Generative RGB-D Face Completion for Head-Mounted Display Removal

Conference Paper · March 2021

DOI: 10.1109/VRW52623.2021.00028

3 authors, including:

Nels Numan
University College London
2 PUBLICATIONS **2** CITATIONS
SEE PROFILE

Pablo Cesar
Centrum Wiskunde & Informatica
285 PUBLICATIONS **2,319** CITATIONS
SEE PROFILE

Generative RGB-D Face Completion for Head-Mounted Display Removal

Nels Numan*
Delft University of Technology
TNO

Frank ter Haar†
TNO

Pablo Cesar‡
Delft University of Technology
CWI

Abstract

Head-mounted displays (HMDs) are an essential display device for the observation of virtual reality (VR) environments. However, HMDs obstruct external capturing methods from recording the user's upper face. This severely impacts social VR applications, such as teleconferencing, which commonly rely on external RGB-D sensors to capture a volumetric representation of the user. In this paper, we introduce an HMD removal framework based on generative adversarial networks (GANs), capable of jointly filling in missing color and depth data in RGB-D face images. Our framework includes an RGB-based identity loss function for identity preservation and several components aimed at surface reproduction. Our results demonstrate that our framework is able to remove HMDs from synthetic RGB-D face images while preserving the subject's identity.

Index Terms: Computing methodologies—Artificial intelligence—Computer vision—Reconstruction; Human-centered computing—Human computer interaction (HCI)—Interaction paradigms—Virtual reality

1 Introduction

Virtual reality (VR) enables users to explore immersive virtual environments (IVEs) with a head-mounted display (HMD), which render a virtual world based on their physical movement. The natural interface this technology offers has enabled a wide range of simulations, which are too complex, hazardous or costly for execution in the real world. Groups of individuals are commonly brought together in a single IVE to create social VR experiences, such as collaborative learning [35,41], treatment of mental disorders [10,26], and teleconferencing [4,11,40]. These experiences are primarily driven by the quality of interpersonal interaction in VR, which is fundamentally impacted by the visual quality of human representations as highlighted by a large body of literature [2,18,27]. There are several approaches to represent humans in IVEs. Some solutions are based on advanced computer generated imagery (CGI), including cartoon-like avatars such as in AltspaceVR[1] and photorealistic avatars [52], which mimic the movement of the user. Other approaches are rooted in the videoconferencing world, aimed at volumetric telepresence. This paper focuses on such methods, where users are captured in real-time using RGB-D sensors, who are reconstructed as volumetric video (point clouds) and rendered in the virtual world with other users [11,25,30,32].

While HMDs are an essential display device for VR, they obstruct external capturing methods from recording the user's upper face, severely impacting the social aspects of VR applications. In this work, we propose an image-based method for the virtual removal of HMDs, which coherently fills in the occluded color and geometric information of a subject's face represented in RGB-D images. This task is referred to as HMD removal.

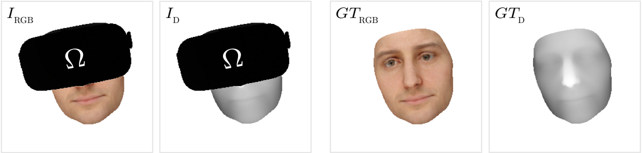


Figure 1: Illustration of our target problem. RGB-D input image I contains a face and masked region Ω. We aim to virtually remove the HMD by filling in the missing color (RGB) and geometric (D) information of image region Ω, seamlessly connecting it with the image region $I - \Omega$. GT denotes the ground truth image.

In general, there are two types of approaches to HMD removal: model-based methods and image-based methods. Model-based methods [13,28,37,47,58] provide realistic results by representing the user with a 3D face model, which is transformed during online usage based on sensor data. However, methods of this kind require an offline capture and calibration process. In contrast, image-based methods [49,57] attempt to synthesize the facial region through image inpainting, which involves resolving occluded pixels of an image in a realistic way. However, existing approaches consider only RGB images, which do not offer the necessary geometric data to accurately represent the user in an IVE.

To address this, we introduce an image-based HMD removal framework that is capable of the removal of HMDs in RGB-D images. To solve this problem (Fig. 1), we build on recent advancements in the field of image inpainting involving the application of generative adversarial networks (GANs). The main contributions of this work are as follows. We propose a GAN-based framework that is capable of inpainting both the color and depth information present in RGB-D face images, which includes an RGB-based identity loss function for identity preservation. In order to train and evaluate our framework, we define a synthesization pipeline for the creation of a large-scale RGB-D face image dataset based on Basel Face Model (BFM) 2017 [16] with random pose, ambient illumination, and expression.

2 Related Work

2.1 Head-Mounted Display (HMD) Removal

HMD removal describes the recovery of missing image information caused by the occlusion of an HMD in a coherent and realistic way. Model-based methods [13,28,37,47,58] use a face model to represent the user's facial geometry and expressions, which typically is recorded or designed prior to online usage. At runtime, coefficients are inferred from sensor data, which in turn are used to transform the face model. These methods provide realistic results with low bandwidth requirements, but often require custom hardware and calibration processes. Image-based approaches [49,57] generally do not use an intermediate model to virtually remove the HMD, but rely on operations in the image space to resolve the occluded area. Consequently, methods of this kind form a subtask of face completion or image inpainting. Zhao et al. [57] proposed a GAN-based framework to virtually remove synthetically placed HMDs from RGB images. This procedure is robust against moderate variations in pose and is able to preserve identity given a reference image of

*e-mail: nels.numan@gmail.com
†e-mail: frank.terhaar@tno.nl
‡e-mail: p.s.cesar@cwi.nl

[1] https://altvr.com

the target subject. However, this approach relies on a pose map that represents the subject's head pose and only considers RGB images, which limits its applications within VR. We address this limitation by proposing a framework that is capable of handling RGB-D images.

2.2 Image Inpainting

In general, image-based HMD removal approaches are driven by research from the field of image inpainting, which describes the task of filling missing image regions with realistic content. Recent works [24,39,53–55] have adopted the concept of GANs [17], which learn a representative estimate of the distribution of the given training data. Various methods demonstrate to perform well on face images, but often do not preserve the subject's identity and only consider RGB color images.

Several approaches for depth image inpainting exist that utilize corresponding RGB data as context for the inference of missing depth information [1, 22, 44, 56]. Other works train models that attempt to minimize the difference between the surface normals of the completed depth image and its ground truth [34,56]. To improve reproduction of surfaces formed by the depth image, we utilize surface normal images and employ several components proposed by Matias et al. [34].

There is limited research on RGB-D image inpainting [12, 14, 15, 36, 48], which is likely due to the fact that RGB-D images contain multimodal information. While the RGB values represent the color of the captured object, the D values represent the distance between the object and the sensor. Therefore, each modality has its own statistical properties, which complicates feature construction. Fujii et al. [14, 15] introduced the first generative approach to RGB-D image inpainting, fusing color and depth features at feature-level. However, the method does not consider large non-rectangular masks, RGB-D face images, or identity preservation.

3 Approach

We adopt the two-stage RGB image inpainting method proposed by Yu et al. [55] as our base architecture. Our decision is based on its state-of-the-art performance and its architectural characteristics. Firstly, this framework uses gated convolution, allowing masks to have any size and to appear anywhere in the input image. Furthermore, it employs the SN-PatchGAN discriminator, which is able to focus on different locations and semantics across image channels. This is particularly relevant for capturing different types of semantics represented in the multimodal images that we aim to inpaint.

Our network architecture, shown in Fig. 2, follows the structure of the base architecture [55] and consists of a coarse-to-fine generator G and discriminator D, which are trained through an adversarial process. Throughout this process, the objective of G is to accurately inpaint a given image, whereas the objective of D is to determine whether the inpainted image is real or not.

3.1 Generator

The coarse-to-fine generator G inpaints the RGB-D image in two stages. The first stage produces a coarse prediction of the masked image region. Subsequently, this prediction is fed to the second stage where it is further refined. G takes an occluded RGB-D image I, a binary mask Ω, and an RGB reference image I_{ref} as its input.

Fusion of color and depth information

All image channels are passed to the generator in concatenative fashion. Consequently, both the color channels (RGB) as well as the depth channel (D) of the RGB-D input image are processed simultaneously without any form of feature-level fusion. This naive fusion strategy is generally referred to as data-level fusion. We have explored and developed several feature-level fusion methods in the coarse stage of the generator, such as fusion through feature summation [20] and residual feature fusion [14, 38]. However, we have not found concrete evidence to prove the benefit of such feature-level fusion strategies over data-level fusion within the context of this work. Therefore, at the input of both stages of G, we employ data-level fusion.

Contextual surface attention (CSA) module

The refinement stage of generator G contains a contextual attention (CA) module, enabling propagation of information originating from any spatial location in the image. Without any modification to the CA module of the base framework [55], it considers absolute depth values in its matching procedure. To enable the generator to interpret the geometric surfaces formed by these values, we adopt the contextual surface attention (CSA) module introduced by Matias et al. [34]. This module relies on a surface normal image, created through the estimation of the normal vectors of each pixel through the analysis of neighborhood depth values. This operation takes place based on the coarse prediction, of which the result is passed to the CA branch of the refinement stage.

3.2 Discriminator

Throughout the model training process, discriminator D learns to distinguish inpainted images from ground truth images. D takes an occluded RGB-D image I and a binary mask Ω as its input.

We employ the SN-PatchGAN discriminator of the base framework, which independently classifies each patch of the input image to be real or fake through convolution. These classification responses are averaged to form the value of SN-PatchGAN loss $\mathscr{L}_{\text{GAN}}$.

3.3 Loss Functions

Our objective function comprises four loss functions, whose value is minimized during training. Firstly, we employ the two loss functions of the base framework, $\mathscr{L}_{\text{GAN}}$ and ℓ_1 reconstruction loss. As previously described, $\mathscr{L}_{\text{GAN}}$ focuses on the channel-wise reproduction of the image information in image patches at different spatial locations, whereas the ℓ_1 loss supervises pixel-wise reconstruction. In addition, we employ our identity loss $\mathscr{L}_{\text{ID}}$ and vectorial loss $\mathscr{L}_{\text{vec}}$ [34], described below.

Identity loss

An image of a human face contains vast information, cognitively distinct to a person's identity [7]. As identity has a profound impact in face-to-face communication [46], its preservation is essential in the context of our framework.

We address this with the identity loss function $\mathscr{L}_{\text{ID}}$, that supervises the model in the identity-preserving reconstruction of faces represented in RGB-D images. Similar to other perceptual loss functions for identity preservation [23,29,31,45,57], we use a pretrained face recognition model. Specifically, we employ a ResNet50 [21] model trained on the VGGFace2 dataset [8]. Moreover, we require an RGB reference face image to be included in the input of the generator, as indicated in Fig. 2. Notably, we do not require the reference image to contain a depth channel. This is due by the fact that highly accurate pretrained RGB-D face recognition models are currently not available. Consequently, the preservation of identity is completely dependent on the color information represented in the inpainted and reference images.

During model training, an identity embedding of the RGB reference image x_{ref} and the inpainted RGB image x_{pred} is computed by passing them through pretrained face recognition model M, of which we take the activation values of the last convolutional layer. We then calculate the mean squared error (MSE) between the two identity embeddings which forms the value of identity loss $\mathscr{L}_{\text{ID}}$, which is defined as:

$$\mathscr{L}_{\text{ID}}(x_{\text{pred}}, x_{\text{ref}}) = \text{MSE}(M(x_{\text{pred}}) - M(x_{\text{ref}})) \tag{1}$$

Vectorial loss

We are faced with a challenge consisting of the joint completion of two spatially-aligned images of differing modalities, an RGB image and a depth image. Notably, the depth image contains pixel values that represent the distance from the face to the RGB-D sensor, which collectively form a surface. Yet, the

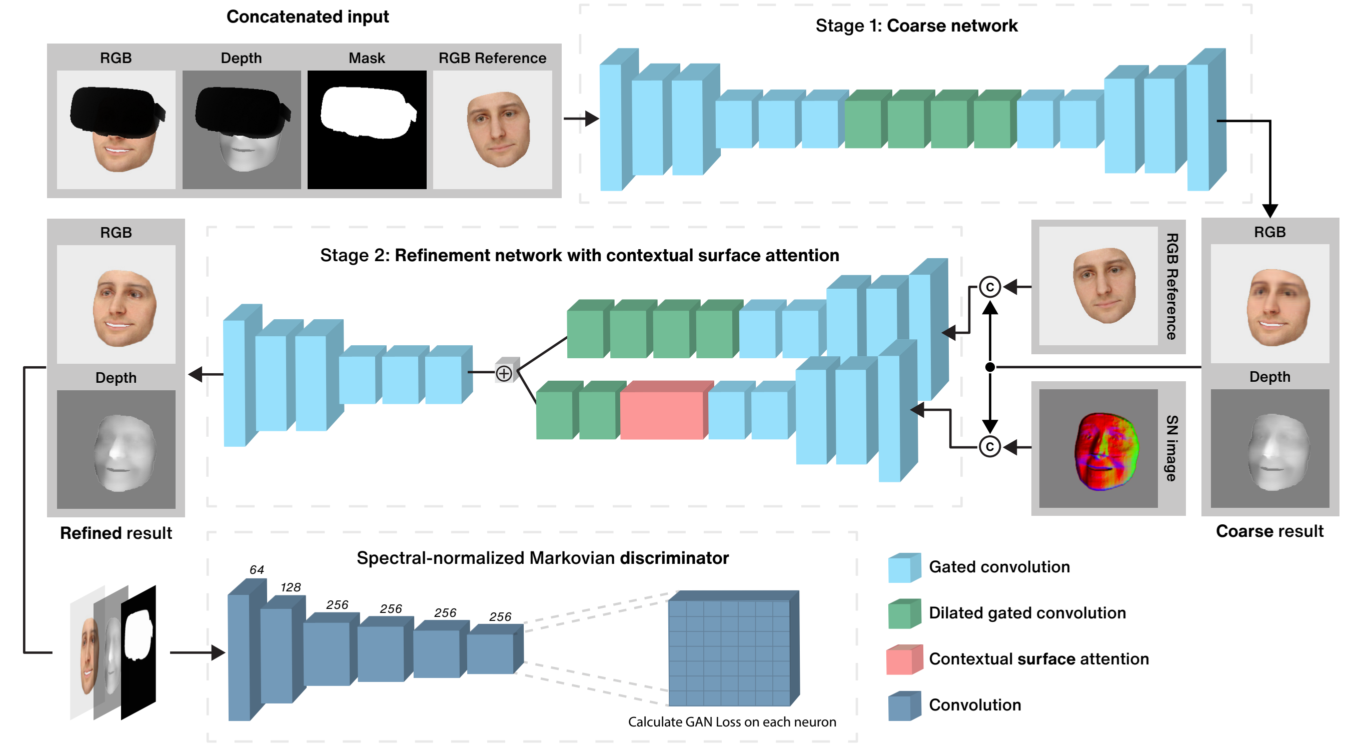

Concatenated input
RGB
Depth
Mask
RGB Reference
Stage 1: Coarse network
Stage 2: Refinement network with contextual surface attention
RGB
Depth
Refined result
RGB Reference
SN image
RGB
Depth
Coarse result
Spectral-normalized Markovian discriminator
64
128
256
256
256
256
Calculate GAN Loss on each neuron
Gated convolution
Dilated gated convolution
Contextual surface attention
Convolution

Figure 2: Overview of our RGB-D image inpainting architecture with data-level fusion, contextual surface attention (CSA), and the required inputs for the identity loss $\mathscr{L}_{\text{ID}}$, vectorial loss $\mathscr{L}_{\text{vec}}$, and SN-PatchGAN loss $\mathscr{L}_{\text{GAN}}$. ⓒ and ⊕ denote concatenation and addition respectively. This figure is adapted from the work of Yu et al. [55].

base architecture does not include a mechanism to interpret the depth values as a surface.

To address this, we employ a loss function that encourages the reproduction of surfaces contained in the depth channel. This loss function is based on surface normal estimation and was introduced as the vectorial loss function by Matias et al. [34].

During training, the surface normal image of the inpainted image and the corresponding ground truth image are computed. To obtain vectorial loss $\mathscr{L}_{\text{vec}}$, we calculate the ℓ_1 distance between these two images. In this way, for each pixel, the error between the ground truth normal vector and the normal vector as inpainted contributes to the value of vectorial loss $\mathscr{L}_{\text{vec}}$.

4 Experimental Results

We evaluate our framework through two types of experiments. Firstly, we perform a qualitative evaluation through a visual examination of the results. Secondly, we define a set of objective metrics to quantitatively evaluate the results.

Since there currently is no RGB-D face completion method that we can directly compare our framework to, *separately* trained RGB image and depth image inpainting models of the base framework [55] form a comparative baseline. These models are trained with the CA module and an objective function comprised of the ℓ_1 loss and $\mathscr{L}_{\text{GAN}}$. The results of these unimodal models demonstrate the level of visual quality we aim to match with our multimodal RGB-D image inpainting models.

Furthermore, we evaluate our framework by comparing the performance of models trained with different sets of components. Specifically, with each model, we add one of the components to the framework to evaluate their impact. Therefore, our evaluation covers models trained with the following configurations: 1) $\ell_1 + \mathscr{L}_{\text{GAN}}$, 2) $\ell_1 + \mathscr{L}_{\text{GAN}} + \mathscr{L}_{\text{ID}}$, 3) $\ell_1 + \mathscr{L}_{\text{GAN}} + \mathscr{L}_{\text{ID}} + \mathscr{L}_{\text{vec}}$, and 4) $\ell_1 + \mathscr{L}_{\text{GAN}} + \mathscr{L}_{\text{ID}} + \mathscr{L}_{\text{vec}}$ with CSA.

4.1 Data Generation

Unlike the wide availability of large RGB face image datasets [8, 19], similarly sized datasets containing RGB-D face images are not available at this time. Due to the dataset size requirements of the training procedure of GANs, we opted to create a synthetic dataset with a high degree of realism and variety.

To facilitate this process, we built a data synthesization pipeline to create a synthetic dataset of RGB-D images of faces based on BFM 2017 [16], a parametric 3D Morphable Model (3DMM) [5] learned from 3D scans of human faces. While the usage of this synthetic dataset reduces the potential of generalization to real-world data, it does provide exact controls over facial expression, pose, and ambient illumination.

Our pipeline starts by taking a random sample across the independent shape, texture, and expression parameters of the BFM 2017 [16] model. Following this, relative to the resulting mesh, we place a predefined mesh of an HMD representing the true dimensions of an Oculus Rift. Next, the compound mesh is placed in world space, at distance d from the camera position, which we sensibly set to 85 centimeters. Finally, a simple ray tracing algorithm is responsible for rendering the color image as well as a corresponding depth image. Effectively, we render three images of 224×224 pixels in size: a color image of the face without the HMD, depth image of the face without the HMD, and binary image of solely the HMD.

We jointly encode the color image and depth image of the face in the four channels of a single 8-bit PNG file. We opt for a depth resolution of 1 mm per pixel for two reasons. Firstly, the approximate resolution of the Microsoft Kinect is 1.3mm per pixel [33], one of the most widely-used commodity RGB-D sensors currently available. Secondly, this choice allows us to encode the depth image in the alpha channel of the 8-bit PNG file. We achieve this by subtracting the scalar d from the inverted depth image, which

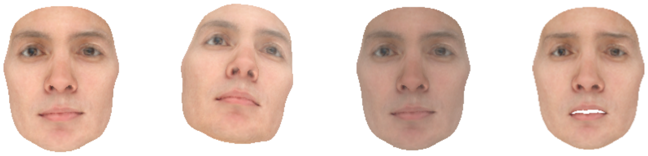


Figure 3: Examples of individual transformations, from left to right: original (frontal view with full ambient illumination), random pose, random ambient illumination, and random expression.

enables the depth values to fit in the 8-bit alpha channel. This is due to the fact that faces will never realistically have a depth that exceeds 256 millimeters. We have made the implementation of the RGB-D rendering segment of this pipeline publicly available as `mesh2rgbd`[2], which is based on `face3d`[3] by Y. Feng. The resulting dataset consists of 48 000 RGB-D face images with the following properties and transformations: 1) *random expression*, 2) *random pose* p: with p_{pitch} and p_{yaw} in range $[-30°, 30°]$ and p_{roll} in range $[-20°, 20°]$, 3) *random ambient illumination* a: with $a_{intensity}$ in range $[80, 110]$. We split this dataset into sets of sizes: 40 000, 4000, 4000, for training, validation, and testing, respectively. Examples of our dataset are shown in Fig. 3.

4.2 Implementation Details and Setup

We implemented our framework in TensorFlow v1.15, based on the source code by Yu et al. [55]. Training of our models was performed on two NVIDIA GeForce RTX 2080 Ti GPUs. This process typically takes approximately 2.5 days, but is continued until convergence of the losses and stabilization of the visual quality of the validation results. We trained and evaluated our models with the aforementioned dataset containing images of 224×224 pixels. For training, we use a learning rate of 0.0001 for both the generator and discriminator and a batch size of 10. Moreover, through hyperparameter tuning based on a combination of visual examination and objective metrics, we determined the default hyperparameter balance of 3:1:1:1, for the ℓ_1 reconstruction loss, SN-PatchGAN loss $\mathscr{L}_{\text{GAN}}$, identity loss $\mathscr{L}_{\text{ID}}$, and vectorial loss $\mathscr{L}_{\text{vec}}$ respectively. We have made the source code of our framework publicly available[4].

At inference with a single NVIDIA GeForce RTX 2080 Ti, our framework achieves an average frame rate of 48 frames per second. While our method currently does not leverage any temporal modality and further research is needed, the performance of our framework theoretically permits real-time RGB-D video inpainting.

4.3 Qualitative Results

Whereas color images can be presented straightforwardly, RGB-D images require additional representations to visualize the geometric characteristics contained in the depth channel. Therefore, we provide three representations for each inpainted RGB-D image: an RGB color image, depth image, and estimated surface normal image. The estimated surface normal image is calculated based on the depth image using the method outlined by Matias et al. [34]. The qualitative results of the specified model configurations are shown in Fig. 4a.

Furthermore, Fig. 5 presents two examples where the model receives a incomplete image of one identity and a reference image of another. The model used for this experiment was trained with an objective function including the identity loss $\mathscr{L}_{\text{ID}}$.

4.4 Quantitative Results

The quantitative evaluation of generative models remains challenging, as consensus has not been reached with respect to a standardized set of objective metrics [6]. Despite this, they do provide us with an empirical base to compare the defined model configurations.

We define our objective metrics to align with the evaluation of existing image inpainting methods, to agree with human perceptual judgement, and to reflect the inherent challenges of HMD removal in RGB-D images. Based on these considerations we report the following metrics: 1) mean ℓ_1 error, 2) mean ℓ_2 error, 3) Peak Signal-to-Noise Ratio (PSNR), 4) Structural Similarity (SSIM) index [51], 5) Visual Information Fidelity (VIF) index [43]. Notably, there is evidence to suggest that the VIF index of depth images is correlated with the quality of experience of 3D video compared to other quality metrics [3]. Therefore, this is an insightful addition to our set of evaluation metrics.

In addition, we define a metric to quantify and compare the degree of identity preservation. This metric uses the FaceNet [42] model N, pretrained on the MS-Celeb-1M [19] dataset which produces a 128-byte vector representing the subject's identity. Similar to the identity loss, the identity error is calculated based on only the RGB channels. To obtain the identity error, we calculate the Euclidean distance between the identity vector of the ground truth image x and inpainted image $\hat{x}$ respectively. Accordingly, the identity error is calculated as follows:

$$\text{ID}(x, \hat{x}) = ||N(x) - N(\hat{x})||_2 \tag{2}$$

We present the quantitative results of each model configuration in Fig. 4b.

5 Analysis and Discussion

In this paper, we set out to jointly inpaint color and depth information in occluded RGB-D face images. As image inpainting approaches typically handle unimodal data, it is of interest to compare the results of the unimodal models (Ⓐ) and the simplest version of our multimodal model (Ⓑ). In both Fig. 4a and Fig. 4b, we observe a notable deterioration of the results of the multimodal inpainting model compared to the results of the unimodal inpainting models. Both the resulting color and depth images contain a significant amount of noise and artifacts. This suggests that the feature construction process in the multimodal inpainting model is unable to jointly capture the information in the color and depth channels, worsening its predictive capabilities.

Nonetheless, considering the difficulty of multimodal feature construction, the results of the multimodal model (Ⓑ) are of notable quality and demonstrate the viability of data-level fusion approaches to RGB-D image inpainting. A natural progression of this work would be to further explore the application of feature-level fusion methods and depth-aware convolution [9, 50] to image inpainting. Moreover, future research might explore how our method compares to the recently introduced RGB-D image inpainting method by Fujii et al. [15].

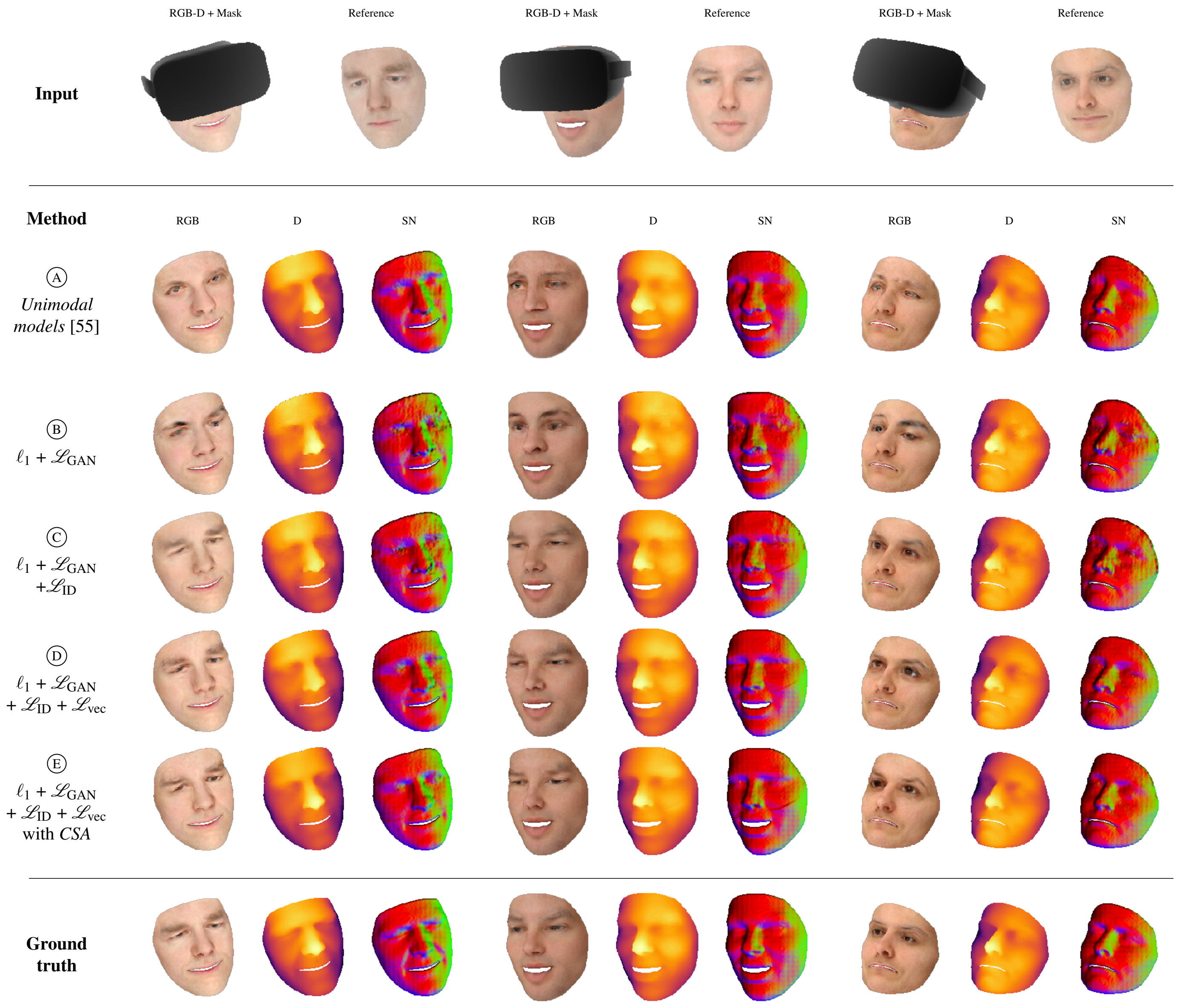
5.1 Identity Preservation

We now discuss the effect of our identity loss $\mathscr{L}_{\text{ID}}$, which stimulates the preservation of identifying facial features. In Fig. 4a, we observe that the model trained with our identity loss (Ⓒ) produces identity features that are consistent with the provided reference image. This effect is reflected empirically in Fig. 4b, where we see improved results across all metrics. Moreover, we observe an improved symmetry of facial features in Fig. 4a. However, this does not extend to eye color, as we regularly observe differing eye colors in a single inpainted image. A possible cause is that an incorrect eye color has a minor impact on the loss values considering the relatively small image region it comprises.

[2] `https://github.com/nsalminen/mesh2rgbd`

[3] `https://github.com/YadiraF/face3d`

[4] `https://github.com/nsalminen/HMDRemoval`

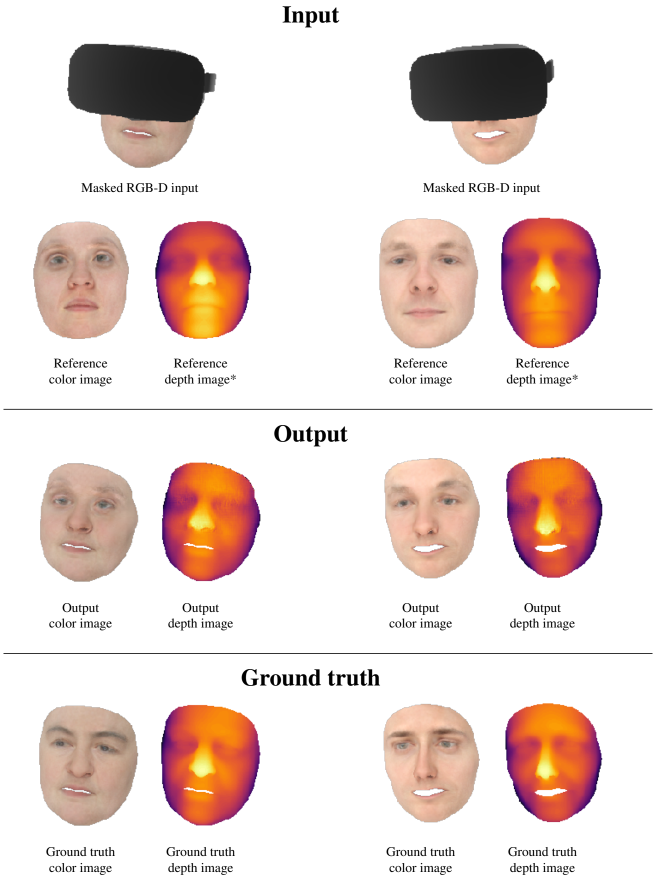

RGB-D + Mask
Reference
RGB-D + Mask
Reference
RGB-D + Mask
Reference
Input
Method
RGB
D
SN
RGB
D
SN
RGB
D
SN
Ⓐ *Unimodal models* [55]
Ⓑ $\ell_1 + \mathscr{L}_{\text{GAN}}$
Ⓒ $\ell_1 + \mathscr{L}_{\text{GAN}} + \mathscr{L}_{\text{ID}}$
Ⓓ $\ell_1 + \mathscr{L}_{\text{GAN}} + \mathscr{L}_{\text{ID}} + \mathscr{L}_{\text{vec}}$
Ⓔ $\ell_1 + \mathscr{L}_{\text{GAN}} + \mathscr{L}_{\text{ID}} + \mathscr{L}_{\text{vec}}$ with *CSA*
Ground truth

(a) Comparative qualitative results, shown for color (RGB), depth (D), and estimated surface normals (SN). For visualization, D is normalized to [0, 1] and displayed with the *inferno* colormap from the `matplotlib` package. The normal vectors (x, y, z) for each pixel in SN are estimated based on D and are visualized with RGB values.

	ℓ_1 **error**			ℓ_2 **error**			**PSNR**			**SSIM**			**VIF**			**ID**
Method	RGB	D	SN	RGB	D	SN	RGB	D	SN	RGB	D	SN	RGB	D	SN	RGB
Ⓐ *Unimodal models* [55]	11.333	4.346	20.991	27.500	17.666	37.086	18.839	23.519	16.832	0.913	0.972	0.884	0.488	0.660	0.490	11.780
Ⓑ $\ell_1 + \mathscr{L}_{\text{GAN}}$	11.813	5.109	26.949	28.318	18.716	45.523	18.587	23.049	15.025	0.915	0.968	0.858	0.490	0.640	0.453	12.157
Ⓒ $\ell_1 + \mathscr{L}_{\text{GAN}} + \mathscr{L}_{\text{ID}}$	**8.155**	3.765	23.928	21.867	15.315	40.202	20.810	24.662	16.101	**0.936**	0.975	0.867	**0.528**	0.664	0.465	7.965
Ⓓ $\ell_1 + \mathscr{L}_{\text{GAN}} + \mathscr{L}_{\text{ID}} + \mathscr{L}_{\text{vec}}$	8.515	**3.500**	19.355	22.161	14.979	33.925	20.686	24.836	17.615	0.933	**0.976**	**0.893**	0.524	0.670	**0.495**	**7.851**
Ⓔ $\ell_1 + \mathscr{L}_{\text{GAN}} + \mathscr{L}_{\text{ID}} + \mathscr{L}_{\text{vec}}$ with *CSA*	8.363	3.612	**19.087**	**21.770**	**14.927**	**33.464**	**20.875**	**24.890**	**17.719**	0.934	**0.976**	**0.893**	0.527	**0.675**	**0.495**	7.891

(b) Comparative quantitative results, shown for color (RGB), depth (D), and estimated surface normals (SN). We report the following metrics: mean ℓ_1 error (lower is better), mean ℓ_2 error (lower is better), Peak Signal-to-Noise Ratio (PSNR) (higher is better), Structural Similarity (SSIM) index [51] (higher is better), Visual Information Fidelity (VIF) index [43] (higher is better).

Figure 4: Results of models with different configurations. Ⓐ shows results originating from two distinct models, each trained to inpaint either color or depth images. Ⓑ-Ⓔ show models with different configurations, where loss functions and components are added to demonstrate their impact on the visual results.



* Depth images are not passed as input to the model and are only shown for comparison purposes.

Figure 5: Results of model input containing differing identities, generated by a model trained *with* our identity loss function $\mathscr{L}_{\text{ID}}$. Results are shown for color (RGB) and depth (D).

Perhaps the most insightful demonstration of the effects of the identity loss function is performed by feeding the model with a masked image of one identity and a reference image of another, as shown in Fig. 5. The inpainted results show faces that are globally consistent with the known region of the image, while also containing distinct facial features from the given reference image. Interestingly, while the identity loss is calculated based on the RGB channels, the depth channel of the inpainted images show similar facial features to the given reference image. This indicates that the model learns the relation between the identity loss and depth image indirectly, based on the feedback it receives regarding the inpainted RGB channels. This effect is visible in both Fig. 4a and Fig. 5.

5.2 Reproduction of Geometric Surfaces

Our framework contains a loss function and module that are focused on the improvement of surface reproduction. The qualitative and quantitative results of the model trained with the vectorial loss function $\mathscr{L}_{\text{vec}}$ (Ⓓ) appear to be in agreement, as a clear improvement of the smoothness of the inpainted depth image and its surface normal representation can be identified in both Fig. 4a and Fig. 4b. This is similar to the findings of Matias et al. [34], who proposed this function for depth image inpainting. However, this comes at a minor cost as seen in Fig. 4b, which shows decreased quality of inpainted RGB channels (Ⓓ). Furthermore, we note an inconsistent connection between the inpainted and known region of the image in some cases. A likely explanation is that an inconsistent connection at the boundary of the inpainted region has a limited impact on the value of the vectorial loss function.

In terms of the CSA module [34], we found no evident effect to indicate improvement of the visual quality of the results of this model (Ⓔ) in Fig. 4a. However, the quantitative results of the model with the CSA module presented in Fig. 4b show improved results across nearly all metrics. The most likely cause of this overall improvement is the addition of the estimated surface normal image to the CA branch of the network, which enhances the feature matching process of the CA module.

5.3 Real-world Data and Applications

GANs aim to learn complex distributions in order to generate samples with a broad diversity and level of detail. In our search for a suitable RGB-D face dataset, we concluded that a sufficiently sized dataset for training a GAN is not currently available. Our choice to train and evaluate our framework with a synthetic dataset came with several benefits, including the ease of dataset creation and full control over identity and recording conditions.

However, this decision also has some serious drawbacks. Firstly, BFM 2017 [16] is based on a total of 200 facial shape and texture captures. While this has no direct consequence to the number of identities we can sample from the model, it does limit the size of its parameter space, which affects the sample diversity. This has a significant impact on the bias and generalizability of our framework. Secondly, our synthetic dataset contains *perfect* RGB-D images, without noise, misalignment, or artifacts, frequently found in real-world RGB-D recordings. This being so, our trained models are not robust with respect to these types of data characteristics, as it has not been made familiar with them during training. This not only affects the applicability of our framework to real-world data, but also forms a problem for its evaluation, as we cannot ensure that the explored components would behave similarly on real-world data.

That being said, our synthetic dataset forms a substantial base for the exploration of joint RGB-D image inpainting. Our findings have important implications for models trained on RGB-D data and we hypothesize that our models can be fine-tuned through training on a real-world dataset. Additional research is needed to investigate socially evocative factors such as the reproduction of eye gaze and expression, which several model-based methods have previously explored [13, 28, 37, 47, 58].

6 Conclusion

HMD removal is a challenging task which has emerged with the increasing usage of IVEs for social VR applications [4, 10, 11, 26, 35, 40, 41]. In this paper, we proposed a framework that is capable of the virtual removal of HMDs in RGB-D images. We formulated this problem as a joint RGB-D image inpainting task and proposed a framework that is capable of simultaneously filling in the missing color and depth information of face images occluded by an HMD. To preserve the identity features of the inpainted faces, we proposed an RGB-based identity loss function. Moreover, to improve surface reproduction in the depth channel, we employed the vectorial loss and CSA module proposed by Matias et al. [34]. In absence of a large-scale RGB-D face dataset, we devised a pipeline to create a synthetic RGB-D face dataset. Based on this dataset, we performed qualitative and quantitative experiments to demonstrate the performance of each component and showed our framework's robustness against expression, pose, and ambient illumination. Despite its exploratory nature and limitations, our research offers unique insights into the design and behavior of a multimodal image inpainting framework that can be of interest to future research.

Acknowledgments

The authors would like to show their gratitude to the members of the Intelligent Imaging group at TNO for their support and to thank the anonymous reviewers for their insightful comments.

References

[1] A. Atapour-Abarghouei and T. Breckon. Depthcomp : real-time depth image completion based on prior semantic scene segmentation. In *28th British Machine Vision Conference (BMVC) 2017*. British Machine Vision Association (BMVA), September 2017.

[2] J. N. Bailenson, N. Yee, D. Merget, and R. Schroeder. The effect of behavioral realism and form realism of real-time avatar faces on verbal disclosure, nonverbal disclosure, emotion recognition, and copresence in dyadic interaction. *Presence: Teleoperators and Virtual Environments*, 15(4):359–372, 2006. doi: 10.1162/pres.15.4.359

[3] A. Banitalebi-Dehkordi, M. T. Pourazad, and P. Nasiopoulos. A study on the relationship between depth map quality and the overall 3d video quality of experience. In *2013 3DTV Vision Beyond Depth (3DTV-CON)*, pp. 1–4. IEEE, 2013. doi: 10.1109/3dtv.2013.6676650

[4] S. Beck, A. Kunert, A. Kulik, and B. Froehlich. Immersive group-to-group telepresence. *IEEE Transactions on Visualization and Computer Graphics*, 19(4):616–625, 2013. doi: 10.1109/tvcg.2013.33

[5] V. Blanz, T. Vetter, et al. A morphable model for the synthesis of 3d faces. In *Siggraph*, vol. 99, pp. 187–194, 1999. doi: 10.1145/311535.311556

[6] A. Borji. Pros and cons of gan evaluation measures. *Computer Vision and Image Understanding*, 179:41–65, 2019. doi: 10.1016/j.cviu.2018.10.009

[7] V. Bruce and A. Young. Understanding face recognition. *British journal of psychology*, 77(3):305–327, 1986. doi: 10.1111/j.2044-8295.1986.tb02199.x

[8] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *International Conference on Automatic Face and Gesture Recognition*, 2018. doi: 10.1109/fg.2018.00020

[9] Y. Chen, T. Mensink, and E. Gavves. 3d neighborhood convolution: Learning depth-aware features for rgb-d and rgb semantic segmentation. In *2019 International Conference on 3D Vision (3DV)*, pp. 173–182. IEEE, 2019. doi: 10.1109/3dv.2019.00028

[10] N. Didehbani, T. Allen, M. Kandalaft, D. Krawczyk, and S. Chapman. Virtual reality social cognition training for children with high functioning autism. *Computers in Human Behavior*, 62:703–711, 2016. doi: 10.1016/j.chb.2016.04.033

[11] S. Dijkstra-Soudarissanane, K. E. Assal, S. Gunkel, F. t. Haar, R. Hindriks, J. W. Kleinrouweler, and O. Niamut. Multi-sensor capture and network processing for virtual reality conferencing. In *Proceedings of the 10th ACM Multimedia Systems Conference*, MMSys '19, p. 316–319. Association for Computing Machinery, New York, NY, USA, 2019. doi: 10.1145/3304109.3323838

[12] D. Doria and R. J. Radke. Filling large holes in lidar data by inpainting depth gradients. In *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pp. 65–72. IEEE, 2012. doi: 10.1109/cvprw.2012.6238916

[13] C. Frueh, A. Sud, and V. Kwatra. Headset removal for virtual and mixed reality. In *ACM SIGGRAPH 2017 Talks*, p. 80. ACM, 2017. doi: 10.1145/3084363.3085083

[14] R. Fujii, R. Hachiuma, and H. Saito. Joint inpainting of rgb and depth images by generative adversarial network with a late fusion approach. In *2019 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, pp. 203–204. IEEE, 2019. doi: 10.1109/ismar-adjunct.2019.00-46

[15] R. Fujii, R. Hachiuma, and H. Saito. Rgb-d image inpainting using generative adversarial network with a late fusion approach. In *International Conference on Augmented Reality, Virtual Reality and Computer Graphics*, pp. 440–451. Springer, 2020.

[16] T. Gerig, A. Morel-Forster, C. Blumer, B. Egger, M. Luthi, S. Schönborn, and T. Vetter. Morphable face models-an open framework. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pp. 75–82. IEEE, 2018. doi: 10.1109/fg.2018.00021

[17] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pp. 2672–2680, 2014.

[18] S. N. Gunkel. [dc] multi-user (social) virtual reality commnunication. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 1359–1360. IEEE, 2019.

[19] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *European conference on computer vision*, pp. 87–102. Springer, 2016. doi: 10.1007/978-3-319-46487-9_6

[20] C. Hazirbas, L. Ma, C. Domokos, and D. Cremers. Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture. In *Asian conference on computer vision*, pp. 213–228. Springer, 2016. doi: 10.1007/978-3-319-54181-5_14

[21] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016. doi: 10.1109/cvpr.2016.90

[22] D. Herrera, J. Kannala, J. Heikkilä, et al. Depth map inpainting under a second-order smoothness prior. In *Scandinavian Conference on Image Analysis*, pp. 555–566. Springer, 2013. doi: 10.1007/978-3-642-38886-6_52

[23] R. Huang, S. Zhang, T. Li, and R. He. Beyond face rotation: Global and local perception gan for photorealistic and identity preserving frontal view synthesis. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2439–2448, 2017. doi: 10.1109/iccv.2017.267

[24] S. Iizuka, E. Simo-Serra, and H. Ishikawa. Globally and locally consistent image completion. *ACM Transactions on Graphics (ToG)*, 36(4):107, 2017. doi: 10.1145/3072959.3073659

[25] J. Jansen, S. Subramanyam, R. Bouqueau, G. Cernigliaro, M. Martos Cabré, F. Pérez, and P. S. César Garcia. A pipeline for multiparty volumetric video conferencing: Transmission of point clouds over low latency dash. In *MMSys 2020 - Proceedings of the 2020 Multimedia Systems Conference*, pp. 341–344, May 2020. doi: 10.1145/3339825.3393578

[26] E. Klinger, S. Bouchard, P. Légeron, S. Roy, F. Lauer, I. Chemin, and P. Nugues. Virtual reality therapy versus cognitive behavior therapy for social phobia: A preliminary controlled study. *Cyberpsychology & behavior*, 8(1):76–88, 2005. doi: 10.1089/cpb.2005.8.76

[27] M. E. Latoschik, D. Roth, D. Gall, J. Achenbach, T. Waltemate, and M. Botsch. The effect of avatar realism in immersive social virtual realities. In *Proceedings of the 23rd ACM Symposium on Virtual Reality Software and Technology*, pp. 1–10, 2017. doi: 10.1145/3139131.3139156

[28] H. Li, L. Trutoiu, K. Olszewski, L. Wei, T. Trutna, P.-L. Hsieh, A. Nicholls, and C. Ma. Facial performance sensing head-mounted display. *ACM Transactions on Graphics (ToG)*, 34(4):47, 2015. doi: 10.1145/2766939

[29] M. Li, W. Zuo, and D. Zhang. Deep identity-aware transfer of facial attributes. *arXiv preprint arXiv:1610.05586*, 2016.

[30] S. Lin, Y. Chen, Y.-K. Lai, R. R. Martin, and Z.-Q. Cheng. Fast capture of textured full-body avatar with rgb-d cameras. *The Visual Computer*, 32(6-8):681–691, 2016.

[31] 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. doi: 10.1109/iccv.2015.425

[32] Z. Liu, H. Qin, S. Bu, M. Yan, J. Huang, X. Tang, and J. Han. 3d real human reconstruction via multiple low-cost depth cameras. *Signal Processing*, 112:162–179, 2015.

[33] Q. Luo and G. Yang. Research and simulation on virtual movement based on kinect. In *International Conference on Virtual, Augmented and Mixed Reality*, pp. 85–92. Springer, 2014. doi: 10.1007/978-3-319-07458-0_9

[34] L. P. Matias, M. Sons, J. R. Souza, D. F. Wolf, and C. Stiller. Veigan: Vectorial inpainting generative adversarial network for depth maps object removal. In *2019 IEEE Intelligent Vehicles Symposium (IV)*, pp. 310–316. IEEE, 2019. doi: 10.1109/ivs.2019.8814157

[35] T. Monahan, G. McArdle, and M. Bertolotto. Virtual reality for collaborative e-learning. *Computers & Education*, 50(4):1339–1353, 2008. doi: 10.1016/j.compedu.2006.12.008

[36] S. Mori, J. Herling, W. Broll, N. Kawai, H. Saito, D. Schmalstieg, and D. Kalkofen. 3d pixmix: Image inpainting in 3d environments. In

2018 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct), pp. 1–2. IEEE, 2018. doi: 10.1109/ismar-adjunct.2018.00020

[37] K. Olszewski, J. J. Lim, S. Saito, and H. Li. High-fidelity facial and speech animation for vr hmds. *ACM Transactions on Graphics (TOG)*, 35(6):1–14, 2016. doi: 10.1145/2980179.2980252

[38] S.-J. Park, K.-S. Hong, and S. Lee. Rdfnet: Rgb-d multi-level residual feature fusion for indoor semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4980–4989, 2017. doi: 10.1109/iccv.2017.533

[39] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2536–2544, 2016. doi: 10.1109/cvpr.2016.278

[40] M. J. Prins, S. N. Gunkel, H. M. Stokking, and O. A. Niamut. Togethervr: A framework for photorealistic shared media experiences in 360-degree vr. *SMPTE Motion Imaging Journal*, 127(7):39–44, 2018. doi: 10.5594/jmi.2018.2840618

[41] D. J. Roberts, A. S. Garcia, J. Dodiya, R. Wolff, A. J. Fairchild, and T. Fernando. Collaborative telepresence workspaces for space operation and science. In *2015 IEEE Virtual Reality (VR)*, pp. 275–276. IEEE, 2015. doi: 10.1109/vr.2015.7223402

[42] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 815–823, 2015. doi: 10.1109/cvpr.2015.7298682

[43] H. R. Sheikh and A. C. Bovik. Image information and visual quality. *IEEE Transactions on image processing*, 15(2):430–444, 2006. doi: 10.1109/tip.2005.859378

[44] J. Shen and S.-C. S. Cheung. Layer depth denoising and completion for structured-light rgb-d cameras. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1187–1194, 2013. doi: 10.1109/cvpr.2013.157

[45] Y. Shen, B. Zhou, P. Luo, and X. Tang. Facefeat-gan: A two-stage approach for identity-preserving face synthesis. *arXiv preprint arXiv:1812.01288*, 2018.

[46] T. L. Taylor. Living digitally: Embodiment in virtual worlds. In *The social life of avatars*, pp. 40–62. Springer, 2002. doi: 10.1007/978-1-4471-0277-9_3

[47] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner. Facevr: Real-time gaze-aware facial reenactment in virtual reality. *ACM Transactions on Graphics (TOG)*, 37(2):1–15, 2018. doi: 10.1145/3182644

[48] L. Wang, H. Jin, R. Yang, and M. Gong. Stereoscopic inpainting: Joint color and depth completion from stereo images. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8. IEEE, 2008. doi: 10.1109/cvpr.2008.4587704

[49] M. Wang, X. Wen, and S.-M. Hu. Faithful face image completion for hmd occlusion removal. In *2019 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, pp. 251–256. IEEE, 2019. doi: 10.1109/ismar-adjunct.2019.00-36

[50] W. Wang and U. Neumann. Depth-aware cnn for rgb-d segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 135–150, 2018. doi: 10.1007/978-3-030-01252-6_9

[51] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli, et al. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. doi: 10.1109/tip.2003.819861

[52] S.-E. Wei, J. Saragih, T. Simon, A. W. Harley, S. Lombardi, M. Perdoch, A. Hypes, D. Wang, H. Badino, and Y. Sheikh. Vr facial animation via multiview image translation. *ACM Transactions on Graphics (TOG)*, 38(4):1–16, 2019. doi: 10.1145/3306346.3323030

[53] R. A. Yeh, C. Chen, T. Yian Lim, A. G. Schwing, M. Hasegawa-Johnson, and M. N. Do. Semantic image inpainting with deep generative models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5485–5493, 2017. doi: 10.1109/cvpr.2017.728

[54] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang. Generative image inpainting with contextual attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5505–5514, 2018. doi: 10.1109/cvpr.2018.00577

[55] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang. Free-form image inpainting with gated convolution. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4471–4480, 2019. doi: 10.1109/iccv.2019.00457

[56] Y. Zhang and T. Funkhouser. Deep depth completion of a single rgb-d image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 175–185, 2018. doi: 10.1109/cvpr.2018.00026

[57] Y. Zhao, W. Chen, J. Xing, X. Li, Z. Bessinger, F. Liu, W. Zuo, and R. Yang. Identity preserving face completion for large ocular region occlusion. *arXiv preprint arXiv:1807.08772*, 2018.

[58] Y. Zhao, Q. Xu, W. Chen, C. Du, J. Xing, X. Huang, and R. Yang. Mask-off: Synthesizing face images in the presence of head-mounted displays. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 267–276. IEEE, 2019. doi: 10.1109/vr.2019.8797925