

Folien zur Vorlesung am 25.03.2025
3D Computer Vision

MOTIVATION, ÜBERBLICK, INHALTE

Literatur und Quellen

- Buch:
 - Computer Vision: Algorithms and Applications, Richard Szeliski → In der 2. Edition vor allem die Kapitel 2, 11, 12 und 13, verfügbar unter <http://szeliski.org/book>
- Folien/Kurse:
 - [Introduction to Computer Vision, CS5670, Spring 2025, Cornell Tech](#) (vor allem) → [Noah Snavely](#) und andere: [Yung-Yu Chuang](#), [Fredo Durand](#), [Alexei Efros](#), [William Freeman](#), [James Hays](#), [Svetlana Lazebnik](#), [Andrej Karpathy](#), [Fei-Fei Li](#), [Srinivasa Narasimhan](#), [Silvio Savarese](#), [Steve Seitz](#), [Richard Szeliski](#), and [Li Zhang](#).
 - Die meisten Folien sind aus diesem Kurs (Version 2022+2024) kopiert.
- Videos:
 - Shree Nayar: <https://fpcv.cs.columbia.edu/> → Reconstruction I + II
 - Die komplette Reihe ist sehr zu empfehlen!

Computer Vision

- Einfache Fragen:
 - Was ist hier zu sehen?



[i:colo07]

Computer Vision

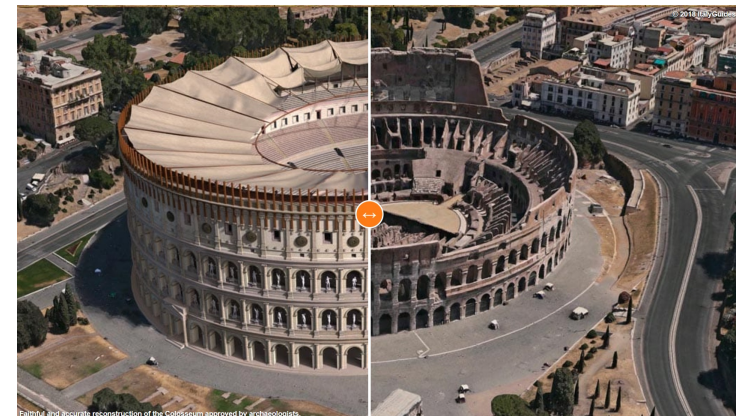
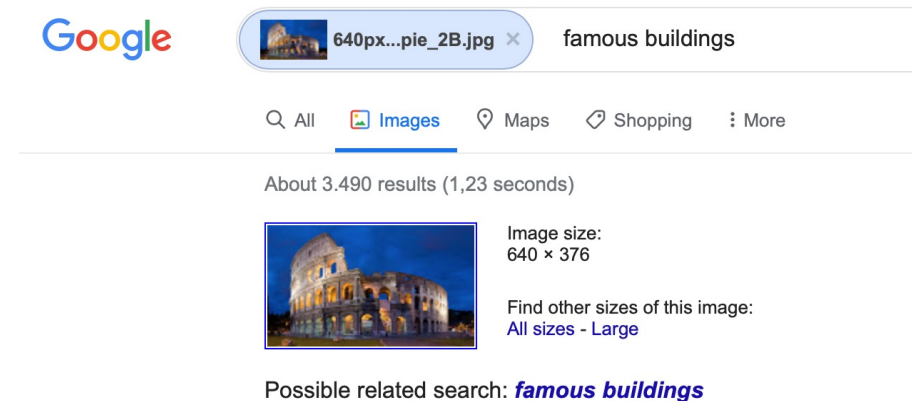
- Einfache Fragen:
 - Was ist hier zu sehen?
 - Wie sieht das Gebäude von oben/hinten aus?



[i:colo07]

Computer Vision

- Einfache Fragen:
 - Was ist hier zu sehen?
 - Wie sieht das Gebäude von oben/hinten aus?
- → Schwierig für einen Computer zu beantworten.

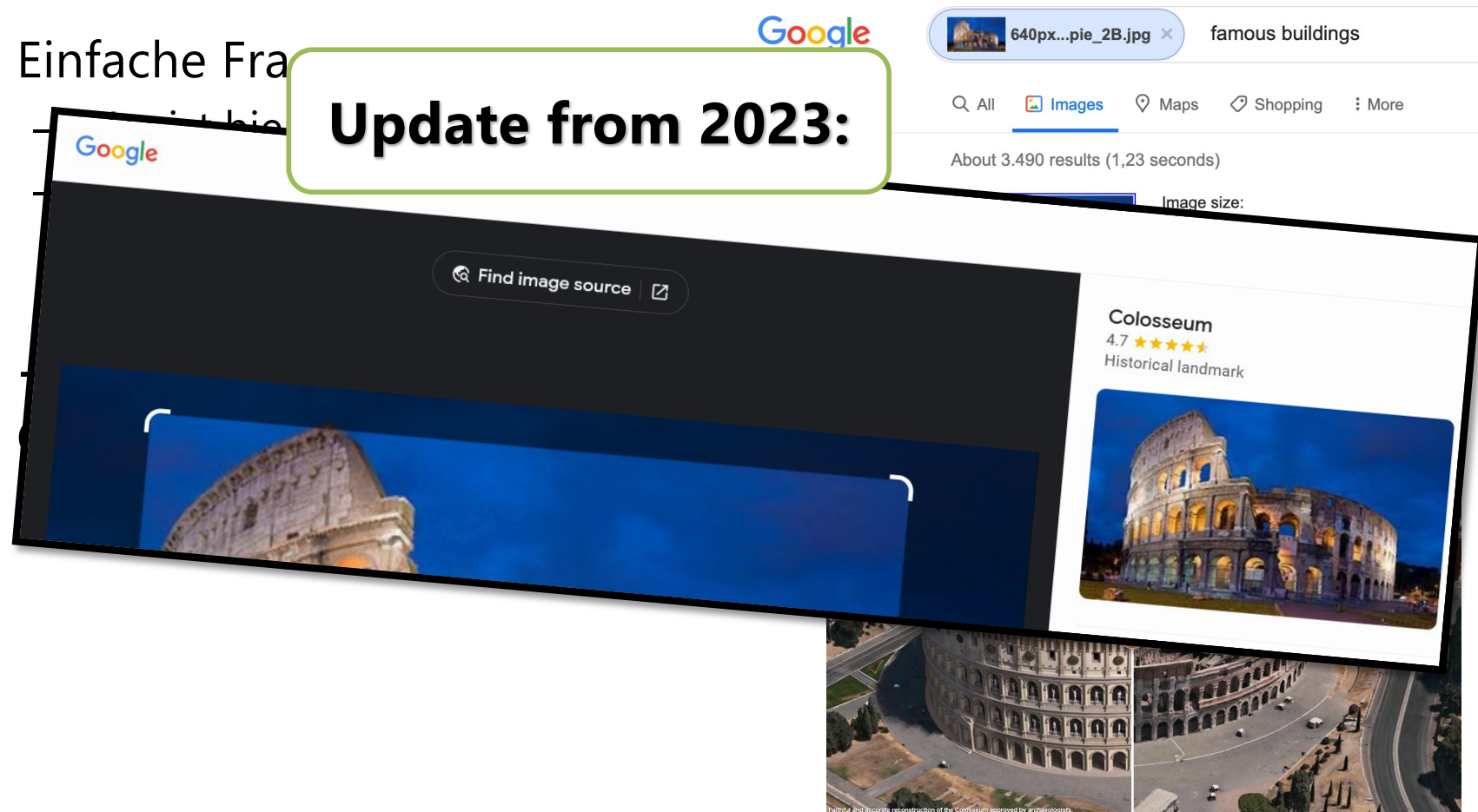


<https://www.italyguides.it/en/lazio/rome/ancient-rome/colosseum>

Computer Vision

- Einfache Fra

Update from 2023:



<https://www.italyguides.it/en/lazio/rome/ancient-rome/colosseum>

Computer Vision

- Einfache Frage
– Was ist hier zu sehen?

Update from 2025:

- → Schwierig
Computer zu

ChatGPT
Gemeinsam nutzen
UW



Was ist hier zu sehen?

Das Bild zeigt das Kolosseum in Rom, Italien. Es handelt sich um ein antikes Amphitheater, das während der römischen Kaiserzeit erbaut wurde. Es ist eines der bekanntesten Wahrzeichen Roms und ein bedeutendes architektonisches und historisches Symbol des Römischen Reiches. Das Kolosseum wurde für Gladiatorenkämpfe, Tierhetzen und öffentliche Spektakel genutzt und ist heute eine der meistbesuchten Touristenattraktionen der Welt.









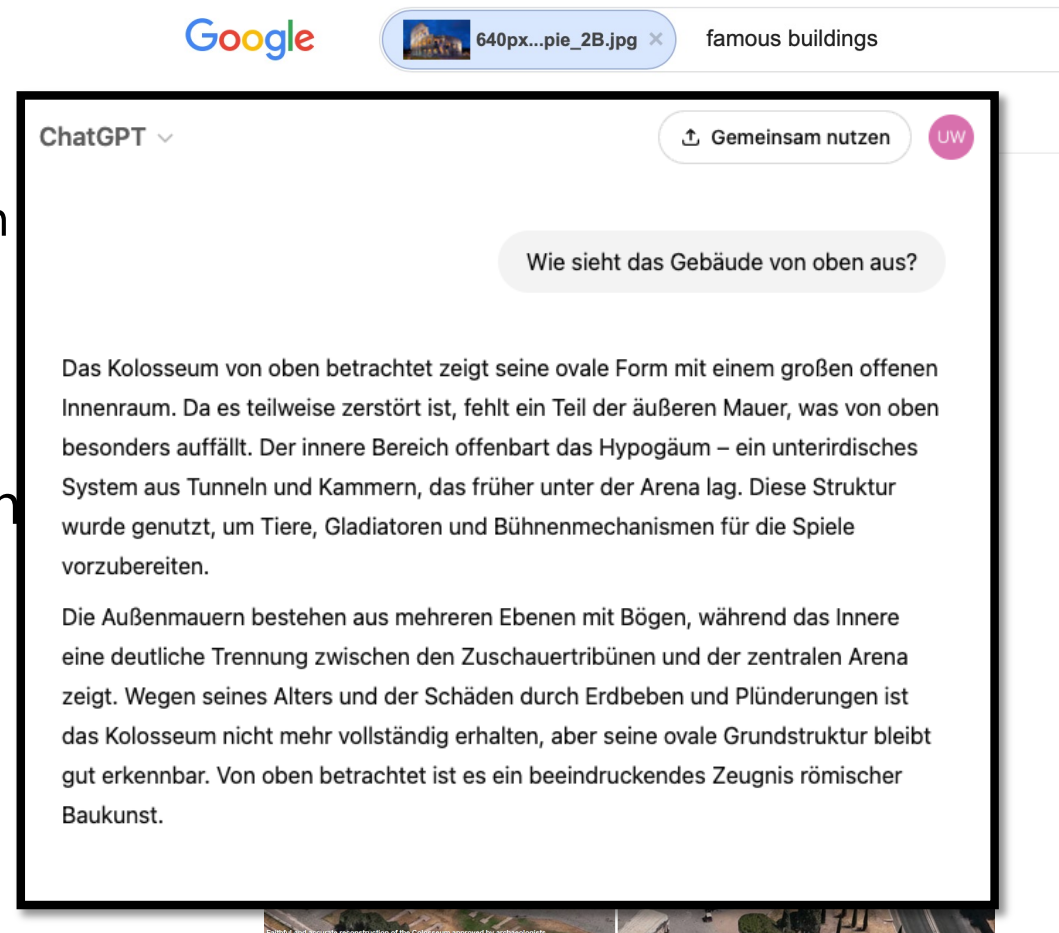
[e/colosseum](#)

Computer Vision

- Einfache Fragen:
 - Was ist hier zu sehen?

Update from 2025:

- → Schwierig für einen Computer zu beantworten



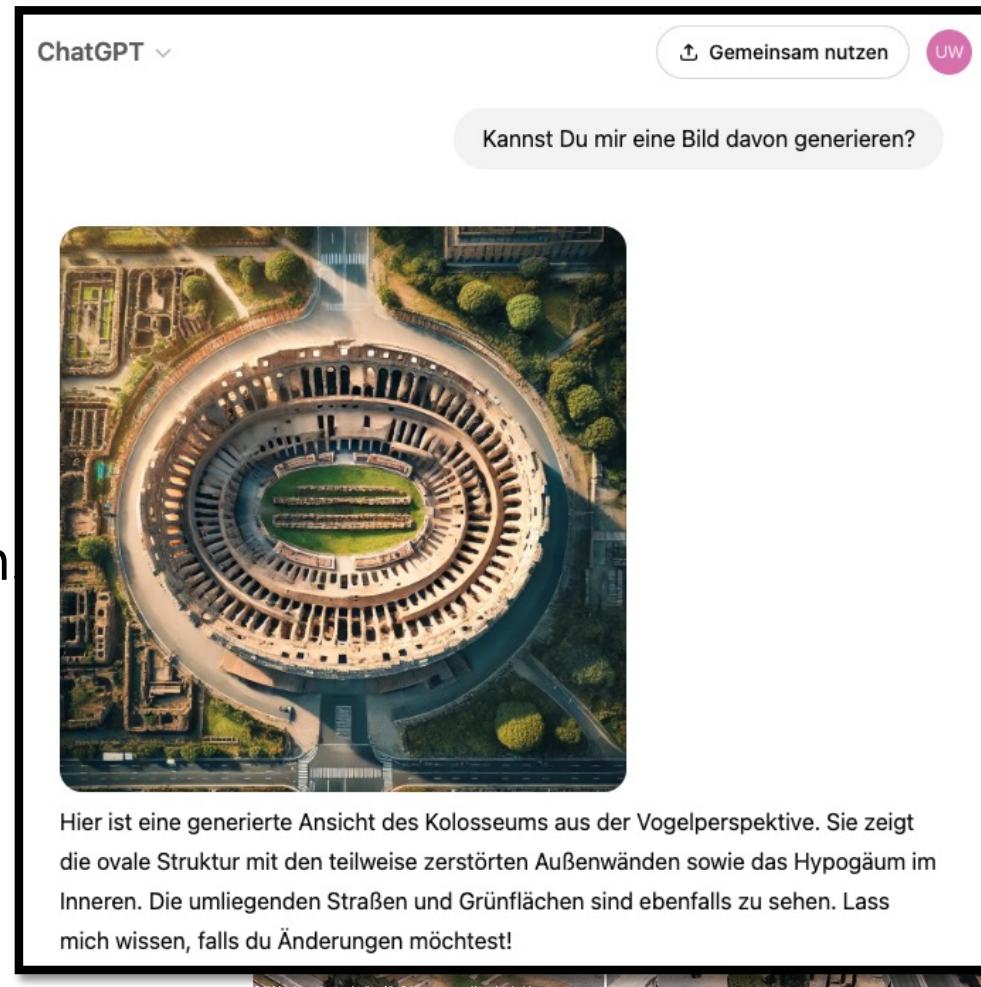
<https://www.italyguides.it/en/lazio/rome/ancient-rome/colosseum>

Computer Vision

- Einfache Fragen:
– Was ist hier zu sehen?

Update from 2025:

- → Schwierig für einen Computer zu beantworten



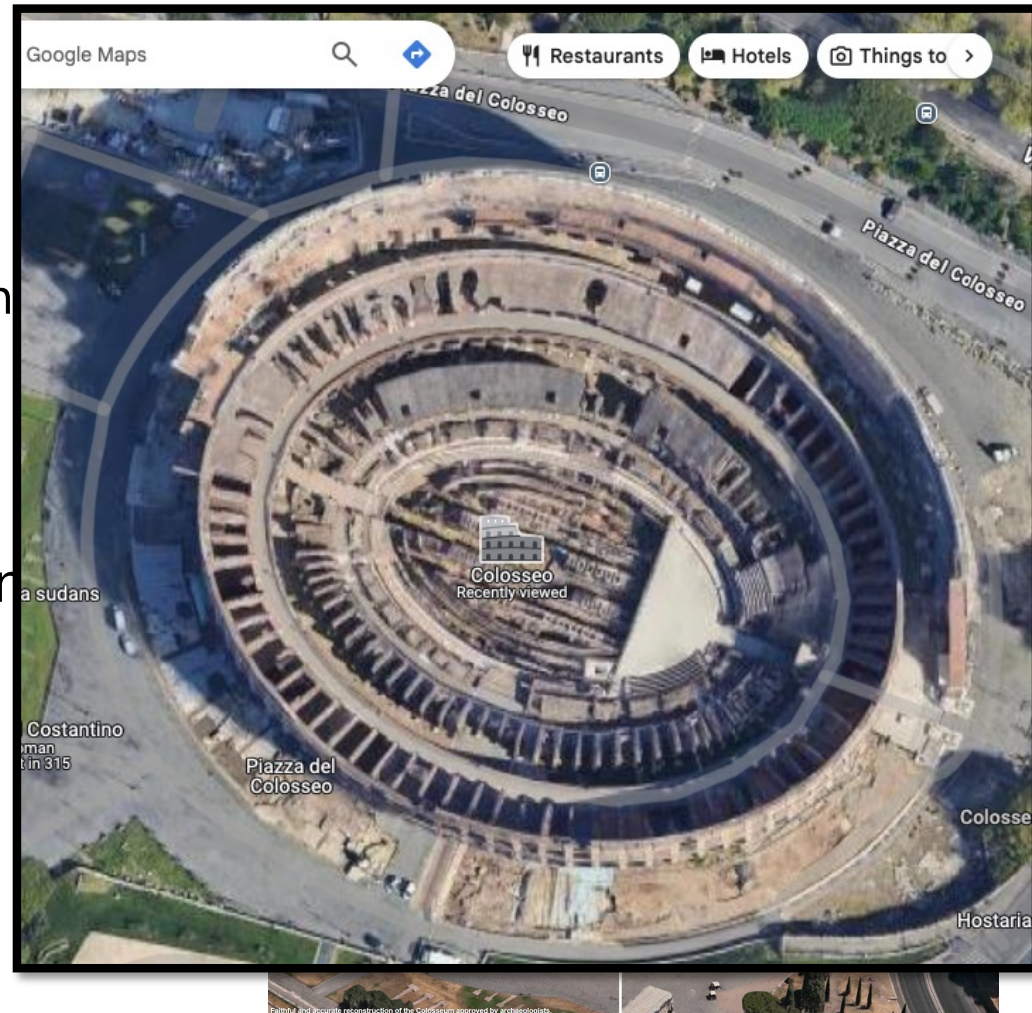
<https://www.italyguides.it/en/lazio/rome/ancient-rome/colosseum>

Computer Vision

- Einfache Fragen:
 - Was ist hier zu sehen?

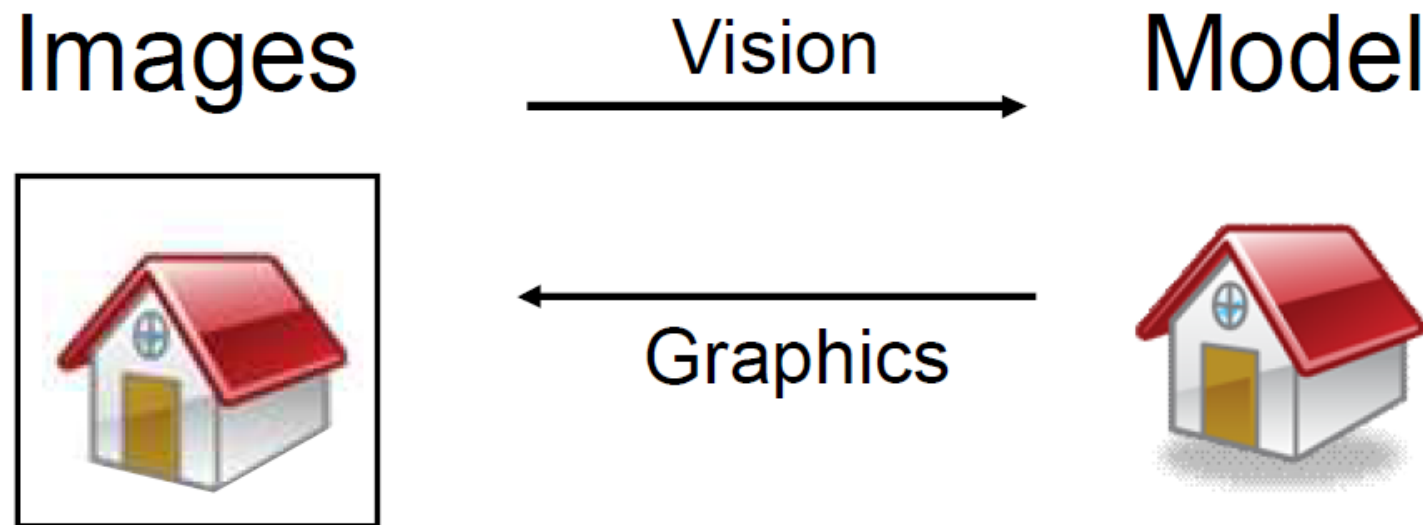
Update from 2025:

- → Schwierig für einen Computer zu beantworten



<https://www.italyguides.it/en/lazio/rome/ancient-rome/colosseum>

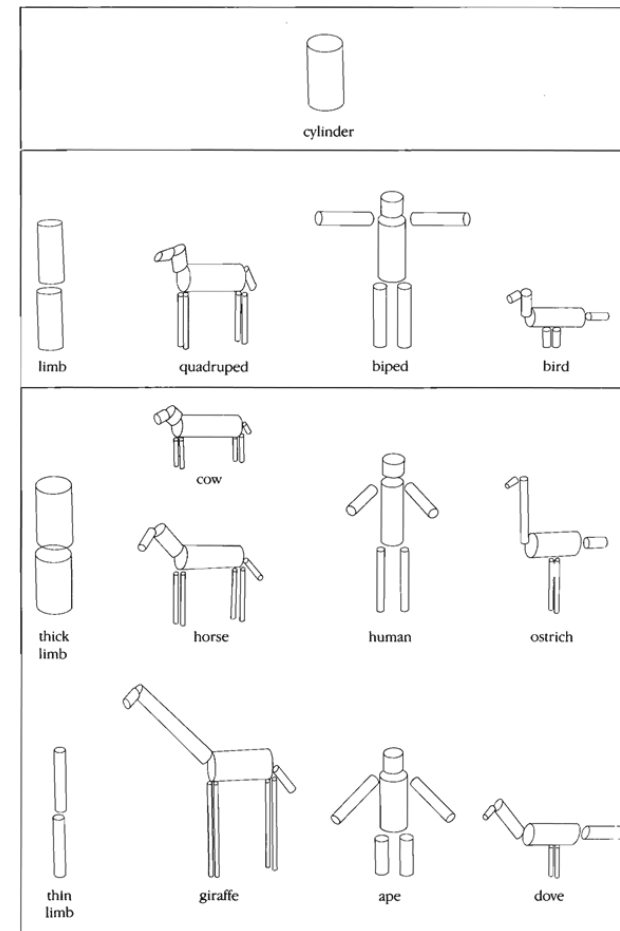
Computer Vision vs Computer Graphics



Slide is from <http://vision.cs.utexas.edu/376-spring2018/>

Computer Vision

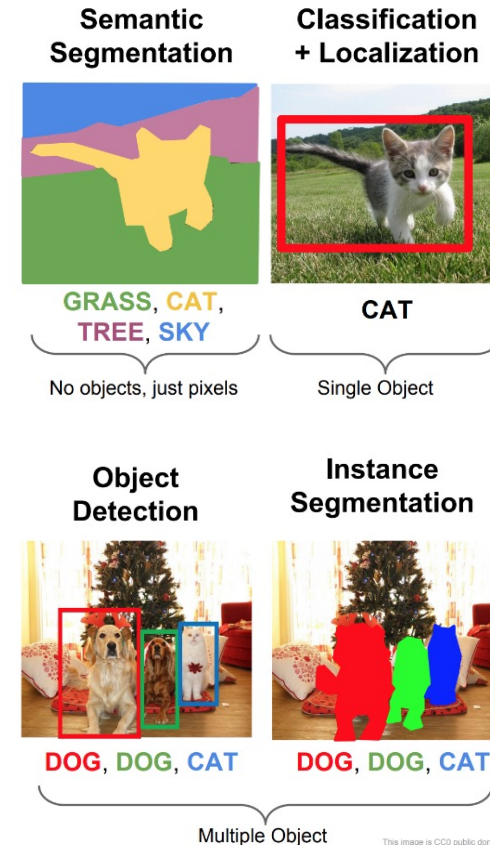
- Ein früherer Ansatz aus den 80er Jahren:
- **Wenn man die Objekte geometrisch rekonstruiert, kann man sie erkennen.**



[i:marr82]

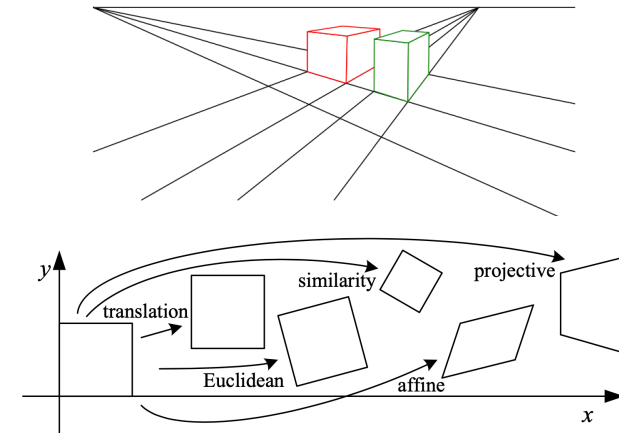
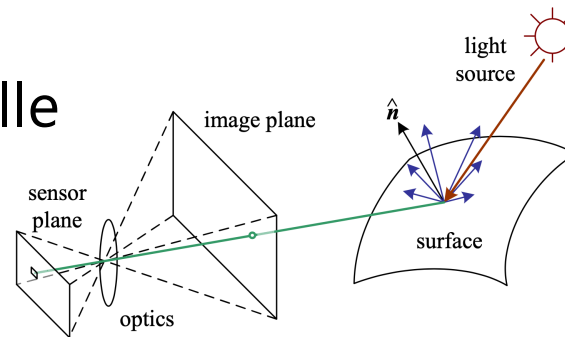
Computer Vision

- Begriffe
 - „Reconstruction“ → welche Geometrie hat das Objekt?
 - „Recognition“ → was ist das für ein Objekt?
 - „Detection“ → wo (im Bild) befindet sich das Objekt?
 - „Object“ vs „Scene“ → Geht es um einzelne Objekte oder um die gesamte Umgebung?

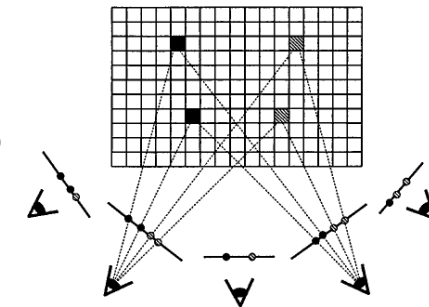
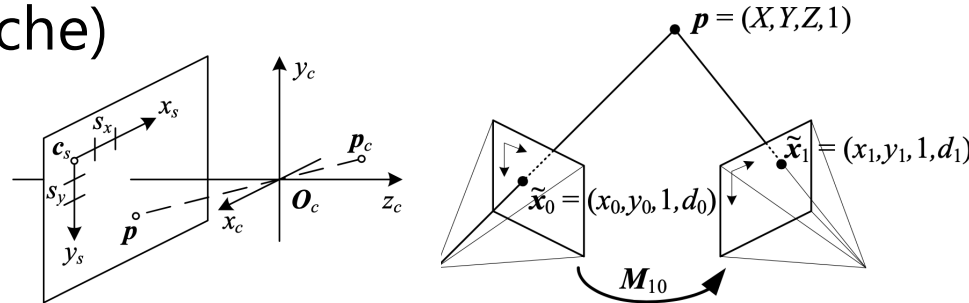


Klassische Computer Vision

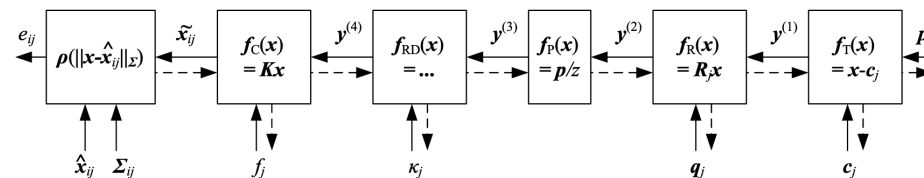
- Physikalische Modelle



- Mathematische (geometrische) Modelle

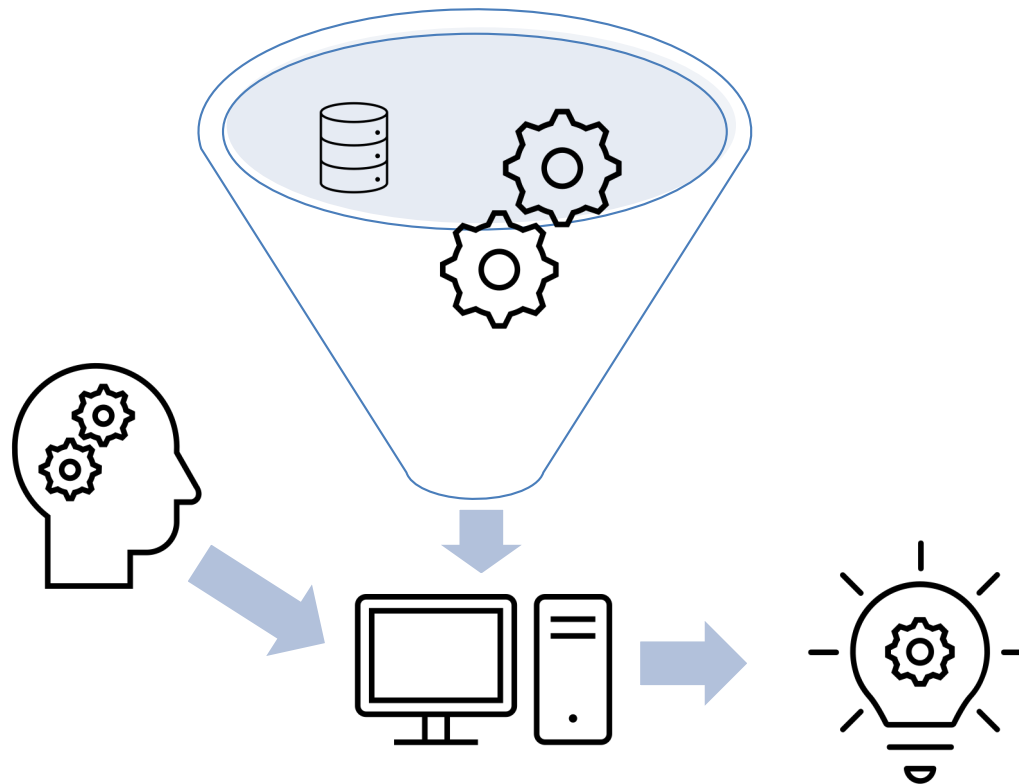


- Komplexe, aber nachvollziehbare Prozesse

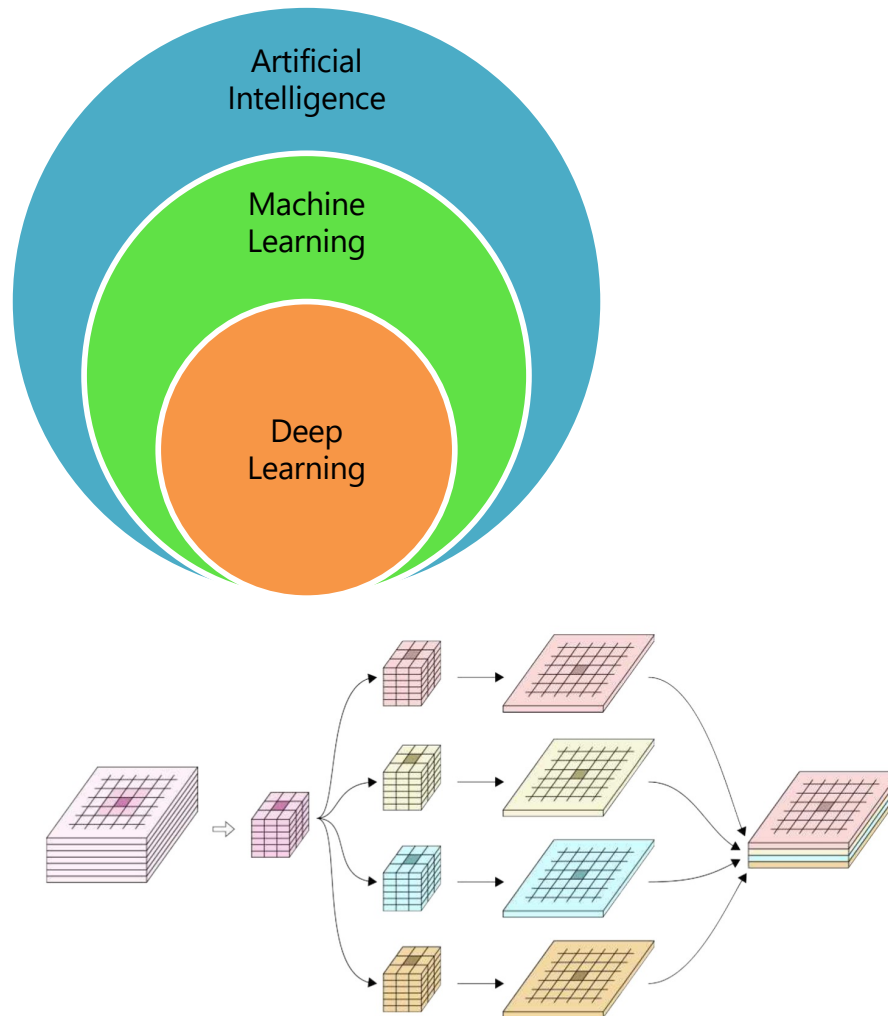


[i:szeliski20]

Klassische Computer Vision

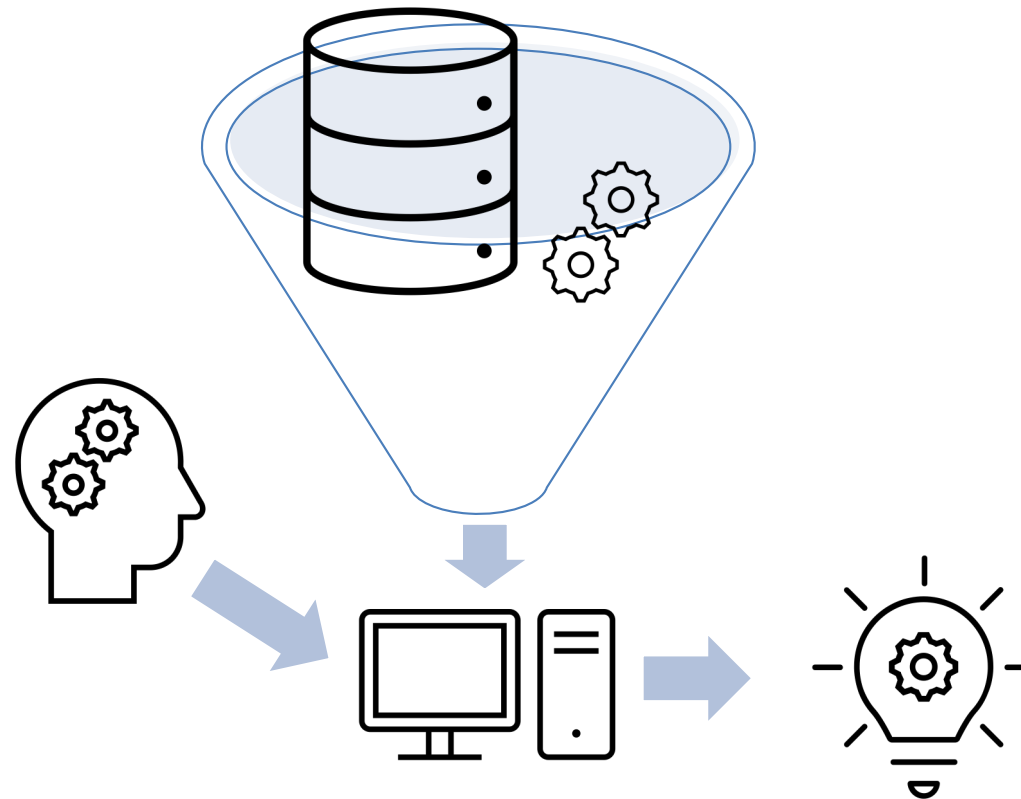


Deep Learning

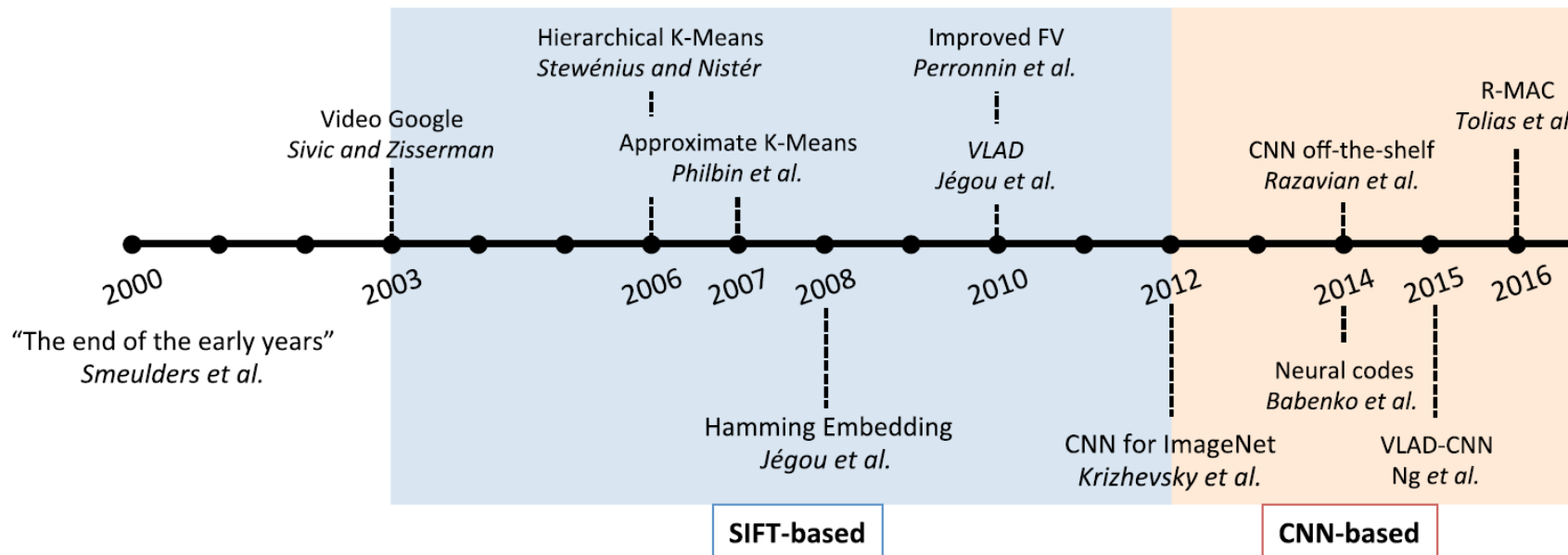


- Künstliche Intelligenz (Artificial Intelligence) als Überbegriff für wissensbasierte Entscheidungsfähigkeit
- Maschinelles Lernen (Machine Learning) als Gruppe von Algorithmen, die ein Modell aus Trainingsdaten erstellen.
 - Überwachtes Lernen (Supervised learning)
 - Unüberwachtes Lernen (Unsupervised learning)
 - Bestärkendes Lernen (Reinforcement learning)
- Deep Learning als Untergruppe, die künstliche Neuronale Netze (Artificial Neural Network) als Modell verwendet.
 - Faltendes Neuronales Netz (convolutional neural network)

Deep Learning



Übergang von klassischer Computer Vision zu Deep Learning



[i:szeliski20]

Entwicklung von Deep Learning

NUMBER of GITHUB STARS by AI LIBRARY (OVER 40K STARS), 2014–21

Source: GitHub, 2021 | Chart: 2022 AI Index Report

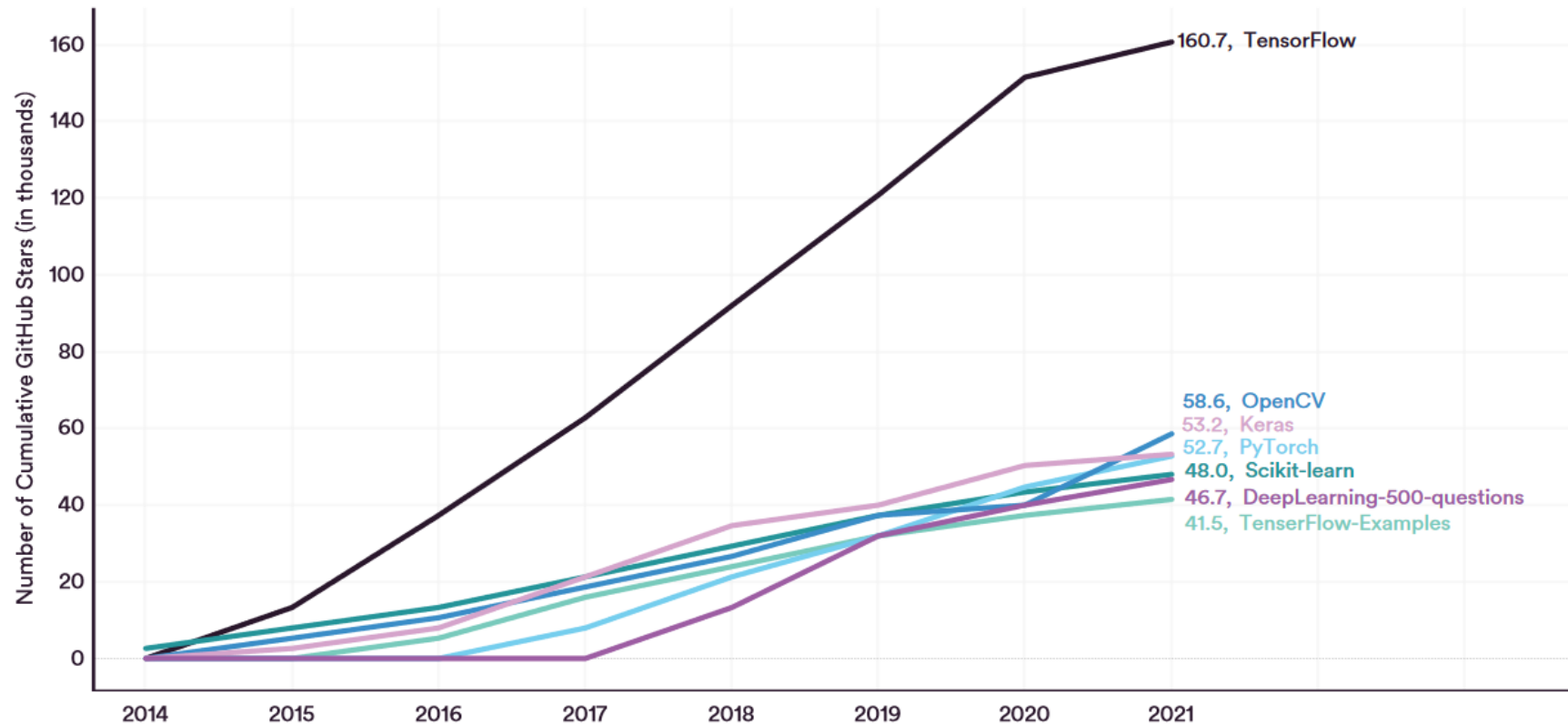
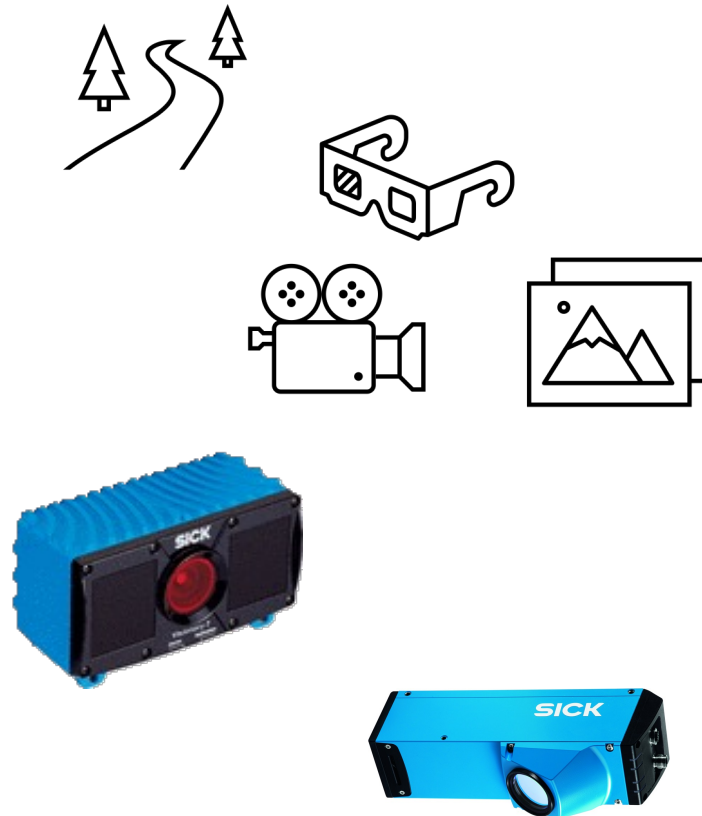


Figure 1.3.1

Image from Stanford Artificial Intelligence Index Report 2022

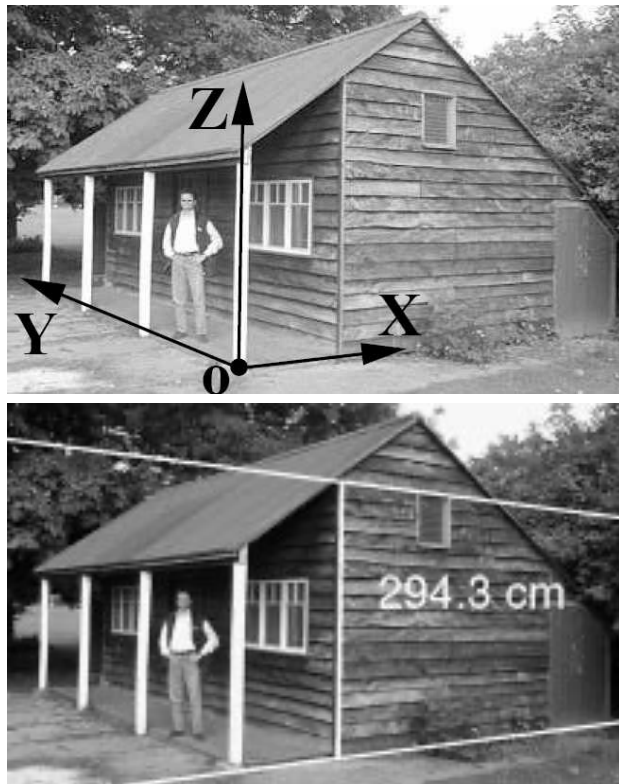
3D Rekonstruktion

- Aus Kamerabildern
 - Monokulare Tiefenkriterien
 - Stereo
 - Structure from Motion
 - ...
- Mit speziellen Sensoren
 - Time of Flight
 - Triangulation



Aus einem einzelnen Bild

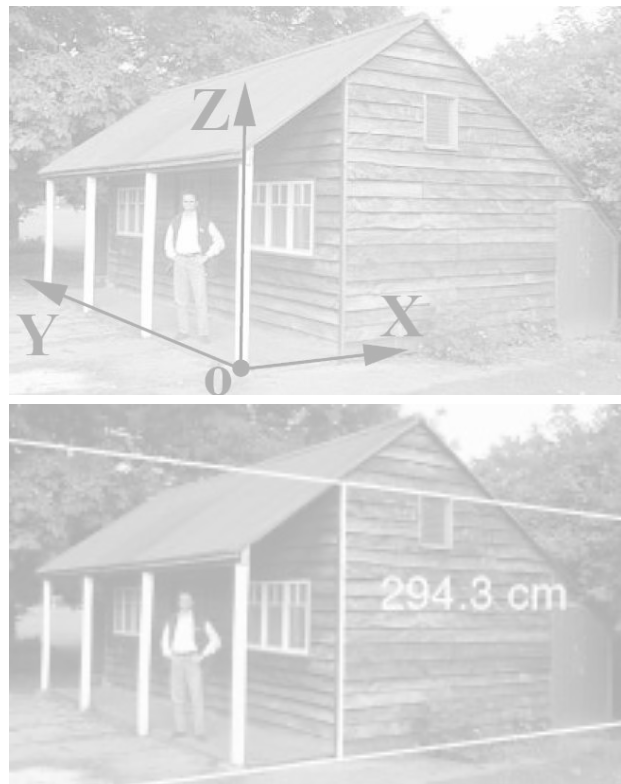
- Klassische Lösung



[i:criminsi99]

Aus einem einzelnen Bild

- Klassische Lösung



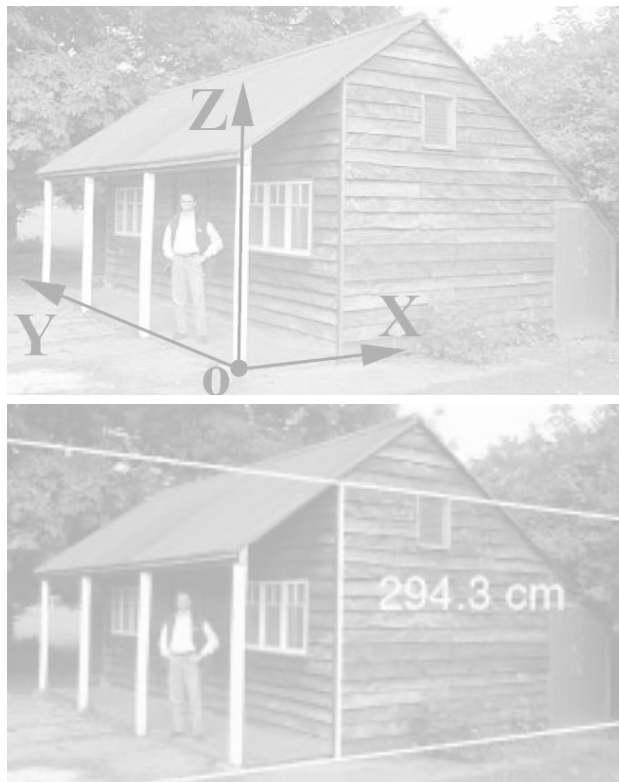
- Deep Learning (2020)



[v:facebook20]

Aus einem einzelnen Bild

- Klassische Lösung



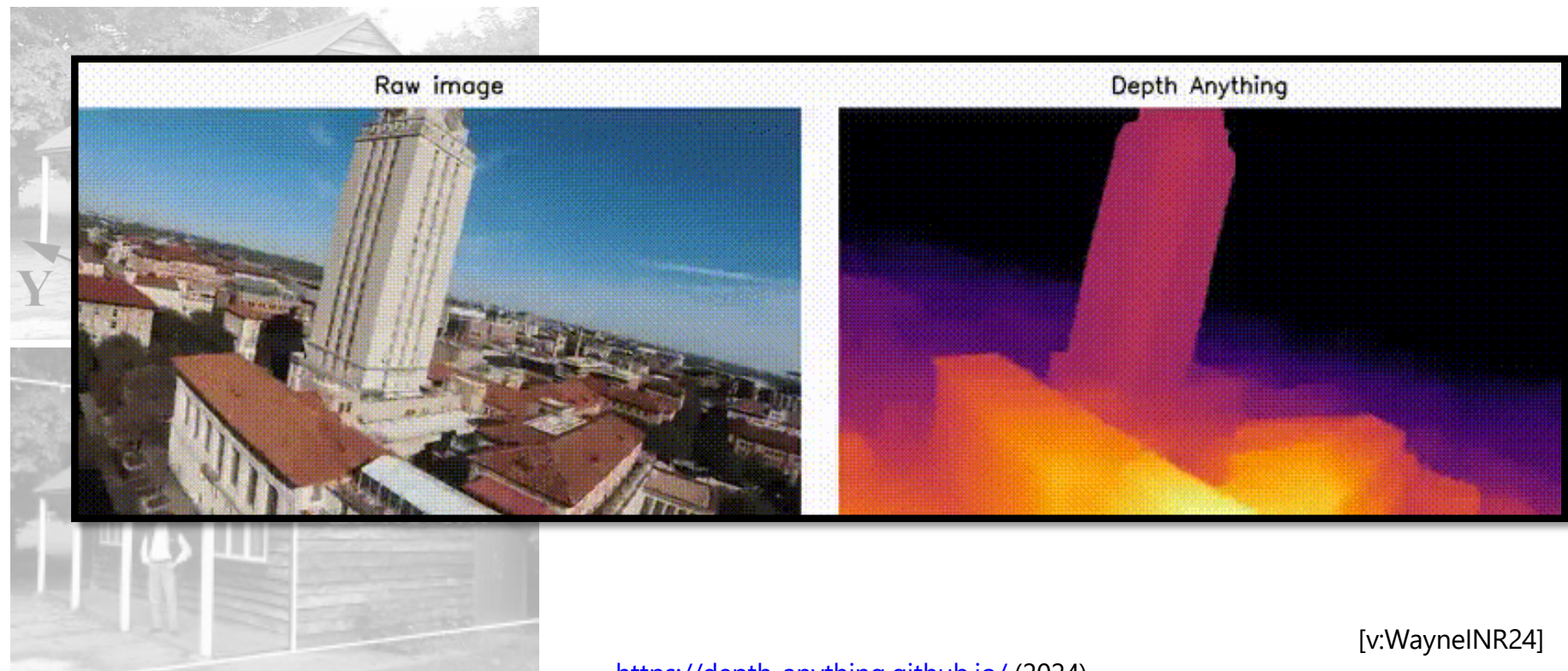
- Deep Learning (2020)



[v:facebook20]

Aus einem einzelnen Bild

- Klassische Lösung
- Deep Learning (2024)

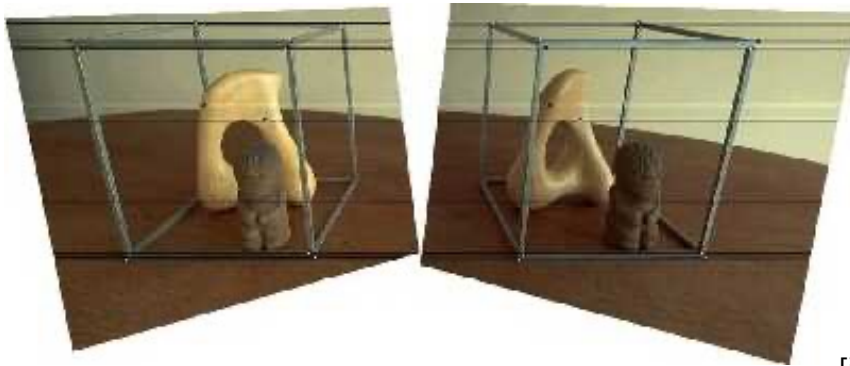


<https://depth-anything.github.io/> (2024)

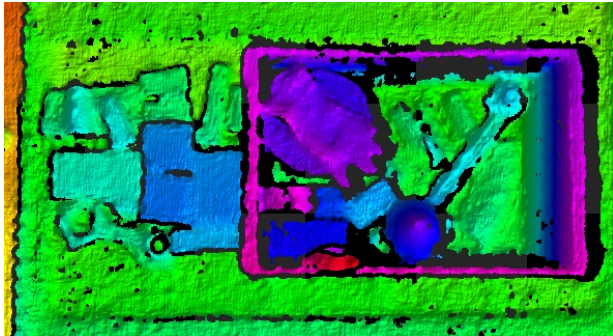
[v:WayneINR24]

Aus zwei Bildern

- Klassische Stereo Vision

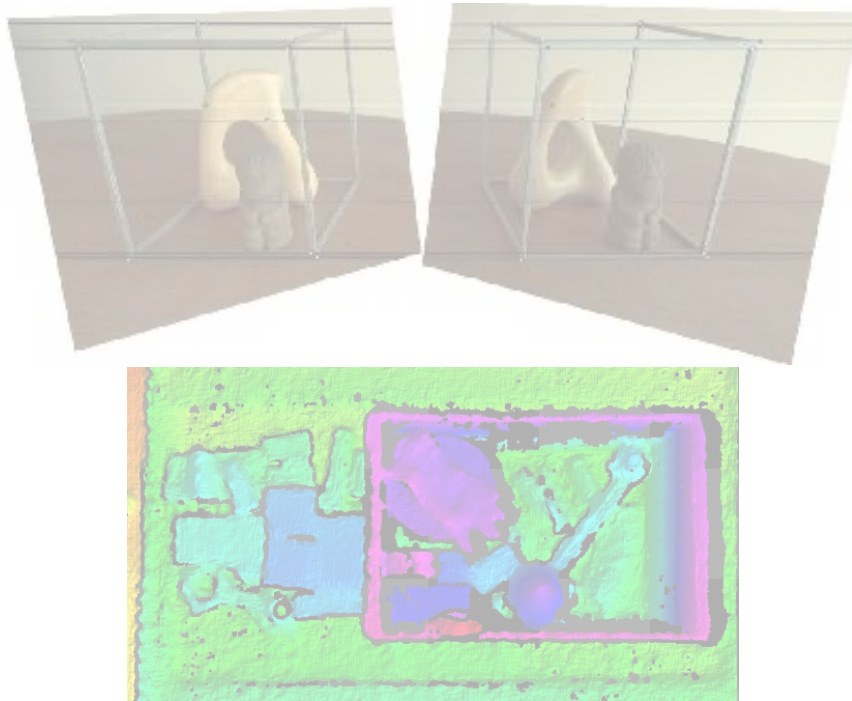


[i:szeliski20]

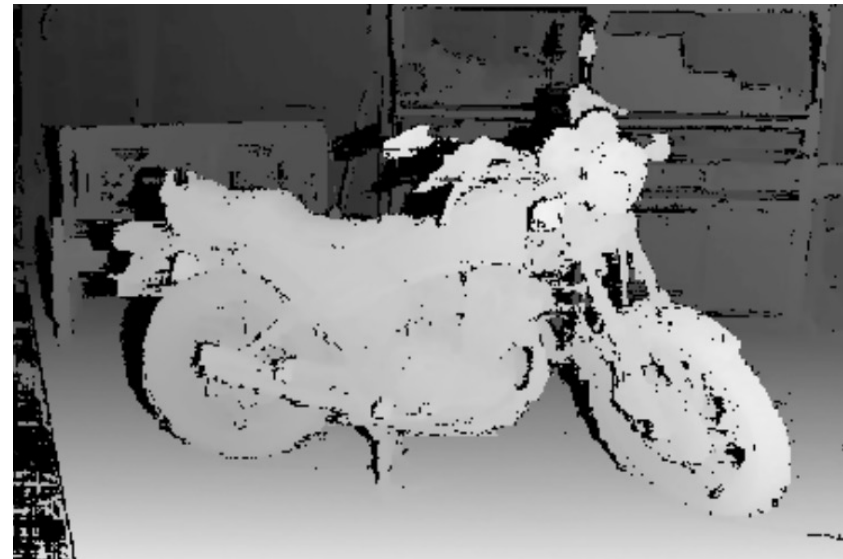


Aus zwei Bildern

- Klassische Stereo Vision

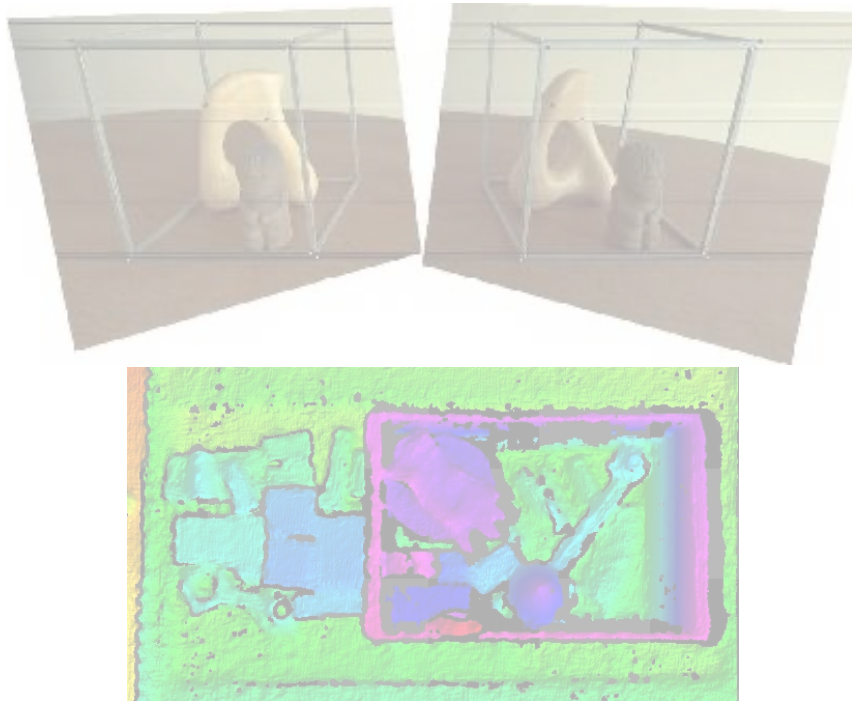


- Deep Learning (2020)
 - [i:puglia20] – Intel SystemOnChip Lösung mit CNN Unterstützung und bis zu 250fps

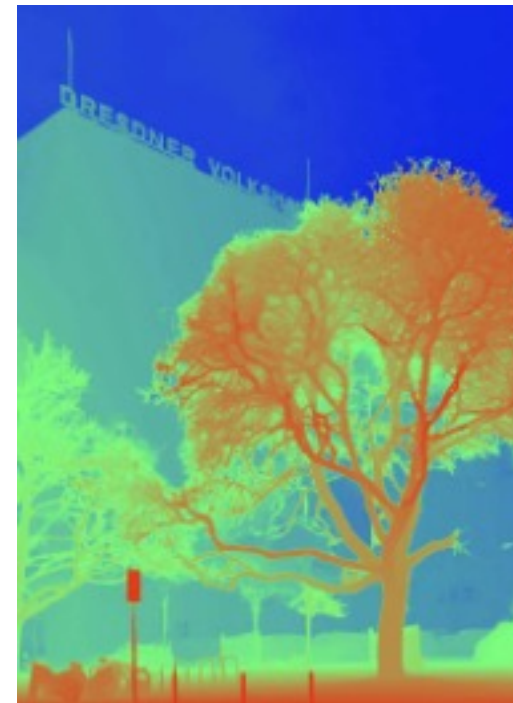


Aus zwei Bildern

- Klassische Stereo Vision



- Deep Learning (2025)
 - DEFOM–Stereo: Depth Foundation Model Based Stereo Matching, [Jiang2025]



Aus vielen Bildern

- Structure from Motion (SfM)



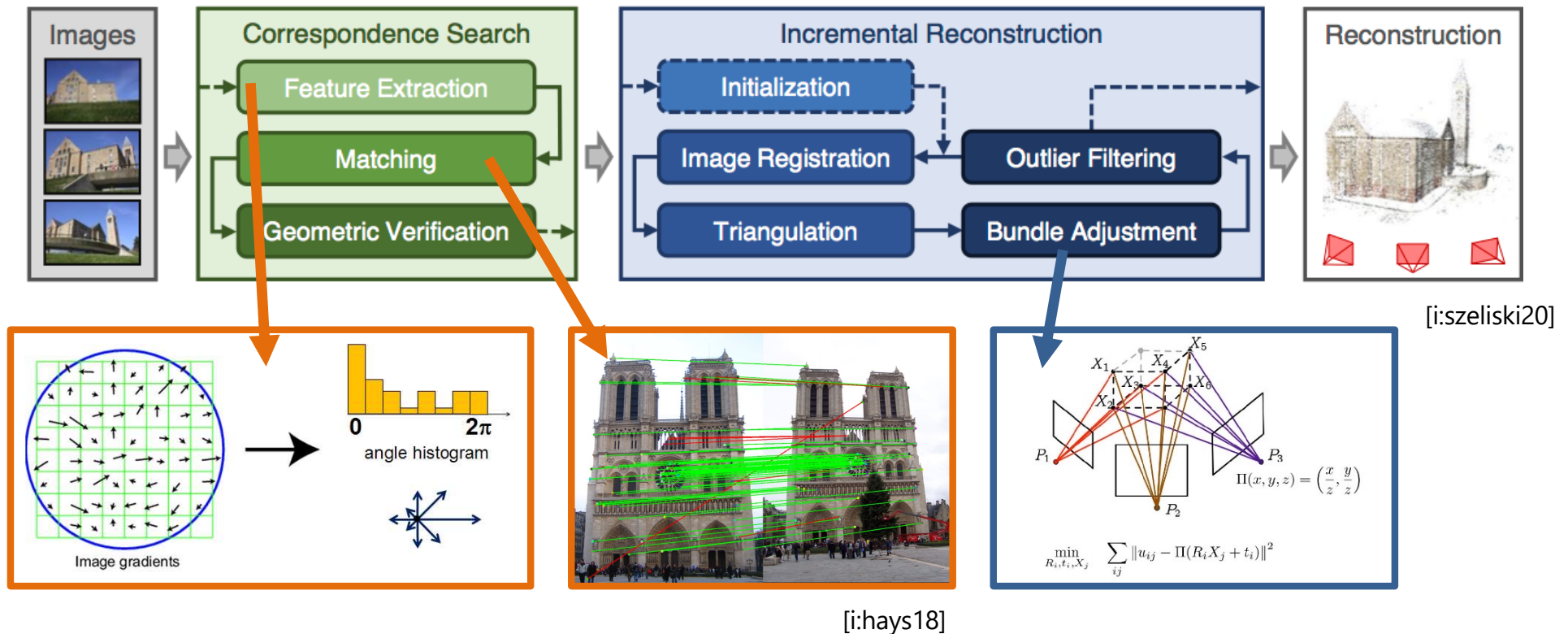
[v:agarwal09]



[i:colo10]

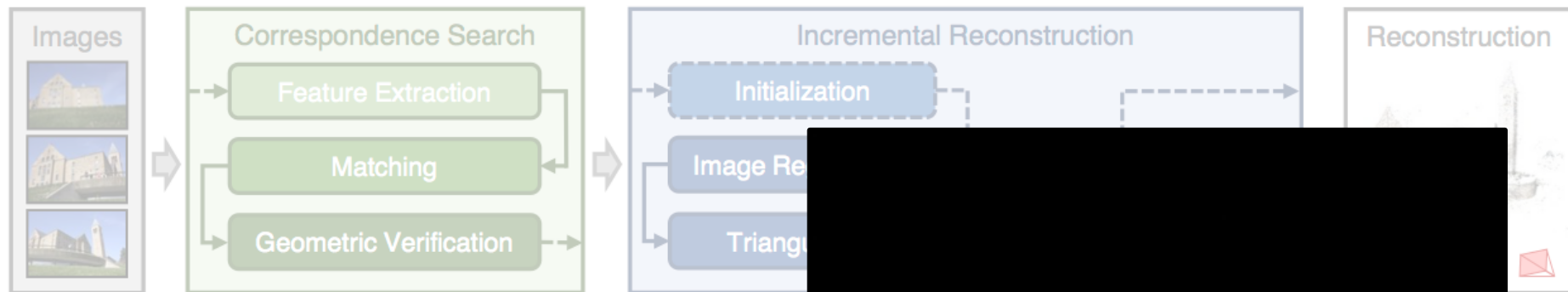
Aus vielen Bildern

- Structure from Motion (SfM)



Aus vielen Bildern

- Structure from Motion (SfM)



- Deep Learning
 - Mit Semantik

[v:cherabier18]

Aus (nicht ganz so) vielen Bildern

- NeRF (2020)
 - Training: 1-2 Tage, Rendering: ca 2-3 Minuten pro Frame



[v:mildenhall20]

Aus (nicht ganz so) vielen Bildern

- Instant-NGP, 2022 (NVIDIA Implementierung)



[v:müller22]

Deep Learning ist eine Revolution. Warum?

Zwei Gründe:

- Unvorstellbar gute Ergebnisse
 - Auf Basis von herkömmlichen Fotos
- Völlig andere Herangehensweise
 - Menschengemachte Modelle gegenüber aus Daten gelernte Modelle
 - Mehr Ausprobieren als Nachdenken
 - Mehr Datenaufbereitung als Algorithmenentwicklung

Erfolgsfaktoren von Deep Learning

- **Moore's law** → exponentielle Entwicklung der verfügbaren Rechenleistung
- **Fotos** → Durch den Einbau von Kameras in Smartphones ist die Menge an verfügbaren Fotos explodiert
- **Open source** → Facebook, Google, IBM, Microsoft sind sehr offen
 - Viele Werkzeuge verfügbar (TensorFlow, PyTorch, Keras usw.)
- **Open education** → [MIT OpenCourseWare](#) hat über 5M Abonnenten auf YouTube
 - [Gilbert Strang, Linear Algebra lectures](#) (fast) alle > 100K Ansichten
 - A. Karpathy, [Stanford CS231n](#) alle Vorlesungen > 30K Ansichten

Inhalte der Vorlesung

Themenüberblick 3D Computer Vision SoSe 2025 (Stand Februar 2025)

1. Was ist Computer Vision?
2. Wie entstehen Kamerabilder?
 - a. Bildentstehung, Kameras, Lochkameramodell
 - b. Perspektivische Projektion
 - c. Kameramodell, intr. und extr. Parameter
 - d. Bildverzerrung
3. Projektive Geometrie
4. Kamerakalibration
5. Perspektivische Projektion und Tiefeninformationen aus einem Bild
6. Einfaches Stereo, Disparität, Korrelationsfunktion, lokale und globale Stereoalgorithmen
7. Triangulation, Strukturierte Beleuchtung
8. Multi-view Stereo, Tiefeninformation aus vielen Bildern mit kalibrierten Kameras
9. Neural Radiance Fields (NeRF)
10. Epipolargeometrie, Fundamentalmatrix
11. Feature Matching, SIFT, RANSAC
12. Structure From Motion, Bundle Adjustment
13. 3D Datenstrukturen, Punktwolken
14. Time-of-Flight Kameras

Lernziele

- Wie kommt man von Bildern zu 3D Modellen?
 - Wie entstehen Kamerabilder? Was passiert geometrisch?
 - Welche Informationen stecken drin? Welche kann man direkt auslesen?
 - Welche Annahmen müssen getroffen werden?
 - Welche Modelle werden gebildet?
 - Welche technischen Lösungen gibt es noch?
 - Inwiefern kann Deep Learning helfen?

Bildquellen

- [i:colo07] - [source](#) [license](#)
- [i:colo10] – [source](#)
- [i:marr82] - Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. W. H. Freeman, San Francisco.
- [i:szeliski20] - Richard Szeliski - Computer Vision: Algorithms and Applications (September 20, 2020 draft)
- [i:hays18] – [source](#) - Project 2: Local Feature Matching, CS 6476: Computer Vision, James Hays, Samarth Brahmbhatt, and John Lambert,
- [i:crimini99] – [source](#) - Criminisi, A., Reid, I. & Zisserman, A. Single View Metrology. *International Journal of Computer Vision* 40, 123–148 (2000)
- [i:puglia20] - <https://arxiv.org/pdf/2001.04552v1.pdf> Deep Learning Stereo Vision at the edge, Luca Puglia and Cormac Brick, Intel R&D
- [i:jiang25]: Hualie Jiang, Zhiqiang Lou, Laiyan Ding, Rui Xu, Minglang Tan, Wenjie Jiang, and Rui Huang. [DEFOM-Stereo: Depth foundation model based stereo matching](#) (RVC version). arXiv:2501.09466, 2025.

Videoquellen

- [v:agarwal09] – [source](#) - Building Rome in a Day, Sameer Agarwal, Noah Snavely, Ian Simon, Steven M. Seitz and Richard Szeliski, International Conference on Computer Vision, 2009, Kyoto, Japan.
- [v:facebook20] [source](#) - Powered by AI: Turning any 2D photo into 3D using convolutional neural nets
- [v:WayneINR24] - <https://twitter.com/WayneINR/status/1750945037863551247> Exploring the capabilities of DepthAnything – see [gallery](#)
- [v:cherabier18] [source](#) - Learning Priors for Semantic 3D Reconstruction, Ian Cherabier, Johannes L. Schönberger, Martin R. Oswald, Marc Pollefeys, Andreas Geiger
- [v:mildenhall20] [source](#) - NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis, Ben Mildenhall and Pratul P. Srinivasan and Matthew Tancik and Jonathan T. Barron and Ravi Ramamoorthi and Ren Ng, 2020, ECCV
- [v:müller22] [source](#) - Instant Neural Graphics Primitives with a Multiresolution Hash Encoding, Thomas Müller, Alex Evans, Christoph Schied, Alexander Keller (all from NVIDIA), ACM Transactions on Graphics (SIGGRAPH 2022, Best paper award)

Textquellen

- Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. W. H. Freeman, San Francisco.
- Richard Szeliski - Computer Vision: Algorithms and Applications (September 20, 2020 draft)
- <https://towardsdatascience.com/everything-you-ever-wanted-to-know-about-computer-vision-heres-a-look-why-it-s-so-awesome-e8a58dfb641e> Ilija Mihajlovic
- <https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf> Krizhevsky2012
- <http://www.lyndonhill.com/opinion-cvlegends.html>