HAW HAMBURG

HAUPTPROJEKT
Juri Zach

Erkennen von Fischen in Unterwasservideos mithilfe von unüberwachten Lernverfahren

FAKULTÄT TECHNIK UND INFORMATIK
Department Informatik

Faculty of Computer Science and Engineering
Department Computer Science

Betreuung durch: Prof. Dr. Tim Tiedemann
Eingereicht am: 18. Oktober 2020

HOCHSCHULE FÜR ANGEWANDTE WISSENSCHAFTEN HAMBURG
Hamburg University of Applied Sciences

Inhaltsverzeichnis

1 Einleitung **1**

2 Ähnliche Arbeiten **2**

3 Erkennen von Fischen anhand ihrer Bewegungen **3**

4 Verwendete Daten **5**

5 Algorithmen **7**

5.1 Berechnung des optischen Flusses 7

5.2 Veränderung des optischen Flusses 7

5.3 Aufbau der neuronalen Netze 8

5.4 Verlustfunktionen 9

5.5 Validierungsmetriken 10

6 Experimente **11**

6.1 Experiment 1: Training auf dem David Datensatz mit Veränderung des optischen Flusses 11

6.2 Experiment 2: Training auf dem David Datensatz ohne Veränderung des optischen Flusses 13

6.3 Experiment 3: Training auf dem virtuellen Tauchdatensatz 15

7 Fazit **18**

8 Ausblick **19**

Literatur **20**

In dieser Arbeit wird ein maschinelles Lernverfahren zum Segmentieren von Fischen in Unterwasservideos vorgestellt. Dieses soll im Rahmen des AuTag BeoFisch Projektes, zum Untersuchen von Fischpopulationen in der Ostsee, eingesetzt werden.
Da die Annotation von Videobildern sehr zeitaufwändig ist, wird in dieser Arbeit ein unüberwachtes Lernverfahren verwendet. Dieses ist in der Lage Fische und andere sich Unterwasser fortbewegende Lebewesen, pixelgenau zu segmentieren.
Zu diesem Zeitpunkt sind die Unterwasserroboter, welche zum Erstellen des Trainingsdatensatzes benötigt werden, noch nicht vorhanden. Aus diesem Grund wird das hier entwickelte Verfahren auf virtuellen Daten getestet, um dessen Nützlichkeit für das AuTag BeoFisch Projekt zu evaluieren.

Keywords: maschinelles Lernen, unüberwachtes Lernen, neuronale Netze, Bildsegmentierung, Unterwasservideos

1 Einleitung

Im Rahmen des Projektes *Autonome Tauchroboter-gestützte Beobachtung von Fischschwärmen* (AuTag BeoFisch) [1] wird eine Methode zum Erkennen von Fischen entwickelt. Das Ziel des AuTag BeoFisch Projektes ist der Einsatz von mehreren Unterwasserrobotern, die kooperativ Untersuchungsgebiete in der Ostsee erforschen, um möglichst genaue Daten für Biologen zu sammeln.
Die am AuTag BeoFisch beteiligten Biologen interessieren sich vorwiegend für Daten über Fische. Momentane Forschungsfragen beziehen sich auf das Vorkommen, Verhalten und die Biomasse von Dorschen, der aufgrund seiner kommerziellen Nutzung von besonderer Relevanz ist. Doch auch das Erforschen andere Fischarten wird nicht ausgeschlossen.

Damit die Unterwasserroboter gezielt Daten über Fische sammeln können, müssen sie über Methoden verfügen um ein Unterwasserobjekt als Fisch zu identifizieren. Zu diesem Zweck wird ein Algorithmus zum Erkennen von Fischen in Videosequenzen entwickelt. Im Bereich des maschinellen Sehens werden vorwiegend Lernverfahren, insbesondere *Deep*

[1] https://autosys.informatik.haw-hamburg.de/project/beofisch/

Learning angewendet. Eine bewerte und erfolgversprechende Methode ist das Trainieren von neuronalen Netzen auf annotierten Trainingsdaten. Leider gibt es nur wenige öffentliche Datensätze von Unterwasser-Fischen, keines davon eignet sich für das AuTag BeoFisch Projekt.
Im Gegensatz zum Sammeln von Videodaten ist das Annotieren derselben sehr zeitaufwändig. Aus diesem Grund wird in dieser Arbeit ein unüberwachtes Lernverfahren zum Erkennen von Fischen in Unterwasservideos vorgestellt. Hiermit können schwimmende Fische, aufgrund ihrer Bewegungen, pixelgenau segmentiert werden. Dies gelingt durch die Tatsache, dass sich Fische, anders als viele andere Unterwasserobjekte und Organismen, autonom bewegen können. Hierdurch sind ihre Bewegungen unabhängig von denen ihrer Umgebung, welche durch Strömungen, Gravitation, Kamerabewegungen oder anderen Umwelteinflüssen bedingt sind.
Mithilfe von zwei neuronalen Netzen, die sich gegenseitig drauf trainieren unabhängige Bewegungen zu erkennen, werden sich autonom bewegende Objekte pixelgenau segmentiert.

Die Arbeit ist wie folgt gegliedert: Im Abschnitt 2 werden ähnliche wissenschaftliche Arbeiten vorgestellt, auf denen diese Arbeit aufbaut. Abschnitt 3 beschreibt das Vorgehen beim unüberwachten Lernen und die Annahmen, auf denen das hier vorgestellte Verfahren basiert. Im Abschnitt 5 werden die verwendeten Algorithmen im Detail beschrieben. Dieser Abschnitt richtet sich an eine Leserschaft mit informatischem Vorwissen und könnte für Leser aus anderen Fachrichtungen schwer zu verstehen sein. Die Ergebnisse der Arbeit werden im Abschnitt 6 vorgestellt und anschließend im Abschnitt 7 zusammengefasst. Ein Ausblick zu weiteren Forschungsfragen gibt der Abschnitt 8.

2 Ähnliche Arbeiten

Zur Vorbereitung dieser Arbeit wurde nach geeigneten Trainingsdatensätzen gesucht. Im Unterwasserbereich gibt es nur wenige Datensätze von Fischen. Hierzu gehört das Projekt *Fish4Knowledge* [4] in dem seit mehreren Jahren Unterwasservideos an neun verschiedenen Standorten in Taiwan aufgenommen werden. Diese sind frei verfügbar und bieten eine große Datenbasis für unüberwachte Lernverfahren. Einige dieser Daten wurden von PeRCeiVe Lab mit pixelgenauen Segmentierungen annotiert [2]. Ein weiterer Unterwasserdatensatz ist *EyeSea* [1], welcher Videos von drei stationären Kameras an Wasserkraftwerken beinhaltet und mit Begrenzungsrahmen um Fische annotiert wurde.

Dieser Datensatz wurde bereits in einer Arbeit von Xu und Matzner [13] verwendet, um Fischerkennung auf neuronalen Netzen zu trainiert. Hierbei hat sich herausgestellt, dass sich neuronale Netze, die auf Daten von nur zwei der Kameras trainiert wurden, schlecht auf die Daten der dritten Kamera generalisieren lassen. Dies lässt vermuten, dass Transferlernen im Unterwasserbereich nicht vielversprechend ist.
Die schlechte Generalisierbarkeit könnte allerdings auch an den stationären Kameras und dem somit konstanten Bildhintergrund liegen. Hierdurch können Trainingsfehler entstehen, bei denen nicht die visuellen Eigenschaften der Fische, sondern sich verändernde Hintergrundeigenschafgen gelernt werden.

Sowohl der *Fish4Knowledge* als auch der *EyeSea* Datensatz haben den Nachteil, dass sie nicht in der Ostsee aufgenommen wurden. Hierdurch unterscheiden sich die Lichtverhältnisse, die Arten der vorkommenden Fische sowie Plankton- und Partikel-Konzentrationen. Außerdem wurden sie mit stationären Kameras aufgenommen. Durch die Bewegung eines Unterwasserroboters entstehen deutlich komplexere Videoaufnahmen, da sich die gesamte Kamera im Raum bewegt. Dies kann die Identifizierung von Fischen erheblich erschweren.

Im Bereich der Bildsegmentierung gibt es mehrere Videodatensätze von bewegten Objekten, welche von nicht stationären Kameras aufgenommen wurden. Diese zeigen verschiedene Alltagsszenen (Überwasser) und sind mit pixelgenauen Annotationen versehen. Hierzu gehören die Datensätze Davis2016 [9], FBMS59 [8] und SegTrackV2 [12].
In der Arbeit Yang et al. [14] wird auf diesen Datensätzen ein unüberwachter Lernansatz zum Segmentieren von bewegten Objekten demonstriert.
Wie dieser Ansatz aufgebaut ist und weshalb er sich besonders gut für die Segmentierung von Fischen eignet, wird im nächsten Abschnitt erklärt.

3 Erkennen von Fischen anhand ihrer Bewegungen

Die Unterwasserwelt der Ostsee bietet einer Vielzahl von Lebewesen einen Lebensraum. Hierzu gehören verschiedene Meeressäuger, Fische, Muscheln, Schnecken, Krebstiere, Quallen, Plankton und viele andere.
Aufgrund ihrer Bewegungsweise können die verschiedenen Meeresbewohner in zwei Kategorien unterteilt werden. Zum einen gibt es Organismen, die sich gar nicht, oder nur sehr langsam fortbewegen können. Zu dieser Gruppe gehören alle Meerespflanzen, Plankton (wozu auch Quallen gehören), Muscheln und Schnecken. Zu der anderen Gruppe gehören

alle Organismen, die sich schnell fortbewegen können. Hierzu gehören Fische, Meeressäuger und Krebstiere.

In der Gruppe der beweglichen Organismen befinden sich viele Lebewesen, welche für die Biologen des AuTag BeoFisch Projektes von Interesse sind. Aus diesem Grund wird ein unüberwachter Lernalgorithmus entwickelt, der Organismen der zweiten Gruppe, anhand ihrer Bewegungen, erkennen und in Videobildern segmentieren kann. Da erwartet wird, dass Fische den Großteil der beweglichen Organismen ausmachen, welche im Rahmen des AuTag BeoFisch Projektes gefilmt werden, wird in dieser Arbeit oft über das Erkennen und Segmentieren von Fischen geschrieben. Dennoch wird der hier beschriebene Algorithmus nicht nur Fische, sondern auch andere Organismen wie Meeressäuger oder auch Krebstiere (sollte nah am Grund gefilmt werden) erkennen.

Der Lernalgorithmus basiert auf einer Arbeit von Yang et al. [14] in der bewegte Objekte anhand ihrer gemeinsamen Bewegungsinformation separiert werden.
Bei bewegten Bildern kann der optische Fluss zwischen zwei aufeinanderfolgenden Bildern errechnet werden. Dies ist ein Vektorfeld, welcher anzeigt inwieweit und in welche Richtung sich einzelne Bildpunkte bewegen.
Der optische Fluss einer Unterwasseraufnahme wird viele gemeinsame Bewegungsinformationen enthalten. Aufgrund der Kamerabewegung, bewegt sich Beispielsweise der gesamte Hintergrund konstant in eine Richtung oder Unterwasserströmungen lassen Quallen und Plankton vorbeitreiben. All diese Bewegungen resultieren aus allgemeinen Gegebenheiten, welche sich auf viele Objekte gleichzeitig auswirken. Anders verhält es sich hingegen mit der Schwimmbewegung eines Fisches. Diese hängt nur von dem Fisch selbst ab und teilt sich keine oder nur wenige Bewegungsinformation mit der Umgebung.

Basierend auf dieser Annahme werden zwei neuronale Netze entwickelt, welche im gegenseitigen Konkurrenzkampf erlernen sollen, Objekte ohne gemeinsame Bewegungsinformation zu erkennen.
Das eine neuronale Netz wir als Rekonstruktions-Netz bezeichnet. Dieses lernt, fehlende Segmente eines unvollständigen optischen Flusses zu rekonstruieren. Hierfür wird dem Netz ein Bild und der unvollständige optische Fluss des Bildes zur Verfügung gestellt. Das Netz wird darauf optimiert, den optischen Fluss so gut wie möglich zu rekonstruieren.
Das zweite neuronale Netz wird als Generator-Netz bezeichnet. Seine Aufgabe besteht darin, bestimmte Segmente aus einem optischen Fluss herauszuschneiden, welche anschießend vom Rekonstruktions-Netz wiederhergestellt werden können. Hierbei ist es die

Aufgabe des Generator-Netzes, Fluss-Segmente auszuwählen, welche besonders schwer wiederherzustellen sind.

Im Laufe des Trainings wird das Rekonstruktions-Netz anhand der gemeinsamen Bewegungsinformation immer besser den optischen Fluss rekonstruieren können. Damit dass Generator-Netz dennoch seine Aufgabe erfüllen kann, muss dieses Fluss-Segmente entfernen, welche sich möglichst wenig gemeinsame Bewegungsinformation mit anderen Objekten im Bild teilen.
In Unterwasserbildern gehören Objekte ohne gemeinsame Bewegungsinformation zu der Gruppe der beweglichen Organismen. Somit lernt das Generator-Netz, sich bewegende Fische, Meeressäuger und ähnliches zu segmentieren um die Rekonstruktion des optischen Flusses so schwer wie möglich zu gestalten.

Durch diesen Algorithmus, der im folgenden Text als *unüberwachtes Segmentieren* bezeichnet wird, können ohne menschlichen Annotationsaufwand, eine Vielzahl von Fischen, Meeressäugern und anderen sich bewegenden Organismen aus Videobildern segmentiert werden. Diese Segmentierungen können zum Erlernen der optischen Merkmale von Fischen genutzt werden, wodurch auch nicht schwimmende Fische erkannt werden können. Mithilfe von Clustering-Algorithmen können die verschiedenen Segmentierungen in Untergruppen unterteilt werden, um beispielsweise zwischen verschiedenen Fischarten zu unterscheiden.

4 Verwendete Daten

Geplant ist es, einen Unterwasservideodatensatz mit einem Tauchroboter der Bauart BlueROV 2 zu erstellen. Hierbei sollen vor allem Videos verwendet werden in denen sich fortbewegende Organismen wie Fische vorkommen. Um den Algorithmus mit schwierigen Szenarien zu testen, sollen aber auch Daten von treibenden und sich in der Strömung bewegenden Objekten wie Quallen und Algen, sowie bewegte Grundaufnahmen im Datensatz enthalten sein.
Ein kleiner Teil des Datensatzes wird per Hand annotiert und als Validierungsdatensatz verwendet. Die übrigen Daten werden zum Trainieren der unüberwachten Segmentierung verwendet.

Leider ist der Tauchroboter zum jetzigen Zeitpunkt noch nicht verfügbar, weshalb auf andere Daten zurückgegriffen werden muss.

Aus diesem Grund wird zum Testen der Qualität des Algorithmus der Davis2016 Datensatz [9] verwendet. Dieser besteht aus verschiedenen Videosequenzen von Alltagsszenarien (Überwasser) und beinhaltet die Segmentierungsmasken der Vordergrundobjekte, welche zum Berechnen der Qualität verwendet werden. Für die unüberwachte Segmentierung wird der Datensatz um den optischen Fluss erweitert. Wie auch in der als Vorlage dienenden Arbeit von Yang et al. [14] wird der optische Fluss mithilfe der Ground Truth Segmentierungsmasken verändert, um sicherzustellen, dass die Bewegungsinformationen vom Vordergrund und Hintergrund keine geteilten Informationen enthalten.

Die Veränderung des optischen Flusses mithilfe der Ground Truth Segmentierungsmasken gewährleistet zwar eine Unabhängigkeit der Bewegungsinformationen, doch sie vereinfacht auch stark das Lernproblem. Um die Schwierigkeit des Lernproblems mit realen Daten besser einschätzen zu können und um herauszufinden inwieweit die Bewegungsinfomaitonen in realen Daten unabhängig voneinander sind, wird die unüberwachte Segmentierung auch auf den Davis Datensatz ohne Veränderungen im optischen Fluss getestet.

Um herauszufinden, inwieweit sich die unüberwachte Segmentierung für Unterwasservideos eignet, wird ein eigener Datensatz erstellt. Optimal wäre es Unterwasservideos mithilfe eines Tauchroboters aufzunehmen. Da dieser allerdings noch nicht verfügbar ist, wird ein simulierter Datensatz erstellt. Hierfür werden Videoaufnahmen aus dem Computerspiel *World of Diving* verwendet. *World of Diving* ist ein Tauchsimulator, in dem der Spieler verschiedene Unterwasserwelten erkunden kann, welche nach realen Vorbildern gestaltet wurden. Für den Spieler gibt es verschiedene Fische, Quallen, Meeressäuger und Unterwasserpflanzen zu entdecken. Zum Erstellen des Datensatz werden ausgewählte Szenen aus dem Spiel mitgeschnitten. Dieser Datensatz wird im folgenden Text als *virtueller Tauchdatensatz* bezeichnet.
Aufgrund mangelnder Ressourcen können für den virtuellen Tauchdatensatz keine Segmentierungsmasken erstellt werden. Diese werden zum späteren Zeitpunkt für einen echten Unterwasserdatensatz erstellt, lohnen sich aber nicht für Daten, welche am Ende gar nicht zum Training des endgültigen Modells verwendet werden. Der virtuelle Tauchdatensatz dient nur einer groben Einschätzung der Nützlichkeit des Algorithmus im Unterwasserbereich. Die Qualität der Ergebnisse wird durch eine menschlich visuelle Einschätzung ermittelt.

Da die unüberwachte Segmentierung auf drei verschiedene Datensätze angewendet wird, unterteilt sich diese Arbeit auch in drei verschiedene Experimente, welche im Abschnitt 6 beschrieben werden. Diese Experimente sind stark voneinander abhängig, da derselbe Algorithmus verwendet wird, dennoch werden sie separat durchgeführt und ausgewertet.

5 Algorithmen

In diesem Abschnitt wird die technische Umsetzung der unüberwachten Segmentierung im Detail erklärt. Der folgende Text richtet sich an eine Leserschaft mit informatischen Vorkenntnissen, insbesondere im Bereich der Bildverarbeitung und des maschinellen Lernens.

5.1 Berechnung des optischen Flusses

Um den optischen Fluss des Davis Datensatzes zu errechnen wird ein vortrainiertes PWC-Netz [10] verwendet. Dieses wurde auf dem *FlyingChairs* [6] und *FlyingThings3D* [7] Datensatz trainiert. Hierbei handelt es sich um synthetische Datensätze welche Bildsequenzen und den optischen Fluss von verschieden, im Raum schwebenden Objekten beinhalten. Die Implementierung sowie die vortrainierten Gewichte des verwendeten PWC-Netzes sind frei verfügbar [5].

Empirisch hat sich gezeigt, dass sich das PWC-Netz nicht für Unterwasseraufnahmen eignet. Aus diesem Grund wird der optische Fluss des virtuellen Tauchdatensatzes mithilfe des Dual TV-L1 Algorithmus [11] berechnet. Hierbei wird die Implementierung der *Open Computer Vision* Programmbibliothek verwendet.

5.2 Veränderung des optischen Flusses

Um ideale Daten zu generieren, in denen die Bewegungsinformationen vom bewegten Objekt und seinem Hintergrund unabhängig sind, wird der optische Fluss des Davis Datensatzes mithilfe der Ground Truth Segmentierungsmaske künstlich verändert. Hierfür wird jeweils beim optischen Fluss des bewegten Objektes sowie beim Fluss des Hintergrundes, die Rotation und Vektorlänge durch die Addition eines Zufallswertes verändert.

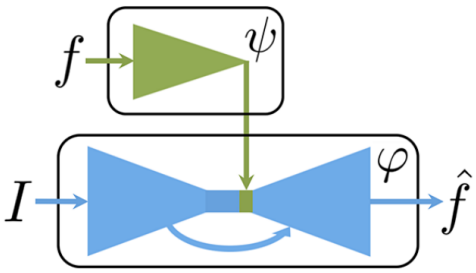

Abbildung 1: Das Rekonstruktions-Netz wird in zwei Teilnetze unterteilt. Hierbei wird der optische Fluss f separat vom Bild I in ψ kodiert. In φ wird das Bild kodiert und mithilfe des komprimierten Flusses, zum ursprünglichen optischen Fluss $\hat{f}$ rekonstruiert. [15]

Die Zufallswerte werden aus einer Verteilung von $\mathcal{U}[-1, 1]$ Radiant für die Rotation und $\mathcal{U}[-30, 30]$ Pixel für die Lageveränderung generiert.

5.3 Aufbau der neuronalen Netze

Generator-Netz: Als Generator-Netz wird eine kleinere Variante des SegNet [3] verwendet. Der Kodierungsteil besteht aus sechs Faltungsschichten, jeweils gefolgt von einer Batchnormalisierung. Diese reduzieren die Größe des Eingangsbildes auf ein Viertel der Originalgröße. Der Kodierungsteil wird gefolgt von vier geweiteten Faltungsschichten (engl.: dilated convolution) mit zunehmendem Radius (2, 4, 8, 16). Der Dekodierungsteil besteht aus sieben Faltungsschichten, die mithilfe von Upsampling Schichten die ursprüngliche Bildgröße wiederherstellen. Die jeweils zweite und vierte Schicht des Kodierungs- und Dekodierungsteils sind mit Skip-Verbindung verbunden. Die letzte Schicht des Generator-Netzes generiert, mithilfe einer Sigmoind-Aktivierung, die Segmentierungsmaske.

Rekonstruktions-Netz: Zum Rekonstruieren des optischen Flusses wird eine angepasste Variante des *Conditional Prior Net* (CP-Netz) [15] verwendet. Dieses besteht aus einem Kodierungsnetz ψ und einem Kodierungs-/Dekodierungsnetz φ (siehe Abbildung 1).
Diese Netzwerkarchitektur wurde dazu entwickelt, jeweils ein Bild I und den dazugehörigen optischen Fluss f, separat in einer komprimierten Form zu kodieren und anschließend die Kombination des komprimierten Bildes und Fluss, zum ursprünglichen optischen

Fluss $\hat{f}$ zu dekodieren. Durch eine getrennte Kodierung des optischen Flusses und des Bildes soll das Netz dazu bewegt werden den semantischen Kontext des Bildes zu lernen, welcher eine präzisere Rekonstruktion des optischen Flusses ermöglicht.

Aufbauend auf dieser Annahme, wird das CP-Netz in dieser Arbeit dazu verwendet einen nur teilweise vorhandenen optischen Fluss zu rekonstruieren. Hierfür wird der unvollständige optische Fluss sowie eine Maske, welche die Segmente des Flusses markiert die entfernt wurden, im Kodierungs-Netz ψ kodiert. Hierdurch sollen die vorhandenen Flussinformationen kompakt zusammengefasst werden. Das zum Fluss gehörende Bild wird in φ kodiert und zusammen mit der Flusskodierung zum ursprünglichen optischen Fluss $\hat{f}$ rekonstruiert.

Ausgenommen der veränderten Eingangswerte wurde die Architektur des CP-Netzes vollständig aus der Arbeit von Yang und Soatto [15] übernommen.

5.4 Verlustfunktionen

Verlustfunktionen des Rekonstruktions-Netzes: Die Verlustfunktion des Rekonstruktions-Netzes $r(I, f_m, m)$ misst, wie gut der optische Fluss f aus dem Bild I, dem maskierten Fluss $f_m = f \odot m$ und der Maske m rekonstruiert wird. Hierfür wird die Ähnlichkeit zweier Fluss-Matrizen benötigt, welche mit der Formel 1 berechnet wird. Diese misst den mittleren normierten und quadrierten Abstand zwischen den einzelnen Elementen der Fluss-Matrizen. Als Vektornorm wird hier die L2-Norm verwendet.

$$\delta(f^1, f^2) = \frac{1}{N} \sum_{n=0}^{N} ||f_n^1 - f_n^2||^2 \tag{1}$$

Der Verlust l_r des Rekonstruktions-Netzes misst sowohl die Qualität der Flussrekonstruktion des mit der Maske m maskierten Flusses $f_{r,m}$, als auch die Flussrekonstruktion des mit der komplementären Maske $m_c = 1 - m$ maskierten Flusses f_{r,m_c}, sowie die Qualität der leeren Flussrekonstruktion $f_{r,m_0} = r(I, f_0, m_0)$.

$$l_r = \delta(f, f_{r,m}) + \delta(f, f_{r,m_c}) + \delta(f, f_{r,m_0}) \tag{2}$$

Die leere Flussrekonstruktion entsteht, indem das Rekonstruktions-Netz den optischen Fluss ausschließlich aus dem Bild I rekonstruiert. Der Fluss f_0 und die Maske m_0 beinhalten nur nullen und somit keine Informationen. Das Verwenden der leeren Flussrekonstruktion soll einen Anreiz zum Lernen des semantischen Kontextes des Bildes geben.
Hierbei ist wichtig, dass beim Verwenden der leeren Flussrekonstruktion genügend Trainingsdaten vorhanden sind oder Datenvermehrungsmethoden auf den optischen Fluss angewendet werden. Andernfalls lernt das Rekonstruktions-Netz den optischen Fluss der Bilder auswendig.

Verlustfunktionen des Generator-Netzes: Der Verlust l_g des Generator-Netzes ergibt sich aus dem Quotienten der Qualität der Flussrekonstruktion und der Qualität der leeren Flussrekonstruktion. Wie auch bei der Verlustfunktion des Rekonstruktions-Netzes fließen sowohl die Rekonstruktionsqualität des maskierten Flusses (Formel 3 erster Term), als auch die des komplementär maskierten Flusses (Formel 3 zweiter Term) in die Verlustfunktion mit ein. Anders als bei der Verlustfunktion des Rekonstruktions-Netzes wird allerdings nicht der gesamte Fluss miteinander verglichen, sondern nur diejenigen Teile, welche von der Maske maskiert und vom Rekonstruktions-Netz wiederhergestellt wurden.

$$l_g = \left(1 - \frac{\delta(f_m, f_{r,m})}{\delta(f_m, f_{r,m_0})}\right) + \left(1 - \frac{\delta(f_m, f_{r,m_c})}{\delta(f_m, f_{r,m_0})}\right) \tag{3}$$

Unter der Voraussetzung, dass das Rekonstruktions-Netz gut funktioniert, approximiert die Funktion 3, die Menge der für die Rekonstruktion hilfreichen Informationen, im maskierten Fluss.
Schafft es das Generator-Netz Beispielsweise den Fluss eines sich bewegenden Objektes zu segmentieren, welches keine Bewegungsinformationen mit seiner Umgebung teilt, so hilft der vorhandene Fluss dem Rekonstruktions-Netz nicht dabei, den maskierten Fluss zu rekonstruieren. In diesem Fall müsste die Qualität der Flussrekonstruktion $\delta(f_m, f_{r,m})$ gleich sein mit der Qualität der leeren Flussrekonstruktion $\delta(f_m, f_{r,m_0})$.

5.5 Validierungsmetriken

Um die Qualität der Segmentierungen zu quantifizieren wird die Überlappung zwischen der Ground Truth Segmentierungsmaske und der vom Generator-Netz erzeugten Maske

berechnet. Hierfür wird sowohl der Dice-Koeffizient DIC als auch der Jaccard-Koeffizient (oder Jaccard-Index) J verwendet. Beide Metriken messen die Überlappung von zwei Flächen auf sehr ähnliche Weise:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{4}$$

$$DIC(A, B) = \frac{2|A \cap B|}{|A| + |B|} \tag{5}$$

Die Experimente wurden anfangs mit dem Dice-Koeffizient ausgeführt, später wurde der Jaccard-Koeffizient hinzugenommen um eine bessere Vergleichbarkeit mit der Arbeit von Yang et al. [14] herzustellen.

6 Experimente

In diesem Abschnitt wird das Training und die Resultate der unüberwachten Segmentierung auf dem David Datensatz, mit und ohne Veränderung des optischen Flusses und dem virtuellen Tauchdatensatz beschrieben.

6.1 Experiment 1: Training auf dem David Datensatz mit Veränderung des optischen Flusses

Um die Qualität der unüberwachten Segmentierung zu überprüfen, wird diese auf dem Davis Datensatz trainiert und validiert. Genau wie in der als Vorlage dienenden Arbeit von Yang et al. [14], wird der optische Fluss mit einem PWC-Netz [10] errechnet und anschließend mithilfe der Ground Truth Segmentierungsmaske verändert (siehe Abschnitt 5.2), um eine Unabhängigkeit der Bewegungsinformationen von Vorder- und Hintergrund zu gewährleisten.

Das Training des Rekonstruktions- und Generator-Netzes stellte nach einigem anpassen der Hyperparameter keine Schwierigkeiten dar.

Die Resultate von Yang et al. [14] konnten im Experiment leider nicht vollständig reproduziert werden. Yang et al. [14] erreichte in seiner Arbeit einen Jaccard-Index von

Abbildung 2: Diese Abbildung zeigt besonders gute Ergebnisse der unüberwachten Segmentierung auf dem Davis Validierungsdatensatz mit Veränderung des optischen Flusses.

0.715 auf dem Davis Validierungsdatensatz, wohingegen in diesem Experiment nur ein Jaccard-Index von 0.565 erreicht wurde.

Bei einer genaueren Analyse der Daten hat sich herausgestellt, dass die unüberwachte Segmentierung bei vielen Videosequenzen sehr gute Ergebnisse liefert (siehe Abbildung 2), während die Ergebnisse bei anderen Videosequenzen auffallend schlecht ausfallen. Im Durchschnitt ergibt sich hierdurch ein mittlerer Jaccard-Index.

Schlechte Ergebnisse liefert die unüberwachte Segmentierung vor allem bei bewegten Wasseroberflächen und verschwommenen Bildern (siehe Abbildung 3 links). Dies scheint an dem Informationsgehalt und der Qualität des optischen Flusses zu liegen. Durch Bewegungen und Spiegelungen im Wasser entstehen sehr viele unterschiedliche Bewegungsinformationen, die schwer aus der umgebenden Bewegung vorherzusagen sind und dementsprechend vom Generator-Netz segmentiert werden. Bei verschwommenen Bildern hat das PWC-Netz Schwierigkeiten einen korrekten optischen Fluss zu errechnen. Auch bei Flächen mit geringer Struktur (Straßenoberflächen, diesigem Himmel oder dunklen Flächen) berechnet das PWC-Netz Bewegungen, welche aus menschlicher Sicht wenig Sinn ergeben (siehe Abbildung 3 Mitte).

Hinzu kommen Eigenschaften der unüberwachten Segmentierung, welche sich negativ auf den Jaccard-Index auswirken, allerdings nicht als Fehler zu werten sind. Hierzu gehört beispielsweise das Segmentieren des Schattens oder der Spiegelung eines bewegten Objektes (siehe Abbildung 3 rechts). Schatten oder Spiegelungen sind nicht in der Ground

Abbildung 3: In den drei Abbildungen werden verschiedene Schwächen des Generator- und PWC-Netzes gezeigt. Die oberen Bilder zeigen jeweils ein Bild inklusive Segmentierung, währen das darunterliegende Bild den zugehörigen optischen Fluss zeigt.
Links ist der schwer zu interpretierende optische Fluss bei bewegten Wasseroberflächen abgebildet. In der Mitte ist der nicht passende optische Fluss gezeigt, den das PWC-Netz für Flächen mit wenig Struktur berechnet. Rechts ist die Eigenschaft des Generator-Netzes zu sehen, Schatten oder Spiegelungen von bewegtem Objekt mit zu segmentieren.

Truth Segmentierungsmaske enthalten, dennoch ist es korrekt diese zu segmentieren, da sie sich viele gemeinsame Bewegungsinformationen mit dem bewegten Objekt teilen.

Dieses Experiment hat gezeigt, dass die unüberwachte Segmentierung funktionsfähig ist und bereits einige Schwächen offengelegt. Inwieweit sie sich für reale Daten eignet, wird im nächsten Experiment erforscht.

6.2 Experiment 2: Training auf dem David Datensatz ohne Veränderung des optischen Flusses

Das erste Experiment zeigt zwar, dass die unüberwachte Segmentierung funktioniert, doch aufgrund der Veränderung des optischen Flusses, mithilfe der Ground Truth Segmentierungsmaske, lässt es keine Schlussfolgerung darüber zu, inwieweit die Überlegungen aus Abschnitt 3 auf reale Daten zutreffen.
In diesem Experiment wird die unüberwachte Segmentierung auf dem Davis Datensatz, ohne künstliche Trennung der Bewegungsinformationen von Vorder- und Hintergrundobjekten angewendet. Um Überanpassung zu verhindern, wird der optische Fluss dennoch

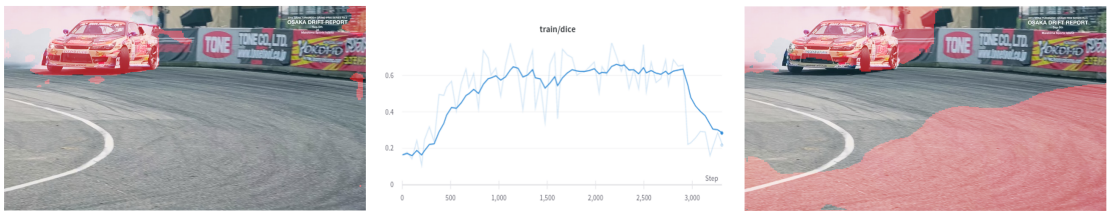

Abbildung 4: Diese Abbildungen zeigen die Instabilität beim Trainieren des Generator-Netzes. Das linke Bild zeigt die Segmentierung im stabilen Zustand und das Rechte die Segmentierung kurz nach dem plötzlichen Qualitätsverlust. Mittig ist der Dice-Koeffizient im Laufe des Trainings abgebildet.

zufällig verändert, doch ohne Verwendung der Ground Truth Segmentierungsmaske. Die hierbei entstehenden Resultate sollen eine Einschätzung geben, wie gut sich die unüberwachte Segmentierung auf reale Daten anwenden lässt.

Im Gegensatz zum Experimenten 1 hat sich das Training der neuronalen Netze in diesem Experiment als schwierig herausgestellt. Während das Rekonstruktions-Netz leicht zu trainieren war und gute Ergebnisse liefert, wies das Generator-Netz Instabilitäten auf. Diese Instabilitäten zeigen sich durch eine schlagartige Erhöhung des Generator Verlustes, welche mit einer Verschlechterung des Dice-Koeffizienten einhergeht (siehe Abbildung 4). Nach dieser plötzlichen Veränderung weist das Generator-Netz ungefähr die Hälfte der Pixel den Hintergrund und die andere Hälfte dem Vordergrund zu. Die bewegten Objekte werden noch erkannt, doch zusätzlich werden ihnen große Bereiche des Hintergrundes zugeordnet.

Es ist leider nicht gelungen die Ursache der Instabilität zu finden. Da sowohl das Verkleinern des Generator-Netzes als auch das Hinzufügen von normalverteiltem Rauschen zu den Gradienten des Generator-Netzes die Instabilität verbessern, wird vermutet, dass es sich um eine Form von Überanpassung handelt.

Nach dem das Problem der Instabilität beseitigt wurde, erreichte dieses Experiment ein Jaccard-Index von 0.438.
Bei einer genaueren Datenanalyse stellt sich heraus, dass sich viele der Beispiele des Davis Datensatzes nicht für die unüberwachte Segmentierung eignen, da (zusätzlich zu den in Experiment 1 beschriebenen Problemen) keine vollständige Unabhängigkeit der Bewegungsinformationen zwischen Vorder- und Hintergrund besteht. Zwar zeigt der Davis Datensatz verschiedene Sequenzen von bewegten Objekten, doch gibt ed immer wieder Abschnitte, in denen sich das Vordergrundobjekt nicht oder nicht vollständig bewegt.

Besonders auffällig zeigt sich dies bei den Aufnahmen von großen behäbigen Tieren wie die des Bären, Elefanten, Nashorn und Kamel, bei denen sich häufig nur ein bis zwei Beine bewegen während die Gesamtkörperbewegung nur sehr langsam ist. In diesen Fällen werden nur die bewegten Körperteile segmentiert, welche einen Bruchteil der Ground Truth Segmentierungsmaske darstellen.
Zudem hat sich gezeigt, dass die neuronalen Netze, ohne Veränderung des optischen Flusses, deutlich stärker von den Schwächen des PWC-Netzes beeinflusst werden.
Bei Datenbeispielen, in denen sich die Bewegungen des Vordergrundobjektes deutlich von dem des Hintergrunds abheben, lieferte die unüberwachte Segmentierung ähnlich gute Ergebnisse wie im Experiment 1.

Dieses Experiment zeigt, dass sich die unüberwachte Segmentierung nur für sehr bestimmte Daten eignet. Ob schwimmende Fische dazugehören wird im nächsten Experiment erforscht.

6.3 Experiment 3: Training auf dem virtuellen Tauchdatensatz

Für eine grobe Einschätzung, inwieweit sich die unüberwachte Segmentierung für die Segmentierung von Fischen eignet, wird der Algorithmus auf dem virtuellen Tauchdatensatz getestet. Aufgrund der Schwächen des PWC-Netzes beim Berechnen des optischen Flusses von Flächen mit wenig Konturen (siehe Ergebnisse aus Experiment 1), welche Unterwasser ein Großteil des Hintergrundes ausmachen, hat sich das PWC-Netz als ungeeignet herausgestellt. Stattdessen wird in diesem Experiment der Dual TV-L1 Algorithmus [11] zum Berechnen des optischen Flusses verwendet.

Wie auch im Experiment 2 weist das Generator-Netz Instabilitäten im Training auf. Diese äußern sich durch das Segmentieren des gesamten Vorder- beziehungsweise Hintergrundes oder durch das Segmentieren des immer gleichen Bildbereiches. Genau wie im Experiment 2 konnte das Problem durch das Hinzufügen von normalverteiltem Rauschen, zu den Gradienten des Generator-Netzes, gelöst werden.

Nachdem das Problem der Instabilität behoben war, lieferte das unüberwachte Segmentieren gute Ergebnisse. Da es für den virtuellen Tauchdatensatz keine Ground Truth Segmentierungsmasken gibt, können hier keine quantitativen Qualitätsaussagen gemacht werden. Die manuelle Evaluierung der Trainings- und Validierungsdaten zeigt aber, dass das Generator-Netz, sich fortbewegende Fische sicher segmentieren kann (siehe Abbildung 5).

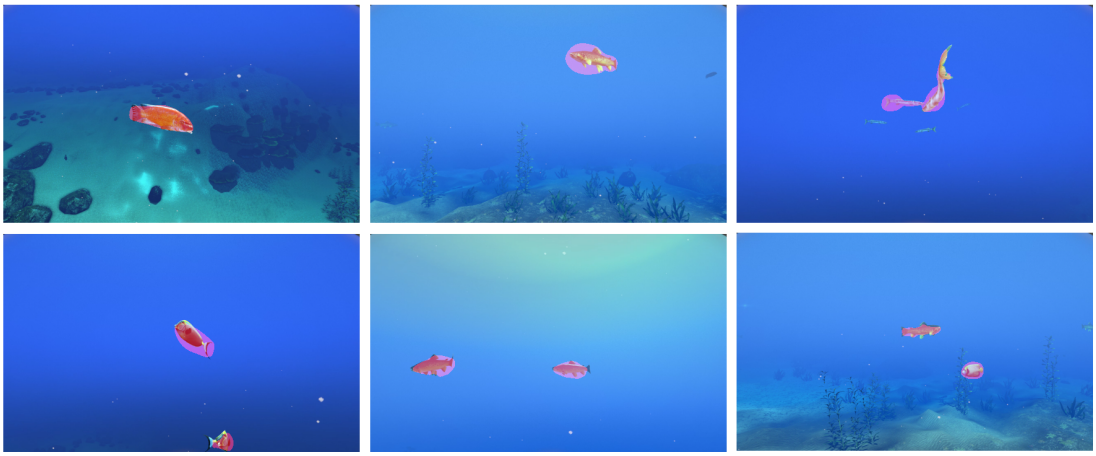

Abbildung 5: In dieser Abbildung sind erfolgreiche Segmentierungen von Fischen des virtuellen Tauchdatensatzes zu sehen.

Sogar Fische die sich nur sehr langsam fortbewegen, werden vom Generator-Netz segmentiert, auch wenn die Bewegung anhand des optischen Flusses kaum ersichtlich ist (siehe Abbildung 6 links). Fische die sich nicht bewegen werden allerdings nicht segmentiert (siehe Abbildung 6 mittig und rechts).
Aufnahmen bei denen Fische nahe über dem Grund schwimmen stellen eine Schwierigkeit dar, können aber größtenteils gut segmentiert werden. Erst bei stark bewegten Aufnahmen mit sehr unübersichtlichem optischen Fluss werden zusätzlich zu den Fischen oft Wasserpflanzen oder Steine segmentiert (siehe Abbildung 7). Bei Aufnahmen mit mehreren Fischen fällt auf, dass oft nicht alle schwimmenden Fische segmentiert werden (siehe Abbildung 8).

Schwimmende Fische scheinen sich sehr gut für eine unüberwachte Segmentierung zu eignen. Eine pixelgenaue Segmentierung ist zwar noch nicht möglich, könnte aber durch eine Verbesserung des optischen Flusses erreicht werden. Bei der hier verwendeten Parametrisierung des Dual TV-L1 Algorithmus geht der optische Fluss von Fischen häufig über dessen Konturen hinaus, wodurch es für den Generator zweckmäßig ist die Segmentierung des Fisches zu vergrößern, um die Bewegungsinformationen des Fisches vollständig zu verdecken.

Eine Verbesserung des unüberwachten Segmentierens dahingehend, dass alle bewegten Fische im Bild segmentiert werden ist schwierig, aufgrund der Eigenschaften des Algorithmus. Sei ein schwimmender Fisch perfekt segmentiert, so ist es auf Basis seiner

Abbildung 6: Diese Abbildung zeigt die Segmentierung von Fischen mit verschiedenen Bewegungsgeschwindigkeiten. Während schnell und langsam schwimmende fische segmentiert werden, können nicht schwimmende Fische nicht segmentiert werden.

Abbildung 7: Eine Schwierigkeit stellt das Segmentieren von bewegten Aufnahmen nahe dem Grund dar. Doch erst bei starken Bewegungen werden zusätzlich zu den Fischen, Algen und Steine segmentiert.

Abbildung 8: In dieser Abbildung sind die Segmentierungen von mehreren Fischen zu sehen. Hierbei fällt auf, dass oft nicht alle sich bewegenden Fische segmentiert werden.

Bewegung nicht möglich die Bewegungen des Hintergrunds oder anderer Fische vorherzusagen. Somit ist es für das Generator-Netz nicht von Bedeutung, ob ein oder mehrere schwimmende Fische segmentiert werden.

7 Fazit

Die Methode der unüberwachten Segmentierung eignet sich nur für sehr bestimmte Daten. In den Alltagsszenarien des Davis Datensatzes sind die Bewegungen des Hintergrundes oft zu komplex und unvorhersehbar, als das Objekte, die aus menschlicher Sicht zum Vordergrund gehören, unüberwacht segmentiert werden können. Auch ist eine Trennung der Bewegungsinformationen von Vorder- und Hintergrundobjekten nicht immer gegeben.

Anders verhält es sich bei den Unterwasserszenarien des virtuellen Tauchdatensatzes. In diesem Datensatz sind Fische und Säugetiere die einzigen beweglichen Objekte. Zudem sind Großteile des Hintergrundes aufgrund der trüben Sicht verdeckt. Hierdurch sind die Bewegungsinformationen von schwimmenden Fischen getrennt von denen des Hintergrundes, wodurch sich der Datensatz sehr gut für eine unüberwachte Segmentierung eignet.

Diese Ergebnisse lassen hoffen, dass sich die unüberwachte Segmentierung auch für Videoaufnahmen aus der Ostsee gut eignet.

Mithilfe eines Validierungsdatensatzes kann die unüberwachte Segmentierung weiter auf Unterwasserdaten optimiert werden. Vor allem der optische Fluss ist ausschlaggebend für die Qualität des Algorithmus.
Diese Arbeit hat sich nur sehr wenig mit der Berechnung des optischen Flusses beschäftigt. Eine Optimierung dieser Komponente könnte die Ergebnisse deutlich verbessern.

8 Ausblick

In den nächsten Schritten des AuTag BeoFisch Projektes wird mithilfe eines BlueROV 2 ein Unterwasserdatensatz erstellt. Um die Qualität von Algorithmen validieren zu können, wird ein kleiner Teil des Datensatzes manuell segmentiert. Der Rest der Daten kann zum Trainieren von unüberwachten Lernalgorithmen, wie der unüberwachten Segmentierung, verwendet werden.

Da das Berechnen des optischen Flusses sehr rechenaufwändig ist, eignet sich die unüberwachte Segmentierung nicht für einen direkten Einsatz auf dem BlueROV 2. Aus diesem Grund wird versucht das Generator-Netz darauf zu trainieren, nur mithilfe eines einzelnen Bildes, die beweglichen Objekte zu segmentieren. Da das Generator-Netz von der Feature-Relevanz des Rekonstruktions-Netzes lernt, könnte es die Verbindung zwischen visuellen Merkmalen und Bewegung erlernen. In diesem Fall können Generator-Netze trainiert werden, die klein genug sind, um auf mobilen Systemen des BlueROV 2 zu laufen und sowohl schwimmende als auch nicht schwimmende Fischen segmentieren können.
Damit dies gelingt muss allerdings der Trainingsdatensatz dahingehend angepasst werden, dass es eine direkte Verbindung zwischen Bewegung und visuellen Merkmalen gibt. Dies bedeutet, dass der optische Fluss möglichst pixelgenau ist und keine nicht schwimmenden Fische im Datensatz vorkommen.

Falls dieser Ansatz nicht gelingt, kann mithilfe der unüberwachten Segmentierung ein Trainingsdatensatz erstellt werden, mit dessen Hilfe die visuellen Merkmale von Fischen durch überwachte Lernverfahren erlernt werden können.

Literatur

[1] *EyeSea Underwater Video*. https://data.pnnl.gov/dataset/12978. – Accessed: 2020-08-10

[2] *Fish Detection and Tracking*. http://www.perceivelab.com/dataset/Fish%20Detection%20and%20Tracking. – Accessed: 2020-08-11

[3] Badrinarayanan, Vijay ; Kendall, Alex ; Cipolla, Roberto: SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. In: *CoRR* abs/1511.00561 (2015). – URL http://arxiv.org/abs/1511.00561

[4] Boom, Bas ; He, Jiyin ; Palazzo, Simone ; Huang, Phoenix ; Beyan, Cigdem ; Chou, Hsiu-Mei ; Lin, Fang-Pang ; Spampinato, Concetto ; Fisher, Robert: A research tool for long-term and continuous analysis of fish assemblage in coral-reefs using underwater camera footage. In: *Ecological Informatics* 23 (2014), 09, S. 83–97

[5] *Optical Flow Prediction with Tensorflow*. – URL https://github.com/philferriere/tfoptflow. – Eingesehen am 15.09.2020

[6] Dosovitskiy, A. ; Fischer, P. ; Ilg, E. ; Häusser, P. ; Hazirbaş, C. ; Golkov, V. ; Smagt, P. v.d. ; Cremers, D. ; Brox, T.: FlowNet: Learning Optical Flow with Convolutional Networks. In: *IEEE International Conference on Computer Vision (ICCV)*, URL http://lmb.informatik.uni-freiburg.de/Publications/2015/DFIB15, 2015

[7] Mayer, N. ; Ilg, E. ; Häusser, P. ; Fischer, P. ; Cremers, D. ; Dosovitskiy, A. ; Brox, T.: A Large Dataset to Train Convolutional Networks for Disparity, Optical Flow, and Scene Flow Estimation. In: *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, URL http://lmb.informatik.uni-freiburg.de/Publications/2016/MIFDB16, 2016. – arXiv:1512.02134

[8] Ochs, Peter ; Malik, Jitendra ; Brox, Thomas: Segmentation of Moving Objects by Long Term Video Analysis. In: *IEEE transactions on pattern analysis and machine intelligence* 36 (2013), 12

[9] Perazzi, F. ; Pont-Tuset, J. ; McWilliams, B. ; Van Gool, L. ; Gross, M. ; Sorkine-Hornung, A.: A Benchmark Dataset and Evaluation Methodology for Video Object Segmentation. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, S. 724–732

[10] Sun, Deqing ; Yang, Xiaodong ; Liu, Ming-Yu ; Kautz, Jan: PWC-Net: CNNs for Optical Flow Using Pyramid, Warping, and Cost Volume. In: *CoRR* abs/1709.02371 (2017). – URL http://arxiv.org/abs/1709.02371

[11] Sánchez, Javier ; Meinhardt-Llopis, Enric ; Facciolo, Gabriele: TV-L1 optical flow estimation. In: *Image Processing On Line* 3 (2013), 07, S. 137–150

[12] Tsai, David ; Flagg, Matthew ; Rehg, James: Motion Coherent Tracking Using Multi-label MRF Optimization, 01 2010, S. 1–11

[13] Xu, Wenwei ; Matzner, Shari: Underwater Fish Detection using Deep Learning for Water Power Applications. In: *CoRR* abs/1811.01494 (2018). – URL http://arxiv.org/abs/1811.01494

[14] Yang, Yanchao ; Loquercio, Antonio ; Scaramuzza, Davide ; Soatto, Stefano: Unsupervised Moving Object Detection via Contextual Information Separation. In: *CoRR* abs/1901.03360 (2019). – URL http://arxiv.org/abs/1901.03360

[15] Yang, Yanchao ; Soatto, Stefano: Conditional Prior Networks for Optical Flow. In: *CoRR* abs/1807.10378 (2018). – URL http://arxiv.org/abs/1807.10378