Inhaltsverzeichnis

Dokumentation zum praktischen Projekt im Kurs KIM

1.0 Einleitung

1.1 Ziel des Projektes

Zweck des Projektes war es, ein selbstausgewähltes KI-Tool auszuprobieren und anhand der vorausgegangenen Präsentationen zur Theorie, aufgestellte Hypothesen und Grenzen des Tools zu erklären. In der Arbeit mit dem Tool Depth-Anything-3 habe ich folgende Hypothesen getestet und bentwortet: 1.: Wo sind die Grenzen und sind sie hardwearabhängig? 2.: Kann das Modell auch gute 3D-Modelle von einzelnen Objekten erstellen? 3.: Wie lassen sich gesammelte Erkenntnisse mit der Architektur/Technologie des Tools erklären und in Verbindung mit der vorausgegangenen Präsentation bringen ?

1.2 Ausgewähltes Tool

In der folgenden Dokumentation geht es um das open Source Projekt Depth-Anything-3. Das Tool dieses Projektes ist ein Transformer-gestütztes KI-Modell, welches aus einem oder mehreren Input Bild/ern eine 3D Szene ertsellt.

2.0 Technologie und Architektur

2.1 Technologie hinter Depth Anything 3

2.1.1 Technologischer Gesamtansatz

Die zugrunde liegende Technologie basiert auf der Annahme, dass räumliche Tiefe und Geometrie statistisch aus visuellen Informationen erschließbar sind, auch wenn nur zweidemensionale Bilder vorliegen. Depth Anything 3 nutzt ausschließlich RGB-Bildinformationen und lernt aus großen Datenmengen (mehrere hundert Millionen Bilder), wie sich die dreidemensionale Strukturen der realen Welt in zweidimensionalen Projektionen widerspiegeln. Kern dieser Technologie ist die monukulare 3D-Geometrieinferenz, mit der Tiefe implizit aus Bildmerkmalen wie Perspektive, realtiver Objektgröße, Schatten, Verdeckungen und semantischen Mustern geschätzt wird. Diese Daten werden statisch gelernt, wodurch das Modell auch in unbekannte Szenen Tiefenschätzungen erzeugen kann.

2.2.2 Monekulare Tiefenschätzung

Das Modell modelliert Tiefe aus Teilen einer konsistenten 3D-Raumstruktur. Wenn mehrere Bilder aus einer Szene vorliegen, erzeugt das Modell nicht für jedes seperate Bild eine Tiefenkarte, sondern eine gemeinsame geometrische Erklärung, die für alle Ansichten konsistent ist. -> Multi-View-Kohärenz.

2.2.3 Depth-Ray-Repräsentation

Zusätzlich zu der Schätzung der Entferung jeden Pixels zur Kamera, erweitert Depth Anything 3 diesen Ansatz , indem zusätzlich die Richtung im 3D-Raum geschätzt wird. Diese Kombination sorgt dafür, dass jedes Pixel eine vollständige 3D-Koordinate im Raum rekonstruiert. Dadurch wird eine grundlegend wichtige Verbindung zwischen bildbasierten Tiefenmodellen und der angezielten 3D-Repräsentation, wie Punktwolken oder Volumenmodellen, gebildet.

2.2.4 Kamerapositionsinferenz

Depth Anything 3 inferiert Kameraposen automatisch. Das Modell ist darauf ausgelegt, räumliche Strukturen, sowie die relative Position und Orientierung der Kamera gleichzeitig zu lernen.

2.3 Architektur von Depth Anything 3

2.3.1 Transformer-Backbone und Tokenisierung

Ein Vision-Transformer-Encoder, der auf dem vortrainierten DINOv2-Modell basiert, werden Eingabebilder zunächst in überlappende oder nicht-überlappende Patches zerlegt und in

Token umgewandelt. Diese Token repräsentieren Bildbereiche und erhalten durch Self-Attention-Mechanismen des Tranformers globale Kontextinformationen.

2.3.2 Muli-View-Token

Bei der Verarbeitung mehrerer Bilder wird ein Multi-View-Token benutzt. Bei diesen Tokens werden unterschiedlichen Ansichten durch Transformer gemeinsam verarbeitet, wodurch das Modell Korrespondenzen zwischen Bildern lernt. Dieser Ansatz erlaubt eine flexible und robuste Integration von Bildsequenzen.

2.3.3 Verlustfunktionen

Depth Anything nutzt mehrere geometrische Loss-Terme, die sowohl die Genauigkeit einzelner Vorhersagen optimieren und zusätzlich explizit die Konsistenz zwischen verschiedener Ansichten und Ausgaben erzwingt. Dadurch wird verhindert, dass rein Visuelles gelernt wird und stattdessen eine stabile, physikalische Raumrepräsentation entsteht.

3.0 Methodik

3.1 lokales Setup

Als ausgangspunkt dient das ofizielle GitHub-Repository, welches die vollständige Implementierung des Modells sowie Beispielskripte enthält. Das Repository wurde zunächst geforkt, um projektspezifische Anpassungen vorzunehmen, ohne die Originalquelle zu verändern und gleichzeitig sicherzustellen, dass zukünftige Updates des Ursprungprojektes weiterhin nachvollziehen zu können.

Nach dem Fork wurde das Repository lokal geklont. Das lokale Setup umfasste die angegeben Abhängigkeit, sowie die Konfiguration der Laufzeitumgebung entsprechend der Hardware- und Betriebsgegebenheiten. Konkret wurden im bestehenden Code die Pfadangaben für Eingabe- und Ausgabendaten angepasst. Hierbei wurden keine Modifikationen der Modellarchitektur vorgenommen. Da jedoch ein Teil der Eingabedaten als Videos vorlagen, diente ein zusätzliches Hilfsskript zur Frame-Extraktion als Erweiterung. Dieses Skript dient zur Bildframes zerlegung einzelner Videos, die anschließend als Eingabe verwendet werden können.

3.2 Tests und Experiemente

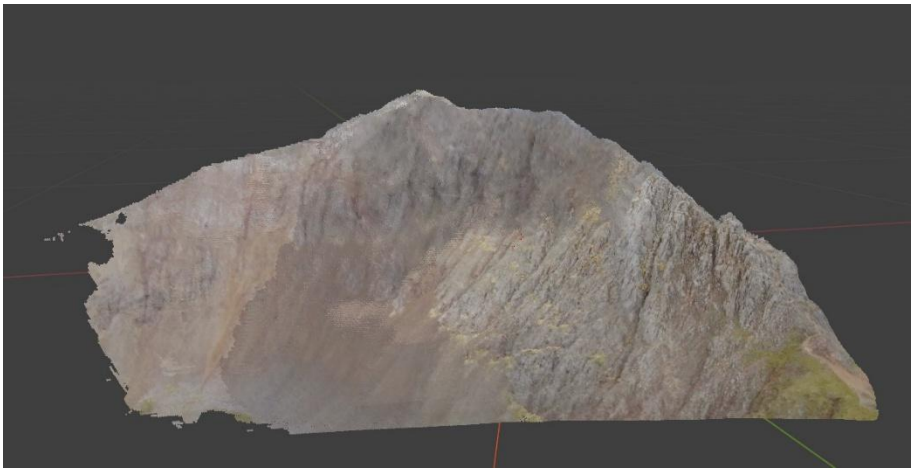
Zur Untersuchung der Leistungsfähigkeit von Depth Anything 3 hinsichtlich rekonstruktion von Objekten wurden verscheidene Tests und Experimente mit unterscheidlichen Datentypen und Szenen durchgeführt. Ziel dieser Experimente war es nicht nur, das Modell im vorgesehenen Einsaztbereich – der Rekonstruktion größräumiger Szenen – zu evaluieren, sondern gezielt auch dessen Grenzen zu untersuchen. Insbesondere wurde alaysiert, inwieweit sich der Technische Ansazt des Modelles auch auf isolierte

Einzelobjekte übertragen lässt, obwohl das Modell primär für die Rekonstruktion von Szenen konzipiert ist.

Als Input kamen sowohl eigene Fotos und Videos als auch öffentlich verfügbare Bilder aus dem Internet zum Einsatz. Die kombination dieser Datenquellen ermöglichte es verschieden kontrollierte Eingabeszenarien zu testen. Für jeden Input wurden jeweils ein repräsentatives Eingabebeispiel sowie das daraus resultierende, nachverarbeitete 3D-Modell dokumentiert, um den Zusammenhang zwischen Eingangsdaten und Ergebnis nachvollziehbar darzustellen.

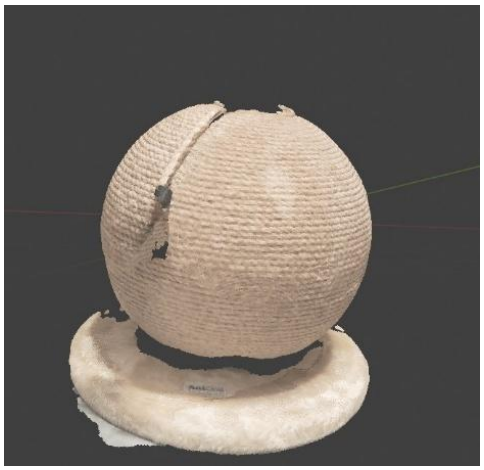
3.2.1 Rekonstruktion von Szenen

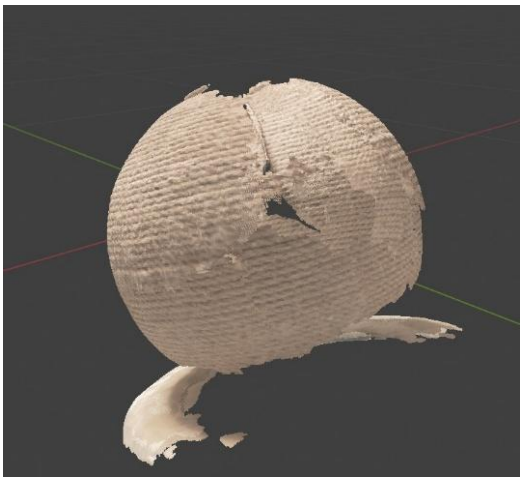
Zum Testen, wie sich das Modell mit der Rekonstruktion von Szenen verhält, wurden einzelne Frames einer Drohnenfahrt um eine Berglandschaft als Eingabedaten verwendet. Diese Szene weist große Tiefenunterscheide, natürliche Strukturen und eine klare räumliche Staffelung auf und eignet sich daher besonders zur Bewertung der szenenbasierten Rekonstruktionsfähigkeit.



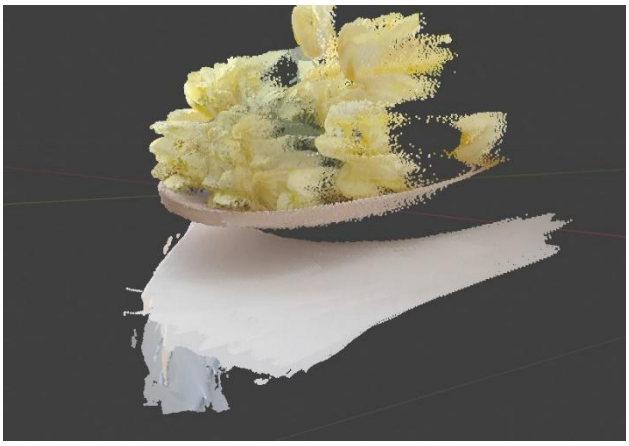
Zum weiteren Testen wurde zunächst ein „walk through“ durch eine Bibiliothek in einzelne Frames extrahiert, welche anschließend als Eingabedaten dienten. Dieses Modell diente der Untersuchung, wie sich das Modell mit schwierigen Lichtverhältnissen und Oberflächen wie Glas verhält.

Um die Grenzen des Modells bezüglich seines primären Einsatzbereiches zu untersuchen, wurden Einzelobjekte getestet. Hierzu zählte unter anderem ein Katzenspielzeug, das aus mehreren Blickwinkeln fotografiert wurde. Die zugehörigen Eingabebilder zeigen ein klar abgegrenztes Objekt. Das daraus erzeugte und remeshte 3D-Modell lässt die grundlegende Form des Objekts erkennen, bleibt jedoch detailarm und sehr löchrig. Dieses Ergebnis verdeutlicht, dass die Rekonstruktion isolierter Objekte mit Depth Anything 3 deutlich instabieler ist als bei szenenbasierten Anwendungen.





Ein weiteres Testobjekt stelle ein Blumengesteck dar, dessem organische und feingliedrige Struktur eine besondere Herausforderung für die Tiefenschätzung darstellt. Die Eingabebilder zeigen ein visuell komplexes Objekt mit mehreren Überlappungen. Im remeshten 3D-Modell sind zar kleinere Details, wie einzelne Blüten zu erkennen, jedoch ist kein insich geschlossenes Modell zu erkennen. Auch dies unterstreicht die eingeschränkte Eignung des Tools für detaillierte Objektrekonstruktionen aus wenigen Ansichten



Alle erzeugten 3D-Repräsentationen wurden im Glb-Format exportiert und anschließend in Blender importiert. Dabei zeigte sich, dass die Rohgeometrie ohne zusätzliche Nachverarbeitung visuell schwer interpretierbar und für eine Weiterverwendung nicht geeignet waren. Aus diesem Grund wurde für alle Modelle ein Remeshing-Schritt durchgeführt. Erst durch diese geometrische Verienheitlichung wurden die rekosntruierten Szenen und Objekte als zusammenhängende, erkennbare 3D-Form sichtbar.

Die Gegenüberstellung der Eingabedaten und der remeshten Ergebnisse zeigt deutlich, dass die Stärken des Tools vor allem bei großräumigen Szenen liegen. Bei isolierten

Einzelobjekten stößt das Tool an klare Grenzen, selbst wenn zusätzlihce Nachverarbeitungsschritte eingesetzt werden.

4.0 Ergebnisse

Die im Rahmen der Experimente erzielten Ergbenissen zeigen, dass Depth Anything 3 grundsätzlich in der Lage ist, aus mehransichtigen Eingabedaten konsistente dreidemensionale Repräsentationen zu erzeugen, jedoch deutliche Grenzen aufweist. Diese Grenzen lassen sich unmittelbar auf die Architektur und den technischen Ansatz des Modelles zurückführen.

4.1Ergebnisse und Fehler bei Einzelobjekten

Die Rekonstruktion einzelner Objekte zeigen besonders ausgepräkte Schwächen. IN den remeshten 3D-Modellen traten starke Lochbildungen auf, welche selbst bei einer hohen Anzahl von Eingabedaten aus verscheidenen Blickwinkeln nicht vollständig verschwanden.

Diese Beobachtungen lassen sich mit der Modellarchitektur erklären. Depth Anything 3 ist kein expliziertes 3D-Rekonstruktionsmodell, sonderl ein Tiefenschätzungsmodell, das Tiefeninformationen bildweise vorhersargt. Die 3D-Geometrie ensteht lediglich indirekt aus diesen Tiefenkarten. Eine Modelierung von geschlossenen Objektoberflächen oder Volumen finden nicht statt. Einzelobjekte werden zudem nicht als geschlossene Enheit erkannt, sonder lediglich als Teil eines Bildes interpretiert.

Auch die Verwendung vieler Eingabebilder aus unterschiedlichen Ansichten konnte dieses Probelem nicht vollständig lösen. Die Architektur des tools sieht keine Multi-view_Fusion auf Objektebene vor. Die Tiefneschätzung mehrerer Ansichten werden nicht unter globalen, geometrischen Konsistenzbedinungen optimiert, wodurch sich Unsicherheiten und inkonsistenzen direkt in Form von Löchern und unvollständigen Oberflächen im 3D-Modell zeigen.

4.2 Ergebnisse und Fehler bei szenenbasierten Rekonstruktionen

Die Ergebnisse der Szenenrekonstruktion zeigen jedoch insgesamt stabilere Rekonstruktion. Die 3D-Modelle der Szenen weisen eine klare zusammenhängende Struktur auf. Zwar treten auch hier lokale Lücken auf, diese beeinträchtigen jedoch nicht die globale räumliche Kohärenz der Szene.

Szenen enthalten zahlreiche Tiefenhinweise wie Größenverhältnisse, wiederkehrende Strukturen und klaw Vorder- und Hintergrundbeziehungen. Diese Informationen können durch die transformerbasierte Architektur effektiv genutzt werden, um eine stabile Tiefenschätzung zu erzeugen. Die resultierenden 3D-Modelle sind daher insgesamt konsistenter.

4.3 Architekturbedingte Ursachen der beobachteten Grenzen

Depth Anything 3 arbeitet mit relativen Tiefenwerten, die nicht absolut skaliert sind. Dies erschwert insbesondere bei Einzelobjekten die präzise Zusammenführung mehrerer Ansichten zu einer geschlossenen Geometrie. Des weiteren ist die Architektur auf die erfassung globaer Kontextinformationen optimiert. Während diese bei großräumigen Szenen von Vorteil ist, führt es bei feinen Sturkturen zu einer geringen Rekonstruktionsgenauigkeit. Dünne Objektteile, komplexe Oberflächen oder organische Formen können dadurch nicht zuverlässig erfasst werden. Zuletzt fehlt eine expliziete 3D-Konsistenzprüfung, wie sie beispielsweise in NeRF-Ansätzen verwendet wied. Die Tifenkarten werden unabhängig voneinander geschätzt. Diese Inkonsistenzen führen in der 3D-Rekonstruktion direkt zu Löchern, wie sie sowohl bei Objekten, aber auch in Abgeschwächter Form bei Szenen sichtbar sind.

Die Gegenüberstellung der remeshten Szenen- und Objektmodelle zeigt deutlich, dass Depth Anything 3 überwiegend bei der Rekonstruktion großräumiger Szenen überwiegt. Einzelobjekte bleiben selbst nach zusätzlicher Nachverarbeitung unvollständig und geometrisch instabil. Die beobachteten Grenzen sind somit keine Folge unzureichender Eingabedaten, sonder eine Konsequenz der zugrunfe liegenden technischen Architektur des Tools. Diese Erkenntnisse definieren klar den sinvolleren Einsatzbereich für die Rekontstruktion von Szenen.

5.0 Quellen

5.1 Wissenschaftliche Hauptquellen:

- https://arxiv.org/abs/2511.10647 -> Depth Anything 3: Recovering the Visual Space from Any Views. Haotong Lin, Sili Chen, Junhao Liew, Donny Y. Chen, Zhenyu Li, Guang Shi, Jiashi Feng, Bingyi Kang. [Submitted on 13 Nov 2025]
- https://depth-anything-3.github.io/ -> Ofizielle Projektseite
- https://github.com/ByteDance-Seed/Depth-Anything-3 -> ofizielles GitHub-Repository

5.2 Repository Fork:

- https://github.com/issiissi/Depth-Anything-3

5.3 Weitere Quellen:

- https://beta.hyper.ai/en/papers/2511.10647 -> HyperAI / Paper Summary Website zu DA3
- https://www.themoonlight.io/en/review/depth-anything-3-recovering-the-visual-space-from-any-views -> [Literature Review] Depth Anything 3: Recovering the Visual Space from Any Views