Auf dem Weg zu Face-to-Face-Telepräsenzanwendungen in Virtual Reality mit generativen neuronalen Netzen

Philipp Ladwig, Alexander Pech, Christian Geiger

Mixed Reality und Visualisierung (MIREVI)
Hochschule Düsseldorf
E-Mail: {philipp.ladwig, geiger}@hs-duesseldorf.de, alexander.pech@study.hs-duesseldorf.de

Abstract: Es existiert derzeit keine Lösung für eine Face-to-Face-Konversation in Virtual Reality (VR), bei der die Teilnehmer Head-mounted-Displays (HMDs) tragen. Die Herausforderung besteht darin, dass das Gesicht der Teilnehmer zu einem signifikanten Teil verdeckt ist und bewährte Technologien zur Gesichtserfassung nicht eingesetzt werden können. Wir stellen ein Teilsystem vor, das auf dem Weg zu dreidimensionalen immersiven Telepräsenzanwendungen das Verdeckungsproblem durch HMDs lösen könnte. Das System nutzt preiswerte Bauteile, eine gängige RGBD-Kamera für die Aufnahme sowie mehrere Verfahren aus dem Bereich der künstlichen Intelligenz. Wir lernen mit einem *Generative Adversarial Network* (GAN) die individuelle Beschaffenheit und Mimik des Gesichtes einer Person und rekonstruieren diese mit VR-fähigen Bildwiederholfrequenzen. Dabei erweitern wir das Konzept der Pix2Pix-GAN-Architektur nach Isola et al. [IZZE17] um die dritte Dimension und Rendern das Ergebnis als texturierte Punktwolke. Wir nennen unser Konzept RGBD-Face-Avatar-GAN und stellen den Code frei zur Verfügung. Der Fokus der Arbeit liegt auf der softwareseitigen Rekonstruktion des Gesichtes, jedoch nicht auf dem Bau oder der Umsetzung eines Sensor-HMDs, das die Mimik einer nutzenden Person erkennt.

Keywords: Generative Adversiral Network, Pix2Pix, Gesichtsrekonstruktion, Telepräsenz, Virtual Reality

Abbildung 1: *Konzeptionelle Idee unseres RGBD-Face-Avatar-GANs: Unser Verfahren besteht aus einem Trainings- und einem Rekonstruktionprozess, in dem ein Neuronales Netz (GAN) die Struktur und Mimik des Gesichtes einer Person lernt. Anhand von Sensoren in einem HMD (nicht Teil dieser Arbeit) könnte das RGBD-Face-Avatar-GAN in Zukunft eine texturierte Punktwolke mit entsprechender Mimik der Person rekonstruieren.*

Veröffentlicht durch die Gesellschaft für Informatik e.V. 2020 in B. Weyers, C. Lürig, D. Zielasko (Hrsg.):
GI VR / AR Workshop 2020, 24. - 25, September 2020, Trier.
Copyright © 2020 bei den Autoren.
http://dx.doi.org/10.18420/vrar2020_15

1 Einleitung

Aktuelle Head-mounted-Displays (HMDs) sind in der Lage, glaubhafte und immersive Erfahrungen zu vermitteln. Dies gilt jedoch nur bedingt für das virtuelle Treffen von realen Personen. Da die Erstellung und das Rendern fotorealistischer Personen in Echtzeit aufwendig sind und heutige HMDs keine adäquaten Sensoren für das Gesichtstracking besitzen, sind authentische Face-to-Face-Konversationen in VR nur schwer bis gar nicht realisierbar. Nur wenige Forschergruppen haben bisher an diesem Problem angesetzt und Verfahren präsentiert, die authentische Face-Avatare für VR erzeugen können [TZS+18, LSSS18]. Diese Ansätze stehen nicht der Öffentlichkeit zur Verfügung und benötigen zum Teil kostspielige Hardware [LSSS18].

Wir präsentieren ein System, das im Vergleich zu anderen mit geringen Hardwarekosten auskommt, moderate Rechenkapazität benötigt und akzeptable Rekonstruktionsergebnisse mit einer Bildwiederholfrequenz von mehr als 90 Hz generiert. Das System ist Open Source: `https://github.com/Alpe6825/RGBD-Face-Avatar-GAN`.

Die Hauptbeiträge dieser Arbeit sind 1.) eine Erweiterung sowie 2.) eine neuartige Anwendung der GAN-Architektur des Pix2Pix-Konzeptes nach Isola et al. [IZZE17].

- Die Erweiterung umfasst den Ausbau der Pix2Pix-GAN-Architektur um einen weiteren Kanal, um nicht nur RGB, sondern auch D (Tiefeninformationen) mit Hilfe einer Punktwolke darstellen zu können.

- Eine neuartige Anwendung des Pix2Pix-Konzeptes für dreidimensionale Telepräsenzanwendungen in VR mit einer Bild- und Punktwolkenrekonstruktion mit mehr als 90Hz.

Das Konzept des Sensor-HMDs, welches die Mimik einer Person erkennen kann, ist in Entwicklung und nicht Teil dieser Arbeit. Die Idee wird jedoch erwähnt, um das Potenzial der Technik aufzuzeigen.

2 Anforderungen

In Bezug auf die Thematik der Telepräsenz für VR werden im Folgenden einige Anforderungen aufgeführt, die verschiedene Entscheidungen innerhalb dieser Arbeit erklären und Einfluss auf die Auswahl der verwandten Arbeiten haben.

Von Natur aus ist eine Face-to-Face-Kommunikation dreidimensional. Eine Unterhaltung beinhaltet in der Regel nicht nur den verbalen Kommunikationskanal, sondern zu einem nicht unerheblichen Teil auch den nonverbalen Kanal. Augenkontakt, Mimik sowie Gesten mit Armen und Händen (Kinesik) und sogar der Körperabstand zueinander (Proxemik) sind wesentliche Informationsträger während eines Gesprächs. Zudem sind Menschen in Konversationen gewohnt, einen gemeinsamen 'Aktionsbereich' zu haben, in dem z. B. durch Zeigen mit dem Finger bestimmte physikalische Objekte referenziert werden können. Die aktuell

wohl immersivsten und gängigsten Mainstream-Telepräsenz-Technologien sind Anwendungen für Videokonferenzen wie z. B. Skype oder Facetime. Diese erlauben zwar das Lesen der Mimik des Gegenübers, jedoch ist z.B. kein ‚echter' Augenkontakt möglich, ausfallende Gesten werden ggf. am Kamerabild abgeschnitten und Körperabstand zwischen den Personen existiert nicht. Im Rahmen dieser Arbeit besteht daher die Anforderung, dass die Darstellung der Personen, und explizit der Gesichter, drei-dimensional statt nur zweidimensional erfolgt.

Menschliche Avatare und die Wahrnehmung solcher wurde in vielfältigen Bereichen untersucht. Ein zentraler Punkt in diesem Kontext ist das Auftreten des Uncanny-Valley-Effekts. Menschen sind ausgesprochen empfindlich für minimale und unnatürliche Abweichungen in Gesichtern. Sobald eine virtuelle Figur in ihrer Erscheinung und Bewegung menschenähnlich ist, jedoch nicht vollkommen der Realität entspricht, wird sie häufig als unsympathisch und unangenehm eingestuft. Im Rahmen eines Telepräsenzsystems gilt es, diesen Effekt zu umgehen. Eine Technik, die in den letzten Jahren das Uncanny Valley erfolgreich überbrückt hat, sind *Deepfakes*. Sie ermöglichen derart authentische Ergebnisse, dass Verfahren und Algorithmen Gegenstand aktueller Forschungen sind, die gefälschte von realen Aufnahmen unterscheiden sollen, da das menschliche Auge nicht mehr dazu in der Lage ist [RCV+19]. Im Rahmen dieser Arbeit werden daher Algorithmen genutzt, die auch für die Erstellung von Deepfakes eingesetzt werden. Wir erweitern diesen Ansatz um eine weitere Dimension (texturierte dreidimensionale Punktwolke statt RGB-Bild), um möglichst realistische Darstellungen von 3D-Face-Avataren zu generieren.

Deepfakes sind rechenintensiv und werden häufig offline berechnet. Für Telepräsenzanwendungen sind jedoch mehrere Bilder pro Sekunde gefordert. Im Bereich von VR ist ein Wert von 90 Bildern pro Sekunde eine gängige Bildwiederholfrequenz und gilt im Rahmen dieser Arbeit als Anforderung an das System.

3 Verwandte Arbeiten

Die Gesichtsrekonstruktion unter einem HMD ist ein junges Forschungsfeld. Olszewski et al. [OLSL16] präsentieren ein System, das mit Hilfe einer RGB-Kamera die Mimik des unteren Gesichtsbereichs auf einen Avatar überträgt. Li et al. [LTO+15] erweiterten diesen Ansatz mit Drucksensoren im Schaumstoff des HMDs, die die Mimik einer Person detailliert erfassen. Die Idee der Verwendung von Sensoren im HMD ist unserem Konzept ähnlich, jedoch nutzen wir personalisierte Avatare, die im Vorfeld trainiert und in einem finalen Schritt synthetisiert werden.

Casa et al. [CFA+16], Früh et al. [FSK17] sowie Thies et al. [TZS+18] nutzten stationäre RGBD-Kameras, um im Vorfeld personalisierte Avatare der Nutzer*innen ohne HMDs zu erstellen. Nachdem die Nutzer*innen ein HMD aufgezogen haben, erkennt dieselbe stationäre RGBD-Kamera die Mimik. Durch die feste Position der Kamera ist der Bewegungsbereich der Nutzer*innen eingeschränkt. Augenbewegungen werden durch Eye-Tracking-Kameras registriert und auf den Avatar übertragen. Alle genannten Systeme verursachen, in unterschiedlichen Stärken, den Uncanny-Valley-Effekt. Um diesen abzuschwächen, entfernten Früh

et al. [FSK17] das HMD nicht vollständig, sondern renderten es als halb-transparentes Objekt. Diese Systeme sind unserem insofern ähnlich, dass sie einen personalisierten Avatar mit Hilfe einer RGBD-Kamera erstellen und annähernd fotorealistische Rekonstruktionen erzeugen.

Das System von Lomardi et al. [LSSS18] erstellt fotorealistische Avatare mit authentischer Mimik. Während das Generieren der personlisierten Avatare mit den Methoden der vorherigen Arbeiten in wenigen Minuten abgeschlossen ist, benötigt das System von Lombardi et al. mehr als einen Tag. In einem aufwendigen Verfahren wird der dreidimensionale Avatar auf Basis einer Vielzahl von Bildern aus unterschiedlichen Perspektiven und Mimiken generiert. Seine finale Version wird im Telepräsenzszenario durch drei RGB-Kameras gesteuert, die am HMD befestigt sind. Eine Schlüsselkomponente dieses Systems ist die Nutzung von *Variational Autoencoders* (VAEs). Es ähnelt unserem Ansatz, der auf Generative Adversarial Networks (GANs) basiert und, abhängig von der Hardware, mit einem Tag eine ähnliche Rechenzeit benötigt. Generell weisen die Architekturen und konzeptionellen Ideen von VAEs und GANs deutliche Ähnlichkeiten auf.

In der Vergangenheit wurde anhand von VAEs und GANs mehrfach bewiesen, dass sie für authentische Gesichtsrekonstruktion geeignet sind und überzeugende Deepfakes generieren können. Da VAEs jedoch häufiger zu unscharfen Resultaten tendieren, nutzen wir GANs [JH19]. Erstmals präsentiert wurde letzteres Konzept von Goodfellow et al. [GPAM$^+$14]. Radfort et al. [RMC15] erweiterten diesen Ansatz um fünf konkrete Aspekte und verbesserten das Konzept nachhaltig. Darüber hinaus erzielten Karras et al. [KALL17] durch das Prinzip des *Progressive Growing GAN* fotorealistische Portraitbilder, die von echten Fotos nicht mehr zu unterscheiden sind. Jedoch besitzt das GAN nach Karras et al. wenig bis keine externe Kontrolle über die Erscheinung des generierten Objektes bzw. Gesichtes, da als Eingabe des Netzes ein latenter Vektor (ein Vektor mit randomisierten Werten) dient, die keine direkte Zuweisung zu einer Eigenschaft wie Haarfarbe, Mimik oder Geschlecht aufweist. Karras et al. ergänzten in einer weiterführenden Arbeit [KLA18] die Architektur des GANs und konnten übergeordnete Attribute (z. B. Pose, Identität) automatisch von stochastischen Variationen (z. B. Sommersprossen, Haar) trennen, jedoch gehörte die Mimik einer Person nicht dazu.

Es hat sich gezeigt, dass Conditional GANs (cGANs) in der Lage sind, konkrete Zusammenhänge zwischen Ein- und Ausgaben zu lernen und zu reproduzieren. So haben Mirza und Osindero [MO14] die Eingabe in Generator und Diskriminator je um ein Label y erweitert, mit dem es möglich ist, Bilder aus einer bestimmten Kategorie y zu erzeugen. Dieser Ansatz der Konditionierung von GANs wurde im Rahmen des *Pix2Pix-GAN* von Isola et al. [IZZE17] aufgegriffen und dahingehend erweitert, dass auf einen Rauschvektor z als Eingabe verzichtet wurde. Ohne Letzteren als Eingangsvektor gibt es keinen latenten Raum Z (da $z \in Z$). Wird der im Rauschvektor enthaltene stochastische Aspekt nicht kompensiert, lernt das GAN nur die Trainingsbeispiele auswendig. Alle Eingaben, die von den Trainingsdaten abweichen, würden zu unzureichenden Ergebnissen führen, wie Isola et al. [IZZE17] beschreiben. Durch die Verwendung einer *U-Net-Architektur* im Pix2Pix-GAN, basierend auf dem *U-Net* von

Ronneberger et al. [RFB15], können in Kombination mit *Dropouts* der stochastische Aspekt sowie der fehlende latente Raum anderweitig in den Generator integriert werden.

Der Diskriminator des Pix2Pix-GANs erhält das selbe Eingangsbild x wie der Generator sowie dessen Ausgangsbild $y_{fake} = G(x)$ bzw. das zu x passende Bild y_{real} aus dem Datensatz. Dies entspricht im Grunde der Idee des cGANs [MO14], bei dem nicht nur das Ergebnis des Generators, sondern auch dessen Unterschied zur Eingabe bewertet wird. Anders als beim cGAN ist die Ausgabe des Diskriminators des Pix2Pix-GANs kein Skalar, das über 'echt' oder 'falsch' entscheidet, sondern eine Matrix. Durch die Verwendung von Convolutional-Layern (vgl. Radford et al. [RMC15]), repräsentiert jeder Eintrag in der Ausgangsmatrix einen $n*m$-großen Bereich des Eingangsbildes. Dadurch ist es möglich, abstrakte Repräsentationen als Matrizen (z.B. Bilder) zur Konditionierung des Netzwerkes zuzulassen, um kontrollierten Einfluss auf die Ausgabe des Generators nehmen zu können. So können z.B. 'Heatmaps' als Eingabe genutzt werden, auf die eine 'konditionierte' und gelernte Ausgabe erfolgt.

4 Systemaufbau

Im Folgenden werden kurz der Ablauf und Aufbau des vorgeschlagenen Systems erklärt, um in den weiteren Abschnitten detaillierter auf die Schritte einzugehen. Unser Verfahren beginnt mit der Aufnahme eines personenbezogenen RGBD-Datensatzes. Anschließend werden die gewonnen Daten durch ein automatisiertes Verfahren normalisiert und in einem nächsten Schritt wird jeweils eine Heatmap pro RGB-Bild extrahiert. Diese Heatmaps beinhalten wesentliche Informationen der Mimik der jeweiligen Person als sogenannte 'Landmarks'. Anschließend wird das GAN, das wir in dieser Arbeit vorschlagen, mit den RGBD-Bildern sowie den zugehörigen Heatmaps trainiert, um die Beziehungen zwischen diesen analysieren zu können. Nach dem Trainingsprozess besteht der finale Schritt aus der Ausführung des Netzes. Dabei erhält der Generator lediglich eine Heatmap (ein entsprechendes Sensor-HMD könnte die Konstruktion der Heatmap durchführen) und erstellt auf Basis dieser ein RGB- und ein D-Bild, die dem Gesicht der 'gelernten' Person entsprechen. Mit Hilfe der RGBD-Daten wird eine Punktwolke des Gesichtes erstellt.

4.1 Trainingsdaten

Der RGBD-Datensatz der jeweiligen Person bildet die Grundlage für den Trainingsprozess. Für die Akquirierung der Daten wird eine RGBD-Kamera (*Mircosoft Azure Kinect*) genutzt, die in fester Position in einer Helmkonstruktion angebracht ist, wie in Abbildung 1 links und Abbildung 2 dargestellt. Die Kamera erfasst die Mimik und speichert die Farbinformationen als 3-Kanal-PNG-Datei (8-Bit-RGB) sowie die Tiefeninformationen als 1-Kanal-PNG-Datei (16-Bit-Graustufen). Während des Trainingsprozesses wird die Person dazu angehalten, eine Vielzahl von verschiedenen Mimiken auszuführen, damit das GAN sie lernen und später rekonstruieren kann. Ein Datensatz einer Person sollte rund 800 bis 1000 RGBD-Bilder beinhalten und die Aufnahme dauert etwa 10 bis 15 Minuten. Nach der Akquirierung der

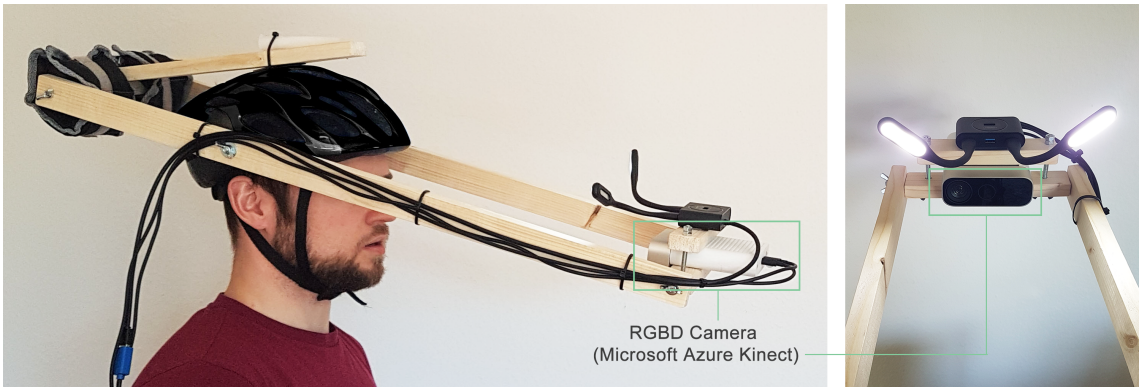

Abbildung 2: *Helm-Konstruktion zur RGBD-Datenaufnahme. Mit dieser Konstruktion fließen Kopfdrehungen nicht mit in den Traningsdatensatz ein, denn sie sind für das System nicht relevant, da Kopfdrehungen im Telepräsenz-Szenario durch das Tracking des HMDs realisiert werden. Dadurch wird die Trainingszeit des GANs verkürzt und die Qualität der Ausgabe verbessert, da weniger Informationen in das Modell einfließen müssen. Der Materialpreis der Helmkonstruktion ohne RGBD-Kamera liegt bei etwa 55 Euro.*

Daten müssen diese angepasst und normalisiert werden. Kontraste lassen sich verschärfen, indem das Histogramm normalisiert wird.

4.2 Bestimmung von Facial Landmarks

Um die Ausgabe des GANs und somit die Mimik des Gesichtes einer Person steuern zu können, muss der Datensatz vor dem Trainingsprozess gelabelt werden. Als Label dienen *Facial Landmarks*, die als zweidimensionale Schwarz-Weiß-Heatmap für jedes Tupel von RGB- und korrespondierendem Tiefenbild im Datensatz bestimmt wird, wie in den Abbildungen 3 bis 6 zu sehen ist. Die Position der Facial Landmarks identifiziert in jedem Bild des Trainingsdatensatzes die Mimik der Person und wird bei der Ausführung des GAN-Modells als Eingabe genutzt. Für die Bestimmung der Facial Landmarks wird das *Face Alignment Network* (FAN) von Bulat und Tzimiropoulos [BT17] verwendet. Dies ist ein *Convolutional Neural Network* (CNN), das in der Lage ist, 2D Landmarks auf Basis eines RGB-Bildes zu bestimmen.

Die Blickrichtung der Person innerhalb eines Bildes kann das FAN hingegen nicht bestimmen. Diese Information ist jedoch essentiell, da anderenfalls das GAN in der Rekonstruktionsphase nicht in der Lage ist, die Blickrichtung einer Person korrekt wiederzugeben. Daher wurden, zusätzlich zu den 68 generierten Facial Landmarks des FANs, zwei weitere Landmarks auf Basis eines Eye-Tracking-Verfahrens implementiert. Mit Hilfe von OpenCV werden die Positionen der Pupillen erfasst und als weitere Landmarks zur Heatmap konkateniert.

4.3 Netz-Architektur

Vorherige Arbeiten mit neuronalen Netzen, wie z. B. Wu et al. [WZX+16] haben gezeigt, dass ein Voxel-basierter Ansatz mit hohen Trainings- und Ausführungszeiten des Modells einhergeht. Daher wurde gezielt eine RGBD-basierte Lösung angestrebt. Der Vorteil liegt in

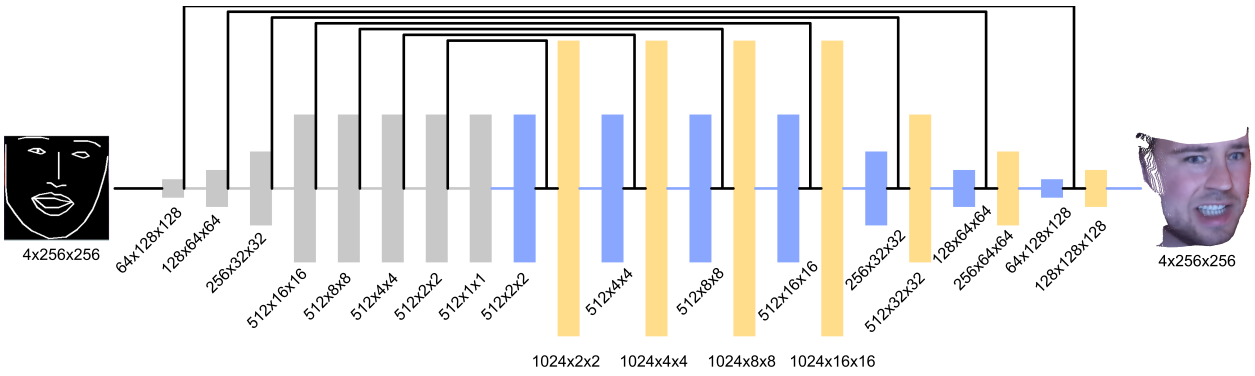

Abbildung 3: *Generator des RGBD-Face-Avatar-GANs.*

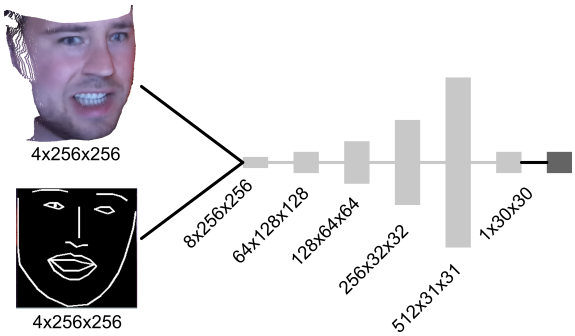

Abbildung 4: *Diskriminator des RGBD-Face-Avatar-GANs.*

der kompakten Repräsentation der Daten als Punktwolke und der Möglichkeit der Adaption von bisherigen RGB-basierten Verfahren. Für die grundlegende Netzwerk-Architektur eignet sich das *Pix2Pix-GAN* nach Isola et al. [IZZE17], wie bereits oben erwähnt. Dieses ist in der Lage, auf Basis von zuvor trainierten Heatmaps entsprechende RGB-Bilder zu generieren. Der Generator des RGBD-Face-Avatar-GANs erhält als Eingabe eine 256 x 256 Pixel große Heatmap der Facial Landmarks, wie in Abbildung 3 gezeigt. Gegenüber des Pix2Pix-GANs wurde die Ausgabe um eine vierte Featuremap erweitert, um RGBD-Bilder erzeugen zu können. Zudem erhält der Diskriminator als Eingabe in Summe acht Featuremaps statt nur drei. Die ersten vier entsprechen den vier Kanälen des RGBD-Bildes, die restlichen vier enthalten jeweils die korrespondierende Landmark-Heatmap, wie in Abbildung 5 gezeigt. Auch wenn die Heatmap durch ein Graustufenbild repräsentiert wird und prinzipiell nur eine Featuremap benötigt, wird sie vierfach in den Diskriminator gegeben, um den Trainingsprozess mit einem ausgeglichenen Verhältnis zwischen RGBD-Bild und Landmarks zu beginnen (vgl. Abbildung 5). Die Architektur des Diskriminators ist in Abbildung 4 dargestellt.

4.4 Trainingsprozess

Vor dem Trainingsprozess werden alle Gewichtungen des RGBD-Face-Avatar-GANs mit einem zufälligen Wert initialisiert. Die Zufallswerte folgen dabei einer Gaußverteilung mit Erwartungswert 0 und Standardabweichung 0.2. Die Batchsize der Eingabedaten in das GAN beträgt 1. Generator und Diskriminator werden je mit einer Lernrate von 0.0002 trainiert,

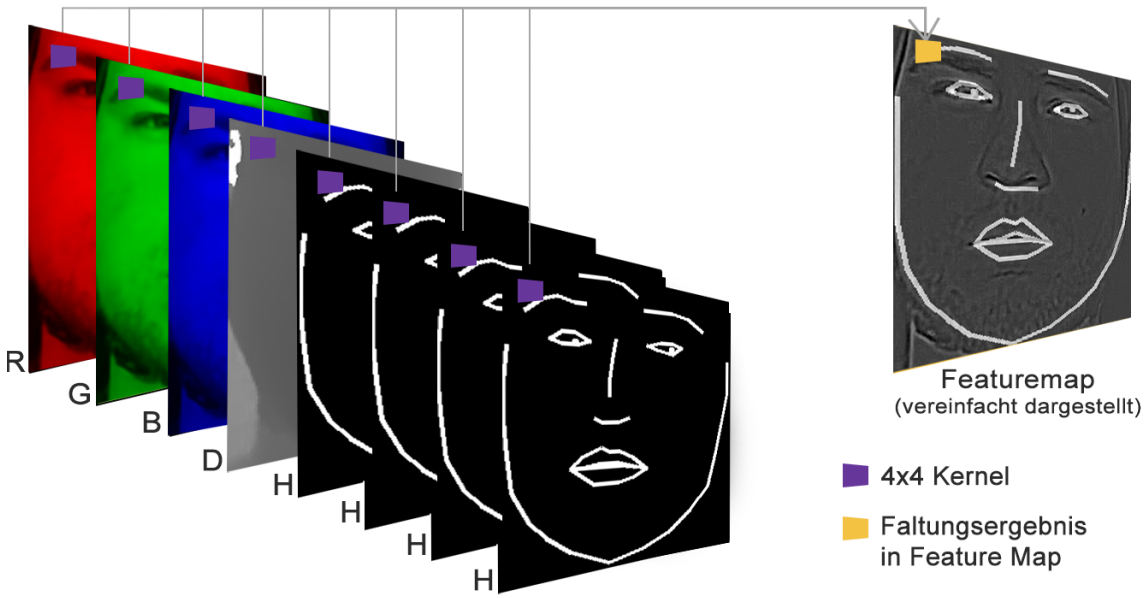

Abbildung 5: *Beispiel Faltung für Diskriminator-Input. Die Kanäle RGBD werden einzeln gewichtet, während die Heatmap (H) vierfach gewichtet wird, damit Generator und Diskriminator mit einem ausgeglichenen Verhältnis den Lernprozess beginnen.*

wobei beim Diskriminator gilt: $Loss_D = (Loss_{Dreal} + Loss_{Dfake}) * 0.5$, was die Lernrate für den Diskriminator reduziert, wodurch dieser im Verhältnis zum Generator langsamer lernt. Dies ist notwendig, da der Diskriminator zu Beginn der Trainingsphase seine Aufgabe mühelos bewältigen kann. Lernt der Diskriminator zu schnell, hat der Generator keine Chance, das Erstellen der gewünschten Bilder zu erlernen.

5 Resultate

Mehrere iterative Testdurchläufe, verbesserte Datennormalisierungen und Anpassungen der Landmarks waren nötig, um authentische Ergebnisse zu erzielen, die in Abbildung 6 aufgeführt sind. Dies sind Ergebnisse, welche die Grundlage für eine Gesichtsrekonstruktion in VR bilden. Anstatt mit einem Sensor-HMD wurden die Landmarks mit einer Webcam erzeugt.

Die Identität der Person sowie auch die persönlichen Züge der Mimik sind gut wiederzuerkennen. Die Grauwertbilder in der 4. Spalte zeigen die Unterschiede zwischen Eingabebild (Webcam-Bild) und vom GAN generiertem Ausgabebild und wurden mit Hilfe der *Strukturellen Ähnlichkeit* (SSIM) nach Wang et al. [ZBSS04] berechnet. Dieser Wert wird für gewöhnlich im Bereich der Quantifizierung der Qualität von Bildkompressionsverfahren genutzt. Ein SSIM-Wert von 1 sagt aus, dass die Vergleichsbilder identisch sind. Da unser GAN ähnlich wie ein Encoder und ein Decoder arbeitet ist der SSIM ein geeignetes Werkzeug, um eine Aussage über die Rekonstruktionsqualität zu treffen. Obwohl die Beleuchtung und der Hintergrund zwischen Eingabe- und Ausgabebild zum Teil starke Unterschiede aufweisen, und dadurch das SSIM-Ergebnis negativ beinflussen, besitzen die Bilder trotzdem eine hohe Ähnlichkeiten im Gesichtsbereich.

Abbildung 6: *Resultate: Die 1. Spalte zeigt RGB-Bilder, die mit einer Webcam aufgenommen wurden. Die 2. Spalte zeigt die entsprechenden Landmarks bzw. Heatmaps, die auf Basis der Bilder aus der 1. Spalte gewonnen wurden. Auf Basis der Heatmap ist das GAN in der Lage, die Ausgaben in der 3. sowie 5. und 6. Spalte zu produzieren. Die Eingabebilder von Zeile A bis E sind nicht Teil des Traningsdatensatzes gewesen und sind daher unbekannte Bilder für das GAN. Die Grauwertbilder in der 4. Spalte zeigen die Ergebnisse der Berechnungen der Strukturellen Ähnlichkeit SSIM [ZBSS04] vom Eingabebild (1. Spalte) im Vergleich zum generierten Bild (3. Spalte). Umso dunkler ein Pixel in Spalte 4 ist, desto größer ist die Differenz zwischen den zwei Bildern. Zeile E zeigt eine Mimik, welche nicht im Trainigsdatensatz vorhanden ist. Zeile F zeigt ein Bild, welches Teil des Trainigsdatensatzes war.*

Negativ fällt auf, dass in den Ausgabebildern ein Verlust an Schärfe zu verzeichnen ist und dass die Rekonstruktion der Augen teilweise fehlerhaft ist. Während der erste Punkt dem GAN zugesprochen werden kann, und durch Parametereinstellungen während des Training beeinflussbar ist, ist der zweite Punkt auf die zum Teil fehlerhafte Blickrichtungserkennung in den Eingabebildern zurück zuführen, welche die Landmarks der Iris gelegentlich außerhalb der Augenkonturen positioniert, wie in Zeile 2 zuerkennen. Präziseres Eye-Tracking würde bessere Ergebnisse liefern.

Die Rekonstruktion der Tiefe ist in Spalte 5 und 6 als Punktwolken dargestellt. Der Detailgrad der Tiefeinformationen ist vergleichbar mit dem der generierten RGB-Bilder. Jedoch weicht die generierte Tiefe um einen unbekannten Skalierungsfaktor ab, welcher die Punktwolke auf der Tiefenachse, im Vergleich zur Realität, verzerrt wiedergibt. Dieser Umstand erlaubt aktuell keine eins-zu-eins Darstellung der Tiefe und lässt sich vermutlich auf einen Fehler während der Normalisierung, Encoding, Decoding oder eigentlichen Rekonstruktion der Werte zurückführen.

Eine Limitierung unseres GANs ist das Unvermögen, vorher nicht trainierte Mimiken zu rekonstruieren. Dies schlägt sich als Bildartefakte nieder, wie sie in Zeile E zuerkennen sind. Erhält das GAN hingegen ein RGB-Bild, welches Teil des trainierten Datensatzen war, wird dieses überdurchschnittlich gut rekonstruiert, wie man es an dem Differenzbild und hohem SSIM-Wert in Zelle F4 erkennen kann.

Die dargestellten Resultate wurden mit 868 RGBD-Bildern über einen Zeitraum von 22 Stunden mit einer Nvidia RTX 2080Ti trainiert. Für die Implementierung und das Training wurde PyTorch 1.5 genutzt. Um die Ausführungsgeschwindigkeit zu optimieren, wurde das Netz in einer C++-Umgebung portiert (LibTorch). Die durchschnittliche Ausführungsgeschwindigkeit der GPU auf einem System mit Intel i9-9880H und Nvidia RTX 2080 liegt unter 9 ms (111 Hz) und ist damit kompatibel für VR-Anwendungen. Innerhalb der 9 ms wird die Heatmap von dem GAN eingelesen und produziert eine RGB-Textur sowie eine Punktwolke mit bis zu 65535 Punkten. Weitere Resultate werden in folgendem Video präsentiert: `https://youtu.be/Z6hsRRjtIN8`

6 Zusammenfassung und zukünftige Arbeiten

Es wurde das RGBD-Face-Avatar-GAN vorgestellt, das die Struktur und Mimik einer Person lernen und als texturierte Punktwolke mit VR-fähigen Bildwiederholfrequenzen reproduzieren kann. Wir haben die Architektur des Pix2Pix-GANs [IZZE17] um eine dritte Dimension erweitert und präsentieren einen weiteren Anwendungsfall für das Pix2Pix-Konzept. Facial-Landmark-Heatmaps dienen zur Konditionierung des Generators, was die Steuerung der Mimik des synthetisierten Gesichtes ermöglicht. Die Nutzung solcher Heatmaps erlaubt eine abstrakte bzw. codierte Darstellung der Mimik einer Person. In Zukunft können die Heatmaps potenziell von Sensoren innerhalb eines HMDs erfasst werden und diese steuern. Die Resultate der Rekonstruktion befinden sich teilweise im Uncanny Valley, überzeugen dennoch mit einer authentischen Wiedergabe der Identität und persönlichen Mimik der Person. Da-

her ist ein wichtiger Teil der zukünftigen Arbeit die qualitative Verbesserung der generierten RGBD-Daten, um das Uncanny Valley vollständig zu verlassen.

Danksagung

Wir danken der MIREVI-Gruppe und dem 'Promotionszentrum Angewandte Informatik' (PZAI) in Hessen, insbesondere Ralf Dörner. Diese Arbeit wird gesponsert von: Deutsches Bundesministerium für Bildung und Forschung (BMBF) unter den Projektnummern 16SV8182 und 13FH022IX6. Namen der Projekte: HIVE-Lab (Health Immersive Virtual Environment Lab) und iKPT 4.0 (Interaktive körpernahe Produktionstechnik 4.0).

Literatur

[BT17] Adrian Bulat and Georgios Tzimiropoulos. How Far are We from Solving the 2D 3D Face Alignment Problem? *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[CFA+16] Dan Casas, Andrew Feng, Oleg Alexander, Graham Fyffe, Paul Debevec, Ryosuke Ichikari, Hao Li, Kyle Olszewski, Evan Suma, and Ari Shapiro. Rapid Photorealistic Blendshape Modeling from RGB-D Sensors. In *Computer Animation and Social Agents*, CASA '16, 2016.

[FSK17] Christian Frueh, Avneesh Sud, and Vivek Kwatra. Headset Removal for Virtual and Mixed Reality. In *ACM SIGGRAPH 2017 Talks*, SIGGRAPH '17, 2017.

[GPAM+14] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Networks. *Advances in neural information processing systems*, 2014.

[IZZE17] P. Isola, J. Zhu, T. Zhou, and A. A. Efros. Image-to-Image Translation with Conditional Adversarial Networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[JH19] Greg Walters John Hany. *Hands-On Generative Adversarial Networks with PyTorch 1.x*. Packt Publishing Ltd., December 2019.

[KALL17] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive Growing of GANs for Improved Quality, Stability, and Variation, 2017. Zugriff am 17.07.2020 auf `https://arxiv.org/abs/1710.10196`.

[KLA18] Tero Karras, Samuli Laine, and Timo Aila. A Style-Based Generator Architecture for Generative Adversarial Networks, 2018. Zugriff am 17.07.2020 auf `https://arxiv.org/abs/1812.04948`.

[LSSS18] Stephen Lombardi, Jason Saragih, Tomas Simon, and Yaser Sheikh. Deep Appearance Models for Face Rendering. *ACM Trans. Graph.*, 37(4), July 2018.

[LTO+15] 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 Trans. Graph.*, 34(4), July 2015.

[MO14] Mehdi Mirza and Simon Osindero. Conditional Generative Adversarial Nets, 2014. Zugriff am 17.07.2020 auf `https://arxiv.org/abs/1411.1784`.

[OLSL16] Kyle Olszewski, Joseph J. Lim, Shunsuke Saito, and Hao Li. High-Fidelity Facial and Speech Animation for VR HMDs. *ACM Trans. Graph.*, 35(6), November 2016.

[RCV+19] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Niessner. FaceForensics++: Learning to Detect Manipulated Facial Images. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[RFB15] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer - Assisted Intervention - MICCAI*, 2015.

[RMC15] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks, 2015. Zugriff am 17.07.2020 auf `https://arxiv.org/abs/1511.06434`.

[TZS+18] Justus Thies, Michael Zollhöfer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. FaceVR: Real-Time Gaze-Aware Facial Reenactment in Virtual Reality. *ACM Trans. Graph.*, 37(2), June 2018.

[WZX+16] Jiajun Wu, Chengkai Zhang, Tianfan Xue, William T Freeman, and Joshua B Tenenbaum. Learning a Probabilistic Latent Space of Object Shapes via 3D Generative-Adversarial Modeling. In *Advances in Neural Information Processing Systems*, 2016.

[ZBSS04] Zhou Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 2004.