HTWK

IM Fakultät Informatik und Medien

Gesture2Robot Season 2
–
Objekterkennung über die Hololens 2 zur Teach-in Programmierung

Teammitglieder

Jan Philipp Seeland

Benedikt Beigang

Elmar Kresse

Christoph Walther

Gliederung

Aufgabenstellung

•

2D Segmentierung

•

3D Punktwolken Segmentierung

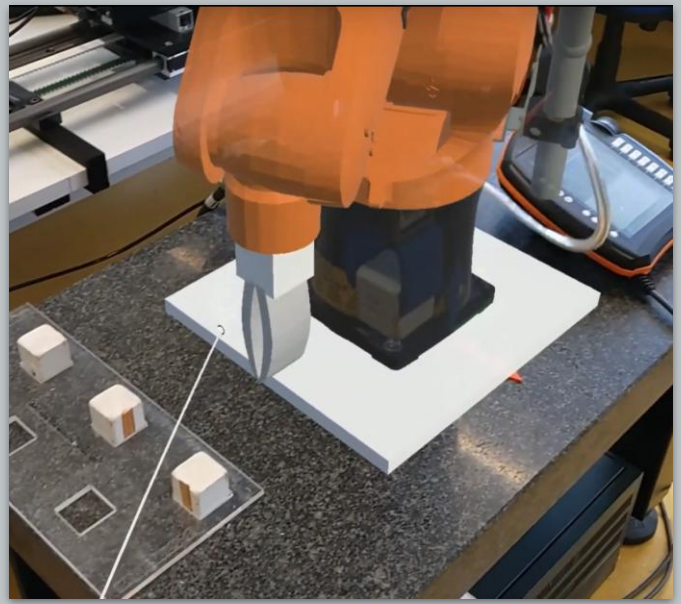
•

Fazit und Zukunft

Benedikt Beigang, Jan Philipp Seeland, Elmar Kresse, Christoph Walther
Hochschule für Technik, Wirtschaft und Kultur Leipzig

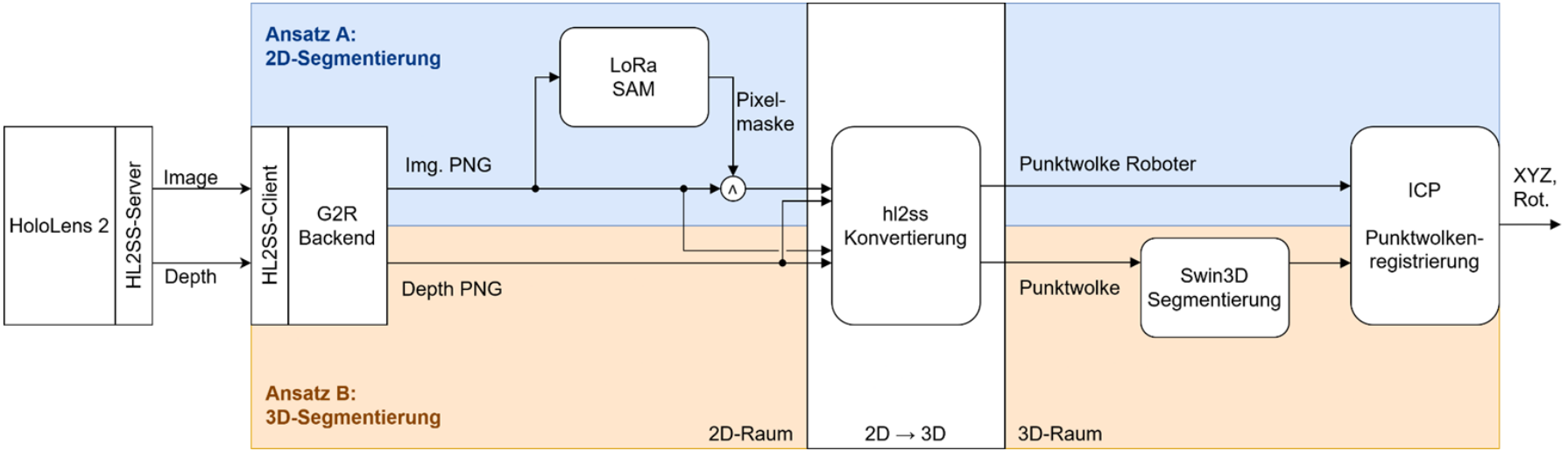
Aufgabenstellung

Automatisierte Platzierung des Roboters



Benedikt Beigang, Jan Philipp Seeland, Elmar Kresse, Christoph Walther
Hochschule für Technik, Wirtschaft und Kultur Leipzig

Zwei parallele Lösungsansätze



Benedikt Beigang, Jan Philipp Seeland, Elmar Kresse, Christoph Walther
Hochschule für Technik, Wirtschaft und Kultur Leipzig

Datenabfrage, Aufarbeitung und Verarbeitung

- visible Light Cameras - Four grayscale cameras 640x480@30FPS Stream
- RM Depth Long-throw - 32bit PNG 320x288@5FPS
- RGB camera (PV) - 1920x1080@30FPS Stream
- Spatial input - Head pose, gaze ray (origin and direction), hand tracking

Benedikt Beigang, Jan Philipp Seeland, Elmar Kresse, Christoph Walther
Hochschule für Technik, Wirtschaft und Kultur Leipzig

Datenabfrage, Aufarbeitung und Verarbeitung

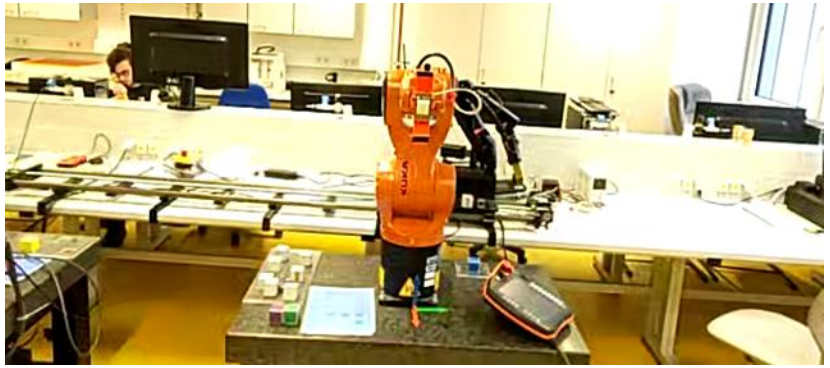
- Forschungsprojekt am Stevens Institute of Technology, gesponsert von der Defense Advanced Research Projects Agency (DARPA)
- Auswahlgrund: Bidirektionale Kommunikation + Unity-Integration auf Brille
- Motivation: Bereitstellung von Hololens-Rohdaten für rechenintensive Applikationen
- Kein Zugriff auf Tiefendaten über Microsoft-Schnittstellen
  hier: Research Mode

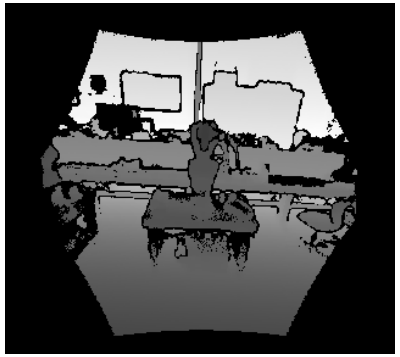
Field	Type	Length in bytes
Focal length	1x2 float	8
Principal point	1x2 float	8
Radial distortion	1x3 float	12
Tangential distortion	1x2 float	8
Projection	4x4 float	64

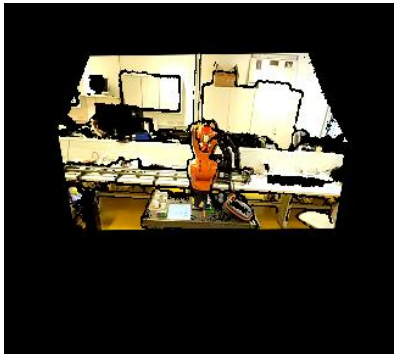
Table 15. PV calibration data.

Datenabfrage, Aufarbeitung und Verarbeitung

- RGBD Bildausrichtung
  der Aufnahmen Zuschnitte aus beiden Sensoraufnahmen
- Farbe kann mithilfe des
  research Modus und der
  Kameras für sichtbares Licht (oben)
  erzeugt werden
- alternativ Erzeugung mithilfe der Foto/Video-Streams unten
- Video + extrinsische Kalibrierungsdaten
- Autofokus deaktiviert, damit Daten valide bleiben







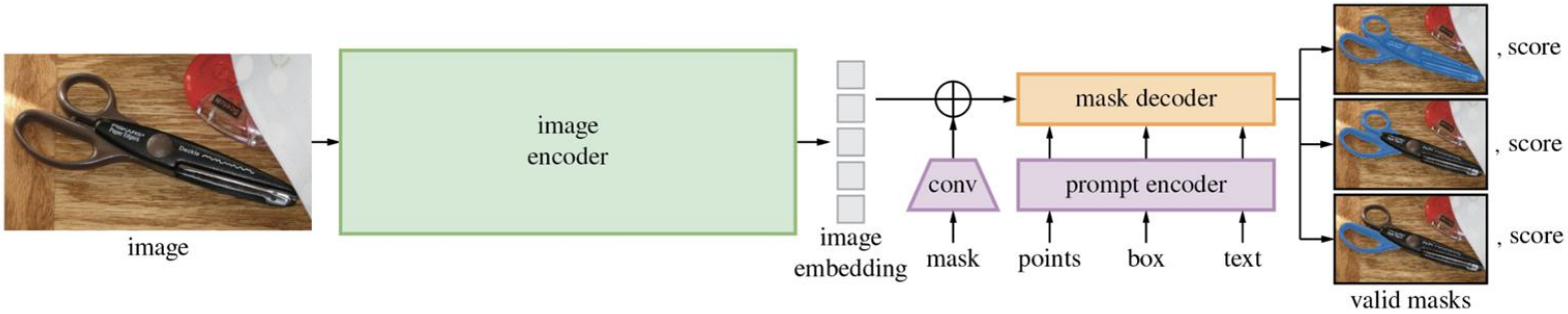
Benedikt Beigang, Jan Philipp Seeland, Elmar Kresse, Christoph Walther
Hochschule für Technik, Wirtschaft und Kultur Leipzig

2D-Segmentierungs Ansatz

SAM mit LoRa

Benedikt Beigang, Jan Philipp Seeland, Elmar Kresse, Christoph Walther
Hochschule für Technik, Wirtschaft und Kultur Leipzig

SAM - Segment Anything Model

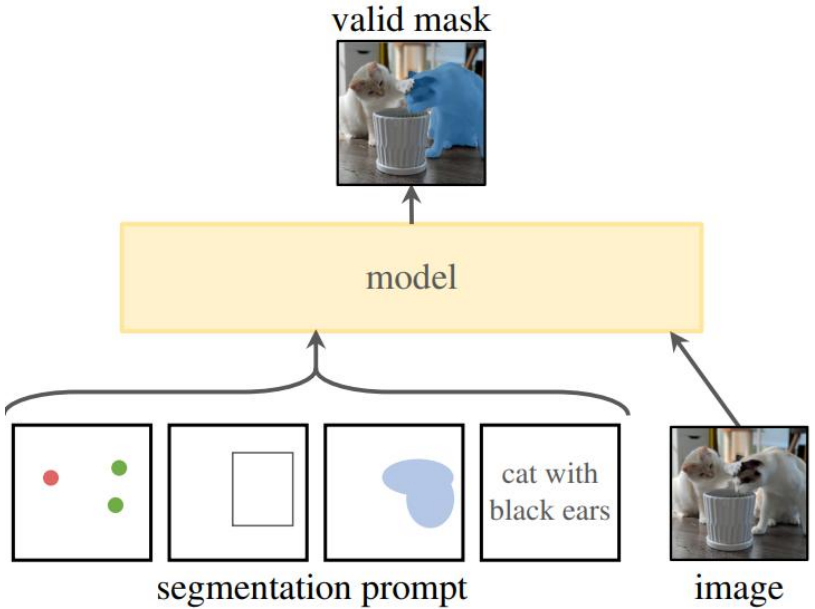


[a]

- trained on
  - a dataset of 11 million images and
  - 1.1 billion high quality object masks
- strong zero-shot performance

Benedikt Beigang, Jan Philipp Seeland, Elmar Kresse, Christoph Walther
Hochschule für Technik, Wirtschaft und Kultur Leipzig

SAM - Segment Anything Model



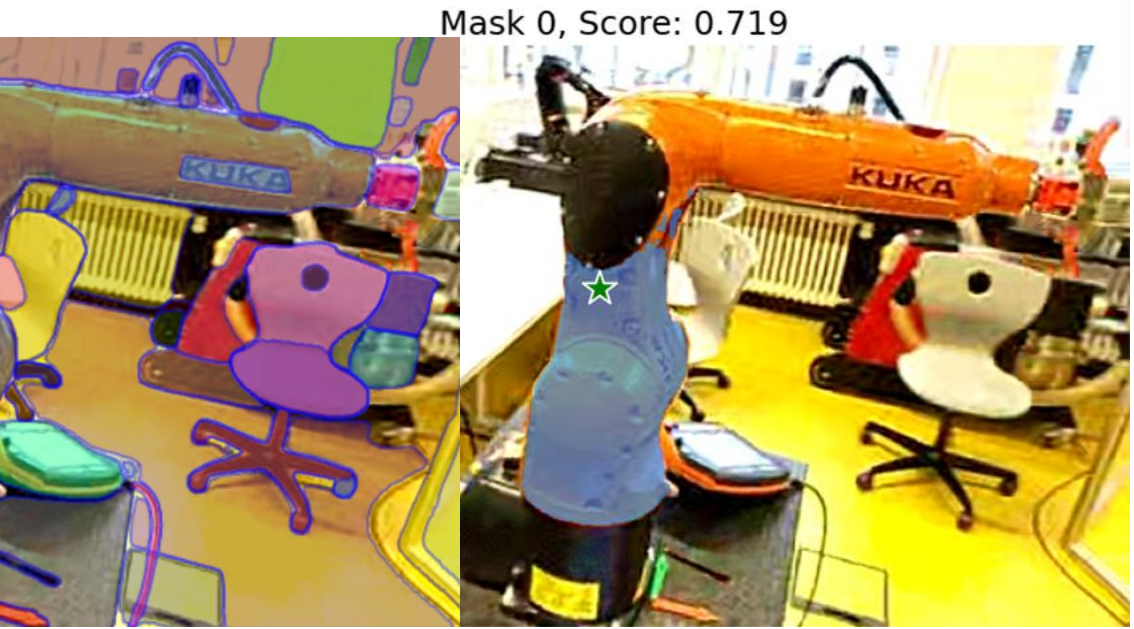
[a]

Promptable Segmentation

- Punkte
- Bounding-Boxen
- Masken
- Text Prompt (noch nicht offiziell veröffentlicht) [g]

Benedikt Beigang, Jan Philipp Seeland, Elmar Kresse, Christoph Walther
Hochschule für Technik, Wirtschaft und Kultur Leipzig

SAM - Sieht doch gut aus?



Probleme:

- Erkennung des Konzeptes Roboter
- Geschwindigkeit: ~3-5s pro Bild
- Größe (2,4GB!)
- Finetuning nicht offiziell unterstützt

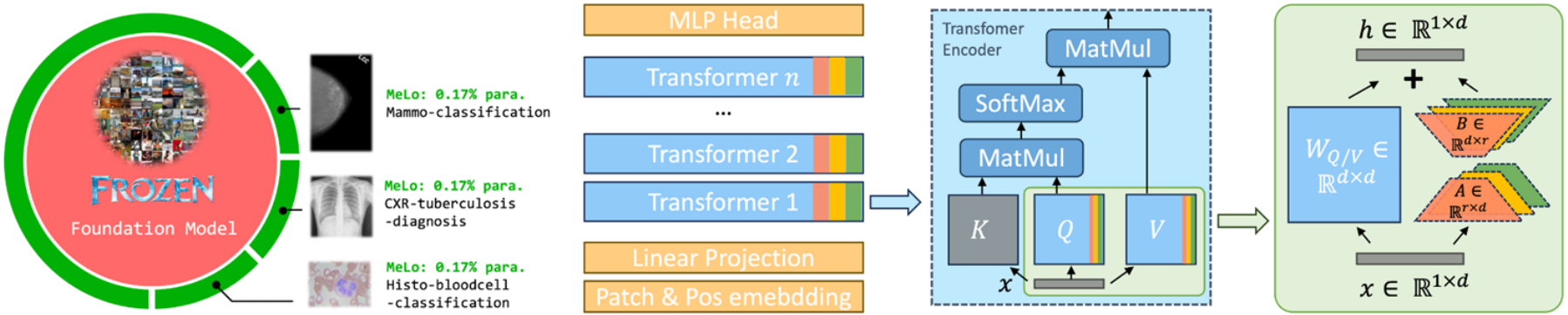
Benedikt Beigang, Jan Philipp Seeland, Elmar Kresse, Christoph Walther
Hochschule für Technik, Wirtschaft und Kultur Leipzig

Finetuning Ansätze

- **Segment Anything Model with 🤗Transformers** [h]
  - sehr aktuell
  - Finetuning für Zellen und Organe
  - Problem: Training verschlechtert Ergebnisse (vermutlich zu kleiner Datensatz)
- **Segment Anything in KerasCV**
  - Maskenerstellung durch Text-Prompt [i]
  - kompliziertes SetUp (jax-backend + GroundingDINO + SAM)
  - nicht für Finetuning ausgelegt
- **LoRA-SAM (Low-rank Adaptation)** [j]
  - Ausgelegt auf fine-tuning
  - schnell im Training und beim Abspielen
  - kleine Dateien

Benedikt Beigang, Jan Philipp Seeland, Elmar Kresse, Christoph Walther
Hochschule für Technik, Wirtschaft und Kultur Leipzig

LoRA - Low-Rank Adaptation



(a) Overview (b) Method

[k]

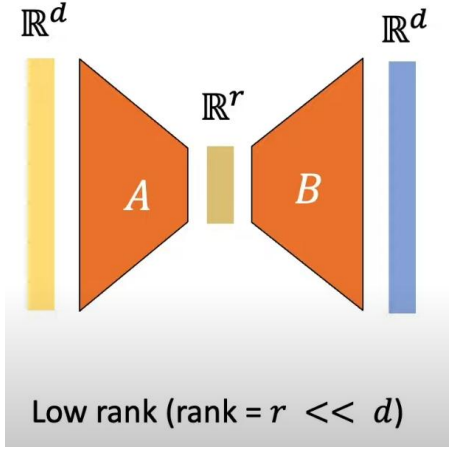
- Parameter effiziente fine-tuning Technik
  - nur wenig Anpassung der Gewichte auf vor-trainierten Modellen nötig
- Eingabe sind “low-rank decomposition” Matrizen
- erstmals in LLMs verwendet -> Übertragung auf Vision Transformer

Benedikt Beigang, Jan Philipp Seeland, Elmar Kresse, Christoph Walther
Hochschule für Technik, Wirtschaft und Kultur Leipzig

Vorteile von LoRA

- SAM => 4,4s
- SAM mit LoRA => 0.7s
- Kleinere Datensätze
- Finetuning über kleine Ränge

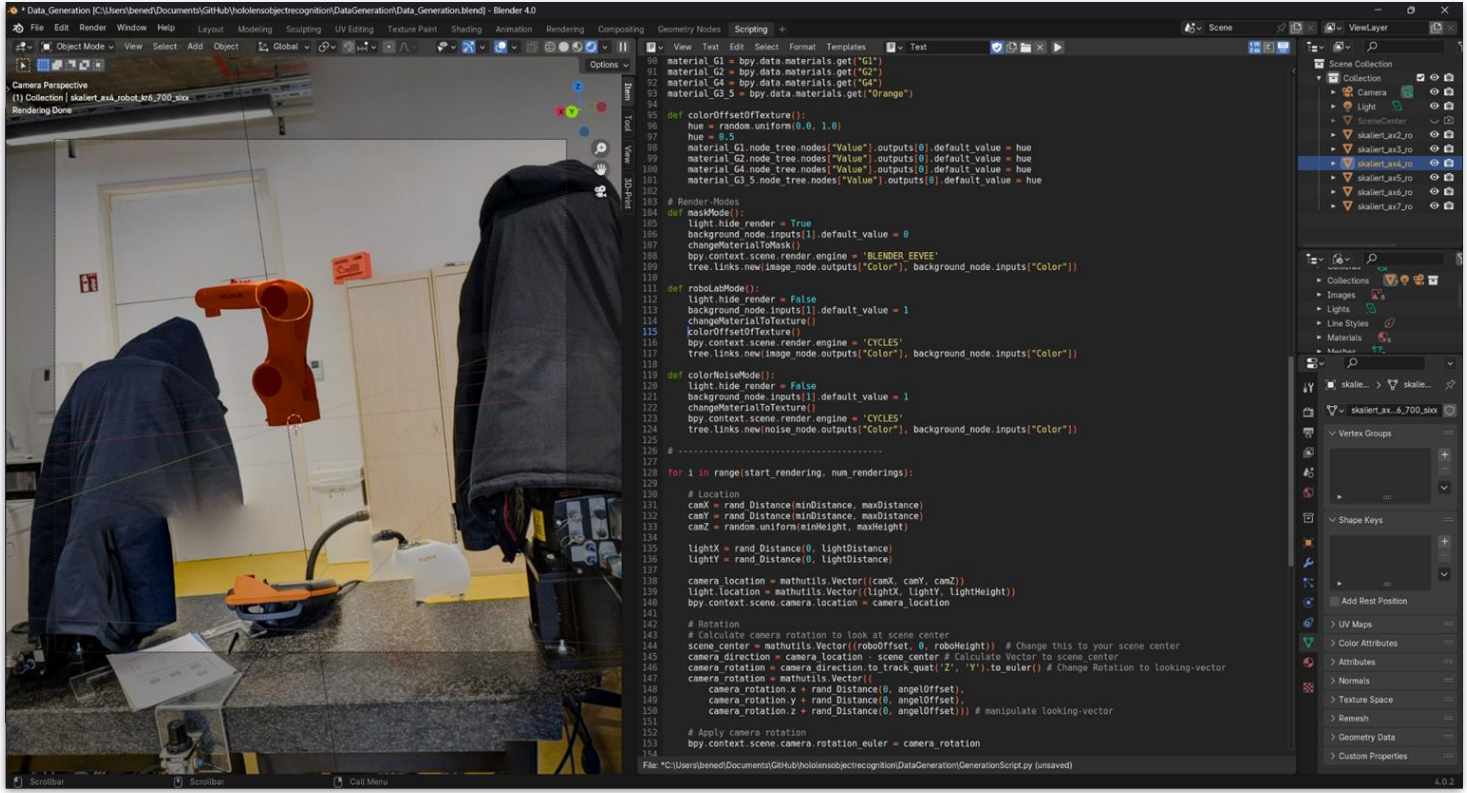
“[LoRA] freezes the original weights of visual foundation models while adding small low-rank plug-ins that can achieve remarkable results using only a small fraction of trainable weights.”



[I]

Benedikt Beigang, Jan Philipp Seeland, Elmar Kresse, Christoph Walther
Hochschule für Technik, Wirtschaft und Kultur Leipzig

Erstellung des Datensatzes in Blender

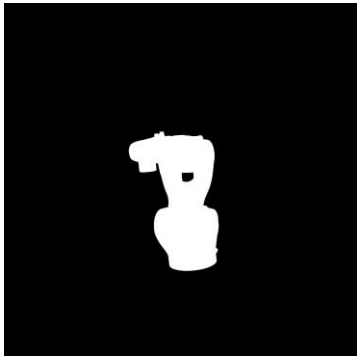


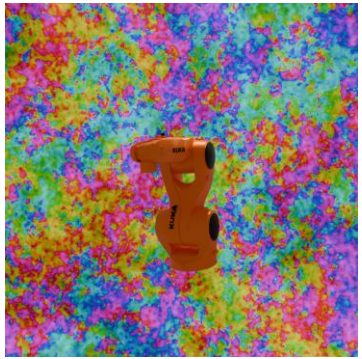
Benedikt Beigang, Jan Philipp Seeland, Elmar Kresse, Christoph Walther
Hochschule für Technik, Wirtschaft und Kultur Leipzig

I-ITWK

Variable Parameter

- Kameraposition
- Kamerawinkel
- Farbe des Roboters
- Beleuchtungswinkel
- Hintergrund



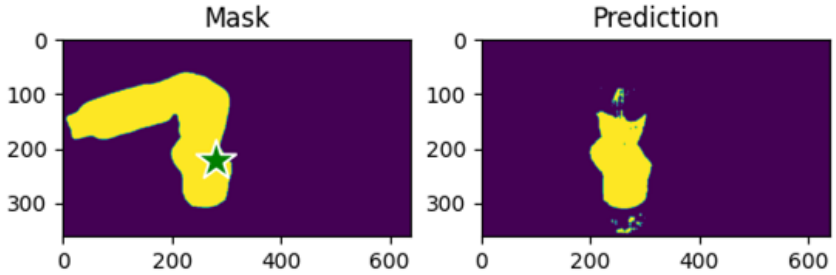


+

Benedikt Beigang, Jan Philipp Seeland, Elmar Kresse, Christoph Walther
Hochschule für Technik, Wirtschaft und Kultur Leipzig

Erstellung des Evaluations-Skripts

- 30 diverse Bildaufnahmen von der HoloLens 2
- Erstellung der Groundtruth-Masken per Hand
- Berechnung der falsch-negativen und falsch-positiven Pixel zur Berechnung der Accuracy und des Overshoots
- Score: 1 - ((fp+fn) / alle Pixel)



Mask-Accuracy: 40.07 %
Background-Overshoot: 0.32 %
Final Scoring: 70.0 %

Benedikt Beigang, Jan Philipp Seeland, Elmar Kresse, Christoph Walther
Hochschule für Technik, Wirtschaft und Kultur Leipzig

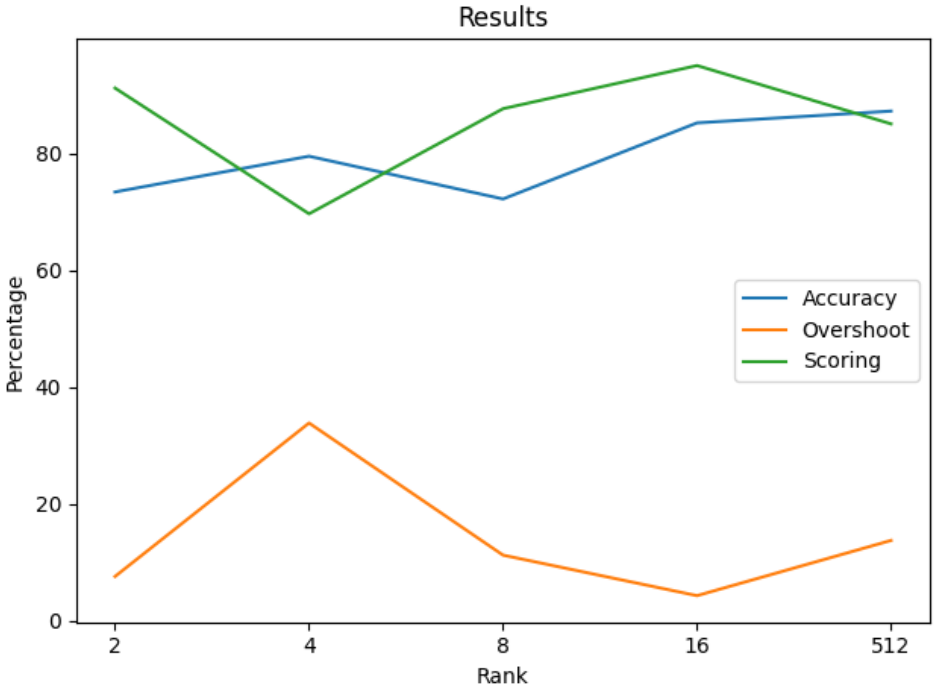
Welchen Rank?

- 500 “realistisch” gerenderte Images
- 5 Epochen
- => Rank 2 und 16 am besten

Rank 2

Rank 4

Rank 512



Benedikt Beigang, Jan Philipp Seeland, Elmar Kresse, Christoph Walther
Hochschule für Technik, Wirtschaft und Kultur Leipzig

Welcher Datensatz?

- Rank 2
- 5 Epochen
- “realistisch” gerenderte Images performen am besten

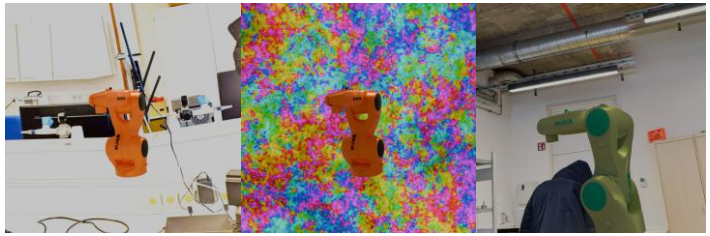
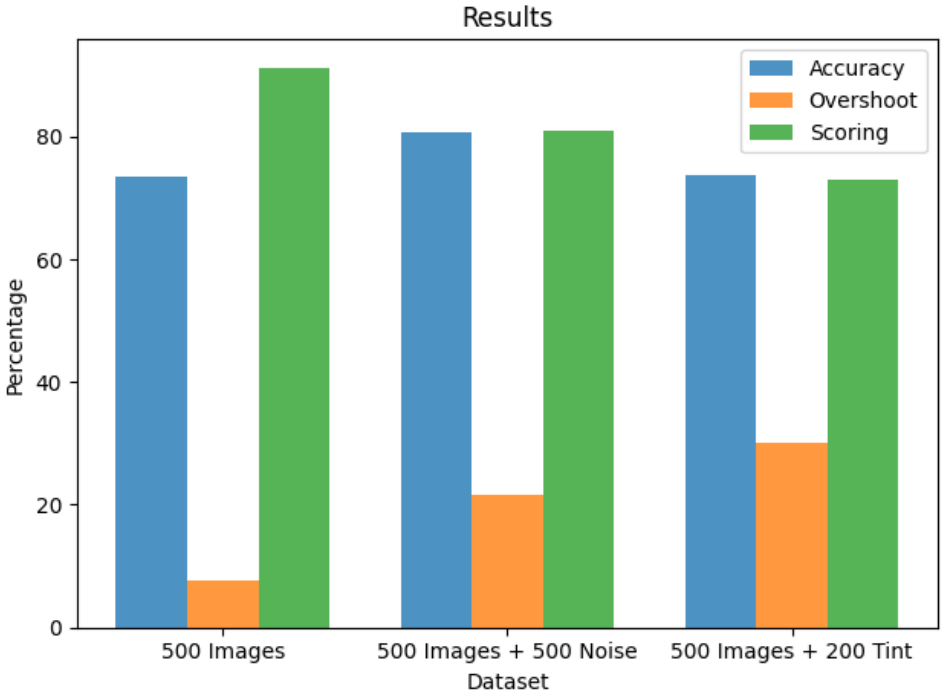


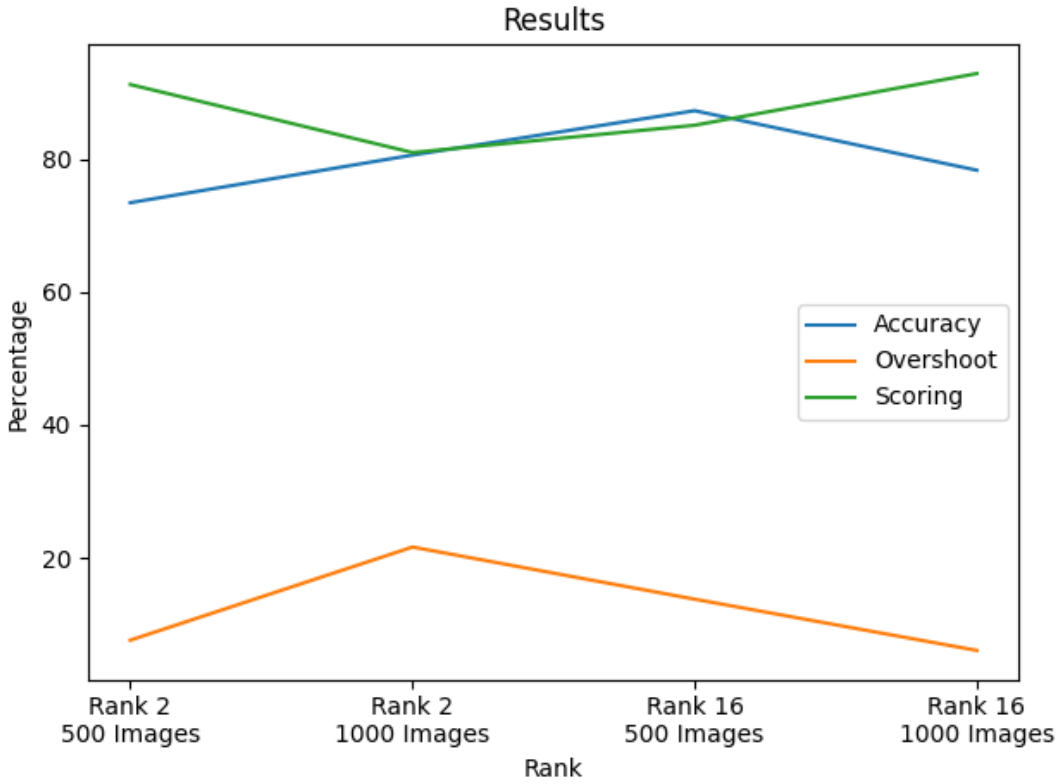
Image Noise Tint



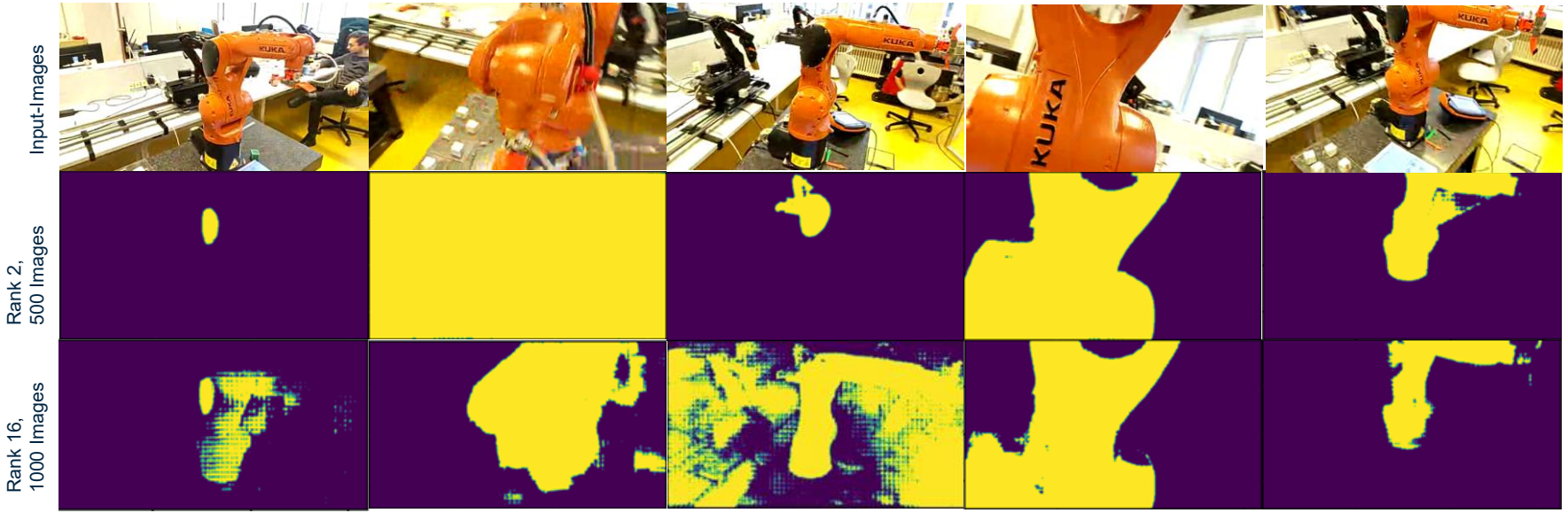
Benedikt Beigang, Jan Philipp Seeland, Elmar Kresse, Christoph Walther
Hochschule für Technik, Wirtschaft und Kultur Leipzig

I-ITWK

Die besten Ergebnisse



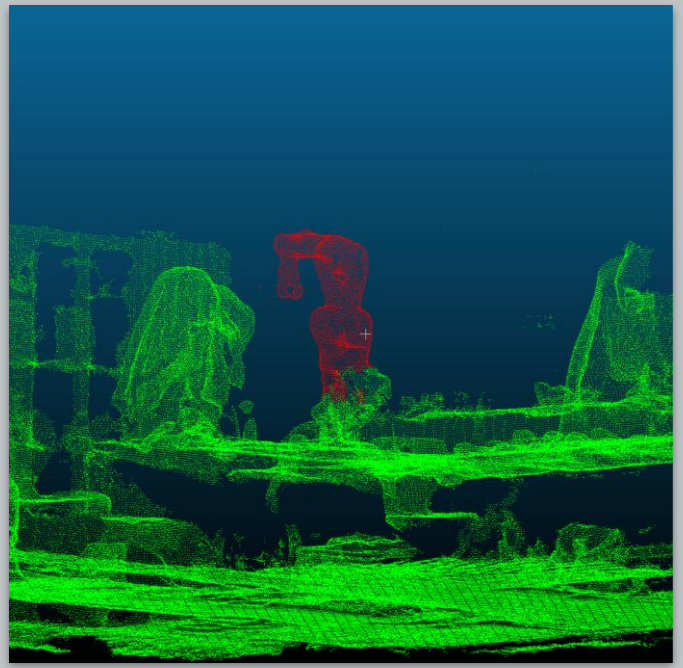
Erfolge und Probleme



Benedikt Beigang, Jan Philipp Seeland, Elmar Kresse, Christoph Walther
Hochschule für Technik, Wirtschaft und Kultur Leipzig

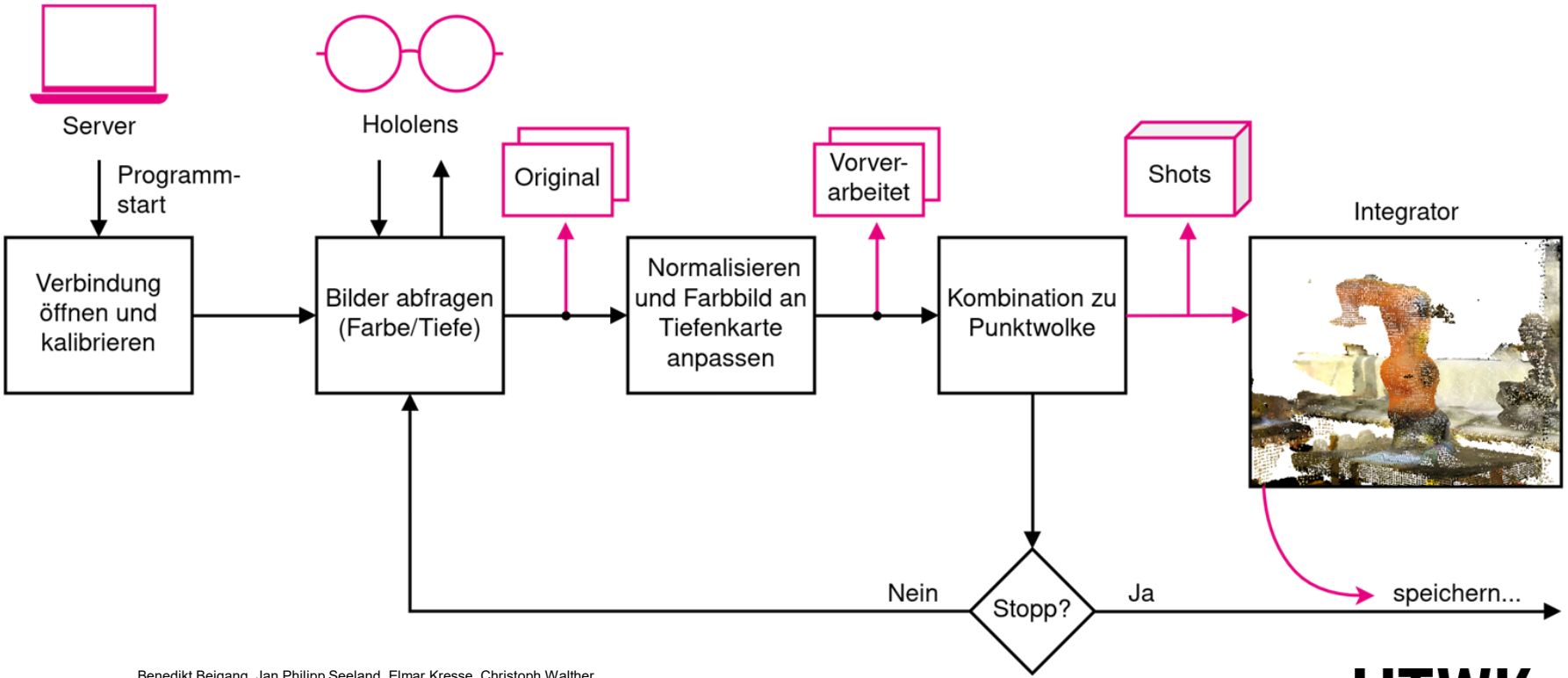
3D RGB Punktwolken-Segmentierungsansatz

Swin3D
shift-window based transformer



Benedikt Beigang, Jan Philipp Seeland, Elmar Kresse, Christoph Walther
Hochschule für Technik, Wirtschaft und Kultur Leipzig

PV Integration Pipeline



Benedikt Beigang, Jan Philipp Seeland, Elmar Kresse, Christoph Walther
Hochschule für Technik, Wirtschaft und Kultur Leipzig

Stanford 3D Indoor Scene Dataset (S3DIS)

- 6 Innenraumbereiche (“Areas”)
- Rund 70.000 Bilder, 700 Mio. Punkte
- 13 Objektklassen

Aufbau:

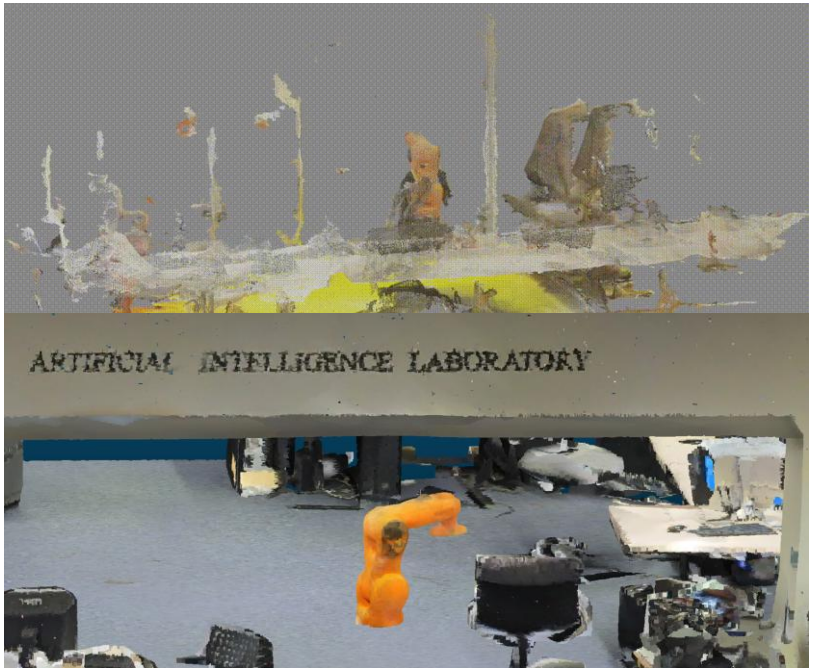
- Area_1
  - conferenceRoom_1
    - Annotations:
      beam, board, chair, …
  - conferenceRoom_2
  - …
- Area_2…6

Benedikt Beigang, Jan Philipp Seeland, Elmar Kresse, Christoph Walther
Hochschule für Technik, Wirtschaft und Kultur Leipzig

Generierung eigener Datensatz

HTWK Robo 3D Dataset (HR3D)

- Shots_00
  - scan_00…24: **Roboterscans**
    - Annotations: robo, other
  - scan_50…51: **Raumscans**
- Shots_01…24
  - pcd_robo_xx_01…343
    Einzelperspektiven
- Shots_50…51
  - pcd_robo_5x_01…180
    Einzelperspektiven Raum
- Shots_666: S3DIS Area 1



Labelling der Daten (Ausschneiden des Roboters)

1. Manuelles Labeling

- CloudCompare:
  2D Schnitt mehrerer Blickwinkel
- Für alle 26 integrierten Punktwolken

Benedikt Beigang, Jan Philipp Seeland, Elmar Kresse, Christoph Walther
Hochschule für Technik, Wirtschaft und Kultur Leipzig

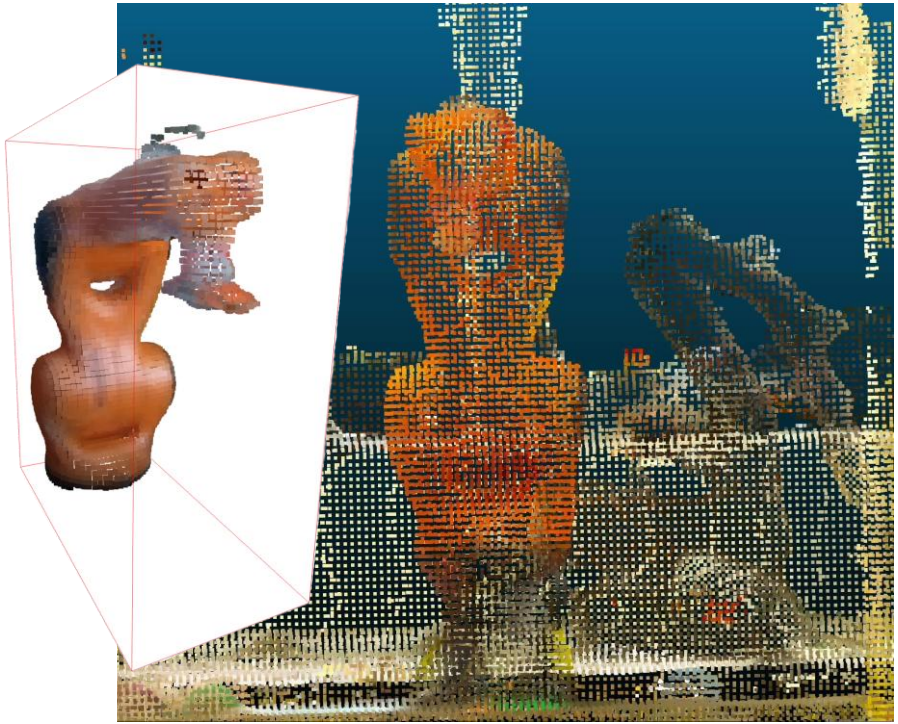
Labelling der Daten (Ausschneiden des Roboters)

1. Manuelles Labeling

- CloudCompare:
  2D Schnitt mehrerer Blickwinkel
- Für alle 26 integrierten Punktwolken

2. Automatisches Labeling

- Bounding-Box um Roboter erstellen
- Alle Teilaufnahmen im Bereich segmentieren



Benedikt Beigang, Jan Philipp Seeland, Elmar Kresse, Christoph Walther
Hochschule für Technik, Wirtschaft und Kultur Leipzig

Datenaugmentierung

- Variierung der Scans
  - Verschiedene Roboterpositionen, Lichtsituationen, Scanqualität

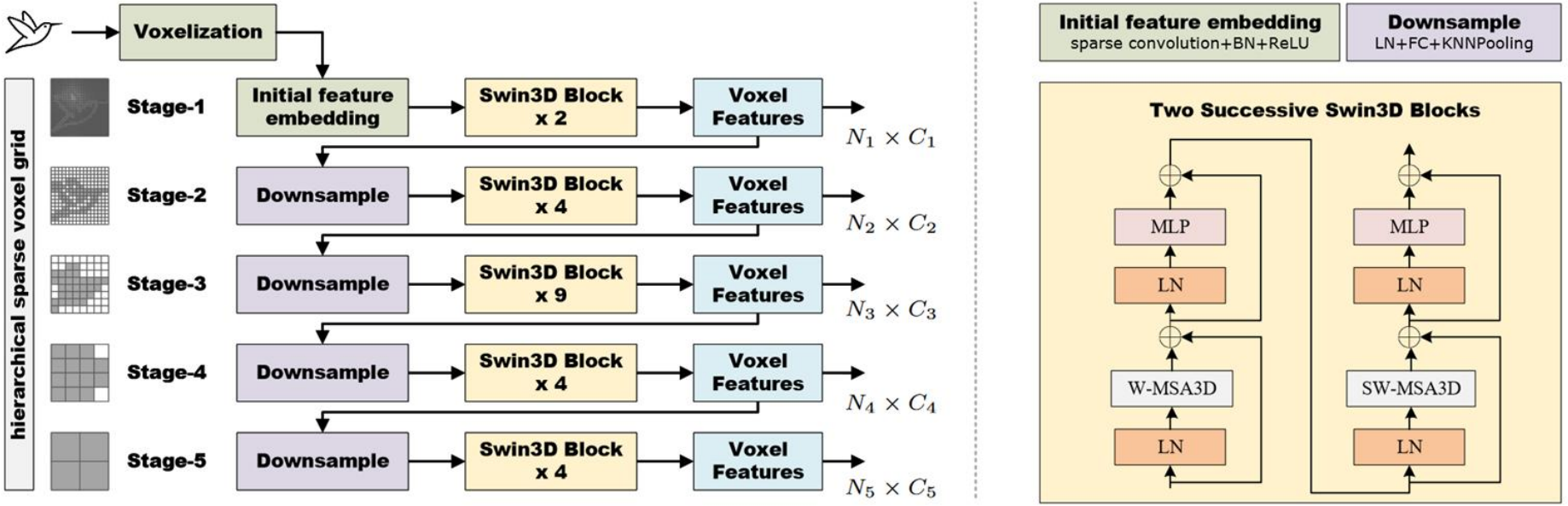
In Swin3D

- Etablierte Datenaugmentation as Papern CAGroup3D und FCAF3D
- Normalisierung und Anwendung verschiedener Transformationen
- Weitere Transformationen vorhanden (z. B. Farbe, etc) → Zukunft

```
train_transform = transform.Compose(
    [
        transform.RandomRotate(along_z=args.get
        ("rotate_along_z", True)),
        transform.RandomScale(
            scale_low=args.get("scale_low", 0.8),
            scale_high=args.get("scale_high", 1.2),
        ),
        transform.RandomJitter(sigma=jitter_sigma,
        clip=jitter_clip),
        transform.RandomDropColor(
            color_augment=args.get("color_augment", 0.0)
        ),
    ]
)
```


Benedikt Beigang, Jan Philipp Seeland, Elmar Kresse, Christoph Walther
Hochschule für Technik, Wirtschaft und Kultur Leipzig

Swin3D



[e] Swin3D: A Pretrained Transformer Backbone for 3D Indoor Scene Understanding

Benedikt Beigang, Jan Philipp Seeland, Elmar Kresse, Christoph Walther
Hochschule für Technik, Wirtschaft und Kultur Leipzig

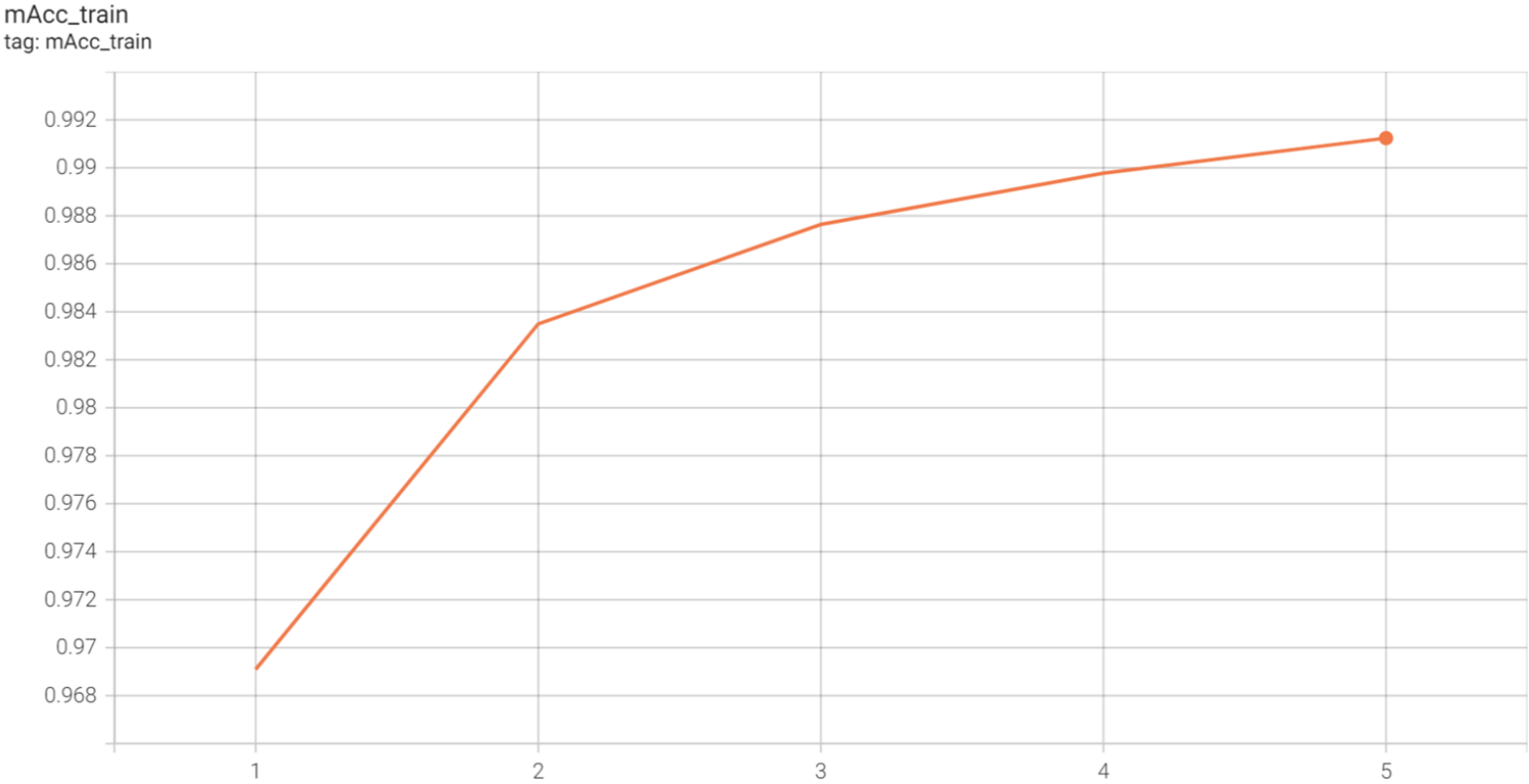
Training

- Ubuntu 22.04 mit pytorch 2.2.1 und CUDA 12.1 nativ auf neuem extra Rechner installiert
- Minkowski Engine selbst kompiliert 40GB Arbeitsspeicher benötigt
- RTX 3070 (eigener Datensatz RoboHTWKLab) mit swin3D_RGB_S gestartet

```
[01/20 17:59:05 main-logger]: Val result: mIoU/mAcc/allAcc 0.8792/0.9597/0.9888.
[01/20 17:59:05 main-logger]: Class_0 Result: iou/accuracy 0.9884/0.9914.
[01/20 17:59:05 main-logger]: Class_1 Result: iou/accuracy 0.7701/0.9280.
[01/20 17:59:05 main-logger]: <<<<<<<<<<<<<<<< End Evaluation <<<<<<<<<<<<<<<<
[01/20 17:59:05 main-logger]: Val result Point: mIoU/mAcc/allAcc 0.8832/0.9610/0.9691.
[01/20 17:59:05 main-logger]: Class_0 Result: iou/accuracy 0.9647/0.9720.
[01/20 17:59:05 main-logger]: Class_1 Result: iou/accuracy 0.8016/0.9500.
[01/20 17:59:05 main-logger]: <<<<<<<<<<<<<<<< End Evaluation <<<<<<<<<<<<<<<<
[01/20 17:59:05 main-logger]: Saving checkpoint to: runs/s3dis_Swin3D_RGB_S_123/model/model_last.pth
[01/20 17:59:13 main-logger]: lr: [6e-05, 6e-06]
[01/20 17:59:18 main-logger]: Epoch: [2/10][10/5073] Data 0.000 (0.065) Batch 0.259 (0.507) Remain 06:25:34 Loss 0.0152 Lr: [6e-05, 6e-06] Accuracy 0.9969.
[01/20 17:59:21 main-logger]: Epoch: [2/10][20/5073] Data 0.000 (0.033) Batch 0.324 (0.443) Remain 05:36:34 Loss 0.0089 Lr: [6e-05, 6e-06] Accuracy 0.9985.
[01/20 17:59:25 main-logger]: Epoch: [2/10][30/5073] Data 0.000 (0.022) Batch 0.317 (0.400) Remain 05:04:23 Loss 0.0119 Lr: [6e-05, 6e-06] Accuracy 0.9970.
[01/20 17:59:28 main-logger]: Epoch: [2/10][40/5073] Data 0.001 (0.017) Batch 0.284 (0.383) Remain 04:51:22 Loss 0.0110 Lr: [6e-05, 6e-06] Accuracy 0.9982.
[01/20 17:59:33 main-logger]: Epoch: [2/10][50/5073] Data 0.001 (0.039) Batch 0.284 (0.418) Remain 05:17:47 Loss 0.0070 Lr: [6e-05, 6e-06] Accuracy 0.9989.
[01/20 17:59:38 main-logger]: Epoch: [2/10][60/5073] Data 0.000 (0.032) Batch 0.257 (0.418) Remain 05:17:48 Loss 0.0354 Lr: [6e-05, 6e-06] Accuracy 0.9876.
[01/20 17:59:41 main-logger]: Epoch: [2/10][70/5073] Data 0.000 (0.028) Batch 0.298 (0.408) Remain 05:09:48 Loss 0.0105 Lr: [6e-05, 6e-06] Accuracy 0.9982.
[01/20 17:59:45 main-logger]: Epoch: [2/10][80/5073] Data 0.000 (0.024) Batch 0.260 (0.400) Remain 05:03:43 Loss 0.0177 Lr: [6e-05, 6e-06] Accuracy 0.9950.
[01/20 17:59:47 main-logger]: Epoch: [2/10][90/5073] Data 0.001 (0.022) Batch 0.251 (0.384) Remain 04:51:50 Loss 0.0086 Lr: [6e-05, 6e-06] Accuracy 0.9984.
```


Benedikt Beigang, Jan Philipp Seeland, Elmar Kresse, Christoph Walther
Hochschule für Technik, Wirtschaft und Kultur Leipzig

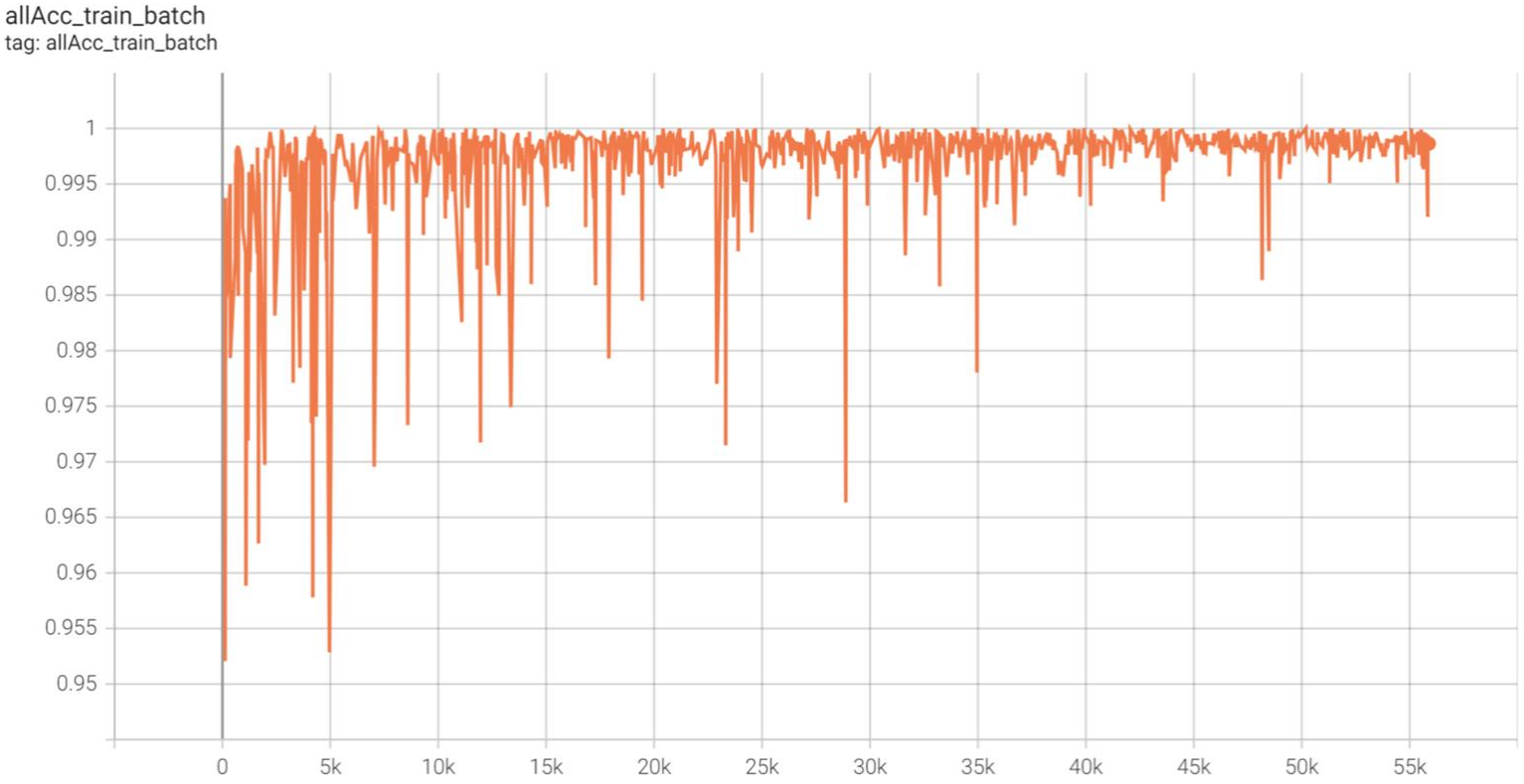
Auswertung Training und Evaluation



Benedikt Beigang, Jan Philipp Seeland, Elmar Kresse, Christoph Walther
Hochschule für Technik, Wirtschaft und Kultur Leipzig

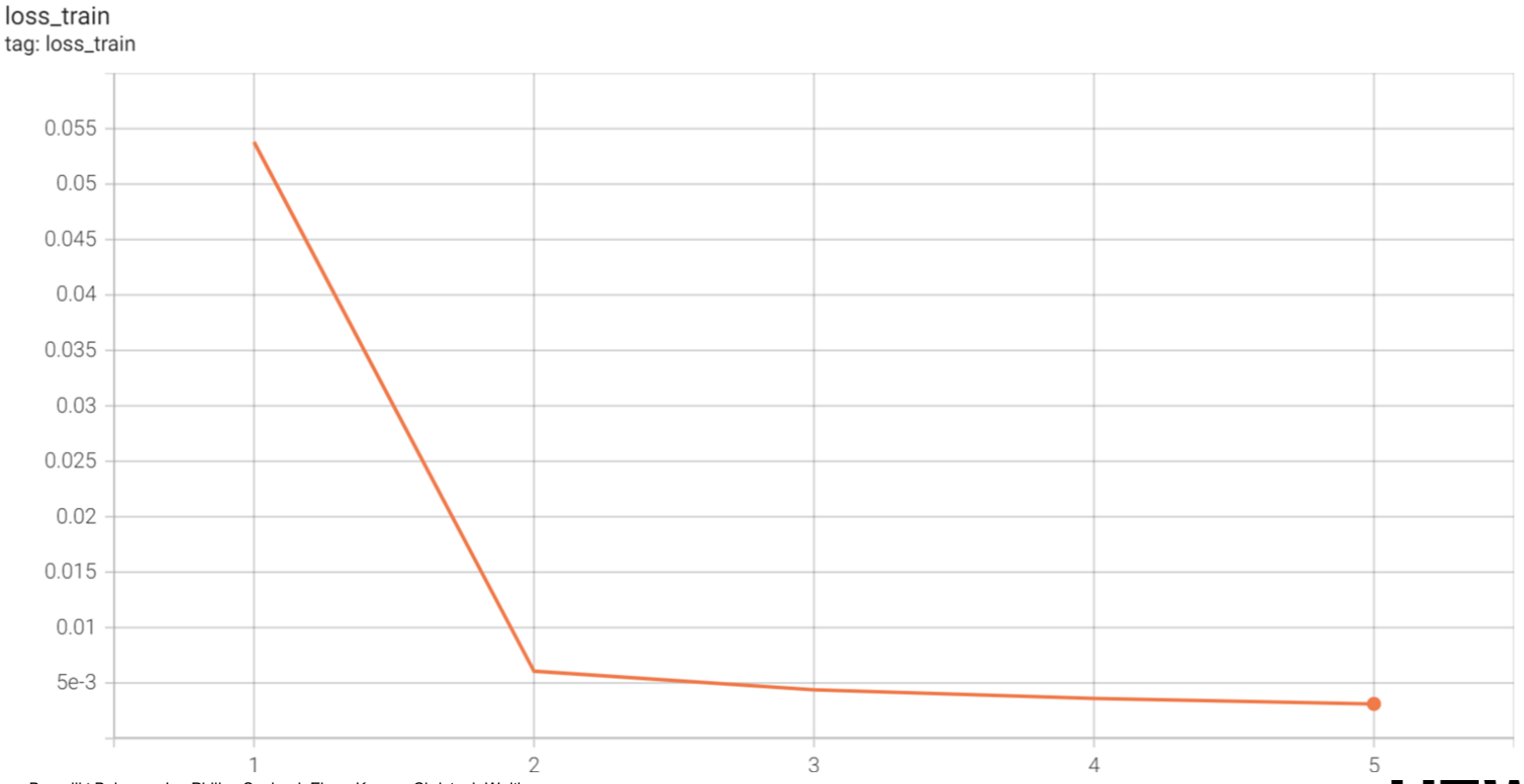
I-ITWK

Auswertung Training und Evaluation

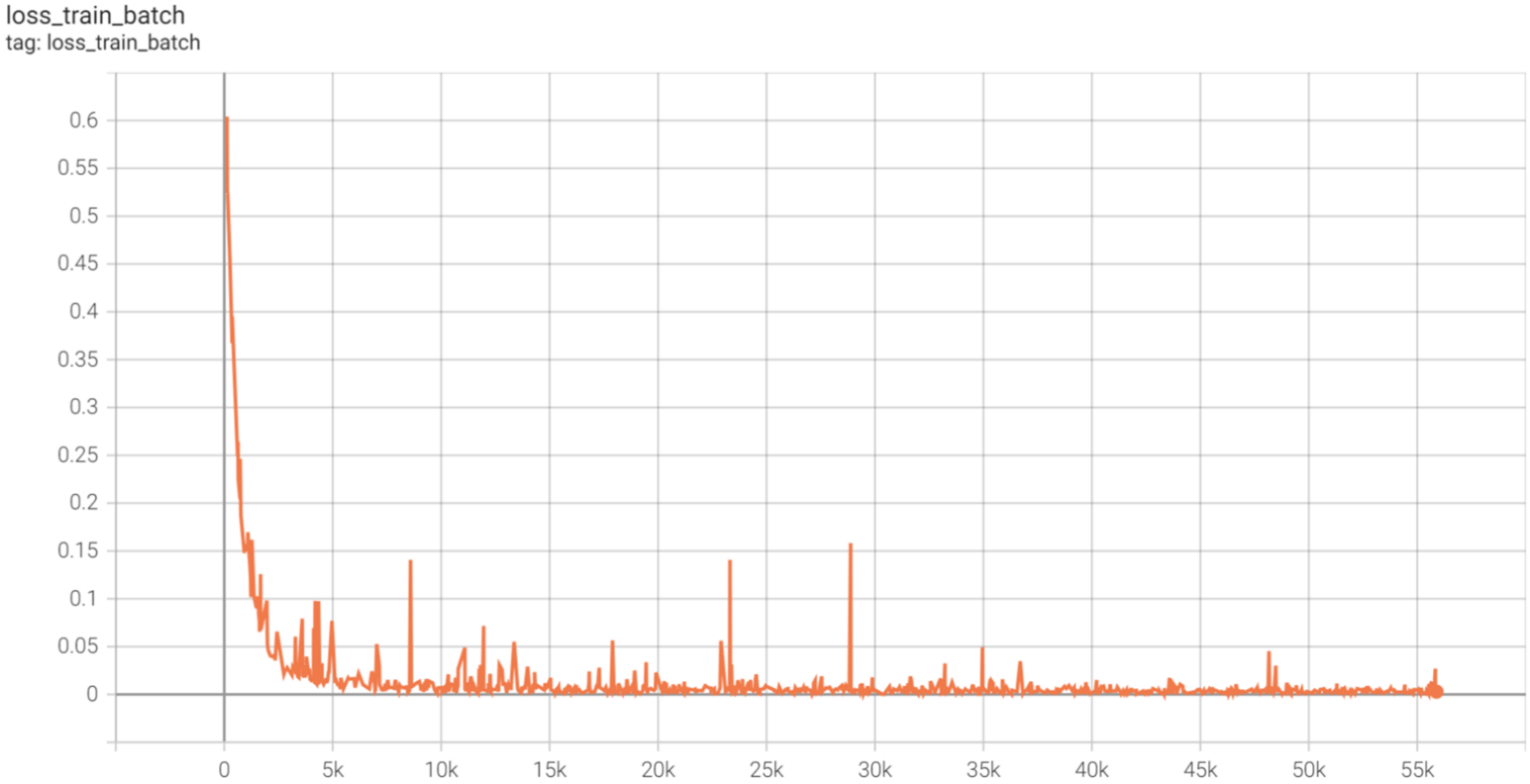


Benedikt Beigang, Jan Philipp Seeland, Elmar Kresse, Christoph Walther
Hochschule für Technik, Wirtschaft und Kultur Leipzig

Auswertung Training und Evaluation



Auswertung Training und Evaluation



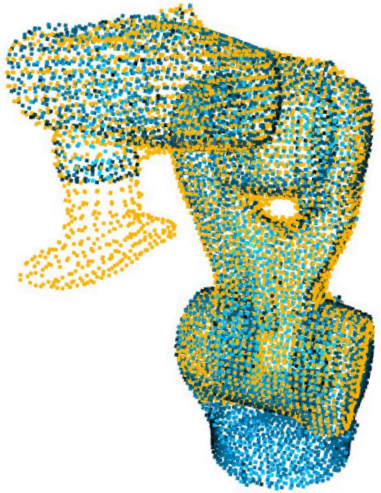
Training & Positionierung

Point Cloud Registration:

- Definition: Der Prozess der Ausrichtung mehrerer Punktwolken, um eine konsistente Referenz zu schaffen

- Anwendungen: 3D-Rekonstruktion, Objekterkennung, Umgebungs-Modellierung

- Herausforderungen: Rauschen, Ausreißer, nicht-überlappende Bereiche, große Datenmengen



Benedikt Beigang, Jan Philipp Seeland, Elmar Kresse, Christoph Walther
Hochschule für Technik, Wirtschaft und Kultur Leipzig

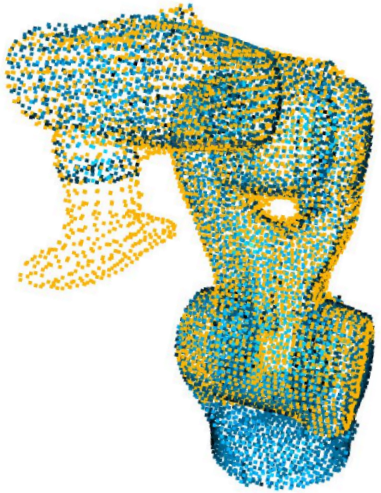
Training & Positionierung

ICP (Iterative Closest Point):

- Grundprinzip: iterativer Algorithmus zur Bestimmung der bestmöglichen Transformation zwischen zwei Punktwolken

- Ziel: Minimierung quadratischer Abstände zwischen den Punkten der beiden Wolken

- Anwendungen: Registrierung von 3D-Modellen in Robotik, Computer Vision, Augmented Reality



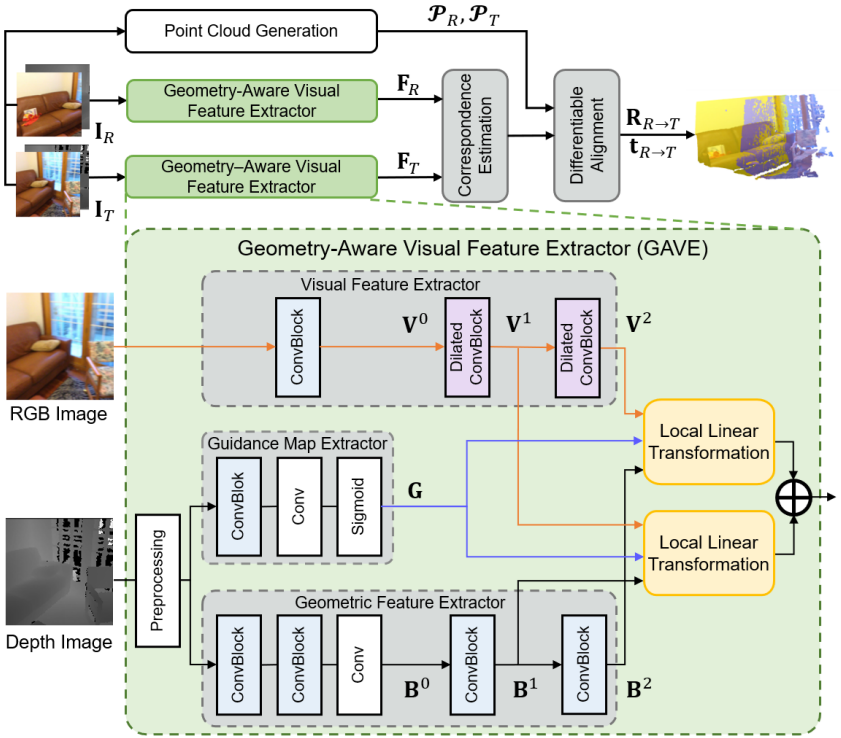
Benedikt Beigang, Jan Philipp Seeland, Elmar Kresse, Christoph Walther
Hochschule für Technik, Wirtschaft und Kultur Leipzig

Fazit und Zukunft

Benedikt Beigang, Jan Philipp Seeland, Elmar Kresse, Christoph Walther
Hochschule für Technik, Wirtschaft und Kultur Leipzig

Zukunft

- Weiteres LoRA/SAM-Training, um Score zu erhöhen
- evtl. Anpassen der Bounding Boxes für Punktwolken Generierung
- Mapping des segmentierten Roboters mit RGB Daten auf das Modell Objekt (ICP)
- Response mit relativer Position an die Hololens
  (SAM on Hololens bzw. externer Server)



Benedikt Beigang, Jan Philipp Seeland, Elmar Kresse, Christoph Walther
Hochschule für Technik, Wirtschaft und Kultur Leipzig

Lernerfahrungen

Jan Philipp	Benedikt
Tools: tensorflow 2 / Keras / Pytorch / Python	Tools: tensorflow 2 / Keras / Pytorch / Python / Blender
Paper: Segment Anything, LoRA	Paper: Segment Anything, LoRA
Einrichtung, Training und Abspielen von Neuronalen Netzen, Änderung von Layer und Parametern, Erstellung von Train/Test Datensätzen (für fine-tuning)	Erstellung von Train/Test Datensätzen (für fine-tuning) und Evaluierung, Training und Abspielen von Neuronalen Netzen

Benedikt Beigang, Jan Philipp Seeland, Elmar Kresse, Christoph Walther
Hochschule für Technik, Wirtschaft und Kultur Leipzig

Lernerfahrungen

Christoph	Elmar
Tensorflow und Pytorch, HL2SS (Research Mode), Jupyter, Open3D, CloudCompare	Tensorflow, Pytorch, Cloudcompare, HL2SS, Hololens 2 Datenextraktion
Paper: Swin3D [e], HL2SS [b], Joint 2D-3D-Semantic Data for Indoor Scene Understanding [d]	Paper: Swin3D [e], HL2SS [b], Uni3DScenes Structured3D [d]
Manuelle und automatische Punktwolken-verarbeitung, -registrierung und Labelling, Finetuning/Evaluation Segmentierung, Punktwolkenregistrierung (ICP)	Datensegmentierung (manuelles Labeling) Pre Trained Modell finetuning Transformer Anwendung

Literaturverzeichnis

[a] https://github.com/facebookresearch/segment-anything
[g] https://github.com/facebookresearch/segment-anything/issues/4
[h] https://keras.io/examples/vision/sam/
[i] https://keras.io/guides/keras_cv/segment_anything_in_keras_cv/
[j] https://github.com/JamesQFreeman/LoRA-ViT
[k] https://arxiv.org/pdf/2311.08236.pdf
[l] https://www.youtube.com/watch?v=DhRoTONcyZE
[b] Dibene, Juan Carlos und Dunn, Enrique: HoloLens 2 Sensor Streaming (2022) https://arxiv.org/pdf/2211.02648.pdf
[c] Joint 2D-3D-Semantic Data for Indoor Scene Understanding
http://buildingparser.stanford.edu/images/2D-3D-S_2017.pdf
[d] Structured3D: A Large Photo-realistic Dataset for Structured 3D Modeling
https://www.ecva.net/papers/eccv_2020/papers_ECCV/papers/123540494.pdf
[e] Swin3D: A Pretrained Transformer Backbone for 3D Indoor Scene Understanding
https://arxiv.org/abs/2304.06906
[f] Armeni, Iro et al.: 3D Semantic Parsing of Large-Scale Indoor Spaces – Supplementary Material (2016)
https://svl.stanford.edu/assets/publications/pdfs/3dsemanticparsing_supp_mat.pdf

Tipps und Anmerkungen von Varanasi

- **Besser Zitieren (mit Autoren etc)**
- **Ist auch ok wenn wir mehr als 8 seiten Projektbericht haben**
- **das projekt soll nicht in der schublade verschwinden → neue Studenten akquirieren**
- **auch auf das automatische Labelling eingehen**
- **wir sollen auch auf die technischen details eingehen, zb auch bei der Erstellung des Datensatzes**
- **das sind quasi 3 Projekte 🚀**
- **wir sollen die Herausforderungen erwähnen, sodass die nach uns, da weiterarbeiten können**

Benedikt Beigang, Jan Philipp Seeland, Elmar Kresse, Christoph Walther
Hochschule für Technik, Wirtschaft und Kultur Leipzig