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

Nels Numan

Delft University of Technology

TU Delft | TNO

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

by

Nels Numan

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Tuesday October 20, 2020 at 02:00 PM.

Student number:	4615107	
Project duration:	July 1, 2019 – October 1, 2020	
Thesis committee:	Dr. Pablo Cesar	TU Delft, supervisor
	Dr. Frank ter Haar	TNO, supervisor
	Dr. Ricardo Marroquim	TU Delft
	Dr. Julian Kooij	TU Delft

An electronic version of this thesis will be available at http://repository.tudelft.nl/.

TU Delft

Abstract

Virtual reality (VR) creates an exceptional experience in which users can explore virtual environments. Wearing a head-mounted display (HMD), users are able to observe a virtual world that is rendered based on their physical movement and actions. A common solution for capturing the visual and geometric information needed for the construction of virtual environments is the use of RGB-D sensors. These sensors not only capture a collection of RGB data like conventional cameras do, but additionally record a depth value for each pixel. Thus, RGB-D sensors are able to capture both the visual and geometric properties of a space, including any objects or people. This makes immersive social VR experiences possible, where people in different physical locations can be placed in the same virtual environment. However, HMDs obstruct the RGB-D sensor from capturing the wearer's upper face, which severely impacts the social aspects of VR applications. To address this, we proposed a framework that is capable of the virtual removal of head-mounted displays in RGB-D images, which is referred to as the task of HMD removal. Due to its novelty, we took an exploratory approach to this task.

We formulated this problem as a joint RGB-D face image inpainting task and proposed a GAN-based coarse-to-fine architecture 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. Leveraging the knowledge of a pretrained identity embedding model, this perceptual loss function stimulates the preservation of identity-specific facial features.

Furthermore, we proposed several architectural structures to explore multimodal feature fusion of the color and depth information contained in RGB-D images. To this end, we introduced data-level fusion, which naively combines the color and depth information at network input. In addition, we introduced hybrid fusion, which involves feature-level fusion in the coarse stage of the architecture and data-level fusion in the refinement stage of the architecture. Within the concept of hybrid fusion, we investigated several fusion strategies, including residual fusion. Our findings suggest that data-level fusion achieves similar performance to hybrid fusion.

Moreover, to improve surface reproduction in the depth channel, we introduced the employment of a surface normal loss function and contextual surface attention module, which both rely on surface normals that are estimated based on the depth channel of the RGB-D image. We also considered the addition of surface normal information to the discriminator input, which we found to have an adverse effect on the visual quality of the results.

In absence of a large scale RGB-D face dataset, we devised a pipeline for the creation of a synthetic RGB-D face dataset for the evaluation of our network. Despite its exploratory nature, our research provides unique insights into the design and behavior of a multimodal image inpainting architecture that can be of interest to future research.

Acknowledgements

I would like to express my sincerest gratitude to the following people.

- My supervisors, Frank ter Haar and Pablo Cesar, for providing invaluable guidance through their constructive feedback, great patience and extensive knowledge.
- My colleagues at the Intelligent Imaging department of TNO, for their exciting suggestions, stimulating feedback and infectious enthusiasm.
- My mom and sister, for their endless love, support and encouragement.
- My grandparents, for always believing in me.
- My friends, including Laure, Alenka, Pia, Lisa, Marinka, Rafi, Kostas, and Bohye, for their help, understanding and unconditional support.

Contents

List of Figures

List of Tables

1 Introduction

Virtual reality (VR) creates an exceptional experience in which users can explore virtual environments. Wearing a head-mounted display (HMD), users are able to observe a virtual world that is rendered based on their physical movement and actions. The natural interface that this technology offers has enabled a wide range of simulations, which are too complex, hazardous or costly for execution in the real world. While HMDs form an essential virtual display device for VR, HMDs obstruct any form of external observation of the wearer's upper face, which severely impacts the social aspects of VR applications. In this thesis, we propose an image-based method for the virtual removal of HMDs, which coherently fills in the occluded color and geometric information of the wearer's face represented in an RGB-D image.

1.1. Background and context

Early works [1–3] explored the social implications of VR and conceptualized the field that we now refer to as social virtual reality. This concept involves the assembly of a group of people in the same VR environment that supports some form of human-to-human communication and collaboration. In subsequent years, technological advances in VR technology have resulted in ongoing research efforts towards the creation of immersive social VR experiences such as collaborative learning [4, 5], entertainment [6], treatment of mental disorders [7, 8] and teleconferencing [9–12].

When compared to face-to-face interaction, computer-mediated interactions in virtual environments inherently convey less social and contextual cues [13]. During a face-to-face interaction, these cues are effortlessly transmitted. However, the effects of similar interactions in VR depend entirely on the capabilities of the mediating technologies [14]. When simulating a real social interaction in a virtual environment, we want the user experience to be as realistic as possible [15]. The virtual representation of a human, also referred to as an avatar, plays a fundamental role in this type of situation [16]. Realism and visual quality of avatars in virtual environments are seen as factors that drive the experience of *being* with another person [17, 18], commonly referred to as copresence or social presence.

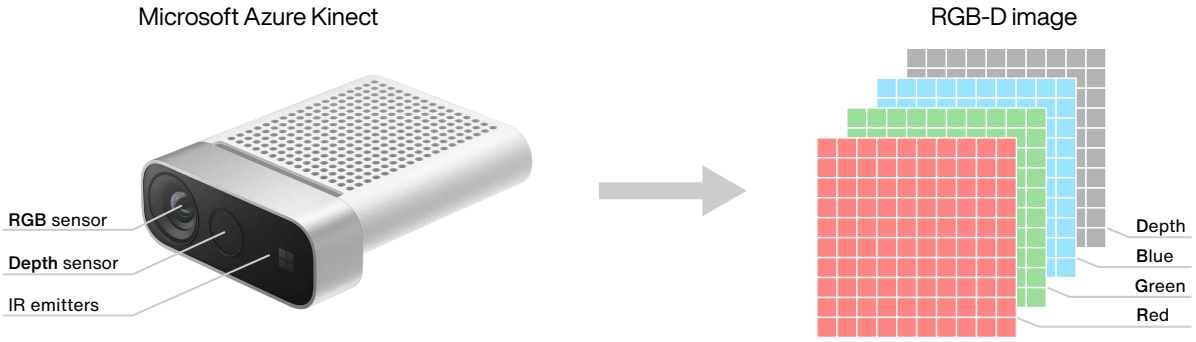


Figure 1.1: An example of an RGB-D sensor: the Microsoft Azure Kinect [26], including an RGB sensor, depth sensor and accompanying IR emitters (left). The outputs of these sensors can be combined and mapped to each other, resulting in an RGB-D image (right).

However, some commercial solutions have been introduced that aim to achieve copresence through a cartoon-like avatar, such as Facebook Horizon [19] and AltspaceVR [20]. By avoiding any attempt to achieve high ecological validity and photorealistic human likeness, solutions of this kind reduce the risk of a potential uncanny valley [21, 22] created by flaws of the mediating technology or representation [23, 24]. Other ways of avoiding an uncanny valley have been explored through research towards the creation of near perfect photorealistic human representations [25]. These methods aim to fully alleviate the negative impact a virtual embodiment may have on social interaction in a virtual environment [14].

A common solution for capturing the visual and geometric information needed for these representations is the use of commercially available RGB-D sensors, such as the Intel RealSense [27] and Microsoft Azure Kinect [26] (Figure 1.1). These sensors not only capture a collection of RGB data like conventional cameras do, but additionally record a depth value for each pixel. This data can in turn be displayed in a shared virtual environment in a visually and geometrically consistent way. Shown in Figure 1.2, an example of such a setup is the TogetherVR platform introduced by Dijkstra-Soudarissanane et al. [12] at the Netherlands Organization of Applied Science (TNO), where our thesis research was carried out. This system accommodates remote communication and collaboration through the creation of a shared virtual environment where up to four people can take place at a virtual table. Each user is captured using a pair of RGB-D sensors, the output of which is transformed into a point cloud and placed in a shared virtual environment. Wearing an HMD, each user is able to observe the shared virtual environment and communicate with up to four other users.

However, the HMD worn by the user occludes the upper part of their face and prevents the sensor from capturing it. As a result, during social interaction in a virtual environment such as the aforementioned, it is a challenge for users to estimate gaze direction [28], make eye contact [29], interpret non-verbal information [11, 30, 31] or to recognize the identity of others. For this reason, we want to reconstruct the occluded region of the user's face in a realistic way, which is referred to as the task of HMD removal.

A critical issue arises when it comes to resolving the missing facial region: how does one know what visual content to replace this occluded facial region with? A number of approaches involve an offline process to record a dynamic 3D face model [32–35]; which require HMDs fitted with internal infrared cameras [34] or RGB-D cameras [32]. Aside from the necessity of

Figure 1.2: Test setup of the TogetherVR platform introduced by Dijkstra-Soudarissanane et al. [12]. In this case, two HMD-wearing subjects are captured with RGB-D sensors (left). In turn, the subjects are placed in a shared virtual environment (right). For testing purposes, the two subjects are located in the same physical space.

custom equipment, many of these methods require elaborate calibration and setup processes. This prompts the question whether there are other ways to fill in the facial region occluded by an HMD.

A number of approaches have been proposed that aim to solve the task of HMD removal by synthesizing the occluded facial region through image inpainting [36]. Image inpainting, also known as image completion, describes the task of filling undesired or unknown pixel regions with realistic content. Recent progress in generative adversarial networks (GANs) [37] has inspired a wide range of image inpainting methods [38–41], which comprise an adversarial training process between a generator network and a discriminator network. This process aims to capture the high-level semantic and low-level pixel information of ground truth images in order to generate realistic content for missing image regions. GAN-based image inpainting methods have achieved state-of-the-art results that contain complex structures such as buildings, landscapes, animals and human faces. Although the aforementioned HMD removal and image inpainting methods have been proven to perform well with RGB image data, research into their application to image data including a depth channel, known as RGB-D data, is lacking. Furthermore, the same frameworks cannot be assumed to be applicable to images that additionally contain a depth channel. This is due to the fact that depth images possess different statistical properties and characteristics than their RGB counterpart, between which a correlation cannot be assumed. This stems from the fact that while the color information of each pixel is based on the color of the captured object, depth information is based on the point distance between the object and the sensor.

Several approaches for depth image inpainting exist, many of which focus on utilizing available corresponding RGB data as context for the inference of the missing depth information [42–45]. Other works approach the problem by training models that attempt to minimize the difference between the surface normals of the completed depth image and its ground truth [45, 46].

Recently, Fujii et al. [47] proposed a proof-of-concept GAN-based approach to the joint inpainting of RGB-D images. This multimodal approach shows the potential of joint RGB-D image

Figure 1.3: Illustration of our target problem. The RGB and D channels of image I are shown separately, in I_{RGB} and I_D respectively. I contains a masked region Ω, which is shaped like an HMD and is placed on the face in the appropriate location. The ground truth image is indicated with GT, of which the RGB and D channels are shown in GT_{RGB} and GT_D.

inpainting methods, but lacks experiments to evaluate the proposed architecture. Moreover, the approach is based on an inpainting framework [38] that is no longer considered state-of-the-art. Aside from this work, to the best of our knowledge, a GAN-based framework that proves to jointly inpaint the channels that are present in an RGB-D image does not exist.

In this thesis, we aim to address this by proposing a joint RGB-D image inpainting architecture for the virtual removal of HMDs from RGB-D images. Due to the novelty of this problem, we take an exploratory approach and scope our research based on the research objectives that are defined below.

1.2. Problem formulation

The problem of RGB-D image inpainting can be generally formulated as follows. Given RGB-D input image I containing the masked region Ω, the aim is to consistently fill in region Ω. Typically, this process involves the extraction and propagation of known image information from $I - \Omega$. For RGB-D image inpainting, this information comprises color and geometric information, represented in the RGB channels and D channel respectively.

The primary goal of this thesis is to virtually remove an HMD from an RGB-D face image. Therefore, in our case, the RGB-D input image I contains a face and the missing region Ω is shaped like an HMD (Figure 1.3). We aim to virtually remove the HMD by filling in the missing color and geometric information of missing image region Ω, seamlessly connecting it with the known image region $I - \Omega$. This brings us to our main research objective.

Research Objective 1 *Define an architecture that is capable of virtually removing the HMD from the wearer's face in RGB-D images.*

We look to define an architecture that is able to perform joint RGB-D image inpainting, of which the input is an occluded RGB-D face image I and a binary mask Ω, and the output is a completed RGB-D image. Our goal involves a wide range of domain-specific challenges including preservation of identity, facial expression, face pose, eye gaze, temporal correctness, and audible correspondence. Due to their breadth and complexity, we consider a subset of these challenges, presented in the research objectives described below. The conclusion derived by each objective will motivate the final definition and configuration of our proposed architecture.

Research Objective 1.1 *Define a module and loss function that stimulates the preservation of the identity features of the wearer's face.*

In a virtual environment, the visual representation of a person's identity is seen as one of the most evocative factors to a social experience [48]. The identity of this embodiment has a major function in social interactions as it provokes several components of copresence, including familiarity, comfort and immersiveness. Considering the importance of the connection between the user's offline and online *self* [30], we aim to propose a loss function that stimulates our architecture to fill in the missing region Ω while preserving the identity features of the respective face of the wearer. While the concept of perceptual loss functions for identity preservation has previously been proposed [36, 49, 50], each differ in their exact definition. Moreover, identity loss functions have not been previously applied to our base framework [41] nor RGB-D images.

Research Objective 1.2 *Define an architecture that is capable of handling the multimodal characteristics of RGB-D images.*

RGB-D images contain color and geometric information, represented in the RGB channels and D channel of the image. While the RGB channels represent the color of the captured object, the D channel represents the point distance between the object and the sensor. Consequently, each modality has its own statistical properties and characteristics, between which a correlation cannot be assumed. This has major consequences for the feature understanding of CNN-based architectures, as convolutional layers commonly construct their output features by combining the layer activations of their input. We aim to explore strategies for learning common and modality-specific features to improve feature understanding of the architecture, and to ultimately improve the visual quality of the architecture output.

Research Objective 1.3 *Define an architecture that stimulates the creation of smooth geometric surfaces.*

We aim to explore the architecture's understanding of the geometric shape and surface that the depth pixels collectively form. Pixel-wise reconstruction loss functions such as the L1 or L2 loss are commonly used during training of image inpainting methods [38, 40, 41]. This type of loss function does not consider the geometric and surface properties of the depth images, which can result in suboptimal feature construction and noisy inpainting results. Moreover, we aim to investigate the addition of modules or other architectural changes to accommodate the understanding of geometric information represented in the RGB-D images.

Research Objective 2 *In absence of a large-scale RGB-D face dataset, create a suitable dataset that is sufficiently sized.*

Unlike the wide availability of large RGB face image datasets [51–54], similarly sized datasets containing RGB-D images of faces are not available at this time. Due to the dataset size requirements of the training procedure of GANs, we aim to create a synthetic dataset with a high degree of realism and variety.

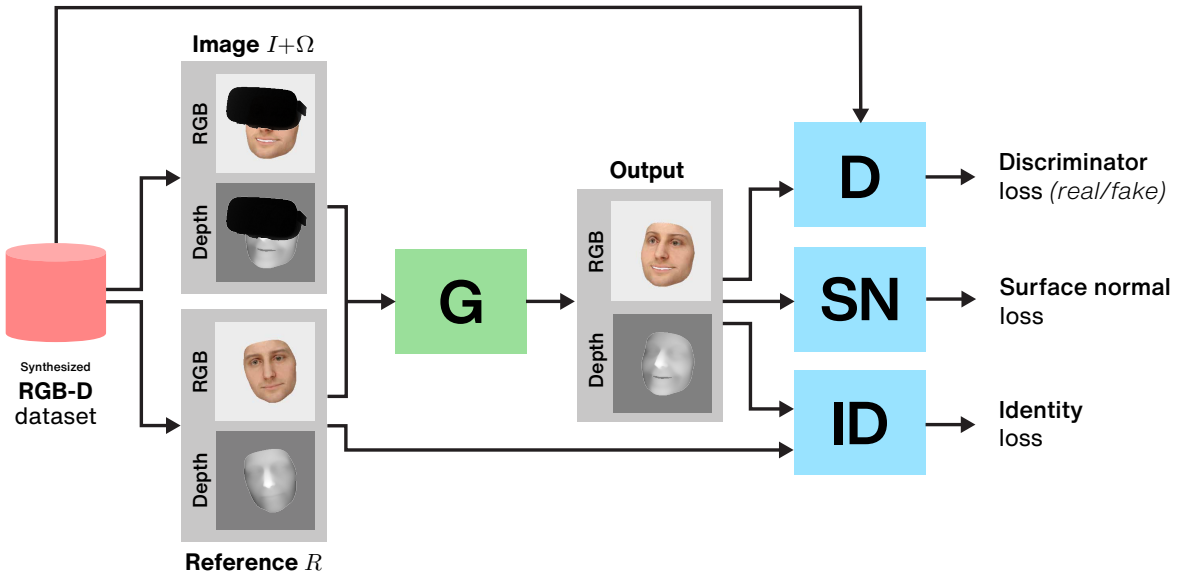


Figure 1.4: Simplified representation of our GAN-based architecture, including a generator (G), discriminator (D), surface normal loss (SN) and identity loss module (ID). The L1 reconstruction loss calculated on the output and the back-propagation connections to the generator are not explicitly shown in this overview.

1.3. Evaluation

In order to evaluate our defined research objectives, we define a number of observable indicators that measure the performance of our proposed method. Specifically, we perform a qualitative and quantitative evaluation of several configurations of our architecture. To evaluate our contributions qualitatively, we perform an elaborate visual examination of several representations of the inpainted results of each compared architectural component or structure. This includes the selected baseline framework by Yu et al. [41], *separately* trained for RGB image inpainting and depth image inpainting.

Moreover, to measure the quality of the inpainting results in a quantitative way, we evaluate the results with several quality metrics: L1 error, L2 error, Peak Signal-to-Noise Ratio (PSNR), Structural Similarity (SSIM) Index [55], and Visual Information Fidelity (VIF) index [56]. To express the preservation of identity of the inpainted results quantitatively, we employ an identity error metric based on the pretrained face embedding model FaceNet [57] trained on the MS-Celeb-1M [54] dataset. This identity preservation metric is independent of our framework's loss function, as it is built on a different model trained on an independent dataset.

The aforementioned metrics provide objective insights on the visual quality of the inpainted results and enable a straightforward examination of the reconstruction capabilities of our architecture. We evaluate our research objectives and motivate our points of future work based on the results of the aforementioned experiments.

1.4. Contribution

A simplified overview of our architecture is shown in Figure 1.4, where components used exclusively during training are shown in blue. The generator, shown in green, is used during model inference. What follows is a summary of the contributions of our work.

Research Objective 1.1 *Define a module and loss function that stimulates the preservation of the identity features of the wearer's face.*

We introduced an identity loss function for the preservation of identity in inpainted images. Similar to existing identity-preserving image inpainting and generation methods [36, 49, 50], we trained our model to minimize the distance between face identity embeddings of our inpainted image and a given reference image during training. To derive the embeddings, we used a pretrained ResNet50 [58] face recognition model trained on the VGGFace2 dataset [53]. This model is invariant to the illumination, pose and expression of the face in the input image. Therefore, the conditions of the reference image do not need to match those of the occluded image.

In our evaluation, we demonstrate that, given an incomplete image I, mask Ω, and reference image R, our architecture can independently extract and propagate the identity features from R to I at inference.

Research Objective 1.2 *Define an architecture that is capable of handling the multimodal characteristics of RGB-D images.*

We investigated the employment of multimodal feature fusion strategies to improve feature learning from RGB-D images. Specifically, we explored several strategies to feature fusion, which can be divided into two types: data-level fusion and hybrid fusion. Whereas data-level fusion consists of the simple concatenation of the RGB and depth image at the start of the network, hybrid fusion combines data-level fusion with fusion at feature-level. We evaluated three types of hybrid fusion, each of which employ different types of feature-level fusion: fusion through summation, single-path residual fusion, and multi-path residual fusion. We explored the viability of each of these strategies and concluded that data-level fusion and residual hybrid fusion produce similar results.

Research Objective 1.3 *Define an architecture that stimulates the creation of smooth depth surfaces.*

Inspired by several works that employ surface normals to depth image inpainting and generation [45, 46, 59–61], we proposed the application of this concept to joint RGB-D image inpainting. In particular, we proposed the usage of a surface normal loss [46] function, which we demonstrated to improve the reproduction of the desired properties of the depth channel such as smoothness. Moreover, we replaced the contextual attention module of the base framework with the contextual surface attention module [46] and showed how this module benefits from auxiliary surface normal information. Lastly, we proposed the addition of surface normal information to the input of the discriminator network, which we determined to cause a deterioration of the visual quality of the inpainted results.

Research Objective 2 *For the training of this architecture, in absence of a large-scale RGB-D face dataset, create a suitable dataset that is sufficiently sized.*

We built a data synthesization pipeline to create a synthetic dataset of RGB-D images of faces based on the parametric Basel Face Model 2017 [62], a 3D Morphable Model [63] (3DMM) model learned from 3D scans of human faces. While the usage of this synthesized dataset

reduces the potential of generalization to real-life data as is, it does allow for large-scale RGB-D face image generation with exact controls over the face's expression, pose and illumination.

1.5. Thesis outline

This thesis is organized as follows: In Chapter 2, we present relevant background theory and related work. In Chapter 3, we describe our proposed architecture for joint RGB-D image inpainting. In particular, we discuss each aspect as presented in our research objectives. In Chapter 4, we present our dataset, its properties and creation process. Furthermore, we present the qualitative and quantitative results of each of the components of our architecture. In Chapter 5, we provide our perspective on the challenges of training a GAN, discuss the limitations of our work, and describe points of future work. Finally, we summarize our research and conclude this thesis.

2

Related Work

This chapter presents background literature regarding the task of HMD removal. We start by discussing the theoretical concepts of HMD removal and present existing approaches. We then move on to providing a theoretical background of deep neural networks and generative adversarial networks (GANs). Subsequently, we present a representative overview of existing image inpainting methods, for color images, depth images, and RGB-D images. Furthermore, we highlight how image inpainting techniques are used towards face completion. We then discuss several multimodal feature learning strategies. Finally, we review the properties of several currently available RGB-D face datasets.

2.1. Head-mounted display removal

Virtual environments can be observed by users through head-mounted displays (HMDs). HMDs are designed to surround the user with three-dimensional visual information that represents the user's virtual perspective. Sutherland [64] pioneered the conceptual design of HMDs as we know them today. This early method employed a mechanical and ultrasonic sensor to present perspective images relative to the wearer's head movement. Modern HMDs such as the Oculus Rift and Microsoft HoloLens have expanded on this concept and accomplish the same goal by utilizing information from sensors such as gyroscopes, accelerometers, and magnetometers.

While HMDs facilitate the observation of virtual environments, they significantly occlude the upper portion of the wearer's face. This forms a major barrier during face-to-face interactions in shared virtual environments, as the obstruction caused by HMDs makes it impossible to fully observe the wearer's face. This forms a key problem for social VR applications such as teleconferencing [9–11] and remote collaboration [4, 5, 65]. HMD removal describes the task of recovering the missing image information caused by the occlusion of an HMD in a coherent and realistic way. We identify two types of HMD removal approaches: model-based approaches, based on cartoon-like or realistic representations, and image-based methods.

(a) Avatars in Facebook Horizon [19].

(b) Avatars in AltspaceVR [20].

Figure 2.1: Cartoon-like avatars in commercial Social VR platforms.

Model-based methods

In an immersive virtual environment, the body and interactions of users are represented through a virtual embodiment, which is commonly referred to as an avatar. A large and growing volume of literature has investigated the concept of virtual embodiment from a technological and psychological perspective.

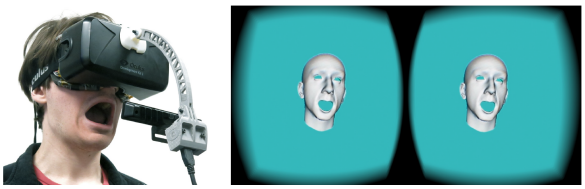
To date, several studies have highlighted how the level of aesthetic and behavioral realism of avatars is related to their acceptability by observers [17, 66, 67]. However, this does not mean one can safely assume a realistic representation is the best option for any social VR application. For instance, if a realistic avatar noticeably deviates from human appearance or behavior, this is likely to distract the observer, causing a decreased level of perceived realism and copresence [21, 22].

In a study by Seyama and Nagayama [22], it was found that when identical abnormalities are applied to artificial and real faces, its impact was greatest for faces with high realism. To that end, a considerable amount of virtual avatar representations, which implicitly target the task of HMD removal, are cartoon-based and thus steer clear from the potential uncanny valley [21, 22]. Shown in Figure 2.1, examples of such cartoon-like representations are used in social VR platforms such as Facebook Horizon [19] and AltspaceVR [20].

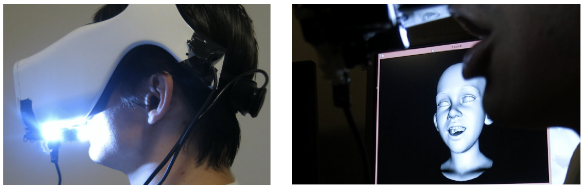
In contrast, other model-based approaches propose the usage of realistic avatars [32, 34, 35, 68]. Despite the greater risk of creating an uncanny valley effect, this choice can be made in favor of the stronger rate of acceptance and copresence of realistic avatars when compared to cartoon-based avatars [69].

Li et al. [32] map the changing geometry coefficients of a user's face to a personalized 3D model that is created offline. The HMD is augmented with a rigidly attached RGB-D sensor to capture the geometry of the visible face region (Figure 2.2a). Surface strain sensors are added to track the facial performance of the upper occluded area of the face. This method requires an accurately recorded or designed 3D model of the user's face prior to online usage. Furthermore, it requires complex calibration before each session, and experiences difficulties regarding capturing eye and lip movement.

Similarly, Olszewski et al. [33] employ a rigidly mounted RGB mouth region camera and an internal IR eye region camera (Figure 2.2b). Two modality-specific CNNs take these data streams as their input to regress facial geometry coefficients to transform an avatar of the user's face. Additionally, the method exploits the coherence between visual and audio recordings,



(a) Online facial reenactment with a head-mounted RGB-D sensor and strain sensors [32].



(b) Online facial reenactment using a head-mounted RGB sensor and CNN regressors [33].

(c) Online facial reenactment with an external RGB-D sensor and interior IR sensor [34].

(d) Online facial reenactment with an external RGB-D sensor and interior IR sensor [68].

Figure 2.2: A collection of existing model-based HMD removal methods setups and results. All pictured methods rely on prior offline 3D face model creation and calibration.

making the method well-suited for speech animation. The proposed system allows users to control an avatar without prior user-specific calibration. However, the resulting method does not consider the texture and identity of the resulting avatar.

Inspired by the aforementioned methods, Thies et al. [34] created the FaceVR framework for HMD removal based on prerecorded footage of the user's face (Figure 2.2c). This method starts by building a parametric 3D model of the user with an exterior stereo RGB camera rig through a number of calibration and regression steps. At runtime, the user - wearing an HMD - is captured with an exterior RGB-D sensor and an interior IR camera. Given these two input sources, the HMD is virtually removed by compositing the prerecorded footage with a rendering of the reconstructed face model and calibration data of the mouth and eyes. This method allows for photorealistic reproduction of human faces, which includes the reenactment of the hidden segment of the user's facial performance such as eye blinking. However, it also requires a calibration and training step for each individual user. Moreover, the method assumes that the pose of the user stays relatively constant throughout the usage of the method.

Zhao et al. [68] proposed a similar approach with internally mounted IR cameras and an external RGB camera where a parametric 3D head model is constructed through a video sequence, capturing several head poses of the user. The constructed 3D head model is aligned at runtime, through the tracking of facial landmarks and visual markers on the HMD. Simultaneously, based on the prerecorded footage, a matched reference image is warped and blended with a colorized version of the internal IR footage. This results in face synthesis that appears realistic and blends in well with the known image region (Figure 2.2d). However, upon inspection of more synthesized frames, clear signs of misalignment of facial features become apparent. Moreover, recalibration is needed for every user and recording environment.

Model-based methods use a virtual character to represent the user's dynamic facial geometry and expressions. In general, this virtual character is either recorded or designed prior to online usage. At runtime, coefficients are inferred from sensor data, which in turn are used

Figure 2.3: Results of RGB image inpainting method for HMD removal by Zhao et al. [36]. The top two rows contain results based on the MS-Celeb-1M [54] dataset, whereas the bottom row contains an inpainting result of real-world footage. For each result set, from left to right: input, inpainted result, ground truth, reference. The input or ground truth of the bottom result is not provided, but can identified based on the inpainted result.

to transform the prerecorded representation. Overall, the aforementioned methods indicate that model-based approaches are capable of producing high quality results. However, each of these methods rely on a controlled environment for setup, of which the conditions are assumed to remain constant during usage. This severely limits the potential widespread application of these systems.

Image-based methods

A number of studies have examined methods that approach HMD removal as an image-based task. As opposed to model-based approaches, image-based approaches typically do not use an intermediate parametric model to virtually remove the HMD. Instead, image-based methods rely on operations in the image or feature space to resolve the masked area. Consequently, methods of this kind can be seen as a subtask of image completion or inpainting.

Zhao et al. [36] explored the application of a generative inpainting method for the purpose of HMD removal in RGB images that have been occluded synthetically. This method is built on the concept of generative adversarial networks, trained to consistently fill in the masked region caused by the HMD. The proposed procedure is robust against moderate variations in pose, and is able to preserve the subject's identity given a reference image of the target subject. Additionally, a target pose map is passed to indicate the intended face orientation. Despite the required target pose map, this method falls short in the case of extreme pose angles and is not robust against expressions. Furthermore, the inpainted results are blurry and do not fully blend in with the known region of the image (Figure 2.3). It is also important to note that this method does not consider depth images or geometric information of any kind. This prevents its application to immersive teleconferencing, in which case RGB-D images are typically used.

Wang et al. [70] aim to alleviate the requirement of a target pose map, and introduced a similar framework that uses a facial landmark detector to predict facial landmarks as a prior step. Successively, the predicted facial landmarks are passed to a GAN architecture, combined with a synthetically occluded RGB image and a reference image. While the facial landmark

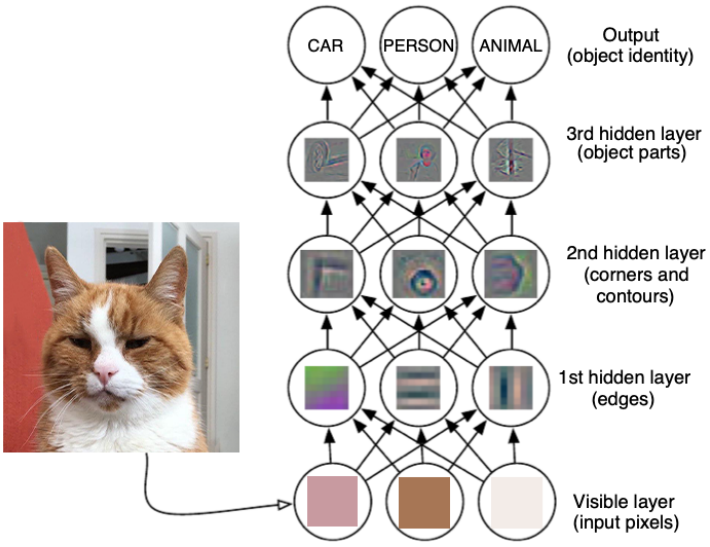


Figure 2.4: Illustration of a deep learning model for image classification, as reproduced from [71] with a custom image sample.

map successfully replaces the target pose map used by Zhao et al. [68] and improves on the method's visual quality, the proposed method does not perform well under severe pose angles and is limited to RGB images as well.

The majority of image-based methods that could theoretically be applied to HMD removal have been defined in the general context of removal of facial occlusions through image inpainting. This being the case, we continue our review of image-based HMD removal methods in Section 2.4.1, where we discuss these methods and elaborate on their expected capabilities when applied to image-based HMD removal.

2.2. Deep neural networks (DNNs)

Before we move on to review existing image inpainting methods, we present the fundamentals of deep learning. Deep learning is an approach to machine learning that is based on the construction of meaningful representations of raw data [71]. Conventional machine learning methods rely on prior feature extraction, which typically requires a substantial amount of human time and domain knowledge. However, in some cases it is nearly impossible to design a feature extractor that extracts all the features that are relevant to our objective.

Deep neural networks (DNNs) do not require prior feature extraction, and construct feature representations from raw data directly. Through a hierarchy of layers of computational neurons with corresponding nonlinear functions, DNNs allow the interpretation of complex representations from simple representations [71]. As we move up this hierarchy, the level of abstraction of representations increases. Each layer further abstracts its input, starting from pixel input, going to simple structures such as edges, contours and corners; and eventually, capturing semantic structures such as objects and their segments. An example of this process is illustrated in Figure 2.4.

As stated by Bengio et al. [72], deep neural networks have two significant advantages as a result of their architecture. Firstly, reusing features is possible as a consequence of hierarchical

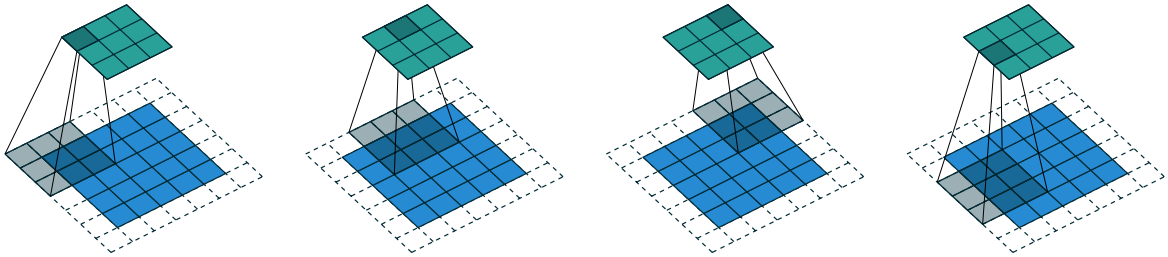


Figure 2.5: Visualization of convolving a 5×5 input with a 3×3 kernel, with padding of size 1, and 2×2 strides. Figure reproduced from [76].

feature learning. Secondly, due to their hierarchical structure, DNNs are able to construct highly abstracted feature representations of the network input data. This enables high-level layers to extract features that are typically invariant to certain local variations in the network input.

2.2.1. Autoencoders

The capability of representation learning with DNNs is best demonstrated based on the paradigm of autoencoders [73]. Autoencoders consist of an *encoder* which encodes input data x into a meaningful latent representation, and a *decoder* which in turn decodes this representation to $\hat{x}$ with the same dimensions as x. Reconstruction loss $\mathcal{L}(x, \hat{x})$ stimulates the encoder to extract the most meaningful features from its input in such a way that the decoder is able to accurately reproduce the original input. In this way, autoencoders are able to learn feature representations in an unsupervised manner.

Generative models build on this concept, and operate in the high-dimensional latent space to perform tasks such as image generation and image inpainting. We will elaborate on the details of generative models in Section 2.3.

2.2.2. Layers

In this section, we will discuss a number of layers that can be used to construct a DNN. Specifically, we will describe the layers that make up our proposed architecture.

Activation layer

At first glance, the network illustrated in Figure 2.4 appears to be a simple linear combination of neurons. This concept can be useful for linear problems, but is less useful for nonlinear problems [71]. Activation functions give DNNs their representational power for problems that behave in a nonlinear manner. Activation layers apply these activation functions to their input and improve the generalization ability of DNNs.

At this point, a singular activation function that works well for all problems does not exist. However, activation functions such as the tanh, sigmoid, ReLU [74], and leaky ReLU [75] have been commonly used due to their desirable properties.

Convolutional layer

Convolutional networks (CNNs) [77] employ a set of learnable filters to extract features from data that have a grid-like topology [71], such as images represented by a multidimensional array

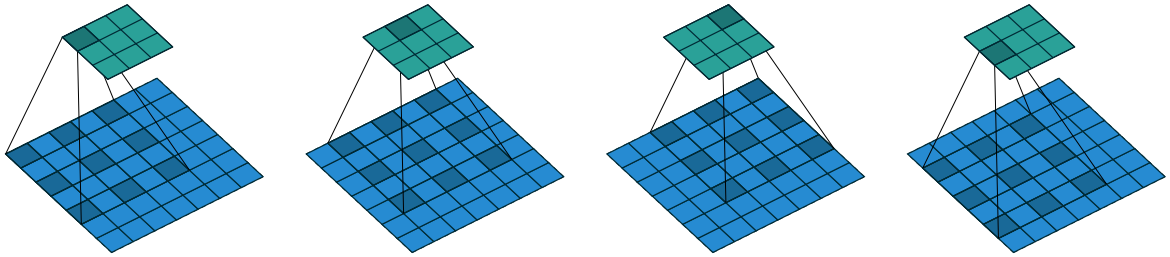


Figure 2.6: Visualization of convolving a 7×7 input with a 3×3 kernel, without padding, 1×1 strides, and a dilation factor of 2. Figure reproduced from [76].

of pixel values.

The convolution operation makes strong assumptions with respect to its input. Firstly, it is assumed that the input values are locally connected and together form feature representations. Secondly, convolution assumes that natural images have invariant statistical properties [71]. As such, features (e.g. edges and corners) can occur at any spatial location, which allows the reuse of filters by employing parameter sharing. As a consequence, parameters can be more efficiently applied to learning better and more varied filters.

The core component of CNNs are convolutional layers, which employ learnable filters that are used for the convolution operation performed on the layer input. To obtain a feature map, the convolution operation convolves the specified filter w with size $K \times K \times D$ over the input x with size $W \times H \times D$. Specifically, the filter slides over the input, calculating the product between each filter and input element in the area that the kernel overlaps with [76] (Figure 2.5). The same operation can be performed with multiple filters to obtain more feature maps.

The size of the sliding step distance when performing convolution is referred to as the stride. Moreover, padding can be used to control the size of the layer output. As such, padding is commonly used to maintain the size of the layer input.

Dilated convolutional layer

Dilated convolution [78] enlarges the effective size of the kernel by inserting zero spaces between the kernel elements. In this way, dilated convolutions have a larger receptive field without having to increase the kernel size and the related parameters and computational requirements. In the context of CNN-based image completion, this for allows the completion of larger missing regions [38].

2.3. Generative adversarial networks (GANs)

Proposed by Goodfellow et al. [37], generative adversarial networks (GANs) have changed the way state-of-the-art methods generate image data and has become the most widely used architecture in this field. Given a set of training images drawn from a distribution p_{data}, GANs learn a representative estimate of this distribution p_{model} [79]. The structure of GANs is shown in Figure 2.7.

The learning process of the original GAN [37] can be put in terms of a minimax two-player game. The players in this game can be described as two functions, generator G and discriminator D. G is stimulated to generate images that resemble samples from distribution p_{data}, while

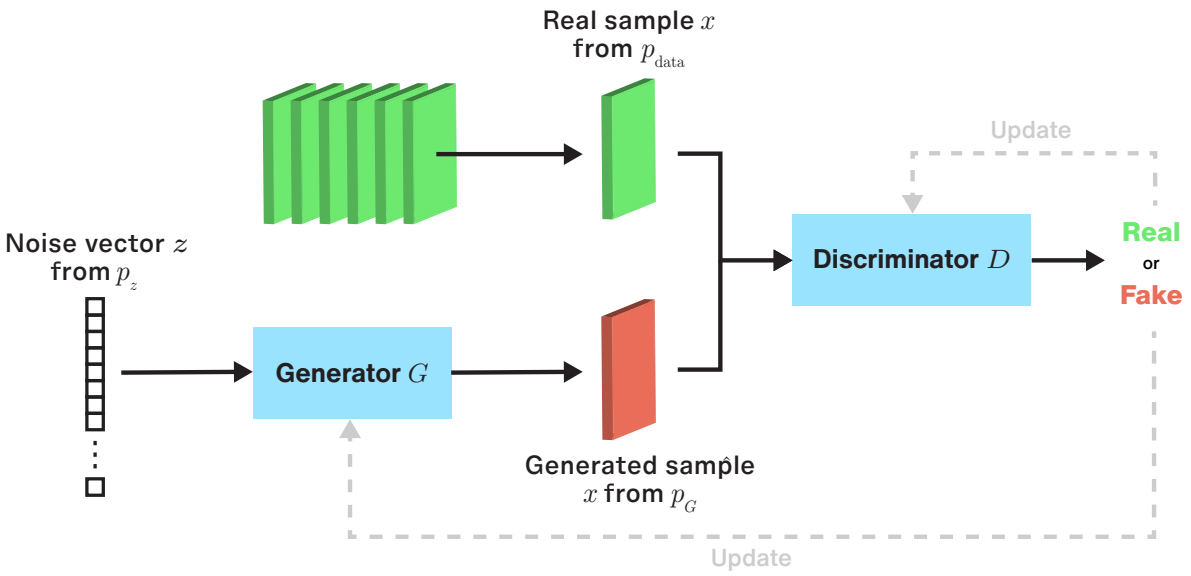


Figure 2.7: Architecture schema of the original GAN [37].

D is encouraged to distinguish these generated (fake) images from real images. In this process, the effective goal of G is to learn how to fool D.

Generator G is typically represented by a deep neural network. Taking a random noise vector z drawn from a distribution p_z (e.g. Gaussian) as its input, G maps z to an image drawn from distribution p_G: $G(z) \rightarrow \hat{x}$. Discriminator D learns to classify images as real or fake, and is defined as $D(x) \rightarrow [0, 1]$.

Generator G and discriminator D are trained in a competing fashion. In this training process, the two models are trained jointly with back propagation, based on the objective function:

$$\min_G \max_D V(G, D) = \mathbb{E}_{x \sim p_{data}(x)}[logD(x)] + \mathbb{E}_{z \sim p_z(z)}[log(1 - D(G(z)))], \tag{2.1}$$

where $V_a(G, D)$ denotes the adversarial loss for G and D, and $D(X)$ refers to the probability that X is a real image. Considering the two models are competing against each other, this function is referred to as the *adversarial loss*.

It should be noted that in the early training stages, when the performance of generator G is poor, discriminator D is able to reject generated images with high confidence. As a result, the term $log(1-D(G(z)))$ saturates and does not provide enough feedback to update G. Therefore, instead of minimizing this term, it is replaced by $logD(G(z))$ which is maximized during training [37].

Goodfellow et al. [37] showed that, given enough capacity, the adversarial training process is theoretically able to reach equilibrium, such that the model distribution equals the distribution of the training data, $p_G = p_{data}$. In this case, the discriminator no longer is able to distinguish generated images from real images, thus $D(X) = 0.5$ for all x.

Conditional GAN (cGAN) Shortly following the introduction of the original GAN, Mirza and Osindero [80] introduced a conditional version of GANs. In this case, the generator and discriminator take an additional input y to which they are conditioned, such as class labels, data

Figure 2.8: Visualization of vector arithmetic as applied to an example of visual concepts.

from other modalities, or any other kind of auxiliary data. As such, it is possible to control the output of the model, which is not the case with the original GAN.

Deep convolutional GAN (DCGAN) Radford et al. [81] proposed DCGAN, which introduced a number of architectural changes to the original GAN to stabilize the training process. These changes involve:

- replacing all pooling layers with strided convolutional layers, to enable spatial correlation;
- removing the hidden fully-connected layers to enable deeper models;
- adding batch normalization in the generator and discriminator, to stabilize training;
- using the *ReLU* activation function in the generator and the *Leaky-ReLU* activation function in the discriminator, to speed up training.

The combination of these modifications have resulted in higher quality output for most situations, as well as a more stable learning process. Moreover, the authors demonstrate the ability of the architecture to learn meaningful representations by showing the ability of interpolation between its points in its latent space. Through vector arithmetic between two points in the latent space, visual concepts can be combined and used for image generation (Figure 2.8). This example provides an interesting insight to the workings of the latent space of GANs.

2.3.1. Evaluation

The objective of generative models is to draw samples from a distribution that closely resembles the distribution of the available data p_{data}. Therefore, in the case where samples are images, generated images are sought to realistically resemble images from the available dataset. To accurately measure the performance of generative models in this regard, it is essential to define *realism* and *resemblance*. Both terms represent concepts that are inherently subjective, making the search for a suitable highly challenging. Furthermore, as mentioned by [82], the choice of the appropriate set of metrics to evaluate a generative model should rest on the application it was intended for.

Given the extensive applications and architectures of generative models, a broad range of strategies for their evaluation exists. Where possible, previous research has predominantly opted for a combination of quantitative and qualitative evaluation, which combines evaluation through a collection of metrics with the addition of user studies or a visual examination of the generated samples. Several studies have outlined how the qualitative and quantitative assessments are currently ill-fitted for a reliable evaluation of GANs [82–84].

Intuitively, qualitative assessment by humans may seem like the most representative method for the evaluation of generative models. While human observers can competently distinguish generated and real images, their evaluation is influenced by the visual quality of the images [83]. As a result, the degree of diversity and generalization of the samples are neglected, favoring models that overfit or memorize the training data [84]. This is particularly harmful when evaluating GANs designed for unconditional image generation [85]. Besides, valid qualitative evaluation is time-consuming, subjective and sensitive to viewing conditions.

To date, various sample-based evaluation metrics have been proposed that aim to quantify the performance of generative models by capturing the correspondence of statistical properties between generated samples and real samples [84]. One of the most widely used metrics for image-based tasks is the Inception Score [86], which assesses the visual quality and diversity of samples in such a way that is consistent with human evaluation. The results of this metric are calculated based on features as produced by the Inception network [87], trained on the ImageNet [88] dataset.

Despite the introduction of several metrics, consensus has not been reached with respect to a standardized set of metrics [83, 84].

2.4. Image inpainting

Image inpainting or completion describes the task of filling undesired or unknown pixel regions with realistic content. In early work, *image inpainting* refers to the act of filling in small or narrow image regions, whereas *image completion* refers to filling in large image regions. More recently, these terms have been used interchangeably to refer to inpainting any size of image region. In this thesis we refer to this task as image inpainting.

Image inpainting has found widespread use in applications such as the removal of unwanted objects or regions [38, 40, 89, 90], image editing [41, 91], image stitching [92], video inpainting [93], and privacy protection [94].

A challenging aspect of this task is to restore the structure and texture of the image in such a way that is undetectable. To achieve this, image inpainting methods commonly utilize known image data to reconstitute the missing regions of an image in a visually plausible and unifying way. Specifically, image inpainting methods rely on a combination of contextual information. Over the years, a large body of strategies have been proposed to localize and use the most relevant contextual information with respect to the missing region.

In this section, we focus on existing image inpainting techniques and present their advantages and disadvantages. Firstly, we discuss methods aimed at inpainting RGB color images, which is the most widely studied. Secondly, we consider methods that are aimed at depth images. Finally, we discuss image inpainting methods that are aimed towards the joint inpainting of both modalities which are represented by RGB-D images.

2.4.1. Color image inpainting

In the last few decades, much research has been committed to image inpainting of RGB color images, resulting in a wide range of methods. Some early approaches focused on filling in missing texture, without any obvious artifacts. Other approaches aimed to propagate the structure

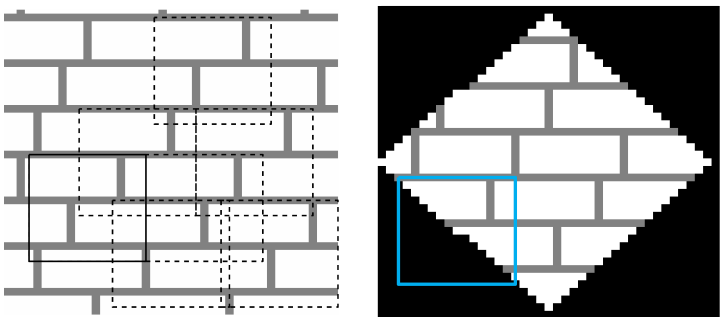


Figure 2.9: Visualization of texture synthesis algorithm by Efros and Leung [96]. Given the texture sample image on the left-hand side, the unknown pixels in the image on the right-hand side are filled. At random, the algorithm picks neighborhood $\hat{x}$ of all neighborhoods that are similar to the neighborhood x of the target pixel. To synthesize the target pixel, the center value of neighborhood $\hat{x}$ is used. Image reproduced from [96]

of the known image region to the missing region. Accordingly, we split early approaches based on their primary reproduction target, which is the texture or structure of the missing region, or a combination of the two. We provide a brief overview of each of these early methods; after which we will progress to the more relevant CNN-based approaches.

Texture-based methods

While texture synthesis is a separate task from image inpainting, texture synthesis methods have been applied to the task of image inpainting with moderate success. Texture-based image inpainting methods [95–98] focus on the reproduction of the texture of the missing region.

Efros and Leung [96] proposed a non-parametric method for texture synthesis based on a given texture sample. This method constructs a new image inwards from an initial seed recursively, in a pixel-wise manner. The value of each of the synthesized pixels in the image is determined based on the nearest neighborhood in the sample texture.

The algorithm has the tendency to end up in the ”wrong” part of the search space when fed with a highly variable texture. As a result, in this scenario it produces visually inconsistent data. Moreover, it is computationally expensive as a full search of the image for each synthesized pixel.

Texture-based methods perform well when inpainting small regions, but fall short when filling in large regions of an image. Moreover, these methods clearly fail to preserve the structure of the image, as they generally do not possess mechanisms to handle visual features such as edges and corners.

Structure-based methods

In contrast, other methods [99, 100] focus on structural continuity when filling in the missing region of an image. To reproduce the missing structural information, directions of the visible structures are attempted to be extrapolated to fill the missing region. This is commonly done based on isophotes, which are lines of equal gray value. Isophotes represent directional information of the image structure, and can be connected to each other to propagate the structure from the surrounding region to pixels located within the missing region. The direction of the isophotes are defined perpendicular to the gradient vectors representing the spatial change of pixel intensity levels.

Generally, structure-based methods handle images with low textural variability and small missing regions well. However, structure-based approaches are not able to accurately produce texture, let alone large patches of it.

Hybrid methods

Drori et al. [101] proposed a method that aims to preserve both image texture and structure. This examplar-based approach synthesizes a complete image through an iterative procedure that approximates unknown regions through the concept of self-similarity. At the start of each iteration, an unknown image fragment is selected based on the highest value in a confidence map. This confidence map represents the vicinity of each pixel to a known region. Subsequently, a similar known image fragment is selected, which is used to fill the respective unknown region. This process repeats itself until the image has been completed. While this method performs well with flat scenes, it is unable to distinguish foreground and background regions, nor moving objects. Moreover, similar to other approaches relying on finding similar image patches [89, 97], this efficiency of this approach is limited by the expensive procedure of optimal patch search.

Barnes et al. [91] introduced an accelerated procedure for the operation of finding similar patches. Their proposed approximation algorithm begins with a randomized or derived guess, which is followed by an iterative process that randomly samples the image to find fitting patches. Coherence is used to propagate such matches quickly through surrounding areas. At the time of its introduction, this method performed an order of magnitude faster than previous patch-based approaches, and enabled the real-time editing of images on a high level. Examplar-based image completion approaches suffer from their lack of knowledge about the anatomy and structure of an image, which causes these methods to be ineffective at filling regions that are surrounded by complicated structures or novel objects. Moreover, the search space for similar image patches is confined to the input image. This can be a problem when the input image does not contain patches that can fill all unknown regions in a realistic way. Hays and Efros [102] demonstrated an image completion procedure that uses a large database of images. Prior to the execution of the image completion algorithm, the image with the highest similarity to the input image is retrieved. This image is then used to complete the input image. However, this type of method relies on the assumption that an image with a similar scene, structure and texture is included in the database which is often not the case.

CNN-based methods

None of the previously discussed image inpainting methods are truly able to capture and reproduce the high-level semantics of images. Consequently, these methods struggle to inpaint missing regions that contain complex and novel visual information. As discussed in 2.2.2, CNNs are able to capture visual information at different levels of abstraction. To achieve better performance on images that contain complex real-world scenes, CNNs have been applied to the task of image inpainting and have defined the state-of-the-art for many years.

The first CNN-based image inpainting methods built on the concept of the autoencoder (Section 2.2.1). In this case, the incomplete input image is first transformed to a latent representation and then transformed back to its original dimensions. With the appropriate architec-

ture and loss functions, autoencoders naturally lend themselves for denoising and inpainting images.

Xie et al. [103] introduced a learning-based method that is founded on the concept of denoising autoencoders, in which case the missing region involves the noise present in the input image. The authors demonstrated the automatic removal of complex patterns such as superimposed text, without requiring prior information of the unknown regions or their location. However, this benefit has a limited impact, as the method is only able to inpaint patterns that have been seen at training time. Moreover, the relatively shallow network merely captures low-level image features.

Recent learning-based approaches [38–41, 104] have adopted concepts of generative adversarial networks (GANs) [37] (Section 2.3). These approaches consider image inpainting as a conditional image generation problem. Instead of feeding a noise vector to the model, the known region of the image is passed as the input. Methods of this kind employ an autoencoder that is commonly trained with an adversarial loss and reconstruction loss to coherently fill in images with missing regions based on relevant features of the known image region.

Pathak et al. [104] introduced an adversarially trained autoencoder architecture that aims to encode the context information provided by the known image region. The authors employ an autoencoder architecture which is trained with a joint loss function consisting of an adversarial loss [37] and reconstruction loss. The adversarial loss stimulates the sampling from the appropriate mode of the learned distribution and makes the synthesized region appear realistic [104]. The reconstruction loss encourages coherency with the known region of the image, by favoring reproduction of structures and texture. In this way, the authors were able to train an autoencoder to complete a fixed 64×64 pixel area in the center of a 128×128 pixel image. However, the approach does not address inpainting regions with an arbitrary shape or location, and does not specify how it can be applied to images with a higher resolution. Moreover, despite the usage of a reconstruction loss, images inpainted by this method lack local coherency with the surrounding known region [38].

To improve overall coherency, Iizuka et al. [38] employed both a global and local discriminator. The global discriminator evaluates whether a scene is coherent in its entirety, whereas the local discriminator assesses the coherency of the area around the generated regions. Moreover, the authors decreased the number of downsampling layers and replaced standard convolutional layers with dilated convolution (Section 2.2.2), enabling the method to use a larger context area around each unknown pixel with the same computational power. As context is a critical factor in this task, this can significantly contribute to the consistency of the generated area and allows the method to process much larger areas. However, this change caused the entire training procedure to take up two months with four NVIDIA Tesla K80 GPUs [38], which forms a major drawback of this framework.

Yu et al. [40] built on the architecture of Iizuka et al. [38]. The authors proposed a fully convolutional model with a contextual attention module that explicitly borrows information from surrounding regions. Whereas convolutional operators typically only process local image features, the proposed contextual attention module learns to extract feature patches from anywhere in the known region of the image. The propagation of contextual information results in more realistic inpainting results with less artifacts. Furthermore, a key contribution of this work

is a two-stage coarse-to-fine network architecture. The coarse stage of the network infers an initial coarse estimation, whereas the refinement stage of the network further refines this prediction. The primary limitation of this method is that its application is constrained to missing regions with rectangular shapes.

To handle unknown regions with irregular shapes, Liu et al. [105] introduced the use of a partial convolutional layer. This layer causes the output of the convolution operation to only depend on the known region of the image. This is achieved by applying masked normalized convolution only on known pixels, given a binary mask indicating the unknown region. Effectively, this means that the output of each layer is multiplied by a binary mask. Yu et al. [41] criticizes this method by stating that categorizing each pixel with the same mask at every layer causes the loss of valuable information such as synthesized pixel data and can result in the receptive field of some neurons to cover only unknown pixels.

Yu et al. [41] addressed this issue by introducing the gated convolution operation, which learns a dynamic learnable feature selection strategy for each image channel at any spatial location across all network layers. This operation allows precise regulation of what pixels are affected by feature information, at every layer of the network. This enables the network's capability of processing irregularly-sized masks, and extension to user-guided image inpainting. Moreover, the authors proposed a spectral-normalized Markovian discriminator, motivated by previously discussed approaches with global and local discriminators [38], Markovian GANs [106] and spectral-normalized GANs [107]. In our work, we use this architecture as our base framework. We elaborate on our the reasons for our choice in Section 3.1.

Image inpainting for face completion

While the concept of face completion is similar to the general task of image inpainting, it is considered more challenging, as it requires the generation of individual facial components which contain large appearance variations. Moreover, it is a major challenge to ensure coherency between these components, while simultaneously maintaining symmetry, realism and preservation of identity. In this subsection we will discuss methods that consider the task of face completion, also known as face inpainting, as a subtask of color image inpainting. We start by discussing a number of early approaches to face completion, which is followed by a representative summary of recent approaches to face completion. Moreover, we discuss identity preservation strategies commonly used in face completion and face synthesization frameworks.

Park et al. [108] focused on the occlusion caused by glasses and showed that missing regions can be inferred through recursive error compensation using PCA reconstruction. This procedure uses color and edge information to extract the region of occlusion, and finally generates an image which compensates for this occlusion. However, the method's sole focus is frontal images, and its usage in real-life scenarios is not evident.

De Smet et al. [109] proposed an algorithm that estimates the parameters of a 3D morphable face model (3DMM) under large occluded areas. Before the estimation of the 3DMM parameters, the occluded area is identified and excluded from computations. De-occlusion is approached by applying a generalized expectation-maximization (GEM) algorithm in which the parameters related to the occluded area are computed iteratively. While the paper presents notable results, the approach relies on fiducial points that have been manually selected. More-

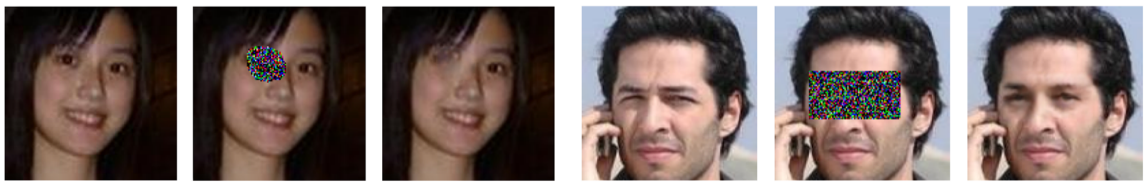


Figure 2.10: Face completion results of approach by Li et al. [111] on the CelebA [52] dataset.
Left to right for each image set: original image, masked input, completion results. *Right set*: model generates realistic results. *Left set*: model fails to generate the eye for an unaligned face. [111]

over, the performance of this method has not been evaluated on any other databases.

Mohammed et al. [110] introduced a statistical method for the inpainting of specified facial regions that consists of a global parametric model with a local non-parametric model that are learned from a database of face images. The authors show the method's ability to generate face images that do not occur in the training data. Yet, the results contain many visible artifacts which are detrimental to their realism.

We now turn to a more recent method by Li et al. [111], who proposed a face completion framework that is based on a generative model. The authors employed a GAN-based network which is trained with a local discriminator that focuses on the realistic appearance of individual face components. In addition, a global discriminator focuses on the contextual faithfulness of the full image. Furthermore, a semantic parsing network is used to compare the synthesized face region with the original image, which enforces a natural shape and size of the completed face. However, as mentioned by the authors, the model does not perform well when input images are not well-aligned, as can be seen in Figure 2.10. Moreover, the model fails to fill in the missing region that is spatially coherent with the pose of the face and the inferred regions appear blurry. Comparable limitations can be observed with the GAN-based approach of Yeh et al. [39], which searches for the closest encoding of a given face image in the learned latent image manifold.

To increase the semantic knowledge of generative models with respect to human faces, Liao et al. [112] proposed a collaborative GAN that splits the learning process into multiple subtasks. In particular, this method aggregates knowledge of the tasks of face landmark detection and semantic segmentation and uses this knowledge for the completion of face images. Results of this method show that incorporating the knowledge learned from these tasks contribute to the model's understanding of the symmetric structure of human faces. However, breaking down the task of face completion requires additional ground truth information for training and complicates hyperparameter tuning. Consequently, it is unlikely that this method is applicable to other datasets without significant re-optimization and training.

A number of recent general-purpose image inpainting frameworks [38, 40, 41, 105] present remarkable results on face images. For instance, Iizuka et al. [38] describe a user study that showed that their approach produces inpainted images of faces that are indistinguishable from real faces 77% of the time. However, visual examination of the application of this method to face completion shows a large number of visual artifacts and overall low visual quality as illustrated in Figure 2.11. Moreover, the aforementioned lengthy training process of this approach forms an obstruction for the application of this method.

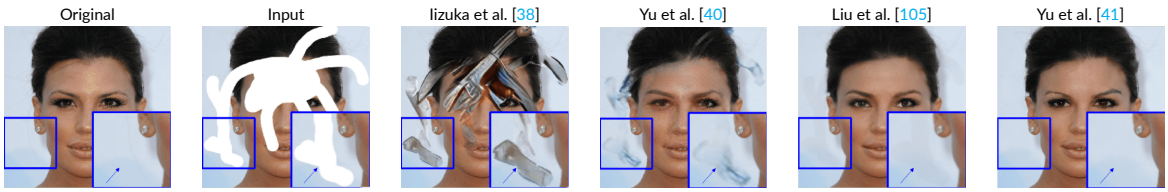


Figure 2.11: Visual comparison of several general-purpose image inpainting methods [38, 40, 41, 105]. This image was reproduced from Yu et al. [41].

In contrast with the method by Iizuka et al. [38], a comparison with the results as produced by the method proposed by Yu et al. [40] reveals that the latter successfully completes the face of the subject in a visually consistent way. Facial features are appropriately positioned and the skin color of the subject is properly propagated to the missing region. Unfortunately, the inpainted face lacks visual symmetry. Moreover, a notable amount of visual artifacts are present in the inpainted image.

Turning to the inpainted results of the method by Liu et al. [105] and Yu et al. [41], we observe two similarly inpainted images. Aside from the consistent inference of the missing region, both methods produce visually symmetric facial features. Nevertheless, closer inspection reveals that the result generated by the method by Liu et al. [105] contains noticeable visual artifacts. Based on these observations, both methods show potential for application to the task of HMD removal.

Identity preservation As mentioned, a major contributor to the applicability of HMD removal methods in a social VR context is the preservation of identity [30, 48], which has motivated Objective 1.1. Aside from the image-based HMD removal methods [36] discussed in Section 2.1, none of the image inpainting methods address this point. In some cases this is related to the fact that their scope is wider than face completion [38, 40, 105], while other studies simply disregard this aspect in the design of their framework [39, 112]. In view of this fact, relatively few image inpainting methods exist that target identity preservation.

Broadening our view to the research in the field of face generation [52, 113] and frontalization [114], it stands out that the vast majority of generative approaches considering identity preservation build on a common type of identity loss function [36, 49, 52, 113, 114]. In particular, these methods typically employ a pretrained face recognition model to obtain a latent identity embedding of the inpainted image and a given reference image. By minimizing the distance between these embeddings during training, the model learns to propagate identity-specific features from the reference image to the inpainted image. While opinions differ on the optimal distance function and face recognition model, this theoretical concept is commonly and successfully applied to methods targeting identity preservation. Moreover, the identity loss of the HMD removal framework by Zhao et al. [36] is based on the same notion.

2.4.2. Depth image inpainting

While depth inpainting has received less attention than color image inpainting, research towards depth image inpainting has been an increasing topic of research. This can most probably be attributed to the widespread increase in availability and usage of depth sensors. It is inter-

esting to note that depth images are finding an increasing numbers of applications such as the creation of immersive virtual environments [5, 9, 10, 12], semantic segmentation [115], and autonomous driving [94]. These and all other applications of depth images benefit from accurate and complete depth information.

Current commodity RGB-D sensors provide an affordable way of recording color images with corresponding depth information. However, the depth information often contains missing information and artifacts. This can be caused by sensor noise and surfaces that are reflective or are either too far or too close to the sensor [116, 117]. Image inpainting methods focused on denoising can be applied for the removal of artifacts of this type [43, 103]. Moreover, depth image inpainting techniques can be applied to object removal and surface completion in depth images [46, 94]. In this way, applications that use depth data can greatly profit from depth image inpainting.

Existing RGB image inpainting methods have previously been applied to depth image inpainting [46, 118–120] with reasonable success. However, in practice, these methods often fail to address the statistical properties of depth image inpainting involving depth continuity, surface relief, and local feature preservation [121]. Acknowledging this, a substantial amount of depth image inpainting methods have been proposed that address these characteristics.

Depth image inpainting approaches can be divided into methods that reconstruct depth images independently [42, 46, 119, 122, 123] and methods that additionally use color images to serve as contextual information [43, 45]. We briefly evaluate and compare methods of both types in the remainder of this section.

Independent depth image inpainting

To improve RGB-D indoor scene estimation, Silberman et al. [119] applied the RGB image inpainting method as proposed by Levin et al. [124]. While the original image inpainting algorithm was designed for color images, the method performed similarly well on small missing regions in depth images. While the insights provided by this work are limited, it is interesting to note that RGB image inpainting show potential for the purpose of depth image inpainting.

Xue et al. [123] proposed a depth image denoising method that does not use any information from additional modalities such as corresponding color images or related depth images. The authors applied the low rank assumption to the completion process of corrupted depth images. However, they found that this assumption does not translate well to depth images due to their textureless and sparse characteristics. Based on these properties, low gradient regularization was combined with low-rank regularization for inpainting noisy depth images. Consequently, this method is focused on the inpainting of a large collection of small regions, and therefore is not well-suited for inpainting larger regions.

Matias et al. [46] proposed a method for object removal for depth images, built on the existing color image inpainting method by Yu et al. [40]. Applied to depth images, this GAN-based framework is able to learn meaningful features from depth images, based on a combination of a reconstruction loss, adversarial loss, and a contextual attention module. As depth images inherently represent surfaces, the authors proposed the vectorial loss function that encourages the coherency of the synthesized region and the rest of the depth image based on estimated surface normal images. Moreover, the authors modify the contextual attention module by pro-

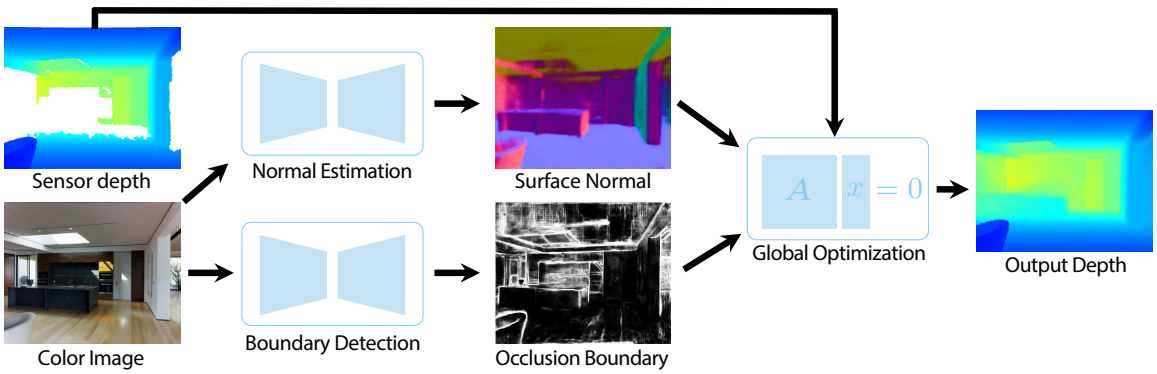


Figure 2.12: System pipeline of depth image inpainting framework by Zhang and Funkhouser [45].

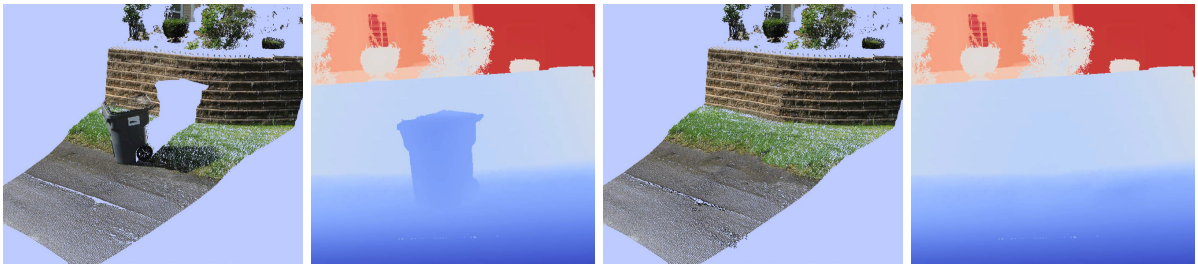
viding it with an estimated surface normal image to support the localization of similar surfaces in the known region of the depth image. The method is evaluated on data depicting street scenes, containing a large amount of sizable surfaces. Considering this fact, it is unclear how this method would perform on depth images with a high level of detail and corresponding sparse gradients.

Color-guided depth image inpainting

Many researchers take advantage of available color information [43–45] to complete the task of depth image inpainting. Bearing in mind that the color and depth image are spatially aligned, color features and depth features are assumed to be strongly related [105]. This type of method takes at least two inputs: an incomplete depth image and a corresponding complete color image.

Herrera et al. [43] proposed an image inpainting method to complete the depth channel of an RGB-D image. This method aims to reproduce visually consistent structures by favoring information that surrounds the boundaries of the missing regions. However, the method assumes that discontinuities between surfaces in the depth image are aligned with discontinuities in the color image. This limits the applicability of this algorithm to specific situations.

Following the successful application of GANs in color image inpainting [38–40, 104], Zhang and Funkhouser [45] applied the concept to the inpainting of the depth channel of RGB-D images, shown in Figure 2.12. The authors indicated that the main challenge of depth image inpainting is related to the lack of strong features in the depth channel. Therefore, this method takes an intermediate step to predict local properties of the depth values before progressing to complete the raw depth map. A deep neural network is used for the estimation of an occlusion boundary image and surface normal image. Concatenating the predictions with the raw depth image, a global optimization step processes and outputs the inpainted depth image containing absolute depth values. While this method solves a different problem than ours, it is to be noted that alternative depth representations can contribute to a joint RGB-D image inpainting method. Specifically, within the context of our work, this corresponds to Objective 1.3.



(a) (b) (c) (d)

Figure 2.13: Example of Doria and Radke [126] RGB-D image inpainting method with corresponding intermediate depth gradient inpainting result. (a) LiDAR scan (RGB-D) of a trashcan in front of a background consisting of concrete, grass, and a brick wall. (b) Magnitude of the depth gradient of ground truth LiDAR scan. (c) Magnitude of the depth gradient of inpainted result. (d) Inpainted result of LiDAR scan (RGB-D).

2.4.3. RGB-D image inpainting

Each pixel in an RGB-D image contains multimodal information of a common entity. Specifically, the RGB channels represent the color modality, whereas the depth channel represents the depth modality. While these modalities are related, there exists a semantic gap between the modalities represented by the color and depth channels, as they each have different signal frequencies and characteristics.

Limited research towards the inpainting of RGB-D images exists [47, 125–127]. This could be attributed to the challenges involved with multimodal data, combined with the high computational cost of image inpainting procedures [128]. As noted in earlier sections of this chapter, the extraction and interpretation of image features form an unavoidable challenge in successfully inpainting an occluded image, which also holds for RGB-D image inpainting. In fact, feature extraction and interpretation from RGB-D images is especially challenging considering the multimodality of the image information. To accurately interpret the information represented by an RGB-D image, a method to extract features that incorporate the complementary relation between the color and depth channels is needed.

Doria and Radke [126] introduced a framework for joint inpainting of RGB-D images as captured by a LiDAR scanner. The proposed patch-based image inpainting method finds similar RGB-D patches in the known region of the RGB-D image. Applying the color image inpainting framework by Criminisi et al. [129], the authors found that it is challenging to find a patch with matching absolute depth values. To enable the identification of patches that are structurally similar but lie at different depths, the depth channel of the RGB-D images is replaced with depth image gradients. Taking the RGB and depth gradient image as its input, this patch-based method is able to extract the geometric information of patches that are similar in structure. In turn, the depth gradient values are resolved to absolute depth values, resulting in an inpainted RGB-D image. However, this method is unable to reproduce structures that are complex or have a structure that is not found in the known region of the image. An inpainted result with the intermediate depth image gradient values can be seen in Figure 2.13. The method performs well with the sample shown in this figure as the required textures and structures are all available in the known region of the image.

While research into multimodal feature learning is actively continuing in the field of seman-

tic segmentation, it has not been as widely applied for the task of image inpainting. One of the first joint RGB-D image inpainting frameworks was proposed by Mori et al. [127], which performs examplar-based inpainting based on several cost functions that indicate the loss of texture and spatial information. Additionally, this method minimizes the loss of geometric information by employing a normal map derived from the depth image.

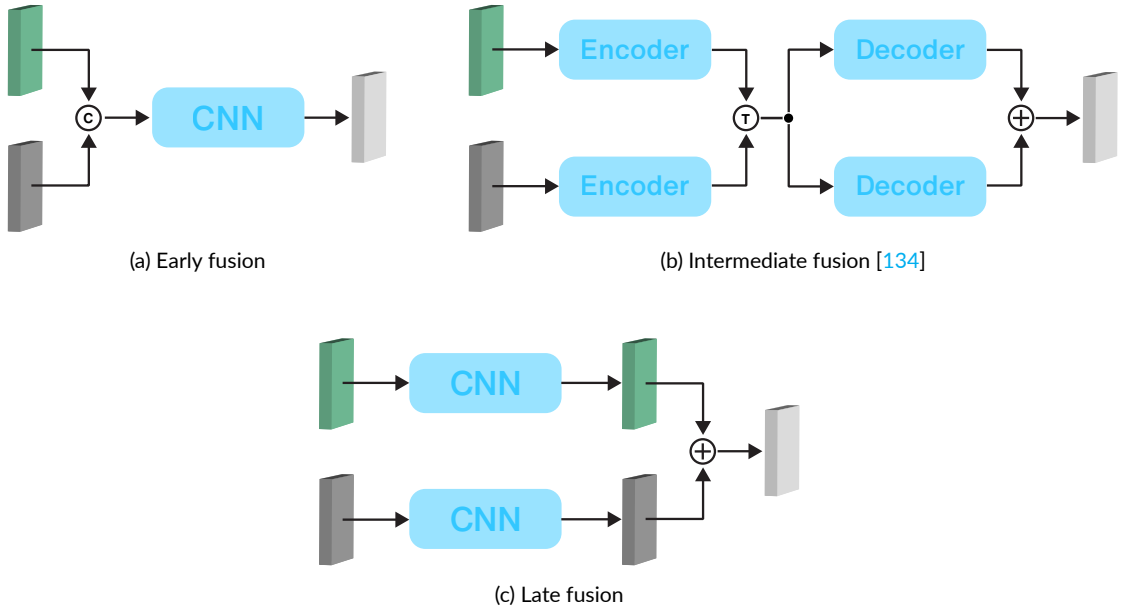
To the best of our knowledge, the first joint inpainting method of RGB-D images based on a generative model was recently introduced by Fujii et al. [47]. This method is built on the GAN-based image inpainting framework of Iizuka et al. [38], which employs two encoding branches that separately encode color and depth information. Once encoded, a feature-level fusion strategy is applied through a multi-input fusion module containing several residual blocks. The fused features are decoded separately, resulting in an inpainted RGB and depth image. During training, the authors use a mean squared error loss for the RGB channels as well as for the depth channel, which leaves depth characteristics such as surface normals unconsidered. Moreover, the proposed method was not quantitatively or qualitatively evaluated, and only two inpainted RGB-D image samples are provided. In addition, corresponding source code has not been published at this time and the publication does not provide sufficient information to reproduce and evaluate the outlined framework.

RGB-D image inpainting has not reached a comparable maturity level to RGB image inpainting, and further research is needed to discover how the geometric characteristics of RGB-D images can be considered in the inpainting process. The studies presented thus far indicate the potential of the usage of geometric representation in the inpainting process, which we elaborate on in the following chapters of this work.

2.5. Multimodal RGB-D feature learning

A different field within computer vision that commonly handles RGB-D information is semantic scene segmentation [130–133]. Humans typically perform this task through color differentiation and perception of depth. In much the same way, RGB-D images enable such methods to combine the visual information provided by the RGB channels with geometric information from the depth channel.

Early image segmentation approaches use hand-crafted features to encode the visual and geometric information of objects, surfaces and regions [119, 135, 136]. In the following years, CNN-based methods have enabled the automatic learning of cross modality feature learning [137]. Couprie et al. [138] proposed one of the first methods that utilizes multimodal image information for image segmentation. The authors made a straightforward modification to an existing CNN-based architecture by concatenating the color and depth image channels at the network input. However, this early feature fusion method (Figure 2.14a) ignores the differing characteristics of the color and depth information, which could possibly even hurt performance [137]. To augment the available information of the depth channel, Gupta et al. [136] introduced a three-channel depth encoding that comprises horizontal disparity, height above ground and the angle of the local surface normal with respect to gravity (HHA). The embedding offered significant improvements for the task of scene segmentation, which inspired a number of other segmentation approaches [115, 139].



(a) Early fusion

(b) Intermediate fusion [134]

(c) Late fusion

Figure 2.14: Types of fusion strategies that have been previously applied in the field of semantic scene segmentation. Symbols ⊕, Ⓣ, and Ⓒ denote element-wise summation, transformation and concatenation respectively.

Long et al. [115] evaluated the performance of their proposed model on solely HHA channels. While this did not show better performance compared to past methods, the authors additionally trained a two-stream end-to-end model that trains on RGB and HHA channels separately, where the predictions from both streams are summed at the final layer. This type of feature fusion is referred to as *late fusion* (Figure 2.14c). While not as significant, it is worth mentioning that Long et al. [115] also reported improved results for a data-level fusion version of their model that was trained on RGB-D data. Overall, these experiments showed significant improvements to the model's performance, demonstrating the benefits of multimodal feature learning with RGB and HHA. For segmentation tasks, the HHA depth encoding provides much needed information about the spatial relation of objects relative to the rest of the scene. However, we argue that not all of this information is as useful for models trained in our domain of HMD removal or face completion. This is due to the fact that the visual and geometric information needed for joint RGB-D image inpainting of faces does not significantly benefit from information describing the face's relation to the rest of the scene. What we do consider of interest with respect to the HHA encoding is the angle of the local surface normal, as this gives valuable information about the surface formed by the depth channel and could support Objective 1.3.

Wang et al. [134] make use of an autoencoder-based architecture with intermediate feature fusion (Figure 2.14b). The procedure starts by separately encoding the RGB and depth information through a two-stream encoder. These informative features are then fed to a feature transformation network, where the color and depth features are correlated by discovering their complementary features. The final result is obtained by fusing decision scores of the two modalities. This process does not only correlate the multimodal features, but also allows each modality to enhance their representation by borrowing features from the other modality. Methods that are built on an autoencoder-based architecture are particularly relevant to

our research, as state-of-the-art image inpainting approaches involve GANs that also contain autoencoder structures. In turn, similar modules could contribute to multimodal feature understanding in our framework.

Following a similar autoencoder-based structure, Hazirbas et al. [140] proposed a method in which the encoder consists of two branches which encode RGB and depth separately. Depth features from the depth branch are combined with the color features at several points in the encoding process. Blocks that are responsible for this are referred to as fusion blocks, which combine the feature maps of both modalities through element-wise summation. The decoder jointly upsamples the color features that have been fused with depth features to finally obtain the semantic segmentation of the input scene.

Inspired by the work of Hazirbas et al. [140], Park et al. [133] implements feature fusion through residual learning with skip-connections. To improve performance on high resolution RGB-D images, the authors built on RefineNet [141], which iteratively refines higher-level features based on several collections of low-level features. Also based on RefineNet is the method's feature fusion block, which downsamples the color and depth features after which they are rescaled before their element-wise summation. The skip-connections in this feature fusion model allows the unconstrained flow of modality-specific features through the network.

We observe that approaches that employ feature fusion successfully improve the interpretation of the complementary relation between the color and depth channels. We identify such strategies as major points of attention during the exploration for the design of our framework. However, caution is warranted in the application of feature fusion methods in our framework, as fusion methods cannot be presumed capable of blind application to image inpainting. Specifically, the tasks of semantic segmentation and image inpainting are inherently different; the former task aims to *condense* feature information into meaningful labels, while the latter aims to *expand* feature information in order to fill in the missing region.

2.6. Datasets

In this section, we give a representative overview of RGB-D face image datasets that are currently available. Specifically, we discuss a number of their characteristics such as their size, recording conditions and variations in face properties.

Originally introduced for the stimulation of research towards three-dimensional face recognition methods, the FRGC v2 dataset was introduced by Phillips et al. [142]. This dataset contains a total of 4007 captures of 466 different subjects, where each subject was captured in a controlled and uncontrolled illumination and approximately half the amount of captures showed some form of facial expression. Considering the dataset was introduced in 2005, the hardware used for capturing this dataset was a Minolta Vivid 900/910 range scanner, which produced images with a size of 640×480. Due to its high quality content and large size, the FRGC v2 dataset continues to be used for the evaluation of 3D face recognition methods to this day.

A similar dataset that was recorded around the same time is the CASIA 3D Face dataset, which was also introduced to contribute to the evaluation of 3D face recognition methods. This dataset contains 4624 scans of 123 subjects, recorded across a wide variation and combination of illumination, poses and expressions.

Name	Sensor	Subjects	Description
FRGC v2 (2005)	Minolta Vivid 900/910	477 (205 ♀, 272 ♂)	Two illumination types (for controlled capture), two facial expressions.
BU-3DFE (2006)	3DMD Digitizer	100 (56 ♀, 44 ♂)	Seven facial expressions with four intensity levels each, except for neutral. Facial texture from two views.
CASIA 3D Face (2009)	Minolta Vivid 910	123	4624 frames with varying pose, illumination and six expressions (incl. neutral).
FaceWarehouse (2013)	Kinect v1	150	Twenty expressions (incl. neutral).
KinectFaceDB (2014)	Kinect v1	52 (14 ♀, 38 ♂)	Nine expressions (incl. neutral), illumination and types of occlusion.
VT-KFER (2015)	Kinect v1	32 (18 ♀, 14 ♂)	Seven expressions (incl. neutral) in scripted and unscripted situations.
IAS-Lab RGB-D Face (2016)	Kinect v2	41	13 different conditions with varying pose, luminance and expression.
4DFAB (2017)	Kinect v1	180 (60 ♀, 120 ♂)	Six expressions, spontaneous reactions, recorded during four sessions over a period of 5 years.

Table 2.1: Representative list of RGB-D face datasets including a description of their key characteristics.

The BU-3DFE dataset was introduced to facilitate the evaluation of approaches aimed at three-dimensional facial expression classification. This dataset contains a total of 2500 scans of 100 subjects, where each subject performed seven different facial expression types across four levels of intensity for all basic emotions except the neural expression.

The FaceWarehouse dataset [145] further expanded the range of facial expressions as it consists of recordings of 150 subjects with 19 unique expression types. Introduced in 2013, this dataset was recorded with Kinect v1 which provides data that is of relatively lower quality when compared to the aforementioned high quality range scanners.

Contributing to the evaluation of Kinect-based face recognition and similar tasks, the Kinect-FaceDB [146] was introduced in 2014. This dataset consists of 936 captures of 52 subjects with nine different facial variations across illumination, pose, occlusion and facial expressions.

The VT-KFER [147] dataset further expanded facial expressions by providing spontaneous and non-spontaneous recordings of 32 subjects recorded using Kinect v1. In particular, this dataset contains 7 labeled facial expressions, in scripted and unscripted scenarios. This dataset distinguishes itself through the addition of spontaneous facial expressions, constituting a closer resemblance to real-life situations.

The IAS-Lab RGB-D Face [148] dataset contains 41 subjects captured in 13 different conditions across illumination, pose and expression. Recorded with the Kinect v2, this dataset provides an RGB image with a corresponding registered point cloud. A major drawback of this

dataset is that it was recorded in several uncontrolled environments.

Particularly remarkable for its recording timeline, the 4DFAB [149] dataset consists of 180 subjects captured in four different sessions over five years. Each subject is recorded with 6 non-spontaneous facial expressions as well as spontaneous reactions. Additionally, the dataset contains recordings of each subject's utterances of 9 different words. The size and wide range of settings of this dataset increase its usefulness for various applications.

In this section, we have aimed to give a brief but representative overview of the RGB-D face datasets that are currently available. The introduction of commodity RGB-D sensors have accelerated the creation of RGB-D datasets and has resulted in a wide range of interesting datasets. However, the amount of unique subjects portrayed in these datasets remain relatively low at this time.

When selecting a dataset for training a GAN-based model, it is essential to note that GANs aim to learn a representative estimate of the distribution of the training set [79]. On this account, it is of absolute importance to select a dataset that is representative of its domain. Bearing this in mind, the unique identities represented in the datasets previously discussed do not form a reliable representation of the domain of human faces as a whole. For this reason, we opt for the usage of a synthesized dataset based on Basel Face Model 2017 [62] which is based on the parametric 3D Morphable Model [63]. Details regarding the construction and characteristics of this dataset can be found in 4.1.

2.7. Research gap

Head-mounted device (HMD) removal is a challenging task which has emerged with the increasing usage of HMDs to observe virtual reality (VR) environments. As discussed in Section 2.1, due to the novelty of this problem, not every research direction has been fully explored. One direction approaches the task of HMD removal in a purely image-based manner, in which only a few methods have been proposed [36, 70]. Leaving out complex intermediate representations, this branch of methods resolves the occluded face region with image inpainting techniques [38]. Aside from the flawed visual quality of results generated by these existing methods, they also fail to consider RGB-D images, which are widely used for the construction of shared immersive virtual environments [6, 10, 12]. For this reason, in this work, we propose a method that aims to perform HMD removal through the joint inpainting of RGB-D images.

Inpainting images in a realistic and consistent manner has been a long-standing goal in the field of computer vision. As discussed in Section 2.4.1, generative adversarial networks (GANs) have been empirically shown to form the base of the current state-of-the-art image inpainting methods. In Section 2.4.2, we discussed how this ability has been demonstrated to be transferable to the inpainting of depth images with a number of modifications [45, 46]. However, despite the large and growing interest in RGB-D data for its wide applications, we found that there is only a small body of research that is concerned with joint RGB-D image inpainting (Section 2.4.3). Moreover, the few studies that do consider joint RGB-D image inpainting are not fully evaluated [47] and do not consider complex image structures such human faces. For this reason, we set out to explore a joint RGB-D face image inpainting framework, with the intended target application of HMD removal.

3

Architecture

In Chapter 2, we reviewed existing approaches to HMD removal, image inpainting, face completion, and the multimodal feature fusion of RGB-D data. We identified the possibility of performing HMD removal through the completion of the task of RGB-D image inpainting with a generative approach. To the best of our knowledge, a joint RGB-D image inpainting method does not currently exist. Consequently, we take an exploratory approach to the definition our method, which is guided by the research objectives defined in Chapter 1.

We selected the state-of-the-art RGB image inpainting framework by Yu et al. [41] to form the base architecture of our joint RGB-D image inpainting method. We based this decision on a number of characteristics of the framework, including its state-of-the-art performance and promising SN-PatchGAN discriminator.

The intuitive concept of our framework is based on a two-stage coarse-to-fine GAN architecture which is fed an incomplete RGB-D image, a mask, and a reference image. The first stage of the architecture produces a coarse prediction of the masked region. Subsequently, the coarse result is fed to the refinement stage of the architecture where it is further refined through two branches, one of which includes a contextual attention module. We use a reconstruction loss function and SN-PatchGAN loss function [41] to attend to the accurate reproduction of pixel values and higher-level perceptual content.

In view of our research objectives, we explore several types of other components and loss functions of our architecture. Firstly, to achieve preservation of identity, we propose a perceptual identity loss function which encourages the reproduction of distinctive facial features based on a given reference image (Objective 1.1). Moreover, as RGB-D images contain a color and depth modality, we explore different methods for fusion of these multimodal features (Objective 1.2), including data-level feature fusion and hybrid feature fusion which combines both data-level and feature-level fusion. Lastly, we discover methods that stimulate the architecture to interpret the depth values as a surface (1.3). Specifically, we discuss the employment of a surface normal loss function [46], contextual surface attention module [46] and surface normal discriminator.

We begin this chapter by elaborating on our choice of base framework and introduce the fundamental aspects of this architecture [41]. Subsequently, we define our identity loss function that is based on a pretrained face recognition model. In the following section, we elaborate on how we establish multimodal feature understanding in the coarse stage of the model and discuss the difference between data-level fusion and several versions of hybrid fusion. Finally, we discuss the aspects of depth surface reproduction and the formulation of the surface loss function, contextual surface attention module, and surface normal discriminator.

3.1. Baseline framework

We adopt the two-stage RGB image inpainting approach proposed by Yu et al. [41] as the base architecture of our framework. This architecture possesses a number of advantageous characteristics in comparison to others:

- **Free-form masks** The architecture of this framework uses gated convolution (Section 3.1.2), allowing masks to have any size and to appear anywhere in the input image. Inadvertently, one may assume that HMDs typically cover a rectangular region of the face [36, 70]. However, during mediated social interaction, the head pose of the user with respect to the RGB-D sensor varies significantly. Accordingly, the mask covering the occluded face region has a variable location and size. This makes the ability to use free-form masks particularly useful in our case.
- **SN-PatchGAN for semantic learning** The authors of the framework propose a variant of generative adversarial networks. The discriminator of this architecture provides a means of focusing on different locations and semantics across image channels. In turn, this method could be particularly useful for capturing different types of semantics represented in the multimodal RGB-D images that we aim to inpaint.
- **State-of-the-art performance** The framework achieves state-of-the-art results on several benchmark datasets such as Places2 [150] and CelebA-HQ [151]. Specifically, the framework has been shown to perform well on inpainting face images, which demonstrates its potential for our objective.
- **Publicly available source code** The authors have made the source code publicly available and actively provide comprehensive answers to anyone's questions with respect to details of the framework.

In the following sections, we discuss the key components that make the architecture of this framework unique. Specifically, we discuss the concept of contextual attention, gated convolution, and the SN-PatchGAN discriminator.

3.1.1. Contextual attention

When looking at an image, humans intuitively pay attention to specific image locations and features. This mechanism assists humans in the identification of relevant objects or regions by selectively increasing the activity of sensory neurons [152]. Recent methods have facilitated the investigation of attention mechanisms to convolutional neural networks [153–155].

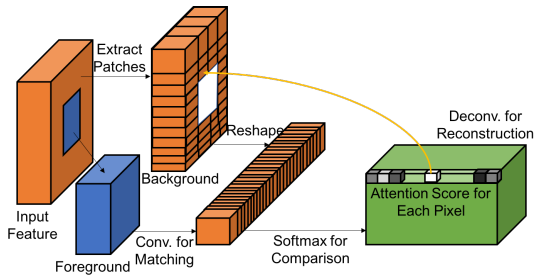


Figure 3.1: Visualization of contextual attention layer. Image reproduced from [40].

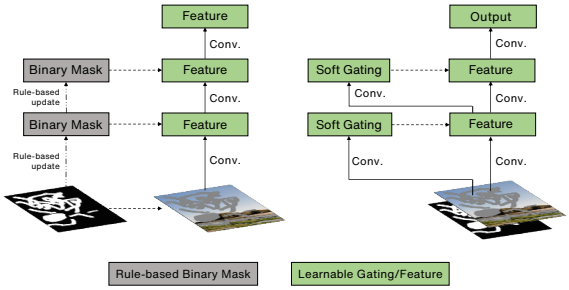


Figure 3.2: Visualization of gated convolution. Image reproduced from [40].

Yu et al. [40] applied this concept to image inpainting and introduced a contextual attention layer that facilitates the propagation of related feature patches at distant spatial locations with respect to the missing region. In this case, the missing and known region of the image are referred to as the foreground and background respectively. Illustrated in Figure 3.1, the first step of this procedure involves matching features of the foreground f and background b. This process starts with the extraction of 3×3 feature patches from the the background segment of the given input feature map, which are then reshaped as convolutional filters. Following this, the similarity between the background patches $\{b_{x',y'}\}$ and foreground patches $\{f_{x,y}\}$ are calculated based on their cosine similarity. The cosine similarity of each combination is then weighed with a scaled softmax function: $s^*_{x,y,x',y'} = \mathrm{softmax}_{x',y'}(\lambda s_{x,y,x',y'})$, where λ is a constant value. Finally, the matched background patches are used as deconvolutional filters to reconstruct the missing image region.

3.1.2. Gated convolution

The significant number of proposed image inpainting methods do not have the ability to inpaint non-rectangular masks [38, 40, 104]. This is a consequence of the workings of two-dimensional convolution. As described in Section 2.2.2, a convolutional layer learns filters for each input channel which can be used for convolution at any spatial location in the image. This type of convolution is suitable for tasks such as image classification and object detection, where all input pixels can be considered as *valid* [41]. However, Yu et al. [41] states that in the case of image inpainting, pixels within the missing region are considered to be *invalid*. The conventional convolution operation does not distinguish *valid* or *invalid* pixels and features, which causes ambiguity and visual artifacts [41, 105]. This issue persists in deep layers, where synthesized pixels and features form an ill-founded context for any further synthesization.

To solve this problem, Yu et al. [41] introduced gated convolutions which learn a dynamic feature selection mechanism for each channel and spatial location. Rather than classifying all spatial locations as either valid or invalid, gated convolutions learn a soft mask based on feature data (Figure 3.2. Yu et al. [41] defines gated convolution as:

$$\begin{aligned} \text{Gating}_{y,x} &= \sum\sum W_g \cdot I \\ \text{Feature}_{y,x} &= \sum\sum W_f \cdot I \\ \text{O}_{y,x} &= \phi(\text{Feature}_{y,x}) \odot \sigma(\text{Gating}_{y,x}) \end{aligned} \tag{3.1}$$

where W_g and W_f denote separate convolutional filters, ϕ is an arbitrary activation function, σ is the *sigmoid* function, and $\odot$ refers to element-wise multiplication.

As such, gated convolutions allow masks to have any shape and location within the boundaries of the input image. By eliminating the ambiguity regarding valid and invalid pixels through gated convolution, the proposed image inpainting framework achieves results with less visual artifacts and inconsistent colors when compared to conventional and partial convolution [105].

3.1.3. SN-PatchGAN

SN-PatchGAN is a GAN loss proposed by Yu et al. [41], designed to be used for training image inpainting framework that handle free-form masks that may appear at any spatial location. The discriminator is a CNN composed of six strided convolutions, and is spectrally normalized [107].

Given an input source image and a binary mask image, the CNN-based discriminator returns a 3D feature map of shape $\mathbb{R}^{h \times w \times c}$, where h, w, c denote height, width and number of channels respectively. In turn, the SN-PatchGAN loss is applied to each of the resulting feature points. Essentially, this defines $h \times w \times c$ GANs, each of which are employed at their respective spatial location and channel. As opposed to previous inpainting methods [38], the usage of a global discriminator is unnecessary as the receptive field of each neuron in the feature map can cover the entire input image.

Moreover, the resulting three-dimensional feature map allows the discriminator to capture different types of semantics occurring within each image channel, which obviates the usage of perceptual losses. We theorize that such capability is particularly useful in the case of RGB-D images, as these contain multiple modalities represented in separate channels. As such, we argue that the SN-PatchGAN discriminator is of great significance in a framework that jointly inpaints RGB-D images.

3.2. Identity preservation

In general, when a deep neural model inpaints an image, it relies on the combination of contextual information provided by the known region of the image. Inferring missing regions becomes more challenging when the regions in question contain highly detailed information. We work with a challenging instance of this, which is the complex appearance and structure of the human face. The human faces contains a large amount of information that is cognitively distinctive to a person's identity [156]. Using their perceptive capacity, humans extract a wide range of information to mediate face recognition. This information forms one of the most evocative factors in social experience [48] and has a profound impact on communication.

When inpainting a human face with the discussed base framework, it produces the most plausible contents of the missing region based on the information it is provided with, which

Figure 3.3: RGB channels of inpainting result without preservation of identity. From left to right: RGB ground truth, masked RGB input, inpainted RGB result.

effectively comprises the low-level and high-level features of the masked input image. While the framework is effective towards the inference of visually and semantically consistent image regions, it lacks knowledge regarding the person's invisible distinctive facial features. Thus, when inpainting a face image with this framework, it is almost certain that the inferred region turns out to have a different perceived identity when compared to the ground truth image. An example of this can be seen in Figure 3.3.

Keeping the projected application of our framework in mind, this forms a major issue in the scenario of face-to-face conversation in a shared VR environment. As a result, the connections between the user's offline and online *self* are compromised and likely form a barrier in any social interaction performed within the shared VR environment [30].

3.2.1. Identity loss

We address this issue through the introduction of the identity loss function $\mathcal{L}_{\text{ID}}$. This loss function supervises the model in the identity-preserving reconstruction of the face represented in an RGB-D image. We achieve this as follows. Similar to other perceptual loss functions for identity preservation [36, 52, 113, 114, 157], we use a pretrained face recognition model for this purpose. Moreover, we require a reference image of the same person for the input of our network. During training, an identity embedding of both the reference image and the inpainted image is computed by passing them through the pretrained face recognition model. Following this, we calculate the L2 distance between the two identity embeddings. This value forms the identity loss value. Throughout the training process, this loss is minimized to reduce the distance between the embeddings of the generated image and the given reference image.

The identity loss function $\mathcal{L}_{\text{ID}}$ is based on the notion of perceptual loss functions that use features extracted by pretrained networks for applications such as style transfer [158], super resolution [158] and image generation [159]. A number of works use perceptual losses for the preservation of identity [52, 113, 114] given a source image and a reference image. We define a similar identity loss function that encourages the preservation of identity features between the inpainted image and a given reference image. The calculation of this loss function starts by obtaining the activations of specific layers for the reference image x_{ref} and the inpainted result image x_{pred} through a forward pass of a pretrained face recognition model M_{ID}. In the

task of face verification, the distance between these two layer activation values represents the likelihood that the images represent the same person. To complete the calculation, we measure the mean squared error (MSE) between the respective activation values. The identity loss function is defined as follows:

$$\mathcal{L}_{\text{ID}}(x_{\text{pred}}, x_{\text{ref}}) = \text{MSE}(M_{\text{ID}}(x_{\text{pred}}) - M_{\text{ID}}(x_{\text{ref}})) \tag{3.2}$$

3.2.2. Selection of a pretrained face recognition model

In the previous subsection, we defined $\mathcal{L}_{\text{ID}}$, which uses the pretrained model M_{ID} to obtain an identity embedding for both the reference image and inpainted image. The concept of this identity loss function is used in several works [52, 113, 114], each which have different reasons for their choice of model M_{ID}. In this section, we will discuss which face recognition model M_{ID} we use in our work, and why we consider it the best fit for our purpose. Moreover, we quantitatively evaluate the performance of each pretrained model on a subset of our test set (Section 4.1).

Face recognition has been a long-standing tasks which has been widely explored over the years. Many state-of-the-art face recognition models use deep neural networks as a backbone such as VGGNet [160], GoogleNet [87] or ResNet [58]. These models are trained with images from massive datasets that are fed to the network to obtain a complex face representation. The resulting representations are then compared through a distance measure, which most commonly are the L2 distance or cosine distance.

To pick a suitable face recognition model for our application, we reviewed several available models. In our case, we assume that the reference image may be captured in a different environment than our source image. Therefore, we require our model to be invariant to facial properties such as expression, illumination and pose. Moreover, we seek a model that requires little to no preprocessing. This is in view of the fact that any preprocessing steps would add additional constraints and computational steps regarding the input of our framework. Furthermore, we look for a model that performs well on our synthesized data as well as real-world data. Since, we want to minimize the required training time while retaining the model's ability to extract meaningful real-world face representations. The latter is particularly important if the synthetically trained model is fine-tuned with real-world data at a later point.

Based on the aforementioned considerations, we select two model candidates: 1) ResNet50 [58] trained on the VGGFace2 dataset [53], and 2) FaceNet [57] trained on the MS-Celeb-1M [54] dataset. The VGGFace2 [53] dataset is a large-scale face dataset containing 3.31 million images of 9131 subjects. Each of these images has large variations in pose, illumination and expression, as well as ethnicity and profession. Trained on this data, it has been demonstrated that ResNet50 achieves state-of-the-art performance on evaluation benchmarks. The MS-Celeb-1M [54] dataset consists of 10 million face images of 100,000 subjects, sourced from the public internet. Trained on this dataset, the face identity embedding network FaceNet similarly obtains state-of-the-art performance on several evaluation benchmarks. One major drawback regarding the usage of the FaceNet network compared to the ResNet50 network trained on VGGFace2 is the requirement of spatially-aligned images based on facial landmarks.

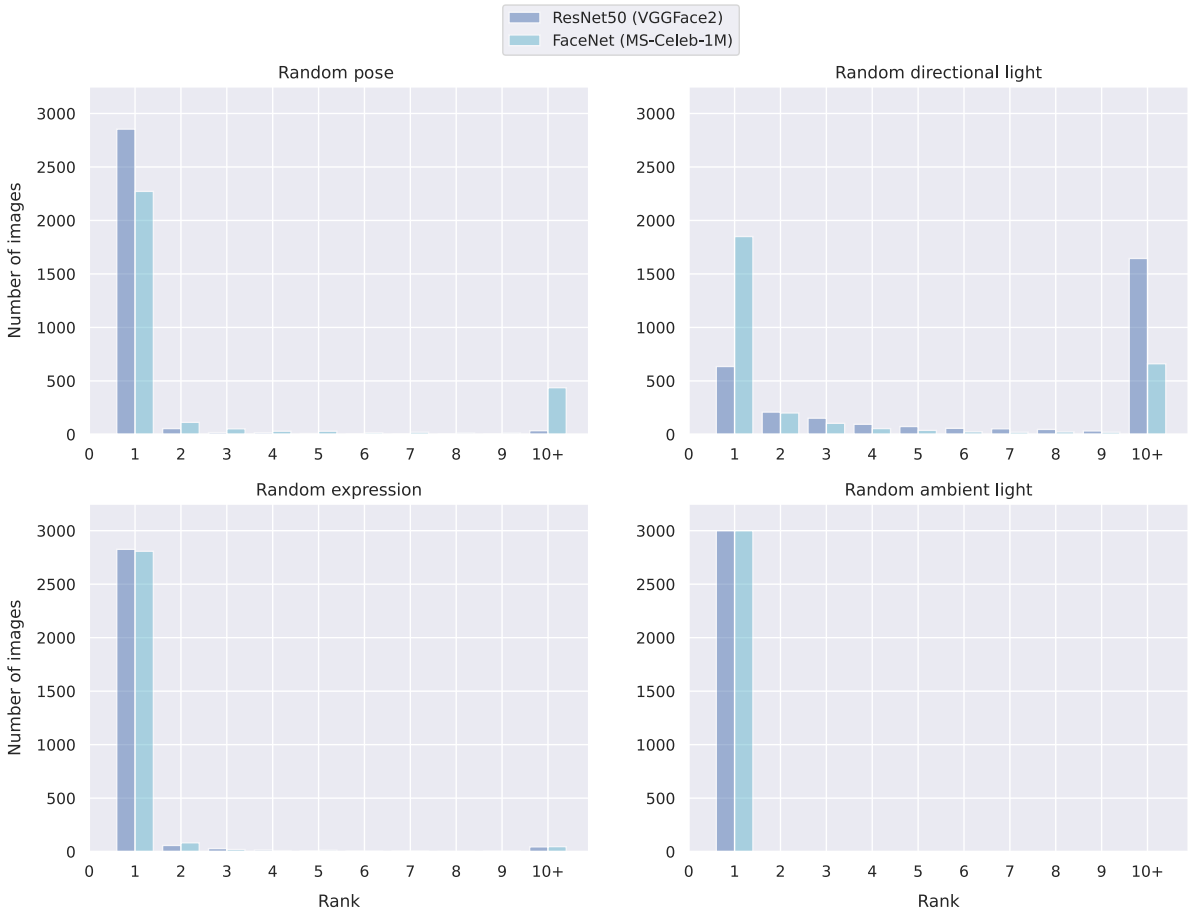


Figure 3.4: Rank-N accuracy results for frontal face images versus several variations: random pose ($\pm 30^\circ$ for p_{pitch}, p_{yaw} and p_{roll}), random directional light (while at eye-level, moved randomly by ± 100 centimeters in left and right direction), random expressions, and random ambient light between 80 and 120.

Validation of application

To validate the applicability of the chosen pretrained model on images from our dataset, we conduct a small experiment. On a dataset containing 3000 RGB face images with a random identity, we measure the model's ability to retrieve identity features when comparing feature vectors extracted from images under different conditions.

In this experiment, we performed a pairwise search between the images with frontal faces and images with one of the following transformations:

- **Random expression**
- **Random pose** p: p_{pitch}, p_{yaw} and p_{roll} in range $[-30^\circ, 30^\circ]$
- **Random directional light**: at eye-level, moved left or right by distance in range $[-100$ cm, $+100$ cm$]$
- **Random ambient illumination** a: in range $[80, 110]$

This dataset was generated based on the parametric 3DMM model, using the pipeline described in Chapter 4.1.

Based on the embeddings generated by the identity model M_{ID}, we performed an approximate nearest neighbor search using the *annoy*[1] Python library. The ranking results of the pairwise searches are shown in Figure 3.4. We observe that the pretrained ResNet50 identity model shows excellent performance when comparing the frontal view images and random poses. Interestingly however, the performance of the identity model in case of the matching of faces with random illumination is significantly worse. We hypothesize that this is a result of the harsh shadows that are cast by the directional illumination. The impact of this shortcoming can be minimized by requiring the captured subject to be situated in a well-lit environment.

We observe similar results in terms of the robustness of the pretrained FaceNet identity model with respect to expression and ambient illumination. However, looking at the result plot in regards to random pose variation, we note a significantly worse performance of FaceNet compared to the ResNet50 model. In contrast, the FaceNet model achieves a remarkably higher performance when it comes to directional illumination.

These results provide important insights into the application of a face identity model trained on real-world datasets to our synthesized dataset. Bearing our target application in mind, we deem robustness against pose to be more important than robustness against directional illumination. Moreover, the ResNet50 model has the clear advantage of not requiring any processing steps prior to the inference step of this model. In addition, the ResNet50 model has a shorter inference time: FaceNet processes 15 images per second, whereas ResNet50 processes 60 images per second. Based on these considerations, we opt for the usage of the pretrained ResNet50 model in our framework.

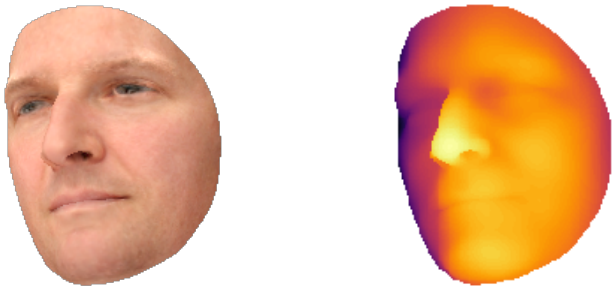
3.3. Fusion of color and depth information

The goal of image inpainting is to fill in a missing image region in such a way that it is undetectable to a human observer. GAN-based approaches build on the concept of convolutional neural networks (CNN) that capture image features at several levels. As discussed in 2.2.2, shallow convolutional layers capture low-level visual and spatial features, whereas deep convolutional layers capture semantic features. The base framework that we build on consists of a CNN that follows an autoencoder structure, which conceptually functions as follows. Firstly, the encoder transforms the model input from image space into a high-level latent feature space. In turn, the decoder uses this feature representation to produce a completed image.

The learning process of our model is designed to learn to construct feature representations that not only capture the low-level visual information, but also the semantics of the visual structures. To successfully complete an image, these representations need to capture the contextual content of the image, as well as define a plausible hypothesis for the missing region [104]. Accordingly, feature representation learning lies at the core of completing the task of image inpainting with CNNs. Therefore, it is crucial to design our image inpainting framework and its respective training process in such a way that it is able to learn a semantically meaningful joint feature representation for RGB-D images.

Consider an RGB-D image of a human face, as pictured in Figure 3.5. When taking a closer look at the RGB color image and the depth image, we observe that they are both full of charac-

[1] https://github.com/spotify/annoy/



RGB Depth

Figure 3.5: RGB color image and depth image generated by our synthesization pipeline presented in Section 4.1.

teristic information. From a human perspective, we can identify several facial features by their appearance and visual structure in the color image, such as the eyes, nose, and mouth. On the other hand, we recognize similar regions in the corresponding depth image, but in a different way. Aside from the visibility of a few facial features, the primary type of information conveyed by the depth image is the shape of the face. For instance, we intuitively observe the relatively large extrusion of the nose and shallow chin area. However, visual texture information of high-frequency components such as the eyebrows, lips and eyes are not represented in the depth image.

Based on these observations, it is clear that both modalities contain information that is partially related, but each are comprised of different characteristics and statistical properties. In other words, each modality contains information that is unique to itself and that cannot be directly derived from information from the other modality. While it may be reasonably straightforward for humans to discover the partial relations between the color and depth modality, it is a major challenge for a CNN-based network to achieve a similar level of understanding.

This brings us to the question of how we should capture the features of each modality, and at what point they should be combined. In Section 2.5, we discussed several approaches to multimodal feature understanding. In general, combining features can be achieved at several points in a network through early fusion, intermediate fusion, or late fusion. Feature fusion between color and depth has been widely explored in the fields of object detection and image segmentation, in which the multimodal RGB-D data contributes to robustness and accuracy. In our work, we discover how feature fusion can be applied to the task of joint RGB-D image inpainting. In this section, we briefly explain each strategy and discuss the design of our architecture in combination with two fusion types: data-level fusion and hybrid fusion. Whereas data-level fusion involves the concatenation of both modalities at the input of each stage, hybrid fusion combines this with feature-level fusion in the coarse stage of the architecture. The rest of this section is organized as follows: in Section 3.3.1, we explain and discuss the concept and effects of data-level fusion. In Section 3.3.2, we explain and discuss the notion of hybrid fusion. Moreover, we describe the employment of a number of proposed candidate feature-level fusion strategies within the concept of hybrid fusion.

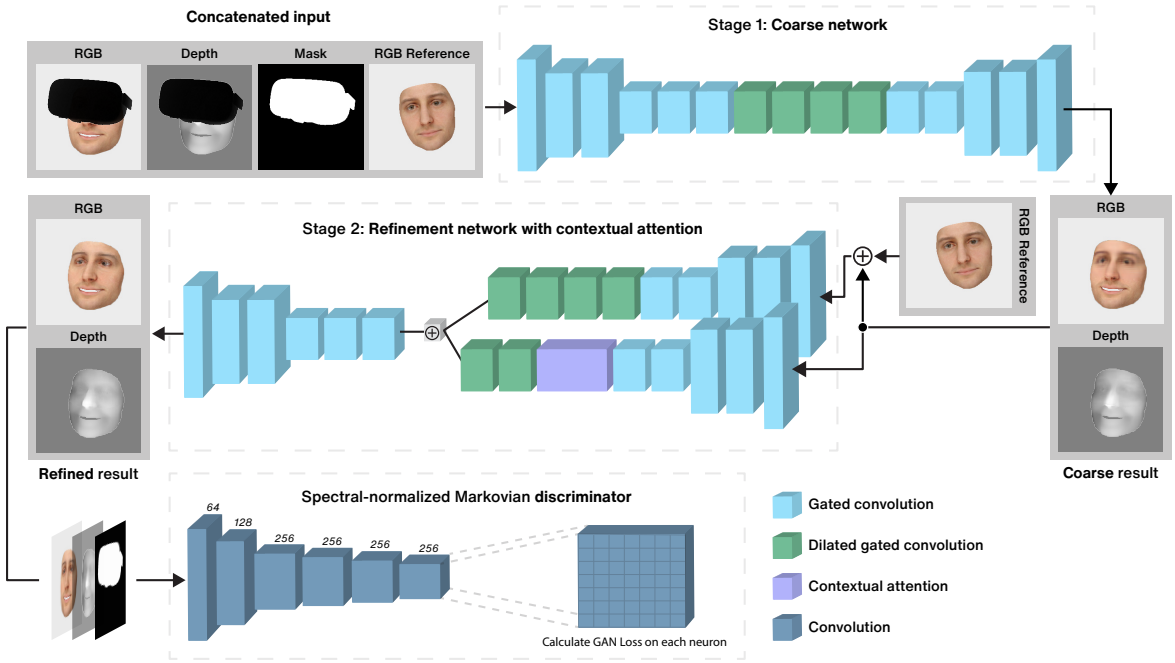


Figure 3.6: Overview of the RGB-D image inpainting architecture with data-level fusion.

3.3.1. Data-level fusion

Data-level fusion, also known as early fusion, refers to a feature fusion strategy that involves the combination of multiple sources of unimodal data at the input level of the network. In the case of an RGB color image and a corresponding depth image, this involves the concatenation of their combined image channels. Considering we may assume these two images are spatially aligned, this operation does not have to rely on any preprocessing or prior feature extraction steps. This makes the naive strategy of data-level fusion a tempting approach.

The fact that data-level fusion allows the joint capture of all image channels has advantages. The main advantage of data-level fusion is that it allows the network to truly capture the multimodal nature of the concatenated input data. Considering the spatial consistency between the color channels and the depth channel of the RGB-D images, the network is able to learn features that aim to capture both modalities jointly. Consequently, data-level fusion has the potential to exploit the strong relation between the color and depth image.

However, data-level fusion also has its disadvantages. Firstly, the color and depth images have different characteristics as well as different statistical properties. In turn, it may be a challenge to reliably construct meaningful joint features. For example, while features consistent across the two modalities will be prioritized, features that occur solely in the depth modality may be overlooked. Meaningful joint features construction becomes even more difficult when there exists noise or missing data in either modalities, which is commonly the case for current commodity RGB-D sensors [26, 27].

We want to investigate the effects of early data-level fusion for joint inpainting of RGB-D data. As such, we define a framework that combines the color image and depth image through early data-level fusion. This involves the concatenation at the network input level, as shown in Figure 3.6. Consequently, the input layer of the network receives one additional image channel.

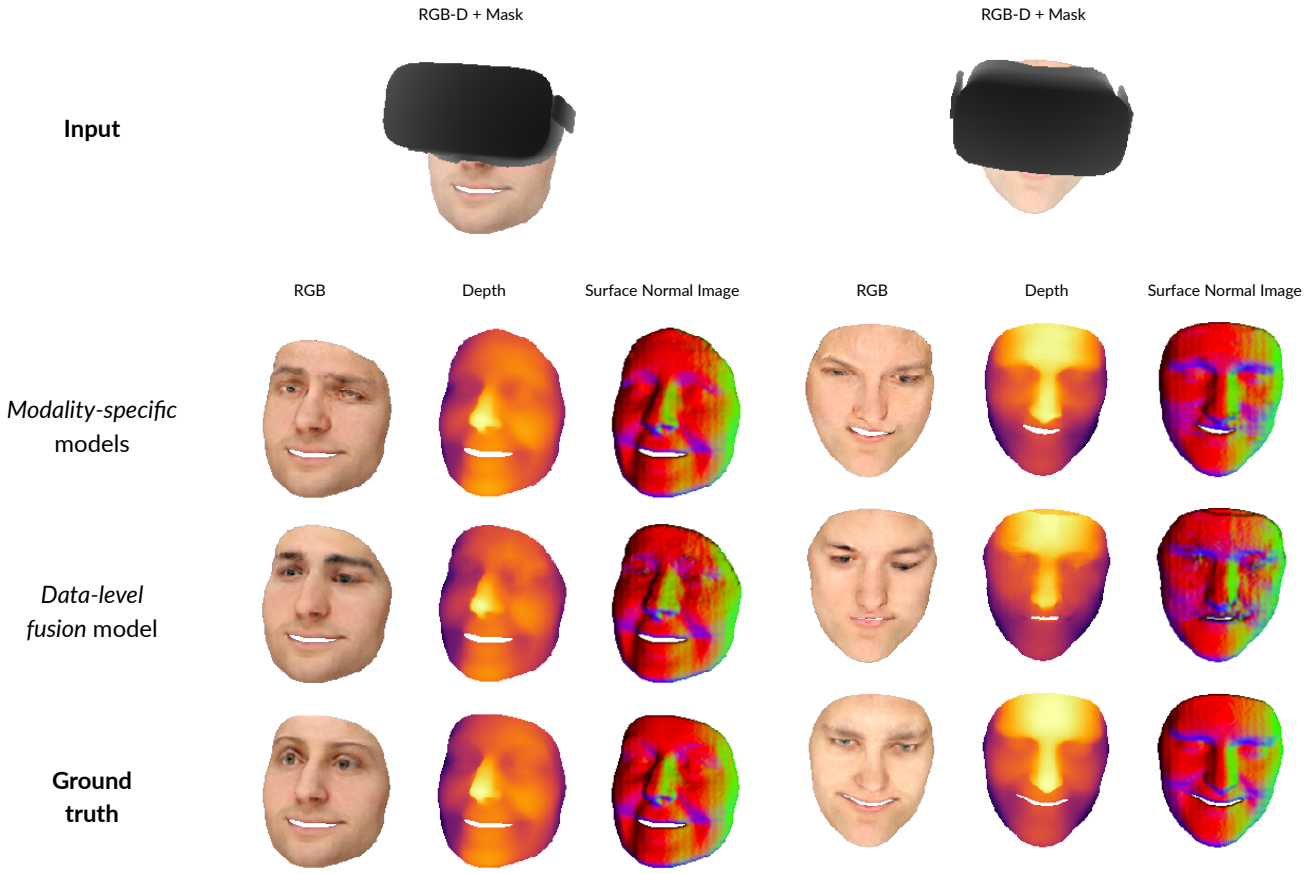


Figure 3.7: Comparative results color images, depth images and surface normal images, generated by modality-specific RGB and depth image inpainting models and our data-level fusion model.

Early evaluation of data-level fusion

We consider data-level fusion as the initial and most straightforward attempt of multimodal feature learning. This simplicity stems from the fact that a single model is used to learn the relation between the statistical properties of the raw color and depth image data. Throughout the training process, the joint representation of all image channels is learned and converted to a joint high-dimensional feature space.

We discuss the evaluation of this fusion strategy extensively in Chapter 4. However, to motivate the steps undertaken in the rest of this chapter, we will briefly touch upon the qualitative performance of data-level fusion through a visual examination.

In Figure 3.7, we provide a comparison between the data-level fusion model described above and two models that have been individually trained for RGB color image inpainting and depth image inpainting. We would like to emphasize that the RGB and depth output of the *modality-specific models* come from two independent models that have no knowledge of each other. An estimated surface normal image is shown in addition, which demonstrates the continuity and smoothness of the depth channel. More information on the estimation of surface normal images can be found in Section 3.4.1. The dataset that was used for this experiment was synthetically generated, of which the process is described in Section 4.1.

The goal of this experiment is not to evaluate whether the inpainted faces look exactly like their respective ground truth image, but to evaluate the visual quality of the inpainted results. On first glance, it is clear that the RGB and depth images lie far from the ground truth images. This is largely due to the fact that none of the models use our proposed identity loss function, as this loss function cannot be used for depth image inpainting models.

The inpainted results in Figure 3.7 contain a significant amount of noise, and facial features often do not appear to be symmetric. Turning to the depth images and their alternative representations as estimated surface normal images, it becomes clear that the resulting depth images inpainted by the data-level fusion model contain a significant amount of noise. This is especially apparent in the surface normal images, which signifies that the surfaces formed by the depth values are irregular and noisy.

Nevertheless, it is interesting to observe the reasonable visual quality of the results of data-level fusion when keeping in mind that no specific steps were carried out to stimulate the construction of complementary features. However, there is ample space for improvement of the jointly trained model with respect to the modality-specific models.

3.3.2. Combining fusion methods: hybrid fusion

In this section, we propose an alternative fusion method, which we refer to as hybrid fusion. In our visual evaluation of the usage of a data-level fusion technique through the concatenation of the color and depth image, we observed that the depth channel is not accurately reproduced (Figure 3.7). We hypothesize that the cause of this is the feature learning process of data-level fusion, which obstructs the network in the construction of features that are both modality-specific and shared among the two modalities. Specifically, as the RGB-D data is fed to the network in a joint manner, there is no mechanism or operation that is explicitly aimed at combining the knowledge of the statistical properties of the RGB and depth modality.

Our objective is to improve the feature obstruction process while retaining the merits of data-level fusion. It is for this reason we propose hybrid fusion. This fusion strategy leverages the fact that our framework consists of two stages: a coarse stage and a refinement stage (Figure 3.6). In the coarse stage, we replace the single-stream encoder-decoder structure with a modality-specific dual-branch encoder-decoder structure. Moreover, to facilitate the construction of complementary features, we add feature-level fusion in the coarse stage of the architecture. We experiment with several types of feature-level fusion at several network depths. As such, the refinement stage of our hybrid fusion architecture remains unchanged.

We refer to this fusion method as *hybrid fusion*, because it uses two different types of fusion within a single architecture. In the first stage of network we perform feature-level fusion, whereas we perform data-level fusion at the start of the second stage.

The intuitive reasoning for the modality-specific coarse stage is to provide the refinement stage of the architecture with a multimodal statistical prior of the color and depth channels of the missing region. In turn, the refinement stage of the architecture is able to further refine the relation between the color and depth modality, building on the coarse prediction.

We explore several types of feature-level fusion in the coarse stage. In particular, we consider fusion through fusion through *summation* [140], *single-path residual* fusion and *multi-path residual* fusion.

Fusion through summation

In this section, we define the feature-level fusion of the coarse stage as fusion through summation. This fusion strategy is commonly used in deep architectures, such as FuseNet [140], which is aimed at semantic segmentation of RGB-D images. The FuseNet architecture employs two encoder branches, for RGB and depth information. These branches each receive input of their respective type, and construct modality-specific features. Features from the depth branch are fused with features in the RGB branch through a fusion layer. Given color features and depth features, the fusion layer performs element-wise summation, of which the result is divided by two. The resulting features are fed to an RGB-D decoder, which produces a semantic segmentation map.

In the evaluation of FuseNet, it was found that the segmentation accuracy improves with the increase of the amount of fusion layers used in the encoders. Hazirbas et al. [140] theorize that the first layers of the encoders benefit from fusion the most, as depth information can provide complimentary information to low-level feature construction, such as edges and corners.

We take inspiration from this fusion strategy, and aim to investigate whether fusion through summation can benefit the joint feature construction of RGB-D images, to ultimately improve the visual quality of the inpainting result of our architecture. We define fusion function F_{sum} as follows:

$$F_{\text{sum}}(\mathcal{X}_{color}, \mathcal{X}_{depth}) = \frac{(\mathcal{X}_{color} + \mathcal{X}_{depth})}{2} \tag{3.3}$$

Where $\mathcal{X}_{color}$ and $\mathcal{X}_{depth}$ are activations with equal dimensions, which represent the features and gates from the color and depth branch respectively. Thus, this operation adds both the feature values and gating values of both modalities.

Moreover, we add batch normalization (BN) to every non-output convolutional layer, before non-linear activation. As outlined by Hazirbas et al. [140], the scale and shift parameters of BN layers learn to combine the color and depth features in the optimal way and aim to prevent features from being overwritten by features from the other modality. A downside of employing BN is the possibility of increased noise in the inpainted output of the model, as found by Yu et al. [40].

We explore the positioning of the fusion layers through the coarse stage of the architecture. The positions of fusion layer is shown in Figure 3.8, of which the effects will be elaborated on in Chapter 4.

Residual fusion

Residual learning was first proposed alongside the introduction of the ResNet architecture by He et al. [58]. The ResNet architecture consists of modular building blocks referred to as residual units that contain a pair of convolutional blocks and a skip connection. Skip connections enable deeper architectures, which commonly enable more accurate results in various computer vision tasks.

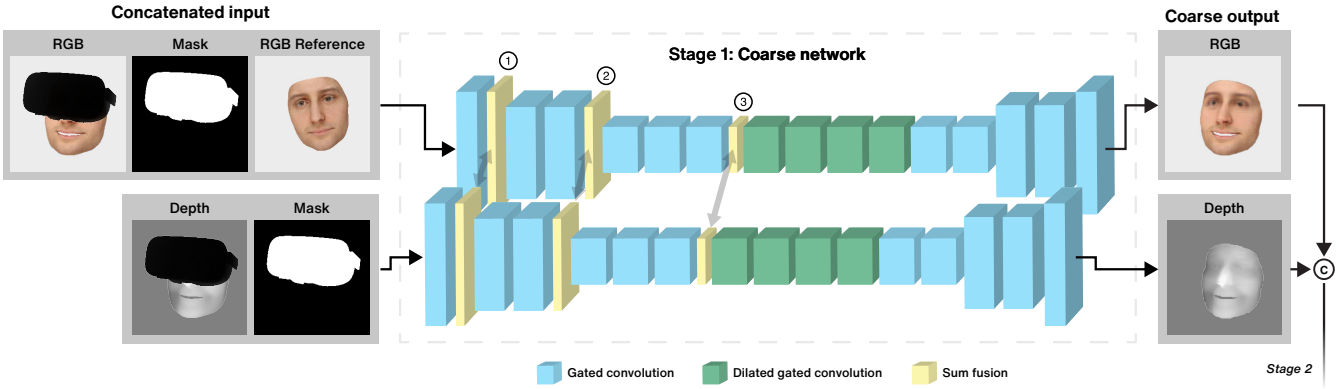


Figure 3.8: Overview of our RGB-D image inpainting architecture with fusion through summation. Three candidate fusion locations throughout the network are marked with numbers.

Aside from the aforementioned benefits, skip connections also find an interesting use case in multimodal feature fusion [47, 132, 133]. Skip connections help to exploit the complementary characteristics of color and depth images. Specifically, they provide an alternative path through the residual unit, allowing multimodal feature enrichment, while retaining meaningful modality-specific features. Fujii et al. [47] outlined a form of residual fusion for joint RGB-D image inpainting, where the generator consists of nine intermediate residual units, surrounded by modality-specific encoders and decoders. As the authors do not provide any evaluation or source code of their work, the performance of this approach is unclear at this moment. We take inspiration from this approach and the field of semantic scene segmentation, and discover applications of residual fusion in our hybrid fusion framework. In this section, we propose two forms of residual fusion, built with two types of residual units: single-path residual units and multi-path residual units. The modular residual fusion module consists of several of such residual units and receives its input from a color- and depth-specific encoder of which the activations are combined with the unit input in additive fashion. While a concatenative skip connection could possibly improve the feature reusability and, ultimately, feature quality, this would add too many network parameters for the GPU memory that is available to us. The position of the residual block with respect to the rest of the architecture is shown in Figure 3.11.

Having defined the residual fusion module, we now turn to the definitions of the single-path residual unit and multi-path residual unit. As our base framework uses dilated gated convolution, we define the r-dilated gated residual unit, which uses gated convolution with dilation rate r (Figure 3.9). Each residual unit contains two dilated gated convolution operations including activation: the *sigmoid* function for the gating values and the *ReLU* function for the features. Before the first residual unit, the features of the depth and color branch are summed in element-wise fashion. As the resulting features are fed into the unit, r-dilated gated convolution is performed twice, after which they are summed with the original input. Before they are fed back, the features pass through a *ReLU* activation function once again. As the input, output and skip connection paths of this module are singular, we refer to this module as *single-path residual fusion*.

The *single-path residual fusion* module combines the features from the color and depth encoder branches through addition. Consequently, the feature sets are refined in a unified fash-

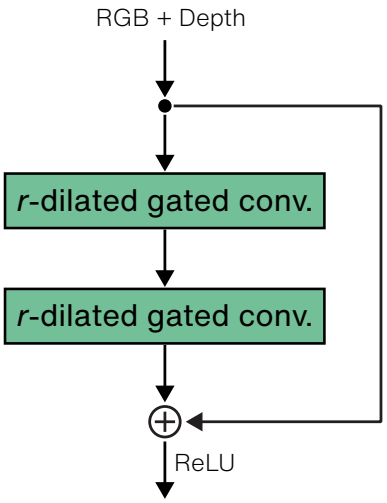


Figure 3.9: Single-path gated r-dilated residual unit, where r represents the dilation rate. As *ReLU* activation is included in gated convolution, we do not visualize intermediate activation.

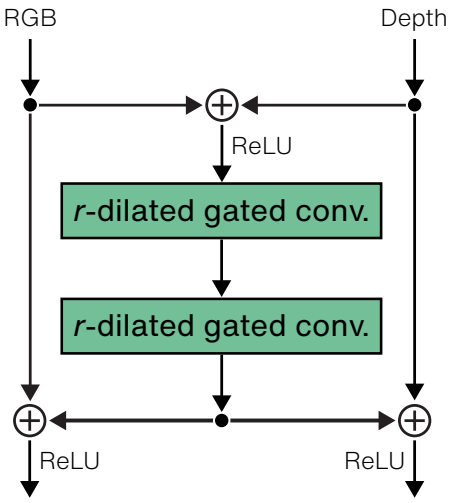


Figure 3.10: Multi-path gated r-dilated residual unit, where r represents the dilation rate. As *ReLU* activation is included in gated convolution, we do not visualize intermediate activation.

ion, after which they are fed back to the color and depth decoder branches. While this stimulates joint feature construction, we expect that some modality-specific features may be sacrificed as a result of the addition operation at the start of the module. For this reason, we investigate a different variation of the module, which consists of **multi-path** residual units. As their name suggests, we define multi-path residual units to have two modality-specific inputs, outputs and skip connections, as shown in Figure 3.10. As the color and depth features are fed into this unit, they are summed in element-wise fashion. Following this, r-dilated gated convolution is performed twice, of which the resulting features are summed element-wise with the features of the two modality-specific skip connections. Finally, each of the resulting features go through a *ReLU* activation function before they continue their journey through the network. We hypothesize that the addition of modality-specific skip connections improves the retention of modality-specific features in multi-path residual units compared to single-path residual units.

The single-path residual fusion module and the multi-path residual fusion module both contain either four or six residual units of their respective type. We denote single-path and multi-path by *SP* and *MP*, respectively, followed by the number of residual blocks used in the coarse stage (e.g., Residual-*SP4*). These modules replace an equal amount of r-dilated gated convolution layers in both the color and depth branch. The rate r of each residual unit equals the dilation rate of the convolutional layer it replaces. We evaluate the effects of single-path versus multi-path residual fusion and the amount of employed residual units in Chapter 4.

3.4. Surface interpretation

In the previous section, we discussed several feature fusion strategies to improve the meaningfulness of feature representations in our framework. While doing so, we noticed that this did not significantly decrease the amount of depth noise that is visible in the output images of the network. Furthermore, we note that while the resulting color image has a near seamless con-

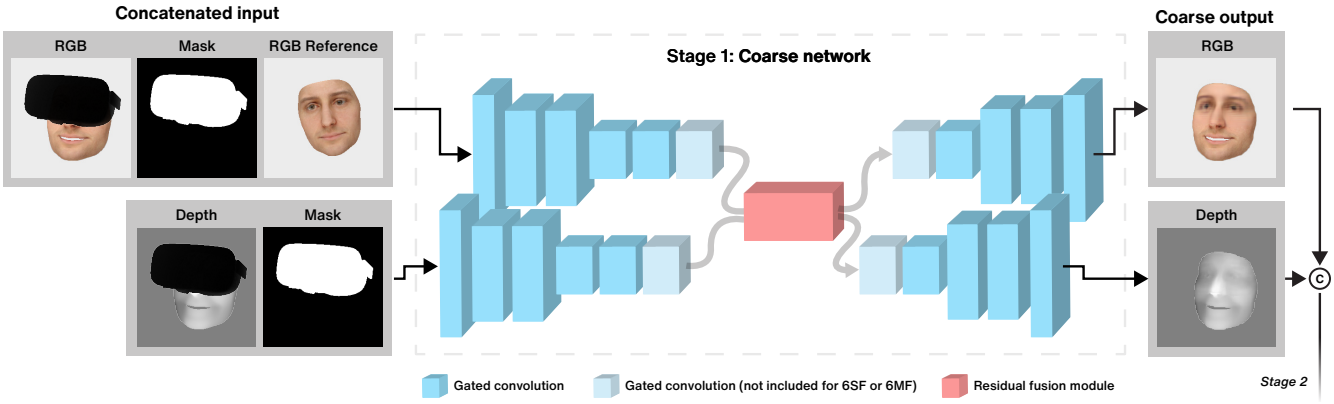


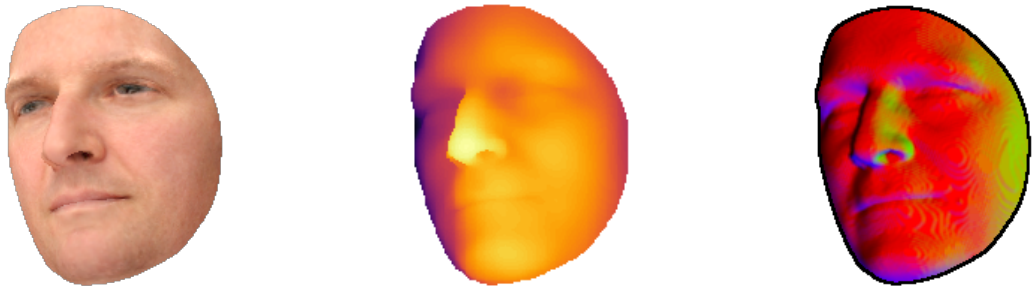
Figure 3.11: Overview of our RGB-D image inpainting architecture with residual fusion. The residual fusion module contains either four or six residual units. In case of single-path residual fusion, the module contains single-path residual units. In case of multi-path residual fusion, the module contains multi-path residual units.

nection with the known region of the color image, the corresponding depth image often does not. In this section, we focus on the definition of an additional representation of the depth channel which could contribute to the quality of reproduction of depth values.

In the context of depth reproduction, the framework currently focuses on the pixel-wise reconstruction of the absolute values in the depth channel of the incomplete RGB-D image. During the training process, the network is taught to reconstruct the depth channel in a pixel-wise manner through the L1 reconstruction loss, as described in Section 3.1. Notably, the depth modality is solely represented by this depth image, which contains the absolute distance from the face to the sensor on pixel level. To this end, the network does not have a mechanism to attend to the reproduction of the depth values as a surface. The continuity of the surface inferred by the network is therefore largely left unconsidered. In addressing this problem, we draw inspiration from several works [45, 46, 59–61] that utilize surface normal representations for the generation and inpainting of depth images. We address this by making several additions throughout the network that provide an improved depth value continuity along the surface of the represented faces. In this section, we discuss these changes, which involve an alternative depth representation, with a corresponding surface normal loss, a surface contextual attention module, and a surface normal discriminator.

3.4.1. Surface normal representation

Zhang and Funkhouser [45] introduced a method to fill in the missing regions in the depth channel of otherwise complete RGB-D images. The paper's approach describes the training of a network to predict local properties of the visible surface at each pixel, which in turn are used to resolve the absolute depth values. These local properties consist of surface normals and occlusion boundaries. The evaluation of this method showed that these intermediate representations contribute to improved performance in the tasks of depth inpainting. Inspired by this work, Matias et al. [46] used estimated surface normals for the purpose of object removal in depth images. The authors based their surface estimation method on the work of Nakagawa et al. [61], which describes the estimation of surface normals based on depth image gradients. The concept of this method is to estimate the normal vectors of each pixel by analyzing their



RGB Depth Surface normal (SN) image

Figure 3.12: Sample with separate visualizations of RGB channels, depth channel and corresponding estimated surface normal (SN) image.

neighborhood depth values, which in turn can be defined through several convolutional operations. We use this relative depth representation in our framework to encourage the accurate representation of the surface represented in RGB-D images.

To obtain the surface normal image, we calculate the depth image gradient for each pixel $p_{i,j}$, in the x and y direction. The direction of the gradient indicates the x and y components of the normal vector, whereas the magnitude of the gradient estimates the z component of the normal vector. Based on this, orthogonal vectors $v_{\vec{\Delta}i}$ and $v_{\vec{\Delta}j}$ are defined for the x-direction and y-direction respectively. To arrive at the surface normal image, we take the cross product between these two orthogonal vectors for each pixel $p_{i,j}$. In turn, the resulting x, y and z values can be visualized in the form of an RGB image. See Equation 3.4 for the corresponding equation as it was outlined by Matias et al. [46] and Nakagawa et al. [61]. An example of a created surface normal image based on a sample from our dataset can be seen in Figure 3.12.

$$
\begin{aligned}
p_{\Delta i} &= \frac{p_{i+1,j} - p_{i-1,j}}{2} \\
p_{\Delta j} &= \frac{p_{i,j+1} - p_{i,j-1}}{2} \\
v_{\vec{\Delta}i} &= (1.0, 0.0, p_{\Delta i}) \\
v_{\vec{\Delta}j} &= (0.0, 1.0, p_{\Delta j}) \\
\vec{v} &= v_{\vec{\Delta}i} \times v_{\vec{\Delta}j} \\
\vec{n} &= \frac{\vec{v}}{||\vec{v}||}
\end{aligned}
\tag{3.4}
$$

Based on this alternative depth representation, we make several additions to our architecture. These will be elaborated on in the following subsections.

3.4.2. Surface normal loss

As mentioned throughout this chapter, we are faced with a unique challenge that consists of the joint completion of two spatially-aligned images of differing modalities. The objective function of the base framework consists of the SN-PatchGAN loss and L1 reconstruction loss. These

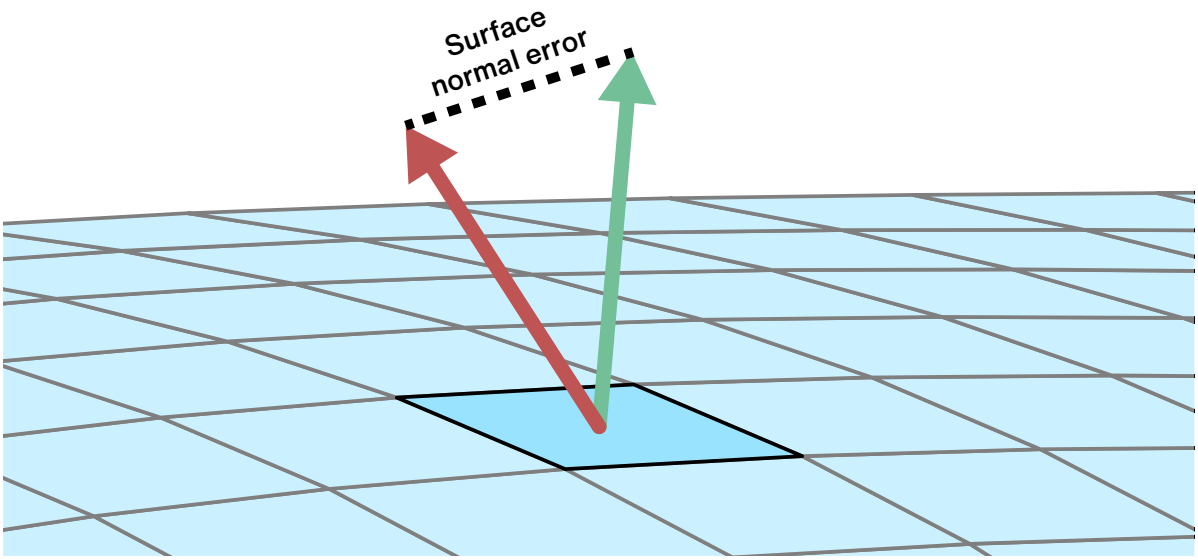


Figure 3.13: Visualization of Surface normal error [46] between a pixel's normal vector as generated (red) and its ground truth (green). The concept of this image was reproduced from Matias et al. [46].

two loss functions are jointly used: the SN-PatchGAN loss focuses on the channel-wise reproduction of the images in a different location, whereas the L1 reconstruction loss measures the pixel-wise reconstruction loss.

In this section, we will discuss an additional loss function that encourages the reproduction of surfaces contained in the depth channel. This loss is based on surface normal image estimation and was introduced as the vectorial loss function by Matias et al. [46]. The depth image gradients used for surface normal estimation can be obtained with a collection of convolutional operations.

The loss function bears high similarity with the L1 reconstruction loss from the base framework, and its concept is straightforward. During training, we calculate the surface normal image of both the inpainted and the corresponding ground truth image. To obtain surface normal loss $\mathcal{L}_{\text{SN}}$, we calculate the L1 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 loss value.

Based on this additional loss function, we hypothesize that the network is stimulated to reproduce the represented surfaces accurately, without affecting the previously mentioned loss functions that are already used by the network.

3.4.3. Contextual surface attention

Aside from the addition of the surface normal loss function, we make an additional change to the base framework based on the work of Matias et al. [46]. The base framework discussed in 3.1 contains a dedicated branch in the refinement stage of the architecture to capture the long-rage spatial dependencies within the image. This branch contains a contextual attention module [40], which enables the extraction of information originating from any spatial location in the image. This module starts with the extraction of 3×3 patches from the known region (background) as well as the unknown region (foreground) of the image. Background patches

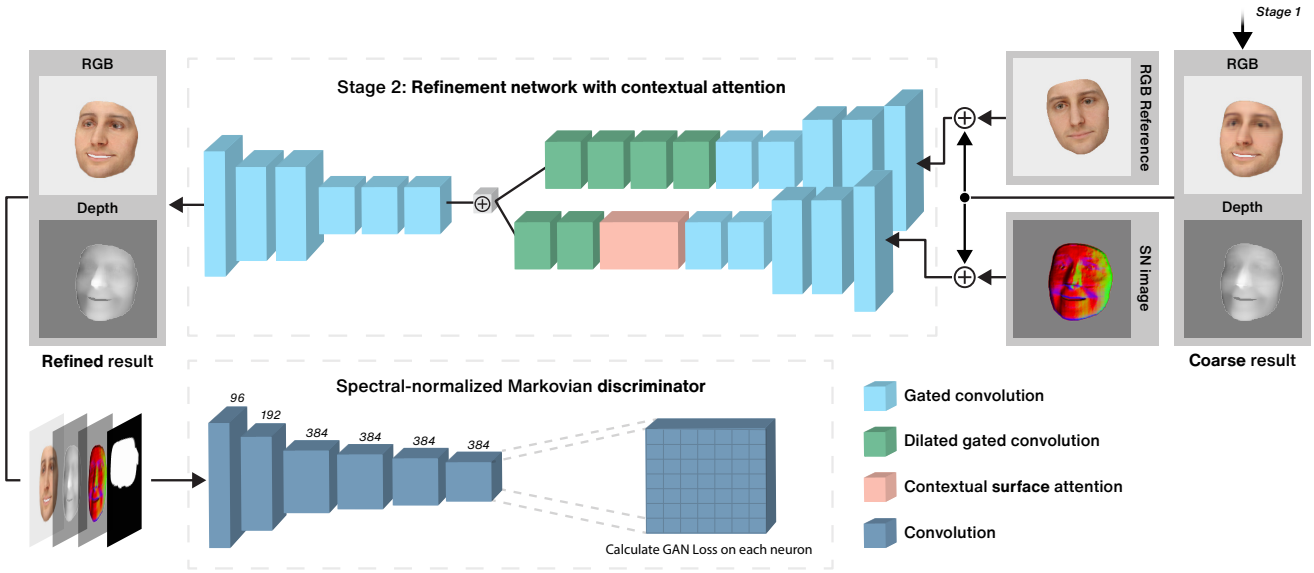


Figure 3.14: Overview of our RGB-D image inpainting architecture with contextual surface attention and the surface normal discriminator.

$b_{x',y'}$ and foreground patches $f_{x,y}$ are matched based on their cosine similarity. As shown in Figure 3.6, we currently pass the RGB-D images to the contextual attention of the network, which looks for similar patches based on similarities across both the depth and color modality that these four channels represent.

Without any modification to the contextual attention module of the base framework, the module considers absolute depth values in its matching procedure. What it does not consider is the relative difference between the pixels, which together form a surface. To improve this, we pass a surface normal image to the contextual attention branch in addition to the RGB-D channels as proposed by Matias et al. [46]. As such, this addition does not make any procedural changes to the contextual attention module itself. We hypothesize that this will improve the contextual attention module's ability to identify patches that are of similar geometric structure. Based on the success of this approach by Matias et al. [46], in which case this was applied to object removal from depth images, we expect this modification to produce inpainting results of higher quality.

3.4.4. Surface normal discriminator

With the introduction of the surface normal loss and contextual surface attention module presented in the previous sections, we stimulate the network to construct a model that not only attends the raw depth values of the depth image, but also considers the surface that it collectively forms. To complete the circle of the usage of estimated surface models in our framework, we introduce a modification to the existing SN-PatchGAN discriminator.

Referring back to Section 3.1, we note that the SN-PatchGAN discriminator actively attends the semantics of the inpainted results. Taking an image with a corresponding mask indicating the missing region as its input, this discriminator calculates a three-dimensional feature map which represents different spatial locations and image channels. In turn, the SN-PatchGAN

loss is calculated for each element of this feature map.

We adapt this process by additionally feeding the estimated three-channel surface normal image to the discriminator. By concatenating the input image with its estimated surface normal image, we give the discriminator direct insight into the complete surface of the respective RGB-D image. We increase the feature depth of the discriminator by 50% to accommodate features that are introduced by the additional input channels, as shown in Figure 3.14. While this intuitive addition could also negatively affect the existing performance of the SN-PatchGAN, we theorize that this change could have a significant impact on the local and global consistency of the inpainted result of the depth image.

4

Results

In this chapter, we qualitatively and quantitatively evaluate our framework and its components, as they were proposed in Chapter 3. We consider the choice of the right evaluation metric for drawing sound conclusions to be equally impactful to the design of a method for a given application [82]. With this in mind, we carefully choose the strategy and metrics for the evaluation of our joint RGB-D image inpainting framework in view of its intended target task of HMD removal.

Research Objective 1 *Define an architecture that is capable of virtually removing the HMD from the wearer's face in RGB-D images.*

We recall our main research objective as introduced in Chapter 1 above. As discussed in Chapter 2, to the best of our knowledge, a method that targets the joint inpainting of RGB-D face images does not currently exist. Throughout this thesis, we have taken an exploratory approach to fulfilling our primary research objective. In Chapter 3, our exploration was guided by a number of sub-objectives. Similarly, we structure our evaluation based on these individual research sub-objectives. As such, in this chapter, we evaluate each solution proposed in Chapter 3, qualitatively and quantitatively, with respect to its objective. We recall these research sub-objectives as they were defined in Chapter 1:

Research Objective 1.1 *Define a module and loss function that stimulates the preservation of the identity features of the wearer's face.*

Research Objective 1.2 *Define an architecture that is capable of handling the multimodal characteristics of RGB-D images.*

Research Objective 1.3 *Define an architecture that stimulates the creation of smooth geometric surfaces.*

As mentioned in Section 2.3.1, a quantitative evaluation metric that properly evaluates the quality of generated images does not exist. This fact has motivated our choice of quantitative evaluation metrics, as well as an elaborate visual examination to provide qualitative insights.

At the end of this chapter, we aim to have provided extensive insights on which approaches satisfy our research objectives, and which do not. Moreover, based on the gathered findings, we determine which framework configuration forms the best solution for our main research objective, which involves HMD removal through joint RGB-D image inpainting. We train and evaluate the model configurations with our synthesized dataset, in accordance with the objective:

Research Objective 2 *In absence of a large-scale RGB-D face dataset, create a suitable dataset that is sufficiently sized.*

We start this chapter by introducing the dataset we use for the training and inference procedure of all network configurations in Section 4.1. This is followed by an outline of the implementation details of the framework in Section 4.2. Section 4.3 presents the qualitative results and a corresponding analysis through visual examination. Finally, Section 4.4 provides an outline of the used objective metrics, followed by the results and a corresponding analysis.

4.1. Dataset

In this section, in accordance with Objective 2, we introduce the dataset used in our work and explain the process of its creation. During the adversarial training process of a GAN, the generator G implicitly learns the data distribution p_x through the training data x observed by the discriminator D. Relying on training data x, the learning process of a GAN thrives from a sufficiently large and diverse training set to uniquely identify its true distribution [161]. Bearing in mind the wide range of facial shapes and appearances in the world, we require a dataset that contains a large number of unique face images with high variability in appearance among them.

While datasets of RGB images of faces are widely available [52, 53], the same cannot be said regarding the availability of RGB-D image datasets. In Section 2.6, we reviewed a representative set of RGB-D face image datasets that are publicly available. Having evaluated the number of identities and images present in these datasets, we concluded that the sizes of RGB-D datasets are insufficient to train a GAN at this time.

Based on the absence of a suitable dataset, we faced the challenge of obtaining a satisfactory amount of data. As it is costly and time-consuming to gather the RGB-D captures of thousands of people, we adopted a different strategy. We created an image synthesis method based on the 3DMM parametric space [63] of the Basel Face Model [62, 162]. This model consists of a statistical shape, texture and expression model, enabling the generation of 3D meshes of faces with high variability in shape, color, and expression. The shape and texture spaces of this model have been defined based on face scans of 100 female and 100 male subjects, the majority of European origin. While this approach allows us to generate a large amount of faces with unique identities, it introduces a number of limitations with respect to model bias and generalization, which is discussed in Chapter 5.

We use our defined synthesization pipeline to create our dataset. Shown in Figure 4.1, the pipeline starts by taking a random sample across the independent shape, texture, and expression parameters. Following this, relative to the resulting mesh, we place a predefined mesh of an HMD representing the true dimensions of an Oculus Rift [163]. We align the HMD by

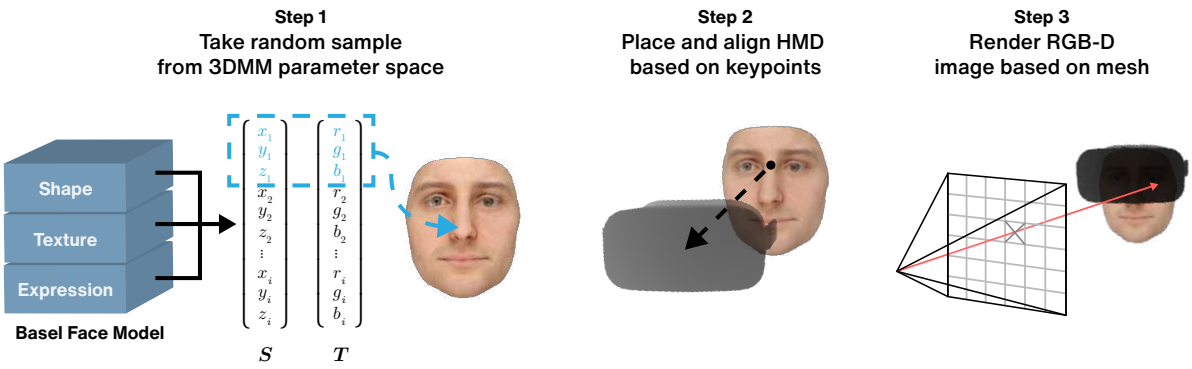


Figure 4.1: Steps of our data synthesization pipeline. S and T denote the resulting shape and texture vector respectively, of which each element is in spatial correspondence.

offsetting the position of the face based on predefined keypoints on the face and the HMD. These keypoints mark the locations of the eyes and the lenses of the HMD. We deem it most practical to use the HMD as a center point for alignment in both synthetic and real-world applications, as it is easiest to track the position of the HMD because of its consistent appearance and the potential use of fiducial markers [164]. Moreover, data of sensors such as gyroscopes, accelerometers, and magnetometers, that some HMDs provide can also be useful for this purpose.

Next, the compound mesh is placed in world space, at distance cd from the camera position. Unless otherwise specified, the value we use for cd is 850 millimeters. As such, the face wearing the HMD is located 850 millimeters, or 85 centimeters, from the sensor. While the exact distance typically depends on the hardware setup and position of the users, this defined distance forms a realistic estimate.

A benefit of data synthesization is that we are in full control of all conditions regarding image rendering. From the camera's perspective, the resulting mesh is rendered based on a set of transformations: pose p with properties p_{pitch}, p_{yaw} and p_{roll}, ambient illumination a with property $a_{intensity}$ and directional illumination d with property $d_{position}$ and intensity $d_{intensity}$. We refer to Figure 4.2 for examples of these transformations.

Finally, a simple ray tracing algorithm is responsible for rendering the RGB color image as well as a corresponding depth image. Effectively, we render three images of size 256×256 pixels: a color image of the face without HMD, depth image of the face without HMD, and binary image of solely the HMD, which indicates the occluded region as caused by the HMD.

We jointly encode the color image and depth image of the face in the four channels of a single file. We opt for a depth resolution of 1 mm per pixel, for which our reason is two-fold. Firstly, the approximate resolution of the Microsoft Kinect is 1.3mm per pixel [165], which is 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 an 8-bit PNG file.

We manage to achieve the latter by subtracting the scalar cd from the inverted depth image, which accordingly enables the depth values to fit in the 8-bit alpha channel. This is a consequence of the fact that faces will never realistically have a depth that exceeds 256 millimeters. Besides the compactness of this representation, the alpha channel serves as a simple way of

Figure 4.2: Examples of individual transformations, from left to right: frontal with full ambient illumination, pose, directional light, ambient light, and expression.

visualizing the depth values of the image. To recover the original values, one can simply invert the depth channel and add scalar cd. We have made the Python/C++ implementation of the RGB-D rendering segment of this pipeline publicly available as *mesh2rgbd*[1], which is based on the *face3d*[2] package by Yao Feng.

Using the defined pipeline, we can generate any number of face identities of highly variable appearance and expression as well as their corresponding RGB-D images. It should be noted that identities generated through this method form a relatively small subset of the facial appearances that are present in the world. Moreover, our synthesization technique generates perfect data, whereas current depth sensors may contain a significant amount of noise. Nonetheless, at this moment, this pipeline provides the best way of generating a high number of faces for RGB-D image synthesis. We elaborate on the limitations of our complete framework in Chapter 5.

We construct our dataset using the pipeline described above. This dataset consists of a total of $40\,000$ RGB-D images, each of which contain a face with a unique identity. For the training and evaluation of our framework, we split our dataset in three: $40\,000$, 4000, 4000, into a training set, validation set, and testing set, respectively. For intermediate and precise evaluation of our framework, we have created several versions of this dataset with step-wise or random values for the following properties: random expression, random ambient illumination, random directional illumination, random pose, and combinations among these properties. We refer to our final dataset as *3DMM-RGBD-40k*, which consist of faces with the following random properties and transformations:

- **Random expression**
- **Random pose** p: with p_{pitch} and p_{yaw} in range $[-30°, 30°]$ and p_{roll} in range $[-20°, 20°]$
- **Random ambient illumination** a: with $a_{intensity}$ in range $[80, 110]$

In regard to the range of head poses, we base our decision on the work by Bartlett et al. [166], who measured the head poses of participants in an interview setting and found that the values of the three pose axes lie within the specified ranges above. Moreover, we limit the roll axis to the range of $[-20°, 20°]$, as this range forms the upper bound of roll angles that the

[1] https://github.com/nsalminen/mesh2rgbd
[2] https://github.com/YadiraF/face3d

human head is capable of [167]. Whereas the relative head pose depends on the sensor position in a real-world situation, we assume that the subject is directly facing the RGB-D sensor in our synthesis.

While our pipeline supports RGB-D image rendering with directional illumination, we choose not to include this transformation in our evaluation datasets at this time. The reason for this decision is based on the results of the validation of pretrained face recognition models from Section 3.2.

4.2. Implementation

Our framework has been implemented in Tensorflow [168], based on the source code by Yu et al. [41]. Training of our models has been performed on two NVIDIA GeForce RTX 2080 Ti GPUs. This process typically takes up around 2.5 days, but is continued until convergence of the losses and stabilization of the visual quality of the validation results. For training, we use a learning rate of 0.0001 for both the generator and discriminator. Moreover, through hyperparameter tuning based on a combination of visual examination and objective metrics, we have determined the default hyperparameter balance of 3:1:1:1, for the reconstruction loss, SN-PatchGAN loss, identity loss and surface normal loss respectively.

The models have been trained and evaluated with datasets containing images of size 224×224, which have been resized from their original size of 256×256. We have opted for this change to be able to train with larger batch sizes, while staying within the memory bounds of the GPUs. At inference with a single NVIDIA GeForce RTX 2080 Ti, the data-level fusion models achieve an average frame rate of 48 frames per second, whereas hybrid fusion models achieve an average frame rate of 41 frames per second. While our method currently does not leverage any temporal modality and further research is needed, the performance of the configurations theoretically permit real-time RGB-D video inpainting with a single high-end GPU.

4.3. Qualitative results

This section presents the results and analysis of the proposed methods through visual examination. Whereas color images can be presented in a straightforward way, RGB-D images require additional representations to reflect their 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.

As the identity loss function is integral to further experiments, we start with the visual examination of results generated by our proposed method for preservation of identity (Objective 1.1). This is followed by the visual examination and comparison of results generated by different fusion methods (Objective 1.2). Lastly, we visually examine the results of several additions to the network focused on surface reproduction (Objective 1.3).

4.3.1. Identity preservation

In this section, we study the effects of the identity loss function, which aims to fulfill Objective 1.1. The identity loss stimulates the extraction and application of facial identity features extracted from the RGB channels of the given reference image. In this section we study the

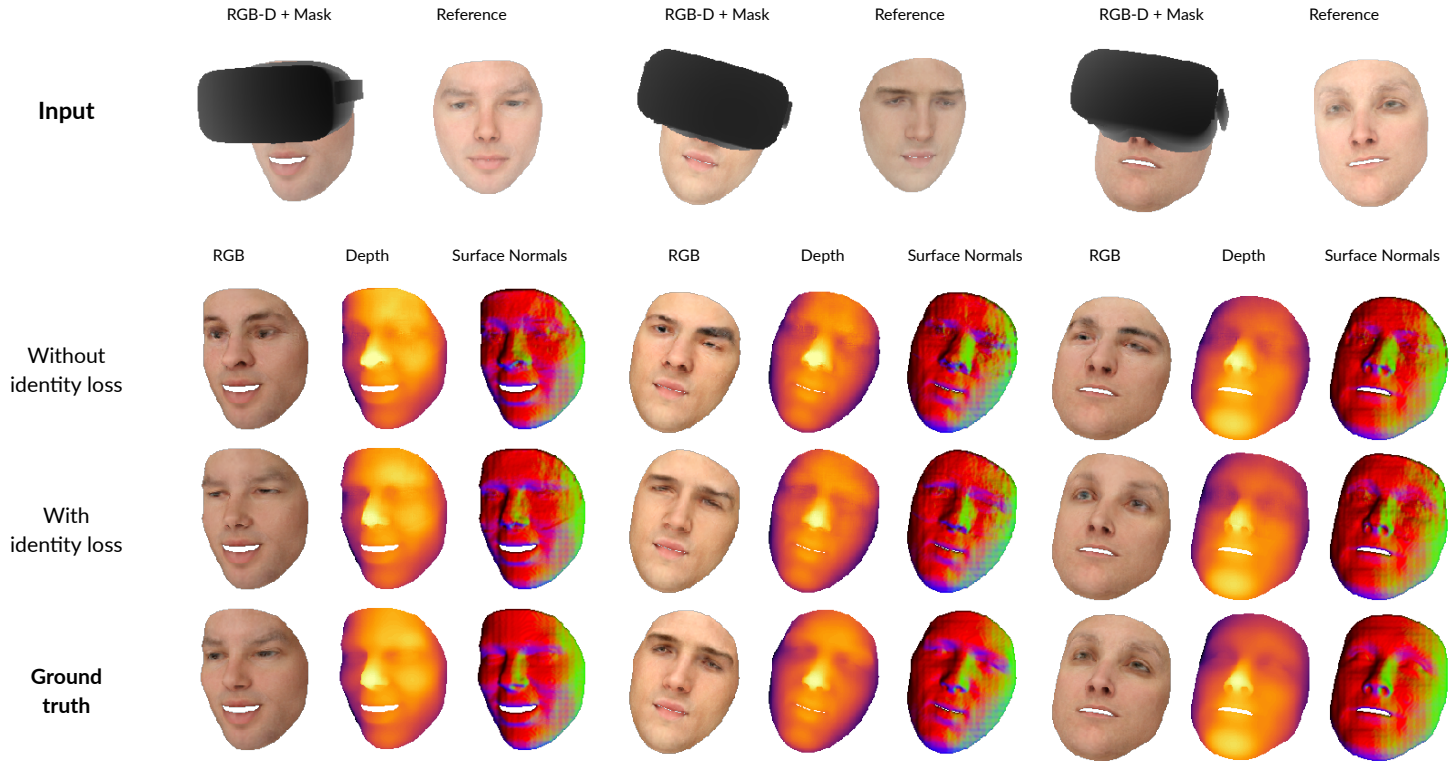


Figure 4.3: Comparative overview of the results of the model trained *without* the identity loss function and the model trained *with* the identity loss function.

effects of the employment of this loss function. We compare two different methods: data-level fusion *without* identity loss and data-level fusion *with* identity loss.

When considering the results of the model trained *without* identity loss, we observe that this method produces a globally consistent synthesization of the masked region with respect to the skin color and overall shape of the input. However, it is apparent that facial identity features do not correspond with the ground truth image. This is not surprising, as this model has no contextual knowledge of the subject's identity. Accordingly, the model provides its best estimation of the facial features of the masked region, which rarely corresponds to the subject's true identity.

Turning now to the results of the model trained *with* identity loss, we observe that the face identity represented in the resulting images bears high similarity with that of the given reference image. In particular, we note that the general appearance of identifying facial features such as the nose, eyes and eyebrows of the subject are properly represented in the RGB channels of the inpainted image. However, we identify several cases in which the eye color is not preserved or symmetric. Moreover, some inpainted results contain visual artifacts, which are most frequently seen around the eye region of the subject. We observe that this effect becomes increasingly prevalent with greater pose angles of the faces in the input image and the reference image. This suggests that extreme pose angles are detrimental to the performance of the identity loss function. We elaborate on this common error in Section 5.1.1.

The effects on the depth channel of the inpainted images requires closer inspection, as its interpretation is not as intuitively familiar. Considering the identity loss solely considers the

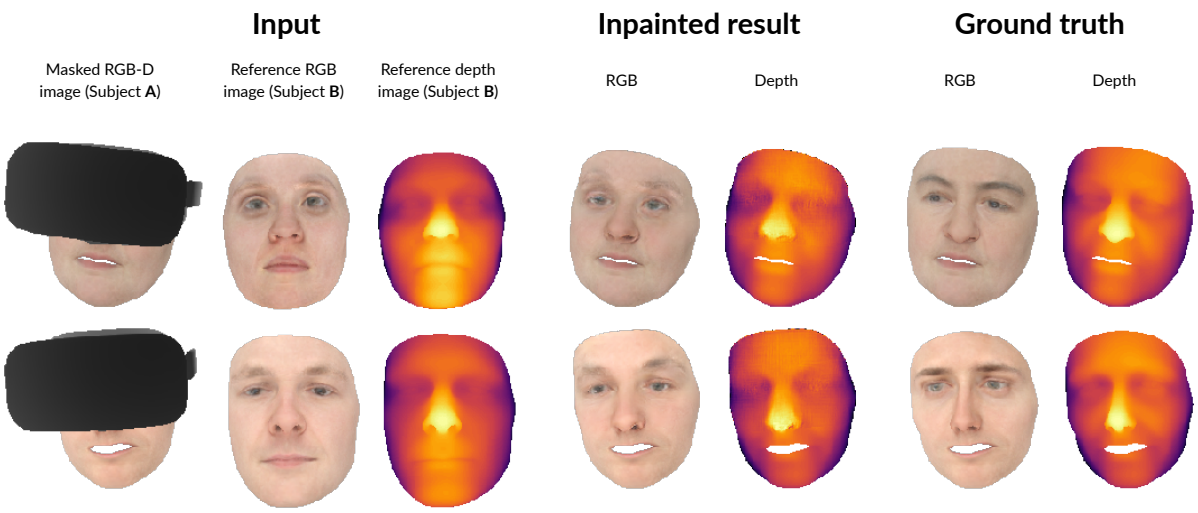


Figure 4.4: Results of model input containing differing identities, generated by a model trained *with* our identity loss function.

RGB channels, it is especially interesting to study the indirect effects of this loss function with respect to the inpainted depth channel. Upon close inspection, we note that the model appears to preserve the overall visual correlation between the RGB channels and depth channel. As such, the facial appearance represented in the RGB channels appears to correspond to the geometric face structure represented in the depth channel.

Perhaps the most insightful demonstration of the effects of the identity loss function is performed by feeding the model with a masked version of one identity and a reference photo of another. Figure 4.4 presents two examples of this kind, where the input consists of a masked RGB-D image and reference image, each of which belong to two different subjects. 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. Similar to the aforementioned observations, we find that visual facial features such as the eyes, nose and eyebrows are correctly preserved in the inpainted image. Further examination of the depth image reveals that the RGB channels and depth channel appear highly correlated. Interestingly, a number of geometric facial features of the reference image appear to have been preserved in the inpainted image. For example, looking at the first example in Figure 4.4, we identify the distinct reproduction of the shape of the reference subject's nose. This is an interesting finding, as the identity loss uses only the RGB channels of the input and reference image for the calculation of the identity loss used during training.

4.3.2. Fusion of color and depth information

In this section, we address the visual differences of the results of the fusion types as described in Section 3.3, aimed at Objective 1.2. A separately trained RGB image and depth image inpainting model [41] will form a comparative baseline for the fusion methods we evaluate in this section. The output of these unimodal models will demonstrate the level of visual quality we aim to match with a jointly trained model.

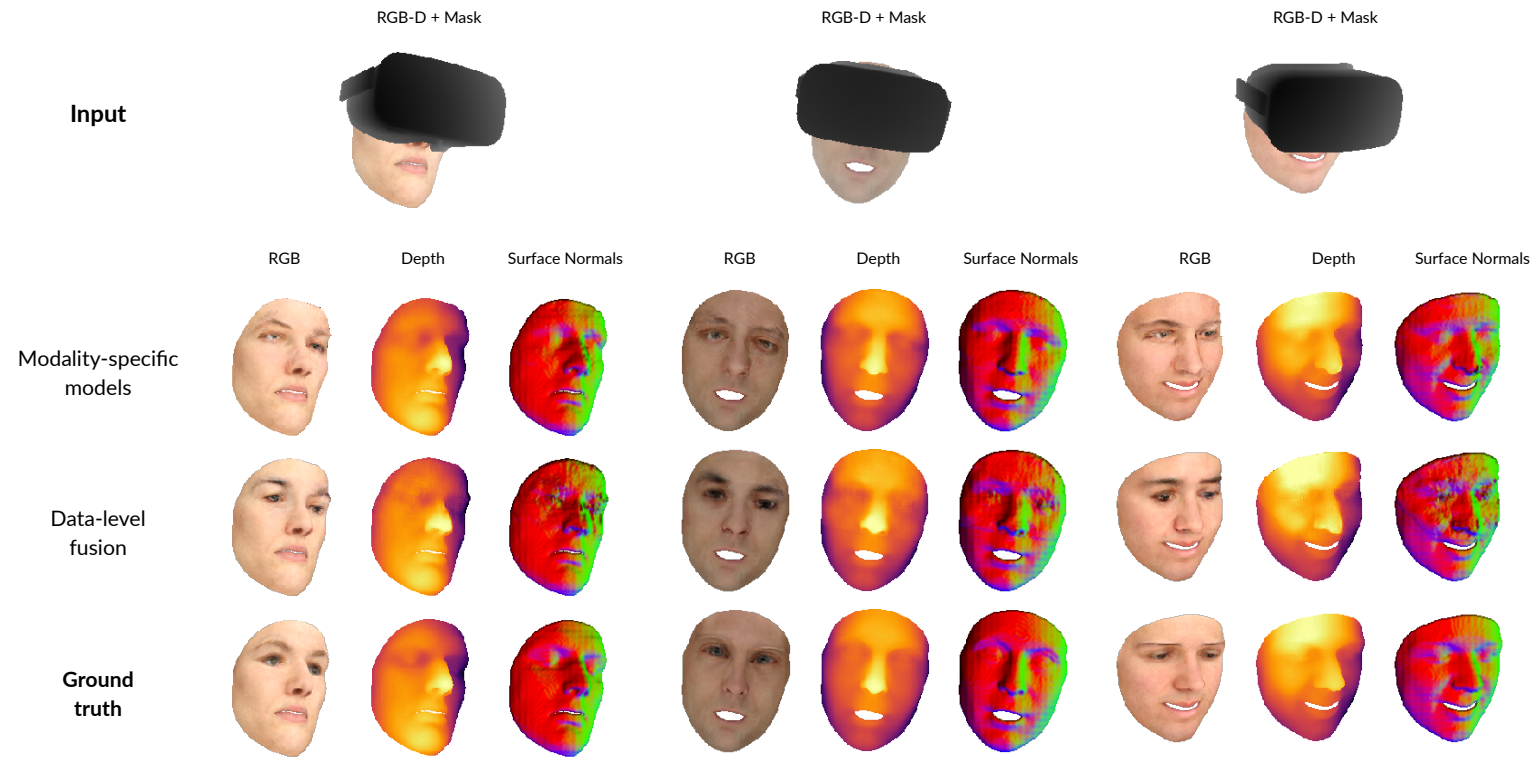


Figure 4.5: Comparative overview of the results of the modality-specific image inpainting models, trained separately on the color and depth channels, and the data-level fusion model.

We will first inspect the results of data-level fusion, which concatenates the RGB and depth image at the network input. Following this, we will examine the results of hybrid fusion, where the following fusion methods are applied in the coarse stage of the network: fusion through addition, single-path residual fusion and multi-path residual fusion.

Data-level fusion

In the same vein as the early evaluation presented in Section 3.3.1, we compare the output of two separately trained modality-specific inpainting models with our data-level fusion model. The goal of this experiment is not to evaluate whether the inpainted faces look exactly like their ground truth, but to evaluate the visual quality of the inpainted results. Moreover, this experiment could provide insights into how complementary information among the modalities contributes to the visual quality of the results of the jointly trained data-level fusion model.

Expanding on our earlier evaluation, we present several additional samples for this comparison in Figure 4.5. Please note that the models used in this comparison do *not* employ our proposed identity loss function, as it is not possible to use this loss function in a modality-specific depth inpainting framework as is.

We firstly evaluate the results in regard to the inpainted RGB channels. The modality-specific model for RGB inpainting produces a result that is consistent with the known region of the image, with respect to both color and shape. However, it is apparent from these results that the modality-specific RGB model does not produce facial features that appear natural or coherent. The data-level fusion model makes a reasonable improvement in this regard, which could be related to information that is available in the respective depth image. Specifically, the

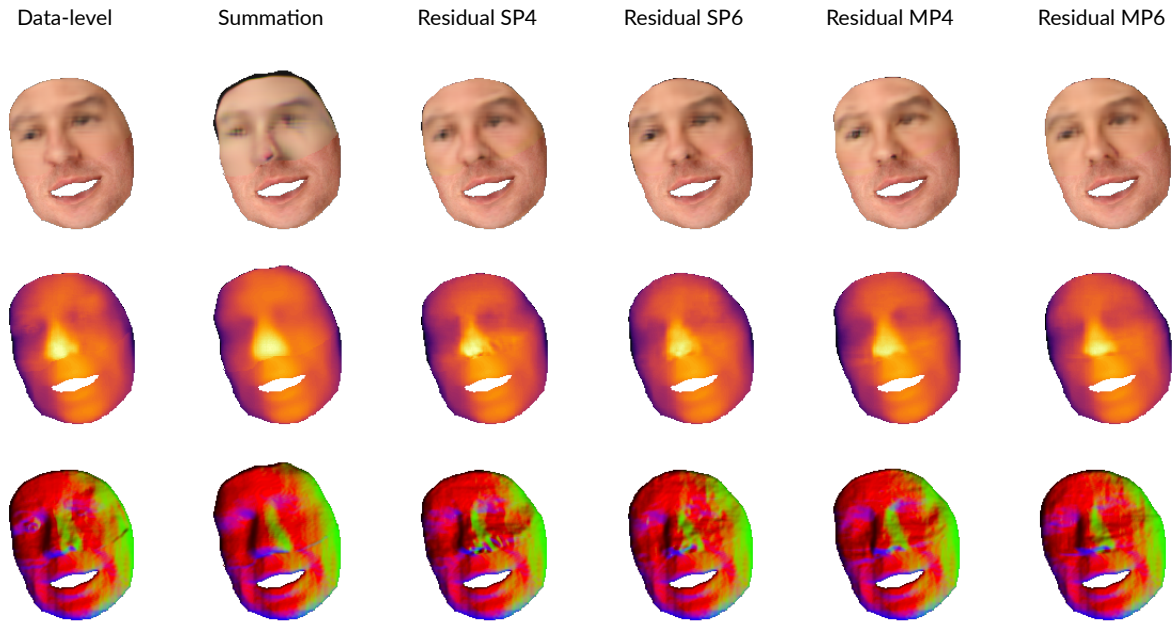


Figure 4.6: Comparative overview of the *coarse* inpainting results of the proposed hybrid fusion types.

data-level fusion model may be able to construct more meaningful features based on the joint color and depth information, which could enable perception of the pose of the face.

As far as the depth images of the presented results are concerned, we note a large difference between the modality-specific depth image inpainting model and the data-level fusion model. This is even more apparent in the irregular and noisy surface normal image representations of the depth information.

Having examined the RGB and depth channels individually, we briefly touch upon the consistency between the inpainted RGB and depth channels. Indisputably, we cannot expect two independently trained models to produce outputs that are consistent among each other. However, as was seen in the previous section, we note a strong consistency between the color and depth images produced by the data-level fusion model. This comes as no surprise, as the full feature construction process in this model is done jointly.

Hybrid fusion

Hybrid fusion refers to a combination of feature-level fusion in the first stage and early fusion in the second stage of the network. The first stage is trained with a reconstruction loss and identity loss and provides a coarse prediction of the masked region. In turn, this prediction is refined by the second stage of the network, which is additionally trained with the SN-PatchGAN loss. Accordingly, we may assume that the visual quality of the refined result directly depends on the visual quality of the coarse result. Therefore, to inspect the effects of different feature-level fusion methods applied in hybrid fusion, we start with the visual examination of the coarse inpainted results of the first stage.

While we have experimented with fusion through summation at several positions throughout the coarse stage of our network as discussed in 3.3.2, the results consistently were of low quality and appeared similar. For this reason, we qualitatively evaluate just one of these

models and do not further include this approach in our evaluation. For residual fusion, we denote single-path and multi-path by *SP* and *MP*, respectively, followed by the number of residual blocks used in the coarse stage (e.g., Residual-*SP4*).

A set of coarse results for each fusion method is shown in Figure 4.6. At first glance, it is immediately clear that feature-level fusion through summation does not provide the results we are looking for. All channels appear blurry and the RGB channels in particular contain a significant amount of noise and lack color coherency with the known image region. The estimated surface normal image of this approach appears very smooth due to the blurriness of the depth channel, but contains little detail and does not seamlessly connect with the known image region.

The coarse output of the feature-level residual fusion look similar, with no dramatic improvement with respect to data-level fusion. However, the connection between the known and missing image region appears more seamless compared to data-level fusion and fusion through summation.

As the visual examination of coarse inpainted results provided limited insights, we continue our evaluation by turning to Figure 4.7, where we show a collection of refined results of the models. As mentioned, we will exclude fusion through summation, as its coarse results indicate the relative failure of this method. Consequently, with respect to hybrid fusion, we continue our evaluation with solely single-path and multi-path residual fusion.

In general, all fusion methods in Figure 4.7 show a good reproduction of the subject's face, as facial features are well-represented in both the color and depth channels. However, in similar correspondence, the inpainted regions contain a reasonable amount of noise and are sometimes not fully connected to the known region of the image. This is particularly noticeable in the depth channel and its surface normal representation.

The RGB channels of each of the fusion methods appear to be of similar quality, where we observe the occasional sign of minor noise. Moreover, as observed during evaluation of our identity loss, asymmetry of eye shape and color is commonly found. The depth image and its surface normal representation show similar results, with no significant differences that can be identified across all image samples. In fact, we observe mixed results of residual fusion methods, which perform slightly better than data-level fusion in the leftmost sample in Figure 4.7 but perform on par or worse in the rightmost sample.

We cannot make a well-founded conclusion on the performance of data-level fusion versus hybrid fusion. While these findings are somewhat disappointing, visual evidence exists to support that residual fusion forms a good candidate for multimodal feature fusion. As such, we refer to the quantitative evaluation of multimodal feature fusion in Section 4.4 for further analysis.

4.3.3. Reproduction of surfaces

Despite the efforts that were evaluated earlier in this chapter, the presented results still contain a striking amount of noise, especially in the depth images and their corresponding surface normal images. Furthermore, we observe strong inconsistencies in the connection between the unknown and known region of the image, which is most prominent in the case of large face pose angles.

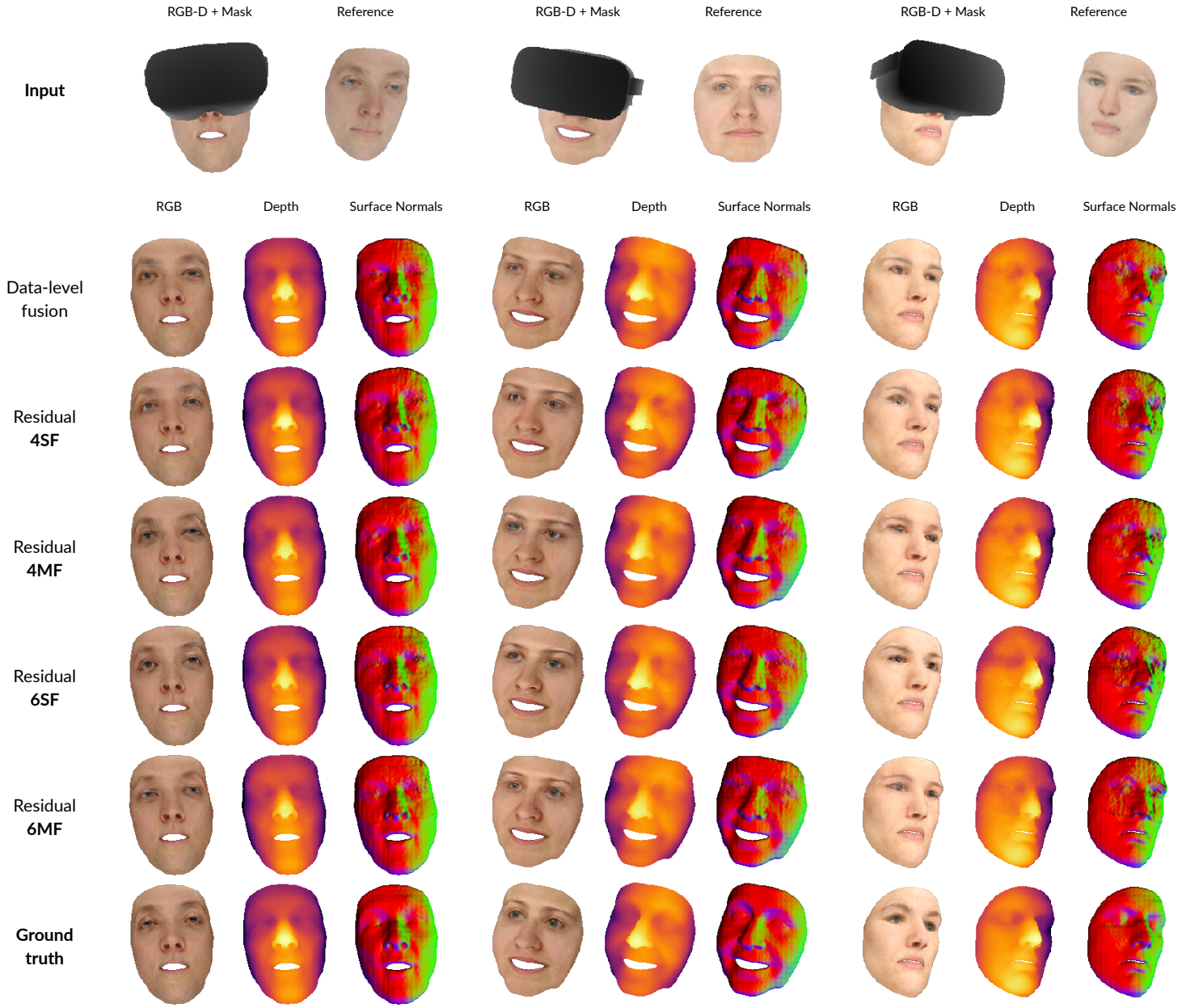


Figure 4.7: Comparative overview of the refined inpainting results of the proposed hybrid fusion types.

Described in Section 3.4, a number of additions have been described that set out to improve the framework's depth channel reconstruction by improving the quality and consistency of the formed surface. In this section, we evaluate their results. The models used for this evaluation are based on our data-level fusion network, trained with a reconstruction loss, SN-PatchGAN loss and identity loss. To avoid confusion and to save space, we denote this network configuration as (M).

Surface normal loss function We first assess the effects of the surface normal loss $\mathcal{L}_{\text{SN}}$ on the visual quality of the results. Upon consideration of the results of the model trained with the surface normal loss in Figure 4.8, it stands out that the inpainted regions of the depth images and surface normal images show a decreased amount of noise compared to aforementioned

Figure 4.8: Comparative overview of the inpainting results of the proposed surface normal loss function, contextual surface attention (CSA) module, and surface normal discriminator (SND). (M) denotes our data-level fusion network, trained with a reconstruction loss, SN-PatchGAN loss and identity loss.

methods. Consequently, the inpainted region appears smoother and facial features and their details are distinctly visible. Furthermore, this improvement appears not to take away from the visual quality of the RGB channels of the image.

However, it is interesting to note that the addition of the surface normal loss does not always fully remove the previously noted inconsistent connection of the known and unknown region of the depth image. An instance of this is most visible in the rightmost sample in Figure 4.8, where the split between the unknown and known image region is clearly visible.

After taking a closer look at the RGB channels of the results, we make a number of observations. Firstly, the symmetry of the facial features seems to have worsened ever so slightly. Moreover, in the middle sample in Figure 4.8, we spot a fair amount of noise around the upper left region of the face, and even detect a hole in the right eyebrow.

Contextual surface attention We now turn to the effects of the contextual surface attention (CSA) module. The contextual attention module, introduced by Yu et al. [40], is built on the concept of finding and reusing known patches that are statistically similar to the coarse prediction of the missing region. We hypothesized that the addition of surface attention aids in this search, as outlined in Matias et al. [46].

Interestingly, comparing the results of our model with and without the CSA module reveals that the addition of surface attention does not have an evident effect on visual quality. The visual quality of the inpainted RGB channels appears consistent with the aforementioned models, including the middle sample of Figure 4.8, which proves to be a challenging instance. However, we do observe a moderate improvement of the connection between the unknown and known region of the depth image, as can be seen in Figure 4.8. Despite this observation, a clear benefit of the usage of this module could not be determined based on our visual assessment.

Surface normal discriminator What follows is an evaluation of the results of the model that was trained with the surface normal discriminator (SND), as shown in Figure 4.8. It is apparent that the RGB channels are largely comparable to the model that was trained without the surface normal discriminator.

Turning now to the depth images and their surface normal representations, we observe similar inpainted results with minor improvements. Upon close examination of these images, we no longer find inconsistent connections between the unknown and known region of the image. However, aside from this observation, the results can be closely compared to the model without the surface discriminator.

4.4. Quantitative results

The quantitative evaluation of generative models remains a challenging task. Despite this difficulty, we define a set of objective metrics that align with the task at hand and provide a strong base to draw conclusions from. In this section, we start by outlining our objective metrics, followed by their results.

4.4.1. Objective metrics

In this section, we outline the metrics we use for the evaluation of our method. These metrics give us an empirical base to compare the defined network configurations.

As discussed in Section 2.3.1, a consensus on a standardized set of objective metrics for the evaluation of GANs does not currently exist. Similarly, this is the case for the evaluation of image inpainting tasks in general. The latter is a consequence of the fact that image inpainting methods do not necessarily seek to create a pixel-perfect reconstruction of the missing area. Rather, their objective is to reconstruct the missing area in a plausible and realistic way, while blending in with the known region of the image. To further break down our requirements, we define a collection of desired characteristics. Specifically, we require our set of metrics to:

- align with the evaluation of existing image inpainting methods;
- provide a clear and reliable ground for derivation of conclusions;

- agree with human perceptual judgement [84];
- reflect the inherent challenges with respect to our application [82, 85], specifically, identity preservation and the depth reconstruction.

Considering image inpainting is a variant of a conditional image generation task, we possess the ground truth pixel values for the synthesized missing region. This permits us to use full reference quality metrics to measure the quality of the inpainted images. Moreover, aside from visual fidelity, these metrics assist in recognizing mode collapse. Therefore, following existing image inpainting methods [40, 41, 91], we report the following metrics: 1) mean L1 error 2) mean L2 error 3) Peak Signal-to-Noise Ratio (PSNR) 4) Structural Similarity (SSIM) index [55]. These metrics assess the visual quality of the inpainted images by comparing the information that is represented in the inpainted image with the ground truth image. As subjective evaluation has suggested that the quality of depth images is correlated to the fidelity of corresponding 3D models [169], this will provide us with insights regarding the application of our method.

Additionally, we employ the Visual Information Fidelity (VIF) index [56]. This metric aims to represent the difference of human-perceivable information between a test image and a reference image based on statistical properties that approximate the human visual system. Notably, there is evidence to indicate that the VIF index of depth images is highly correlated with the quality of experience of 3D video compared to other quality metrics Banitalebi-Dehkordi et al. [169]. For this reason, this metric is a particularly insightful addition to our set of evaluation metrics.

In order to quantify and compare the degree of identity preservation, we define a metric based on the FaceNet face identity embedding model [57] pretrained on the MS-Celeb-1M [54] dataset. The usage of this model has been validated in Section 3.2. This metric will be calculated on the RGB channels exclusively, as a reliable RGB-D face representation model is currently not readily available.

Mean L1 and mean L2 error

The mean L1 error, also known as the mean absolute error, measures the absolute difference between the estimated values and target values. For predicted image $\hat{x}$ and ground truth reference image x, where N denotes the size of the mask, the L1 error is calculated as follows:

$$L1(x, \hat{x}) = MAE(x, \hat{x}) = \frac{1}{N} \sum_{i=1}^{N} |x_i - \hat{x}_i| \tag{4.1}$$

The mean L2 error, also known as the mean squared error, measures the absolute difference between the estimated values and target values. Similar to the L1 error, the L2 error is calculated as follows:

$$L2(x, \hat{x}) = \text{MSE}(x, \hat{x}) = \frac{1}{N} \sum_{i=1}^{N} (x_i - \hat{x}_i)^2 \tag{4.2}$$

The L2 error is highly sensitive to outliers when compared to the L1 error. Furthermore, it is important to note that whereas these errors represent the color difference in points (0-255) in the RGB channels, they will reflect the difference in millimeters (mm) in the depth channel.

Peak Signal-to-Noise Ratio (PSNR)

Peak Signal-to-Noise Ratio (PSNR) is a simple image quality metric that measures the difference of two images at pixel level. It expresses the difference between the maximum value of a signal and the maximum value of a distorting noise. In our case, the signal is formed by the ground truth image and the noisy signal is an inpainted version of the same image. The mathematical definition of PSNR, based on the mean squared error (MSE), is as follows:

$$\text{PNSR}(x, \hat{x}) = 10 \cdot \log_{10}(\frac{R^2}{\text{MSE}(x, \hat{x})}) \tag{4.3}$$

Where x and $\hat{x}$ denote the reference image and test image respectively, M and N represent the number of rows and columns of the input signals respectively, and R is the maximum signal value. In the case of our evaluation, R is equal to 255 considering we compare 8-bit unsigned integer data.

Although image quality is a subjective matter [170], this metric remains to be a common image quality metric for its clear mathematical definition. Considering perceived image quality is highly subjective, we will use this evaluation metric to measure the effects of methods in a strictly empirical way, while remaining aware of its limited correlation to perceptual quality [55].

The PSNR metric is represented by a ratio with a range of $[0, 1]$, where $\text{SSIM}(x, x) = 1$.

Structural Similarity (SSIM) Index

The Structural Similarity (SSIM) index [55] measures the structural similarity between a ground truth image and the respective inpainted image. As opposed to PSNR, SSIM does not estimate absolute errors but estimates the perceived change in structural image information. The SSIM algorithm compares local patterns of pixel intensities that have been normalized for luminance and contrast, as the structure of objects in a scene is assumed to be independent of these effects. The algorithm is split in stages for luminance comparison, contrast comparison and structure comparison between the given input images. We refer to the paper by Wang et al. [55] for a detailed description of the procedure.

The SSIM index metric is represented by a ratio with a range of $[0, 1]$, where $SSIM(x, x) = 1$.

Visual Information Fidelity (VIF) Index

The Visual Information Fidelity (VIF) [56] index builds on the concept of the human visual system. This metric assesses the quality of an image by measuring the information represented in the test image and reference image. Measuring the respective image information is done based on natural scene statistics [171] represented by a wavelet-domain Gaussian scale mixture model.

The VIF index metric is represented by a ratio with a range of $[0, 1]$, where $VIF(x, x) = 1$.

Identity error

In order to evaluate the effectiveness of the identity loss function, we define an identity error metric. This metric is based on the same notion as our identity loss function. Firstly, a pretrained model is used to calculate a feature embedding of the subject's facial appearance. In turn, the Euclidean distance between embeddings represents a measure of identity similarity.

To ensure an independent evaluation, we employ a different architecture and different dataset. Specifically, this metric uses the FaceNet [57] architecture, pretrained on the MS-Celeb-1M [54] dataset which produces a 128-byte vector representing the subject's identity. Accordingly, the identity error is calculated as follows:

$$\text{ID}(x_{\text{rgb}}, \hat{x}_{\text{rgb}}) = ||M(x)_{\text{rgb}} - M(\hat{x}_{\text{rgb}})||_2 \tag{4.4}$$

Where x_{rgb} and $\hat{x}_{\text{rgb}}$ represent the RGB channels of the ground truth image and inpainted image respectively, and M denotes the pretrained face identity embedding model.

4.4.2. Results

In this section, we present the results of the quantitative experiments. To provide insights on the quality of results across the color and depth modality, we calculate the metrics the color image represented in the RGB channels and the depth image represented in the depth channel. Moreover, to provide a means of evaluating the surface formed by the depth image, we additionally calculate each metric on the estimated surface normal image. The estimation procedure of the surface normal image is described in Section 3.4.1.

Identity preservation

We start our evaluation of the results by considering preservation of identity, in accordance with Objective 1.1. While we evaluate the explored fusion methods at a later point in this section, we evaluate our identity loss within the data-level fusion version of our network. Comparing the results of these models as presented in Table 4.1, we note a consistent and significant improvement of metric values based on the addition of the identity loss. This result indicates that the identity loss effectively assists the model in inpainting the masked region in such a way that is more consistent with the ground truth image. This is in line with our visual examination in Section 4.3.1. Remarkably, consistent with our visual examination, an improved quality of the depth channel is achieved across all metrics.

However, the proposed identity loss is not directly concerned with the visual quality of the inpainted RGB-D images, but focuses on the visual loss of identity on a perceptual level. To assist in the evaluation of this matter, Table 4.1 additionally provides the average identity error of the models based on the RGB channels of their output. As expected, the model that is trained with our identity loss significantly outperforms the model that is trained without identity loss in this regard. While one may anticipate these values to be even further apart based on the

	L1 error			L2 error			PSNR			SSIM			VIF			ID
Method	RGB	D	SN	RGB	D	SN	RGB	D	SN	RGB	D	SN	RGB	D	SN	RGB
(A) *Without* ID loss	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
(B) *With* ID loss	**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**	**7.698**

Table 4.1: Quantitative results (L1 error, L2 error, PSNR, SSIM, and VIF) of the model trained *without* identity loss and the model trained *with* identity loss. The results are split for the color channels (RGB), depth channel (D) and surface normal image channels (SN).

	L1 error			L2 error			PSNR			SSIM			VIF		
Method	RGB	D	SN	RGB	D	SN	RGB	D	SN	RGB	D	SN	RGB	D	SN
Modality-specific	**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**
Data-level fusion	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

Table 4.2: Quantitative results (L1 error, L2 error, PSNR, SSIM, and VIF) of the modality-specific models for RGB and depth image inpainting and the data-level fusion model. The results are split for the color channels (RGB), depth channel (D) and surface normal image channels (SN).

presented qualitative results, it is essential to note that the identity error is calculated with the combination of the known region and inpainted region. As such, a reasonably sized section of the 128-byte embedding is based on the lower known half of the face.

Fusion of color and depth information

In this section, we present a quantitative breakdown of the results of the fusion types as described in Section 3.3, aimed at Objective 1.2. Similar to our qualitative evaluation of the proposed fusion types, a separately trained RGB image and depth image inpainting model [41] will form a comparative baseline for the fusion methods we evaluate in this section. The output of these modality-specific models will give an indication of the level of visual quality we aim to match with a jointly trained model in terms of objective metric values.

We will first look into the results of data-level fusion and how this compares to modality-specific models. Following this, we will examine the results of hybrid fusion, where one of three fusion methods is applied, namely: fusion through addition, single-path residual fusion and multi-path residual fusion.

We firstly look into how the results of data-level fusion compare to results produced by modality-specific models. Considering the results in Table 4.2, we consistently observe that the modality-specific models provides better results compared to the results of the data-level fusion model. Moreover, in the results for the data-level fusion model for the metrics SSIM and VIF, which both represent perceived quality, we note a large relative degradation of the results for the depth channel. In contrast, we do not identify the same effect for the color channels of the results of the data-level fusion model. This is consistent with our observations in the corresponding visual examination of the results of these models.

	L1 error			L2 error			PSNR			SSIM			VIF			ID
Method	RGB	D	SN	RGB	D	SN	RGB	D	SN	RGB	D	SN	RGB	D	SN	RGB
Data-level fusion	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
Residual-*SP4*	**7.951**	**3.532**	24.604	**21.214**	14.684	41.062	**21.080**	25.042	15.926	**0.937**	0.975	0.865	0.528	0.668	0.462	7.547
Residual-*SP6*	8.046	3.641	24.120	21.370	**14.649**	40.288	21.031	**25.062**	16.093	**0.937**	**0.976**	0.866	**0.534**	**0.669**	0.464	7.521
Residual-*MP4*	8.231	3.857	**23.912**	22.083	15.445	40.381	20.714	24.573	16.066	0.936	0.975	**0.869**	0.527	0.662	**0.466**	7.515
Residual-*MP6*	8.166	3.786	24.512	21.776	15.271	40.714	20.806	24.665	15.990	0.934	0.975	0.862	0.529	0.660	0.461	**7.436**

Table 4.3: Quantitative results (L1 error, L2 error, PSNR, SSIM, and VIF) of our data-level fusion model, hybrid single-path residual fusion and hybrid multi-path residual fusion. The results are split for the color channels (RGB), depth channel (D) and surface normal image channels (SN).

	L1 error			L2 error			PSNR			SSIM			VIF			ID
Method	RGB	D	SN	RGB	D	SN	RGB	D	SN	RGB	D	SN	RGB	D	SN	RGB
(M)	**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
(M) + $\mathcal{L}_{SN}$	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**

Table 4.4: Quantitative results (L1 error, L2 error, PSNR, SSIM, and VIF) of model (M) and model (M) with the addition of $\mathcal{L}_{SN}$. The results are split for the color channels (RGB), depth channel (D) and surface normal image channels (SN).

We now move on to discuss the results of the different types of fusion, presented in Table 4.3. In general, each method presented in this table achieves results that are similar to each other, which is in correspondence with our visual examination. While the margins are slim, digging deeper into the presented results, we find some evidence to indicate that residual fusion typically outperforms data-level fusion. Moreover, we find that single-path residual fusion outperforms multi-path residual fusion. Despite these findings, which should be interpreted with care, we cannot prove a clear benefit of hybrid fusion over data-level fusion in the context of our architecture and data.

Reproduction of surfaces

We now move on to consider the performance of models containing additions that are aimed at improving the surfaces formed by the inpainted depth values, as introduced in Section 3.4. The models used for this evaluation are based on our data-level fusion network, trained with a reconstruction loss, SN-PatchGAN loss and identity loss. Once again, to avoid confusion and to save space, we denote this network configuration as (M).

We evaluate the following aspects in order: surface normal loss $\mathcal{L}_{SN}$, contextual surface attention (CSA), and surface discriminator. As described in Section 3.4, these components are intended to work together by facilitating the interpretation of surface in several parts of our network. For this reason, we evaluate these components by cumulatively adding them to the aforementioned network configuration (M) one-by-one, which gives us the opportunity to evaluate their added value.

The results in Table 4.4 reveal that the addition of the surface normal loss has a great impact on the visual quality of the inpainted surfaces. Across all metrics, we observe a significant improvement in regard to the reproduction of the depth image, which corresponds to our observation during visual examination of the results. However, this does appear to come at a minor cost, as the inpainting of RGB channels show to have degraded in the wake of the addition of

	L1 error			L2 error			PSNR			SSIM			VIF			ID
Method	RGB	D	SN	RGB	D	SN	RGB	D	SN	RGB	D	SN	RGB	D	SN	RGB
(M) + $\mathcal{L}_{SN}$	**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**
(M) + $\mathcal{L}_{SN}$ + *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

Table 4.5: Quantitative results (L1 error, L2 error, PSNR, SSIM, and VIF) of model (M) with the addition of $\mathcal{L}_{SN}$ and model (M) with the addition of $\mathcal{L}_{SN}$ and CSA. The results are split for the color channels (RGB), depth channel (D) and surface normal image channels (SN).

Method	L1 error			L2 error			PSNR			SSIM			VIF			ID
	RGB	D	SN	RGB	D	SN	RGB	D	SN	RGB	D	SN	RGB	D	SN	RGB
(M) + $\mathcal{L}_{SN}$ + *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
(M) + $\mathcal{L}_{SN}$ + CSA + SN disc.	9.697	4.245	20.485	24.275	16.953	36.415	19.875	23.778	16.994	0.928	0.972	0.890	0.509	0.662	**0.496**	8.325

Table 4.6: Quantitative results (L1 error, L2 error, PSNR, SSIM, and VIF) of model (M) with the addition of $\mathcal{L}_{SN}$ and CSA, and model (M) with the addition of $\mathcal{L}_{SN}$, CSA and the SN discriminator. The results are split for the color channels (RGB), depth channel (D) and surface normal image channels (SN).

the surface normal loss.

The contextual surface attention (CSA) module builds on top of the base framework's contextual attention (CA) module (Section 3.1) by adding surface normal information. Based on the additional information, the modified module aims to improve the search of feature patches that are appropriate in the refinement of the inpainted region. Table 4.5 represents the results of this change, which evidently shows an improved performance for the model that contains the CSA module in place of the CA module of the base framework.

The addition of supplementary information regarding the depth values of the RGB-D image do not compromise the inpainting performance of the color channels. This is somewhat unexpected, as we observed a moderate deterioration of the quality of the inpainted RGB channels upon the addition of the surface normal loss in Table 4.4.

We now move on to the evaluation of the surface normal discriminator (SND), of which the comparative results are presented in Table 4.6. What is striking about the quantitative results of the model trained with the SND is that they are worse for every metric, with the exception of the VIF of the surface normal image. This significantly deviates from our qualitative evaluation, where we found that the visual quality of the of the model trained with and without the SND were nearly identical.

Breaking down the results further, we find that both the RGB and depth channels show remarkably deteriorated results, whereas the metrics in regard to the surface normal present results that have declined more mildly.

4.5. Implications of the results

Thus far, we have provided a qualitative and quantitative look of each of the proposed network components and losses based on our defined research objectives. As this work is exploratory, we have provided minor points of intermediate analysis throughout this chapter. To conclude this chapter, we summarize our findings and analyze their implications.

4.5.1. Identity preservation

We studied the effect of our proposed identity loss, which was aimed at the preservation of identifying facial features, in accordance with Objective 1.1. In our qualitative evaluation, we found that models trained with our identity loss produce identity features that are consistent with the provided reference image. An interesting finding is that, while the identity loss is calculated based on the RGB channels, the depth channel of the inpainted images shows similar

facial features. This indicates that the model is able to indirectly learn the relation between the identity loss and depth image, based on the feedback it receives regarding the inpainted RGB channels. This effect was most clearly visible in Figure 4.4, where masked images and reference images of two different identities were fed to the model. Moreover, we observed an improved symmetry of facial features in the inpainted images. However, this symmetry does not extend as far as eye color, as we regularly observed differing eye colors in a single inpainted image. This is likely due to the fact that an incorrect eye color makes a minor impact to our employed loss functions considering the relatively small image area it comprises. It can thus be suggested that a discounted reconstruction loss function or stricter identity loss function could alleviate this problem. Upon the visual examination of the results, we note that inpainted results often differ from the ground truth with respect to expression. This effect was foreseeable, as the model does not have any information that indicates the expression of the subject. However, it is worth noting that our model demonstrates robustness against different types of expressions that are visible in the known image region, such as opened mouths and smiles. Lastly, we observe that the impact of the identity loss function deteriorates with great pose angles of the faces in the input image and the reference image. This suggests that extreme pose angles are detrimental to the performance of the identity loss function, which could be attributed to the limitations of the face embedding model that the identity loss is built on.

4.5.2. Fusion of color and depth information

We also considered the performance of different types of fusion, corresponding to Objective 1.2. To gain insights on the performance of modality-specific models and jointly trained models, we compared the results of a separate color image inpainting model and depth image inpainting model with the results of a data-level joint RGB-D model. We found a notable deterioration of the results of the jointly trained model compared to the results of modality-specific models. This suggests that the joint feature construction in the joint RGB-D model is not able to properly capture the features of both modalities. Moreover, this implies that the geometric information represented in the depth channel does not support the improvement of the inpainting of the color image, and vice-versa. We observed the largest relative deterioration of the visual quality of the data-level fusion model in the depth images, which contained a high amount of noise (Figure 4.5). The worsened performance of the data-level fusion model could be attributed the fact that the construction of features is compromised, as the model input contains multiple modalities with different statistical properties. And as the depth modality is only represented by a single output channel, it is outmatched by the 3-channel color modality in the calculation of the reconstruction loss and SN-PatchGAN loss. This may explain why we mainly observe the degradation of the visual quality of the results of the data-level fusion model in the depth channel.

What followed was a comparison of the fusion types that have been proposed in this work. Aside from data-level fusion, we evaluated several types of feature-level fusion within the proposed concept of hybrid fusion (Section 3.3.2), specifically, fusion through summation, single-path residual fusion and multi-path residual fusion. At an early point in the evaluation process, we concluded that fusion through summation performed poorly (Figure 4.7), as it resulted in blurry and noisy output from the coarse network stage. Therefore, it was excluded from any fur-

ther evaluation. This poor result was somewhat expected, as the summation operation applied to the color and depth features most probably overrules modality-specific features. Nevertheless, it is interesting that this fusion type resulted in moderate success in the work of Hazirbas et al. [140], which indicates that this fusion type may just be ill-fitted for the task of RGB-D image inpainting.

In both the qualitative and quantitative evaluation we found that data-level fusion and both types of residual fusion provide comparable results throughout the coarse and refined stage of the network. Specifically, we observed a significant amount of noise in the inpainted region of the depth image and its surface interpretation, which we commonly found to be visibly disconnected from the known image region. While the quantitative results of the models with residual fusion were slightly better when compared to data-level fusion, there exists insufficient proof to suggest that residual fusion actively contributes to results of higher visual quality. There are two likely causes for this observed effect. Firstly, the usage of residual fusion within the concept of hybrid fusion may not sufficiently stimulate and accommodate multimodal feature understanding. Or secondly, we may have overestimated the impact of the coarse result on the refinement stage of the network. We find the latter explanation to be most likely as in preliminary experiments we have found that, while of much lower visual quality, the second stage of the network is able to produce fairly good results even with a blank input image. We hypothesize that larger differences in performance between the fusion types may occur with a more challenging real-world dataset, in which the model has to actively deal with missing and noisy information.

4.5.3. Reproduction of surfaces

We are now moving on to consider the insights obtained based on the results of the components and loss function focused on surface reproduction, following Objective 1.3. We firstly considered the effects of the surface normal loss function, which calculates the L1 loss between the estimated surface normal images. The qualitative and quantitative results of this loss function appeared to be in agreement, as a clear improvement of the smoothness and consistency of the inpainted depth image and its surface normal representation was identified (Figure 4.8, Table 4.4). This is similar to the finding of Matias et al. [46], who proposed this function for the depth image inpainting. As mentioned, the improvement of surface reproduction due to the surface normal loss function comes at a minor cost as shown in Table 4.4, which is a decreased quality of inpainted RGB channels. We expect that this cost could possibly be minimized through further hyperparameter tuning.

Interestingly, we note that there still exists an inconsistent connection between the unknown and known region of the image. A likely explanation for this is that an inconsistent connection at the boundary of the synthesized region has a limited impact on the value of the surface loss function, making this an affordable flaw during model training.

In terms of the contextual surface attention (CSA) module that replaces the contextual attention (CA) module of the base network, we found no evident effect to indicate any improvement in the visual quality of results (Figure 4.8). However, in contrast, the quantitative results of the model with the CSA module presented in Table 4.5 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 contextual attention branch of the network, enhancing the feature matching process.

Having outlined our analysis of the results of the surface normal loss and CSA module, we now turn to consider the surface normal discriminator (SND). Remarkably, we found a significant disagreement between our qualitative and quantitative evaluation of this discriminator. Specifically, in our visual examination (Figure 4.7), we observed that the results of the model trained *with* the SND are predominantly similar to the results of the model *without* the SND. On the other hand, the quantitative results presented in Table 4.6 tell a vastly different story, as the results of the model trained *with* the SND are consistently worse than the model trained *without* the SND. This highlights the importance of each of our evaluation methods, and simultaneously prompts caution in their interpretation. Based on this uncertainty, we cannot safely state whether SND is a worthwhile addition the network.

Furthermore, we make a notable observation regarding the quantitative results with respect to the inpainted depth images of all discussed loss functions and components in this section. Specifically, as the quantitative results of the surface normal image improve, the quantitative results of the depth images often remain the same or even deteriorate. This is somewhat counter-intuitive, as both representations are sourced from the same depth channel. However, there is a likely explanation for this effect. As the surface-oriented losses and modules stimulate the model to prioritize smooth and consistent surfaces represented in the depth channel, the model has a limited freedom with respect to the estimation of depth values of single pixels. In this case, if the inpainted surface does not exactly correspond to the ground truth, this error will be propagated to a large part of the neighboring depth values.

4.5.4. Concluding remarks

This chapter set out to evaluate the proposed components and loss functions that were explored and proposed in Chapter 3. At this point, we have gained sufficient insights to select the final model that forms the best solution with respect to our research objectives.

Firstly, it has become clear that the identity loss function is an indispensable loss function of our network. Furthermore, aside from hybrid fusion with fusion through summation, all proposed fusion methods demonstrate similar performance in both the qualitative and quantitative evaluation. Although we hypothesize that residual fusion will outperform data-level fusion on real-world data, we therefore pick the simplest fusion strategy, which indisputably is data-level fusion. Furthermore, we found that the surface normal loss function significantly improves the visual quality and objective metric values of the depth channel and its surface representation. In addition, the CSA module demonstrated improved results across all RGB-D image channels. Lastly, the SND caused visual improvements of the surface normal image, but also brought about a significant deterioration of the results. Taken together, we determine our best performing framework configuration to be: 1) identity loss, 2) data-level fusion, 3) surface normal loss, and the 4) CSA module. A full overview of this architecture is available in Section A.2.

5

Conclusion

The main objective of this thesis was to design a framework for head-mounted display removal in RGB-D images. Based on our review of existing work in Chapter 2, we identified that relatively few joint RGB-D image inpainting frameworks [47, 101] have been introduced, and those undertaken do not focus on images of the human face. To guide our exploration, we defined a set of research objectives, in which we prioritized the preservation of identity and the degree of realism of the color and depth image information contained in RGB-D images. In correspondence with our research objectives, we proposed a number of architectural structures, loss functions and components in Chapter 3. In Chapter 4, we presented the results and analysis of all proposed architectural aspects. Following our results analysis, we determined our best performing framework configuration to include: 1) identity loss, 2) data-level fusion, 3) surface normal loss, and the 4) CSA module. In this chapter, we summarize our findings and discuss their implications with respect to past work and future work.

5.1. Discussion

We start by discussing our framework's robustness against pose in Section 5.1.1. We then summarize and discuss our findings with respect to representation learning from RGB-D images (Section 5.1.2) and the employment of surface normal estimation (Section 5.1.3). In Section 5.1.4, we discuss several challenging aspects of the training procedure of our framework and GANs in general. Lastly, in Section 5.1.5, we elaborate on the limitations of our work.

5.1.1. Pose robustness

In Chapter 4, we observed that our proposed framework demonstrates robustness against a wide range of pose angles, without the need of a target pose map [36, 70]. This may indicate that the geometric information provided by the depth channel provides the model with a reliable means of determining the pose of the occluded face.

However, as our training set and test set contain faces with randomized poses, we gained little insight on how the performance of our framework deteriorates as the pose angles increase. In this section, we aim to gather more insights in this regard. To achieve this, we evaluate the

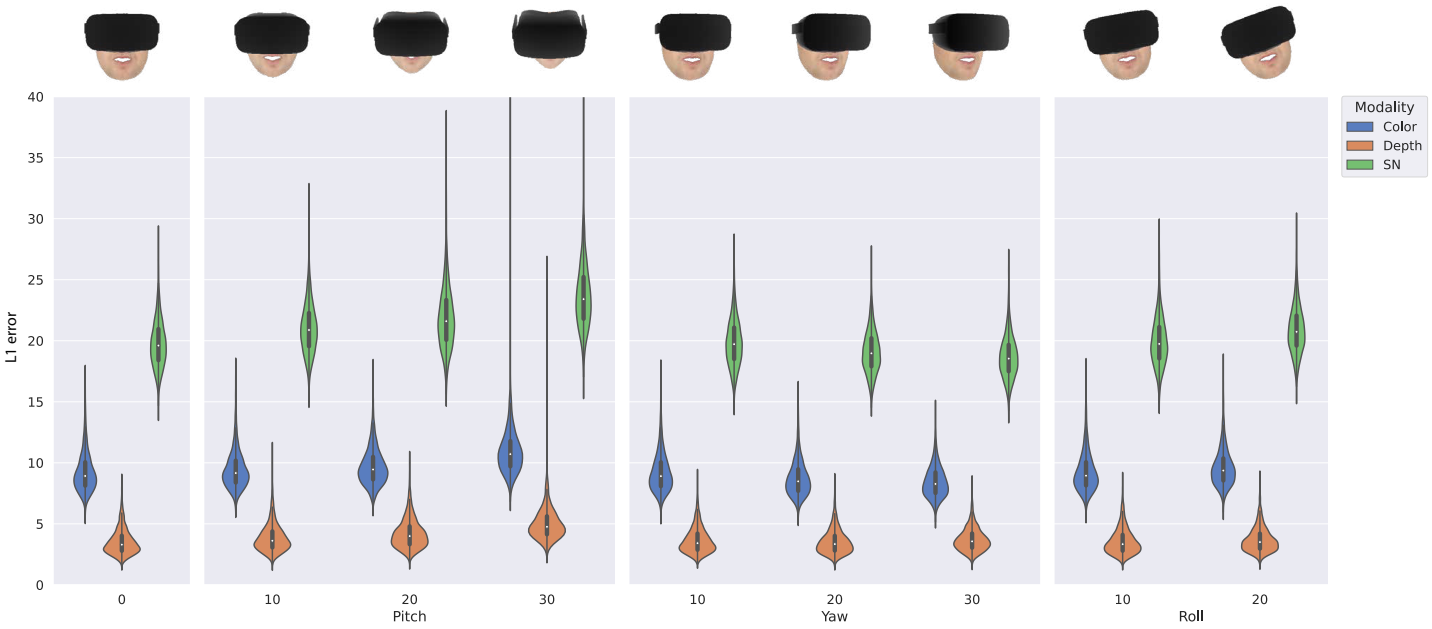


Figure 5.1: Violin plots that show the distribution of the L1 error of a set of specified pose angles (pitch, yaw, roll).

quantitative performance of our best-performing framework, as discussed in Section 4.5.4. The premise of this experiment is equivalent to the quantitative experiments conducted in Section 4.4, with the exception of the datasets used for the evaluation. In this experiment we use nine separate test sets, each of which contain the same faces but with differing specified pose angles.

The resulting L1 error is shown in Figure 5.1, presenting the mean results as well as the distribution of the L1 error values in a violin plot. Let us first consider the pitch, with values $[0°, 10°, 20°, 30°]$. We observe a relation between the increase of the mean error and the increase of the pitch angle. Notably, we find that the distribution of the L1 error widens simultaneously. This appears to be a consequence of the decreasing known image area available to our framework, which in turn is a consequence of the extrusion of the HMD.

Interestingly, compared to the observed effect regarding the pitch axis, we identify an inverse effect for the yaw axis, with values $[0°, 10°, 20°, 30°]$. Specifically, we find that the mean L1 error decreases as the yaw angle increases. This is somewhat counterintuitive, as similar to the pitch axis; the size of the known image region decreases with the increase of the yaw angle. However, in the case of the yaw axis, the region that is to be inpainted effectively also decreases, as less of the face is visible. This is a direct consequence of the fact that we measure our L1 error based on the full mask region, which in the case of high yaw angles, largely contains blank pixels.

Turning now to the effects of difference in roll angles, for the values $[0°, 10°, 20°]$, we observe a similar effect to the pitch axis. While the impact is weaker, we determine that the increased roll angles negatively affect the measured L1 error of the inpainted results. This is particularly interesting, as neither the surface area of the mask nor the balance between the amount of non-blank and blank pixels is affected.

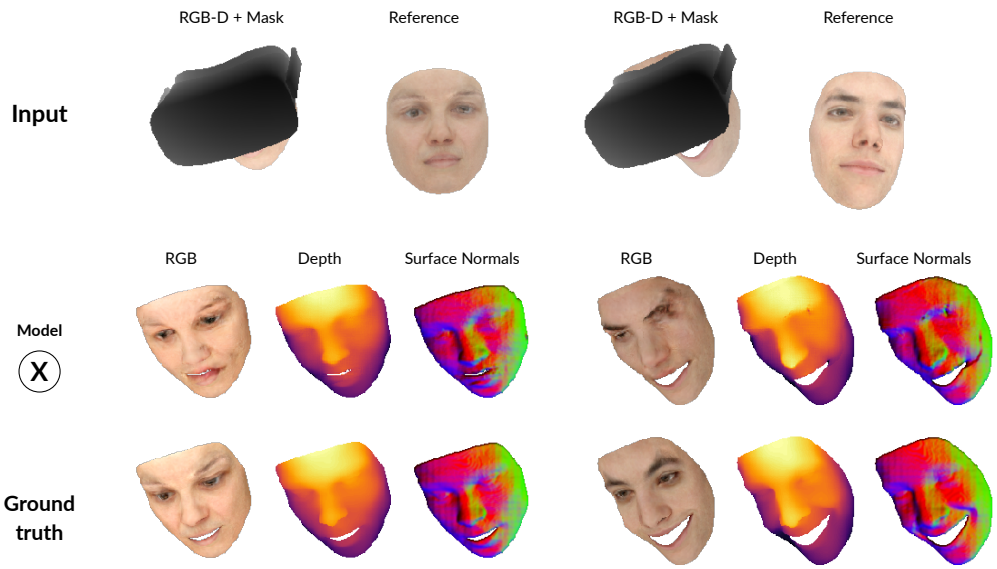


Figure 5.2: Failure cases with extreme pose angles. Model (X) refers to our best-performing model, which includes: 1) identity loss, 2) data-level fusion, 3) surface normal loss, and the 4) CSA module.

We found that combining several large angles for the pitch, yaw and roll axes result in more dramatic changes in objective metric values. However, lacking a straightforward visualization method for the simultaneous change of three different angles and having to keep our evaluations consolidated and relevant, we opt to limit ourselves to the above ranges. Additional visualizations of the other objective metric values are available in the appendix of this thesis (Section A.1).

Failure cases Lastly, we present a number of cases where the model fails to accurately inpaint the masked region, which are shown in Figure 5.2. While the depth images of the inpainted results appear reasonable, clear artifacts and noise are visible in the RGB channels of both results. We note that both input images have an extremely large masked region, which gives the model a minimal amount of contextual information to inpaint the image. Therefore, it is somewhat anticipated that the model would perform poorly in these cases.

5.1.2. RGB-D feature learning

In this thesis, various architectural structures have been introduced to stimulate multimodal feature learning. Having conducted our analysis of these aspects in Section 4.5, we now discuss our key learnings.

Our initial solution to fusion comprised the naive approach of data-level fusion. While the performance of this architecture is considerably worse than modality-specific models (Section 4.2), the results of data-level fusion were of reasonable visual quality. This is surprising, as we hypothesized that the feature construction process would be severely impacted when provided with multimodal data. Moreover, existing work outlined that data-level fusion may even hurt performance [137]. We speculate that the reasonable performance of data-level fusion could be a consequence of the flawless RGB-D images in our dataset. Specifically, our model may

learn to recognize cross modal features that are not as consistently found in real-world RGB-D data. Additional experiments with a sufficiently sized real-world RGB-D dataset would provide the means to investigate this theory.

Our proposed hybrid fusion strategy was aimed at the removal of the artifacts and noise introduced by the multimodality of the input data. Following previous work in semantic scene segmentation [140], we applied fusion through summation in the coarse stage with two modality-specific branches. However, during its evaluation in Section 4.3, we found that the coarse results of this fusion strategy are poor in quality. Perhaps this is related to the fact that we use separate decoders for the color and depth modality, as opposed to the singular decoder employed by Hazirbas et al. [140]. Moreover, this fusion strategy may simply be nontransferable to the task of joint RGB-D image inpainting.

Furthermore, we explored the application of residual fusion within the concept of hybrid fusion. Despite the consistent but marginal improvement in its quantitative evaluation over data-level fusion (Section 4.3), we could not identify a clear improvement of visual quality. However, similar to data-level fusion, we expect to make different observations in combination with a sufficiently sized real-world RGB-D dataset. Within the context of this work, considering the additional resource requirements of hybrid fusion and the presented results, we deemed data-level fusion to be the best choice among the two fusion strategies. Nonetheless, if computational resources are abundant, hybrid fusion may be favored for its marginally improved visual quality of inpainted results. Furthermore, we believe that the results of hybrid fusion warrant further exploration, as we have not exhaustively explored this concept due to the lengthy training time of our framework. Specifically, we recommend the exploration of more complex feature transformation operations [134] or a modality interaction path [132].

5.1.3. Surface normal representation

In addition to fusion strategies, we introduced several architectural changes to improve the interpretation of the depth values as a surface, inspired by a number of studies that focus on depth images [45, 46, 136]. In general, the addition of surface normal information demonstrated a significant improvement of the visual quality of the depth image. The surface loss function [46] successfully supervised the reproduction of the desired properties of the depth channel such as smoothness and the contextual surface attention module [46] showed to benefit from the auxiliary surface normal information.

In contrast, the surface normal discriminator demonstrated to have an adverse effect on the objective metric results (Section 4.4). Even so, it should be noted that the surface normal discriminator showed to remove the inconsistent connection between the known and unknown region of the image in our visual examination. We hypothesize that further hyperparameter tuning can alleviate the negative impact of this discriminator. Moreover, we speculate that the usage of a separate discriminator for surfaces could ease the process of hyperparameter tuning and may improve the visual quality of the results.

Figure 5.3: Example of failure mode with blank background. Left: input I with mask Ω, right: output.

5.1.4. Training our GAN-based architecture

In this section we discuss challenges we faced during the design and training of our architecture, including the creation of our dataset, dealing with failure modes of GANs, and balancing hyperparameters.

Dataset creation As typically is the case with neural models, GANs heavily rely on the statistical properties and characteristics of the provided training and evaluation data. In this section, we elaborate on a number of findings with respect to the design of a suitable dataset.

In a social VR setting, users are typically captured with an RGB-D sensor [10, 12]. Given the color and depth image provided by this sensor, it is a relatively easy task to extract the image region that contains the subject's face. Consequently, we created a database with face images that do not contain any background information (Section 4.1). Initially, we intuitively set the background region of the images to zero for each of the four image channels. However, upon training with this dataset, it became clear that some property of the data caused a severe instability of the training process. In particular, we observed that the coarse stage of the network always failed, whereas the refinement stage of the network showed an increased amount of instability and frequently failed as well. An example of such a result can be seen in Figure 5.3, where the inpainted region is blank.

By adjusting our dataset, we found that the instability of training is a consequence of the black background of the images in our dataset. When replacing the black background with a dark grey background, the stability and reliability of our architecture increased significantly. While we identified the cause through this observation, it is not immediately clear why this interaction exists, as black is a valid color that has no harmful interactions with any operations performed within our network. Moreover, while one may think that this could be related to our masked region, our framework marks the masked region with a solid grey value in a channel dedicated to the image mask. Consequently, this also does not seem to provide a clear explanation of the identified behavior.

We found that the coarse stage is guaranteed to fail when training with images with a black background, while the refinement stage is able to provide a valid output in some cases. This observation leads us to take a closer look at the design of the coarse stage of our architecture. Specifically, our attention is drawn to the L1 reconstruction loss. In the most basic version of our architecture, this is the only loss function used by the coarse stage, while the refinement stage additionally employs the SN-PatchGAN loss. This indicates that the problem may be related to

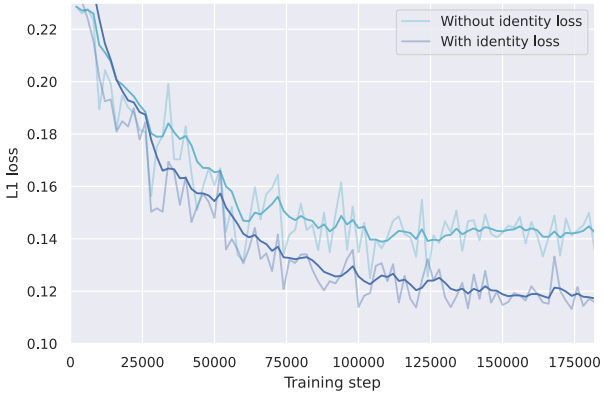


Figure 5.4: Reconstruction loss curve during training of a model training *without* identity loss and a model trained *with* identity loss.

the reconstruction loss function. Taking into account the large portion of our masked region that is *blank* (indicating background), in terms of reconstruction loss, it is very expensive for the model to predict a foreground pixel in the position of a ground truth background pixel. In turn, this results in unstable gradients during training, which could destabilize learning to the point that the model fails completely. We consider this to be the most likely cause of this issue.

Convergence and hyperparameters GANs employ a unique training procedure that is unlike the training procedure of other machine learning models. Training two models with opposite objectives at once, their training process is impacted with issues regarding transparency and reliability [86]. As mentioned in Section 2.3, GANs are trained through an adversarial process between generator G and discriminator D, where the theoretical target outcome is to find a Nash equilibrium of the common objective function. Thus, the discriminator countermeasures the objective of the generator, and vice versa. Since this causes a large oscillation of gradient updates during training, this minimax problem forms a complicated task for optimization methods, causing instability and high sensitivity to the adjustment of hyperparameters.

While several optimization methods have been proposed to improve this problem [86], hyperparameter tuning of GANs remains a challenging task that demonstrates unstable behavior. Accordingly, we faced similar difficulties in our work.

The output of the objective function of a GAN is difficult to interpret during training, making it a challenge to recognize model convergence based on it. This due to the fact that each of the models update their parameters for their respective objective, while completely disregarding the objective of the other. During training, we monitor a collection of aspects that are indicative of the status of the training procedure. Firstly, while the SN-PatchGAN loss value itself is not interpretable, the convergence of its loss curve and its respective gradients can suggest model convergence. In terms of training loss curves, we primarily monitored the convergence of the curves of the reconstruction loss, identity loss and surface normal loss. The values of these loss functions form a more straightforward base to recognize model convergence. An example of a converged reconstruction loss curve is shown in Figure 5.4.

However, the clearest indicator of convergence of our architecture, and GANs in general, is the observed visual quality of the output. While loss functions form a reasonable base for

recognizing model convergence, they are not always reliable. For this reason, evaluation of hyperparameter configurations of GANs nearly always includes visual examination.

The difficulty of recognizing convergence aside, a major challenge in the design and hyperparameter tuning of our models was the lengthy training time. GANs currently take a significant amount of time to train, ranging from a few days to as much as two months [38]. While some signs of failure may be noticed early in the training process at which point it can be stopped, termed *early stopping* [71], improvements in the visual quality of the results may only appear at a later point in the training process. Therefore, in view of hyperparameter tuning, it is generally beneficial to keep the amount of hyperparameters of a network down to a minimum.

5.1.5. Limitations

Research scope As discussed in Section 2.1, HMD removal is a broad task that comprises many challenges across multiple dimensions. The task of HMD removal is concerned with resolving the occlusion caused by an HMD, but can be further broken down into several functional aspects including: 1) eye gaze detection, 2) speech synchronization, and 3) expression detection. These individual tasks are not fully solved, which makes combining them towards solving HMD removal even more difficult. Moreover, these tasks require additional sensors such as internal IR cameras, which we did not have readily available throughout this research. Having taken these aspects into consideration, we have limited ourselves to the challenge of joint completion of visual and geometric information with a unique approach. This limits the application of our proposed framework, as it lacks a large quantity of functionalities for usage in real-world situations, which other model-based HMD removal methods do provide [34].

Aside from the mentioned subtasks, HMD removal also involves a wide range of non-functional aspects including: 1) the degree of realism of the occlusion removal, 2) time performance, and 3) temporal correctness. In this case, realism concerns humanlikeness and identity, which both have a profound impact on social experiences [30, 48]. In this work, we have taken the degree of realism of the inpainted results as our primary objective, and have prioritized this factor in both our framework design and evaluation. While we expect the results of our efforts to be transferable to real-world data, we are unable to demonstrate or prove this due to the lack of large-scale RGB-D datasets.

While speed and resource requirements are of absolute importance for a real-world HMD removal system, we have not considered this as a primary objective in this work. This choice was made to avoid conflict with our objective of achieving high visual quality and realism, which arguably is more important. Large deep neural networks currently form the state-of-the-art in image completion, which have high resource requirements. To leverage these advancements, we chose not to include specific speed or resource requirements, especially considering the fact that we process an additional channel compared to the default three RGB channels. Another argument against setting premature speed and resource requirements is the fact that hardware is continuously evolving and improving, allowing for larger workloads during model training and inference. Despite the fact that speed was not a primary objective in this thesis, the proposed models achieve real-time performance on a single NVIDIA GeForce RTX 2080 Ti GPU, as described in Section 4.2.

In a real-world immersive teleconference system, HMD removal would be applied to real-time videos. In this case, temporal correctness refers to the consistency between the inpainted frames. To bring down our scope to an attainable level for a master's thesis, we opted to solely focus on image inpainting. Therefore, our framework does not take temporal correctness into consideration, making it unfit for applications in a real-world situation as is. While closely related to image inpainting, video inpainting [93, 172] is a task with its own set of challenges and remains unsolved. Our framework could potentially form the base for a video inpainting framework, as the identity loss demonstrates the potential of information extraction from additional frames.

Lack of suitable datasets As discussed in Section 2.3, GANs aim to learn complex distributions in order to generate samples with sufficient diversity and level of detail. Therefore, the training process of GANs requires a large amount of data.

In Section 2.6, we provided a representative overview of the RGB-D face datasets that are currently available. We concluded that RGB-D face image datasets that are sufficiently sized for training a GAN currently do not exist, which motivated the decision to create a synthetic dataset based on the parametric Basel Face Model [62, 173]. This choice comes with the benefit of having full control over both the properties of the faces and the synthetic recording conditions. Moreover, it allows us to train our models with a large number of different identities, without the need of a controlled recording environment or the recruitment of thousands of people.

However, this choice also has a range of serious drawbacks. Firstly, the statistical Basel Face Model was constructed based on a total of 200 recorded face shape and color samples. While this has no direct consequence in regard to the number of identities we can sample from the model, it does limit the size of the parameter space of the model, which affects the diversity of sampled identities. This has a significant impact on the bias and generalizability of our model.

Secondly, our synthesized data set contains *perfect* RGB-D images, without any noise, misalignment, or artifacts that are frequently found in real-world RGB-D recordings. This being so, our model is not robust with respect to these types of data characteristics, as it has not been made familiar with them during training. An example that demonstrates this fact is shown in Figure 5.5, in which the performance of the model on a real-world recorded sample is shown. While our framework is able to figure out the correct pose of the face, the inpainted face region appears blurry and incoherent. Moreover, the identity of the subject's face is not preserved and does not appear realistic. The latter is a consequence of the fact that our model has only learned how to transfer identity features from and to faces with data characteristics from our synthetic dataset. Since, during training, any knowledge from the identity embedding model that is not applicable to the training data is not transferred to our model. Therefore, we hypothesize that the quality as well as the identity preservation of the inpainted result will improve dramatically after fine-tuning with real-world recordings.

Moreover, this not only affects the applicability of our model to real-world data, but also forms a major problem in the evaluation of our model. One of the main issues in this regard is that our proposed fusion methods may behave entirely differently with real-world data. Consequently, observations made based on real-world data may result in contrasting findings with respect to the conclusions in this work. For instance, we expect data-level fusion to perform

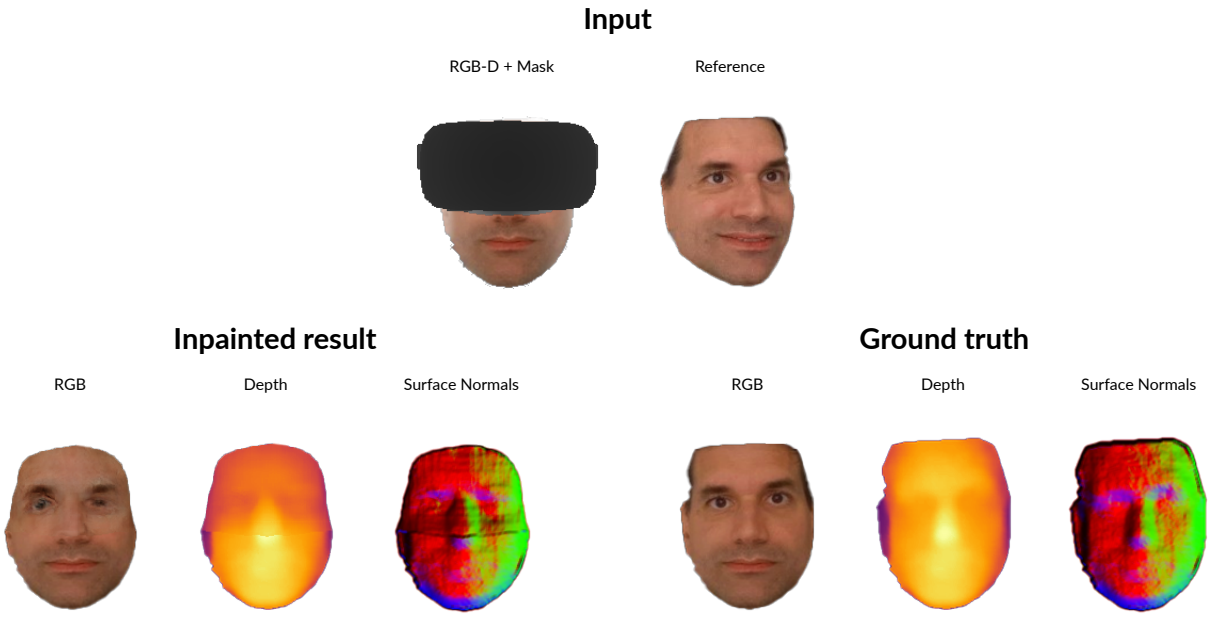


Figure 5.5: Example of an inpainted result of our framework (as described in Section 4.5.4) when given a real-world RGB-D image input. This image was recorded with an Azure Kinect [26], which was then preprocessed through depth-wise background removal and facial contour detection, without manual steps.

less well with real data, as modalities will not be as closely aligned and will contain noise and missing values. On the other hand, our proposed hybrid fusion strategy may facilitate the aggregation of complementary features, reducing the impact of missing and noisy data.

Despite the aforementioned issues, our synthesized data forms a substantial base for the exploration of joint RGB-D inpainting and our findings have important implications for models trained on real-world RGB-D data. Moreover, we expect that our models can be fine-tuned by retraining them on a real-world dataset with a lowered learning rate. Hence, it may be conceivable to apply our framework in real-world situations.

Evaluation strategy As outlined in 5.1.4, the interpretation of the performance of GANs is complex and difficult to quantify, which makes their evaluation challenging. We employed a qualitative and quantitative evaluation to provide a reliable assessment of a GAN-based framework, which is common practice for the evaluation of GAN-based methods. Specifically, we performed an elaborate visual examination of the results and selected a collection of representative objective metrics. Although our evaluation provided interesting insights, it is important to bear in mind that there exist potential inaccuracies or biases in this type of evaluation.

- Our evaluation does not include a large-scale human subjective study. Consequently, the findings of our qualitative experiment should be interpreted with caution.
- The visual quality of geometric data is difficult to interpret based on the two-dimensional format of this thesis. Therefore, some effects may have been overlooked or misinterpreted.

- Despite our efforts in the creation of a representative set of objective metrics, consensus has not been reached with respect to a standardized collection of metrics for the evaluation of GANs [83, 84]. In addition, the faces in our dataset contain expressions that are concealed behind the HMD, which our framework is in no way aware of. Consequently, our framework may incorrectly predict the occluded expressions, which affects the objective metric results.

Taken together, these points highlight the importance of careful interpretation of the results presented in this work. Notwithstanding these limitations, our evaluation demonstrated the effectiveness of our framework and provides important insights for future work.

5.2. Future work

There exist numerous potential paths for future research that can build upon the contributions of our work. Throughout our analysis and discussion, we have outlined a number of directions of future research. In this section, we provide a condensed overview of the directions that we find most promising.

- To validate the application of our framework in a real-world system, we suggest fine-tuning our trained models with a sufficiently sized set of real RGB-D sensor data of human faces once this becomes available. In this way, the valuable knowledge that was gained by training on synthetic data could be transferred to models trained with real RGB-D images.

- To improve the faithfulness of our synthetic dataset, we recommend the modification of our synthesization pipeline to augment the generated images. This augmentation would ideally involve the usage of a statistical model that simulates the noise of an RGB-D sensor [174].

- To improve the RGB-D feature learning process, we suggest the exploration of depth-aware convolution [130, 175]. This type of convolution allows the incorporation of the geometric information represented in the depth channel into two-dimensional convolution. At this moment, usage of depth-aware convolution is not widespread and is mainly aimed at semantic scene segmentation. Despite this, we believe that there is great a potential in the application of depth-aware convolution in our framework. We have taken the first steps in the implementation of this convolution type in our framework, but were unable to complete and evaluate this within the time frame of this research. Accordingly, we deem this to be a promising direction of future work, as it would potentially obviate the need of other fusion strategies.

- To enable the reproduction of occluded emotion, we recommend the addition of model input indicating the appearance of the occluded face region. Inspired by existing approaches to HMD removal [32–35], this input could originate from sensors such as strain sensors or internal IR cameras.

5.3. Conclusion

Head-mounted display (HMD) removal is a challenging task which has emerged with the increasing usage of HMDs to observe shared virtual reality (VR) environments. In this work, we propose a method that performs HMD removal through the joint inpainting of RGB-D images.

We determined that, despite the growing interest in the usage of RGB-D images, there does not exist an established RGB-D image inpainting framework that leverages generative adversarial networks (GANs). Due to the novelty of this problem, we took an exploratory approach to the design of a framework of this kind, guided by a set of defined research objectives. In this concluding section, we present our findings with respect to these objectives.

Research Objective 1 *Define an architecture that is capable of virtually removing the HMD from the wearer's face in RGB-D images.*

Due to its favorable properties and performance, we built our framework on top of the RGB image inpainting framework by Yu et al. [41]. Our coarse-to-fine architecture generates a coarse prediction of the masked region in its first stage, after which the coarse prediction is refined in its second stage. During training, the loss consists of a L1 reconstruction loss and SN-PatchGAN loss. In view of our research objectives, we explored several components and loss functions to enable joint RGB-D image inpainting.

Research Objective 1.1 *Define a module and loss function that stimulates the preservation of the identity features of the wearer's face.*

Firstly, to achieve preservation of identity (Objective 1.1), we propose a perceptual identity loss function which encourages the reproduction of distinctive facial features based on a given reference image. During training, the identity loss is calculated based on the distance between the identity embedding of the reference image and the inpainted image. These identity embeddings are retrieved through inference of a pretrained identity embedding model. Our quantitative and qualitative results clearly demonstrate that the addition of the identity loss successfully stimulates the preservation of identity-specific facial features.

Research Objective 1.2 *Define an architecture that is capable of handling the multimodal characteristics of RGB-D images.*

Secondly, we proposed several architectural structures to explore multimodal feature fusion of the color and depth information contained in RGB-D images. To this end, we introduced data-level fusion, which naively combines the color and depth information at network input, resulting in reasonable inpainted results. In addition, we introduced hybrid fusion, which involves feature-level fusion in the coarse stage of the architecture and data-level fusion in the refinement stage of the architecture. Within the concept of hybrid fusion, we investigated several fusion strategies, and proposed single-path and multi-path residual units. Our findings suggest that data-level fusion achieves similar performance to hybrid fusion. Therefore, we do not possess conclusive evidence to suggest that hybrid fusion outperforms data-level fusion in the context of our synthetic dataset.

Research Objective 1.3 *Define an architecture that stimulates the creation of smooth geometric surfaces.*

Thirdly, to improve the visual quality of the surfaces produced by the framework, we evaluated the usage of several architectural components. We found that the addition of the surface normal loss [46], provides a significant improvement of the smoothness of the surfaces represented in the inpainted RGB-D image. Furthermore, enhancing the matching process of the base framework's contextual attention module with the addition of a surface normal information demonstrated an improved overall performance. Moreover, we evaluated the addition of surface normal information to the SN-PatchGAN discriminator, which turned out to have an adverse effect and significantly deteriorated the results.

Research Objective 2 *In absence of a large-scale RGB-D face dataset, create a suitable dataset that is sufficiently sized.*

As a large RGB-D face image dataset is currently not available, we resorted to the creation of a synthetic dataset based on the Basel Face Model [62]. Accordingly, we were in full control of the face poses, expressions and synthetic recording conditions such as lighting. We built a full pipeline for the generation of faces with random expressions, random pose and random ambient lighting. In turn, we trained and evaluated our models with the resulting dataset.

In summary, we proposed a framework that is capable of the virtual removal of head-mounted display in RGB-D images. We formulated this problem as a joint RGB-D face image inpainting task and proposed a coarse-to-fine architecture 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. We further proposed a data-level fusion and hybrid fusion strategy and demonstrated their viability. Moreover, to improve surface reproduction in the depth channel, we introduced the employment of a surface normal loss function and contextual surface attention module. In absence of a large scale RGB-D face dataset, we devised a pipeline for the creation of a synthetic RGB-D face dataset. Based on the resulting dataset, we performed both qualitative and quantitative experiments to demonstrate the performance of each of the proposed architectural components and showed our framework's robustness against pose and expression. To conclude our exploration, we finally compared all evaluated model configurations and selected our final solution to our main research objective. Despite its exploratory nature and limitations, our research offers unique insights into the design and behavior of a multimodal image inpainting architecture that can be of interest to future research.

A

Appendix

A.1. Additional objective metric plots regarding pose robustness

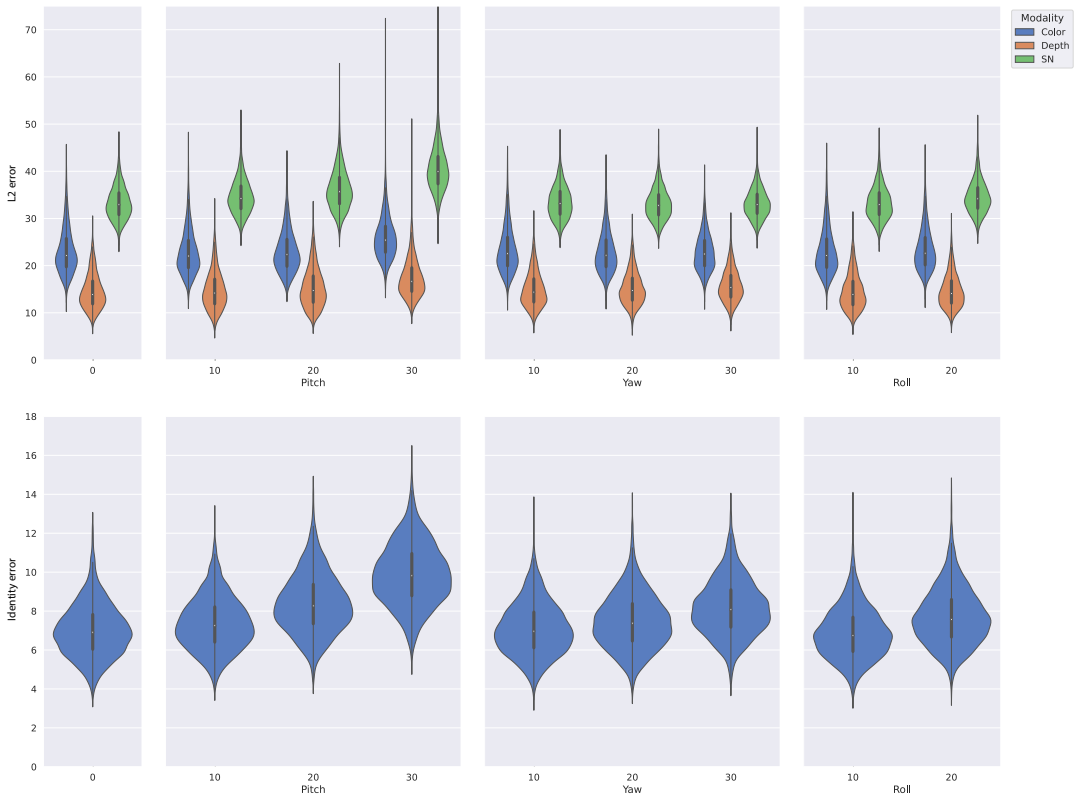


Figure A.1: Violin plots that show the distribution of the L2 error and identity error of a set of specified pose angles (pitch, yaw, roll).

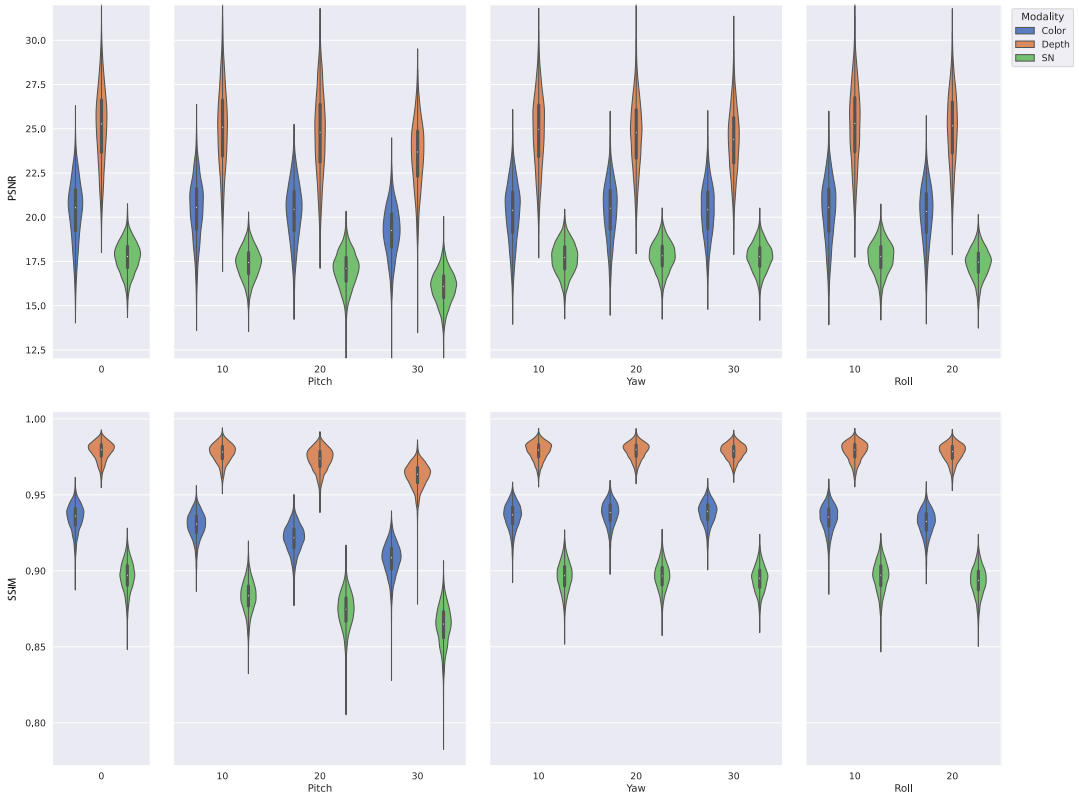


Figure A.2: Violin plots that show the distribution of the PSNR and SSIM of a set of specified pose angles (pitch, yaw, roll).

A.2. Full overview of the final architecture

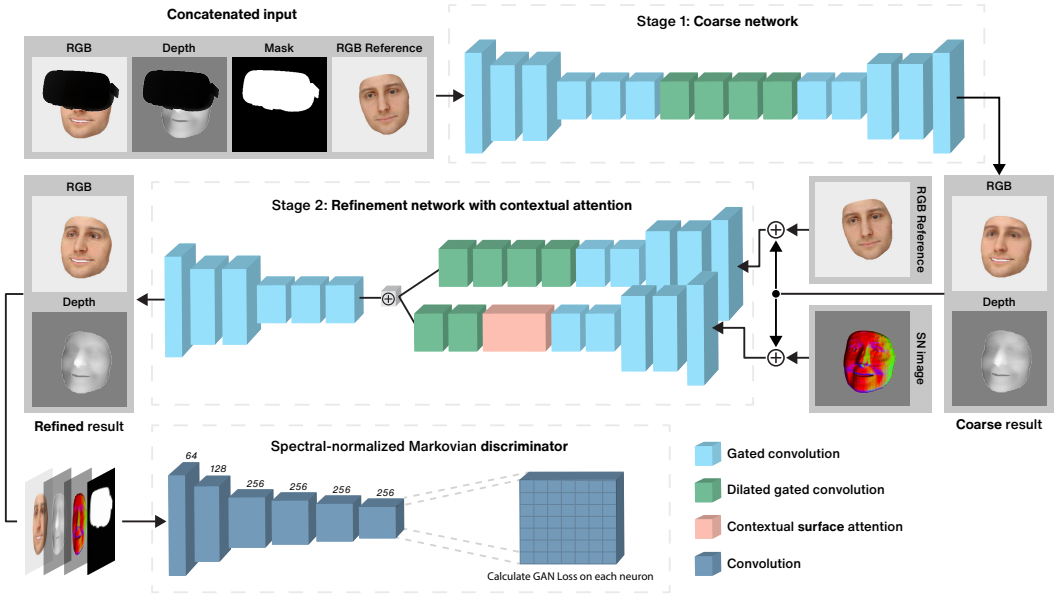


Figure A.3: Overview of our RGB-D image inpainting data-level fusion architecture with identity loss, surface normal loss, and contextual surface attention.

Bibliography

[1] Jack M Loomis. Distal attribution and presence. *Presence: Teleoperators & Virtual Environments*, 1(1):113–119, 1992. doi:10.1162/pres.1992.1.1.113.

[2] Frank Biocca and Mark R Levy. Communication in the age of virtual reality, 1995.

[3] Ralph Schroeder. *Possible worlds: the social dynamic of virtual reality technology*. 1996.

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

[5] David J Roberts, Arturo S Garcia, Janki Dodiya, Robin Wolff, Allen J Fairchild, and Terrence Fernando. Collaborative telepresence workspaces for space operation and science. In *2015 IEEE Virtual Reality (VR)*, pages 275–276. IEEE, 2015. doi:10.1109/vr.2015.7223402.

[6] Simon Gunkel, Hans Stokking, Martin Prins, Omar Niamut, Ernestasia Siahaan, and Pablo Cesar. Experiencing virtual reality together: Social vr use case study. In *Proceedings of the 2018 ACM International Conference on Interactive Experiences for TV and Online Video*, pages 233–238. ACM, 2018. doi:10.1145/3210825.3213566.

[7] Evelyne Klinger, Stéphane Bouchard, Patrick Légeron, Stéphane Roy, Françoise Lauer, Isabelle Chemin, and Pierre 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.

[8] Nyaz Didehbani, Tandra Allen, Michelle Kandalaft, Daniel Krawczyk, and Sandra 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.

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

[10] Simon NB Gunkel, Martin Prins, Hans Stokking, and Omar Niamut. Social vr platform: Building 360-degree shared vr spaces. In *Adjunct Publication of the 2017 ACM International Conference on Interactive Experiences for TV and Online Video*, pages 83–84. ACM, 2017. doi:10.1145/3084289.3089914.

[11] Martin J Prins, Simon NB Gunkel, Hans M Stokking, and Omar 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.

[12] Sylvie Dijkstra-Soudarissanane, Karim El Assal, Simon Gunkel, Frank ter Haar, Rick Hindriks, Jan Willem Kleinrouweler, and Omar Niamut. Multi-sensor capture and network processing for virtual reality conferencing. In *Proceedings of the 10th ACM Multimedia Systems Conference*, pages 316–319, 2019. doi:10.1145/3304109.3323838.

[13] Bradford S Bell and Steve WJ Kozlowski. A typology of virtual teams: Implications for effective leadership. *Group & Organization Management*, 27(1):14–49, 2002. doi:10.1177/1059601102027001003.

[14] Saniye Tugba Bulu. Place presence, social presence, co-presence, and satisfaction in virtual worlds. *Computers & Education*, 58(1):154–161, 2012. doi:10.1016/j.compedu.2011.08.024.

[15] Mel Slater and Sylvia Wilbur. A framework for immersive virtual environments (five): Speculations on the role of presence in virtual environments. *Presence: Teleoperators & Virtual Environments*, 6(6):603–616, 1997. doi:10.1162/pres.1997.6.6.603.

[16] Brid O'Conaill, Steve Whittaker, and Sylvia Wilbur. Conversations over video conferences: An evaluation of the spoken aspects of video-mediated communication. *Human-computer interaction*, 8(4):389–428, 1993. doi:10.1207/s15327051hci0804_4.

[17] Jeremy N Bailenson, Nick Yee, Dan Merget, and Ralph 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.

[18] Frank Biocca, Chad Harms, and Judee K Burgoon. Toward a more robust theory and measure of social presence: Review and suggested criteria. *Presence: Teleoperators & virtual environments*, 12(5):456–480, 2003. doi:10.1162/105474603322761270.

[19] *Facebook Horizon*, 2020 (accessed February 3, 2020). URL https://www.oculus.com/facebookhorizon/.

[20] *AltspaceVR*, 2020 (accessed June 27, 2020). URL https://altvr.com.

[21] Masahiro Mori, Karl F MacDorman, and Norri Kageki. The uncanny valley [from the field]. *IEEE Robotics & Automation Magazine*, 19(2):98–100, 2012. doi:10.1109/mra.2012.2192811.

[22] Jun'ichiro Seyama and Ruth S Nagayama. The uncanny valley: Effect of realism on the impression of artificial human faces. *Presence: Teleoperators and virtual environments*, 16 (4):337–351, 2007. doi:10.1162/pres.16.4.337.

[23] Debaleena Chattopadhyay and Karl F MacDorman. Familiar faces rendered strange: Why inconsistent realism drives characters into the uncanny valley. *Journal of Vision*, 16(11): 7–7, 2016. doi:10.1167/16.11.7.

[24] Valentin Schwind, Katrin Wolf, and Niels Henze. Avoiding the uncanny valley in virtual character design. *interactions*, 25(5):45–49, 2018. doi:10.1145/3236673.

[25] Shih-En Wei, Jason Saragih, Tomas Simon, Adam W Harley, Stephen Lombardi, Michal Perdoch, Alexander Hypes, Dawei Wang, Hernan Badino, and Yaser Sheikh. Vr facial animation via multiview image translation. *ACM Transactions on Graphics (TOG)*, 38(4): 1–16, 2019. doi:10.1145/3306346.3323030.

[26] Microsoft Azure Kinect. https://azure.microsoft.com/en-us/services/kinect-dk/. Accessed: 2020-05-20.

[27] Intel RealSense. https://realsense.intel.com. Accessed: 2020-05-20.

[28] Stephen RH Langton. The mutual influence of gaze and head orientation in the analysis of social attention direction. *The Quarterly Journal of Experimental Psychology: Section A*, 53(3):825–845, 2000. doi:10.1080/027249800410562.

[29] David M Grayson and Andrew F Monk. Are you looking at me? eye contact and desktop video conferencing. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 10(3): 221–243, 2003. doi:10.1145/937549.937552.

[30] Steve Benford, John Bowers, Lennart E Fahlén, Chris Greenhalgh, and Dave Snowdon. User embodiment in collaborative virtual environments. Citeseer. doi:10.1145/223904.223935.

[31] Marc Fabri, David J Moore, and Dave J Hobbs. The emotional avatar: Non-verbal communication between inhabitants of collaborative virtual environments. In *International gesture workshop*, pages 269–273. Springer, 1999. doi:10.1007/3-540-46616-9_24.

[32] Hao Li, Laura Trutoiu, Kyle Olszewski, Lingyu Wei, Tristan Trutna, Pei-Lun Hsieh, Aaron Nicholls, and Chongyang Ma. Facial performance sensing head-mounted display. *ACM Transactions on Graphics (ToG)*, 34(4):47, 2015. doi:10.1145/2766939.

[33] Kyle Olszewski, Joseph J Lim, Shunsuke Saito, and Hao 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.

[34] Justus Thies, Michael Zollhöfer, Marc Stamminger, Christian Theobalt, and Matthias 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.

[35] Christian Frueh, Avneesh Sud, and Vivek Kwatra. Headset removal for virtual and mixed reality. In *ACM SIGGRAPH 2017 Talks*, page 80. ACM, 2017. doi:10.1145/3084363.3085083.

[36] Yajie Zhao, Weikai Chen, Jun Xing, Xiaoming Li, Zach Bessinger, Fuchang Liu, Wangmeng Zuo, and Ruigang Yang. Identity preserving face completion for large ocular region occlusion. *arXiv preprint arXiv:1807.08772*, 2018.

[37] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

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

[39] Raymond A Yeh, Chen Chen, Teck Yian Lim, Alexander G Schwing, Mark Hasegawa-Johnson, and Minh N Do. Semantic image inpainting with deep generative models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5485–5493, 2017. doi:10.1109/cvpr.2017.728.

[40] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5505–5514, 2018. doi:10.1109/cvpr.2018.00577.

[41] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Free-form image inpainting with gated convolution. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4471–4480, 2019. doi:10.1109/iccv.2019.00457.

[42] Amir Atapour-Abarghouei and Toby P Breckon. Depthcomp: real-time depth image completion based on prior semantic scene segmentation. 2017. doi:10.5244/c.31.58.

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

[44] Ju Shen and Sen-Ching 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*, pages 1187–1194, 2013. doi:10.1109/cvpr.2013.157.

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

[46] Lucas PN Matias, Marc Sons, Jefferson R Souza, Denis F Wolf, and Christoph Stiller. Veigan: Vectorial inpainting generative adversarial network for depth maps object removal. In *2019 IEEE Intelligent Vehicles Symposium (IV)*, pages 310–316. IEEE, 2019. doi:10.1109/ivs.2019.8814157.

[47] Ryo Fujii, Ryo Hachiuma, and Hideo 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)*, pages 203–204. IEEE, 2019. doi:10.1109/ismar-adjunct.2019.00-46.

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

[49] Yujun Shen, Ping Luo, Junjie Yan, Xiaogang Wang, and Xiaoou Tang. Faceid-gan: Learning a symmetry three-player gan for identity-preserving face synthesis. In *Proceedings of*

the IEEE Conference on Computer Vision and Pattern Recognition, pages 821–830, 2018. doi:10.1109/cvpr.2018.00092.

[50] Zhigang Li and Yupin Luo. Generate identity-preserving faces by generative adversarial networks. *arXiv preprint arXiv:1706.03227*, 2017.

[51] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014.

[52] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou 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.

[53] 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.

[54] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *European conference on computer vision*, pages 87–102. Springer, 2016. doi:10.1007/978-3-319-46487-9_6.

[55] Zhou Wang, Alan C Bovik, Hamid R Sheikh, Eero 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.

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

[57] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015. doi:10.1109/cvpr.2015.7298682.

[58] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. doi:10.1109/cvpr.2016.90.

[59] Jiaxiong Qiu, Zhaopeng Cui, Yinda Zhang, Xingdi Zhang, Shuaicheng Liu, Bing Zeng, and Marc Pollefeys. Deeplidar: Deep surface normal guided depth prediction for outdoor scene from sparse lidar data and single color image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3313–3322, 2019. doi:10.1109/cvpr.2019.00343.

[60] Jun Li, Reinhard Klein, and Angela Yao. A two-streamed network for estimating fine-scaled depth maps from single rgb images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3372–3380, 2017. doi:10.1109/iccv.2017.365.

[61] Yosuke Nakagawa, Hideaki Uchiyama, Hajime Nagahara, and Rin-Ichiro Taniguchi. Estimating surface normals with depth image gradients for fast and accurate registration. In *2015 International Conference on 3D Vision*, pages 640–647. IEEE, 2015. doi:10.1109/3dv.2015.80.

[62] Thomas Gerig, Andreas Morel-Forster, Clemens Blumer, Bernhard Egger, Marcel Luthi, Sandro Schönborn, and Thomas Vetter. Morphable face models-an open framework. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 75–82. IEEE, 2018. doi:10.1109/fg.2018.00021.

[63] Volker Blanz, Thomas Vetter, et al. A morphable model for the synthesis of 3d faces. In *Siggraph*, volume 99, pages 187–194, 1999. doi:10.1145/311535.311556.

[64] Ivan E Sutherland. A head-mounted three dimensional display. In *Proceedings of the December 9-11, 1968, fall joint computer conference, part I*, pages 757–764, 1968. doi:10.1145/280811.281016.

[65] Boram Yoon, Hyung-il Kim, Gun A Lee, Mark Billinqhurst, and Woontack Woo. The effect of avatar appearance on social presence in an augmented reality remote collaboration. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pages 547–556. IEEE, 2019. doi:10.1109/vr.2019.8797719.

[66] Mel Slater and Anthony Steed. Meeting people virtually: Experiments in shared virtual environments. In *The social life of avatars*, pages 146–171. Springer, 2002. doi:10.1007/978-1-4471-0277-9_9.

[67] Dario Bombari, Marianne Schmid Mast, Elena Canadas, and Manuel Bachmann. Studying social interactions through immersive virtual environment technology: virtues, pitfalls, and future challenges. *Frontiers in psychology*, 6:869, 2015. doi:10.3389/fpsyg.2015.00869.

[68] Yajie Zhao, Qingguo Xu, Weikai Chen, Chao Du, Jun Xing, Xinyu Huang, and Ruigang 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)*, pages 267–276. IEEE, 2019. doi:10.1109/vr.2019.8797925.

[69] Marc Erich Latoschik, Daniel Roth, Dominik Gall, Jascha Achenbach, Thomas Waltemate, and Mario Botsch. The effect of avatar realism in immersive social virtual realities. In *Proceedings of the 23rd ACM Symposium on Virtual Reality Software and Technology*, pages 1–10, 2017. doi:10.1145/3139131.3139156.

[70] Miao Wang, Xin Wen, and Shi-Min Hu. Faithful face image completion for hmd occlusion removal. In *2019 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, pages 251–256. IEEE, 2019. doi:10.1109/ismar-adjunct.2019.00-36.

[71] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

[72] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35 (8):1798–1828, 2013. doi:10.1109/tpami.2013.50.

[73] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006. doi:10.1126/science.1127647.

[74] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010.

[75] Andrew L Maas, Awni Y Hannun, and Andrew Y Ng. Rectifier nonlinearities improve neural network acoustic models.

[76] Vincent Dumoulin and Francesco Visin. A guide to convolution arithmetic for deep learning. *arXiv preprint arXiv:1603.07285*, 2016.

[77] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989. doi:10.1162/neco.1989.1.4.541.

[78] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015.

[79] Ian Goodfellow. Nips 2016 tutorial: Generative adversarial networks. *arXiv preprint arXiv:1701.00160*, 2016.

[80] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[81] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[82] Lucas Theis, Aäron van den Oord, and Matthias Bethge. A note on the evaluation of generative models. *arXiv preprint arXiv:1511.01844*, 2015.

[83] Qiantong Xu, Gao Huang, Yang Yuan, Chuan Guo, Yu Sun, Felix Wu, and Kilian Weinberger. An empirical study on evaluation metrics of generative adversarial networks. *arXiv preprint arXiv:1806.07755*, 2018.

[84] Ali 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.

[85] Ishaan Gulrajani, Colin Raffel, and Luke Metz. Towards gan benchmarks which require generalization. *arXiv preprint arXiv:2001.03653*, 2020.

[86] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Advances in neural information processing systems*, pages 2234–2242, 2016.

[87] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015. doi:10.1109/cvpr.2015.7298594.

[88] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009. doi:10.1109/cvpr.2009.5206848.

[89] Antonio Criminisi, Patrick Pérez, and Kentaro Toyama. Region filling and object removal by exemplar-based image inpainting. *IEEE Transactions on image processing*, 13(9):1200–1212, 2004. doi:10.1109/tip.2004.833105.

[90] Jing Wang, Ke Lu, Daru Pan, Ning He, and Bing-kun Bao. Robust object removal with an exemplar-based image inpainting approach. *Neurocomputing*, 123:150–155, 2014. doi:10.1016/j.neucom.2013.06.022.

[91] Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan B Goldman. Patchmatch: A randomized correspondence algorithm for structural image editing. In *ACM Transactions on Graphics (ToG)*, volume 28, page 24. ACM, 2009. doi:10.1145/1576246.1531330.

[92] Anat Levin, Assaf Zomet, Shmuel Peleg, and Yair Weiss. Seamless image stitching in the gradient domain. In *European Conference on Computer Vision*, pages 377–389. Springer, 2004. doi:10.1007/978-3-540-24673-2_31.

[93] Alasdair Newson, Andrés Almansa, Matthieu Fradet, Yann Gousseau, and Patrick Pérez. Video inpainting of complex scenes. *SIAM Journal on Imaging Sciences*, 7(4):1993–2019, 2014. doi:10.1137/140954933.

[94] Ries Uittenbogaard, Clint Sebastian, Julien Vijverberg, Bas Boom, Dariu M Gavrila, et al. Privacy protection in street-view panoramas using depth and multi-view imagery. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10581–10590, 2019. doi:10.1109/cvpr.2019.01083.

[95] David J Heeger and James R Bergen. Pyramid-based texture analysis/synthesis. In *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*, pages 229–238. Citeseer, 1995. doi:10.1145/218380.218446.

[96] Alexei A Efros and Thomas K Leung. Texture synthesis by non-parametric sampling. In *Proceedings of the seventh IEEE international conference on computer vision*, volume 2, pages 1033–1038. IEEE, 1999. doi:10.1109/iccv.1999.790383.

[97] Marcelo Bertalmio, Luminita Vese, Guillermo Sapiro, and Stanley Osher. Simultaneous structure and texture image inpainting. *IEEE transactions on image processing*, 12(8):882–889, 2003. doi:10.1109/tip.2003.815261.

[98] Li-Yi Wei, Sylvain Lefebvre, Vivek Kwatra, and Greg Turk. State of the art in example-based texture synthesis. 2009.

[99] Marcelo Bertalmio, Guillermo Sapiro, Vincent Caselles, and Coloma Ballester. Image inpainting. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 417–424. ACM Press/Addison-Wesley Publishing Co., 2000. doi:10.1145/344779.344972.

[100] Antoni Buades, Bartomeu Coll, and J-M Morel. A non-local algorithm for image denoising. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pages 60–65. IEEE, 2005.

[101] Iddo Drori, Daniel Cohen-Or, and Hezy Yeshurun. Fragment-based image completion. In *ACM Transactions on graphics (TOG)*, volume 22, pages 303–312. ACM, 2003. doi:10.1145/1201775.882267.

[102] James Hays and Alexei A Efros. Scene completion using millions of photographs. *ACM Transactions on Graphics (TOG)*, 26(3):4, 2007. doi:10.1145/1275808.1276382.

[103] Junyuan Xie, Linli Xu, and Enhong Chen. Image denoising and inpainting with deep neural networks. In *Advances in neural information processing systems*, pages 341–349, 2012.

[104] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016. doi:10.1109/cvpr.2016.278.

[105] Guilin Liu, Fitsum A Reda, Kevin J Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. Image inpainting for irregular holes using partial convolutions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 85–100, 2018. doi:10.1007/978-3-030-01252-6_6.

[106] Chuan Li and Michael Wand. Precomputed real-time texture synthesis with markovian generative adversarial networks. In *European Conference on Computer Vision*, pages 702–716. Springer, 2016. doi:10.1007/978-3-319-46487-9_43.

[107] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018.

[108] Jeong-Seon Park, You Hwa Oh, Sang Chul Ahn, and Seong-Whan Lee. Glasses removal from facial image using recursive error compensation. *IEEE transactions on pattern analysis and machine intelligence*, 27(5):805–811, 2005. doi:10.1109/tpami.2005.103.

[109] Michael De Smet, Rik Fransens, and Luc Van Gool. A generalized em approach for 3d model based face recognition under occlusions. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1423–1430. IEEE, 2006.

[110] Umar Mohammed, Simon JD Prince, and Jan Kautz. Visio-lization: generating novel facial images. *ACM Transactions on Graphics (TOG)*, 28(3):57, 2009. doi:10.1145/1576246.1531363.

[111] Yijun Li, Sifei Liu, Jimei Yang, and Ming-Hsuan Yang. Generative face completion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3911–3919, 2017. doi:10.1109/cvpr.2017.624.

[112] Haofu Liao, Gareth Funka-Lea, Yefeng Zheng, Jiebo Luo, and S Kevin Zhou. Face completion with semantic knowledge and collaborative adversarial learning. In *Asian Conference on Computer Vision*, pages 382–397. Springer, 2018. doi:10.1007/978-3-030-20887-5_24.

[113] Mu Li, Wangmeng Zuo, and David Zhang. Deep identity-aware transfer of facial attributes. *arXiv preprint arXiv:1610.05586*, 2016.

[114] Rui Huang, Shu Zhang, Tianyu Li, and Ran 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*, pages 2439–2448, 2017. doi:10.1109/iccv.2017.267.

[115] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. doi:10.1109/cvpr.2015.7298965.

[116] Hamed Sarbolandi, Damien Lefloch, and Andreas Kolb. Kinect range sensing: Structured-light versus time-of-flight kinect. *Computer vision and image understanding*, 139:1–20, 2015. doi:10.1016/j.cviu.2015.05.006.

[117] Leandro Cruz, Djalma Lucio, and Luiz Velho. Kinect and rgbd images: Challenges and applications. In *2012 25th SIBGRAPI Conference on Graphics, Patterns and Images Tutorials*, pages 36–49. IEEE, 2012. doi:10.1109/sibgrapi-t.2012.13.

[118] Alexandre Hervieu, Nicolas Papadakis, Aurélie Bugeau, Pau Gargallo, and Vicent Caselles. Stereoscopic image inpainting: distinct depth maps and images inpainting. In *2010 20th International Conference on Pattern Recognition*, pages 4101–4104. IEEE, 2010. doi:10.1109/icpr.2010.997.

[119] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *European conference on computer vision*, pages 746–760. Springer, 2012. doi:10.1007/978-3-642-33715-4_54.

[120] Junyi Liu, Xiaojin Gong, and Jilin Liu. Guided inpainting and filtering for kinect depth maps. In *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, pages 2055–2058. IEEE, 2012.

[121] Amir Atapour-Abarghouei and Toby P Breckon. A comparative review of plausible hole filling strategies in the context of scene depth image completion. *Computers & Graphics*, 72:39–58, 2018. doi:10.1016/j.cag.2018.02.001.

[122] Jonas Uhrig, Nick Schneider, Lukas Schneider, Uwe Franke, Thomas Brox, and Andreas Geiger. Sparsity invariant cnns. In *2017 International Conference on 3D Vision (3DV)*, pages 11–20. IEEE, 2017. doi:10.1109/3dv.2017.00012.

[123] Hongyang Xue, Shengming Zhang, and Deng Cai. Depth image inpainting: Improving low rank matrix completion with low gradient regularization. *IEEE Transactions on Image Processing*, 26(9):4311–4320, 2017. doi:10.1109/tip.2017.2718183.

[124] Anat Levin, Dani Lischinski, and Yair Weiss. Colorization using optimization. In *ACM SIGGRAPH 2004 Papers*, pages 689–694. 2004. doi:10.1145/1186562.1015780.

[125] Liang Wang, Hailin Jin, Ruigang Yang, and Minglun Gong. Stereoscopic inpainting: Joint color and depth completion from stereo images. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008. doi:10.1109/cvpr.2008.4587704.

[126] David Doria and Richard 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*, pages 65–72. IEEE, 2012. doi:10.1109/cvprw.2012.6238916.

[127] Shohei Mori, Jan Herling, Wolfgang Broll, Norihiko Kawai, Hideo Saito, Dieter Schmalstieg, and Denis Kalkofen. 3d pixmix: Image inpainting in 3d environments. In *2018 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, pages 1–2. IEEE, 2018. doi:10.1109/ismar-adjunct.2018.00020.

[128] Shohei Mori, Sei Ikeda, and Hideo Saito. A survey of diminished reality: Techniques for visually concealing, eliminating, and seeing through real objects. *IPSJ Transactions on Computer Vision and Applications*, 9(1):1–14, 2017. doi:10.1186/s41074-017-0028-1.

[129] Antonio Criminisi, Patrick Perez, and Kentaro Toyama. Object removal by exemplar-based inpainting. In *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, volume 2, pages II–II. IEEE, 2003.

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

[131] Nicolas Audebert, Bertrand Le Saux, and Sébastien Lefèvre. Semantic segmentation of earth observation data using multimodal and multi-scale deep networks. In *asian conference on computer vision*, pages 180–196. Springer, 2016. doi:10.1007/978-3-319-54181-5_12.

[132] Liuyuan Deng, Ming Yang, Tianyi Li, Yuesheng He, and Chunxiang Wang. Rfbnet: deep multimodal networks with residual fusion blocks for rgb-d semantic segmentation. *arXiv preprint arXiv:1907.00135*, 2019.

[133] Seong-Jin Park, Ki-Sang Hong, and Seungyong Lee. Rdfnet: Rgb-d multi-level residual feature fusion for indoor semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4980–4989, 2017. doi:10.1109/iccv.2017.533.

[134] Jinghua Wang, Zhenhua Wang, Dacheng Tao, Simon See, and Gang Wang. Learning common and specific features for rgb-d semantic segmentation with deconvolutional networks. In *European Conference on Computer Vision*, pages 664–679. Springer, 2016. doi:10.1007/978-3-319-46454-1_40.

[135] Xiaofeng Ren, Liefeng Bo, and Dieter Fox. Rgb-(d) scene labeling: Features and algorithms. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2759–2766. IEEE, 2012. doi:10.1109/cvpr.2012.6247999.

[136] Saurabh Gupta, Pablo Arbelaez, and Jitendra Malik. Perceptual organization and recognition of indoor scenes from rgb-d images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 564–571, 2013. doi:10.1109/cvpr.2013.79.

[137] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. 2011.

[138] Camille Couprie, Clément Farabet, Laurent Najman, and Yann LeCun. Indoor semantic segmentation using depth information. *arXiv preprint arXiv:1301.3572*, 2013.

[139] Zhen Li, Yukang Gan, Xiaodan Liang, Yizhou Yu, Hui Cheng, and Liang Lin. Lstm-cf: Unifying context modeling and fusion with lstms for rgb-d scene labeling. In *European conference on computer vision*, pages 541–557. Springer, 2016. doi:10.1007/978-3-319-46475-6_34.

[140] Caner Hazirbas, Lingni Ma, Csaba Domokos, and Daniel Cremers. Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture. In *Asian conference on computer vision*, pages 213–228. Springer, 2016. doi:10.1007/978-3-319-54181-5_14.

[141] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1925–1934, 2017. doi:10.1109/cvpr.2017.549.

[142] P Jonathon Phillips, Patrick J Flynn, Todd Scruggs, Kevin W Bowyer, Jin Chang, Kevin Hoffman, Joe Marques, Jaesik Min, and William Worek. Overview of the face recognition grand challenge. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 947–954. IEEE, 2005.

[143] Lijun Yin, Xiaozhou Wei, Yi Sun, Jun Wang, and Matthew J Rosato. A 3d facial expression database for facial behavior research. In *7th international conference on automatic face and gesture recognition (FGR06)*, pages 211–216. IEEE, 2006.

[144] Chenghua Xu, Stan Li, Tieniu Tan, and Long Quan. Automatic 3d face recognition from depth and intensity gabor features. *Pattern recognition*, 42(9):1895–1905, 2009. doi:10.1016/j.patcog.2009.01.001.

[145] Chen Cao, Yanlin Weng, Shun Zhou, Yiying Tong, and Kun Zhou. Facewarehouse: A 3d facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics*, 20(3):413–425, 2013. doi:10.1109/tvcg.2013.249.

[146] Rui Min, Neslihan Kose, and Jean-Luc Dugelay. Kinectfacedb: A kinect database for face recognition. *Systems, Man, and Cybernetics: Systems, IEEE Transactions on*, 44(11): 1534–1548, Nov 2014. ISSN 2168-2216. doi:10.1109/TSMC.2014.2331215.

[147] Sherin Aly, Andrea Trubanova, Lynn Abbott, Susan White, and Amira Youssef. Vt-kfer: A kinect-based rgbd+ time dataset for spontaneous and non-spontaneous facial expression

recognition. In *2015 International Conference on Biometrics (ICB)*, pages 90–97. IEEE, 2015. doi:10.1109/icb.2015.7139081.

[148] Giorgia Pitteri, Matteo Munaro, and Emanuele Menegatti. Depth-based frontal view generation for pose invariant face recognition with consumer rgb-d sensors. volume 531, 07 2016. doi:10.1007/978-3-319-48036-7_67.

[149] Shiyang Cheng, Irene Kotsia, Maja Pantic, and Stefanos Zafeiriou. 4dfab: A large scale 4d facial expression database for biometric applications. *arXiv preprint arXiv:1712.01443*, 2017.

[150] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017. doi:10.1109/tpami.2017.2723009.

[151] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

[152] Yuri B Saalmann, Ivan N Pigarev, and Trichur R Vidyasagar. Neural mechanisms of visual attention: how top-down feedback highlights relevant locations. *Science*, 316(5831): 1612–1615, 2007. doi:10.1126/science.1139140.

[153] Volodymyr Mnih, Nicolas Heess, Alex Graves, et al. Recurrent models of visual attention. In *Advances in neural information processing systems*, pages 2204–2212, 2014.

[154] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *Advances in neural information processing systems*, pages 2017–2025, 2015.

[155] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057, 2015.

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

[157] Yujun Shen, Bolei Zhou, Ping Luo, and Xiaoou Tang. Facefeat-gan: A two-stage approach for identity-preserving face synthesis. *arXiv preprint arXiv:1812.01288*, 2018.

[158] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016. doi:10.1007/978-3-319-46475-6_43.

[159] Alexey Dosovitskiy and Thomas Brox. Generating images with perceptual similarity metrics based on deep networks. In *Advances in neural information processing systems*, pages 658–666, 2016.

[160] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[161] Pengchuan Zhang, Qiang Liu, Dengyong Zhou, Tao Xu, and Xiaodong He. On the discrimination-generalization tradeoff in gans. *arXiv preprint arXiv:1711.02771*, 2017.

[162] Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. A 3d face model for pose and illumination invariant face recognition. In *2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 296–301. Ieee, 2009. doi:10.1109/avss.2009.58.

[163] Oculus Rift. https://www.oculus.com/rift/. Accessed: 2020-05-20.

[164] Sergio Garrido-Jurado, Rafael Muñoz-Salinas, Francisco José Madrid-Cuevas, and Manuel Jesús Marín-Jiménez. Automatic generation and detection of highly reliable fiducial markers under occlusion. *Pattern Recognition*, 47(6):2280–2292, 2014. doi:10.1016/j.patcog.2014.01.005.

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

[166] Marian Stewart Bartlett, Gwen Littlewort, Mark G Frank, Claudia Lainscsek, Ian R Fasel, Javier R Movellan, et al. Automatic recognition of facial actions in spontaneous expressions. *Journal of multimedia*, 1(6):22–35, 2006. doi:10.4304/jmm.1.6.22-35.

[167] Virgilio F Ferrario, Chiarella Sforza, Graziano Serrao, GianPiero Grassi, and Erio Mossi. Active range of motion of the head and cervical spine: a three-dimensional investigation in healthy young adults. *Journal of orthopaedic research*, 20(1):122–129, 2002. doi:10.1016/s0736-0266(01)00079-1.

[168] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: A system for large-scale machine learning. In *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, pages 265–283, 2016.

[169] Amin Banitalebi-Dehkordi, Mahsa T Pourazad, and Panos 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)*, pages 1–4. IEEE, 2013. doi:10.1109/3dtv.2013.6676650.

[170] Quan Huynh-Thu and Mohammed Ghanbari. Scope of validity of psnr in image/video quality assessment. *Electronics letters*, 44(13):800–801, 2008. doi:10.1049/el:20080522.

[171] Eero P Simoncelli and Bruno A Olshausen. Natural image statistics and neural representation. *Annual review of neuroscience*, 24(1):1193–1216, 2001. doi:10.1146/annurev.neuro.24.1.1193.

[172] Ya-Liang Chang, Zhe Yu Liu, Kuan-Ying Lee, and Winston Hsu. Free-form video inpainting with 3d gated convolution and temporal patchgan. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9066–9075, 2019. doi:10.1109/iccv.2019.00916.

[173] Sami Romdhani and Thomas Vetter. Efficient, robust and accurate fitting of a 3d morphable model. In *ICCV*, volume 3, pages 59–66, 2003. doi:10.1109/iccv.2003.1238314.

[174] Chuong V Nguyen, Shahram Izadi, and David Lovell. Modeling kinect sensor noise for improved 3d reconstruction and tracking. In *2012 second international conference on 3D imaging, modeling, processing, visualization & transmission*, pages 524–530. IEEE, 2012. doi:10.1109/3dimpvt.2012.84.

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