

Bachelorarbeit

# **3D-Gaussian Splatting für die virtuelle Filmproduktion im visuellen Vergleich mit eigenständig produzierten Filmsequenzen**

Abschlussarbeit im Studiengang Medientechnik (B.Sc.)

verfasst von Simon Budde ( )

vorgelegt am 11. Mai 2026

Erstprüfer: Prof. Dr.-Ing. Marco Grimm

Zweitprüferin: Christina Becker

**HOCHSCHULE FÜR ANGEWANDTE  
WISSENSCHAFTEN HAMBURG**

Fakultät Elektro-, Medien- und Informationstechnik

Kunst- und Medien-campus

Finkenau 35

22081 Hamburg

# Zusammenfassung

Bildbasierte 3D-Rekonstruktionsverfahren sind in der Lage, aus einer Reihe digitaler Fotografien eines Objekts oder einer Umgebung dreidimensionale Computermodelle zu berechnen. Die im Jahr 2023 publizierte Forschungsarbeit „3D Gaussian Splatting for Real-Time Radiance Field Rendering“ nutzt einen effizienten Ansatz in der Repräsentation und Bildsynthese, um erstmals fotorealistische Modelle mit Echtzeit-Wiedergabe zu generieren. Aufgrund der hohen visuellen Qualität der Rekonstruktionen ergeben sich potenzielle Anwendungsmöglichkeiten in der Filmproduktion, insbesondere für die Erstellung virtueller Sets im Kontext von Virtual Production. In dieser Arbeit wird anhand einer Parameterstudie mit objektiven Vergleichskriterien und einer wahrnehmungsbasierten Gegenüberstellung systematisch untersucht, welche Faktoren im Prozess der Modellgenerierung einen Einfluss auf die visuelle Qualität haben und unter welchen Bedingungen die besten Rekonstruktionsergebnisse erzielt werden können. Daran schließt sich die Ausgangsfrage an, ob „Gaussian Splatting“ dem visuellen Anspruch an Bewegtbildformate gerecht werden und als Werkzeug zur Generierung digitaler Szenen für die virtuelle Filmproduktion genutzt werden kann. In einem Modellversuch werden Filmsequenzen einer U-Bahn-Station sowohl vor Ort als auch virtuell produziert und miteinander verglichen. Zudem wird eine Person in einem virtuellen Studio-Setup mit Green-Screen-Keying in die digitale Szene eingefügt. In der Auswertung zeigt sich, dass „Gaussian Splatting“ unter kontrollierten Bedingungen und mit dem erarbeiteten Workflow Rekonstruktionsergebnisse erzielen kann, die zum Großteil als fotorealistisch einzustufen sind. Die virtuell produzierten Aufnahmen im Modellversuch weisen zwar noch Probleme bei der visuellen Darstellung auf, zeigen aber auf anschauliche Weise, dass sich „Gaussian Splatting“ zur Erstellung digitaler Szenen für die virtuelle Filmproduktion gewinnbringend einsetzen lässt.

## Abstract

Image-based 3D reconstruction methods can generate three-dimensional computer models from a series of digital photographs of an object or environment. The research paper “3D Gaussian Splatting for Real-Time Radiance Field Rendering,” published in 2023, employs an efficient image synthesis approach to generate photorealistic models with real-time rendering. Due to the high visual quality of the reconstructions, there is potential for applications in film production, particularly for creating virtual sets in virtual production contexts. In this thesis, a parameter study with objective comparison criteria and perception-based comparisons are used to systematically investigate which factors in the model generation process influence visual quality, and under which conditions the best reconstruction results can be achieved. This leads to the central question of whether 'Gaussian Splatting' meets the visual demands of moving image formats and can be used to generate digital scenes for virtual film production. In a pilot study, film sequences of a subway station were produced both on location and virtually, and the results were compared. Additionally, a person is inserted into the digital scene in a virtual studio setup using green-screen keying. Evaluation shows that under controlled conditions and using the developed workflow, 'Gaussian Splatting' can produce reconstructions that are largely

classified as photorealistic. While the virtually produced footage in the pilot study still exhibits issues with visual representation, it clearly demonstrates that 'Gaussian Splatting' can be used effectively to create digital scenes for virtual film production.

## Danksagung

Ich möchte an dieser Stelle allen danken, die mich im Kontext dieser Arbeit während des gesamten Prozesses von der Ideenfindung über die Planung bis zur Umsetzung unterstützt haben.

Ein großer Dank gilt Marco Grimm, der mich als Erstprüfer in einem Umfang unterstützte, der weit über das hinausgeht, was man von einer betreuenden Lehrperson erwarten kann. Seine wertvollen Impulse in der Themenfindung und Konzeption sowie die tatkräftige Unterstützung beim VP-Projekt im Produktionslabor tragen zu einem Gesamtergebnis bei, das unter anderen Umständen nicht möglich gewesen wäre.

Mein Dank gilt auch Christina Becker, die mir als Zweitprüferin in organisatorischen und konzeptionellen Fragen stets zur Seite stand. Außerdem möchte ich Thinh Nguyen für seine umfangreiche Beratung im Aufbau des Virtual-Production-Setups danken.

# Inhaltsverzeichnis

|       |                                                                       |    |
|-------|-----------------------------------------------------------------------|----|
| I.    | ABBILDUNGSVERZEICHNIS .....                                           | 5  |
| 1     | EINLEITUNG .....                                                      | 6  |
| 1.1   | ZIEL DER ARBEIT .....                                                 | 6  |
| 1.2   | MOTIVATION .....                                                      | 6  |
| 2     | THEORETISCHE GRUNDLAGEN .....                                         | 7  |
| 2.1   | DISZIPLINÄRE EINORDNUNG .....                                         | 7  |
| 2.2   | ENTWICKLUNG BILDBASIERTER 3D-REKONSTRUKTIONSVERFAHREN .....           | 7  |
| 2.2.1 | Prinzipien der Photogrammetrie .....                                  | 7  |
| 2.2.2 | Structure-from-Motion .....                                           | 10 |
| 2.2.3 | Dreidimensionale Rekonstruktion mit Photogrammetrie .....             | 12 |
| 2.3   | DIE GRUNDLAGEN VON 3D-GAUSSIAN SPLATTING .....                        | 15 |
| 2.4   | LiDAR-SCANNING .....                                                  | 17 |
| 3     | SZENENREKONSTRUKTION MIT GAUSSIAN SPLATTING .....                     | 18 |
| 3.1   | VORBEREITENDE ÜBERLEGUNGEN .....                                      | 18 |
| 3.2   | VERWENDETES EQUIPMENT & SOFTWARE .....                                | 19 |
| 3.3   | TEST- UND OPTIMIERUNGSPHASE .....                                     | 20 |
| 3.3.1 | Aufbau der Test-Szene .....                                           | 21 |
| 3.3.2 | Capture-Prozess und RAW-Entwicklung .....                             | 22 |
| 3.3.3 | Methodik der Parameterstudie .....                                    | 23 |
| 3.3.4 | Auswahl der Einflussfaktoren .....                                    | 27 |
| 3.3.5 | Objektiver Vergleich mit PSNR, SSIM und VMAF .....                    | 30 |
| 3.3.6 | Subjektiver Vergleich der visuell wahrgenommenen Modellqualität ..... | 32 |
| 3.4   | DISKUSSION UND AUSWERTUNG DER ERGEBNISSE .....                        | 38 |
| 3.5   | ZWISCHENFAZIT .....                                                   | 42 |
| 4     | VERGLEICH VIRTUELL UND REAL PRODUZierter FILMSEQUENZEN .....          | 43 |
| 4.1   | REKONSTRUKTION EINER SZENE FÜR DIE VIRTUELLE PRODUKTION .....         | 43 |
| 4.2   | KONZEPT UND AUFBAU DES VIRTUELLEN STUDIO-SETUPS .....                 | 46 |
| 4.3   | PRODUKTION VIRTUELLER FILMSEQUENZEN .....                             | 47 |
| 4.4   | VERGLEICH DER VIRTUELLEN UND VOR ORT PRODUZIERTEN FILMSEQUENZEN ..... | 48 |
| 5     | FAZIT .....                                                           | 51 |
| 6     | LITERATURVERZEICHNIS .....                                            | 52 |
| 7     | ANHANG .....                                                          | 54 |
| 8     | VERWENDETE HILFSMITTEL .....                                          | 54 |

# I. Abbildungsverzeichnis

|                                                                                                                                                   |    |
|---------------------------------------------------------------------------------------------------------------------------------------------------|----|
| Abbildung 1: Tiefenberechnung durch Triangulation.....                                                                                            | 8  |
| Abbildung 2: Rekonstruktionsergebnisse des "Shape-and-Motion"-Algorithmus .....                                                                   | 10 |
| Abbildung 3: Programmablaufplan COLMAP.....                                                                                                       | 12 |
| Abbildung 4: Der Trainingsprozess von 3D-Gaussian Splatting.....                                                                                  | 16 |
| Abbildung 5: Beispielfoto aus dem Bilddatensatz (DSCF0858.jpg) .....                                                                              | 22 |
| Abbildung 6: Kamerapositionen in COLMAP.....                                                                                                      | 22 |
| Abbildung 7: Block-Diagramm der SSIM-Berechnung .....                                                                                             | 25 |
| Abbildung 8: Kamerafahrt des gerenderten Vergleichsvideos .....                                                                                   | 26 |
| Abbildung 9: Structure-from-Motion mit 16 Winkeln (113 Fotos), 8 Winkeln (57 Fotos) und 4 Winkeln (29 Fotos) .....                                | 27 |
| Abbildung 10: Extraktion eines Bilddatensatzes mit kleinerem Sichtfeld.....                                                                       | 28 |
| Abbildung 11: Exemplarische Bildausschnitte der gerenderten Videosequenzen (oben: FoV $\approx 60^\circ$ , unten: FoV $\approx 100^\circ$ ) ..... | 33 |
| Abbildung 12: Verschiedene SfM-Initialisierungen, v.l.n.r. COLMAP, Postshot, Postshot mit „Single Lens & Focal Length“ .....                      | 34 |
| Abbildung 13: Einfluss der Anzahl an Trainingsiterationen auf die Modellqualität.....                                                             | 36 |
| Abbildung 14: Anordnung der Säulen von oben (links), Anordnung der Kamerapositionen als Säule (rechts) .....                                      | 45 |
| Abbildung 15: Green-Screen-Setup im Produktionslabor (links), Kamera-Setup mit Tracking-System (rechts).....                                      | 46 |
| Abbildung 16: Standbild der Studiokamera mit Hervorhebung des Green Spill .....                                                                   | 48 |
| Abbildung 17: Notausgangsschild im Vergleich (links: virtuelles Bild, rechts: Kamerabild) .....                                                   | 49 |
| Abbildung 18: Reflexionen im Vergleich (links: virtuelles Bild, rechts: Kamerabild).....                                                          | 49 |
| Abbildung 19: Blick über den Bahnsteig (links: virtuelles Bild, rechts: Kamerabild) .....                                                         | 50 |

# 1 Einleitung

## 1.1 Ziel der Arbeit

„3D-Gaussian Splatting“ bietet im Kontext der bildbasierten Rekonstruktionsverfahren zur Erstellung dreidimensionaler Modelle von realen Objekten und Umgebungen für viele Anwendungsbereiche eine vielversprechende Alternative zu den konventionellen Photogrammetrie-Workflows der letzten Jahrzehnte. Es ist das erste Verfahren, das visuell hochqualitative 3D-Modelle in Echtzeit darstellen kann und erfreut sich daher seit Veröffentlichung des Papers im Jahr 2023 größer werdender Beliebtheit. Diese Arbeit soll die Frage beantworten, ob Gaussian Splatting in der Erstellung virtueller Sets für die Filmproduktion eingesetzt werden kann. Dafür wird zunächst untersucht, welche Werkzeuge, Techniken und Parameter zielführend sind, um ein möglichst fotorealistisches 3D-Modell eines real existierenden Ortes mit 3D-GS zu erstellen. Unter visuellen Gesichtspunkten sollen in einem Modellversuch virtuell produzierte Filmsequenzen beurteilt und mit vor Ort realisierten Aufnahmen verglichen werden.

## 1.2 Motivation

Die Filmproduktion konnte durch die Digitalisierung der audiovisuellen Inhalte eine enorme Effizienzsteigerung verzeichnen, die darauf ausgerichtet ist, die Produktionsabläufe durch den Einsatz von Computern schneller, flexibler und kostengünstiger zu gestalten. Mit der Entwicklung virtueller Produktionsmethoden eröffneten sich in den letzten zehn Jahren Möglichkeiten, audiovisuelle Inhalte auf eine neue Weise unter Studio-Bedingungen zu produzieren. Statt eines realen Sets können Computermodele mit einer Visualisierungstechnik wie dem Green-Screen-Keying oder einer LED-Wand den Hintergrund einer Szene darstellen. Möchte man ein real existierendes Set in den virtuellen Workflow einbinden, konnte bisher nur ein digitaler Nachbau in einem 3D-Programm zu den gewünschten Ergebnissen führen. Mit der Publikation von Gaussian Splatting wurde ein Verfahren vorgestellt, das reale Umgebungen in frei begehbare 3D-Modelle umwandeln kann, die eine hohe visuelle Qualität besitzen und gleichzeitig in einem überschaubaren Zeitrahmen mit einfachen Mitteln generiert werden können. Diese Entwicklung hat das Potenzial, der Filmproduktion als wertvolles Werkzeug in der Inhaltserstellung zu dienen und insbesondere im Feld der virtuellen Produktion aber auch im VFX-Bereich fortschrittliche und effiziente Lösungen zu ermöglichen.

## 2 Theoretische Grundlagen

### 2.1 Disziplinäre Einordnung

Den Kern dieser Arbeit bildet ein Verfahren zur digitalen Rekonstruktion und Darstellung einer real existierenden Umgebung anhand von Fotos. Die Transformation von zweidimensionalen Bildern in ein dreidimensionales Computermodell ist der erste Abschnitt des Prozesses und wird allgemein als 3D-Rekonstruktion bezeichnet. Die Methode ist ein Teilgebiet der Disziplin „Computer Vision“, welche sich mit der maschinellen Analyse und Interpretation von Bilddaten auseinandersetzt. Im zweiten Schritt wird der zu projizierende Ausschnitt des dreidimensionalen Modells in eine zweidimensionale Darstellung umgewandelt, damit er als digitales Bild ausgegeben werden kann. Dieser Prozess fällt in den Bereich des volumetrischen Renderings innerhalb der „Computergrafik“.

Das Verfahren hinter 3D-Gaussian-Splatting adressiert Problemstellungen, die vorwiegend zwei Disziplinen des „Visual Computing“ zuzuordnen sind. Neben „Computergrafik“ und „Computer Vision“ zählen darüber hinaus auch „Visualisierung“ und „Virtuelle Realität“ zu den Inhalten des interdisziplinären Forschungsfeldes. Mit der erstmaligen Durchführung des „International Symposium on Visual Computing“ (ISVC) im Jahr 2005 in Nevada, USA, etablierte sich die Erfassung, Verarbeitung, Modifizierung und Darstellung visueller Daten als eigenständiges Teilgebiet der Informatik (Computer Science).

### 2.2 Entwicklung bildbasierter 3D-Rekonstruktionsverfahren

#### 2.2.1 Prinzipien der Photogrammetrie

Photogrammetrie im ursprünglichen Sinn ist eine messtechnische Methode, mit der sich aus fotografischen Aufnahmen anhand geometrischer Berechnungen quantitative Informationen über die Position und Form von Objekten gewinnen lassen. Der Begriff wird erstmalig im Jahr 1867 in einem Artikel des „Wochenblatts des Architektenvereins zu Berlin“ erwähnt<sup>1</sup>. Der Verfasser entwickelt ein photogrammetrisches Verfahren zur Gebäudevermessung, ein deutscher Architekt namens Albrecht Meydenbauer. Seine und die Arbeit weiterer Pioniere im Bereich der Messbildverfahren legen den Grundstein für die bildbasierte Rekonstruktion dreidimensionaler Eigenschaften.

Die Augen sind das visuelle Werkzeug zur Wahrnehmung der dreidimensionalen Welt. Die Voraussetzung dafür, Räumlichkeit wahrnehmungspsychologisch erfassen zu können, ist die binokulare Sicht. Das menschliche Gehirn verarbeitet die optischen Reize beider Augen zu einem visuellen Eindruck, der die Umgebung nicht mehr nur als Ebene

---

<sup>1</sup> (Meydenbauer, 1867)

interpretiert. In der Stereophotogrammetrie wird dieses Prinzip technisch nachgebildet, indem aus zwei Fotografien mit sich überschneidenden Bildinhalten Tiefeninformationen geometrisch berechnet werden. Das mathematische Fundament dieses Verfahrens bildet die Triangulationsfunktion.

Die untenstehende Abbildung soll anhand eines stark vereinfachten Beispiels zeigen, wie die Tiefeninformation eines Punktes im Raum mit der Triangulation berechnet wird. In (a) sind zwei Bildebenen zu sehen, in denen jeweils eine Projektion des Objektpunktes  $P$  mit den Koordinaten  $[x_L, y_L]$  und  $[x_R, y_R]$  dargestellt ist. Weil die Translation  $T$ , also die Verschiebung der Kameraposition, in diesem Fall nur entlang der x-Achse stattfindet, kann für die Berechnung von  $Z$ , also dem Abstand des Punktes  $P$  zu den Projektionszentren der Kamerabilder, die Draufsicht (b) gewählt werden. Aufgrund der geradlinigen Ausbreitung des Lichts liegen die Bildpunkte auf dem Strahl (Ray) zwischen Objektpunkt und jeweiligem Projektionszentrum. Dadurch lässt sich das Verhältnis der Grundseite  $T$  zur Höhe  $Z$  des orangenen Dreiecks mit dem Verhältnis der Grundseite zwischen den Bildpunkten zur Höhe  $Z$  abzüglich der Brennweite  $f$  des gelben Dreiecks gleichsetzen (siehe Formel 1). Löst man die Gleichung nach  $Z$  auf, erhält man die Distanz vom Objektpunkt zur Kameraposition (siehe Formel 2).

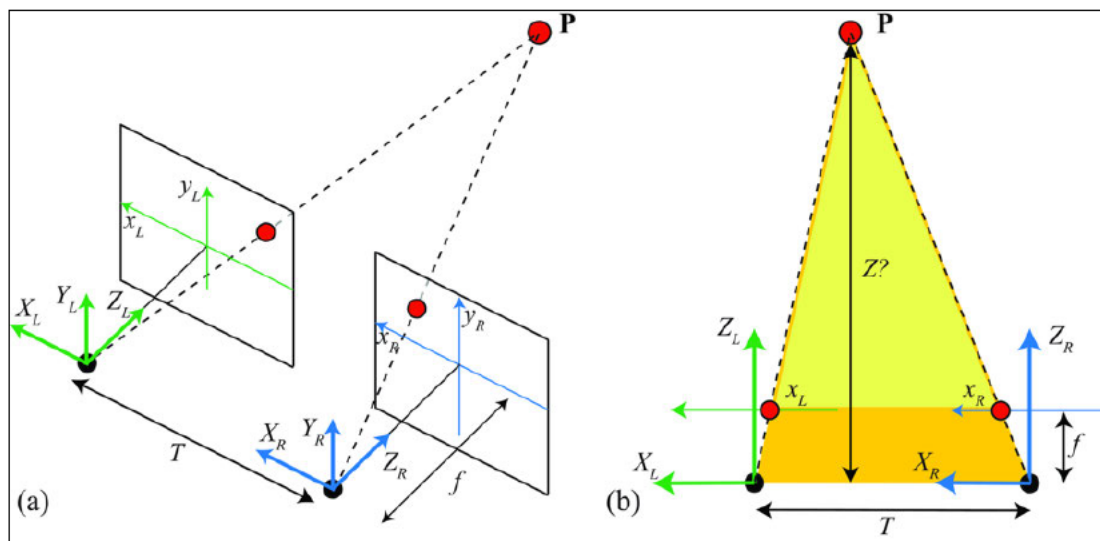


Abbildung 1: Tiefenberechnung durch Triangulation

$$\frac{T}{Z} = \frac{T - x_L + x_R}{Z - f} \quad \equiv \quad Z = \frac{f * T}{x_L - x_R}$$

Die Differenz der Bildpunktpositionen, in diesem Fall in horizontaler Ebene zwischen  $x_L$  und  $x_R$ , gibt Aufschluss über die Tiefe des Objektpunktes und verhält sich antiproportional zu dieser. Der Wert wird als Disparität bezeichnet.

Die gesamte Berechnung basiert auf folgenden Annahmen:

- Es liegt für beide Bildebenen eine Transformationsmatrix vor, die eine vollständige mathematische Beschreibung dafür liefert, wie welche Punkte in der Umgebung auf die zweidimensionale Bildebene projiziert werden. Man spricht in dem Fall von einer kalibrierten Kamera. Es wird zwischen zwei voneinander unabhängigen Parameter-Typen unterschieden.
  - **Extrinsische Parameter** definieren die Position (Translation) und Ausrichtung (Rotation) der Kamera im Koordinatensystem.
  - **Intrinsische Parameter** beschreiben die Projektion der dreidimensionalen Welt auf die Bildebene der Kamera. Je nach Komplexität des Modells gehören neben der Kamerakonstante (Distanz zwischen Bildebene und Projektionszentrum), der vertikalen und horizontalen Brennweite (bei quadratischen Pixeln identisch) sowie dem Hauptpunkt (Pixelkoordinate, durch die die optische Achse verläuft) auch Parameter zur Beschreibung geometrischer Abbildungsfehler (Verzeichnung) des optischen Systems zum intrinsischen Kameramodell.
- Die Bildpunkte wurden als Projektion des Objektpunktes in beiden Bildebenen identifiziert und lokalisiert.
- Die Translation der Kamera ist auf eine Dimension beschränkt und es gibt keine Rotation zwischen den Aufnahmen, d.h. die optischen Achsen verlaufen parallel.

Die Distanzberechnung mit dem vorgestellten stereophotogrammetrischen Verfahren bildet die Grundlage für eine bildbasierte, dreidimensionale Rekonstruktion. Zwei Perspektiven reichen jedoch kaum aus, um eine Umgebung oder ein Objekt räumlich zu erfassen. Mit einer steigenden Anzahl an Aufnahmen erhöht sich allerdings auch die Komplexität des Rekonstruktionsverfahrens, sodass manuelle Lösungswege schnell an ihre Grenzen kommen. Davon abgesehen gibt es bis in die 1990er Jahre keine Möglichkeit, die Triangulationskalkulation ohne Informationen zu Position und Rotation jedes einzelnen Bildes durchzuführen, weshalb das Mehrbild-Verfahren nur in kontrollierten Szenarien praktikabel ist. Einen Aufschwung erlebt die Photogrammetrie mit der kommerziellen Verbreitung der Digitalfotografie, denn die Repräsentation eines Fotos als Pixelmatrix ist für Computer direkt lesbar und ermöglicht eine Automatisierung der Berechnungsschritte. Es entsteht ein Forschungsfeld, das sich dem maschinellen Sehen (Computer Vision) widmet und die Entwicklung von Algorithmen zur Analyse und Interpretation von Bilddaten vorantreibt.

### 2.2.2 Structure-from-Motion

Der Begriff „Structure-from-Motion“ geht auf die Erkenntnisse psychologischer Wahrnehmungsexperimente des schwedischen Professors Gunnar Johansson in den 1970er Jahren zurück<sup>2</sup>. Mit seinen Versuchen zeigt er eindrücklich auf, dass die Struktur eines Objekts nur anhand weniger markanter Punkte erkannt werden kann, sobald das Objekt sich bewegt<sup>3</sup>. Eine dreidimensionale Repräsentation wird also imaginativ durch die Interpretation der sich bewegungsbedingt ändernden Distanzen zwischen den dargestellten Punkten konstruiert.

In „The Interpretation of Structure from Motion“<sup>4</sup> greift Shimon Ullmann 1976 die Idee der Punktrepräsentation als Basis der dreidimensionalen Rekonstruktion auf und formuliert die entscheidende Bedingung für den theoretischen Ansatz von Structure-from-Motion (kurz SfM). Eine eindeutige Rekonstruktion ist demnach nur möglich, wenn man von der Starrheit der Objekte ausgeht. Folgt man dieser Einschränkung, so lässt sich nach einer von Ullmann aufgestellten Berechnung bereits mit jeweils vier nicht-koplanaren (nicht auf einer Ebene liegenden) Punkten aus drei verschiedenen orthografischen Blickwinkeln die Struktur, also die relative Raumorientierung der vier Punkte zueinander, und die Bewegung ermitteln. Ullmans Annahme ermöglicht es, Structure-from-Motion als mathematisches Optimierungsproblem zu formulieren: Gesucht ist die 3D-Struktur und Bewegung, die am besten erklärt, warum sich die 2D-Punkte starr verhalten.

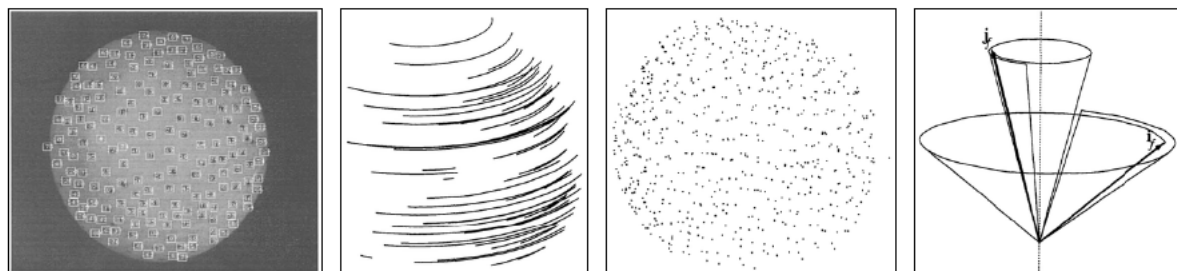


Abbildung 2: Rekonstruktionsergebnisse des "Shape-and-Motion"-Algorithmus

Im Jahr 1992 veröffentlichen Tomasi und Kanade<sup>5</sup> ein computergestütztes Rekonstruktionsverfahren auf der theoretischen Grundlage Ullmanns und schaffen damit den ersten robusten Algorithmus, der aus einer Bildsequenz dreidimensionale Strukturen berechnen kann. Ein zentraler Bestandteil der Arbeit ist die eigens entwickelte Merkmalerkennung (Feature Detection), die nach den Kriterien möglichst hoher Kontraste zu benachbarten Bildpunkten und einer gleichmäßigen Verteilung über das Bild vorgeht (a). Die ausgewählten Punkte werden über die Sequenz hinweg verfolgt (Feature Tracking), um die resultierenden Flugbahnen (Trajectories) (b) mit einer Faktorisierungsmethode in eine Struktur (c) und eine Bewegungsmatrix (d) zu überführen.

Das von Tomasi und Kanade entwickelte Verfahren gilt als Meilenstein in der 3D-Rekonstruktion, ist aber nur der Beginn einer langen Reihe von Entwicklungsschritten

<sup>2</sup> (Johansson, 1973)

<sup>3</sup> (BioMotionLab, 2011)

<sup>4</sup> (Ullmann, 1976)

<sup>5</sup> (Tomasi & Kanade, 1992)

bis hin zu den Structure-from-Motion-Algorithmen von heute. In einer Arbeit von Sturm und Triggs<sup>6</sup> aus dem Jahr 1996 wird das Verfahren zunächst um die perspektivische Projektion erweitert. Im Gegensatz zur orthografischen Projektion kann diese Darstellungsvariante Tiefenverkürzung visualisieren und kommt somit der menschlichen Wahrnehmung sowie den Eigenschaften einer Kamera deutlich näher. Einen großen Fortschritt für die Robustheit und Effizienz des Verfahrens bringt ein Paper aus dem Jahr 2004<sup>7</sup>, in dem „Scale-Invariant Feature Transform“ (SIFT) als skalierungs- und rotationsunabhängige Methode der Merkmalsextraktion vorgestellt wird. Sie avanciert in kurzer Zeit zum Standard für SfM-Anwendungen. Während die Forschung bereits Ansätze zur Berechnung der Kameraposition- und orientierung (extrinsische Parameter) liefert, beruhen bisherige Verfahren auf a priori vorhandenem Wissen über die intrinsischen Kameraeigenschaften. Techniken zur Selbst-Kalibrierung der Kamera adressieren dieses Problem zunehmend. Die Veröffentlichung eines browserbasierten Explorers zur Betrachtung touristischer Orte aus verschiedenen Blickwinkeln zeigt, dass sich mit einer Kombination aus „Direct Linear Transformation“ (DLT) zur Initialisierung und einem Optimierungsalgorithmus (Bundle Adjustment) eine Merkmalsrekonstruktion mit Fotos von verschiedenen Kameras realisieren lässt. Das Kameramodell ist in diesem Fall simpel und auf drei Parameter beschränkt: die Brennweite, die Bildhöhe und die Bildbreite.

Ein moderner und weit verbreiteter Structure-from-Motion-Algorithmus ist „COLMAP“<sup>8</sup>. Die Open-Source-Anwendung vereint die zuvor beschriebenen Entwicklungsstufen mit eigenen Optimierungen zu einem granularen und vielseitigen Tool, das sich sowohl für wissenschaftliche Anwendungen als auch für End-to-End-Prozesse eignet und in viele moderne Anwendungen integriert ist. Die Rekonstruktion wird inkrementell, also Bild für Bild, aufgebaut und weist dabei eine hohe Robustheit auf. Der Ablauf ähnelt früheren SfM-Methoden, wird aber mit dem Ziel maximaler Präzision optimiert. Nach der Erkennung und Zusammenführung von Merkmalen überprüft COLMAP die Korrektheit der Überlappung von Bildern durch eine Rückverfolgung der Merkmalspunkte mit einer angenäherten geometrischen Transformationsmatrix. Somit können als identisch interpretierte Merkmale in nicht korrespondierenden Fotos für den weiteren Verlauf ignoriert werden. Zu Beginn der eigentlichen Rekonstruktion wird ein Bildpaar mit hoher Merkmalskorrespondenz für die Initialisierung ausgewählt, um ein geometrisch akkurates Grundgerüst zu etablieren. Mit einem optimierten Triangulationsalgorithmus nach dem Grundprinzip der Stereophotogrammetrie werden die Fotos nach und nach zum Modell hinzugefügt, wobei die Reihenfolge den Kriterien eines „Next-Best-View“-Entsiders unterliegt. Um den Reprojektionsfehler (Reprojection Error) - also die Abweichung der Merkmalsposition in der aktuellen Rekonstruktionsschätzung von der tatsächlichen Position des Merkmals im Bild - klein zu halten, wird ein iteratives Bundle Adjustment ausgeführt. Diese Korrektur verhindert Folgefehler im Aufbau des Modells und ist ausschlaggebend für eine präzise Rekonstruktion. Eine Filterung abweichender Merkmale (Outlier Filtering) als letzter Schritt trägt ebenfalls zur Verbesserung der Rekonstruktionsqualität bei. Nach der Fertigstellung liegt das Ergebnis in Form einer

---

<sup>6</sup> (Sturm & Triggs, 1996)

<sup>7</sup> (Lowe, 2004)

<sup>8</sup> (Schonberger & Frahm, 2016)

Wolke aus Merkmalspunkten mit geringer Dichte (Sparse Point Cloud) sowie den Kameraperspektiven mit der Referenz zum jeweiligen Input-Bild vor.

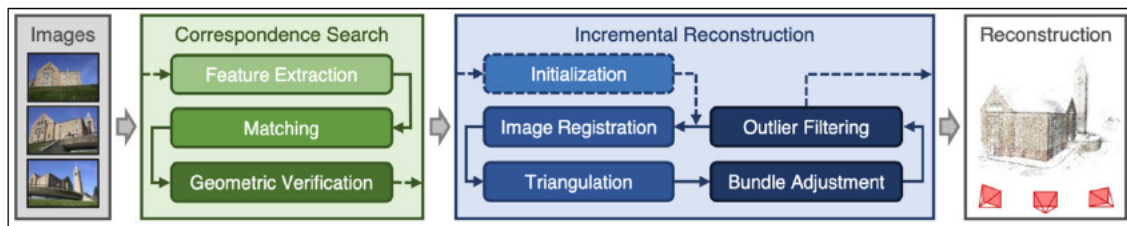


Abbildung 3: Programmlaufplan COLMAP

Im Laufe der Zeit hat sich Structure-from-Motion zu einem beliebten Standardwerkzeug für viele Anwendungsbereiche entwickelt. Vermessungsmethoden in der Medizin, Geologie, Robotik, Architektur sowie in industriellen Anwendungen beruhen auf den Algorithmen von SfM, zum Teil sogar in Echtzeit mit SLAM (Simultaneous Location and Mapping). Die Vorteile der Kosteneffizienz und Praktikabilität bei gleichzeitig hoher Präzision der Rekonstruktion liegen dabei auf der Hand. Die Punktwolke als Ergebnis der SfM-Verfahren hat sich darüber hinaus als geometrisches Grundgerüst für 3D-Rekonstruktionen mit dem Anspruch visueller „Echtheit“ bewährt, so auch für Gaussian Splatting. Lichtbildaufnahmen sind wie kein zweites Medium in der Lage, die visuelle Wahrnehmung kontrolliert und präzise zu imitieren. Gerade dieser Umstand macht bildbasierte SfM-Verfahren zu einem geeigneten Werkzeug in der Entwicklung von digitalen 3D-„Kopien“ realer Szenen.

### 2.2.3 Dreidimensionale Rekonstruktion mit Photogrammetrie

Die Photogrammetrie blickt auf eine lange Geschichte im Vermessungswesen zurück. Ihre heutige Popularität verdankt sie jedoch den Fortschritten der letzten Jahrzehnte, die zu einer Erweiterung der klassischen Begriffsdefinition führten. In den 2010er Jahren etablieren sich algorithmische Verfahren, die aus den Ergebnissen von Structure-from-Motion - also einer groben Punktwolke, einer Bildsequenz und den dazugehörigen Kameraparametern - erstmals realitätsnahe 3D-Modelle berechnen, indem sie aus Punktdaten Flächen modellieren. Photogrammetrie nach heutigem Verständnis umfasst also nicht nur Messmethoden, sondern auch einen weitgehend automatisierten Prozess der oberflächenbasierten 3D-Rekonstruktion aus Fotografien.

Der Ansatz moderner photogrammetrischer Rekonstruktionsverfahren basiert auf der Frage, wie sich die Oberflächen von Objekten aus unorganisierten Punkten im dreidimensionalen Raum rekonstruieren lassen. Da die „Sparse Point Cloud“ von SfM für die Interpretation der Punktdaten als Bestandteil flächiger Strukturen zu wenige Details aufweist, wird die Punktwolke in einem ersten Schritt durch die Rückkopplung mit dem Bilddatensatz iterativ verdichtet, bis eine ausreichende Punktdichte für die Weiterverarbeitung vorliegt. Zudem erhält jeder Punkt einen Orientierungsvektor, der sich nach Faktoren wie der Sichtgeometrie und dem Bildgradienten richtet. Verschiedene Ansätze in diesem Feld werden unter der Bezeichnung „Multi-View Stereo“ zusammengefasst und das Ergebnis begreift man in Abgrenzung zur SfM-Punktwolke als eine sogenannte „Dense Reconstruction“.

Die Oberflächenmodellierung, auch Mesh-Building genannt, ist der entscheidende Schritt beim Aufbau eines photogrammetrischen 3D-Modells. In mehreren Stufen wird aus der Punktrepräsentation ein Festkörpermodell mit zusammenhängenden, texturierten Flächen. Diese setzen sich aus Polygonen, meist aus Dreiecken, zusammen und bilden somit ein Polygon Mesh. Die Grundlage dafür schafft der russische Mathematiker Boris Delaunay bereits im Jahr 1934 mit der Delaunay-Triangulation<sup>9</sup>. Dabei werden benachbarte Punkte so zu Dreiecken verknüpft, dass sich innerhalb des Umkreises von drei Punkten, die ein Dreieck bilden, kein weiterer Punkt befindet. Diese Bedingung gewährleistet eine gleichmäßige und für Rechenprozesse optimierte Flächendarstellung beliebig großer Punktmengen, wird in heutigen Anwendungen allerdings nur noch für die Initialisierung genutzt. Moderne Algorithmen wie die „Poisson Surface Reconstruction“<sup>10</sup> modellieren die Oberfläche eines Objekts als implizite Indikatorfunktion. Diese Funktion unterteilt den Raum in eine Innen- und eine Außenseite, die aus den Orientierungsvektoren (Normalen) der Punktwolke abgeleitet werden. Durch die numerische Approximation dieser Funktion entstehen glatte, rauscharme und wasserdichte bzw. lückenlose Oberflächen, die eine akkurate Rekonstruktion starrer Objekte ermöglichen. Anschließend werden die generierten Flächen den Bildinformationen entsprechend mit Texturen versehen. Eine gängige Methode ist die Erstellung einer „UV-Map“, einer Art digitalem Schnittmuster, mit dem sich die 3D-Geometrie in die 2D-Ebene überführen lässt. Die Farb- und Detailinformationen lassen sich dadurch in einem Textur-Atlas zusammenzufassen. Beim Rendering projiziert die Software die Texturen anhand der UV-Koordinaten präzise auf die dreidimensionale Oberfläche, sodass ein realistisches, texturiertes Modell entsteht.

Zur Erstellung bildbasierter 3D-Modelle haben sich im Laufe der Zeit verschiedene Programme etabliert, darunter Meshroom, Agisoft Metashape und RealityScan. Meshroom ist ein flexibel einsetzbares Open-Source-Tool der Non-Profit-Organisation AliceVision, die aus einem Zusammenschluss europäischer Forscher der Computer-Vision-Community besteht. Die Software Metashape des russischen Unternehmens Agisoft gilt aufgrund der vielen technischen Features und ausgereiften Messfunktionen als Marktführer und Standard für viele Anwendungen. RealityScan vom US-amerikanischen Software-Unternehmen Epic Games stellt eine kostenlose und in der Funktionalität reduzierte Alternative mit dem Fokus auf der Anwendung im Gaming-Bereich dar, unter anderem mit direkter Integration in Unreal Engine. Der beschriebene Workflow, bestehend aus der Berechnung der SfM-Punktwolke, der anschließenden Verdichtung mit Multi-View-Stereo-Algorithmen sowie der Modellierung und Texturierung von Oberflächen, ist ein Standard-Verfahren in jeder der genannten Photogrammetrie-Software. Auch wenn das photogrammetrische Oberflächenmodell in der Praxis weit verbreitet ist, weist es einige spezifische Probleme auf, die gegen eine Verwendung in Szenarien mit einem hohem Realitätsanspruch sprechen, wie z.B. in der virtuellen Filmproduktion.

---

<sup>9</sup> (Delaunay-Triangulation - Geoinformatik Lexikon, o. J.)

<sup>10</sup> (Kazhdan et al., o. J.)

Die drei größten Schwachstellen der konventionellen Photogrammetrie-Pipeline sind:

- **Merkmalsarme Oberflächen**

Die Qualität der Meshes hängt davon ab, wie detailliert Strukturen durch die Punktwolke dargestellt werden. Algorithmen zur Merkmalsextraktion erkennen im Bild vor allem Kontraste, die von SfM dann als Punkte repräsentiert werden. Weil MVS-Methoden auf SfM aufbauen, können schwach texturierte Flächen mit wenigen Merkmalspunkten nur unzureichend verdichtet werden. Die Folge sind ungenaue, löchrige oder sogar fehlende Mesh-Bereiche.

- **Filigrane Strukturen**

Je dünner und detailreicher ein Objekt ist, desto höher muss die Bildauflösung sein, um die Strukturen erfassen zu können. Außerdem kommt es im Prozess der Rekonstruktion häufig vor, dass eine filigrane Struktur fälschlicherweise mehreren Merkmalen zugeordnet oder als Rauschen fehlinterpretiert wird, was in beiden Fällen zu Artefakten führt.

- **Transparenz und Reflektion**

Bei der Texturierung wählt die Software in der Regel frontal aufgenommene Bilder aus und mittelt die Farbwerte der Merkmalspunkte, um einen repräsentativen Farbton berechnen zu können. Allerdings ignoriert diese Vorgehensweise blickwinkelabhängige Farbveränderungen in der realen Welt, sodass reflektierende und transparente Oberflächen entweder falsch rekonstruiert oder gar nicht dargestellt werden.

## 2.3 Die Grundlagen von 3D-Gaussian Splatting

Das Original-Paper zu „3D Gaussian Splatting for Real-Time Radiance Field Rendering“ wurde im Juli 2023 in der Fachzeitschrift „ACM Transactions on Graphics“ veröffentlicht (Kerbl et al., 2023). Die Autoren sind Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler und George Drettakis. Gaussian Splatting basiert auf der Darstellung von Szenen durch eine Ansammlung von 3D-Gaußschen Verteilungen, den sogenannten Gaussians bzw. Splats. Diese Voxels beschreiben nicht nur die Farbe und Transparenz eines Punktes im Raum, sondern auch dessen Position, Größe, Rotation und komplexe Lichtreflexionen. Jeder Gaussian wird durch folgende Parameter definiert.

- Position in x-, y- und z-Richtung (3 Variablen)
- Skalierung (3 Variablen)
- Rotation (4 Variablen)
- Opazität (1 Variable)
- Spherical Harmonics für die Farbdarstellung (3 Variablen für Grad 0 und 45 für die Grade 1–3).

Das Training von Gaussian Splatting beginnt mit einer Initialisierungspunktwolke, die durch Structure-from-Motion aus den Input-Bildern generiert wird (siehe Abbildung 4). Aus dieser Punktwolke werden die Gaussians im dreidimensionalen Raum angelegt. Sobald die Gaussians initialisiert sind, beginnt das Training, bei dem die Parameter der Gaussians fortlaufend mit den Input-Bildern als Ground-Truth verglichen werden. Das Ziel ist es, die Differenz zwischen der gerenderten Szene und den tatsächlichen Bildern zu minimieren. Ein zentraler Prozess in dieser Phase ist die Verdichtung bzw. Densification. Dabei werden die Gaussians, die unterrepräsentiert sind – also zu wenige Details abdecken – geklont. Gleichzeitig werden überrepräsentierte Gaussians, die zu groß oder ungenau sind, geteilt, um die Detailtreue der Szene zu erhöhen. Alle 100 Iterationen findet darüber hinaus eine Pruning-Abfrage statt, bei der zu große oder irrelevante Gaussians eliminiert werden. Dies stellt sicher, dass das Modell effizient bleibt und sich auf die relevanten Bereiche der Szene konzentriert.

Der „Operation Flow“ zeigt, wie die SfM-Punkte zunächst in die Initialisierungsphase fließen, um die 3D-Gaussians zu erzeugen. Diese werden dann als Projektion auf die 2D-Kameraebene abgebildet. Anschließend übernimmt der sogenannte „Differentiable Tile Rasterizer“ die Aufgabe, die Gaussians in ein finales Bild umzuwandeln. Dieser Rasterizer ist differenzierbar, was bedeutet, dass Gradienten zurückberechnet werden können, um die Parameter der Gaussians weiter zu optimieren. Das fertige Bild wird generiert und kann mit den Ground-Truth-Bildern verglichen werden.

Der „Gradient Flow“ veranschaulicht, wie die Informationen aus den Input-Bildern der Kamera danach zurück zu den 3D-Gaussians fließen, um deren Parameter anzupassen. Dieser Zyklus aus Projektion, „Adaptive Density Control“ und Rasterisierung wird wiederholt, bis die gerenderte Szene möglichst genau mit den Input-Bildern übereinstimmt oder das Training gestoppt wird. Die Adaptive Density Control spielt dabei eine entscheidende Rolle, da sie sicherstellt, dass die Gaussians optimal verteilt und angepasst werden, um sowohl Effizienz als auch Genauigkeit zu gewährleisten. Am Ende

dieses Prozesses steht ein hochwertiges, gerendertes Bild, das die ursprüngliche Szene präzise abbildet. Der gesamte Ablauf kombiniert also die Initialisierung, Anpassung und Rendering der Gaussians, um eine Echtzeit-Darstellung von 3D-Szenen zu ermöglichen.

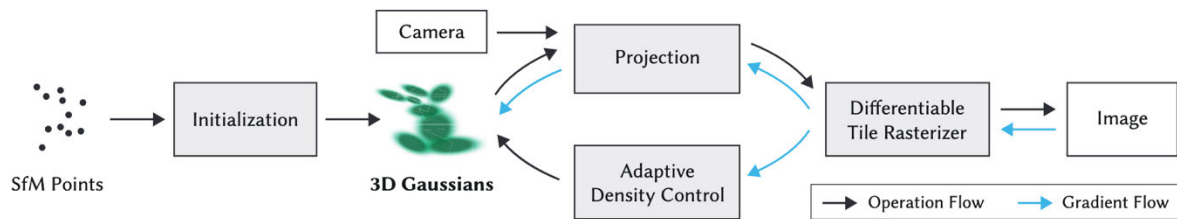


Abbildung 4: Der Trainingsprozess von 3D-Gaussian Splatting

Die Render-Technik von Gaussian Splatting basiert auf dem Splatting-Ansatz, einer Feed-Forward-Methode. Hier werden die dreidimensionalen Voxeldaten direkt auf die zweidimensionale Bildebene abgebildet. Dies unterscheidet sich grundlegend von Feed-Backward-Methoden wie bei Neural Radiance Fields (NeRF), die durch Raycasting die Bildebene auf Volumendaten abbilden. Beim Splatting müssen die Datenpunkte neu angeordnet und mathematisch korrekt verrechnet werden. Dies geschieht durch eine Transformation in den kontinuierlichen Bereich, die als Rekonstruktion bezeichnet wird. Nach der Bearbeitung im kontinuierlichen Bereich wird das Signal durch Sampling wieder in den diskreten Bereich überführt, um den Anforderungen von Displays mit diskreten Pixelwerten gerecht zu werden.

Es lässt sich zusammenfassend sagen, dass eine Szene mit Gaussian Splatting fotorealistisch rekonstruiert werden kann, weil der Bildinhalt durch anisotrope, also richtungsabhängige, 3D-Gaussians mit sphärischen Kugelfunktionen (Spherical Harmonics) modelliert wird, die komplexe Lichtreflexionen und feine Details präzise abbilden. Die Echtzeitfähigkeit rührt daher, dass die Gaussians explizit gespeichert werden und durch Feed-Forward-Splatting mit dem Tile-Rasterizer effizient auf die Bildebene projiziert werden.

## 2.4 LiDAR-Scanning

Im Rahmen der Recherche zu den verschiedenen Methoden der 3D-Rekonstruktion stößt man unweigerlich auf Light Detection and Ranging – kurz LiDAR. Hierbei handelt es sich um eine Messmethode zur Erfassung dreidimensionaler Daten der Umgebung mit einem Laserlichtimpuls. Die Reflexionen der ausgesendeten elektromagnetischen Wellen werden durch einen Sensor im Messgerät detektiert, wodurch sich die Laufzeiten ermitteln und die Positionen der reflektierenden Punkte im Raum berechnen lassen. Somit können hoch-präzise Tiefeninformationen zu Punktwolken verarbeitet werden, die vor Allem für Anwendungen im Mobilitätswesen, Baugewerbe, und in der Geodäsie, also der Wissenschaft der Vermessung der Erde, relevant sind.

Das Trainieren von Gaussians setzt wie zuvor beschrieben die Initialisierung mit einer Punktwolke und den darin enthaltenen Kamerapositionen voraus. Ein Ansatz aktueller Forschungsarbeiten untersucht daher die Kombination von LiDAR-Messdaten mit der SfM-Point-Cloud, um eine noch präzisere Repräsentation der 3D-Szene zu erhalten. Untersuchungen<sup>11</sup> zeigen, dass die geometrische Genauigkeit und damit auch die visuelle Qualität einer Rekonstruktion durch die Kombination beider Datensätze zum Teil deutlich verbessert werden kann. Gerade in Bereichen mit einer unzureichenden Abdeckung durch Fotos, die dem SfM-Algorithmus als Grundlage für die geometrische Rekonstruktion dienen, bietet LiDAR somit eine wertvolle Ergänzung.

Um die Punktwolken von LiDAR und SfM miteinander zu verschmelzen, sind einige Schritte notwendig, die sich in einer 3D-Anwendung wie RealityScan oder CloudCompare umsetzen lassen. Abgesehen von einer Größenangleichung und der Anpassung der Punktdichte durch Herunterskalieren der LiDAR-Daten müssen die Modelle exakt zueinander ausgerichtet werden. Der Aufwand einer manuellen Nachbearbeitung steht dabei nicht immer im Verhältnis zum Mehrwert einer verbesserten Rekonstruktion. Mit der PortalCam von XGRIDS, einem chinesischen Unternehmen mit dem Fokus auf Spatial Computing, ist in der zweiten Jahreshälfte 2025 ein Produkt auf dem Markt platziert worden, das ein 360°-umfassendes 4-Kamera-System und einen 180°-LiDAR-Scanner in einem Gerät vereint. Der Hersteller verspricht die erste Plug&Play-Lösung für die Erstellung von fotorealistischen 3DGS-Modellen und hebt die Verwendung für die virtuelle Filmproduktion hervor.

In weitergehenden Untersuchungen wäre ein Vergleich zwischen den Rekonstruktionsergebnissen mit der PortalCam und denen der „manuellen“ Aufnahmemethode mit einer Digitalkamera durchaus vielversprechend, um das Potenzial des Konzepts zu bewerten. Fraglich ist beispielsweise, ob die automatisierte Belichtung und Bildselektion, die einen wesentlichen Einfluss auf die Rekonstruktionsqualität haben, in jedem Szenario das optimale Ergebnis hervorrufen. Das geschlossene System bestehend aus dem Gerät, einer Handy-App und der Trainings-Software mit proprietärer Datenstruktur erschwert allerdings die Beurteilung des LiDAR-basierten Ansatzes und den Vergleich mit anderen Methoden.

---

<sup>11</sup> (Ijaz, o. J., S. 70)

## 3 Szenenrekonstruktion mit Gaussian Splatting

### 3.1 Vorbereitende Überlegungen

In diesem Kapitel werden zunächst die grundlegenden Voraussetzungen für die Aufnahme eines qualitativ hochwertigen Bilddatensatzes diskutiert, um im weiteren Verlauf gelungene Rekonstruktionen mit Gaussian Splatting erstellen zu können. Die Überlegungen basieren auf den Inhalten der Forschungsarbeiten sowie eigenen Versuchen.

Aus den theoretischen Grundlagen lässt sich ein wichtiges Kriterium für die Eignung einer Umgebung oder eines Objekts als Inhalt eines 3D-GS-Modells ableiten. Zurückgehend auf die Grundannahme der Starrheit der Objekte sollte die Szene für eine in sich konsistente Rekonstruktion innerhalb der Aufnahmen keinen erkennbaren Veränderungen unterliegen. Dazu zählen sowohl variierende Lichtbedingungen als auch sich bewegende Elemente. Um darüber hinaus die visuelle Kontinuität innerhalb der Aufnahmen zu gewährleisten, sollte die verwendete Kamera über einen manuellen Einstellungsmodus verfügen.

Für die Verwendung von Fotos statt Videos in der Erstellung der Aufnahmen sprechen die folgenden zwei Gründe. Erstens basiert das Training der 3D-GS-Modelle auf den Kameraperspektiven von Einzelbildern, weshalb jedes in eine Software importierte Video intern in Fotos umgewandelt wird. Das Training mit allen Frames eines Videos durchzuführen, führt zu hohen Rechenzeiten und wirkt sich nachteilig auf die Rekonstruktionsqualität aus, weil im Bilddatensatz aller Wahrscheinlichkeit nach unscharfe Aufnahmen vorhanden sind. Die nachträgliche Selektion ist hingegen mühsam und aufwendig. Der zweite Grund für die Präferenz von Fotos liegt im Grundkonzept der Belichtung begründet. Für gelungene Aufnahmen sollte weder Tiefen- noch Bewegungsunschärfe erzeugt werden. Eine große Schärfentiefe erreicht man mit einer geschlossenen Blende und Motion Blur lässt sich durch die Wahl einer kurzen Verschlusszeit minimieren. Beide Belichtungseinstellungen führen dazu, dass sich die vom Sensor ausgelesene Lichtmenge stark reduziert, weshalb die ISO-Empfindlichkeit für ein korrekt belichtetes Bild erhöht werden muss. Dieses Vorgehen erhöht allerdings das Bildrauschen und kann nur bis zu einem gewissen Grad ohne massive Qualitätsverluste durchgeführt werden. Nutzt man Fotos statt eines Videos, kann die Verschlusszeit verlängert werden, ohne Motion Blur zu erzeugen. So ist es möglich, auch weniger stark beleuchtete Szenen mit einer ausgeglichenen Helligkeitsverteilung aufzunehmen.

Die optimalen Lichtverhältnisse sind eine weitere wichtige Komponente für eine hochwertige Rekonstruktion. Eine Szene sollte wegen der genannten kameraseitigen Limitationen auf einem möglichst hohen Niveau ausgeleuchtet sein, um scharfe und ausreichend belichtete Aufnahmen erzeugen zu können. Weil die Kameraeinstellungen über den gesamten Bilddatensatz hinweg konstant bleiben müssen und Clipping sowohl in den Highlights als auch in den Schatten zu vermeiden ist, sollte darüber hinaus auf eine gleichmäßige Lichtverteilung geachtet werden. Sind die Helligkeitsunterschiede in einer Szene stark ausgeprägt, kann die Wahl eines lichtempfindlicheren Sensors und die Aufnahme im RAW-Format den Dynamikumfang der Aufnahmen deutlich erweitern und zur Erhaltung der Details in Licht und Schatten beitragen.

Die folgenden drei Gütekriterien können die Qualität einer Rekonstruktion nach der eigener Auffassung am treffendsten beschreiben:

- Genauigkeit der räumlichen Anordnung der Objekte (geometrische Präzision)
- Detailauflösung (Schärfe)
- Dichte der Blickwinkelabdeckung

Die aufgestellten Kriterien dienen dazu, die Rekonstruktionsqualität differenzierter beschreiben und beurteilen zu können. Sie bedingen sich allerdings auch gegenseitig.

## 3.2 Verwendetes Equipment & Software

Für die Erstellung eines 3D-Modells mit Gaussian Splatting werden grundsätzlich zwei Dinge benötigt: ein Kamerasystem für die Aufnahmen und ein Rechner mit Software zur Berechnung des Modells. Bei cloudbasierten Softwarelösungen entfällt die Notwendigkeit der lokalen Berechnung an einem PC, sodass ein Smartphone mit der passenden App schon ausreichen kann, um zufriedenstellende Ergebnisse zu erzielen. Für den Zweck dieser Arbeit ist es jedoch wichtig, die Kontrolle über den gesamten Prozess zu behalten. So kann der Einfluss verschiedener Faktoren auf die Rekonstruktionsqualität untersucht und das Verfahren schrittweise optimiert werden. Bei der Wahl des Equipments sowie bei der Entscheidung für eine geeignete Trainingssoftware muss die Variabilität der Einstellungen daher im Vordergrund stehen.

Für den Capture-Prozess wird eine spiegellose Systemkamera von Fujifilm verwendet. Diese ist in der Lage, hochauflösende Bilder mit einem hohen Dynamikumfang zu fotografieren und bietet umfangreiche Einstellungsmöglichkeiten inklusive der Aufnahme im RAW-Format. Die Kamera hat folgende Spezifikationen:

- Modell: Fujifilm X-H2S
- Sensor: X-Trans CMOS 5 HS mit 26,16 Megapixel
- Sensorgröße: 23,5 x 15,6mm (APS-C)
- Maximale Auflösung: 6240 x 4160 Pixel

Wie bereits in den Vorüberlegungen diskutiert, sollte das horizontale Sichtfeld der Kamera (Field of View) für ein effizientes und gleichzeitig robustes Aufnahme-Verfahren je nach Art und Größe der Szene zwischen 60° und 80° liegen. Folgende Formel gilt für die Berechnung der resultierenden Brennweite:

$$f = \frac{\text{Sensorbreite}}{2 * \tan(\frac{FoV}{2})}$$

Daraus ergibt sich für den Sensor der Kamera ein horizontaler Brennweitenbereich von etwa 14mm bis 20mm, der mit dem „Sigma 10-18mm F2.8 DC DN Contemporary“ überwiegend abgedeckt werden kann. Das Objektiv verfügt über ein X-Bajonett - notwendig für den Anschluss an die Fujifilm X-H2S - und zeichnet sich durch eine verzeichnungsarme, neutrale Charakteristik aus. Für die Verarbeitung der Bilder und das Training der Modelle wird der „Unreal PC1“, ein leistungsfähiger Windows-PC aus dem Videolabor am Campus Finkenau, mit einer NVIDIA GeForce RTX 4090 als Grafikeinheit genutzt.

Angeichts der Vielzahl verfügbarer Programme für die 3D-Rekonstruktion mit Gaussian Splatting erfordert die Wahl der geeigneten Software einige Zeit der Recherche und des Ausprobierens. Zu Beginn des Jahres 2026 existieren bereits rund 20 verschiedene Anwendungen allein für das Training der Modelle, mit steigender Tendenz. Einen

aktuellen Überblick über bestehende Software-Tools sowie die neuesten Entwicklungen zum Thema „Gaussian Splatting“ bietet die Website [www.radiancefields.com](http://www.radiancefields.com). Im wissenschaftlichen Kontext spielen die Vergleichbarkeit und Transparenz der Datenverarbeitung eine zentrale Rolle, weshalb sich die Nutzung eines Tools mit lokaler Verarbeitung auf dem eigenen PC empfiehlt. Die Researcher des Original-Papers von „Gaussian Splatting“ verweisen auf der Homepage des Forschungsprojektes<sup>12</sup> auf den eigenen GitHub-Code, der ein kommandozeilenbasiertes Training und die Darstellung mit einem Viewer ermöglicht. Die Nutzung setzt allerdings ein umfassendes technisches sowie inhaltliches Verständnis voraus und ist für den Einstieg in das Thema eher ungeeignet. Außerdem fehlen einige Funktionen, die den Workflow der Erstellung von „Digital Twins“ im Kontext virtueller Filmproduktionen erleichtern und effizienter gestalten, wie z.B. das Editieren der Modelle, ein ausgereifter Rendering-Modus oder die Integration in „Unreal Engine“.

Programme wie „Postshot“, „LichtFeld Studio“, „nerfstudio“ (für Windows) aber auch „ANNX Studio“ und „Brush“ (für MacOS) basieren auf den Trainings- und Renderingverfahren der 3D-GS-Entwickler, sind im Gegensatz zum GitHub-Projekt aber von vornherein als End-to-End-Software konzipiert und benutzerfreundlicher gestaltet. Für diese Arbeit wird das Programm „Postshot“ verwendet, da es in ersten Tests die besten Rekonstruktionsergebnisse erzielte, eine intuitive und robuste Bedienoberfläche mit wichtigen Funktionen bietet und die Integration der Modelle in „Unreal Engine“ dank eines eigenen Plug-Ins komfortabel gestaltet.

- „Postshot“, Version v1.0.110 (2025-12-04)
- Entwickler: „Jawset Visual Computing“
- Systemanforderungen:
- Software: Windows 10 oder höher
- Hardware: NVIDIA-GPU mit Compute Capability Version 7.5 oder höher („Postshot“ nutzt die Programmierschnittstelle „CUDA“ von NVIDIA, um die Prozesse der Grafikeinheit zu optimieren)
- Software-Lizenzierungstyp: Subscription-Modell
  - „Free“ (kostenlose Basis-Version, mit Wasserzeichen)
  - „Indie“ (kostenpflichtig, für die kommerzielle Nutzung)
  - „Studio“ (kostenpflichtig, ohne Limitierung in Auflösung und Quellformat)
- Verwendete Lizenz: „Studio“-Lizenz

### 3.3 Test- und Optimierungsphase

Nachdem in den vorangegangenen Kapiteln die theoretischen Grundlagen sowie der technische Ablauf einer Rekonstruktion mit Gaussian Splatting erläutert wurden, können auf dieser Basis nun eigene Versuche zur Erprobung und Optimierung des Workflows durchgeführt werden. Dafür wird zunächst eine geeignete Szene aufgebaut und ein

---

<sup>12</sup> (3D Gaussian Splatting for Real-Time Radiance Field Rendering, o. J.)

Capture-Prozess erarbeitet, der die Erkenntnisse aus den Vorüberlegungen berücksichtigt, um optimale Aufnahmen zu erzielen. Im nächsten Schritt wird eine Parameterstudie durchgeführt, die die Auswirkungen der Veränderung ausgewählter Variablen und Einstellungen auf die visuelle Qualität der Rekonstruktionsergebnisse analysiert. Um einen objektiven Vergleich zwischen den Modellen zu ermöglichen, werden gängige Metriken zur Bewertung der Videoqualität herangezogen. Ergänzend dazu wird ein subjektiver Vergleich mit eigenen Beobachtungen in die Beurteilung eingebracht. So lassen sich die rechnerischen Ergebnisse mit der menschlichen Wahrnehmung abgleichen und kritisch einordnen.

### 3.3.1 Aufbau der Test-Szene

Um die Bilddichte der Modelle unter kontrollierten Bedingungen variieren zu können, wird eine Szene gewählt, die halbkugelförmig aus allen Perspektiven aufgenommen werden kann. Ein runder Tisch, auf dem sich weitere Objekte platzieren lassen, bietet sich aus praktischen Gründen an. Damit eine hohe Bildabdeckung in horizontaler und vertikaler Ebene erreicht werden kann, wird aus verschiedenen Höhen mit einer zur Tischmitte ausgerichteten Neigung der Kamera um den Tisch herum fotografiert. Zur Orientierung im Raum dienen Winkelmarkierungen auf dem Tisch und ein Zollstock sorgt für eine gleichbleibende Höhenposition der Kamera. Die Lichtsituation der Szene ist durch die Platzierung des Tisches im Innenbereich und einer von mehreren Quellen ausgehenden Beleuchtung relativ diffus und weitgehend kontrolliert. Die Strahlung der Abendsonne wird durch einen weißen Vorhang ausreichend gestreut, sodass keine harten Schatten auf dem Tisch entstehen.

Auf dem Tisch werden verschiedene Objekte platziert, die die Komplexität der Szene erhöhen sollen und im weiteren Verlauf für die Beurteilung der Rekonstruktionsqualität hilfreich sind. Ein besonderes Augenmerk soll dabei auf blickrichtungsabhängige Lichtphänomene gelegt werden, die durch bildbasierte Rekonstruktionsverfahren typischerweise nur mit Einschränkungen nachgebildet werden können. Laut Original-Paper sind Transparenz, Reflexion und sogar der Extremfall der Spiegelung theoretisch durch Kugelflächenfunktionen (Spherical Harmonics) darstellbar. Ein mit Wasser gefülltes Weinglas, eine Edelstahlkanne und ein Kosmetikspiegel weisen die genannten optischen Phänomene auf und werden auf dem Tisch platziert. Die Schrägstellung des Spiegels ist intendiert und soll Aufschluss darüber geben, ob Spiegelungen mit 3D-GS auch dann adäquat dargestellt werden können, wenn der Bereich hinter dem Spiegel aus anderen Blickwinkeln betrachtet kein Spiegelbild zeigt. In der Regel ist ein Spiegelbild im Modell nämlich nichts anderes als eine Erweiterung des Raumes, da hinter die Spiegelfläche eine Nachbildung der Szene projiziert wird. Sobald die Spiegelung wie in diesem Fall nicht mehr als räumliche Extension interpretiert werden kann, sind die Möglichkeiten und Grenzen der Darstellung mit Kugelflächenfunktionen am deutlichsten erkennbar. Darüber hinaus stellen Details bzw. filigrane Strukturen sowie glatte, monotone Oberflächen im Allgemeinen eine Herausforderung für die bildbasierte Rekonstruktion dar. In welcher Präzision diese Bereiche im Test-Szenario nachgebildet

werden, lässt sich exemplarisch anhand eines Geldscheins, einer kleinen Vase mit Tulpen und eines weißen Blatt Papiers in den Ergebnissen beurteilen.

### 3.3.2 Capture-Prozess und RAW-Entwicklung

Für den Aufnahmeprozess wird die „Fujifilm X-H2S“ mit dem „Sigma 10-18mm F2.8 DC DN Contemporary“ verwendet. Aus 16 Winkelpositionen werden in einer halbkugelförmigen Anordnung je sieben in der Höhe variierende Fotos mit Ausrichtung der optischen Achse zur Tischmitte geschossen. Zusätzlich wird eine Aufnahme aus der zentralen Position der „Kugelspitze“ nach unten blickend realisiert. So entstehen insgesamt 113 Fotos, die mit folgenden Kameraeinstellungen fotografiert wurden:

- Format: Fuji RAW (.raf)
- Auflösung: 6240x4160 Pixel
- Brennweite: 10mm (APS-C)
- Film-Emulation (Farbprofil): PROVIA/Standard
- Fokus-Einstellung: manuell
- Blende: f/10
- ISO: 2500
- Belichtungszeit: 1/100 Sekunde

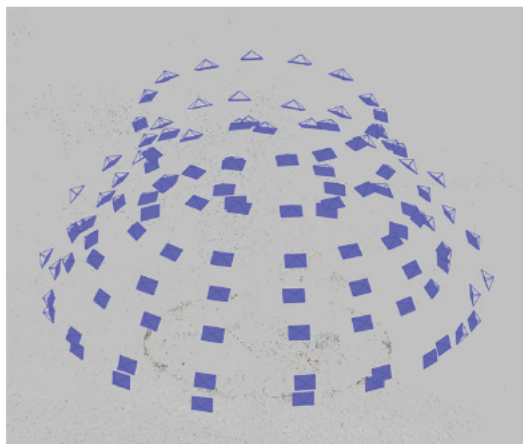
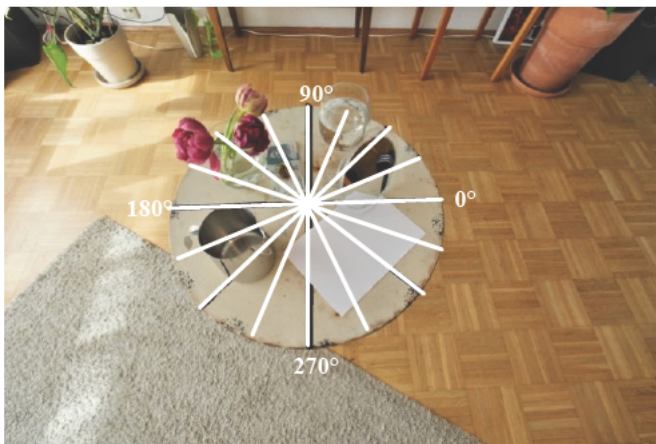


Abbildung 5: Beispielfoto aus dem Bilddatensatz (DSCF0858.jpg)    Abbildung 6: Kamerapositionen in COLMAP

Bei einer Brennweite von 10mm ergibt sich ein Sichtfeld von knapp 100°, was außerhalb des empfohlenen Bereichs von 60° bis 80° liegt. Vor der Testphase wurde abgewogen, ob eine Vergrößerung des Sichtfelds und die damit mutmaßlich einhergehende höhere geometrische Robustheit des Modells durch größere Überlappungsbereiche innerhalb der Bilder angesichts der ohnehin hohen Bildauflösung einen positiveren Effekt auf die Qualität der Rekonstruktion haben würde als ein kleineres Sichtfeld mit höherer räumlicher Auflösung. Die Entscheidung fiel zugunsten der Maximierung des Sichtfelds, also der Verwendung der kleinsten möglichen Brennweite von 10 mm, um die Ergebnisse mit denen eines zugeschnittenen Bilddatensatzes mit kleinerem Sichtfeld vergleichen zu können. Die Frage nach dem optimalen Sichtfeld wird im weiteren Verlauf wieder aufgegriffen.

Der entstandene Bilddatensatz hat eine Größe von 6,1 Gigabyte, jede Aufnahme umfasst über 50 Megabyte. Da es sich um ein herstellerspezifisches Datei-Format handelt, können die Fotos nicht ohne Weiteres für das Training in Postshot verwendet werden. Mit der Software „RAW FILE CONVERTER EX 3.0 powered by SILKYPix“ (Version 8.1.20.0) lassen sich die Fuji-RAW-Fotos bearbeiten und in ein lesbares Format umwandeln. Unter Berücksichtigung der Belichtung und Farbtemperatur wird ein ausgeglichenes Bildprofil für den Datensatz erstellt. Die folgenden Einstellungen werden auf alle Aufnahmen gleichermaßen angewendet:

- Farbtemperatur: 5000 Kelvin
- Belichtung: -1.0 EV
- HDR: 50%

### 3.3.3 Methodik der Parameterstudie

Das Ziel der Parameterstudie besteht darin, ein Verständnis dafür zu erlangen, welche Faktoren entlang der gesamten Verarbeitungskette einer 3D-Rekonstruktion mit der Gaussian-Splatting-Methode wesentlichen Einfluss auf die visuelle Qualität des Modells haben. Einige Faktoren konnten bereits im Vorfeld identifiziert werden, andere lassen sich wiederum nur im Modelltraining herausarbeiten. Dabei wird nicht der Anspruch erhoben, die Auswirkungen der Veränderung einzelner Parameter quantifizieren zu können. Es geht lediglich darum, aus dem direkten Vergleich abzuleiten, ob sich einzelne Einstellungen vor- oder nachteilig auf die Modellqualität auswirken. Die deterministische Charakteristik des Systems als Prozess der maschinellen Datenverarbeitung – von der Auslesung des Kamerasensors bis zum Rendering des 3D-Modells – macht die direkte Vergleichbarkeit erst möglich. Mit anderen Worten kann die Tatsache, dass bei gleichem Input und gleicher Verarbeitung immer das gleiche Resultat erfolgt, genutzt werden, um die Folgen der Änderung eines Faktors isoliert zu betrachten.

Zum Zeitpunkt der Verfassung dieser Arbeit existiert nach allen vorliegenden Kenntnissen noch kein standardisiertes Benchmark-Verfahren zur Ermittlung der Rekonstruktionsqualität von 3D-GS-Modellen unter Berücksichtigung der dreidimensionalen Gesamtheit der Szene. Stattdessen werden in der Forschung sogenannte "Full-Reference"-Metriken genutzt, die zur Quantifizierung der Ähnlichkeit eines verlustbehafteten bzw. komprimierten Signals in Relation zum verlustfreien Signal entwickelt wurden. Im Kontext der bildbasierten 3D-Rekonstruktionsverfahren ist es gängige Praxis, aus dem Modell ein synthetisches Bild mit denselben perspektivischen Eigenschaften der Aufnahme zu rendern und dieses mit dem „Ground-Truth“ zu vergleichen. Um den globalen Qualitätsverlust in Bezug auf den gesamten Bilddatensatz zu ermitteln, wird anschließend der Durchschnitt der Einzelbild-Metriken berechnet.

Die Forschung hat im Laufe der Zeit zahlreiche „Full-Reference“-Metriken zur objektiven Bewertung der visuellen Qualität digitaler Bilder entwickelt. Die älteste und immer noch häufig verwendete Metrik ist das Spitzen-Signal-Rausch-Verhältnis, besser bekannt als PSNR (Peak signal-to-noise ratio). Eine Metrik, die sich stärker an den Eigenschaften der menschlichen Wahrnehmung orientiert, ist der strukturelle Ähnlichkeitsindex (SSIM,

Structural Similarity Index Measure)<sup>13</sup>. Weniger etabliert ist der LPIPS (Learned Perceptual Image Patch Similarity Index), der auf dem Training eines neuronalen Netzwerkes basiert. Die Entwickler bildbasierter Rekonstruktionsverfahren stützen sich in der Evaluation ihrer Forschungsergebnisse überwiegend auf diese drei Metriken. Aus Praktikabilitätsgründen fiel die Entscheidung nach der passenden Berechnungssoftware auf das Open-Source-Tool „FFMetrics“<sup>14</sup>, das auf dem Multimedia-Framework „FFmpeg“ basiert. Das GitHub-Projekt bietet eine simple grafische Oberfläche mit den Funktionen Import, Berechnung, Plotten und Export von CSV-Dateien. Es kann allerdings nur die Standard-Metriken PSNR und SSIM für Videos kalkulieren, LPIPS wird nicht unterstützt. Stattdessen ist es möglich, die Berechnung von VMAF (Video Multimethod Assessment Fusion) durchführen. Dabei handelt es sich um eine von Netflix Research entwickelte „Full-Reference“-Methode zur objektiven Bewertung von Videoqualität. Um die Aussagekraft der Messergebnisse zu stärken, wird im Rahmen der Untersuchung neben PSNR und SSIM auch VMAF mit „FFMetrics“ berechnet.

In der nachfolgenden Auflistung wird dargelegt, wie die Vergleichskriterien (objektiv und subjektiv) zu den jeweiligen Ergebnissen gelangen.

### (1) PNSR (objektiv)

Der PSNR ist eine normalisierte und logarithmierte Variante des „Mean squared error“ (MSE), also der mittleren quadratischen Pixelwertabweichung des verlustbehafteten Bildes vom Referenz-Bild. In der untenstehenden Formel ist **H** die Höhe und **B** die Breite des Bildes in Pixeln. Die Luminanz **Y** und die Farbdifferenzwerte **U** und **V** bilden die drei Kanäle eines Komponentensignals, wobei die Indizes **r** und **d** für das Referenzbild und das „distorted“-Bild stehen. Die Bittiefe ist mit dem Buchstaben **b** abgekürzt. Je höher der dB-Wert, desto ähnlicher sind sich die Bilder.

$$MSE = \frac{1}{3 * H * B} \sum_{i=1}^H \sum_{j=1}^B [(Y_r(i,j) - Y_d(i,j))^2 + (U_r(i,j) - U_d(i,j))^2 + (V_r(i,j) - V_d(i,j))^2]$$

$$PNSR = 10 * \log_{10} \left( \frac{(2^b - 1)^2}{MSE} \right) \quad [dB]$$

---

<sup>13</sup> (Zhou Wang et al., 2004)

<sup>14</sup> (fifonik, 2020/2026)

## (2) SSIM (objektiv)

Der Algorithmus des „Structural Similarity Index Measurement“ vergleicht nicht einzelne Pixel, sondern kleine Bildausschnitte hinsichtlich der Luminanz **l**, dem Kontrast **c** und der Struktur **s**.

$$l(x, y) = \frac{2\mu_x\mu_y + c_1}{\mu_x^2 + \mu_y^2 + c_1}$$

$$c(x, y) = \frac{2\sigma_x\sigma_y + c_2}{\sigma_x^2 + \sigma_y^2 + c_2}$$

$$s(x, y) = \frac{\sigma_{xy} + c_3}{\sigma_x\sigma_y + c_3}$$

Die Ergebnisse dieser Berechnungen, bei denen menschliche Wahrnehmungskurven berücksichtigt werden, werden anschließend zu einem Wert kombiniert. Der globale Index ergibt sich aus der Mittelung der lokalen Werte. Ein SSIM-Score liegt zwischen 0 und 1, wobei 1 nur mit identischem Referenz- und Input-Bild erreicht werden kann.

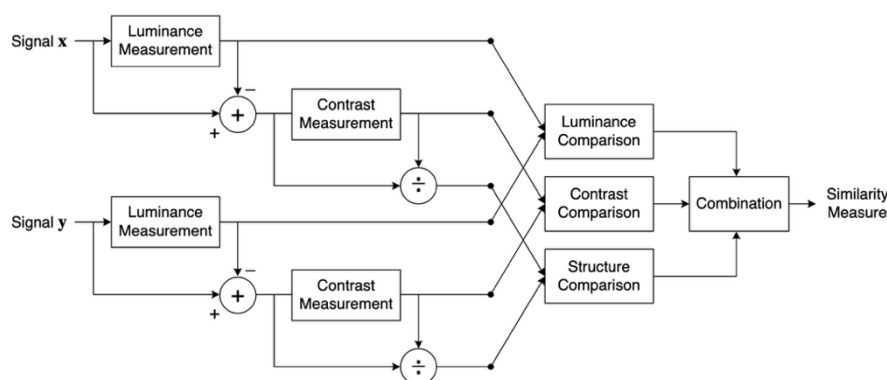


Abbildung 7: Block-Diagramm der SSIM-Berechnung

## (3) VMAF (objektiv)

VMAF ist ein Algorithmus, der wahlweise bis zu 38 verschiedene Metriken mithilfe von maschinellem Lernen gewichtet und zu einem finalen Score zusammenfasst. Die drei elementaren und in jedem VMAF-Modell vorhandenen Metriken sind die „Visual Information Fidelity“ (VIF), die „Detail Loss Metric“ (DLM) und Motion<sup>15</sup>. Sie orientieren sich konzeptionell stark an der menschlichen Perzeption von Bildqualität und wurden entwickelt, um die subjektive Messmethode DMOS (Differential Mean Opinion Score), also eine Form des „Human Benchmarking“ nachzubilden. Der Vorteil der VMAF-Methode liegt in der Adaptivität der Bewertungsgrundlage durch die Möglichkeit, sowohl die Auswahl der Metriken als auch deren Gewichtung durch maschinelles Lernen auf die eigenen Inhalte abzustimmen. Das in dieser Arbeit verwendete Modell „vmaf\_4k\_v0.6.1“ enthält nur die genannten Basismetriken und ist für Videos mit einer 4K-Auflösung optimiert.

<sup>15</sup> (Blog, 2017)

#### (4) Visueller Vergleich (subjektiv)

Der Vorteil der visuellen Methode besteht darin, dass die Vergleichsgrundlage nicht auf Einzelbilder reduziert ist, die perspektivisch exakt aufeinander abgestimmt sein müssen, um Unterschiede in Qualität und Struktur mathematisch detektieren zu können. Bei zwei verschiedenen Modellen einer einzigen 3D-GS-Szene ist ein subjektiver Vergleich den Metriken in der Aussagekraft daher grundsätzlich überlegen, auch wenn er keine numerischen Ergebnisse liefert. Zur Systematisierung des Vergleichs wird aus jedem Modell ein Video mit der immer gleichen Kamerafahrt gerendert (siehe Abbildung 7). Dabei kreist das Sichtfeld mit unterschiedlicher Höhe, Distanz und Brennweite einmal um den Tisch herum, um möglichst viele Perspektiven auf die Szene zu zeigen. Mithilfe dieser Videos können die Objekte und ihre Eigenschaften, wie beispielsweise blickwinkelabhängige Phänomene oder detaillierte Strukturen, im dreidimensionalen Kontext beurteilt werden. In der Beurteilung wird der Fokus weniger auf die allgemeine Rekonstruktionsqualität als auf die Unterschiede zwischen den Modellen gelegt. Die Videos sind im Ordner „Test-Szene“ des Anhangs abgelegt.

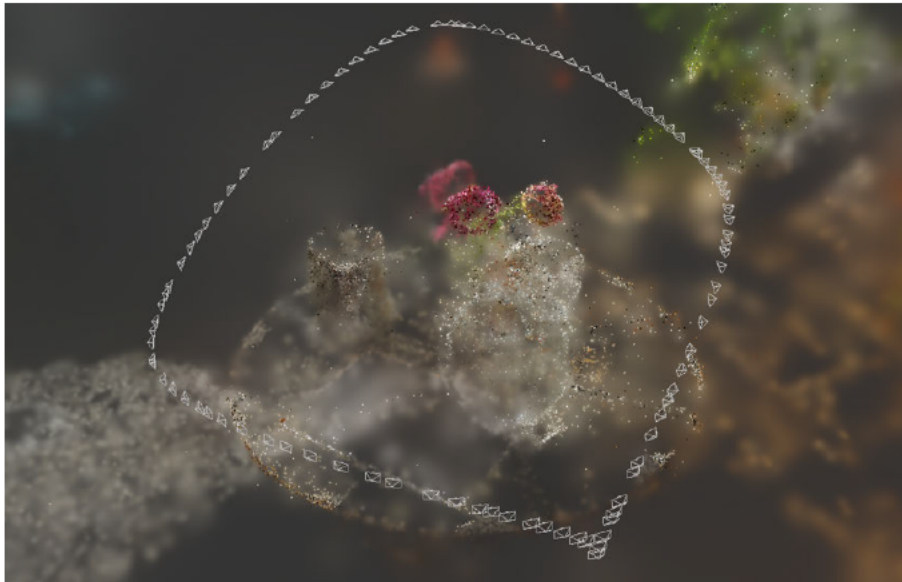


*Abbildung 8: Kamerafahrt des gerenderten Vergleichsvideos*

### 3.3.4 Auswahl der Einflussfaktoren

#### a) Anzahl der Aufnahmen

Im Capture-Prozess stellt sich die Frage, wie eine Szene idealerweise fotografiert werden muss, um eine optimale Repräsentation im dreidimensionalen Raum zu erhalten. Die Vorüberlegungen haben ergeben, dass die Modellqualität grundsätzlich steigt, wenn die Zahl unterschiedlicher Perspektiven auf ein Objekt oder eine Szene erhöht wird, solange die Aufnahmebedingungen unverändert bleiben. Diese These soll anhand eines einfachen Szenarios überprüft werden. Um zu verstehen, wie stark sich eine veränderte Anzahl der Input-Fotos auf die Rekonstruktion auswirkt, werden Datensätze erstellt, die nur die Hälfte bzw. ein Viertel der Fotos aus dem Ursprungsmodell mit 113 Aufnahmen enthalten. Die halbkugelförmige Anordnung der Kamerapositionen ermöglicht es, den Datensatz entsprechend der Abbildungen systematisch so zu reduzieren, dass die grundlegende Struktur der Szene bestehen bleibt.

Leider bietet Postshot keine Funktion, mit der sich in ein bestehendes Modell nachträglich Kamerapositionen importieren lassen, um aus diesen Perspektiven Bilder zu rendern. Auch die manuelle Übertragung der Positionsdaten ist keine Option, denn der SfM-Algorithmus richtet jede Rekonstruktion anders aus. Wenn sich aus einem Modell also nur die Blickwinkel extrahieren lassen, die bereits für das Training genutzt wurden, ist ein Metrik-Vergleich mit dem 16-Winkel-Modell genau dort unmöglich, wo er am meisten Aufschluss bringen würde. Somit bleibt die Beurteilung der Rekonstruktionsqualität an dieser Stelle allein der subjektiven Wahrnehmung überlassen. Nichtsdestotrotz lassen sich die Modelle in den bestehenden Kamerapositionen miteinander vergleichen. Dafür werden zwei 8-Winkel-Modelle und vier 4-Winkel-Modelle mit den entsprechenden Referenz-Bildern trainiert, um anschließend den Durchschnitt der Einzelergebnisse der Metrik-Berechnungen zu ermitteln.

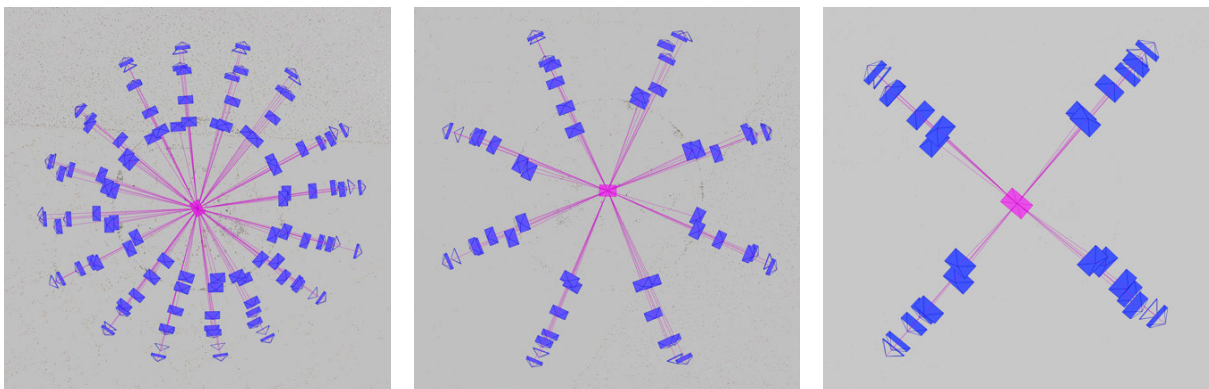


Abbildung 9: Structure-from-Motion mit 16 Winkeln (113 Fotos), 8 Winkeln (57 Fotos) und 4 Winkeln (29 Fotos)

## b) Sichtfeld

Wie im vorherigen Kapitel angemerkt, wird das Rekonstruktionsergebnis maßgeblich vom „Field of View“ der Kamera beeinflusst. Im Rahmen der Parameterstudie soll die Annahme, dass ein großes Sichtfeld zwar zu einer präziseren Modellgeometrie, jedoch auch zu weniger Schärfe in den Details führt, während es sich bei kleinen Blickwinkeln genau andersherum verhält, überprüft werden. Im Rahmen einer exemplarischen Untersuchung werden zwei Modelle auf Basis von Datensätzen mit unterschiedlichen Bildwinkeln miteinander verglichen. Würde man die Szene mit zwei verschiedenen Brennweiten fotografieren, wäre ein subjektiver Vergleich der Modellqualität möglich, doch der objektive Vergleich mit den Metriken fehlerbehaftet, da die Kamerapositionen der beiden Bildreihen nicht exakt übereinstimmen. Zur Vermeidung dieser Ungenauigkeiten ist es sinnvoll, die Aufnahmen des bestehenden Datensatzes mit einem Sichtfeld von  $100^\circ$  so zuzuschneiden, dass bei gleichbleibender Perspektive ein kleinerer Bildwinkel entsteht. Es empfiehlt sich, den Datensatz mit einem Crop-Faktor von 200% mittig aus dem Original zu extrahieren, sodass die halbe Bildauflösung und nach der umgestellten Formel zur Berechnung der Brennweite ein Sichtfeld von ca.  $60^\circ$  resultiert.

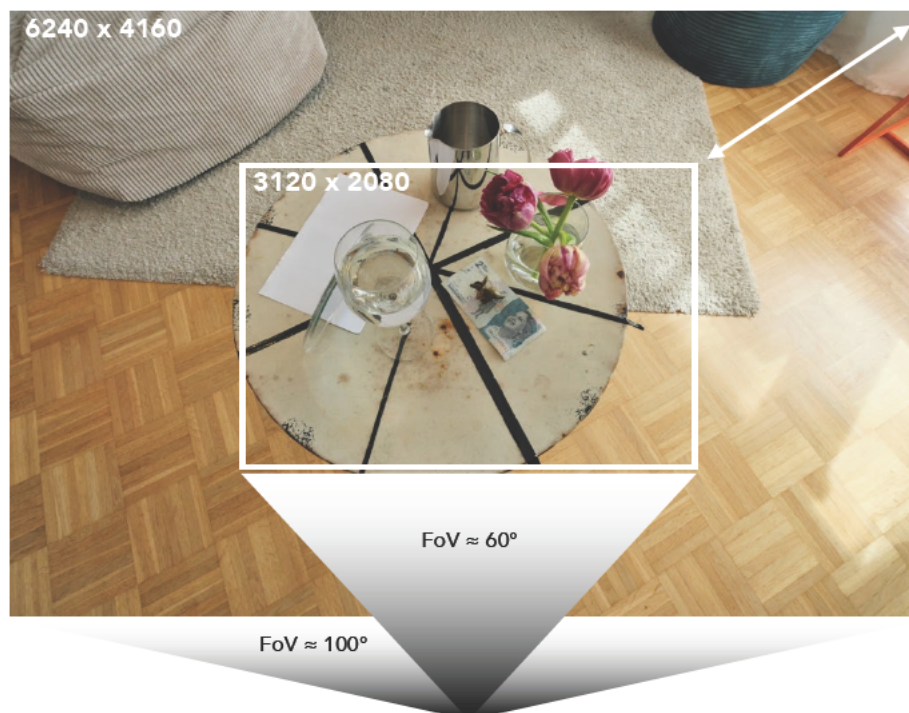


Abbildung 10: Extraktion eines Bilddatensatzes mit kleinerem Sichtfeld

Um einen fairen Vergleich der Resultate zu gewährleisten, muss der Datensatz des größeren auf die Bildauflösung des kleineren Sichtfelds herunterskaliert werden, bevor die Modelle trainiert werden. Da die Ergebnisse der metrischen Berechnungen erst dann aussagekräftig sind, wenn gleiche Bildwinkel und gleiche Auflösungen dafür verwendet wurden, wird vor dem Render-Export aus dem  $100^\circ$ -Modell durch „Heranzoomen“ der gleiche Bildausschnitt wie im  $60^\circ$ -Modell erzeugt. So können die Metriken beider Modelle in Relation zum  $60^\circ$ -Datensatz kalkuliert und direkt verglichen werden. Weil davon

auszugehen ist, dass beide Bildwinkel Vor- und Nachteile in Bezug auf die Rekonstruktionsqualität mit sich bringen, soll in einem dritten Modell untersucht werden, ob sich die jeweiligen Stärken der Sichtfelder durch die Kombination der Datensätze gegenseitig ergänzen können und zu einem besseren Gesamtergebnis führen.

### c) Bildkontrast

Um nachzuvollziehen, wie sich der Kontrast im Bilddatensatz auf die Rekonstruktion auswirkt, werden zwei Modelle mit verschiedenen Kontrastprofilen miteinander verglichen. In der Software, mit der die RAW-Bilder bearbeitet und entwickelt werden, gibt es sieben voreingestellte Kontrastprofile. Um ausreichend Details in den Lichtern und Schatten zu erhalten, werden die eher moderaten Profile „Wenig Kontrast“ & „Hoher Kontrast“ verwendet, während alle anderen Einstellungen unverändert bleiben. Der RMS-Kontrast (Root-mean-square), der als das Verhältnis der Standardabweichung der Luminanz zur mittleren Luminanz definiert ist, beträgt im ersten Datensatz durchschnittlich 0,165 und im zweiten 0,208.

### d) Farbtiefe & Dateiformat

Ein weiterer Einflussfaktor auf die Qualität der Rekonstruktion wird in der Farbtiefe vermutet. Bei der Entwicklung des Fuji-RAW-Formates mit dem „RAW File Converter“ gibt es die Möglichkeit, Fotos im JPEG- oder TIFF-Format zu exportieren. Dabei ist die Farbtiefe bei JPEG-Bildern auf 8 Bit pro Farbkanal festgelegt, bei TIFF-Bildern lässt sich zwischen 8 und 16 Bit wählen. Folgende Datensätze werden für den Vergleich exportiert:

| JPEG 8 Bit                          | TIFF 8 Bit                          | TIFF 16 Bit                          |
|-------------------------------------|-------------------------------------|--------------------------------------|
| Farbtiefe: 24 Bit (8 Bit pro Kanal) | Farbtiefe: 24 Bit (8 Bit pro Kanal) | Farbtiefe: 48 Bit (16 Bit pro Kanal) |
| Farbraum: sRGB IEC61966-2.1         | Farbraum: sRGB IEC61966-2.1         | Farbraum: sRGB IEC61966-2.1          |
| Codierung: YCbCr 4:2:2              | Codierung: YCbCr 4:4:4              | Codierung: YCbCr 4:4:4               |
| Kompression: Qualität 97/100        | Kompression: unkomprimiert          | Kompression: unkomprimiert           |

### e) Auflösung

Die Auflösung ist ein entscheidender Faktor für den Detailgrad eines Bildes. Beim 3D-GS-Algorithmus passen sich die Gaussian Surfels in jeder Trainingsiteration durch Klonen oder Teilen weiter an die Pixelinformationen der Input-Bilder an, wodurch sich die Pixelanzahl auch auf die Rekonstruktion auswirkt. Gilt also: Je höher die Auflösung des Bilddatensatzes, desto besser das 3D-Modell? Für einen systematischen Modellvergleich wird der originale Datensatz mit 6240 x 4160 Pixeln sowohl auf 50% (3120 x 2080) als auch auf 25% (1560 x 1040) der ursprünglichen Auflösung reduziert. Das Downsampling wird über die Postshot-eigene Funktion im Importfenster ausgeführt.

#### **f) SfM-Initialisierung**

Die Grundlage für jedes 3D-GS-Modell ist Structure-from-Motion. In Postshot kann eine Sparse Point Cloud mitsamt der dazugehörigen Kamera- und Bildinformationen entweder aus einem anderen Programm importiert oder direkt in der Software erstellt werden. Die interne SfM-Berechnung läuft im Hintergrund ab, sodass die Prozesse nicht einsehbar sind. Zwar werden die Kamerapositionen auf der Basis von COLMAP berechnet, jedoch stellt sich die Frage, ob eine Vorberechnung mit dem SfM-Tool selbst möglicherweise ein geometrisch präziseres Grundgerüst und damit bessere Rekonstruktionsergebnisse liefern kann als der Default-Modus von Postshot. Es wird ein SfM-Datensatz in COLMAP (Version 3.12.3) generiert, wobei die Voreinstellungen weitgehend beibehalten werden. Lediglich das Kameramodell „Radial“ mit fünf intrinsischen Parametern, das Feature-Matching im „Exhaustive“-Modus, bei dem alle Bilder zueinander relationiert werden, sowie eine nachträgliche Positionskorrektur mit dem „Bundle Adjustment“ werden manuell eingestellt. Ein weiterer Aspekt, der in diesem Zuge untersucht werden soll, ist die Auswirkung der Einstellung „Single Lens & Focal Length“ (SFL) in den „Advanced-Settings“ des Import-Fensters von Postshot.

#### **g) Anzahl der Trainingsiterationen**

Das Modelltraining von Gaussian Splatting ist ein iterativer Optimierungsprozess, bei dem die Anzahl, Position, Form und Farbe der Gaussian Surfels schrittweise an die Bilddaten angepasst werden, um eine fotorealistische Rekonstruktion zu erzielen. Wird das Modell also kontinuierlich besser, je mehr Trainingsiterationen durchlaufen werden? Um dieser Frage nachzugehen, werden Modelle mit 5.000 bis 150.000 Trainingsschritten zu einem Modell rekonstruiert. Dafür werden Bilddaten mit voller Auflösung und mit  $\frac{1}{4}$  der Auflösung genutzt.

#### **h) Maximale Splatanzahl**

Die Standard-Einstellung für die maximale Anzahl der Gaussian Surfels bzw. Splats eines Modells liegt in Postshot bei drei Millionen. Um die Auswirkungen der Splat-Anzahl auf die Datenmenge, die Trainingszeit und die generelle Modellqualität zu untersuchen, werden Rekonstruktionen mit einem Limit von einer, zehn und zwanzig Millionen Splats erstellt.

### **3.3.5 Objektiver Vergleich mit PSNR, SSIM und VMAF**

Die Software FFMetrics ist auf die Metrikberechnung von Videos ausgelegt. In Postshot lässt sich mit der „Image Set Camera“ zwar aus den aneinandergereihten Kamerablickwinkeln des Bilddatensatzes ein Video generieren, allerdings nur in einer Videostandard-Auflösung. Will man die ursprüngliche Auflösung des Inputs beibehalten, müssen JPEG-Bilder exportiert und in einem FFmpeg-basierten Python-Skript zu einem Video zusammengefügt werden. Dabei erweist sich der Codec „FFV1“ (FFmpeg Video

Codec 1) in Kombination mit der Farbcodierung „YUV444“ im Container „.avi“ als eine verlustfreie und mit FFMetrics kompatible Variante. Die Referenz der Metrikberechnung ist, wenn nicht anders angegeben, ein Video bestehend aus 113 JPEG-Aufnahmen mit hohem Kontrastprofil und Original-Auflösung.

| Einflussfaktoren in der Aufnahme & Bearbeitung |           |               |                    |             | Trainingsparameter * |                             |                    | Ergebnisse der Rekonstruktion |                      |                   |                     |              |
|------------------------------------------------|-----------|---------------|--------------------|-------------|----------------------|-----------------------------|--------------------|-------------------------------|----------------------|-------------------|---------------------|--------------|
| Anzahl der Aufnahmen                           | Sichtfeld | Bild-kontrast | Farbtiefe & Format | Auflösung   | SfM-Initialisierung  | Anzahl Trainingsiterationen | Limit Splat-Anzahl | Dauer Training (mm:ss)        | Anzahl Splats (Mio.) | Average PSNR (dB) | Average SSIM (x100) | Average VMAF |
| 29 (4L)                                        | 100°      | HK            | 8Bit (JPEG)        | Quarter     | Postshot             | 10K                         | 10M                | ** 01:48                      | ** 1,85              | ** 29,64          | ** 86,30            | ** 93,82     |
| 57 (8L)                                        |           |               |                    |             |                      |                             |                    | ** 02:05                      | ** 2,65              | ** 28,64          | ** 85,83            | ** 91,03     |
| 113 (16L)                                      |           |               |                    |             |                      |                             |                    | 03:08                         | 3,33                 | 27,93             | 85,47               | 77,27        |
| 113 (16L)                                      | 60°       | HK            | 8Bit (JPEG)        | Half        | Postshot             | 10K                         | 10M                | 06:23                         | 8,44                 | 25,15             | 84,91               | 83,92        |
|                                                | 100°      |               |                    |             |                      |                             |                    | 06:08                         | 9,26                 | 24,51             | 84,44               | 72,79        |
| 226 (16L)                                      | 60°+100°  |               |                    | Full + Half |                      |                             |                    | 15:39                         | 9,0                  | 27,15             | 84,92               | 79,95        |
| 113 (16L)                                      | 100°      | WK            | 8Bit (JPEG)        | Quarter     | Postshot             | 10K                         | 10M                | 03:06                         | 2,78                 | 30,20             | 88,05               | 67,01        |
|                                                |           | HK            |                    |             |                      |                             |                    | 03:08                         | 3,33                 | 27,93             | 85,47               | 77,27        |
| 113 (16L)                                      | 100°      | HK            | 8Bit (JPEG)        | Full        | Postshot             | 30K                         | 10M                | 35:15                         | 10,0                 | 28,43             | 85,50               | 82,58        |
|                                                |           |               | 8Bit (TIFF)        |             |                      |                             |                    | 35:19                         | 10,0                 | 27,70             | 85,40               | 76,68        |
|                                                |           |               | 16Bit (TIFF)       |             |                      |                             |                    | 36:39                         | 10,0                 | 28,79             | 86,15               | 83,64        |
| 113 (16L)                                      | 100°      | HK            | 8Bit (JPEG)        | Quarter     | Postshot             | 10K                         | 10M                | 03:08                         | 3,33                 | 27,93             | 85,47               | 77,27        |
|                                                |           |               |                    | Half        |                      |                             |                    | 06:08                         | 9,26                 | 28,14             | 86,12               | 80,17        |
|                                                |           |               |                    | Full        |                      |                             |                    | 08:02                         | 9,28                 | 27,97             | 85,44               | 79,39        |
| 113 (16L)                                      | 100°      | HK            | 8Bit (JPEG)        | Quarter     | COLMAP               | 10K                         | 10M                | 03:16                         | 3,56                 | 26,23             | 83,24               | 69,65        |
|                                                |           |               |                    |             | Postshot             |                             |                    | 03:08                         | 3,33                 | 27,93             | 85,47               | 77,27        |
|                                                |           |               |                    |             | Postshot ***         |                             |                    | 03:38                         | 4,0                  | 28,61             | 86,22               | 77,57        |
| 113 (16L)                                      | 100°      | HK            | 8Bit (JPEG)        | Full        |                      | 5K                          |                    | 02:42                         | 5,93                 | 26,11             | 83,41               | 69,46        |
|                                                |           |               |                    |             |                      | 10K                         |                    | 08:01                         | 9,28                 | 26,24             | 83,03               | 70,56        |
|                                                |           |               |                    |             |                      | 30K                         |                    | 35:02                         | 10,0                 | 26,39             | 82,65               | 71,46        |
|                                                |           |               |                    |             |                      | 60K                         |                    | 01:17:05                      | 10,0                 | 26,46             | 82,57               | 71,71        |
|                                                |           |               |                    |             |                      | 100K                        |                    | 02:19:38                      | 10,0                 | 25,65             | 82,41               | 71,39        |
|                                                |           |               |                    |             |                      | 150K                        |                    | 03:33:14                      | 10,0                 | 20,47             | 80,29               | 68,02        |
| 113 (16L)                                      | 100°      | HK            | 8Bit (JPEG)        | Full        | Postshot             | 30K                         | 1M                 | 12:15                         | 1,0                  | 28,60             | 86,12               | 79,86        |
|                                                |           |               |                    |             |                      |                             | 10M                | 35:15                         | 10,0                 | 28,43             | 85,50               | 82,58        |
|                                                |           |               |                    |             |                      |                             | 20M                | 01:36:50                      | 20,0                 | 28,36             | 85,31               | 83,11        |

\* Das Training wurde in Postshot mit dem Radiance Field Profile „Splat3“ durchgeführt

\*\* Es wird jeweils der Durchschnittswert aus vier (4L) bzw. zwei (8L) Modellen errechnet

\*\*\* Für die SfM-Berechnung wurde die Einstellung „Single Lens & Focal Length“ aktiviert

### 3.3.6 Subjektiver Vergleich der visuell wahrgenommenen Modellqualität

#### a) Anzahl der Aufnahmen

Solange eine Perspektive eingenommen wird, die sich auf einer der Aufnahmeachsen des Vier-Winkel-Modells befindet, werden die Objekte in ihrer Form und Position korrekt dargestellt. Wenn sich das Sichtfeld allerdings von einer der vier Achsen wegbewegt, verlieren die Gegenstände an Plastizität und die Struktur der Szene fällt in sich zusammen. Dies äußert sich in unorganisierten, frei im Raum schwebenden Splats, insbesondere im Umfeld der Glas- und Metallflächen. Es zeigt sich, dass die fehlenden Perspektiven durch den Algorithmus von Gaussian Splatting für eine vollständige 3D-Repräsentation nicht ausreichend kompensiert werden können, wenn die vertikalen Aufnahmewinkel eine Schrittweite von  $90^\circ$  aufweisen.

Mit acht Aufnahmewinkeln und einer Schrittweite von  $45^\circ$  verbessert sich die Qualität der Rekonstruktion deutlich. Die Objekte sind aus allen Perspektiven erkennbar, sodass die Repräsentation als geometrisch konsistent bezeichnet werden kann. Betrachtet man die Gegenstände allerdings im Detail, lassen sich Ungenauigkeiten in der Position und Ausrichtung der Splats beobachten, was zu einer allgemeinen Unschärfe führt. Zudem trüben freistehende Splats im leeren Raum den Gesamteindruck.

Das 16-Winkel-Modell erreicht wie zu erwarten die beste Modellqualität. Die Objekte auf dem Tisch werden aus allen Perspektiven geometrisch korrekt dargestellt und nur noch von kleinen schwebenden Splats im nahen Umfeld ummantelt. Detaillierte Strukturen wie die Zahlen und Symbole auf dem Geldschein wirken allerdings nach wie vor unscharf. Die in Teilen korrekt dargestellten Spiegelungen des Vier-Winkel-Modells sind einer grauen, flächigen Darstellung mit geringfügig blickwinkelabhängiger Farbänderung gewichen, die mit der Vielzahl von Reflexionen aus verschiedenen Winkeln zu begründen ist.

#### b) Sichtfeld

Das Modell auf Basis der Bilddaten mit einem  $60^\circ$ -Sichtfeld erreicht einen enorm hohen Detailgrad und ist im Gegensatz zum  $100^\circ$ -Modell dazu imstande, filigrane Strukturen beinahe fotorealistisch zu rekonstruieren. Besonders deutlich wird der Qualitätsunterschied bei der Darstellung des Geldscheins, aber auch transparente und reflektierende Objekte wie das Weinglas oder die Metallkanne werden deutlich besser rekonstruiert. Betrachtet man vom Tisch aus allerdings die umliegenden Bereiche im Raum, so zeigt sich, dass aus dem  $100^\circ$ -Datensatz eine geometrisch korrekte Darstellung der Umgebung hervorgeht, während die Rekonstruktion mit einem Sichtfeld von  $60^\circ$  zu geometrischen Fehlern und Artefakten tendiert. Strukturen, die nur aus wenigen Blickwinkeln erfasst wurden, laufen Gefahr, je nach Perspektive an der falschen oder an mehreren Positionen im Raum dargestellt zu werden. Die Ursache ist weniger auf den Trainingsprozess als vielmehr auf die fehlerhafte Berechnung der Kamerapositionen durch den SfM-Algorithmus zurückzuführen.



Abbildung 11: Exemplarische Bildausschnitte der gerenderten Videosequenzen (oben:  $FoV \approx 60^\circ$ , unten:  $FoV \approx 100^\circ$ )

Der Vergleich macht deutlich, dass ein kleineres Sichtfeld für eine fotorealistische Rekonstruktion grundsätzlich von Vorteil ist. Allerdings sorgt ein weites Sichtfeld dafür, dass die Aufnahmen besser zueinander ausgerichtet werden können, wodurch die geometrische Robustheit der Szene steigt.

In einer weiteren Rekonstruktion werden die Datensätze beider Sichtfelder für das Training eines Modells kombiniert, wobei die Aufnahmen mit einem Sichtfeld von  $100^\circ$  in diesem Fall in der originalen Auflösung verwendet werden. Die Vermutung, dass sowohl die Schärfe in den Details als auch die geometrische Robustheit erhalten bleiben und somit zu einem noch besseren Gesamtergebnis führen könnten, lässt sich in weiten Teilen bestätigen. Strukturen werden nahezu mit der gleichen Präzision und Schärfe dargestellt wie im  $60^\circ$ -Modell, während falsche Positionierungen und Artefakte in der Umgebung des Tisches ausbleiben.

### c) Bildkontrast

Die Modelle mit Aufnahmen verschiedener Kontrastprofile zeigen nur minimale Unterschiede in der Rekonstruktionsqualität. Sowohl die geometrische Präzision als auch die Darstellung feiner Strukturen weisen in beiden Varianten die gleiche Ausprägung auf. Der kontrastreiche Bilddatensatz führt zu einer natürlichen Farbdarstellung, während die Szene im kontrastarmen Modell flach und ausgewaschen wirkt.

#### d) Farbtiefe & Dateiformat

Ein Vergleich der Ergebnisse aus den Bilddatensätzen mit einem Farbraum von 8 Bit zeigt, dass das JPEG-Format detaillierte Strukturen trotz der Bildkompression deutlich präziser darstellt als das TIFF-Format. Ohne genauere Untersuchungen lässt sich das beobachtete Phänomen nicht erklären. Die Verdopplung der Farbtiefe des TIFF-Formats von 8 auf 16 Bit pro Kanal hat zur Folge, dass die Rekonstruktion sich wiederum verbessert und eine vergleichbar gute Qualität wie das Modell des JPEG-Datensatzes erreicht. Die Darstellung der Farben ist in allen drei Modellen konstant.

#### e) Auflösung

Der auf ein Viertel der ursprünglichen Auflösung reduzierte Datensatz generiert ein Modell mit erwartungsgemäß geringer Detailschärfe und Splats in einer visuell wahrnehmbaren Größenordnung. Es ist im Vergleich zu den anderen Modellen zweifelsfrei als das mit der geringsten Rekonstruktionsqualität einzustufen. Betrachtet man die Modelle mit voller und halber Auflösung, fällt die Beurteilung etwas differenzierter aus. Während bei halber Auflösung insbesondere die feingliedrige, aber dennoch erkennbare Struktur der Splats ins Auge fällt, verschwimmen die Splats im Modell mit voller Auflösung zu einer homogenen Masse, die im direkten Vergleich allerdings einen etwas unschärferen Eindruck erzeugt.

#### f) SfM-Initialisierung

Die untersuchten Methoden zur Initialisierung der Rekonstruktion mit Structure-from-Motion zeigen keine gravierenden Unterschiede in ihrer Auswirkung auf die Modellqualität der Testszene. Allerdings kann festgestellt werden, dass die Einstellung „Single Lens & Focal Length“ (SFL) aus den meisten Betrachtungswinkeln zu einer Verschlechterung gegenüber dem Modell mit den Standardeinstellungen von Postshot führt. Der Qualitätsverlust äußert sich unter anderem in einer breiter gestreuten Verteilung der Splats um die Objekte herum, die Schärfe nimmt ebenfalls etwas ab. Im direkten Vergleich der Ergebnisse mit COLMAP- und Postshot-Initialisierung lassen sich nur marginale Unterschiede feststellen, die auf die Verwendung desselben Algorithmus mit angepassten Einstellungen zurückzuführen sind.



Abbildung 12: Verschiedene SfM-Initialisierungen, v.l.n.r. COLMAP, Postshot, Postshot mit „Single Lens & Focal Length“

#### **g) Anzahl der Trainingsiterationen**

Bereits eine Anzahl von 5.0000 Trainingsschritten führt dazu, dass das Modell geometrisch konsistent ist, während die Gegenstände allerdings unscharf dargestellt werden und wenig Struktur enthalten. Mit der Erhöhung der Trainingsiterationen erhöht sich die Schärfe in den Details sukzessive. Wie in der Gegenüberstellung in Abbildung 12 zu sehen ist, verbessert sich die Darstellung feiner Strukturen am stärksten im Bereich von 5.000 zu 30.000 Trainingsschritten. Im weiteren Verlauf kann die Repräsentation der Details bis zu einer Schrittzahl von 150.000 nur noch wenig optimiert werden und das unter erheblicher Verlängerung der Trainingszeit. Im Allgemeinen hat die Steigerung der Detailtreue den Nebeneffekt, dass nicht nur die gewünschten Bereiche, sondern auch die frei im Raum schwebenden Splats nachgeschärft und deutlicher sichtbar werden. In der Gesamtbetrachtung darf allerdings nicht unerwähnt bleiben, dass die maximale Anzahl an Splats für das Training auf 10 Millionen begrenzt wurde. Bereits bei 30.000 Trainingsiterationen ist dieses Limit erreicht, sodass für dieses und die folgenden Modelle davon ausgegangen werden muss, dass das Potenzial in der Rekonstruktionsqualität nicht gänzlich ausgeschöpft wurde.



Abbildung 13: Einfluss der Anzahl an Trainingsiterationen auf die Modellqualität

### **i) Maximale Splat-Anzahl**

Es zeigt sich, dass das Modell mit einer Millionen Splats bereits ausreicht, um eine geometrisch konsistente Rekonstruktion in akzeptabler Qualität zu generieren. Die strenge Limitierung führt allerdings dazu, dass im Trainingsfortschritt schon sehr früh die maximale Splat-Anzahl erreicht wird. In den nachfolgenden Iterationen können die Splats dann nicht mehr weiter aufgeteilt werden, wodurch der Detailgrad des Modells begrenzt ist und der Eindruck einer weichgezeichneten Darstellung entsteht. Bei einem Limit von zehn Millionen Splats entsteht eine feinere Auflösung und die Detailinformationen aus dem Bilddatensatz bleiben deutlich besser erhalten. In diesem Stadium der Verdichtung haben die meisten Gaussian Surfels bereits eine Größe erreicht, die sie nur aus kurzer Distanz als einzelne Splats erkennbar werden lässt. Eine Verdopplung auf 20 Millionen Splats wirkt sich dementsprechend in geringem Maße auf die Darstellung der Objekte aus und verbessert das Gesamtergebnis marginal.

### 3.4 Diskussion und Auswertung der Ergebnisse

Aus den Ergebnissen der metrischen Berechnungen und des visuellen Vergleichs lassen sich für einige der untersuchten Faktoren eindeutige Schlussfolgerungen hinsichtlich ihres Einflusses auf die Rekonstruktionsqualität ziehen. Es zeigen sich allerdings auch Widersprüche in den Resultaten und Ungenauigkeiten in der Methodik, weshalb für andere Faktoren die abschließende Bewertung der Einflussnahme nur eingeschränkt gültig ist. Insbesondere bei der Gegenüberstellung der beiden Vergleichsmethoden kommt es zu Diskrepanzen in der Deutung. Die Limitationen des Vergleichs werden abschließend diskutiert.

#### a) Anzahl der Aufnahmen

Die Anzahl der Aufnahmen bestimmt die geometrische Konsistenz der Modelle maßgeblich, wie aus dem visuellen Vergleich hervorgeht. Wenn mehr Aufnahmen aus verschiedenen Winkeln in das Modelltraining eingebracht werden, kann die Geometrie und Struktur der Objekte durch die Splats besser repräsentiert werden. Blickt man auf die Ergebnisse der Metrik-Berechnung, laufen diese der subjektiven Bewertung zuwider, denn das Vier-Winkel-Modell erreicht für alle drei Metriken die höchste Übereinstimmung mit dem originalen Bilddatensatz. Dies lässt sich durch die Betrachtung der gerenderten Bilder bestätigen. Es muss bei der Deutung der metrischen Kalkulationen jedoch berücksichtigt werden, dass der Vergleich auf die fotografierten Kameraperspektiven reduziert ist und die Gesamtheit der dreidimensionalen Szene nicht erfasst werden kann. Beim Vergleich von Modellen mit unterschiedlich hoher Anzahl an Aufnahmen erscheint die Metrikberechnung daher wenig zielführend. In einer kugelförmigen Bildabdeckung, die wie im Test-Szenario für eine Rekonstruktion von Objekten konzipiert ist, sollte die horizontale Schrittweite entsprechend dem Acht-Winkel-Modell maximal  $45^\circ$  und idealerweise weniger als  $30^\circ$  betragen. Bei der Verkleinerung des Sichtfelds ist eine weitere Reduzierung der Schrittweite empfehlenswert, um eine ausreichend große Überlappung in den Aufnahmen sicherzustellen.

Aus den Untersuchungsergebnissen lässt sich für die Erstellung virtueller Sets keine direkte Anleitung für den optimalen Aufnahmeprozess schlussfolgern, da statt eines Objekts ein gesamter Raum bzw. eine Umgebung flächig erfasst werden muss, wodurch das Verfahren komplexer wird. Dennoch kann das Prinzip der maximalen Winkelschrittweite sowohl in horizontaler als auch vertikaler Richtung als Orientierung dienen und wird in der Erstellung der virtuellen Szene wieder aufgenommen. Abschließend lässt sich hinsichtlich der Anzahl an Aufnahmen für die Rekonstruktion einer beliebigen Szene durch den Versuch bestätigen, was bereits vorher angenommen worden war.

Jeder neue Blickwinkel fördert die geometrische Konsistenz und somit die Qualität eines 3D-Modells, sofern

- die Prämisse der Starrheit der Körper berücksichtigt wird, d.h. zwischen den Aufnahmen keine Verformung, Drehung oder Positionsänderung der Objekte stattfindet
- die Lichtbedingungen unverändert bleiben
- alle Kameraeinstellungen beibehalten werden
- die Bewegungs- und Tiefenunschärfe so gering wie möglich gehalten wird
- das Foto genügend Bildmerkmale enthält, die dem SfM-Algorithmus eine geometrisch korrekte Ausrichtung der Kameraposition in Bezug zu den anderen Aufnahmen ermöglichen

Je mehr Aufnahmen für das Training verwendet werden, die den genannten Kriterien entsprechen, desto besser kann das Rekonstruktionsergebnis werden.

### **b) Sichtfeld**

Die Wahl des Sichtfelds der Kamera hat einen starken Einfluss auf die Qualität der Rekonstruktion. Ein größerer Winkel führt zu mehr Überlappung zwischen den Aufnahmen und steigert die geometrische Präzision, ein kleineres Sichtfeld erhöht hingegen die Schärfe in den Details. Es ist daher sinnvoll, den spezifischen Anwendungsfall und die räumliche Größe der Rekonstruktion bei der Wahl des Sichtfelds zu berücksichtigen, um einen Kompromiss aus robuster sowie effizienter Erfassung auf der einen und hoher Detailauflösung auf der anderen Seite zu finden. Ein Bildwinkelbereich zwischen  $60^\circ$  und  $100^\circ$  erscheint in jedem Fall angemessen. Der doppelte Import der Fotos in die Trainingssoftware, bei der einer der Bilddatensätze einen kleineren Ausschnitt des anderen zeigt, stellt sich als eine geschickte Möglichkeit heraus, hohe Robustheit und Präzision ohne Mehraufwand in einem Modell zu vereinen. Dieses Phänomen ist vermutlich darauf zurückzuführen, dass der Verdichtungsprozess des Gaussian-Splatting-Algorithmus innerhalb des Bilddatensatzes keine Hierarchisierung vornimmt und jede Aufnahme dieselbe Relevanz für das Modelltraining besitzt. Durch die wiederholte Einspeisung von Bildinformationen lässt sich der Fokus des Algorithmus bei der Rekonstruktion gewissermaßen auf bestimmte Bildbereiche lenken.

### **c) Bildkontrast**

Der Kontrast der Bilddaten hat im untersuchten Rahmen mit zwei verschiedenen Profilen weder auf die geometrische Robustheit noch auf die Schärfe der Details einen sichtbaren Einfluss. Die Metriken weisen jedoch in widersprüchlicher Weise darauf hin, dass es zwischen den Modellen Unterschiede in der Ähnlichkeit zu den jeweiligen Bilddaten, also Differenzen in der Rekonstruktionsqualität, gibt. Während die Berechnung von PSNR und SSIM für das kontrastarme Profil einen höheren Ähnlichkeitswert zu den Referenzaufnahmen ergibt, entsprechen nach dem VMAF-Score die kontrastreichen Aufnahmen deutlich eher der Referenz. Der Widerspruch lässt sich ohne weitere Untersuchungen weder auflösen noch erklären, weshalb die Schlussfolgerung auf dem wahrnehmungsbasierten Vergleich beruht. Allgemein lässt sich in Bezug auf den Kontrast

der Bilddaten festhalten, dass dieser keinen erkennbaren Einfluss auf die Rekonstruktionsqualität hat. Wichtiger ist hingegen, dass bei der Aufnahme und Bearbeitung der Bilddaten bereits die Kontrast- und Farbgebung erarbeitet wird, die für das spätere Modell gewünscht ist, denn die Pixelinformationen aus den Bilddaten werden im Modelltraining eins zu eins übernommen.

#### **d) Farbtiefe & Dateiformat**

Dem subjektiven Vergleich ist zu entnehmen, dass das JPEG-Format dem TIFF-Format bei einer Farbtiefe von acht Bit überlegen ist. Gestützt wird diese Beobachtung durch die berechneten Metriken, insbesondere durch den VMAF-Score, der im Modell mit Bilddaten im JPEG-Format deutlich höher ausfällt. Auch wenn das Modell des TIFF-Formats in der Verdopplung der Farbtiefe auf 16 Bit in allen drei Metriken das insgesamt beste Resultat hervorruft, ist der Unterschied zu JPEG visuell nicht wahrnehmbar. In Anbetracht der Tatsache, dass das TIFF-Format zehnmal größere Bilddateien und eine dreifach höhere Dateigröße im Modell erreicht, ist die Verwendung von 8-Bit-Dateien im JPEG-Format zu empfehlen.

#### **e) Auflösung**

Aus dem metrischen Vergleich geht hervor, dass Bilddaten mit einem Viertel der ursprünglichen Auflösung das schlechteste Resultat erzeugen, während die halbe Auflösung zu den besten Ergebnissen führt. Die Metrik-Werte decken sich größtenteils mit den Beobachtungen, sodass sich die Frage stellt, weshalb die volle Bildauflösung nicht zu einem präziseren Modell führt. Die Ursache wird darin vermutet, dass die mit 10.000 Schritten verhältnismäßig niedrige Anzahl der Trainingsiterationen den Verdichtungsprozess beendet, bevor überhaupt der Detailgrad erreicht werden kann, ab dem das Modell von der doppelten Auflösung profitiert. Dies wäre in weiteren Untersuchungen zu überprüfen. Grundsätzlich ist eine höhere Auflösung also förderlich, kann aber im Rahmen dieses Versuchs nur für den Bildauflösungsbereich bis ungefähr 3000x2000 Pixeln nachgewiesen werden. Für das Training mit zugeschnittenen Datensätzen als Ergänzung zu den Original-Aufnahmen ist eine hohe Auflösung in jedem Fall hilfreich.

#### **f) SfM-Initialisierung**

Wählt man in den Einstellungen im Modelltraining von Postshot aus, dass für die Aufnahmen nur eine Optik und Brennweite genutzt wurde (SFL), ist in den Rekonstruktionsergebnissen rechnerisch eine höhere Übereinstimmung mit den Originalaufnahmen festzustellen als ohne die Einstellung oder mit extern berechneten Kamerapositionen aus COLMAP. Visuell ist im Gegenteil zu beobachten, dass die SFL-Einstellung zu einer Verschlechterung der Ergebnisse führt. Dies widerspricht der Vermutung, die Einstellung würde die Kamerapositionen präziser berechnen können, weil die Gleichungssysteme des SfM-Algorithmus mit konstanten Parametern gelöst werden. Ob sich dieser Effekt bei Aufnahmen mit voller Auflösung relativieren würde, kann an dieser Stelle nur vermutet werden. Abschließend betrachtet liegen im Rahmen der

Untersuchung keine empirischen Belege vor, dass die Einstellung „Single Lens & Focal Length“ die Rekonstruktionsqualität verbessert. Das COLMAP-Modell erreicht den Metriken nach zu urteilen eine geringere Modellqualität als das Postshot-Modell, während sich visuell nur marginale Unterschiede erkennen lassen. Aus diesem und nicht zuletzt auch aus praktischen Gründen ist die SfM-Initialisierung in Postshot daher zu bevorzugen.

#### **g) Anzahl der Trainingsiterationen**

Aus dem subjektiven Vergleich geht hervor, dass eine höhere Anzahl an Trainingsschritten die Modellqualität verbessert. Die Detailschärfe nimmt allerdings nicht linear zu, sondern erhöht sich bis zu dem Modell mit 30.000 Trainingsiterationen relativ stark und im weiteren Verlauf immer weniger. Der Einfluss des gesetzten Splat-Limits von zehn Millionen auf die Modellqualität kann zwar nicht abschließend beurteilt werden, würde allerdings aufgrund immens steigender Trainingszeiten ohnehin nicht praktikabel sein. Die metrischen Berechnungen zeigen hingegen wenig Übereinstimmung mit den Beobachtungen. Der PSNR- und VMAF-Score lassen zwar auf einen moderaten Anstieg der Qualität bei Zunahme der Trainingsschritte schließen, jedoch wird der deutliche Qualitätsanstieg von 5.000 auf 30.000 Schritte bei gleichzeitiger Detektion eines nicht erkennbaren Qualitätsverlusts bei 150.000 Schritten vernachlässigt. Schlussendlich steht fest, dass insbesondere die Schärfe von Details und filigranen Strukturen mit zunehmender Anzahl der Trainingsiterationen verbessert werden kann. Ein Wert zwischen 30.000 und 60.000 resultiert in visuell hochwertigen Ergebnissen, die eine ausreichend gute Auflösung in den Details aufweisen. Unter Inkaufnahme einer deutlich steigenden Trainingsdauer kann eine weitere Schritterhöhung zu noch präziseren Rekonstruktionen führen.

#### **h) Maximale Splat-Anzahl**

Die Werte von PSNR und SSIM decken sich nicht mit dem subjektiven Vergleich, sondern verorten die beste Qualität im Modell mit einer Millionen Splats. Der VMAF-Score stützt hingegen die Beobachtung, dass die Szene mit steigender Splat-Anzahl besser repräsentiert werden kann und eine höhere Qualität erreicht. Gleichzeitig muss jedoch berücksichtigt werden, dass sich die Trainingsdauer des Modells durch die Verdopplung der Splat-Anzahl von zehn auf 20 Millionen verdreifacht und die Dateigröße sogar versechsfacht. Im Optimalfall wird also ein Kompromiss zwischen einer detailreichen Szenenrepräsentation und einer angemessenen Ressourcennutzung gefunden, der sich je nach Szenengröße in unterschiedlichen Splat-Limits äußert. Eine Anzahl von zehn Millionen stellt einen soliden Wert für eine Rekonstruktion in der Größenordnung der Testszene dar.

Betrachtet man die zum Teil signifikanten Unterschiede in der Bewertung der Rekonstruktionsqualität mit subjektiven und objektiven Methoden, muss die Aussagekraft hinter den verwendeten Metriken durchaus in Frage gestellt werden. In der Vorstellung der Metriken wurde bereits darauf hingewiesen, dass die zugrundeliegende Art des Vergleichs die Modelle auf die Perspektiven reduziert, die sich am Bilddatensatz orientieren. Dieser Ground-Truth-Ansatz ist gängige Praxis im Benchmarking von 3D-Rekonstruktionsverfahren und im Prinzip die einzige Möglichkeit, verschiedene Modelle in Relation zueinander zu setzen. Die eigenen Untersuchungen kommen zu dem Ergebnis, dass sich die visuelle Vergleichsmethode besser eignet

### 3.5 Zwischenfazit

Es wurden die Auswirkungen der Veränderung verschiedener Parameter und Einstellungen auf die Rekonstruktionsqualität von 3D-GS-Modellen untersucht. Es stellt sich heraus, dass die subjektive Methode, die die Ergebnisse des objektiven Vergleichs mit den Metriken PSNR, SSIM und VMAF lediglich validieren sollte, zur Beurteilung der Qualität eines 3D-GS-Modells besser geeignet ist als die metrischen Vergleiche.

Die Untersuchungen haben ergeben, dass in erster Linie eine hohe Aufnahmedichte entscheidend für ein geometrisch robustes und aus allen Blickwinkeln möglichst fotorealistisches Rekonstruktionsergebnis ist. Das für die Aufnahmen optimale Sichtfeld sollte einen Kompromiss aus hoher Robustheit bei 100° und scharfen Details bei 60° darstellen. Werden hingegen Aufnahmen mit großem Sichtfeld genutzt und der ursprüngliche Datensatz zusammen mit einer mittig zugeschnittenen Kopie dessen trainiert, lassen sich die Vorteile beider Bildwinkel zu einer deutlich besseren Rekonstruktion kombinieren. Ein weiterer Faktor mit großem Einfluss auf die Modellqualität ist die Anzahl der Trainingsschritte, die bei mindestens 30.000 liegen sollte. Die Splat-Anzahl sollte nicht zu klein gewählt werden, ein Wert um die zehn Millionen erscheint sinnvoll. Die Auflösung der Aufnahmen sollte möglichst hoch sein, hat aber weniger Einfluss als zunächst angenommen und müsste in weiteren Untersuchungen genauer betrachtet werden. Die Änderung des Bildkontrasts und die Variation der SfM-Initialisierung hat auf die Rekonstruktionsqualität einen eher geringen Einfluss. Die Aufnahmen sollten im JPEG-Format vorliegen. Mit der Einhaltung aller genannten Kriterien ist es möglich, fotorealistische Rekonstruktionen zu erstellen.

## 4 Vergleich virtuell und real produzierter Filmsequenzen

### 4.1 Rekonstruktion einer Szene für die virtuelle Produktion

In der Regel werden virtuelle Produktionsmethoden dazu genutzt, Charaktere und Objekte in einer Umgebung zu zeigen, die nicht als physischer Ort im materiellen Sinne existiert, sondern als Computermodell konstruiert wurde. Die Gestaltungsfreiheit im Digitalen ist an dieser Stelle das Hauptmotiv für die Entscheidung, virtuelle Techniken zur Produktion von Inhalten zu verwenden. Gaussian Splatting als Werkzeug zur Erstellung virtueller Sets verfolgt hingegen den Ansatz, bereits existierende Umgebungen digital zu kopieren, was auf den ersten Blick keinen gestalterischen Freiraum schafft. Die Vorteile für den kreativen Prozess liegen vielmehr in der Befreiung von zeitlichen und physischen Zwängen begründet. Damit ist zum einen gemeint, dass ein virtueller Dreh im Studio nicht durch drehortbedingte Einschränkungen wie die Wetter- und Lichtverhältnisse, Personen- und Straßenverkehr, Restriktionen durch den Gesetzgeber oder die Motivgeber\*innen etc. beeinflusst wird. Zum anderen entstehen durch die Auflösung physischer und zeitabhängiger Grenzen eines realen Sets neue Möglichkeiten in der Bildgebung. Das Beispiel eines Remake des „Bullet-Time“-Effekts aus dem Film „Matrix“ (Corridor Crew, 2025) mit einer virtuellen Gaussian-Splatting-Szene zeigt auf anschauliche Weise, wie Kamerafahrten und Slow-Motion-Effekte realisiert werden können, die vor Ort nur unter großem Aufwand oder gar nicht umsetzbar wären.

Der Anspruch des eigenen Modellversuchs besteht darin, ein virtuelles Set zu generieren, das die Vorzüge des Konzepts aufzeigt. Dafür erscheint eine Szene sinnvoll, die in einem realen Set nur unter erschwerten Bedingungen zu realisieren wäre. Eine U-Bahn-Station als Ort des öffentlichen Lebens ist ein geeignetes Beispiel für eine Location, die nur mit hohem organisatorischen, logistischen und finanziellen Aufwand für einen Dreh nutzbar gemacht werden kann. Abgesehen davon liefert eine solche Umgebung gute Testbedingungen durch konstante Lichtverhältnisse, permanente Zugänglichkeit und zu später Stunde wenig Störfaktoren. Nachdem in den vergangenen Kapiteln anhand der Testszene gezeigt wurde, wie sich die visuelle Qualität eines Modells durch verschiedene Einstellungen optimieren lässt, sollen diese Erkenntnisse für die Rekonstruktion der U-Bahn-Station als Orientierung dienen.

Bezogen auf das Capture-Verfahren liegt der wesentliche Unterschied eines virtuellen Sets im Vergleich zur untersuchten Testszene darin, dass der Fokus nicht auf einen bestimmten Bereich im Raum liegt, sondern dieser in seiner Gesamtheit repräsentiert werden soll. Die Strategie der halbkugelförmigen Positionierung der Aufnahmeperspektiven führt also im Fall der U-Bahn-Station nicht zum gewünschten Ergebnis. Im Vorfeld wurde ein Testdurchlauf durchgeführt, bei dem mit einer Schrittweite von etwa einem halben Meter alle Flächen eines Bahnsteigbereichs der U-Bahn-Station auf verschiedenen Höhen abfotografiert wurden. Die Ergebnisse erreichen bereits eine akzeptable Rekonstruktionsqualität, enthalten allerdings viele frei schwebende Gaussians in näherer Distanz zu den Flächen, wodurch der Eindruck von Unschärfe entsteht. Ein Erklärungsansatz dafür ist, dass dem Algorithmus für eine

geometrisch präzisere Rekonstruktion räumlich variierte Perspektiven auf denselben Bereich fehlen. Deutlich wird dieses Problem durch die Vorstellung, in einer Rekonstruktion stünden die Bildebenen aller Aufnahmeperspektiven parallel zu der Fläche, die modelliert werden soll. In diesem Fall kann der Algorithmus die Splats aufgrund fehlender Tiefeninformationen nicht in der korrekten Distanz im Raum positionieren. Die Anzahl variierteter Blickwinkel auf jeden Bereich des Modells sollte also erhöht werden, um die frei schwebenden Splats zu minimieren. Diese Überlegung lässt sich mit den Ergebnissen belegen, die bei der Untersuchung der Anzahl der Aufnahmen in der Testszene erzielt wurden. Nun ist es weder effizient noch praktikabel, alle Bereiche nach dem „Halbkugelprinzip“ abzufotografieren. Für eine bessere Rekonstruktionsqualität ist daher ein intelligenteres Verfahren erforderlich.

Die erwünschte Effizienz kann an dieser Stelle durch einen Strategiewechsel herbeigeführt werden. Der bisherige Ansatz im Capture-Prozess folgte der intuitiven Annahme, eine Kamera könne die Umgebung wie ein LiDAR-Scanner dreidimensional erfassen. Diese Auffassung führt dazu, dass die Blickrichtung und Position der Kamera so ausgerichtet werden, dass die in der Rekonstruktion gewünschten Bereiche in den Aufnahmen zu sehen sind. Zielführender ist hingegen der Ansatz, die Perspektive danach zu wählen, von wo aus die Umgebung in der Rekonstruktion betrachtet werden soll. Gibt es einen festgelegten Bereich, aus dem die virtuelle Kamera das Modell sehen soll, so kann das Aufnahmeverfahren auf diesen begrenzt werden. Im Idealfall müsste in diesem definierten Raum von jedem Punkt aus eine digitale 360°-Rundumsicht erstellt werden. Eine Möglichkeit, sich dieser Idee anzunähern und gleichzeitig eine Systematik zu etablieren, die praktisch umgesetzt und reproduziert werden kann, ist die Aufnahme in Form von Säulen.

In der U-Bahn-Station ist der Bereich für die Aufnahmen durch den Bahnsteig bereits begrenzt. Um Punkte mit gleicher Distanz zueinander zu bestimmen, hilft es, sich ein Raster auf dem Boden vorzustellen. Von diesen Punkten aus werden in mehreren Höhen acht Fotos mit einer rotierenden Winkelschrittweite von 45° erstellt. Aus einer mittleren Höhenposition werden zusätzlich jeweils acht Fotos mit einer Neigung nach schräg oben und schräg unten gemacht, um die Decke und den Boden zu erfassen. Der gesamte Raum wird somit aus einer Vielzahl an Rundumsichten sukzessive erfasst. In Postshot können die Kamerapositionen auf Basis des Structure-from-Motion-Algorithmus dargestellt werden.

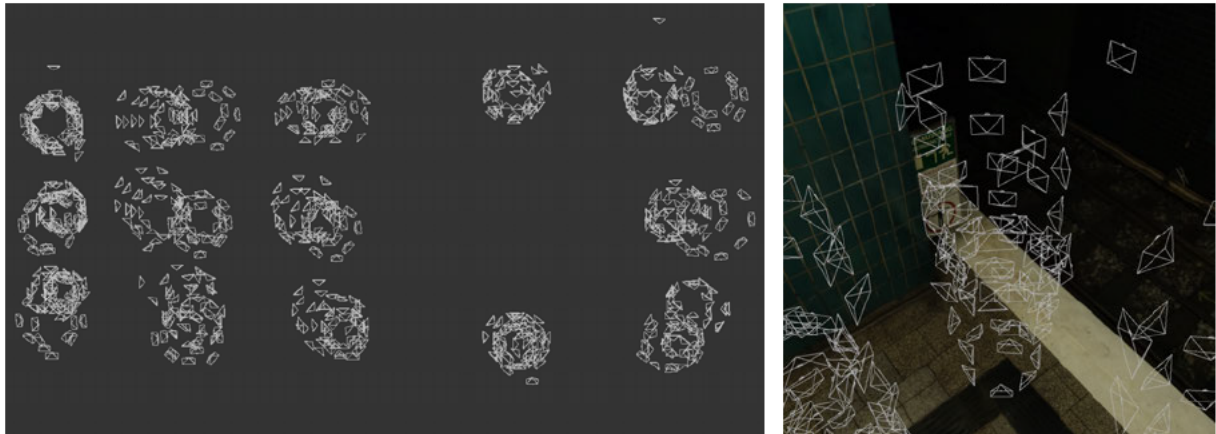


Abbildung 14: Anordnung der Säulen von oben (links), Anordnung der Kamerapositionen als Säule (rechts)

Die Fotos für das Modelltraining wurden mit folgenden Spezifikationen aufgenommen:

- Kamera: Fujifilm X-H2S
- Objektiv: Sigma 10-18mm F2.8 DC DN Contemporary
- Format: JPEG mit 8 Bit Farbtiefe pro Kanal
- Auflösung: 6240 x 4160 Pixel
- Brennweite: 10mm (APS-C)
- Film-Emulation (Farbprofil): PROVIA/Standard
- Fokus-Einstellung: Einzel-AF
- Blende: f/6.4
- ISO: 2500
- Belichtungszeit: 1/100 Sekunde

Die Rekonstruktion hat folgende Spezifikationen:

- Anzahl der Fotos: 838
- Radiance Field Profile: „Splat3“
- Anzahl der Trainingsiterationen: 30.000
- Splat-Limit: 3 Millionen
- SfM-Initialisierung: Postshot mit Standardeinstellungen

Der Bilddatensatz und die Rekonstruktion sind im Anhang unter „Modell\_U-Bahn-Station“ zu finden.

## 4.2 Konzept und Aufbau des virtuellen Studio-Setups

In einem virtuellen Studio verschmelzen das reale und das digitale Set mit verschiedenen Technologien zu einem finalen Output. Die bekannteste Methode im Kontext der virtuellen Filmproduktion verwendet eine Wand aus LED-Panels für die Darstellung des digitalen Inhalts und nimmt diese als Szenenhintergrund direkt in der Kamera auf. Eine Alternative, die in dieser Arbeit Anwendung findet, ist eine klassische Green-Screen-Produktion gepaart mit einer Echtzeit-Visualisierung des finalen Programmbildes.

Der Modellversuch stellt den einfachsten Anwendungsfall im Kontext von Virtual Production dar. Eine Schauspielerin steht im Studio und wird so gefilmt, dass die Umgebung um sie herum als Teil der Szene erkennbar ist. Hinter ihr befindet sich ein Green-Screen, der durch die gleichmäßige Ausleuchtung ein Keying bzw. Freistellen der Person ermöglicht. Die Kamerapositionen werden mithilfe eines Tracking-Systems über das Netzwerk in Echtzeit an die Game-Engine eines Rechners übertragen, die aus den Daten ein Live-Bild der virtuellen Szene mit den entsprechenden Kamerapositionen rendert. Die virtuelle Kamera wurde in der Vorbereitung mit den Kalibrierungsdaten des Studiokamera-Setups konfiguriert, weshalb sie die gleichen Bildeigenschaften besitzt. In einer Compositing-Hardware laufen die beiden Videosignale zusammen und das maskierte Kamerabild wird auf das virtuelle Bild gelegt, um ein Programmbild zu erhalten.

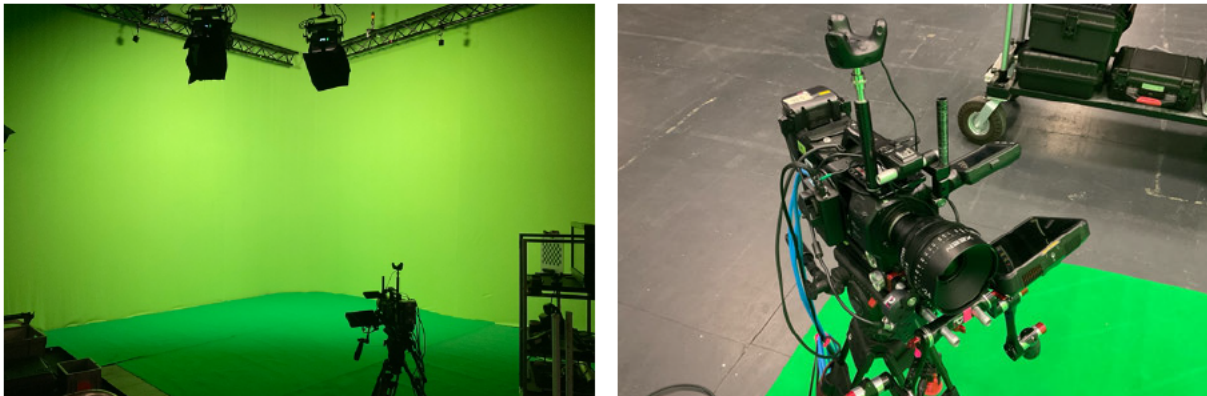


*Abbildung 15: Green-Screen-Setup im Produktionslabor (links), Kamera-Setup mit Tracking-System (rechts)*

Folgende Hardware wurde im Aufbau des Setups verwendet:

- Kamera: Sony PXW-FS7 II
- Objektiv: XEEN Cine Prime 24mm T1.5
- Video-Rekorder: Odyssey 7Q+ in Kombination mit RAW-Modul (XDCA-FS7)
- Trackingsystem: VIVE Mars CamTrack & FIZTrack
- Compositing-Hardware: Blackmagic Design Ultimatte HD
- Rechner: „Unreal PC 1“

## 4.3 Produktion virtueller Filmsequenzen

Das Live-Compositing wurde mit dem Zweck eingerichtet, die Szene vor Ort anpassen und erste Ergebnisse zur Kontrolle einsehen zu können. Für die finale Beurteilung wird das Keying allerdings in der Postproduktion mit DaVinci Resolve durchgeführt. Die Entscheidung rührt daher, dass der Live-Betrieb nur mit Videosignalen in Full-HD-Auflösung realisierbar ist, während der Standard in der Filmproduktion Ultra HD ist. Zudem lassen sich mit einer höheren Auflösung bessere Ergebnisse im Keying erzielen. Für eine hohe Videoqualität wird der visuell verlustfreie Codec ProRes HQ verwendet. Dieser lässt sich allerdings nicht direkt mit der Sony FS7 aufnehmen, weshalb ein RAW-Modul in Kombination mit einem externen Videorekorder angeschlossen werden muss. Im Ergebnis ist leider festzustellen, dass dieser in der Konvertierung des RAW-Inputs der Kamera Bildrauschen hinzufügt. Dazu kommt eine unzureichende Ausleuchtung der Person, weshalb in der Postproduktion eine Rauschreduzierung auf das Video gelegt werden muss.

Das Kamerabild hat folgende Spezifikationen:

- Auflösung: 4096 x 2160 Pixel (DCI-4K)
- Bildrate: 25 Bilder pro Sekunde
- Codec: Apple ProRes 422 HQ
- Farbraum: SGamut3.C
- Log-Profil: S-Log3

Für den Teil, der die virtuelle Produktion des Modellversuchs ausmacht, wurde die Software „Unreal Engine“ in der Version 5.5 genutzt. Postshot bietet für verschiedene Versionen der Game-Engine ein eigenes Plug-in an, mit dem sich die Modelle im proprietären Datenformat .psht in ein Projekt importieren lassen. Der Nachteil gegenüber anderen Plug-ins, wie denen von „Volinga“ oder „XGRIDS“, ist allerdings, dass die Modelle in Unreal Engine nicht weiter bearbeitet werden können. Das Rendern virtueller Hintergründe für die Zusammenführung mit den Kameraaufnahmen ist in Unreal Engine komfortabel gestaltet. Bei jedem Take müssen dafür die Kamerapositionen aufgezeichnet werden, die dann für das Rendering eines Videos mit beliebigen Spezifikationen verwendet werden.

Es wurde bereits darauf hingewiesen, dass die gefilmten Sequenzen, in denen eine Person im Vordergrund zu sehen ist, schlecht ausgeleuchtet sind. Im Nachgang können drei Fehler in der Beleuchtung identifiziert werden, die zu suboptimalen Ergebnissen führten. Zum einen wurde die Schauspielerin nicht stark genug ausgeleuchtet, was zu einem erhöhten Bildrauschen führt. Zum anderen stellte die starke Reflexion grünen Lichts (Green Spill) auf dem Körper der Person ein Problem für die Maskierung dar (siehe Abbildung 15). Eine fehlende Separation von Vorder- und Hintergrund sowie eine zu starke Beleuchtung des Green-Screens können als Ursachen ausgemacht werden. Das dritte Problem liegt in der fehlenden Anpassung der Personenbeleuchtung an die Lichtbedingungen der virtuellen Szene. Der Ordner „Virtual\_Production“ im Anhang beinhaltet exemplarisch zwei Sequenzen, die eine Person im virtuellen Set der U-Bahn-Station zeigen. Die angesprochenen Probleme in der Beleuchtung sowie eine zeitliche Asynchronität zwischen Vorder- und Hintergrund sind zu erkennen.



Abbildung 16: Standbild der Studiokamera mit Hervorhebung des Green Spill

#### 4.4 Vergleich der virtuellen und vor Ort produzierten Filmsequenzen

Um die Rekonstruktionsqualität des U-Bahn-Modells abschließend beurteilen zu können, werden virtuell generierte mit vor Ort produzierten Filmsequenzen in ihren visuellen Eigenschaften verglichen. Der Vordergrund der Bildkomposition in Form der schauspielenden Person spielt für die Bewertung von Gaussian Splatting als Werkzeug zur Erstellung virtueller Sets keine direkte Rolle und wird daher im Vergleich nicht berücksichtigt. Um die virtuelle U-Bahn-Station mit authentischen Kamerabewegungen aus unterschiedlichen Perspektiven zu zeigen, werden die Positions- und Fokusdaten eines realen Kamera-Setups durch das Tracking in der virtuellen Szene gespiegelt und für das spätere Rendering aufgezeichnet. Zur räumlichen Orientierung wird an die Kamera ein Monitor angebracht, der die virtuelle Szene in Echtzeit darstellt und dem Operator die Möglichkeit gibt, durch reale Bewegungen die virtuelle Kameraperspektive nach seinen Vorstellungen zu verändern. Die virtuelle Kamera übernimmt dabei die Eigenschaften des Kamera-Setups, mit dem auch die Sequenzen vor Ort aufgenommen wurden.

- Kamera: Fujifilm X-H2S
- Objektiv: Tamron 17-70mm f/2.8 Di III-A VC RXD
- Brennweite: 17mm (APS-C)
- Bildauflösung: 3840 x 2160 Pixel (4K UHD)
- Bildrate: 25 Bilder pro Sekunde
- Codec: Apple ProRes 422 HQ

Es wurden vier Sequenzen mit einer Dauer von etwa zehn Sekunden erstellt, die daraufhin vor Ort nachgedreht wurden. Die Ergebnisse sind als Anhang im Ordner „4\_Virtual\_Production“ in einem Video gegenübergestellt. Aus dem direkten Vergleich wird deutlich, dass die virtuellen Sequenzen nicht dieselbe Schärfe wie die realen Aufnahmen erreichen. Es scheint so, als würde das virtuelle Video alles in einer reduzierten Auflösung darstellen, obwohl beide Videos die gleichen Spezifikationen besitzen. Wenn man das Modell in Postshot betrachtet, ist eine etwas schärfere Darstellung zu beobachten, was zu der unbestätigten Vermutung führt, dass Render-Einstellungen in Unreal den Schärfebereich des Bildes möglicherweise negativ beeinflussen könnten. Betrachtet man hingegen detailreiche Zonen wie die Notausgangsschilder am Ende des Bahnsteigs, kann das reale vom virtuellen Bild nur bei genauem Hinsehen oder durch die Spiegelung der Deckenbeleuchtung im Kamerabild unterschieden werden (siehe Abbildung 16) und ein hoher Detailgrad ist zu beobachten.

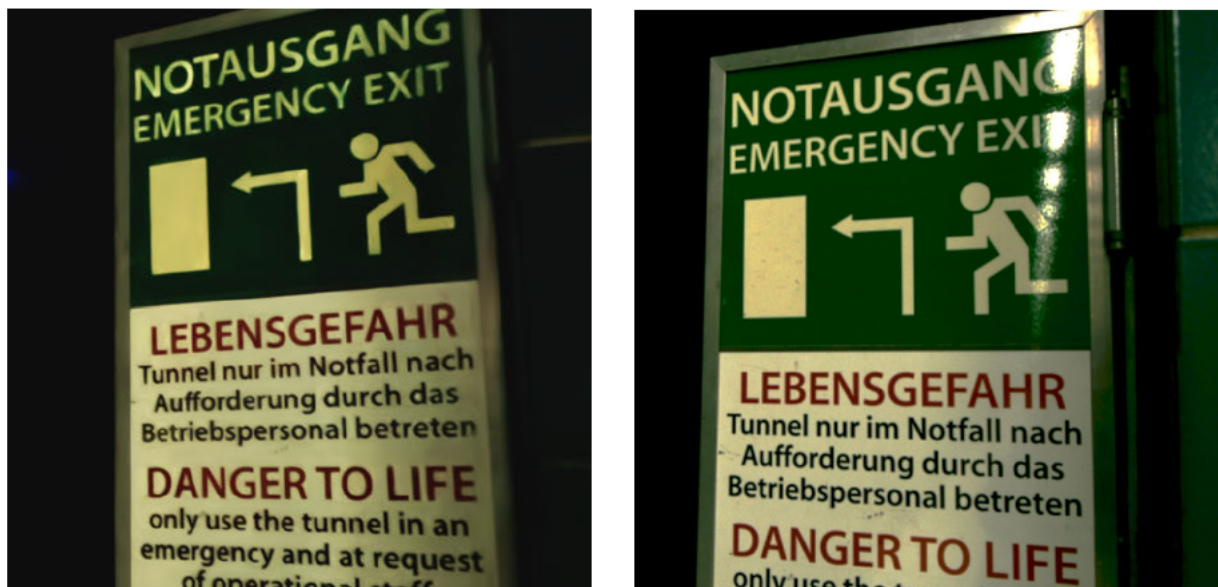


Abbildung 17: Notausgangsschilder im Vergleich (links: virtuelles Bild, rechts: Kamerabild)



Abbildung 18: Reflexionen im Vergleich (links: virtuelles Bild, rechts: Kamerabild)

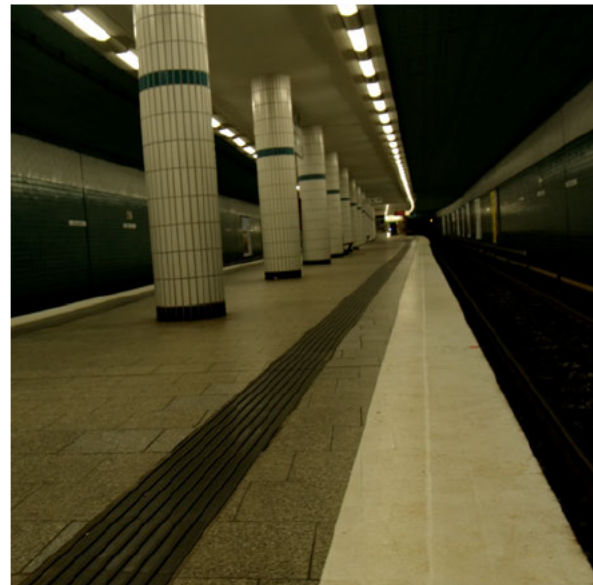
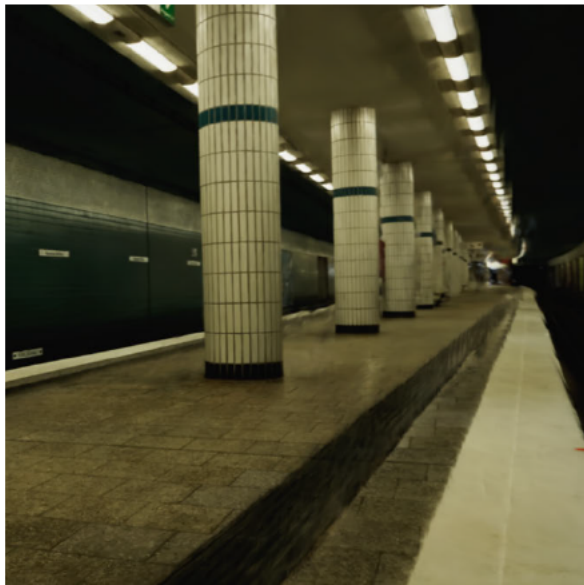


*Abbildung 19: Blick über den Bahnsteig (links: virtuelles Bild, rechts: Kamerabild)*

Die Reflexionseigenschaften können sowohl auf den farbigen Fliesen am frontalen Ende des Bahnsteigs als auch auf dem Boden der Bahnstation beobachtet werden. Die 3D-GS-Rekonstruktion kann diese in den besser repräsentierten Bereichen präzise und mit blickwinkelabhängiger Änderung darstellen (Abbildung 17).

Aus den Beobachtungen beider Szenen wird in der Gegenüberstellung deutlich, dass die virtuelle Szene mit dem verwendeten Modell nicht in der Lage ist, die Bildschärfe der realen Aufnahmen zu erreichen. Allerdings liefert das Modell in anderen Aspekten wie der geometrischen Präzision, den guten Reflexionseigenschaften und der umfangreichen Bildabdeckung im Bereich des Capture-Verfahrens bereits gute Ergebnisse.

Es muss der Transparenz halber erwähnt werden, dass die Rekonstruktion der U-Bahn-Station zeitlich vor der Parameterstudie durchgeführt wurde und die Erkenntnisse aus den Untersuchungen erst im späteren Verlauf erarbeitet werden konnten. In der Modellgenerierung wurde also nicht das volle Potenzial ausgeschöpft, weshalb eine erneute Rekonstruktion der Szene mit anderen Konfigurationen aller Wahrscheinlichkeit nach bessere Ergebnisse erzielen würde. Für ein neues Modell sollten im Capture-Prozess RAW- statt JPEG-Bilder verwendet werden, um in der Bildentwicklung die Dynamik anpassen zu können. Die Splat-Anzahl des Modells betrug drei Millionen, was für die Größe der Szene nicht ausreichend ist, wenn man sich die Anzahl von zehn Millionen für die deutlich kleinere Test-Szene vergegenwärtigt. Eine höhere Anzahl an Trainingsiterationen sowie ein kleineres Sichtfeld würden ebenfalls zur Erhöhung der Rekonstruktionsqualität vor allem in Bezug auf die Bildschärfe beitragen.

## 5 Fazit

In der vorliegenden Arbeit sollte die Frage beantwortet werden, ob sich das Verfahren „Gaussian Splatting“ zur Erstellung digitaler Sets in virtuellen Filmproduktionen eignet. Um dieser Frage nachzugehen, wurde im ersten Teil der Arbeit untersucht, wie sich der Workflow zur Generierung der Modelle optimieren lässt, um die bestmögliche visuelle Qualität zu erreichen. In diesem Zusammenhang haben sich subjektive Vergleiche als die robustere Bewertungsmethode erwiesen, während metrische Vergleiche zum Teil nicht nachvollziehbare Ergebnisse lieferten. Es zeigte sich, dass vor allem die Anzahl, die Perspektive und das Sichtfeld der Aufnahmen darüber entscheiden, ob ein Modell geometrisch präzise und mit scharfen Details nachgebildet werden kann. Zudem haben die Aufnahmebedingungen und Kameraeinstellungen einen großen Einfluss auf die Qualität der Rekonstruktion. Die optimale Anzahl der Splats und Trainingsschritte zeichnet sich durch einen guten Kompromiss aus hoher räumlicher Auflösung und moderatem Zeit- und Speicherbedarf aus. Andere Einstellungen haben hingegen wenig bis keinen Effekt auf das Modell. Unter den genannten Voraussetzungen ist es möglich, qualitativ hochwertige Rekonstruktionen zu erstellen, die dem visuellen Ideal der fotorealistischen Darstellung sehr nahekommen.

Die Produktion gerenderter Szenen in einem virtuellen Versuchsaufbau sollte die visuelle Qualität von „Gaussian Splatting“ unter Beweis stellen. Die gewonnenen Erkenntnisse zur Modelloptimierung konnten nur teilweise in die erstellte Rekonstruktion einfließen. Es gelang, das Capture-Verfahren für virtuelle Sets zu systematisieren. Durch eine säulenartige Anordnung der Kamerapositionen kann ein großes Raumvolumen auf effiziente Weise perspektivisch abgedeckt werden, sodass eine geometrisch robuste und lückenarme Rekonstruktion entsteht. In den Aufnahmeergebnissen zeigt sich, dass es im durchgeführten Modellbeispiel der U-Bahn-Station Probleme mit der Schärfe der virtuellen Szene gibt. Durch eine bessere Abstimmung der Trainingsparameter ließe sich dieser Effekt sicherlich minimieren, was jedoch in weiteren Untersuchungen überprüft werden müsste. Nichtsdestotrotz kann mit dem Modellversuch insgesamt aufgezeigt werden, dass „Gaussian Splatting“ ein leistungsstarkes Werkzeug für die Erstellung virtueller Filmsets sein kann, das nie dagewesene fotorealistische Detailtreue, blickwinkelabhängige Darstellung von Reflexionen und hohe Performance in einem Verfahren vereint.

## 6 Literaturverzeichnis

- 3D Gaussian Splatting for Real-Time Radiance Field Rendering.* (o. J.). Abgerufen 21. April 2026, von <https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting/>
- BioMotionLab. (2011, November 11). *Johansson: Motion Perception part 1* [Videoaufnahme]. <https://www.youtube.com/watch?v=1F5ICP9SYLU>
- Blog, N. T. (2017, April 19). *Toward A Practical Perceptual Video Quality Metric.* Medium. <https://netflixtechblog.com/toward-a-practical-perceptual-video-quality-metric-653f208b9652>
- Corridor Crew. (2025, Oktober 5). *ONE YEAR of Work for TEN SECONDS of Film* [Videoaufnahme]. <https://www.youtube.com/watch?v=iq5JaG53dho>
- Delaunay-Triangulation—Geoinformatik Lexikon.* (o. J.). Abgerufen 1. April 2026, von <https://www.geoinformatik.uni-rostock.de/einzel.asp?ID=477>
- fifonik. (2026). *Fifonik/FFMetrics* [Software]. <https://github.com/fifonik/FFMetrics> (Ursprünglich erschienen 2020)
- Ijaz, T. (o. J.). *Gaussian Splatting with LiDAR Data.*
- Johansson, G. (1973). Visual perception of biological motion and a model for its analysis. *Perception & Psychophysics*, 14(2), 201–211. <https://doi.org/10.3758/BF03212378>
- Kazhdan, M., Bolitho, M., & Hoppe, H. (o. J.). *Poisson surface reconstruction.*
- Kerbl, B., Kopanas, G., Leimkuehler, T., & Drettakis, G. (2023). 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Transactions on Graphics*, 42(4), Artikel 4. <https://doi.org/10.1145/3592433>
- Lowe, D. (2004). *Distinctive Image Features from Scale-Invariant Keypoints.*
- Meydenbauer, A. (1867). *Wochenblatt des Architektenvereins zu Berlin.* (49).
- Schonberger, J. L., & Frahm, J.-M. (2016). Structure-from-Motion Revisited. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4104–4113. <https://doi.org/10.1109/CVPR.2016.445>
- Sturm, P., & Triggs, B. (1996). A factorization based algorithm for multi-image projective structure and motion. In B. Buxton & R. Cipolla (Hrsg.), *Computer Vision—ECCV ’96* (Bd. 1065, S. 709–720). Springer Berlin Heidelberg. [https://doi.org/10.1007/3-540-61123-1\\_183](https://doi.org/10.1007/3-540-61123-1_183)
- Tomasi, C., & Kanade, T. (1992). Shape and motion from image streams under orthography: A factorization method. *International Journal of Computer Vision*, 9(2), 137–154. <https://doi.org/10.1007/BF00129684>
- Ullmann, S. (1976, Oktober). *The Interpretation of Structure from Motion.*

Zhou Wang, Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4), 600–612. <https://doi.org/10.1109/TIP.2003.819861>

## 7 Anhang

Sämtliche Datensätze, Videodateien und 3D-Modelle befinden sich auf einem beigelegten USB-Datenträger (64GB).

## 8 Verwendete Hilfsmittel

### **Le Chat**

Der KI-gestützte Chatbot von „Mistral AI“ wurde für die Unterstützung im Brainstorming, in der Recherche und für allgemeine Fragen im Zusammenhang mit der Bachelorarbeit verwendet. Die generierten Antworten wurden mithilfe der ausgewiesenen Quellen oder durch weitere Internetrecherchen auf ihre Korrektheit überprüft. Es wurde die kostenlose Version im Modus „Schnell“ genutzt.

### **DeepL Write**

Der KI-Schreibassistent wurde für einen Großteil der vorliegenden Arbeit zur Überprüfung der grammatikalischen und sprachlichen Korrektheit und in geringem Maße für die Neu- oder Umformulierung von Sätzen mit dem Zweck der sprachlichen Präzisierung genutzt. Es wurde die kostenlose Version verwendet.

### **Google Gemini 3.0**

Zur Erstellung von Videodateien aus Bilddatensätzen wurde mithilfe des KI-Chatbots von Google schrittweise ein Python-Skript erstellt. Dabei kam die kostenlose Version in den Modi „Fast“ und „Thinking“ zum Einsatz.

# Eigenständigkeitserklärung

Hiermit versichere ich, dass ich die vorliegende Bachelorarbeit mit dem Titel

**3D-Gaussian Splatting für die virtuelle Filmproduktion im visuellen Vergleich mit eigenständig produzierten Filmsequenzen**

selbständig und nur mit den angegebenen Hilfsmitteln verfasst habe. Alle Passagen, die wörtlich aus der Literatur oder aus anderen Quellen übernommen wurden, habe ich deutlich als Zitat mit Angabe der Quelle kenntlich gemacht.

---

Datum

---

Unterschrift