$, $\lambda_4 = 100$. Additionally, we weighted the shape reconstruction loss $\mathcal{L}_{S-rec}$ more heavily at pixels close to the center of the modification by multiplying the error $\mathcal{L}_{S-rec}$ at each pixel with a 128 × 128 pixel weight map W. This weight map W was calculated by projecting the Euclidean distance of each vertex in 3D between S^n and S^n_{mod} and subsequent normalization between zero and one. Using this weight map W, we aimed to both increase convergence speed and facilitate the prediction task by weighting shape reconstruction errors more lightly at regions of the neck, the forehead, and the ear. For optimization we used Adam with $\beta_1 = 0.5$, $\beta_2 = 0.999$ and a batch-size of 16. We trained G, D, and D_{Roi} with a learning rate of $lr = 10^{-5}$ over 1 million iterations. For initial experimental runs, we experienced heavy difficulties to stabilize the training and in general, we observed slow convergence speed and poor image quality of the modifications. To achieve a better initialization, we pretrained our model over another 1 million iterations using supervised learning on a synthetic dataset which comprised corresponding ground-truth images I^n_{mod}. To create this synthetic dataset we used the OpenGL library to render random faces using the BFM2009 face model [23], random expressions that we randomly drew from the 300W-LP dataset, and random background images from the indoor dataset by Quattoni and Torralba [35]. Example images of our synthetic pretraining can be seen in Fig. 4. This dataset enabled supervised learning on synthetic face templates to pretrain our model. For this, we optimized $\mathcal{L}_G$ in (10) except that we replaced $\mathcal{L}_{I-rec}$ with a synthetic loss $\mathcal{L}_{Syn}$:

$$\mathcal{L}_{Syn} = \mathbb{E}_{I^n_{mod}, \tilde{I}^n_{mod}} \left[\mathcal{L}_1 \left(I^n_{mod}, \tilde{I}^n_{mod} \right) \right] \tag{11}$$

After pre-training our model on synthetic data we trained our model 3DDFA dataset by linearly increasing the roi-size of D_{Roi} from 16 × 16 pixels to 48 × 48 pixels over the first 500 000 iterations. For the remaining 500 000 iterations we used a random uniform roi-size between 32 × 32 pixels and 48 × 48 pixels. In total, we trained our model for 1 million iterations on the synthetic data and another 1 million iterations on the 300W-LP dataset which took approximately fifteen days on a Nvidia RTX2080ti.

2) Network Architectures: The detailed architectures of our models are given in Appendix B. For the generator G we used the tiramisu U-Net [36]. G received six channels with an image resolution of 128 × 128 pixels as input which comprised the unmodified image I^n as well as the PNCC of the modified 3D shape S^n_{mod}. The output of G comprised three RGB channels for the predicted modified image $\tilde{I}^n_{mod}$. For the discriminators we used PatchGAN [37] architectures. For the shape estimator G_S we used another tiramisu U-Net which received I^n as input and predicted the projected PNCC

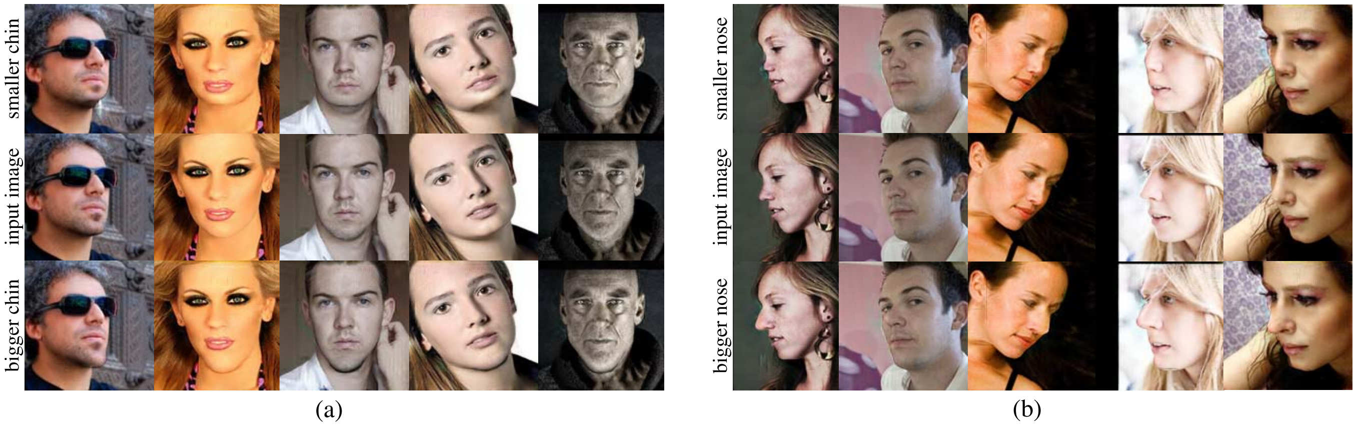


Fig. 5. Selected predictions on the AFLW2000 dataset. The middle row shows the original images I^n which were used as input for G. The top rows show predictions of a smaller chin (a) and nose (b), respectively. The bottom rows show predictions of a larger chin (a) and nose (b), respectively.

of S^n. As described in Section III-B.2, G_S predicted both a mask and a color map of the PNCC. Therefore, the output of G_S comprised five output channel: two channels for the mask (background and face pixels) and three channels for the color values of the PNCC.

IV. Experiments and Results

We conducted three experiments to evaluate the performance of our model G on both "in-the-wild" images and on a clinical example. Hereby, we evaluated G qualitatively on selected samples of the AFLW2000 dataset in experiment IV-A and quantitatively in experiment IV-B. Lastly, we aimed to predict the post-operative face using G in experiment IV-C.

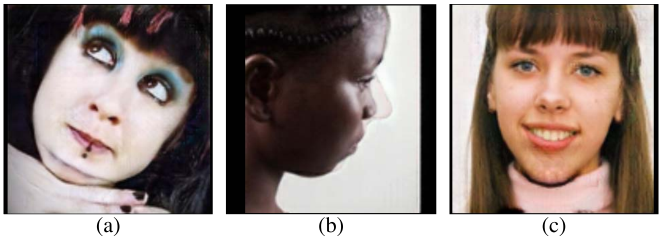


Fig. 6. Three most frequent types of "failure" that we observed for predictions on the AFLW2000 dataset. In (a) our model was tasked to predict a smaller chin. However, the original chin is still visible which results in an unrealistic prediction of the background. (b) shows the prediction of a nose enlargement where the model only generated the outlines of the desired shape of the nose. In (c) the enlarged chin of the woman yields an unnatural dark texture which was frequently observed for chin enlargements of female faces.

A. Qualitative Results

1) Experiment: We evaluated G on the AFLW2000 dataset [9] which yields 2000 pairs of images and corresponding shape parameters of the statistical point distribution model as well as the camera parameters to project the 3D shape to the 2D image plane. Like the 300W-LP dataset, these 3D faces were estimated using a semi-automatic fitting procedure [23] and were assumed as ground-truth in this study. For inference, we tested our model on all 2000 images of the AFLW2000 dataset using all four different modifications S^n_{mod} as input that were proposed in this study and are visualized in Fig. 2 (b): larger chin, smaller chin, larger nose, and smaller nose.

2) Results: Fig. 5 (a) and Fig. 5 (b) show the predictions of G for selected images that we visually judged to be both realistic and accurate compared with the given input S^n_{mod}. In detail, the top rows show the predictions $\tilde{I}^n_{mod}$ for a smaller chin and a smaller nose, respectively while the bottom row shows the predictions for a larger chin and nose. For comparison, the original unmodified images I^n are given in the middle row. As can be seen on our best examples, G was able to modify the desired region for images with varying head pose and illumination settings. The applied modification appeared to be realistic and the integration of the modified face with the rest of the face was plausible. Notably, our model was also able to predict a plausible background of regions that were previously occluded by face as seen in the top row of the figures. However, the overall performance was moderate as our model did not consistently predict realistic and accurate facial modifications on all images of the dataset. As an example, Fig. 6 shows three of the most frequent types of "failure" that we observed on the AFLW2000 dataset. In Fig. 6 (a), G was tasked to predict a smaller chin. However, the model did not remove the previously larger part of the chin which resulted in an unrealistic prediction of the background. Fig. 6 (b) shows the prediction of a nose enlargement. However, the model only generated the outlines of the desired shape of the nose which was sufficient to fool the shape estimator G_S and achieve a low shape reconstruction error $\mathcal{L}_{S-rec}$. These cases of failure could be found for both chin enlargements and nose enlargements and suggest a weak adversarial loss $\mathcal{L}_{Adv}$. Lastly, we observed a specific case of failure that mostly affected predictions of large chins in women. As seen in the example in Fig. 6 (c), the enlarged chin of the female face yielded an unnatural dark texture at the tip of the chin which could be interpreted as either artifacts, facial hair, or heavy shading and generally resulted in chin predictions that appeared more manly compared to the overall appearance of the original

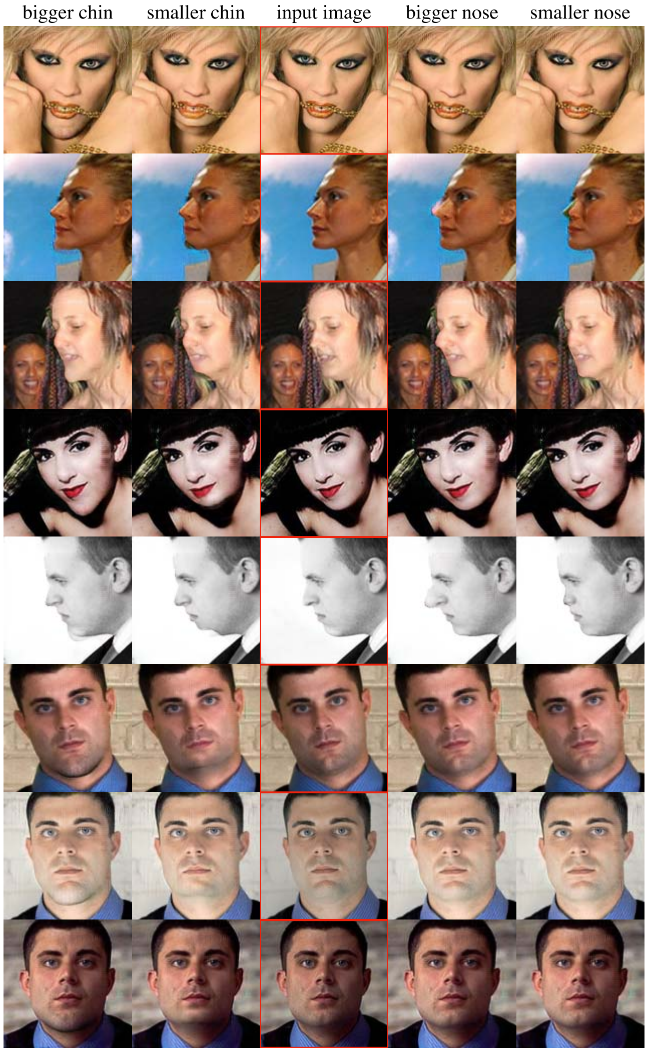


Fig. 7. Nose and chin modifications on the first eight AFLW2000 dataset samples. Third column shows the baseline input images I^n for the generator G. First and second column show the predictions of G for a bigger chin and a smaller chin, respectively. Fourth and fifth column show the predictions of G for a bigger nose and a smaller nose, respectively.

face. A potential explanation for this observation is given in Section V. The described cases can also be seen in Fig. 7 in which we show the predictions for all four modifications on the first eight samples of the AFLW2000 dataset to provide the reader with an unbiased selection of images.

B. Quantitative Results

1) Experiment: In this section, we aim to analyze the accuracy of the predicted modified regions in $\tilde{I}^n_{mod}$. To enable a fully automatic approach, we used a facial landmark predictor [38] and calculated the normalized Euclidean distance between landmarks of $\tilde{I}^n_{mod}$ and the projected landmarks of the corresponding shape S^n_{mod}.

To generate $\tilde{I}^n_{mod}$ we rerun the pipeline on all 2000 images of the AFLW2000 as described in section IV-A. To calculate the "ground-truth" facial landmarks, we annotated 68 landmarks on the 3D shape S^n_{mod} via their indices provided by [39].

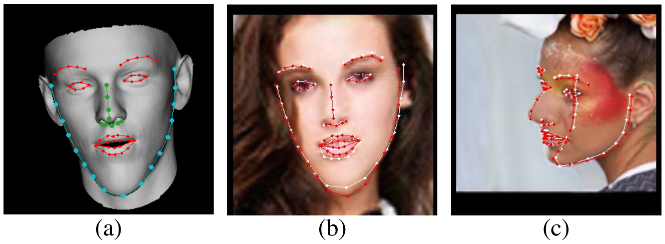


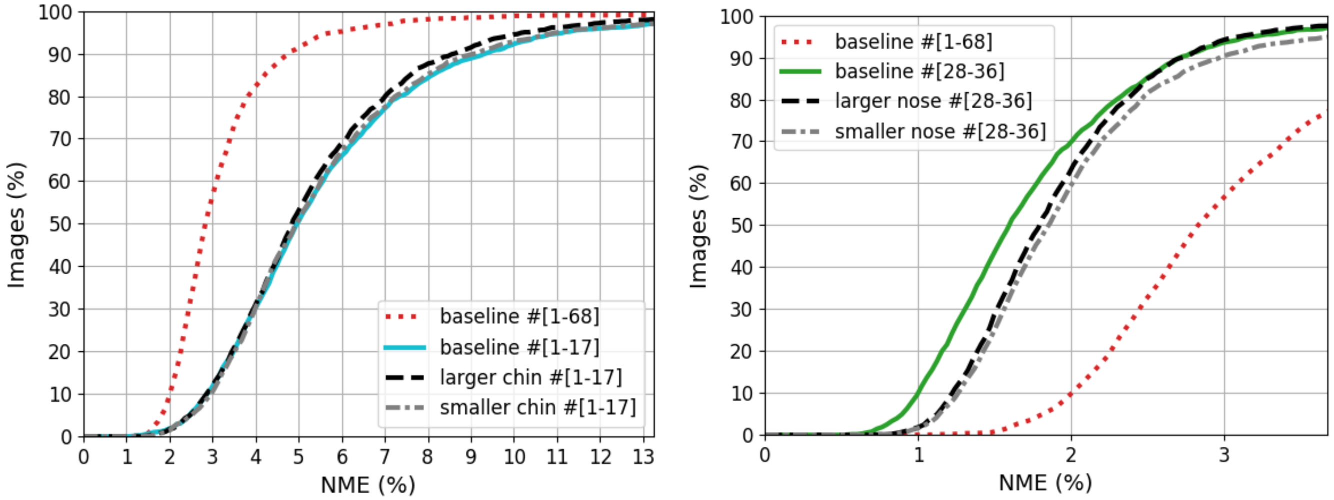
Fig. 8. Facial landmark annotations to evaluate the accuracy of our model. (a) shows the 68 facial landmarks that we annotated on the 3D shape via the vertex indices. Landmarks of the chin region #[1-17] are annotated in cyan. Landmarks of the nose region #[28-36] are annotated in green. (b) shows the predicted landmarks in white on the prediction $\tilde{I}^n_{mod}$ for an enlarged chin. The ground-truth landmarks derived from the modified shape S^n_{mod} are shown in red. (c) shows the landmark annotations on the prediction $\tilde{I}^n_{mod}$ for an enlarged nose.

After that we projected the landmarks in 3D to the 2D image plane as visualized in Fig. 8 (a). Additionally, we annotated the first 17 landmarks (#[1-17]) to belong to the chin region and nine landmarks (#[28-36]) to belong to the nose region as annotated in Fig. 8 (a). We then predicted all 68 facial landmarks of the modified faces $\tilde{I}^n_{mod}$ using the face-alignment network by Bulat and Tzimiropoulo [38]. For comparison, we also predicted the facial landmarks of the original images I^n and calculated the ground-truth facial landmarks using S^n. Examples of the predicted 68 landmarks and the corresponding ground-truth landmarks are given in Fig. 8 (b), (c). To create a comparable setting to [38], we also up-scaled all images from 128×128 pixel to 450×450 pixel resolution before applying the landmark predictor and calculating the normalized mean error (NME) proposed by [38]:

$$\text{NME} = \frac{1}{N}\sum_{k=1}^{N}\frac{\|x_k - y_k\|_2}{d} \tag{12}$$

Hereby, x were the landmark predictions, y were the projected landmarks of the 3D shape, and $d = \sqrt{w \times h}$ was a normalization factor derived by the width w and height h of the bounding boxes given by the AFLW2000 dataset for each unmodified image I^n. Additionally, we separately calculated the NME for the chin region using the landmarks #[1-17] with $N = 17$ and for the nose region using the landmarks #[28-36] with $N = 9$ as shown in Fig. 8 (a).

2) Results: Fig. 9 shows the cumulative distribution functions (CDFs) of the NME across all samples of the AFLW2000. In both figures, a baseline CDF is provided as a dotted line in red which was calculated on the original baseline images I^n using all 68 landmarks #[1-68]. When compared to the reported results in [38], we were able to reproduce similar baseline CDFs and thus, we are confident that we correctly implemented the face-alignment framework and the NME calculation described by Bulat and Tzimiropoulo [38]. For a meaningful analysis of the prediction accuracy of G, the CDFs of the modified images $\tilde{I}^n_{mod}$ must not be interpreted on their own since prediction errors of G might be confused with prediction errors of the landmark predictor or fitting errors of the 3D shapes in the dataset. Instead, we compared the CDFs of the baseline images I^n with the CDFs of



(a) (b)

Fig. 9. Cumulative distribution functions (CDFs) of the normalized mean error (NME) between the landmark predictions and the ground-truth landmarks. In both figures, the red dotted lines give the CDFs for the unmodified baseline images I^n of the AFLW2000 dataset calculated on all facial landmarks #[1-68]. (a) shows the CDFs calculated using the chin landmarks #[1-17] on the baseline images I^n in solid cyan, on the predictions $\tilde{I}^n_{mod}$ of a larger chin as a dashed black line, and on the predictions $\tilde{I}^n_{mod}$ of a smaller chin as a dash-dotted gray line. (b) shows the CDFs calculated using the nose landmarks #[28-36] on the baseline images I^n in solid green, on the predictions $\tilde{I}^n_{mod}$ of a larger nose as a dashed black line, and on the predictions $\tilde{I}^n_{mod}$ of a smaller nose as a dash-dotted gray line. The x-axis was capped at 13.25% and 3.68% to ignore the worst 3% of the baseline images #[1-17] and the baseline images #[28-36], respectively. A higher area under the curve (AUC) means a better overall landmark accuracy on the dataset.

TABLE I

AREA UNDER THE CURVE OF THE BASELINE AND THE MODIFIED IMAGES

head pose θ (°)	number of images	baseline #[1-17] AUC (%)	larger chin AUC diff. (%)	smaller chin AUC diff. (%)	baseline #[28-36] AUC (%)	larger nose AUC diff. (%)	smaller nose AUC diff. (%)
[0 - 30)	1312	61.16	+1.06	-1.08	54.76	-4.62	-5.02
[30 - 60)	383	57.60	+0.73	+3.86	51.56	-11.81	-2.84
[60 - 90]	305	45.11	-2.04	+7.95	40.68	-6.52	-0.96

the modified images $\tilde{I}^n_{mod}$ in an attempt to compensate the landmark prediction errors and the fitting errors of the dataset. Fig. 9 (a) shows a CDF in cyan solid line which was calculated using the landmarks of the chin #[1-17] on the unmodified baseline images I^n. The CDFs of the modified images $\tilde{I}^n_{mod}$ are given for a larger chin as a black dashed line and a smaller chin as a gray dash-dotted line. Likewise, Fig. 9 (b) shows the CDFs using the landmarks of the nose #[28-36] on the baseline images I^n as a green solid line and the predictions $\tilde{I}^n_{mod}$ of the modifications (larger nose as a black dashed line, smaller nose as a gray dash-dotted line). For a better visual comparison, we cropped both figures on the x-axis at 13.25% and 3.68% to exclude the worst 3% of all calculated NME errors that belonged to the baseline CDFs #[1-17] shown as a cyan solid line and #[28-36] as a green solid line, respectively. As reported by [38], these high NMEs that we excluded were mostly attributed to either poor ground-truth annotations of the AFLW2000 dataset or faces in the background that led to wrong landmark predictions. As seen in Fig. 9 (a), the CDFs for larger or smaller chains were comparable to the CDF of the baseline #[1-17]. Quantitatively, the normalized area under the curve (AUC) of the baseline #[1-17] was slightly worse with an AUC of 57.13% compared to the larger chin predictions with an AUC of 59.11% and smaller chin predictions with an AUC of 57.43%. For the nose modifications, the CDFs were worse compared to the baseline #[28-36] which led to a baseline AUC of 51.39%, a larger nose AUC of 47.62%, and a smaller nose AUC of 45.58%. Thus, our quantitative results on our in-the-wild dataset suggested that our model predictions were more accurate for chin modifications compared to nose modifications. These quantitative findings are in accordance with our qualitative findings where we visually observed that the predictions of the chin modifications appeared to be both more realistic and more accurate compared to the modifications of the nose.

The AFLW2000 dataset yielded a high variation of head pose rotations around vertical axis with the angle θ which might have been an additional challenge to G. To analyze the effect of such head rotations around θ, we provide the AUCs for three different absolute ranges of the vertical axis in Table I. Hereby, we calculated the AUCs on all baseline images and modified images using trapeze integration and using the same boundaries for the x-axis that are given in Fig. 9 (a) and 9 (b), respectively. Additionally, we also normalized each AUC by dividing by the respective length of the x-axis. When comparing the baseline AUCs in Table I for different angles, one can see that the AUCs strongly decrease for larger θ which can be attributed to a worse prediction accuracy of the landmark predictor as previously reported by [38]. To allow a better comparison with the baseline, Table I shows the difference between the AUC of the modified images $\tilde{I}^n_{mod}$ and the AUC of the baseline images I^n for each modification and angle θ. For the chin modifications, the AUCs show no clear indication that large head rotations impede the prediction accuracy of G when compared to the baseline AUCs. While the AUC for large angles ($\theta \geq 60°$) decreased for larger chins by 2.04%, the AUC for smaller chins even improved by 7.95% compared to the baseline. This improvement of the AUC compared to the baseline might be explained by a better alignment of the prediction $\tilde{I}^n_{mod}$ with S^n_{mod} compared to the given sample pairs I^n and S^n of the AFLW2000 dataset. For the nose modifications in contrast, larger head rotations with $30° \leq \theta < 60°$ resulted in a strong decrease of the AUC by 11.81% and 6.52% with $60° \leq \theta \leq 90°$. On the other hand, the AUCs for the predictions of smaller noses showed no clear tendency for large θ. Hereby, one should keep in mind that for a frontal view of the head with $\theta = 0°$, the landmark predictions are highly insensitive against modifications of the nose length. Therefore, inaccurate predictions of the nose by G might still yield low NMEs for small θ. As a consequence, the AUCs for small head rotations $\theta < 30°$ should be interpreted with care when regarding the nose modifications. However overall, the lower AUCs for the nose modifications with $\theta \geq 30°$ suggest that the accuracy of the "larger nose" predictions was poor compared to the baseline while the landmark accuracy of the "smaller nose" predictions was comparable to the baseline.

C. Prediction of the Post-Operative Face

1) Experiment: In this section, we aim to demonstrate the potential of our approach to predict 2D images of the post-operative face. For this, we evaluated our model on pre-operative and post-operative measurements of four patients that underwent orthognathic surgery to serve as a proof-of-concept. The images of the patients before surgery are given in Fig. 10 (b) for a frontal and lateral head position, respectively. Before surgery, patient P1, P2, and P3 suffered from a class III malocclusion, and P4 suffered from a class II malocclusion. All four patients were treated by bimaxillary surgery and the resulting post-operative faces can be seen in Fig. 10 (d) which were captured eight weeks after surgery. To test our model for the prediction of post-operative outcome, we passed two inputs to our model G: A cropped image of the pre-operative face as seen in Fig. 10 (b) and a simulation of the post-operative 3D shape that we derived from the surgical planning tool IPS CaseDesigner® [1] as seen in Fig. 10 (a). More precisely, we used a CT image scan from the pre-operative face, segmented soft-tissue and bone tissue and subsequently applied a bimaxillary surgery to the virtual bone structure of the jaw. We used IPS CaseDesigner®to simulate the soft-tissue deformations induced by the correction of the underlying bone structure. Based on this prediction of the post-operative 3D shape, we iteratively fitted a surface template of the BFM2009 model [23] on the 3D virtual face by adopting the approach described in [40]. Next, we estimated the camera matrix to project the fitted surface template onto the pre-operative face in Fig. 10 (b) by aligning the surface mesh to the upper half of the pre-operative face. The resulting projection of the simulated post-operative face to the image plane can be seen in Fig. 10 (a). We converted the projected

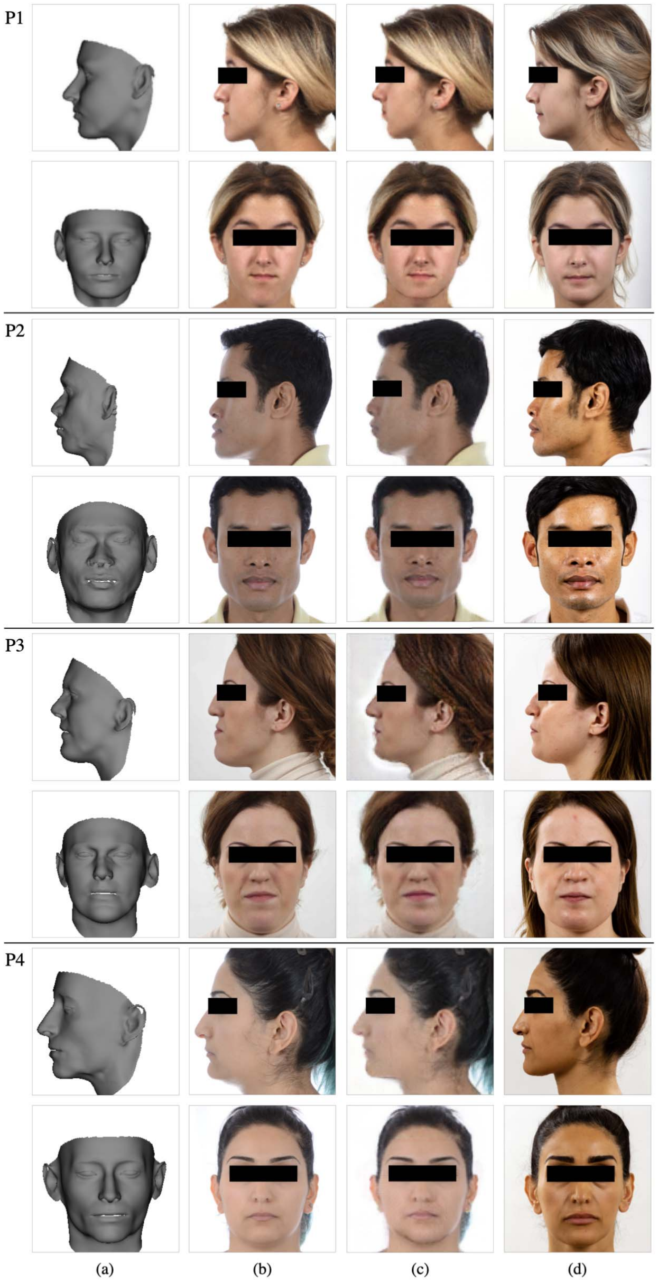


Fig. 10. Prediction of the post-operative face on four clinical examples. The upper row shows the inputs, predictions and ground-truths of each patient in lateral view. The lower row shows the same patient in a frontal view. (a) shows the fitted and projected face templates derived from a simulation of a surgery planning tool to predict the 3D shape of the post-operative face. The visualized 3D shapes were converted to PNCCs and passed as input to our model. (b) shows the images of the patient's face before cranio-maxillofacial surgery which were passed as a second input to our model. (c) shows the predictions of the post-operative face generated by our model. For better visibility, the images were up-scaled from the original output resolution of 128 × 128 pixels. (d) shows the patient's face eight weeks after cranio-maxillofacial surgery as a ground-truth. All images were published with the patient's consent.

TABLE II

FACE DISTANCE AND SSIM BETWEEN THE POST-OPERATIVE IMAGE (GROUND-TRUTH) AND THE PRE-OPERATIVE FACE OR THE POST-OPERATIVE PREDICTION, RESPECTIVELY

patient	view	face distance to the post-operative image		SSIM to the post-operative image	
		pre-operative	prediction	pre-operative	prediction
P1	lateral	0.924	0.894	0.9847	0.9851
	frontal	0.758	0.632	0.9735	0.9750
P2	lateral	0.718	0.656	0.9853	0.9838
	frontal	0.398	0.398	0.9708	0.9705
P3	lateral	1.092	1.051	0.9893	0.9907
	frontal	0.683	0.676	0.9112	0.9112
P4	lateral	0.647	0.639	0.9866	0.9867
	frontal	0.747	0.819	0.9823	0.9824

surface mesh to a PNCC by encoding the color of the surface template via their vertex indices. Lastly, we passed the resulting PNCC and the pre-operative image to our model G to predict a 128 × 128 pixel image of the post-operative face shown in Fig. 10 (c). Note that we only trained G on the 3DDFA dataset i.e. the model was never shown images or shape modifications from our clinical test case.

To compare our predictions in Fig. 10 (c) with the ground-truth post-operative face in Fig. 10 (d), we aimed to calculate the distance between two images of a face which we call face distance in the following. Hereby, we used a face-recognition neural network to mitigate the effect of changing illumination, skin color, hair color, haircut and other changes that were not related to bimaxillary surgery when comparing the prediction of the post-operative face with the post-operative ground-truth. More precisely, we calculated an embedding of each image using the InceptionResnetV1 by Esler [41] which was trained on the VGGFace2 dataset for face recognition [42]. Then, we measured the frobenius norm between two embeddings to calculate the distance between two images of a face. For every row in Fig. 10 we measured the face distance between the model's prediction of the post-operative face in Fig. 10 (c) and the ground-truth post-operative face in Fig. 10 (d). For reference, we also calculated the face distance between the pre-operative face in Fig. 10 (b) and the post-operative face Fig. 10 (d). Next, we also calculated the structural similarity index (SSIM) [43] between the prediction and the post-operative image. Hereby, we first manually aligned the predicted image with the post-operative face, converted the image to gray scale, and performed a histogram matching with the post-operative face to mitigate differences in illumination and skin color. Again, we also calculated the SSIM between the pre-operative face and the post-operative face to serve as a reference. Lastly, we showed our predictions to two clinician experts with years of experience regarding cranio-maxillofacial surgery to form a combined statement about the context of this study as well as the perceived quality of the predictions.

2) Results: The resulting predictions $\tilde{I}^n_{mod}$ of the post-operative face are shown in Fig. 10 (c) for the frontal and the lateral view. Additionally, the face distance and the SSIM are given in Table II. Hereby, two identical faces would ideally lead to a face distance of 0 and two identical images would lead to a SSIM of 1. 4 and 0 correspond to the opposite, respectively.

When comparing the lateral view of the pre-operative face in Fig. 10 (b) with the predictions for the lateral view, one can see a clear upward and right shift of the chin for all three patients P1, P2, and P3. For the lateral view of P4, the difference between the pre-operative face and the prediction was less clear. Note hereby, that P4 suffered from a different class of malocclusion which made the difference between the pre-operative face and the post-operative face visually less obvious. Nonetheless, the face distance in Table II was smaller for all lateral predictions which indicates a closer resemblance to the ground-truth image of the prediction compared to the pre-operative face. For the frontal views, the prediction of P1 yielded a clear upwards shift while the effect of cranio-maxillofacial surgery on the frontal prediction of P2 and P3 was less pronounced. For P4 however, the frontal prediction of the chin appeared unnatural and strongly differed from the ground-truth in Fig. 10 (d). Accordingly, the face distance between the prediction and the ground-truth in Table II increased compared to the face distance between the pre-operative face and the ground-truth. Thus, the face distance measurements suggests the frontal prediction of P4 to be a failed example of our approach. In contrast, the SSIM scores between the predictions and the ground-truth was slightly higher compared to the reference for all patients except for the frontal and lateral predictions of P2. This suggests that the prediction of the post-operative face of P2 was less accurate. On the other hand, the differences between the SSIM scores of the prediction and the reference were only minor i.e. less than 0.2 %.

Overall, the facial appearance of a majority of the predictions in Fig. 10 (c) were similar to the facial appearance of the ground-truth images of the post-operative face in Fig. 10 (d). Notably, our model also predicted a white background on the lateral views in Fig. 10 (c) and plausible backgrounds of the throat on the frontal views in Fig. 10 (c). On the other hand, minor artifacts were present in all predictions which included locally blurred regions of the face (in particular at the lips) as well as smaller artifacts on the skin, the lips, and the throat region. Hereby, one should keep in mind that the model G was never trained on modifications S_{mod} that substantially altered regions of the mouth or the throat. Additionally, when comparing the silhouette of the pre-operative face of P1 in the top row in Fig. 10 (b) with the simulated post-operative face in Fig. 10 (a), one can see a shape difference at the throat which had to be adapted by G. A particular reason for this deviation of the simulated shape compared to the pre-operative face in Fig. 10 (b) might be that the simulated post-operative face was derived from a CT scan which means that the patient was lying on the CT table. Therefore, the head position and the tissue of the throat might have varied compared to the up-right position of the patient during the capturing of the image in Fig. 10 (b). Conveniently, our model learned to almost fully ignore shape variations of the neck since a) the ground-truth shapes of the neck were poorly aligned with the images in our 3DDFA training set and b) the shape reconstruction loss $\mathcal{L}_{S-rec}$ for necks was only lightly weighted as described in Section III-D.

Lastly, we provide the statement of our clinical experts on the context of this study and the perceived image quality of the predictions: "In most cases, patients want to see realistic photographs of their facial appearance after correction of a skeletal deformity. Commercial software offers a 3D mesh with the possibility of texture overlay, still most patients are not able to identify with the displayed data. The provided predictions appear realistic and arguably closer to a natural image of a face a patient can relate to. As there is a risk of body dysmorphophobic disorder in severe changes to the facial appearance, preparing the patient with a relateable prediction and adequate counseling before obtaining informed consent for the procedure. The contours of the predicted faces appear smooth while the predictions of our commercial software produces uneven contours at the jaw after simulating the planned surgery protocol. [Authors note: These uneven contours can be seen at the jaw in Fig. 10 (a). on the lateral view of e.g. subject P2 simulated using [1] or for instance in Fig. 6 of [6].] On the other hand, the applied face modifications partly appear a bit extreme (compare e.g. the lateral prediction of P1 and P2) and the contours of some predictions are locally ambiguous, in particular at the lips. The main advantage we see however, is that the proposed approach does not require a 3D texture scanner. While every clinical site has a CT scanner and a camera, the availability of compatible 3D camera systems remains limited. In such cases, patients would have to make their decision based on a prediction which looks similar to Fig. 10 (a) which is inadequate. Using the approach proposed in this study however, we might be able to provide the patient in the future with a fast, non-committal, and natural-looking prediction of her/his face after only one CT scan."

V. Discussion

In the results in Section IV-A we showed that our CycleGAN was capable of predicting realistic and recognizable modifications of the chin and the nose on selected examples. Subsequently, we aimed to measure the accuracy of our predictions in Section IV-B by evaluating the Euclidean distance of facial landmarks on the AFLW2000 dataset. Hereby, we found the accuracy of the chin modifications to be similar compared to the accuracy of the matched 3D shapes of the AFLW2000 dataset. For the nose modifications, we found worse accuracy across the dataset compared to the baseline which was particularly pronounced for large head rotations. Lastly, we showed our proof-of-concept on four clinical patients where we predicted 2D images of the post-operative face according to the soft-tissue deformations of a surgery planning tool. Through this, we demonstrated that our model was indeed able to apply realistic modifications on four clinical patients without requiring additional training. With regard to the desired use case in clinical practice, one could argue that the task to train on "in-the-wild" images was a much more difficult task (in particular background prediction, illumination, and head pose) compared to the expected more controlled environment of the clinical use case. Thus, the results of our model might improve for a training dataset that is closer to the clinical test case. Likewise, one could also argue that the training on modifications of the nose was not required for the prediction of the post-operative face in Section IV-C which only affected the jaw. Therefore, we like to note that cranio-maxillofacial surgeries in general are not only concerned with jaw deformations but other regions of the face as well and, in particular, nose modifications can have a strong impact on the identity or appearance of the patient's face. Thus, our motivation was to propose a more generalized approach which theoretically enables modifications of any facial region that can be represented by both the statistical shape model and the dataset. Such an approach would require a model G that learned a continuous understanding of the desired 3D shape and accordingly, applied facial modifications wherever the 3D shape differed from the given input image. However, we found both qualitatively and quantitatively that our model performed worse on nose modifications than on chin modifications in terms of both robustness and accuracy. To explain the worse performance of our model on nose manipulations, we suggest the following reasons: First, we hypothesize that modifying noses is a much more complex task to solve compared to chins as noses yield arguably more fine-detailed textures and vary more strongly across different head poses. Consequently, this would suggest that our proposed approach to predict the post-operative outcome of faces might be limited at the moment to spatially less complex structures like the chin. On the other hand, our training procedure might still be biased in favor of chins. Although the number of chin and nose modifications was balanced and we used a weight map in the reconstruction loss to account for the differences of size between noses and chins, our discriminator might have been more sensitive to detect unrealistic chins due to the larger affected area in the image.

From a theoretical point of view, the suitability of our proposed training strategy for manipulating faces using CycleGANs can be discussed controversially. While our method has the advantage of not requiring ground truth images I^n_{mod} or knowledge of physical models, one could argue that a training strategy based on GANs will always bias the predicted face towards the mean face to maximize the expected reward from the discriminator. Therefore, attempts to manipulate facial properties that are "far away" from the mean face like an extremely enlarged chin might result in predictions that are closer to the mean face i.e. less "extreme" compared to the desired modification. Similarly, the model might have learned that extremely enlarged chins are far more likely to belong to male faces. This hypothesis might explain why we found some predictions for enlarged chins of female faces to yield facial hair or artifacts which resulted in a more manly appearance of the chin in Fig. 6 (c). Additionally, our approach using neural networks might also be vulnerable to ethnic imbalances of the training dataset. Consequently, applying our trained model to predict post-operative outcome might end up favoring e.g. patients of light skin color by providing a higher prediction quality compared to faces of dark skin color. However, such a racial bias is unacceptable for a clinical use case and would have to be ruled out by thorough testing before considering a model for a clinical application.

In the above passages we highlighted both empirical and theoretical limitations of our current model to modify facial regions and to predict the post-operative face. Despite these current limitations however, we are confident that these challenges can be overcome in the near future, especially in light of the rapid advances of GANs in recent years. In more detail, we particularly would like to improve the training and regularization strategy of the adversarials, the accuracy of the shape estimator G_S, and the image quality, accuracy, and ethnic balance of the dataset used for training. Afterwards, we would like to test our approach for the prediction of medical outcome in a thorough clinical study. Hypothetically, one might go even further and replace our current representation of the 3D soft-tissue with a 3D bone structure of the jaw. To achieve this, one would have to train a model to directly estimate the bone structure of the jaw from 2D images and subsequently train a CycleGAN to manipulate 2D images based on a modification of the bone structure provided by the physician. Having such a model, physicians would have a fast and cheap means to directly predict the post-operative face from 2D images without the need for expensive and time-consuming tomography scans.

VI. Conclusion

In this study we introduced a novel idea to predict the post-operative face using a neural network. Hereby, we showed that our prototype model was indeed capable of generating realistic predictions of the patient's face after cranio-maxillofacial surgery according to a given soft-tissue simulation. To train our model, we proposed a novel CycleGAN strategy to learn modifying facial regions of "in-the-wild" images according to a 3D plan of facial shape. Compared to current approaches to render the post-operative face, our approach can directly translate and manipulate the facial texture of a patient in 2D and therefore does not require the acquisition of 3D texture scans. Moreover, we achieved this prediction by merely training our model on open-source images without requiring clinically relevant face modifications or hand-crafted physical models. Based on our preliminary results and the rapid improvements of GANs in recent years, we believe that our proposed approach has a high potential to help the patient in their decision process in favor or against surgery. In future work, we aim to both increase the robustness of our model and test our model for the prediction of the post-operative face in a clinical follow-up study.

TABLE III
Network Architectures

Architecture of G

description	input shape→output shape	layer details
input layer	$(h,w,6)\to(\frac{h}{2},\frac{w}{2},64)$	CONV-(N64, K4×4, S2, P1), LeakyReLU
down-sampling	$(\frac{h}{2},\frac{w}{2},64)\to(\frac{h}{4},\frac{w}{4},128)$	CONV-(N128, K4×4, S2, P1), BN, LeakyReLU
down-sampling	$(\frac{h}{4},\frac{w}{4},128)\to(\frac{h}{8},\frac{w}{8},256)$	CONV-(N256, K4×4, S2, P1), BN, LeakyReLU
down-sampling	$(\frac{h}{8},\frac{w}{8},256)\to(\frac{h}{16},\frac{w}{16},512)$	CONV-(N512, K4×4, S2, P1), BN, LeakyReLU
down-sampling	$(\frac{h}{16},\frac{w}{16},512)\to(\frac{h}{32},\frac{w}{32},512)$	CONV-(N512, K4×4, S2, P1), BN, LeakyReLU
up-sampling	$(\frac{h}{32},\frac{w}{32},512)\to(\frac{h}{16},\frac{w}{16},512)$	DECONV-(N512, K4×4, S2, P1), BN, LeakyReLU
up-sampling	$(\frac{h}{16},\frac{w}{16},1024)\to(\frac{h}{8},\frac{w}{8},256)$	DECONV-(N256, K4×4, S2, P1), BN, LeakyReLU
up-sampling	$(\frac{h}{8},\frac{w}{8},512)\to(\frac{h}{4},\frac{w}{4},128)$	DECONV-(N64, K4×4, S2, P1), BN, LeakyReLU
up-sampling	$(\frac{h}{4},\frac{w}{4},256)\to(\frac{h}{2},\frac{w}{2},64)$	DECONV-(N64, K4×4, S2, P1), BN, LeakyReLU
output layer	$(\frac{h}{2},\frac{w}{2},128)\to(h,w,3)$	DECONV-(N3, K4×4, S2, P1), Tanh

Architecture of D

description	input shape→output shape	layer details
input layer	$(h,w,3)\to(\frac{h}{2},\frac{w}{2},48)$	CONV-(N48, K4×4, S2, P1), LeakyReLU
hidden layer	$(\frac{h}{2},\frac{w}{2},48)\to(\frac{h}{4},\frac{w}{4},96)$	CONV-(N96, K4×4, S2, P1), LeakyReLU
hidden layer	$(\frac{h}{4},\frac{w}{4},96)\to(\frac{h}{8},\frac{w}{8},192)$	CONV-(N192, K4×4, S2, P1), LeakyReLU
hidden layer	$(\frac{h}{8},\frac{w}{8},192)\to(\frac{h}{16},\frac{w}{16},384)$	CONV-(N384, K4×4, S2, P1), LeakyReLU
hidden layer	$(\frac{h}{16},\frac{w}{16},384)\to(\frac{h}{32},\frac{w}{32},768)$	CONV-(N768, K4×4, S2, P1), LeakyReLU
hidden layer	$(\frac{h}{32},\frac{w}{32},768)\to(\frac{h}{64},\frac{w}{64},1536)$	CONV-(N1536, K4×4, S2, P1), LeakyReLU
hidden layer	$(\frac{h}{64},\frac{w}{64},1536)\to(\frac{h}{64},\frac{w}{64},1)$	CONV-(N1, K3×3, S2, P1), LeakyReLU
output layer	$(\frac{h}{64},\frac{w}{64},1)\to(1)$	mean

Architecture of D_{Roi}

description	input shape→output shape	layer details
input layer	$(h,w,3)\to(h,w,48)$	CONV-(N48, K3×3, S1, P1), LeakyReLU
hidden layer	$(h,w,48)\to(\frac{h}{2},\frac{w}{2},96)$	CONV-(N96, K4×4, S2, P1), LeakyReLU
hidden layer	$(\frac{h}{2},\frac{w}{2},96)\to(\frac{h}{2},\frac{w}{2},192)$	CONV-(N192, K3×3, S1, P1), LeakyReLU
hidden layer	$(\frac{h}{2},\frac{w}{2},192)\to(\frac{h}{4},\frac{w}{4},384)$	CONV-(N384, K4×4, S2, P1), LeakyReLU
hidden layer	$(\frac{h}{4},\frac{w}{4},384)\to(\frac{h}{4},\frac{w}{4},768)$	CONV-(N768, K3×3, S1, P1), LeakyReLU
hidden layer	$(\frac{h}{4},\frac{w}{4},768)\to(\frac{h}{8},\frac{w}{8},1536)$	CONV-(N1536, K4×4, S2, P1), LeakyReLU
hidden layer	$(\frac{h}{8},\frac{w}{8},1536)\to(\frac{h}{8},\frac{w}{8},1)$	CONV-(N1, K3×3, S1, P1), LeakyReLU
output layer	$(\frac{h}{8},\frac{w}{8},1)\to(1)$	mean

Architecture of G_S

description	input shape→output shape	layer details
input layer	$(h,w,3)\to(\frac{h}{2},\frac{w}{2},48)$	CONV-(N64, K4×4, S2, P1), LeakyReLU
down-sampling	$(\frac{h}{2},\frac{w}{2},48)\to(\frac{h}{4},\ \frac{w}{4},\ 96)$	CONV-(N128, K4×4, S2, P1), BN, LeakyReLU
down-sampling	$(\frac{h}{4},\ \frac{w}{4},\ 96)\to(\frac{h}{8},\ \frac{w}{8},\ 192)$	CONV-(N256, K4×4, S2, P1), BN, LeakyReLU
down-sampling	$(\frac{h}{8},\ \frac{w}{8},\ 192)\to(\frac{h}{16},\ \frac{w}{16},\ 384)$	CONV-(N512, K4×4, S2, P1), BN, LeakyReLU
down-sampling	$(\frac{h}{16},\ \frac{w}{16},\ 384)\to(\frac{h}{32},\ \frac{w}{32},\ 384)$	CONV-(N512, K4×4, S2, P1), BN, LeakyReLU
up-sampling	$(\frac{h}{32},\ \frac{w}{32},\ 384)\to(\frac{h}{16},\ \frac{w}{16},\ 384)$	DECONV-(N512, K4×4, S2, P1), BN, LeakyReLU
up-sampling	$(\frac{h}{16},\ \frac{w}{16},\ 768)\to(\frac{h}{8},\ \frac{w}{8},\ 192)$	DECONV-(N256, K4×4, S2, P1), BN, LeakyReLU
up-sampling	$(\frac{h}{8},\ \frac{w}{8},\ 384)\to(\frac{h}{4},\ \frac{w}{4},\ 96)$	DECONV-(N64, K4×4, S2, P1), BN, LeakyReLU
up-sampling	$(\frac{h}{4},\ \frac{w}{4},\ 192)\to(\frac{h}{2},\ \frac{w}{2},\ 48)$	DECONV-(N64, K4×4, S2, P1), BN, LeakyReLU
output layer	$(\frac{h}{2},\ \frac{w}{2},\ 96)\to(h,\ w,\ 5)$	DECONV-(N5, K4×4, S2, P1), Tanh

Appendix A
Implementation Details to Calculate S_{mod}

In this section, we describe the optimization algorithm to automatically find the shape modifications S_{mod} in more detail. Based on the statistical point distribution model (BFM2009) by [23], the coordinates x_i (x,y, and z) of each vertex i of each 3D face S^n can be described by

$$x = V\alpha + \overline{x} \tag{13}$$

with V being a matrix of 199 eigenvectors to maximally describe the variance of faces, α being a parameter vector with 199 elements, and $\overline{x}$ being the mean face. To define the locally modified region, we manually selected a facial region e.g. the nose or the chin and labeled all vertices x_i within the selected region to belong to $Mask$. Consequently, all other vertices were labeled to belong to $\overline{Mask}$. We aimed to find such an $\hat{\alpha}$ that maximally deflects all vertices within $Mask$ while minimally deflecting all other vertices within $\overline{Mask}$. To achieve this, we optimized the following objective:

$$\min_{\alpha} \sum_{x_i \in \overline{Mask}} \|x_i\|_F - \lambda_1 \sum_{x_i \in Mask} \|x_i\|_F + \lambda_2(\|\alpha\|_F - 1)^2 \quad (14)$$

with $\|.\|_F$ being the Frobenius norm, $\lambda_1 = 4$ to control the deflection, and $\lambda_2 = 1000$ to regularize the solution $\hat{\alpha}$ to a constant length. Having such an optimized $\hat{\alpha}$, we were able to create a linearly scalable local modification S_{mod} using a scalar λ:

$$S_{mod} = \lambda \hat{\alpha} \quad (15)$$

As an example, we generated an enlarged nose modification by choosing $\lambda > 0$ and a shrunken nose modification by choosing $\lambda < 0$. On one hand, this approach to derive S_{mod} can generate local deflections on any region of the face as long as these deflections can be represented by the point distribution model. On the other hand, this approach has the disadvantage that it cannot generate specific local modifications since the optimization algorithm only focuses on a maximal deflection of the selected vertices.

Appendix B
Network Architectures

In the Table III, the detailed architectures are given for all neural networks of this study using the following abbreviations: CONV = 2D convolutional layer, DECONV = 2D transposed convolutional layer, BN = batch normalization, N = number of output channels, K = kernel size, S = stride size, P = padding size. The width and height are set to the image resolution i.e. $h = 128$, $w = 128$ except for the local discriminator D_{Roi} where we set the width and height between 16 and 48. All leaky rectifying linear units (LeakyReLU) were implemented using a negative slope of 0.01.

Acknowledgment

The authors would like to thank Zhu *et al.* [9], [44], Feng [39], and Choi *et al.* [10] for sharing their code.

References

[1] G. R. J. Swennen, *3D Virtual Treatment Planning of Orthognathic Surgery*. Berlin, Germany: Springer, 2017, pp. 217–277, doi: 10.1007/978-3-662-47389-4_3.

[2] A. H. Salazar, J. A. J. Méndez, and F. P. Vázquez, "DOLPHIN 3D," in *Proc. 2nd Int. Conf. Technol. Ecosyst. Enhancing Multiculturality (TEEM)*, Jan. 2014, pp. 35–40.

[3] M. Meehan, M. Teschner, and S. Girod, "Three-dimensional simulation and prediction of craniofacial surgery," *Orthodontics Craniofacial Res.*, vol. 6, pp. 102–107, Aug. 2003.

[4] A. Westermark, S. Zachow, and B. L. Eppley, "Three-dimensional osteotomy planning in maxillofacial surgery including soft tissue prediction," *J. Craniofacial Surgery*, vol. 16, no. 1, pp. 100–104, Jan. 2005.

[5] P. G. M. Knoops *et al.*, "A novel soft tissue prediction methodology for orthognathic surgery based on probabilistic finite element modelling," *PLoS ONE*, vol. 13, no. 5, May 2018, Art. no. e0197209.

[6] W. Mollemans, F. Schutyser, N. Nadjmi, F. Maes, and P. Suetens, "Predicting soft tissue deformations for a maxillofacial surgery planning system: From computational strategies to a complete clinical validation," *Med. Image Anal.*, vol. 11, pp. 282–301, Jun. 2007.

[7] L. H. Cevidanes *et al.*, "Three-dimensional surgical simulation," *Amer. J. Orthodontics Dentofacial Orthopedics*, vol. 138, no. 3, pp. 361–371, 2010.

[8] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2223–2232.

[9] 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.

[10] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 8789–8797.

[11] Y. Choi, Y. Uh, J. Yoo, and J.-W. Ha, "StarGAN v2: Diverse image synthesis for multiple domains," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8188–8197.

[12] J. Bao, D. Chen, F. Wen, H. Li, and G. Hua, "Towards open-set identity preserving face synthesis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 6713–6722.

[13] S. Liu, D. Li, T. Cao, Y. Sun, Y. Hu, and J. Ji, "GAN-based face attribute editing," *IEEE Access*, vol. 8, pp. 34854–34867, 2020.

[14] Z. He, W. Zuo, M. Kan, S. Shan, and X. Chen, "AttGAN: Facial attribute editing by only changing what you want," *IEEE Trans. Image Process.*, vol. 28, no. 11, pp. 5464–5478, Nov. 2019.

[15] E. Gladilin and A. Ivanov, "Computational modelling and optimisation of soft tissue outcome in cranio-maxillofacial surgery planning," *Comput. Methods Biomech. Biomed. Eng.*, vol. 12, no. 3, pp. 305–318, Jun. 2009.

[16] B. T. Harris, D. Montero, G. T. Grant, D. Morton, D. R. Llop, and W.-S. Lin, "Creation of a 3-dimensional virtual dental patient for computer-guided surgery and CAD-CAM interim complete removable and fixed dental prostheses: A clinical report," *J. Prosthetic Dentistry*, vol. 117, no. 2, pp. 197–204, Feb. 2017.

[17] J. Rubio-Palau *et al.*, "Three-dimensional planning in craniomaxillofacial surgery," *Ann. Maxillofacial Surgery*, vol. 6, no. 2, pp. 281–286, Jan. 2016.

[18] P. Premjani, A. Al-Mulla, and D. Ferguson, "Accuracy of 3D facial models obtained from CBCT volume wrapping," *J. Clin. Orthod*, vol. 49, pp. 641–646, Oct. 2015.

[19] C. Lane and W. Harrell, "Completing the 3-dimensional picture," *Amer. J. Orthodontics Dentofacial Orthopedics, Off. Amer. Assoc. Orthodontists, Constituent Societies, Amer. Board Orthodontics*, vol. 133, no. 4, pp. 612–620, Apr. 2008.

[20] S. Guan. (2018) *TL-GAN: Transparent Latent-Space GAN*. Accessed: Jul. 13, 2020. [Online]. Available: https://github.com/SummitKwan/transparent_latent_gan/

[21] W. Shen and R. Liu, "Learning residual images for face attribute manipulation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4030–4038.

[22] A. Bansal, S. Ma, D. Ramanan, and Y. Sheikh, "Recycle-GAN: Unsupervised video retargeting," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 119–135.

[23] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter, "A 3D face model for pose and illumination invariant face recognition," in *Proc. 6th IEEE Int. Conf. Adv. Video Signal Based Surveill.*, Sep. 2009, pp. 296–301.

[24] A. Gokaslan, V. Ramanujan, D. Ritchie, K. I. Kim, and J. Tompkin, "Improving shape deformation in unsupervised image-to-image translation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 649–665.

[25] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 586–595.

[26] J. Booth, E. Antonakos, S. Ploumpis, G. Trigeorgis, Y. Panagakis, and S. Zafeiriou, "3D face morphable models 'in-the-wild,'" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5464–5473.

[27] P. Huber *et al.*, "A multiresolution 3D morphable face model and fitting framework," in *Proc. 11th Joint Conf. Comput. Vis., Imag. Comput. Graph. Theory Appl.*, 2016, pp. 79–86.

[28] A. S. Jackson, A. Bulat, V. Argyriou, and G. Tzimiropoulos, "Large pose 3D face reconstruction from a single image via direct volumetric CNN regression," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1031–1039.

[29] M. Sela, E. Richardson, and R. Kimmel, “Unrestricted facial geometry reconstruction using image-to-image translation,” in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1576–1585.

[30] A. T. Tran, T. Hassner, I. Masi, and G. Medioni, “Regressing robust and discriminative 3D morphable models with a very deep neural network,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5163–5172.

[31] B. Gecer, S. Ploumpis, I. Kotsia, and S. Zafeiriou, “GANFIT: Generative adversarial network fitting for high fidelity 3D face reconstruction,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1155–1164.

[32] M. Arjovsky, S. Chintala, and L. Bottou, “Wasserstein GAN,” 2017, *arXiv:1701.07875*. Accessed: Jul. 13, 2020. [Online]. Available: http://arxiv.org/abs/1701.07875

[33] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, “Improved training of wasserstein GANs,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5767–5777.

[34] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *CoRR*, vol. abs/1412.6980, pp. 1–15, Dec. 2015.

[35] A. Quattoni and A. Torralba, “Recognizing indoor scenes,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2009, pp. 413–420.

[36] S. Jegou, M. Drozdzal, D. Vazquez, A. Romero, and Y. Bengio, “The one hundred layers tiramisu: Fully convolutional DenseNets for semantic segmentation,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 11–19.

[37] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, “Image-to-image translation with conditional adversarial networks,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1125–1134.

[38] 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),” in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1021–1030.

[39] Y. Feng. (2018). *Face3d: Python Tools for Processing 3D Face*. Accessed: Jul. 13, 2020. [Online]. Available: https://github.com/YadiraF/face3d

[40] H. Dai, N. Pears, W. Smith, and C. Duncan, “Statistical modeling of craniofacial shape and texture,” *Int. J. Comput. Vis.*, vol. 128, no. 2, pp. 547–571, Feb. 2020.

[41] T. Esler. *Pretrained Pytorch Face Detection (MTCNN) and Recognition (Inceptionresnet) Models*. Accessed: Mar. 20, 2021. [Online]. Available: https://github.com/timesler/facenet-pytorch

[42] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, “VGGFace2: A dataset for recognising faces across pose and age,” in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2018, pp. 67–74.

[43] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: From error visibility to structural similarity,” *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[44] J. Guo, X. Zhu, and Z. Lei. (2018). *3DDFA*. Accessed: Jul. 13, 2020. [Online]. Available: https://github.com/cleardusk/3DDFA