Schlussbericht IRRW:
Skalierung des inversen Ansatzes zur Bildanalyse

Zuwendungsempfänger: Friedrich-Alexander-Universität Erlangen-Nürnberg

Förderkennzeichen: 16IS22082

Laufzeit: 01.08.2022-31.07.2025

Vorgelegt beim Projektträger des BMBF: VDI/VDE Innovation + Technik GmbH

Berichtsautor: Bernhard Egger

Aufgabenstellung

Computergestützte Lernverfahren zur Bildanalyse waren bereits zu Projektbeginn in bestimmten Anwendungsfeldern schneller oder zuverlässiger als Menschen. Die gleichen Verfahren scheiterten aber in komplexeren Umgebungen, insbesondere in Situationen, die während des Trainings nicht verfügbar waren.

Solche Verfahren werden auch heute noch hauptsächlich mit enormen Datenmengen und manuellen Annotationen trainiert. Das Projekt baute darauf auf, dass die computergestützte Bildanalyse und insbesondere deren Generalisierung nicht ausschließlich als Lernproblem angesehen werden darf, sondern stark von dateneffizienten Ansätzen profitieren kann, um die automatische Verarbeitung in komplexeren Umgebungen zu verwirklichen und eine Generalisierung ermöglichen. In diesem interdisziplinären Projekt haben wir einen Ansatz verfolgt, der die bestehenden Limitierungen basierend auf generativen Modellen und einem inversen Ansatz angeht.

Inverse Methoden zur Bildanalyse zielen darauf ab, alle Teile der Szene zu rekonstruieren. Dieses beinhaltet die 3D Form der Objekte, ihre Materialeigenschaften, die Position sowie die Beleuchtung. Zu Projektbeginn existierende inverse Ansätze funktionierten für einzelne Objekte oder auf synthetischen Daten, aber bisher nicht ausreichend in komplexen und realistischen Umgebungen. Die Theorie von David Marr postuliert, dass dies im Menschen in einem Objekt-spezifischen und einem Objekt-agnostischen Schritt abläuft. Wir haben erstmals neuronale Evidenz für diese Hypothese und möchten in diesem Projekt zeigen, dass diese Separierung elementar ist, um inverse Methoden auf echte Bilder anwenden zu können.

In diesem Projekt stießen wir insbesondere auf drei Herausforderungen: Die erste Herausforderung liegt auf der Modellierungsseite. Inverse Methoden waren nicht ausreichend (foto)realistisch und während wir für spezifische Objekte im Bereich der Analyse-durch-Synthese, z. B. Gesichter, große Fortschritte gemacht haben, skalierte dieser Ansatz nicht für generelle Objekte. Deshalb funktionieren aktuelle z. B. implizite Methoden oft nur auf synthetischen Daten oder sehr limitierten Datensätzen.Die zweite Herausforderung war die Skalierung vom einzelnen Objekt zur Szene. Dies beinhaltet nun auch das Detektionsproblem von Objekten und dies wurde im generativen Sinne mit einem Vorschlagen-und-Verifizieren Ansatz angegangen. Die dritte Herausforderung lag darin herauszufinden, ob unser menschliches visuelles System Methoden umsetzen kann, die auf der Idee von generativen Modellen und inversen Methoden basieren oder nicht. Zu

Projektbeginn gab es hier Indizien für beide Standpunkte, um diese Frage aber besser zu beantworten, müssen neue Modelle erstellt und evaluiert werden und diskriminative Experimente durchgeführt werden. Im Verlauf des Projektes haben wir hinsichtlich all dieser drei Herausforderungen Fortschritte erzielt und konkrete Methoden implementiert und publiziert. Für die meisten Forschungsergebnisse liegt neben einer Open Access Publikation auch eine Veröffentlichung des Quellcodes vor.

Bei diesem Projekt handelt es sich um Grundlagenforschung, um ein Computer Vision System zu erstellen, welches ähnliche Generalisierungseigenschaften wie das menschliche visuelle System besitzt. Während Computer Vision Systemen bereits in unserem Alltag angewendet werden, genügt die Qualität für viele Anwendungen, inklusive dem autonomen Fahren noch nicht. Aus unserer Sicht können die aktuellen Limitierungen nur mittels eines fundamental anderen Ansatzes gelöst werden - wenn der inverse Ansatz auf echten Daten besser generalisiert, wird er auch sehr schnell seinen Weg in Anwendungen finden und dafür wurden im Rahmen dieses Projektes weitere konkrete Forschungsergebnisse erzielt.

Der inverse Ansatz wurde in diesem Projekt durch zwei Kernideen auch für komplexere Umgebungen erweitert. Erstens wurden objektspezifische und objektunspezifische Bestandteile der Bildverarbeitung, basierend auf David Marrs Theorie , getrennt. Wir haben dazu eine 2.5D-Repräsentation verwendet (Tiefenkarten) und diese mit generativen Modellen mit variabler Komplexität kombiniert. Zweitens haben wir uns der Modellierung von Oberflächen gewidmet, speziell in Bezug auf deren Dateneffizienz und deren Fähigkeit zu generalisieren. Wir haben dazu neuartige Ansätze im Bereich der 3D-Erzeugung neuer Ansichten (novel-view synthesis) basierend auf impliziten Methoden mit dateneffizienten Deformationsmodellen verwendet. Unsere Forschung baut zudem auf dem visuellen System des Menschen auf, welches vermutlich einen inversen Ansatz implementiert. Wir haben dabei versucht zu verstehen, inwiefern wir mit Methoden der Computergrafik das menschliche visuelle System besser verstehen können.

Voraussetzungen, unter denen das Vorhaben durchgeführt wurde

Das Projekt wurde an der Friedrich-Alexander-Universität Erlangen-Nürnberg an der Professur für Cognitive Computer Vision, am Lehrstuhl für Graphische Datenverarbeitung, Department Informatik der Technischen Fakultät durchgeführt. Zusätzlich zur Unterstützung durch das BMBF wurden auch Rechenressourcen der Universität genutzt, welche über das National High Performance Computing Center (NHR@FAU) zur Verfügung standen. Diese

Rechenressourcen wurden durch das Land Bayern und den Bund gefördert, teilweise über das DFG Projekt 440719683.

Planung und Ablauf des Vorhabens

Es wurden drei wissenschaftliche MitarbeiterInnen angestellt, um an diesem Projekt zu arbeiten. Eine Stelle wurde zwecks einem Industriepraktikum kurzfristig durch einen anderen Mitarbeiter besetzt und zu Projektende wurde das Team aufgestockt, um möglichst viele der Projektziele innerhalb der bewilligten Laufzeit zu erreichen.

Wir haben den Förderbescheid am 22.06.2022 erhalten, die Projektlaufzeit war vom 01.08.2022 bis 31.07.2025 geplant. Nur durch die frühzeitige Rekrutierung von Personal war ein effektiver Projektstart am 01.10.2022 möglich. Daraus ergab sich eine Verzögerung, die wir am Ende des Projektes durch zusätzliches Personal kompensiert haben.

Zur Publikation der später genannten Artefakte und zur Weiterbildung haben folgende projektbezogenen Reisen (wie im Proposal geplant) stattgefunden:
CVPR Konferenz 2023, ICVSS Summer School 2023, ICCV Konferenz 2023, CogSci Konferenz (virtuell) 2023, NEAT Workshop 2023, Besuch York University 2023, ICCV Konferenz 2024, CogSci Konferenz 2024, NEAT Workshop 2024, WACV Konferenz 2024, ECCV Konferenz 2024, MLx Summer School 2024, VMV 2024, Eurographics Konferenz 2024, ICLR 2025, 3DV 2025, BVM 2025, Dagstuhl 2025.
Bei außereuropäischen Reisen wurde gemäß des Zuwendungsbescheids Rücksprache mit dem Projektträger gehalten.

Das Projekt wurde wie geplant kurzfristig von zwei Gastwissenschaftler der York University und University of Chicago unterstützt.

Wissenschaftlicher und technischer Stand, an den angeknüpft wurde

Zu Beginn des Projekts gewannen inverse Rendering-Verfahren im Zuge des Aufkommens neuronaler Netzwerkmodelle und moderner Inferenztechniken erneut stark an Bedeutung. Im Rahmen unseres Vorhabens lag der Schwerpunkt auf inversen Rendering-Ansätzen, die auf 3D-Repräsentationen basieren.

Auf der Modellierungsseite existierte bereits eine große Vielfalt generativer Modelle, die darauf abzielten, Form und Erscheinungsbild von Objekten aus einer niedrigdimensionalen parametrischen Darstellung zu erzeugen. Diese Modelle unterschieden sich hinsichtlich der

gewählten Repräsentation – von voxel- und meshbasierten Ansätzen über hierarchische bis hin zu impliziten Repräsentationen sowie deren Kombinationen. Modelle mit nahezu fotorealistischer Rekonstruktionsqualität wie NeRF waren jedoch szenenspezifisch und verfügten nur über eingeschränkte Generalisierungsfähigkeiten. Klassenspezifische Modelle wie SRNs oder 3D-Morphable Models generalisierten innerhalb einer Objektklasse gut, setzten jedoch entweder 3D-Daten oder sehr große Mengen an 2D-Daten voraus und ließen sich nicht auf neue Objektklassen übertragen. Insgesamt zeigte sich, dass bestehende Modelle Schwierigkeiten hatten, auf neue Objektklassen oder Few-Shot-Szenarien zu generalisieren. Darüber hinaus litten viele dieser Ansätze unter einer Realismuslücke zwischen synthetischen Trainingsdaten (bedingt durch Datengenerierung oder Modellierungstechnik) und realen Bildern.

Parallel dazu gab es eine umfangreiche Forschung zu Methoden, die Tiefe und andere Szeneneigenschaften aus 2D-Bildern schätzten, etwa InverseRenderNet. Diese Ansätze verfügten jedoch nicht über einen interpretierbaren, szenengraphenähnlichen, latenten Raum, stellten aber einen vielversprechenden Ausgangspunkt für unsere geplante 2.5D-Skizzenschätzung dar. Sie zielten darauf ab, 3D-Eigenschaften einer Szene ohne explizite Objektmodellierung zu erschließen, was die Datenerhebung vereinfachte und sogar Lernen mit schwacher Supervision, etwa aus Videos, ermöglichte.

Neuere Entwicklungen im Bereich Neural Rendering konzentrierten sich vor allem auf die Verbesserung der Rechenleistung entsprechender Frameworks oder auf die Generierung kompletter Szenen. Inverse Problemstellungen wurden dagegen kaum behandelt oder nur für Beispiele, die stark durch die – überwiegend synthetischen – Trainingsdaten abgedeckt waren.

Menschliche Gesichter stellten bereits damals einen zentralen Anwendungsbereich inverser Rendering-Techniken dar. Einerseits waren sie aufgrund ihrer Bedeutung für menschliche Interaktion besonders relevant, andererseits stand eine große Menge an Daten zur Verfügung. In den vorangegangenen zwei Jahrzehnten wurden zahlreiche Ansätze zur Rekonstruktion von 3D-Gesichtern aus 2D-Bildern entwickelt, wobei 3D-Morphable Models als besonders vielversprechend galten. Hierbei diente ein vergleichsweise einfaches statistisches Modell, das auf dicht korrespondierenden 3D-Scans basierte, als Prior in einem stark unterbestimmten Rekonstruktionsproblem. Mit der Zeit wurden vermehrt Bottom-up-Methoden integriert, um die Robustheit zu erhöhen und detailreiche 3D-Rekonstruktionen zu ermöglichen. Wesentliche Einschränkungen bestanden jedoch in der Notwendigkeit hochwertiger 3D-Daten, der Annahme dichter Korrespondenzen sowie umfangreicher indirekter Supervision über komplexe Verlustfunktionen. Insbesondere die

starke Korrespondenzannahme und der hohe manuelle Vorverarbeitungsaufwand erschwerten die Übertragung dieses Konzepts auf andere Objektklassen. Da diese Methoden jedoch von einer niedrigdimensionalen und gut kontrollierbaren Repräsentation profitierten, erschien eine Kombination dieser Ideen mit modernen Neural-Rendering-Techniken besonders vielversprechend.

Eine weitere für das Projekt relevante Entwicklung war die zunehmende Verfügbarkeit von 3D-Daten verschiedenster Objekte. Diese resultierte sowohl aus einfach zugänglichen Aufnahmegeräten in mobilen Endgeräten als auch aus Algorithmen, die 3D-Objektformen aus Videos oder schwach überwachtem Material rekonstruieren konnten sowie aus großen Datensätzen wie Objaverse.

Neben der Modellierung stellte die Inferenz der Modellparameter eine große Herausforderung dar. Für meshbasierte statistische Modelle existierte eine Vielzahl von Inferenzverfahren, darunter Sampling- und gradientenbasierte Methoden. Differenzierbare- und neuronale Renderer brachten zudem lernbasierte Ansätze mit Encodern oder Autoencodern hervor. Nahezu alle bestehenden Verfahren waren jedoch auf eine Form der Poseninitialisierung angewiesen, meist auf Basis erkannter Landmarken, und reagierten empfindlich auf Fehler in dieser Initialisierung. Viele Methoden deckten das Detektionsproblem nicht ab und zahlreiche implizite Repräsentationsansätze setzten bekannte Kameraparameter voraus. Eine weitere Kritik an 3D-Rekonstruktionsnetzwerken lautete, dass sie oft eher eine Klassifikationsaufgabe lösten, indem sie den nächsten Nachbarn rekonstruierten, statt tatsächlich Form und Erscheinungsbild zu erschließen. Frühere Arbeiten zur Integration objektbasierter inverser Rendering-Ansätze in ein Szenenverständnis nutzten Bildparsing-Strategien, waren jedoch durch die damals begrenzte Leistungsfähigkeit generativer Modelle und die geringe Verfügbarkeit geeigneter 3D-Modelle eingeschränkt.

Auch Erkenntnisse aus der Erforschung des menschlichen visuellen Systems spielten eine Rolle. Das damalige Verständnis beschrieb die visuelle Verarbeitung als Kombination aus Feed-forward-Prozessen und sparsamer top-down Rückkopplung. Elektrophysiologische Studien lieferten Hinweise auf inverse-graphics-ähnliche Mechanismen bei der biologischen Gesichtsverarbeitung. Zudem gab es Evidenz für 3D-Formrepräsentationen und 2.5D-Darstellungen in Areal V4 . Unklar war jedoch, welche konkrete Ausgestaltung eines Inverse-Graphics-Ansatzes – sowohl hinsichtlich Modellierung als auch Inferenz – dem Gehirn am ehesten entsprach. Da feed-forward-Objekterkennungssysteme zu dieser Zeit stark im Fokus standen, existierte vergleichsweise wenig Forschung zu inverse-graphics-basierten Mechanismen. Ein wesentlicher Grund dafür war, dass inverse-

graphics-basierende Mechanismen noch nicht ausreichend auf realweltliche Bilddaten generalisierten und daher schwer mit bestehenden Datensätzen vergleichbar waren.

Zum Projektstart waren uns zudem keine Patente oder sonstigen Schutzrechte bekannt, die einer Verwertung oder Veröffentlichung der im Projekt entstehenden Forschungsergebnisse entgegenstanden.

Angabe bekannter Konstruktionen, Verfahren und Schutzrechte, die für die Durchführung des Vorhabens benutzt wurden

Es wurde auf bestehende Verfahren im Forschungsbereich wie im vorherigen Kapitel beschrieben aufgebaut und diese wurden in den jeweiligen Publikationen konkret zitiert. Es wurden dabei keine Schutzrechte verwendet.

Angabe der verwendeten Fachliteratur sowie der benutzten Informations- und Dokumentationsdienste

Die Methodik in diesem Projekt ist auf die aktuelle Literatur im Bereich Computer Vision, Computer Grafik, des maschinellen Lernens sowie der Kognitions- und Neurowissenschaften aufgebaut. Die Ergebnisse wurden in Form von Publikationen dokumentiert und die konkret verwendete Fachliteratur entsprechend zitiert.

Zusammenarbeit mit anderen Stellen

Im Rahmen des Projektes haben wir mit Forschern der OTH Regensburg, der TU Delft, der University of York und dem Massachusetts Institute of Technology kollaboriert.

Das Projekt profitierte zudem von diversen studentischen Arbeiten, welche sich auch in den entstandenen Publikationen durch (Co-)Autorenschaft manifestierte.

Darstellung des erzielten Ergebnisses

Bezüglich der Zielsetzung der Szenenrekonstruktion anhand der Trennung von Szenen- und Objektspezifischer Verarbeitung sind uns entscheidende Fortschritte gelungen. Das Resultat ist eine Publikation „Generalizable 3D Scene Reconstruction via Divide and Conquer from a Single View”. Hier ist insbesondere die Skalierung in Richtung Szene, mehrere Objekte und auch der Einbezug von Tiefenschätzung wie im Projekt vorgeschlagen eingegangen.

Darüber hinaus haben wir die Implementierung von Gen3DSR Open Source gemacht, um die einfache Nutzung durch andere Forschende der Community zu ermöglichen. Diese Arbeit steht im Einklang mit einem der Hauptziele des Projekts: der Entwicklung eines 3D-Rekonstruktionssystems, das sich besser auf reale Umgebungen übertragen lässt und Computer-Vision-Anwendungen in weniger kontrollierten Umgebungen ermöglicht. Wir haben ein Framework zur Szenenanalyse entwickelt, das ein Bild segmentiert und dessen Komponenten über generative Modelle erklärt. Dabei werden Inferenz und 3D-Darstellungen skaliert, um mehrere Objekte zu unterstützen. Zu diesem Zweck trennen wir die objektagnostische Verarbeitung von der objektspezifischen 3D-Modellierung. Dieser Ansatz ist dateneffizienter, da jedes Modul in den Phasen der Szenenanalyse und Instanzverarbeitung isoliert mit geeigneten Datensätzen für die jeweiligen Teilaufgaben trainiert wird.

Im Rahmen unserer Arbeit konzentrierten wir uns auch auf die in realen Szenen auftretenden Okklusionsprobleme. Für die 3D-Szenenrekonstruktion mit einer einzelnen Ansicht entwickelten wir ein amodales Vervollständigungsmodell, das plausible Vervollständigungen für die verdeckten Teile der Objekte in der Szene generieren kann. Für die Multi-View-Umgebung, in der Okklusionen vorübergehend auftreten können, entwickelten wir eine auf dem RANSAC-Paradigma basierende Methode, die Rekonstruktionsinkonsistenzen in objektzentrierten Szenen effektiv reduziert. Das Ergebnis mit dem Titel „RANRAC: Robust Neural Scene Representations via Random Ray Consensus“ wurde auf der European Conference on Computer Vision 2024 veröffentlicht.

Wir haben uns des Weiteren auf die Entwicklung eines hybriden Rekonstruktionsmodells konzentriert, das die Stärken expliziter 3D-Darstellungen mit den Fähigkeiten impliziter neuronaler Darstellungen kombiniert. Aufbauend auf der vorherigen Arbeit innerhalb der Gruppe „Sphere-Guided Training of Neural Implicit Surfaces“ entwickelten wir eine Methode, die sphärische Grundelemente verwendet, um die neuronalen impliziten Oberflächen zu begrenzen, die die Geometrie und das Erscheinungsbild der rekonstruierten Instanz darstellen. Durch die Verwendung der expliziten Geometrie eines Gesichts-3D-Morphable Models (3DMM), das an die Eingabebilder angepasst wird, führen wir eine fundierte Initialisierung der Grundelemente durch, was die Trainings- und Rendering-Prozeduren der impliziten neuronalen Darstellung beschleunigt und die Rekonstruktion von Details verbessert. Wir wenden diese Methode für die Aufgabe einer neuartigen Ansichtssynthese menschlicher Gesichter in einem spärlichen Kameraaufbau an und erzielten wettbewerbsfähige Ergebnisse. Der vorgeschlagene Ansatz erreichte den dritten Platz in der Kategorie „To NeRF or not to NeRF: A View Synthesis Challenge for Human Heads“ auf der

„International Conference on Computer Vision 2023“ und wird im Rahmen der ICCVW 2023-Proceedings veröffentlicht.

Obwohl die Methode für die 3D-Rekonstruktion und -Darstellung bei mehreren Ansichten des Objekts konzipiert ist, glauben wir, dass die vorgeschlagenen Ideen auf generative Modelle und die Rekonstruktion aus einem einzigen Bild ausgeweitet werden können.

Zudem wurde die Zusammenführung von Gaussian Process Morphable Models (GPMMs) und impliziten Modellen bearbeitet. Zunächst wurde ein passendes Framework gesucht, um dies umzusetzen; es stellte sich heraus, dass sich hierfür Neural Kernel Fields (NKFs) besonders gut eignen. NKFs sind methodisch ähnlich zu Gaussian Process Regression und haben daher bereits eine probabilistische Interpretation. NKFs verwenden den sogenannten Neural Spline (NS) Kernel, dieser lässt sich relativ leicht durch GPMMs augmentieren. Dadurch lassen sich implizite und explizite Modelle verbinden. Jedoch stellte sich heraus, dass NKFs noch Probleme haben detailreiche Oberflächen, wie zum Beispiel Gesichter, darzustellen. Aus diesem Grund wurde an einem neuen Kernel gearbeitet, der diese Details besser darstellen soll. Der Kernel basiert auf einer spektralen Darstellung, wobei die Fourier Koeffizienten gelernt werden. Dieser Kernel, im Vergleich zu NS, ist adaptiv (hat Parameter); sein Spektrum lässt sich gezielt manipulieren. Das macht ihn insbesondere auch für sehr verrauschte Daten attraktiv.

Die Umsetzung erfolgt mithilfe des Neural Kernel Field (NKF) Frameworks, welches den sogenannten Neural Spline (NS) Kernel verwendet. Dazu wurde an einem neuen, spektralen Kernel gearbeitet, mit dem Ziel dessen Fourier Koeffizienten zu lernen. Leider konnte dieser Kernel nicht den gewünschten Effekt erzielen (detailreichere Oberflächen, robuster gegenüber verrauschten Daten, skalierbarer). Es stellte sich jedoch heraus, dass stationäre Matérn Kernel eine signifikanten Verbesserung gegenüber NS erzielen. Die Familie der Matérn Kernel vereint mehrere bekannte Kernel, wie zum Beispiel den Laplace und Gaussian Kernel, ist stationär (damit translations- und rotations-invariant), adaptiv (besitzt einen Parameter), skalierbar, numerisch stabil und extrem einfach zu implementieren. Auf Basis dieser Erkenntnisse wurde eine Publikation angefertigt, die 2025 auf der International Conference on Learning Representations (ICLR) angenommen wurde. Diese Arbeit beinhaltet auch einen starken theoretischen Anteil, der Matérn Kernels tiefgehend mathematisch analysiert und, unter anderem, neue Verbindungen zu neuronalen Netzwerken und dem NS Kernel aufzeigt.

Kernel Methoden sind aufwändig in der Berechnung (haben kubische Laufzeitkomplexität und quadratische Platzkomplexität), weshalb das NKF Framework nur für Punktewolken bis

zu 10.000 Punkten anwendbar ist. Für hochauflösende Rekonstruktionen sind das oft zu wenig Punkte, insbesondere im Gesichtsbereich. Aus diesem Grund wurde an einer skalierbaren Methodik gearbeitet. Die neu entwickelte Methode verwendet ein lokales und hierarchisches Regressionsmodell und erlaubt die Anwendung von NKF auf beliebig große Punktewolken.

Ursprünglich wurde vorgeschlagen, dass wir extreme Reize wie Mooney-Bilder verwenden werden, um die Feedforward-Verarbeitung zu untersuchen, da ein Modell des menschlichen Gehirns oder des Menschen in der Schleife erforderlich ist. Wir haben unsere Hypothese für unser Verhaltensexperiment angepasst und stattdessen versucht, mithilfe des gleichen Mechanismus wie beim statistischen Gesichtsmodell festzustellen, ob unser Gehirn Gesichter aus beeinträchtigten Reizen rekonstruiert. Durch erste Iterationen haben wir einige gute Ideen gefunden, um Experimente zu Mooney-Gesichtern zu erstellen, wie etwa die Interpretation von Mooney-Gesichtern mittels Strichzeichnungen (z. B. wenn ein Mooney-Gesicht ein Extremfall einer Strichzeichnung ist). Wir haben auch andere Effekte wie die Kontrastinvertierung detaillierter untersucht. Leider haben unsere Ergebnisse diesbezüglich zu keinen publizierbaren Ergebnissen geführt.

Wir wollten zudem beantworten, ob inverses Rendering das bessere Modell des menschlichen visuellen Systems für die Kernobjekterkennung ist als das aktuelle CNN-basierte System. Wir haben einen ersten Schritt dieses Problems bearbeitet und versucht, die Objekterkennung im Gehirn zu analysieren. Wir haben Datensätze unter verschiedenen Bedingungen gerendert (diese basieren nicht auf ImageNet, wie im Vorschlag vorgeschlagen, sondern auf Meshes aus dem Objaverse-Datensatz), und trainieren Modelle für jeden dieser Datensätze, die sich minimal voneinander unterscheiden (jeweils eine Bedingung z. B. keine Schatten oder eine zufällige Textur) und bewerten die Modelle anhand des Brainscores. Wir nutzten Computergrafik als Werkzeug, um zu untersuchen, wie das menschliche Gehirn Szeneneigenschaften verarbeitet. Insbesondere untersuchten wir die Verarbeitung von Erscheinungs- und Rotationseigenschaften in visuellen Szenen. Unsere Experimente lieferten signifikante Ergebnisse, die zu einer Veröffentlichung auf dem ICLR 2025 Re-Align² Workshop on Representational Alignment sowie einem Flash Talk und einer Posterpräsentation auf dem SfN Neuroscience Chapter 2025 führten: Die Darstellung von Merkmalen wie Textur in den frühen visuellen Bereichen, die Darstellung von Merkmalen wie Schatten, Schattierung und Umgebungsbeleuchtung in den frühen visuellen Bereichen und die Darstellung von Merkmalen wie Rotation in den späteren visuellen Bereichen des ventralen visuellen Stroms.

Verschiedene Ausrichtungsmetriken, die zur Messung der Ausrichtung zwischen neuronalen Netzen und den repräsentationalen Unähnlichkeitsmatrizen des Gehirns verwendet wurden, korrelierten nicht miteinander, was mit früherer Literatur wie Soni et al. 2024 übereinstimmt. In einem projektbezogenen studentischen Projekt zur Bestimmung objekt- und blickpunktzentrierter Repräsentationen im menschlichen Gehirn konnten wir eine gewisse Trennung von Objektklassen wie Menschen und Tieren im dorsalen Strom im Vergleich zum visuellen ventralen Strom des NSD-Datensatzes feststellen. Mit dieser Arbeit wollen wir weitere Fragen zur menschlichen Blickrichtung klären.

Darstellung des voraussichtlichen Nutzens, insbesondere der Verwertbarkeit des Ergebnisses und der Erfahrungen

Wie bereits im Antrag beschrieben, ist es selbst bei erfolgreicher Durchführung dieses Projektes im Bereich der Grundlagenforschung ein weiter Weg in die praktische Anwendung. Während wir verwertbare Erkenntnisse stets publiziert haben, planen wir nach wie vor keine schutzrechtliche Anmeldung unserer Ergebnisse.

In dem Projektantrag haben wir beschrieben, dass wir planen, an einschlägigen Konferenzen Publikationen einzureichen. Insgesamt sind bereits 9 Publikationen entstanden und angenommen. Mit den Publikationen an der WACV, ICLR, ECCV und 3DV sind uns vier Publikationen an Top-Konferenz gelungen. Wie im Projektantrag vorgeschlagen, haben wir den jeweils relevanten Quelltext und wenn sinnvoll möglich auch Daten publiziert und insbesondere bei der Publikation von Gen3DSR hat dies bereits Anklang in der Community gefunden. Dies führt zu einem vereinfachten Zugang zu Informationen, ermöglicht den Transfer zu anderen Nutzern, stellt die Reproduzierbarkeit unserer Forschungsergebnisse sicher und ermöglicht die Erkenntnisse einfach in der Industrie einzusetzen.

Darstellung des während der Durchführung des Vorhabens dem Zuwendungsempfänger bekannt gewordenen Fortschritts auf diesem Gebiet bei anderen Stellen

“Skalierung des inversen Renderings auf die reale Welt“ ist ein Thema von sehr großem Interesse in der Computer-Vision- und Grafik-Community und relevante Arbeiten werden kontinuierlich veröffentlicht. Die folgenden Ergebnisse werden sich voraussichtlich positiv auf die Qualität der Umsetzung der Projektziele auswirken:

Ke, Bingxin, et al. "Repurposing diffusion-based image generators for monocular depth estima-tion." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.

Xu, Katherine, Lingzhi Zhang, and Jianbo Shi. "Amodal completion via progressive mixed context diffusion." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogni-tion. 2024.

Yuan, Haobo, et al. "Open-vocabulary SAM: Segment and recognize twenty-thousand classes interactively." Proceedings of the European Conference on Computer Vision. 2024.

Phongthawee, P., Chinchuthakun, W., Sinsunthithet, N., Raj, A., Jampani, V., Khungurn, P., & Suwajanakorn, S. (. “Diffusionlight: Light probes for free by painting a chrome ball.”arXiv preprint arXiv:2312.09168. 2023.

Liu, R., Wu, R., Van Hoorick, B., Tokmakov, P., Zakharov, S., & Vondrick, C. “Zero-1-to-3: Zero-shot one image to 3d object.” In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 9298-9309). 2023.

Tang, J., Ren, J., Zhou, H., Liu, Z., & Zeng, G. . “Dreamgaussian: Generative gaussian splatting for efficient 3d content creation.” arXiv preprint arXiv:2309.16653. 2023.

Wang, Ruicheng, et al. "Moge: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision." Proceedings of the Computer Vision and Pattern Recognition Conference. 2025.

Xiang, Jianfeng, et al. "Structured 3d latents for scalable and versatile 3d generation." Proceedings of the Computer Vision and Pattern Recognition Conference. 2025.

Huang, Zixuan, et al. "Spar3d: Stable point-aware reconstruction of 3d objects from single images." Proceedings of the Computer Vision and Pattern Recognition Conference. 2025.

Wu, Tianhao, Chuanxia Zheng, and Tat-Jen Cham. "Panodiffusion: 360-degree panorama outpainting via diffusion." arXiv preprint arXiv:2307.03177. 2023.

Jin, Haian, et al. "Neural gaffer: Relighting any object via diffusion." Advances in Neural Information Processing Systems 37 : 141129-141152. 2024.

NoKSR: Kernel-Free Neural Surface Reconstruction via Point Cloud Serialization}, Zhen Li and Weiwei Sun and Shrisudhan Govindarajan and Shaobo Xia and Daniel Rebain and Kwang Moo Yi and Andrea Tagliasacchi, 3DV2025

Auch im Bereich der Neurowissenschaften und Kognitionswissenschaften gab es relevante Publikationen aufgrund derer wir unsere Ansätze leicht angepasst haben:

Neural Taskonomy: Inferring the Similarity of Task-Derived Representations from Brain Activity: Wang et. al 2024

NEST: Neural Encoding Simulation Toolkit: for using encoding models to synthesize respons es to synthetic images https://github.com/gifale95/NEST/blob/main/README.md

Natural Scenes Dataset: for studying the dorsal and ventral visual stream Allen, E.J., St-Yves, G., Wu, Y., Breedlove, J.L., Prince, J.S., Dowdle, L.T., Nau, M., Caron, B. Pestilli, F., Charest, I., Hutchinson, J.B., Naselaris, T.*, Kay, K.* A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. Nature Neuroscience (2021)
Qi, L., Kuen, J., Guo, W., Shen, T., Gu, J., Jia, J., ... & Yang, M. H. (2022). High-quality entity segmentation. arXiv preprint arXiv:2211.05776.

Ke, B., Obukhov, A., Huang, S., Metzger, N., Daudt, R. C., & Schindler, K. (2023). Repurposing diffusion-based image generators for monocular depth estimation. arXiv preprint arXiv:2312.02145.

Jin, L., Zhang, J., Hold-Geoffroy, Y., Wang, O., Blackburn-Matzen, K., Sticha, M., & Fouhey, D. F. (2023). Perspective fields for single image camera calibration. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 17307-17316)

Darstellung der erfolgten oder geplanten Veröffentlichung des Ergebnisses

Im Rahmen des Projektes sind die folgenden Veröffentlichungen entstanden. Diese wurden an Konferenzen und Workshops präsentiert und alle neben der Publikation durch Publisher auch auf arXiv oder openreview als OpenAccess Publikationen veröffentlicht. Die meisten Publikationen wurden zudem mit der Veröffentlichung des Quellcodes mit einer Open Source Lizenz bereichert.

Matérn Kernels for Tunable Implicit Surface Reconstruction
Maximilian Weiherer, Bernhard Egger
The Thirteenth International Conference on Learning Representations (ICLR), April 2025

Generalizable 3D Scene Reconstruction via Divide and Conquer from a Single View
Andreea Ardelean, Mert Özer, Bernhard Egger
International Conference on 3D Vision (3DV), March 2025

Computer Graphics from a neuroscientist's perspective
Shreya Kapoor & Bernhard Egger
Workshop on Representational Alignment (Re-Align) at The Thirteenth International Conference on Learning Representations (ICLR), April 2025

iRBSM: A Deep Implicit 3D Breast Shape Model
Maximilian Weiherer, Antonia von Riedheim, Vanessa Brébant, Bernhard Egger, Christoph Palm
German Conference on Medical Image Computing / Bildverarbeitung für die Medizin (BVM), February 2025
Best Paper Award 2nd Price

RANRAC: Robust Neural Scene Representations via Random Ray Consensus
Benno Buschmann, Andreea Ardelean (Dogaru), Elmar Eisemann, Michael Weinmann, Bernhard Egger
European Conference on Computer Vision (ECCV), October 2024

ArCSEM: Artistic Colorization of SEM Images via Gaussian Splatting
Takuma Nishimura, Andreea Ardelean (Dogaru), Martin Oeggerli, Bernhard Egger
AI for Visual Arts Workshop, European Conference on Computer Vision (ECCV), October 2024

Exploring Multi-modal Neural Scene Representations With Applications on Thermal Imaging
Mert Özer*, Maximilian Weiherer*, Martin Hundhausen, Bernhard Egger
Workshop on Vision-based InduStrial InspectiON (VISION), European Conference on Computer Vision (ECCV), October 2024

Approximating Intersections and Differences Between Linear Statistical Shape Models Using Markov Chain Monte Carlo
Maximilian Weiherer, Finn Klein, Bernhard Egger
IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024

erscheint noch:

Learning Neural Parametric 3D Breast Shape Models for Metrical Surface Reconstruction From Monocular RGB Videos
Maximilian Weiherer, Antonia von Riedheim, Vanessa Brébant, Bernhard Egger, Christoph Palm
Machine Learning for Biomedical Imaging: MELBA 2026