RUB

LEHRSTUHL FÜR
INFORMATIK
IM BAUWESEN

Abdullah Sahin

24.09.2019

Kolloquim zur Bachelorarbeit

Pose Estimation in Gebäuden anhand von Convolutional Neural Networks und simulierten 3D-Daten

Einleitung

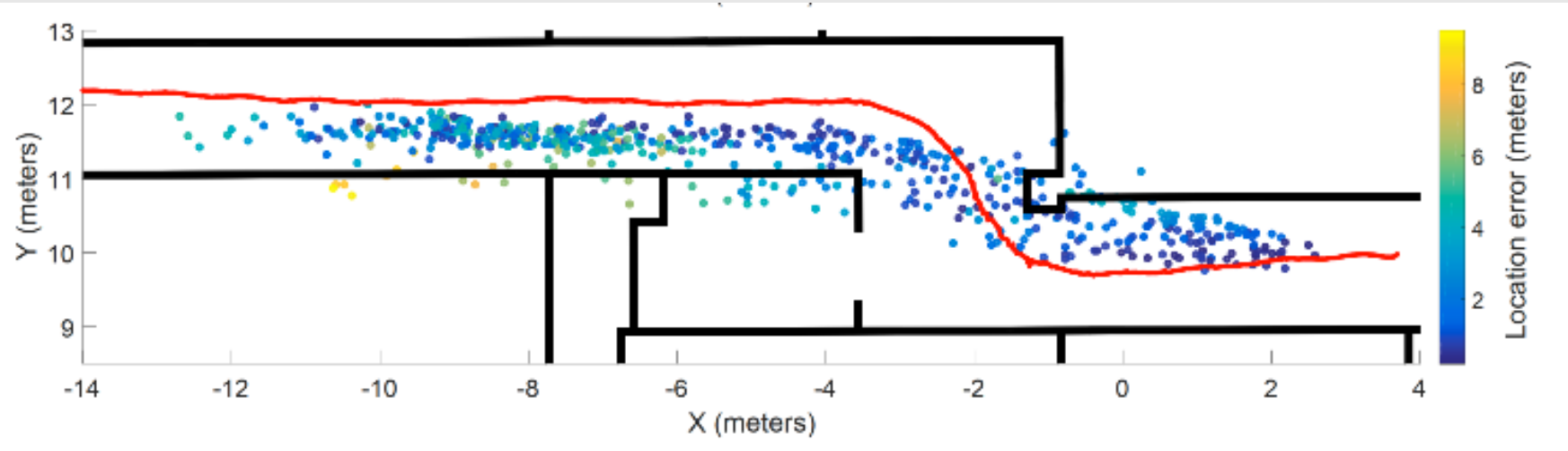
- Pose Estimation in Gebäuden verschafft im Bauwesen z.B.:
  - automatische Baufortschritterfassung
  - Facility-Management & Navigation über Augmented Reality

- visuelle Lokalisierungsverfahren
  - VO oder SLAM sind relativ zum Ausgangspunkt
  - Absolute Bestimmung möglich durch:
    - Suchen eines korrespondierendes Bildes in einer Bildergalerie
    - Pose Regression über Bild-Features

- Ansätze über KNN wie z.B. *PoseNet*
  - Ermittlung der *Ground-Trurth-Daten* über *SfM* genügt

S. der Forschung & Grundlagen

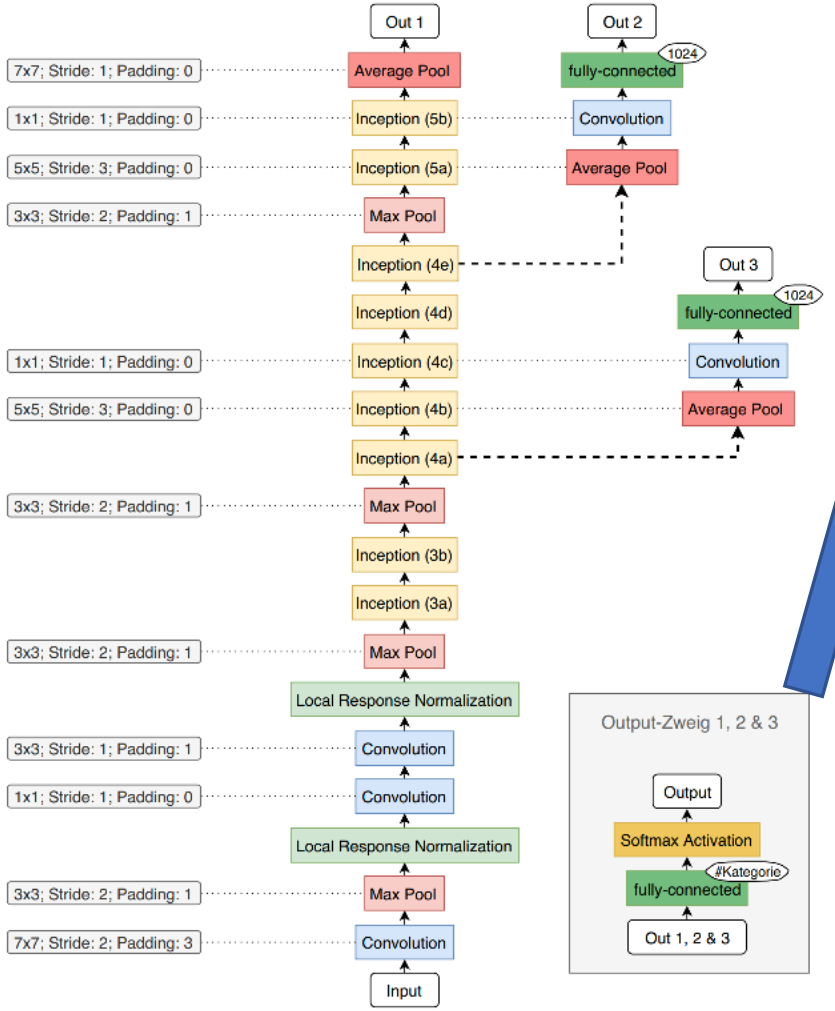
- Acharya et al. erhoben Daten aus Simulation statt über SfM
  - Trainiert mit unterschiedlichen synth. Datentypen => 5m, 20°
  - Trainiert mit Gradientenbildern der synth. Daten => 2m, 7°



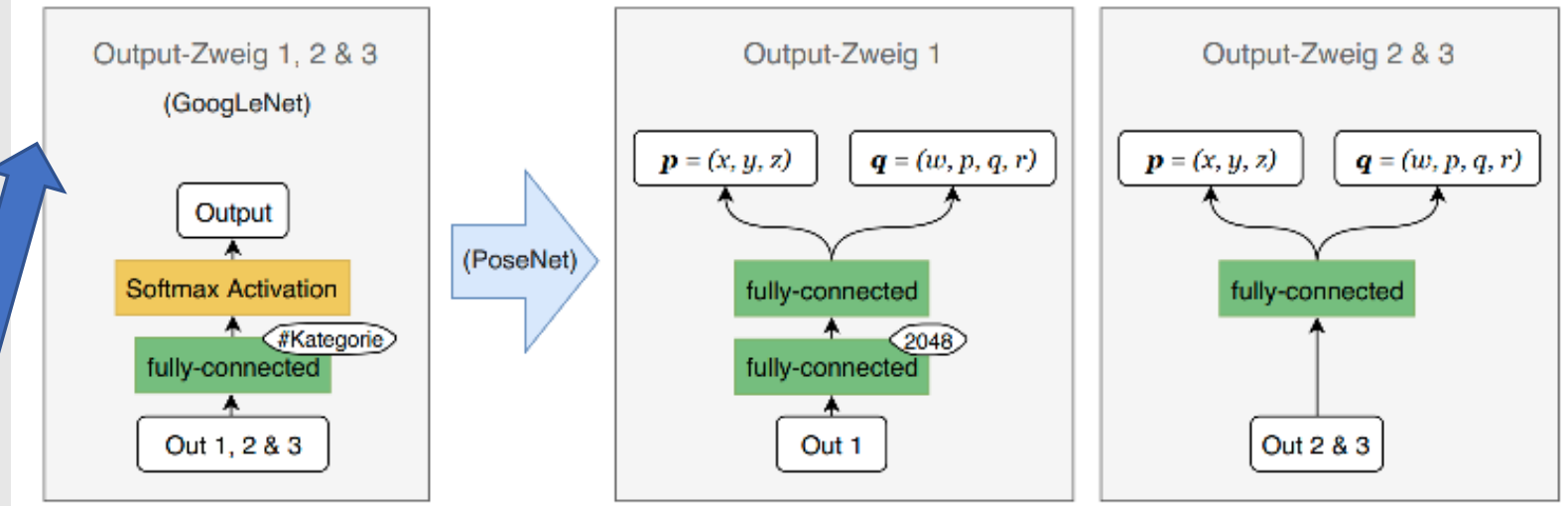
- **Ziel**: diesen Ansatz in größeren Gebäuden auf längeren Strecken zu untersuchen
  - mehrere Richtung & Etagenebenen

CNN Architektur

GoogLeNet, ILSVRC ’14 Sieger



PoseNet, 1. CNN zur Pose-Regression

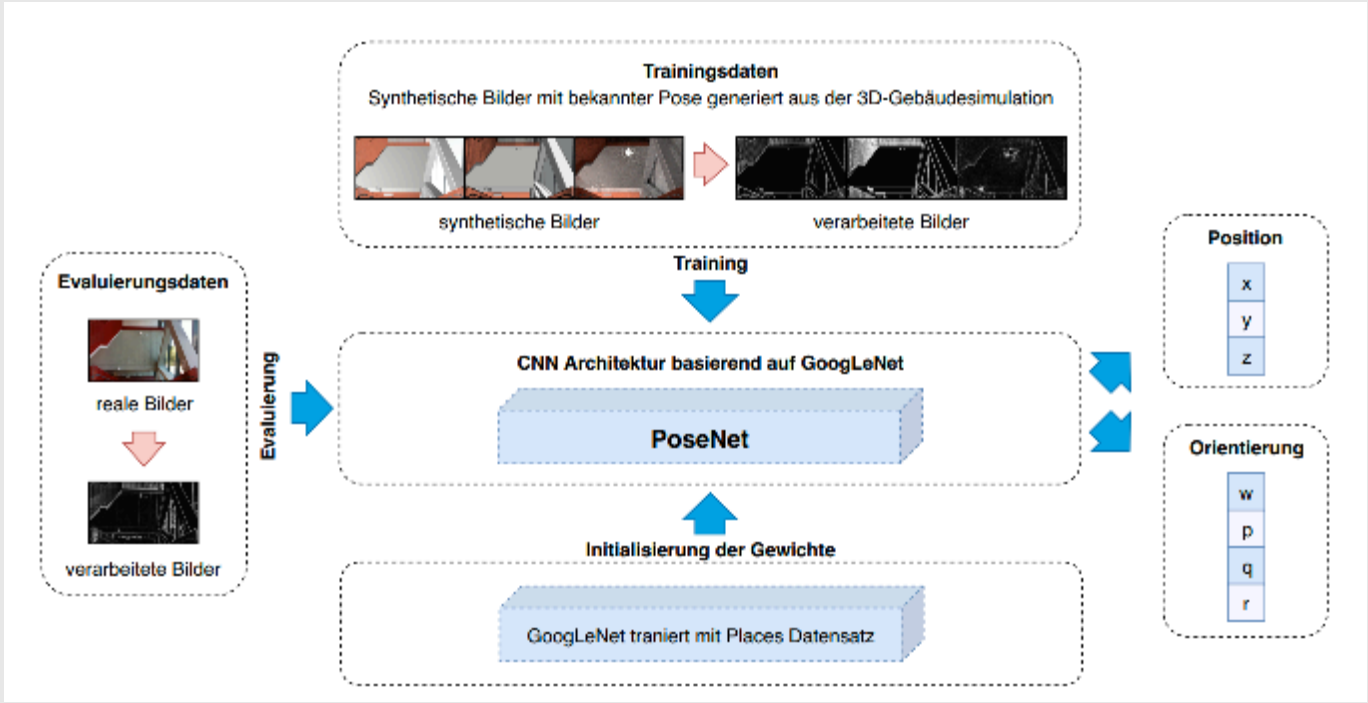


- **ß ist Balance zwischen Position p und Quaterion q**
  - Empfohlen zwischen 120 und 750 in Gebäuden

$$loss(I) = \|\hat{\boldsymbol{p}} - \boldsymbol{p}\| + \beta \left\| \hat{\boldsymbol{q}} - \frac{\boldsymbol{q}}{\|\boldsymbol{q}\|} \right\|$$

Methodik

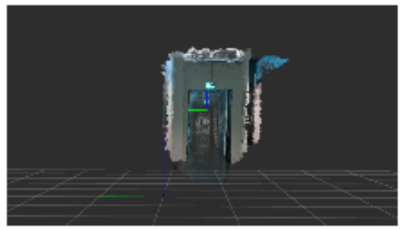
- Erhebung der realen Daten
- Generierung der synth. Daten
- Verarbeitung der Daten
- Datensätze
- Trainingsparameter



Erhebung der realen Daten

- beliebige Kameras
  - SfM-Methoden
- Intel Realsense T265 & D435
  - versicherte bei gegebenen Bestkoniditionen einen Drift von 1%
- Roboter Operating System (ROS)
  - Hardware gesteuert und Datenfluss synchronisiert

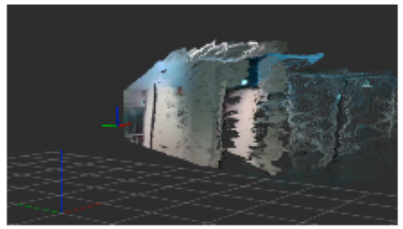
T265
D435



(a) Pose (T265) + 3D-Punktwolke (D435)

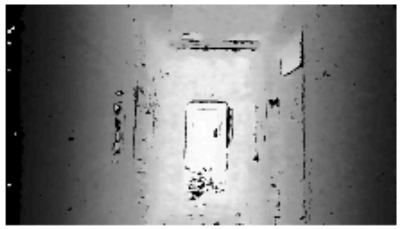
(b) Fischaugenkamera 1 (T265)

(c) Fischaugenkamera 2 (T265)



(d) Pose (T265) + 3D-Punktwolke (D435)

(e) RGB-Bild (D435)



(f) Tiefenbild (D435)

Generierung der synth. Daten

- 3D-Gebäudesmodell aus BIM in Blender v. 2.79b simuliert
- (*bestmögliche*) Imitation der Aufnahmestrecken
  - 0.05m Intervallen mit +/- 10° Neigung in je y- und z-Achse
- Insgesamt 3 Typen (*angelehnt an Acharya et al.*)
  - *cartoon, edge, photoreal*

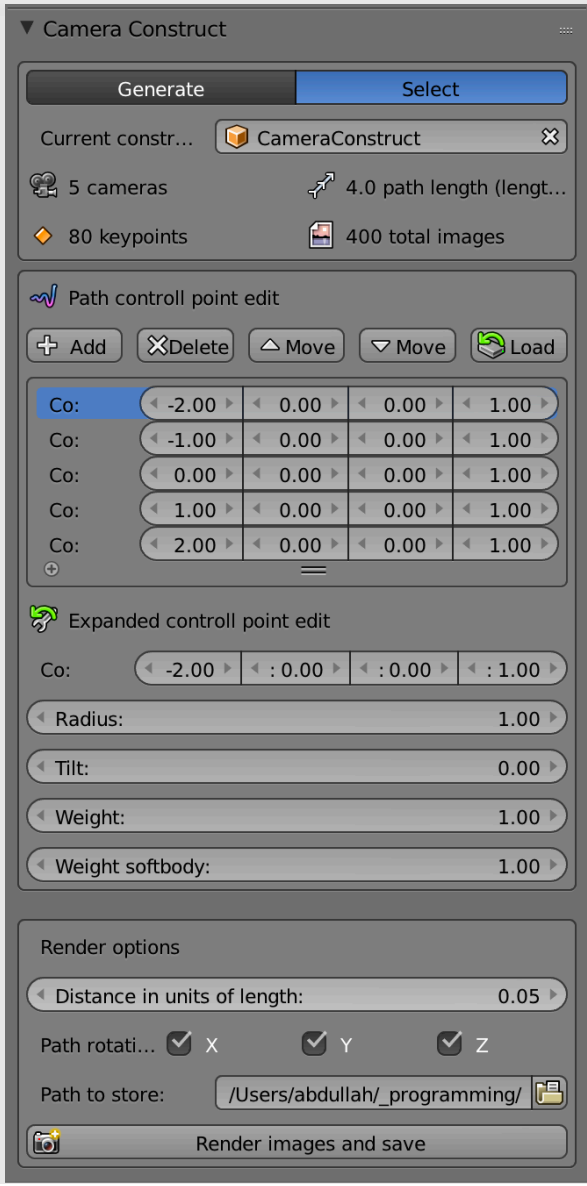
(a) Orginal Pose

(b) -10° um die y-Achse

(c) +10° um die y-Achse

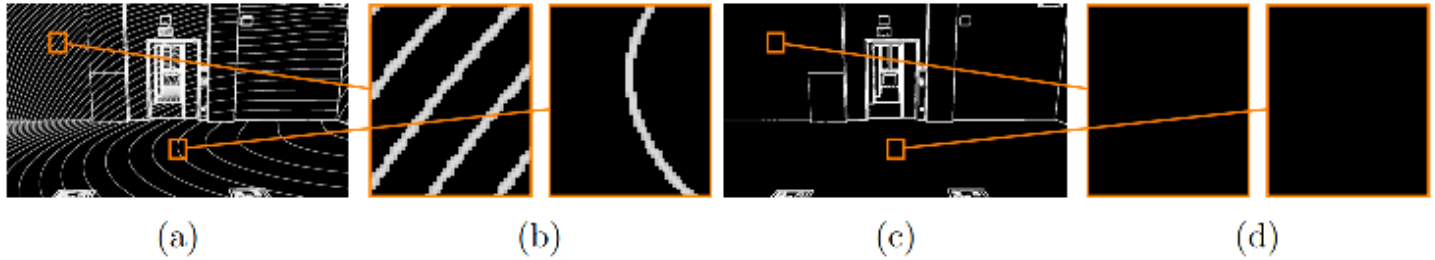
(d) -10° um die z-Achse

(e) +10° um die z-Achse

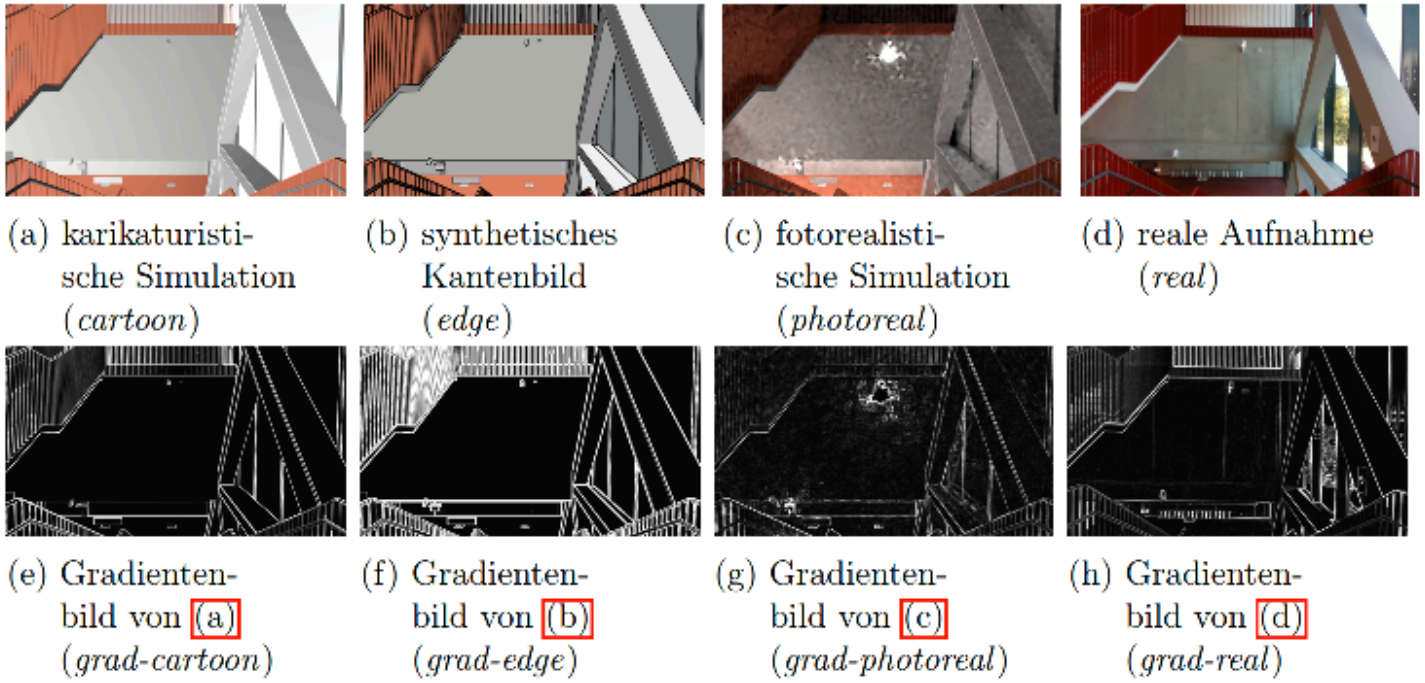


Verarbeitung der Daten

- Gradientenbild
  - reale Bilder auf die Größe der synth. Daten skaliert



- Treshold-Verfahren
  - Unterdrückung der durch die synth. Lichter entstandenen Artefakte



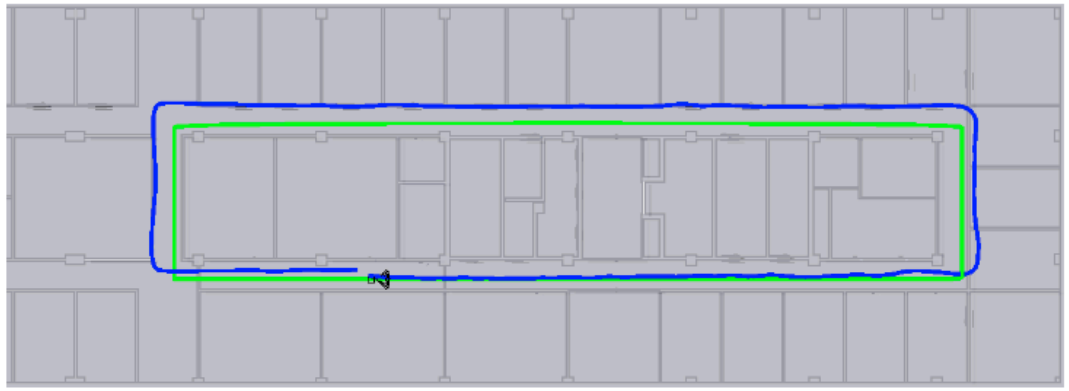
Datensätze

- länger, in mehrere Richtungen verlaufend und auf mehrere Etagenebenen erstreckend

- Daten erhoben:
  - nördliche Hälfte des 6. Stockwerkes des IC-Gebäudes der RUB (IC)
  - Seminargebäude der Hochschule Bochum (HS)

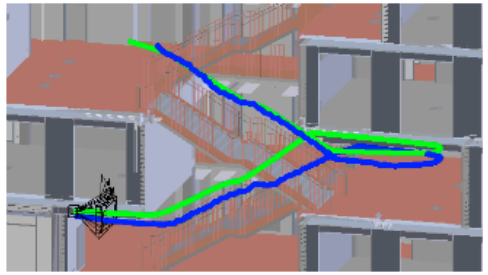
(a) IC-Simulation

(b) HS-Simulation

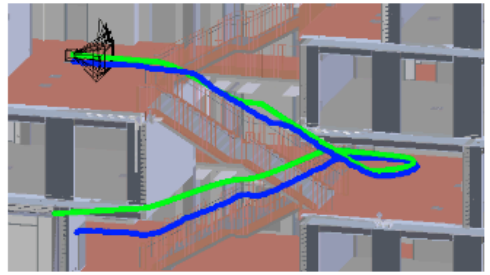


(a) *IC-loop*

(b) *HS-gamma*



(c) *HS-stairs-up*



(d) *HS-stairs-down*

Trainingsparameter

- Hyperparameter wurden **übernommen** bzw. **gleichermaßen bestimmt** oder **im selben Verhältnis** zum Datensatz **gewählt**.

- Hyperparameter Beta wurde wie empfohlen via *Grid-Search* bestimmt
  - Trainiert und Evaluiert mit realen Daten
  - Ergebnis dient als Referenzwert

Hyperparameter	**Wert**
Architektur	PoseNet (s. Abschn. 2.4.2)
Implementierung	*Caffe*
Batchgröße	40
Anzahl der Epochen	160
Datenaufteilung	50% Trainingsdaten 50% Evaluationsdaten
Bildskalierung	480×270
Bildausschnitt	224×244 (*Training*: zufällig, *Evaluation*: zentriert)
Datensatznormierung	Subtraktion des Durchschnittsbildes der Trainingsdaten
β der Kostenfunktion (s. Gleichung 5)	*IC-loop*: 680 *HS-gamma*: 120 *HS-stairs-up*: 470 *HS-stairs-down*: 610
Loss-Optimierer	*AdaGrad*
Lernrate	10^{-3}
Initialisierung der Gewichte	Gewichte eines mit dem *Places Datensatz* trainierten Modells auf GoogLeNet

Ergebnisse

- Evaluationsergebnis
  - Abweichung der Position in Meter und den Orientierungsfehler in Grad
  - werden anhand der Positionsfehler verglichen

- Akkuratesse gibt an:
  - ➢ Median der Evaluationsergebnisse

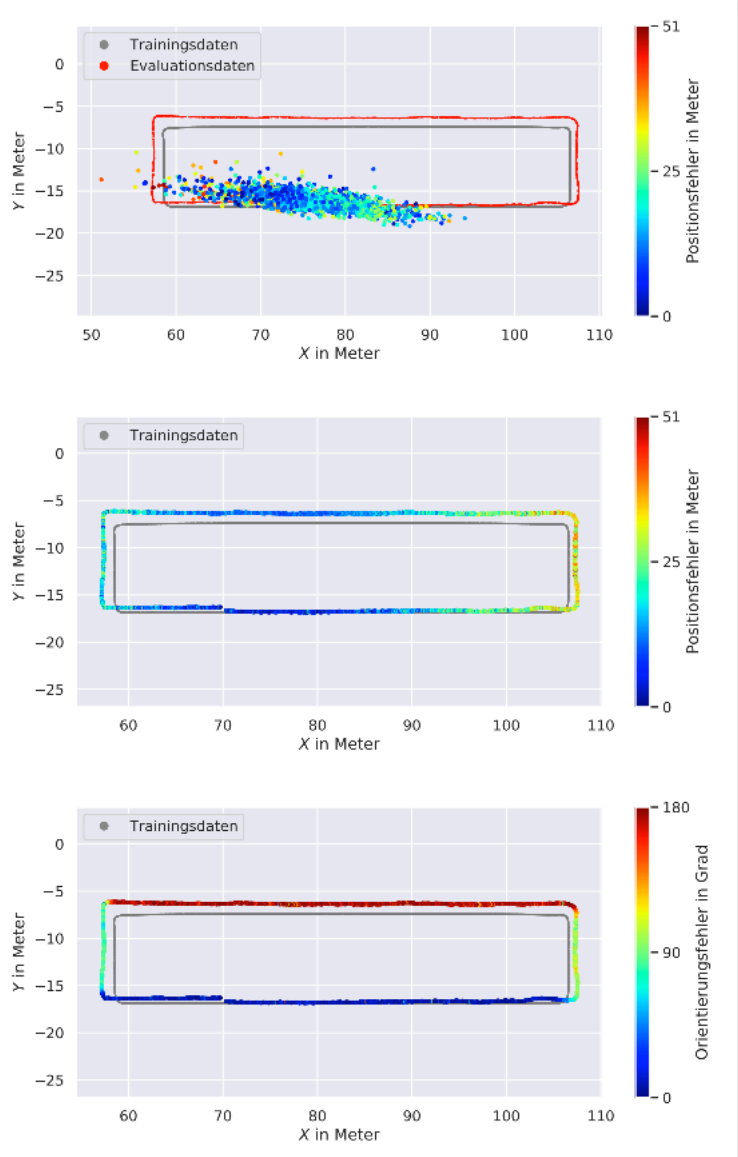
- Evaluation
  - 1.: Gradientenbilder der korrespondierenden synthetischen Evaluationsdaten
  - 2.: Gradientenbilder der realen Evaluationsdaten

IC-loop

Referenzwert: 1.93m, 4.26°

Trainingsdatensatz (Gradientenbild)	**synthetische Daten** (Position, Orientierung)	**reale Daten** (Position, Orientierung)
grad-cartoon	1.61m, 8.17°	23.56m, 51.30°
grad-edge	2.00m, 8.29°	32.91m, 59.17°
grad-photoreal	1.80m, 7.70°	16.68m, 73.25°
∅ Durchschnitt	1.80m, 8.05°	24.38m, 61.24°

- Das Netzwerk bestimmte:
  - alle Evaluationsdaten in einem ca. 30m x 5m großen Teilbereich
  - Die Orientierung als die Aufnahmerichtung der unteren horizontalen Strecke

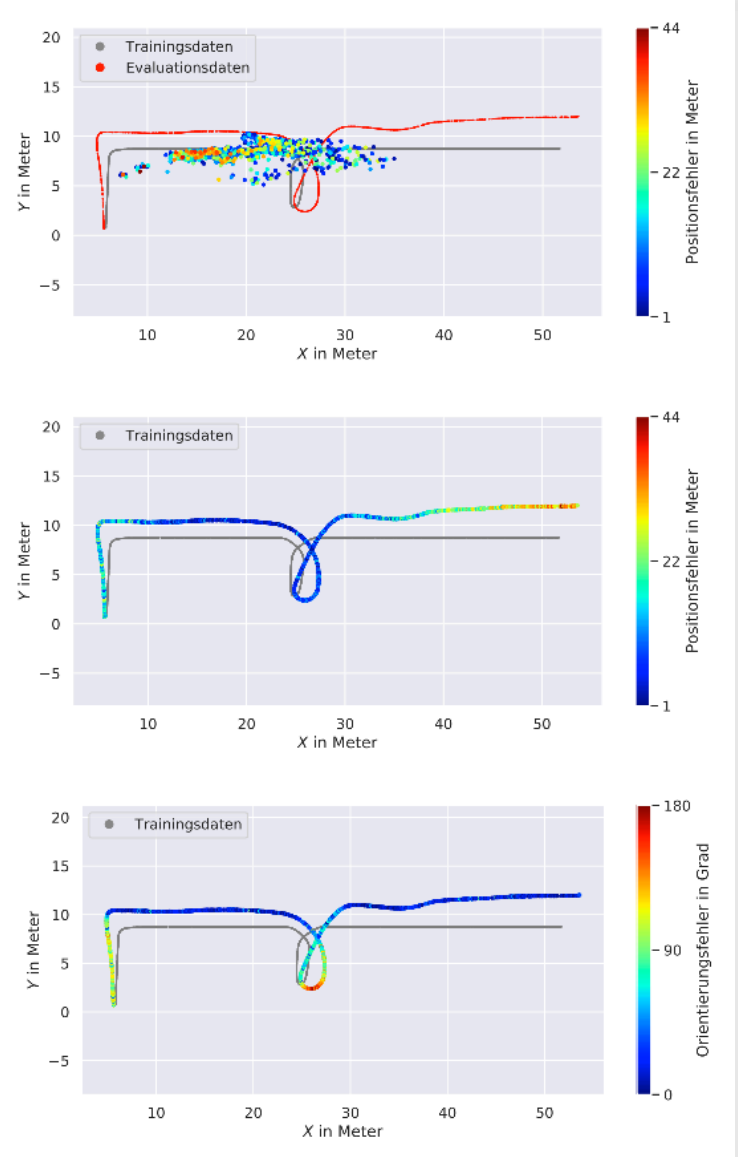


HS-gamma

Referenzwert: 0.95m, 7.53°

Netzwerk (Trainingsdatensatz)	synthetische Daten (Position, Orientierung)	reale Daten (Position, Orientierung)
grad-cartoon	1.00m, 9.92°	8.60m, 19.59°
grad-edge	1.07m, 8.69°	10.15m, 35.11°
grad-photoreal	1.45m, 9.17°	10.27m, 41.60°
∅ Durchschnitt	1.17m, 9.26°	9.67m, 32.10°

- Das Netzwerk bestimmte:
  - alle Evaluationsdaten in einem ca. 20m x 5m großen Teilbereich
  - Die Orientierung als die Aufnahmerichtung der horizontalen Strecken

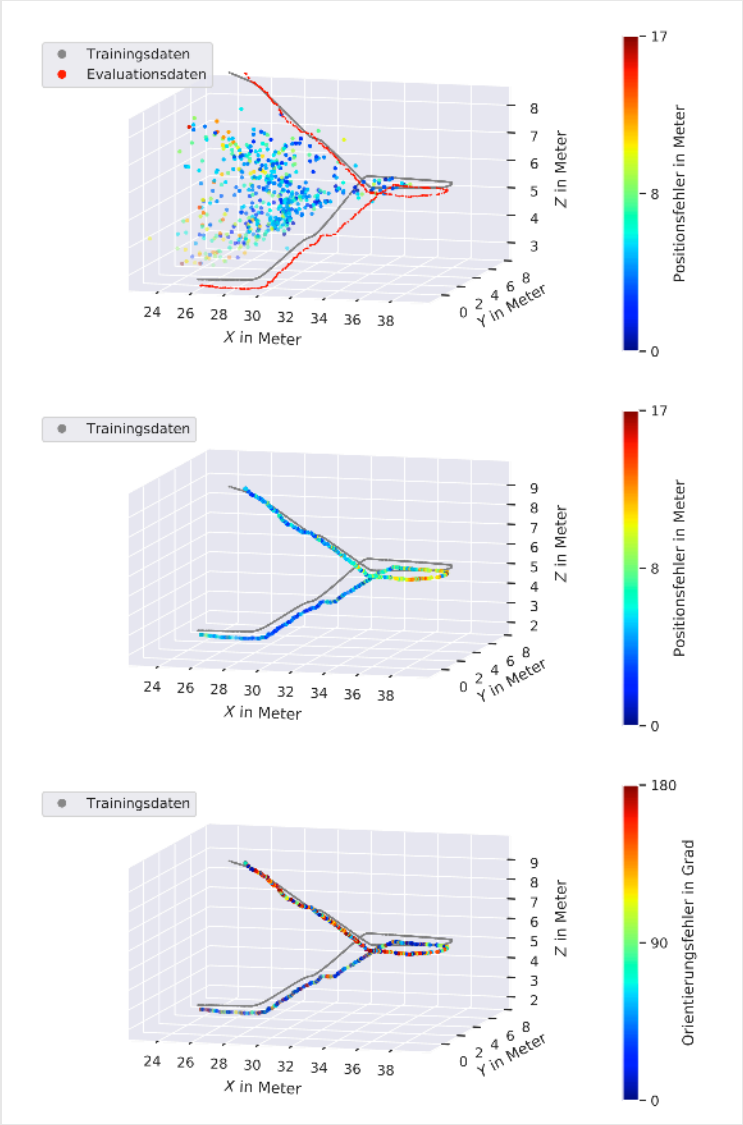


HS-stairs-up

Referenzwert: 0.94m, 8.33°

Netzwerk (Trainingsdatensatz)	**synthetische Daten** (Position, Orientierung)	**reale Daten** (Position, Orientierung)
grad-cartoon	0.82m, 7.76°	4.77m, 23.43°
grad-edge	0.82m, 8.48°	4.33m, 51.64°
grad-photoreal	0.92m, 7.98°	5.16m, 93.38°
∅ Durchschnitt	0.85m, 8.07°	4.75m, 56.15°

- Das Netzwerk bestimmte:
  - alle Evaluationsergebnisse zwischen der dem oberen und unteren Treppenlauf
  - abwechselnd größere Positionsfehler
  - Die Orientierung wurde abwechselnd in der entgegengesetzten Orientierung bestimmt

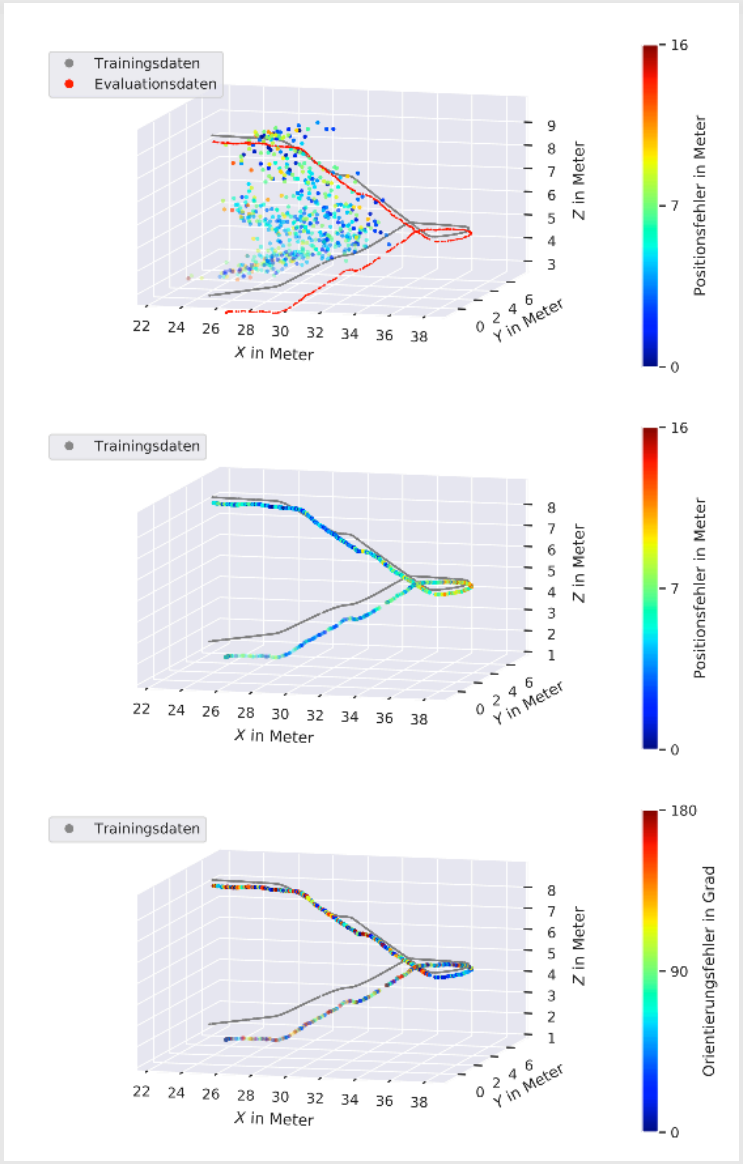


HS-stairs-down

Referenzwert: 0.87m, 9.25°

Netzwerk (Trainingsdatensatz)	**synthetische Daten** (Position, Orientierung)	**reale Daten** (Position, Orientierung)
grad-cartoon	0.91m, 8.01°	4.20m, 47.83°
grad-edge	0.85m, 7.50°	5.59m, 67.34°
grad-photoreal	1.02m, 8.57°	5.25m, 32.70°
∅ Durchschnitt	0.93m, 8.03°	5.01m, 49.29°

- Das Netzwerk bestimmte:
  - alle Evaluationsergebnisse zwischen der dem oberen und unteren Treppenlauf
  - abwechselnd größere Positionsfehler (*sichtbarer*)
  - Die Orientierung wurde abwechselnd in der entgegengesetzten Orientierung bestimmt (*sichtbarer*)



Zusammenfassung

- **Referenzwert**: durch die Bestimmung des Hyperparameters Beta der Kostenf.
- **Evaluation 1**: Gradientenbilder der korrespondierenden synth. Daten
- **Evaluation 2**: Gradientenbilder der realen Daten

Strecke	Referenzwert	∅ Evaluation 1	∅ Evaluation 2
IC-loop	$1.93m$, $4.26°$	$1.80m$, $8.05°$	$24.38m$, $61.24°$
HS-gamma	$0.95m$, $7.53°$	$1.17m$, $9.26°$	$9.67m$, $32.10°$
HS-stairs-up	$0.94m$, $8.33°$	$0.85m$, $8.07°$	$4.75m$, $56.15°$
HS-stairs-down	$0.87m$, $9.25°$	$0.93m$, $8.03°$	$5.01m$, $49.29°$
∅ Durchschnitt	$1.17m$, $7.34°$	$1.19m$, $8.35°$	$10.95m$, $49.69°$

Diskussion

- Diskussion der angewandten Methodik
- Diskussion der Ergebnisse
- Empfehlungen für weiterführende Forschung

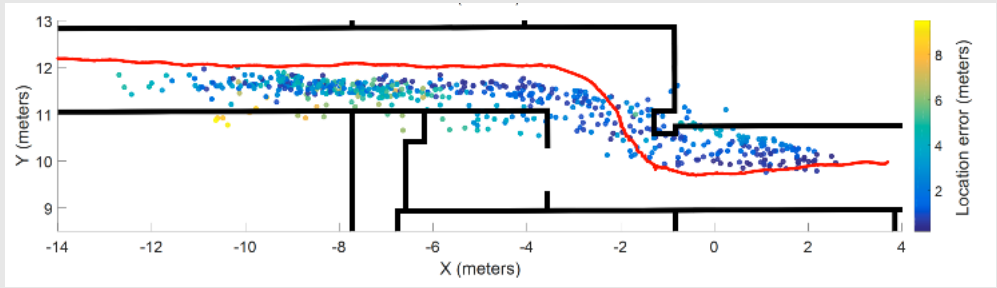
D. der angewandten Methodik

- reale Daten wiesen bis zu 5% Drift auf
  - negativer Einfluss auf die domänenübergreifende Evaluation
- Akkuratesse ist vom Zufall abhängig
  - 5 Trainingsprozesse sind wenig
- Hyperparameter
  - wurden nicht optimiert, könnten auf den Datensatz von Acharya et al. optimiert sein

D. der Ergebnisse

- HS-stairs-down & HS-stairs-up
  - keine Generalisierungsfähigkeit zu erkennen (perceptual-aliasing)

- IC-loop & HS-gamma
  - ca. 5m breiten und 20m bis 30m langen Teilbereich
  - nur eine Richtung
  - Parallelen zur Acharya et al.‘s Datensatz



- PoseNet ist nicht begrenzt
  - mit Daten der gleichen Domäne konnte ca. 1m Positionsakkuratesse erzielt werden
  - Walch et al. erzielten mit TUM-Datensatz (größer als die Obigen) ca. 2m Positionsakkuratesse
  - => domänenübergreifende Training mit Gradientenbildern nur auf 5m x 30m in einer Richtung

Empf. für weiterf. Forschungen

- Anzahl der Trainingsprozesse bei gleichen Hyperparameter erhöhen
  - bessere Ergebnisse erzielen oder ausschließen => bestes synth. Datentyp bestimmen
- Optimierung der Hyperparameter
  - führt zu besseren Ergebnissen
- Nachfolger von PoseNet
  - versichern Verbesserung

Fazit

- Insgesamt wurde der Ansatz mit 2 Gebäuden auf 4 Strecken untersucht
- Durchschnittliche Akkuratesse von ca. 1m, 8° bei Daten der gleichen Domäne
- Domänenübergreifend: 10.95m, 49.69°
  - Parallelen ließen schlussfolgern, dass der Ansatz begrenzt ist
- Lokalisierungsverfahren undenkbar
  - Potenzielle Akkuratesse von ca. 1m im direkten Gebrauch ungeeignet, allerdings durch Kaskadeneffekt verbesserbar

Fazit

- Unzureichende Akkuratesse bei domänenübergreifende Evaluation
  - liegt den Simulationsdefiziten und domänenspezifische Artefakte zugrunde

- Lohnenswerte Untersuchung:
  - Diskrepanzminimierung zwischen synth. und realen Daten durch z.B. GANs
  - Ferner: Beschränkungen der möglichen Posen im Trainingsprozess


Vielen Dank
für eure Aufmerksamkeit!