Schlussbericht

Verbundprojekt:
Cyber-Sicherheit für hochautomatisierte Systeme und das autonome Fahren - SECREDAS

Teilvorhaben:
Entwicklung und Validierung robuster Bildsegmentierungsalgorithmen im Kontext des autonomen Fahrens

Zuwendungsgeber:	Bundesministerium für Bildung und Forschung
Zuwendungsnehmer:	Merantix AG
Förderkennzeichen:	16ESE0322
Laufzeit:	01.06.2018 - 31.10.2021
Autoren:	Sebastian Gerres, Alp Aribal, Adrian Loy

GEFÖRDERT VOM

Bundesministerium
für Bildung
und Forschung

ECSEL
Joint Undertaking

Berlin, März 2022

Inhaltsverzeichnis

I. Kurze Darstellung

1 Aufgabenstellung

Autonome Systeme sind einer Vielzahl von Bedrohungen ausgesetzt, die deren Sicherheit und/oder Leistung beeinträchtigen können. Mutwillige Angriffe auf Sensoren (so genanntes „Sensor-Spoofing“) wurden im Vorhaben SECREDAS als eines von zehn zentralen Bedrohungsszenarien, für die Lösungen entwickelt werden sollten, definiert.

Im Bereich der Bilderkennung (Sensor = Kamera) hat die Technologie des Maschinellen Lernen in den letzten Jahren erstaunliche Ergebnisse mit einer bemerkenswerten Performanz gezeigt. Mit Hilfe neuronaler Netze können Eingabedaten wie Videostreams erfolgreich verarbeitet und z.B. Objekte erkannt werden. Angriffe auf Sensoren – insbesondere jene Angriffe, bei denen kein visueller Unterschied zwischen dem Original- und manipulierten Bild erkennbar ist – schränken die Fähigkeit der Algorithmik, z.B. andere Fahrzeuge zu erkennen, jedoch stark ein. Dies kann beim Einsatz von autonomen Systemen z.B. im Verkehrsbereich (innerhalb von SECREDAS wurden die drei ausgewählte Anwendungsdomänen Automotive, Rail und Health betrachtet) wie dem autonomen Fahren zu lebensgefährlichen Situationen führen. Daher ist es von größter Bedeutung, die Robustheit solcher intelligenter Verfahren für den Einsatz auf realen Systemen und im Produktivbetrieb sicherzustellen.

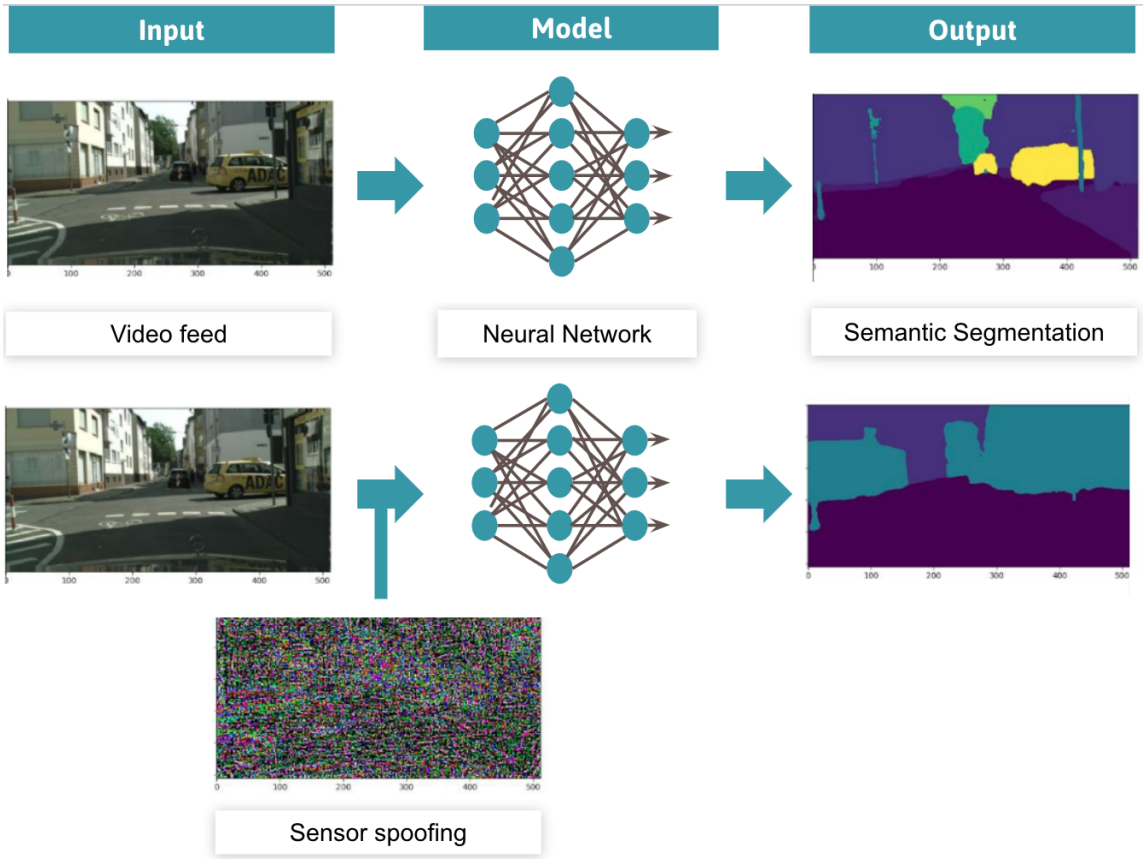


Abbildung 1: Semantische Segmentierung durch Neuronale Netzwerke und der Effekt von Sensor Spoofing

Die Motivation für die Forschungsarbeit der Merantix AG ist es, robuste Methoden für die intelligente Bilderkennung zu entwickeln. Im Bereich der Bilderkennung hat maschinelles Lernen, v.a. modelliert in Form von neuronalen Netzen, in den letzten Jahren erstaunliche Ergebnisse gezeigt. Für die Anwendung auf reale Systeme - wie dem des autonomen Fahrens - ist es jedoch wichtig, nicht nur deren Genauigkeit, sondern auch deren Robustheit zu betrachten. Also beispielsweise den Umgang der Algorithmik mit Daten, auf die nicht trainiert wurde und welche vorab nicht bekannt sind.

Zusammengefasst lag die Aufgabenstellung der Merantix AG im eigenen Teilvorhaben für das Verbundprojekt SECREDAS also in der Entwicklung und Validierung robuster Bildsegmentierungsalgorithmen im Kontext des autonomen Fahrens.

2 Voraussetzungen, unter denen das Vorhaben durchgeführt wurde

Merantix konnte für das Vorhaben SECREDAS auf eine Reihe von Vorarbeiten zurückgreifen. Mit dem Tool “Picasso” verfügte Merantix zu Projektstart über ein Tool, welches neuronale Netzwerke versteht, verbessert und vergleicht und dadurch die Robustheit und die Sicherheit dieser Algorithmen erhöht.[1] Dieses Tool wurde speziell für den Automobilmarkt und die Entwicklung autonomer Fahrsysteme entwickelt. Picasso ist in der Lage, Fehleinschätzungen der Algorithmen zu erkennen, Klassifizierungen zu einzelnen Stichproben zu visualisieren und zu verstehen, Trainingssätze mit schwierigen oder widersprüchlichen Stichproben zu erweitern und die Leistung der Algorithmen auf erweiterten Daten zu quantifizieren. Darüber hinaus konnte Merantix zu Projektstart auf eine umfassende Expertise im Bereich des maschinellen Lernens zurückgreifen, welche durch eine Vielzahl von Publikationen dokumentiert ist.

In dem Vorhaben SECREDAS hat Merantix erstmalig in einem größeren Forschungskonsortium mitgewirkt. Expertise in der Durchführung von nationalen und europäischen Forschungsprojekten lag innerhalb von Merantix zu Projektstart bereits bei einer Vielzahl der Mitarbeitenden - die zu einem Gros ihre universitäre Laufbahn mit einer Promotion abgeschlossen hatten - vor.

Für eine erfolgreiche Projektdurchführung wurden vor Projektstart eine Reihe von Annahmen formuliert und mit den Projektpartnern abgestimmt. So war Merantix u.a. auf die Bereitstellung von Videostreams durch Projektpartner, welche über eigene Fahrzeuge mit eigener Sensorik verfügen, angewiesen.

Für die Durchführung der Entwicklungsarbeiten war Merantix auf ausreichende Speicher- und Rechenkapazitäten angewiesen. Diese hält Merantix nur in sehr begrenztem Bedarf in-house vor und kauft diese zumeist als Service von Dritten ein. Aufgrund einer langjährigen strategischen Partnerschaft mit Google[2] setzt Merantix hier auf die Dienste der Google Cloud Platform (GCP).

---

[1] Ryan Henderson and Rasmus Rothe, “Picasso: A Modular Framework for Visualizing the Learning Process of Neural Network Image Classifiers”, arXiv preprint arXiv:1705.05627, May 2017

[2] https://cloud.google.com/customers/merantix

3 Planung und Ablauf des Vorhabens

Die inhaltliche und organisatorische Koordination des Vorhabens fand auf drei Ebenen statt: der Gesamtprojektebene, der Arbeitspaketebene (Work packages, WP) und der Aufgaben (Tasks). Der Austausch mit den Projektpartnern fand im Kern in Form regelmäßig stattfindenden Telefonkonferenzen (Frequenz: wöchentlich bis monatlich) zwischen den beteiligten Partnern statt.

Wenige Monate nach der Projekthalbzeit begann die COVID-19-Pandemie. Dies führte dazu, dass einzelne Projektpartner nicht mehr in der Lage waren, ihre Arbeiten (und Zulieferungen an andere Projektpartner) im geplanten Rahmen durchzuführen. Darüber hinaus konnten die Vorbereitungen für die Abschlussdemonstration inklusive Integrationstests nicht zeitgerecht durchgeführt werden. Die Auswirkungen beider Aspekte haben sich auch auf die Arbeiten von Merantix niedergeschlagen: Relevante Informationen von Projektpartnern wurden nicht bzw. nur stark verzögert geliefert (z.B. Information über die API der OBU durch den Projektpartner Commsignia). Integrationstests für die Demonstration am Standort Helmond fanden nicht statt. Die Projektleitung seitens NXP hat daher im Dezember 2020 gegenüber der Europäischen Kommission eine Projektverlängerung um sechs Monate bis Oktober 2021 beantragt. Merantix hat an der Vorbereitung der Abschlussdemonstration, die auf den Juni 2021 verschoben wurde, mitgewirkt, konnte aber aufgrund der unsicheren pandemischen Lage im Frühsommer 2021 selbst nicht vor Ort sein. An der Abschlusskonferenz im Oktober 2021 in Helmond (Niederlande) nahm das Team von Merantix teil.

Die Verzögerungen im Projektablauf hatten auch - wenngleich nur geringe - Auswirkungen auf die zeitgerechte Fertigstellung der Deliverables (siehe Tabelle 2).

4 Wissenschaftlicher und technischer Stand, an den angeknüpft wurde

Über die gesamte Projektlaufzeit hinweg wurde der aktuelle Stand der Wissenschaft und Technik gemonitort und relevante Forschungsarbeiten in SECREDAS aufgegriffen. Das Feld des Maschinellen Lernens hat in den vergangenen Jahren seitens der Wissenschaft und Wirtschaft eine enorme Aufmerksamkeit auf sich gezogen, die sich u.a. in einer Vielzahl von Publikationen widerspiegelt. Maßgeblich mit Hilfe der Plattform arXiv hat sich Merantix fortlaufend einen Überblick zu den für die Arbeiten insbesondere in WP4 verschafft.

Ausgewählte, für die Forschungsarbeiten von Merantix relevante Arbeiten sind nachfolgend dargestellt.

Bilderkennung und -segmentierung

S. Mehta, M. Rastegari, A. Caspi, L. Shapiro, and H. Hajishirzi, “ESPNet: Efficient Spatial Pyramid of Dilated Convolutions for Semantic Segmentation,” ArXiv180306815 Cs, Jul. 2018, [Online]. Available: http://arxiv.org/abs/1803.06815.

L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, “DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs,” ArXiv160600915 Cs, May 2017, [Online]. Available: http://arxiv.org/abs/1606.00915.

L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, “Rethinking Atrous Convolution for Semantic Image Segmentation,” ArXiv170605587 Cs, Dec. 2017, [Online]. Available: http://arxiv.org/abs/1706.05587.

K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” ArXiv151203385 Cs, Dec. 2015, [Online]. Available: http://arxiv.org/abs/1512.03385.

M. Cordts et al., “The Cityscapes Dataset for Semantic Urban Scene Understanding,” in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, Jun. 2016, pp. 3213–3223, doi: 10.1109/CVPR.2016.350.

Anomaliedetektion

I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and Harnessing Adversarial Examples,” ArXiv14126572 Cs Stat, Mar. 2015, [Online]. Available: http://arxiv.org/abs/1412.6572.

F. Yu, V. Koltun, and T. Funkhouser, “Dilated Residual Networks,” ArXiv170509914 Cs, May 2017, [Online]. Available: http://arxiv.org/abs/1705.09914.

M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, “MobileNetV2: Inverted Residuals and Linear Bottlenecks,” ArXiv180104381 Cs, Mar. 2019, [Online]. Available: http://arxiv.org/abs/1801.04381.

S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, “Aggregated Residual Transformations for Deep Neural Networks,” ArXiv161105431 Cs, Apr. 2017, [Online]. Available: http://arxiv.org/abs/1611.05431.

S. Zagoruyko and N. Komodakis, “Wide Residual Networks,” ArXiv160507146 Cs, Jun. 2017, [Online]. Available: http://arxiv.org/abs/1605.07146.

A. Nguyen, J. Yosinski, and J. Clune, “Deep neural networks are easily fooled: High confidence predictions for unrecognizable images,” in 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, Jun. 2015, pp. 427–436, doi: 10.1109/CVPR.2015.7298640.

K. Eykholt et al., “Robust Physical-World Attacks on Deep Learning Models,” ArXiv170708945 Cs, Apr. 2018, [Online]. Available: http://arxiv.org/abs/1707.08945.

Bewertung der Robustheit

C. Shorten and T. M. Khoshgoftaar, “A survey on Image Data Augmentation for Deep Learning,” J. Big Data, vol. 6, no. 1, p. 60, Dec. 2019, doi: 10.1186/s40537-019-0197-0.

A. Antoniou, A. Storkey, and H. Edwards, “Data Augmentation Generative Adversarial Networks,” ArXiv171104340 Cs Stat, Mar. 2018, [Online]. Available: http://arxiv.org/abs/1711.04340.

C. Bowles et al., “GAN Augmentation: Augmenting Training Data using Generative Adversarial Networks,” ArXiv181010863 Cs, Oct. 2018, [Online]. Available: http://arxiv.org/abs/1810.10863.

E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le, “AutoAugment: Learning Augmentation Strategies From Data,” Jun. 2019.

C. Szegedy et al., “Intriguing properties of neural networks,” ArXiv13126199 Cs, Feb. 2014, [Online]. Available: http://arxiv.org/abs/1312.6199.

C. M. Bishop, “Novelty detection and neural network validation,” IEE Proc.-Vis. Image Signal Process., vol. 141, no. 4, pp. 217–222, 1994.

D. Hendrycks and T. G. Dietterich, “Benchmarking Neural Network Robustness to Common Corruptions and Surface Variations,” ArXiv180701697 Cs Stat, Apr. 2019, [Online]. Available: http://arxiv.org/abs/1807.01697.

Robustifizierung von Algorithmen

Z.-H. Zhou, J. Wu, and W. Tang, “Ensembling neural networks: Many could be better than all,” Artif. Intell., vol. 137, no. 1–2, pp. 239–263, May 2002, doi: 10.1016/S0004-3702(02)00190-X.

M. Amin-Naji, A. Aghagolzadeh, and M. Ezoji, “Ensemble of CNN for multi-focus image fusion,” Inf. Fusion, vol. 51, pp. 201–214, Nov. 2019, doi: 10.1016/j.inffus.2019.02.003.

J. Guo and S. Gould, “Deep CNN Ensemble with Data Augmentation for Object Detection,” ArXiv150607224 Cs, Jun. 2015, [Online]. Available: http://arxiv.org/abs/1506.07224.

G. Huang, Y. Li, G. Pleiss, Z. Liu, J. E. Hopcroft, and K. Q. Weinberger, “Snapshot Ensembles: Train 1, get M for free,” ArXiv170400109 Cs, Mar. 2017, [Online]. Available: http://arxiv.org/abs/1704.00109.

H. Zheng et al., “A New Ensemble Learning Framework for 3D Biomedical Image Segmentation,” ArXiv181203945 Cs, Dec. 2018, [Online]. Available: http://arxiv.org/abs/1812.03945.

Modellkompression

Blalock, D. W., Ortiz, J. J. G., Frankle, J., and Guttag, J. V. What is the state of neural network pruning? ArXiv, abs/2003.03033, 2020. URL https://arxiv.org/abs/2003.03033.

Li, J., Cotterell, R., and Sachan, M. Differentiable Subset Pruning of Transformer Heads. arXiv:2108.04657 [cs], August 2021. URL http://arxiv.org/abs/2108.04657. arXiv: 2108.04657 version: 1.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. Scaling Laws for Neural Language Models. arXiv:2001.08361 [cs, stat], January 2020. URL http://arxiv.org/abs/2001.08361. arXiv: 2001.08361.

Frankle, J. and Carbin, M. The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks. arXiv:1803.03635 [cs], March 2019. URL http://arxiv.org/abs/1803.03635. arXiv: 1803.03635.

He, W., Wu, M., Liang, M., and Lam, S.-K. Cap: Context-aware pruning for semantic segmentation. In Proceedings of the IEEE/CVF Winter Conference on Applications of

Computer Vision (WACV), pp. 960–969, January 2021. URL https://openaccess.thecvf.com/content/WACV2021/papers/He_CAP_Context-Aware_Pruning_for_Semantic_Segmentation_WACV_2021_paper.pdf.

Gajurel, A., Louis, S. J., and Harris, F. C. Gpu acceleration of sparse neural networks. ArXiv, abs/2005.04347, 2021.

5 Zusammenarbeit mit anderen Stellen

Bereits kurz nach dem Projektstart wurden WP-übergreifende Arbeitsgruppen konzipiert, um besser sicherstellen zu können, dass die Konsortialpartner gemeinsam an der Integration der von ihnen entwickelten Technologien arbeiten. Dieses Konzept wurde zur Projektlaufzeit dahingehend weiterentwickelt, dass alle von den Projektpartnern zu entwickelnden Technologien, die sich in Form von Common Technology Elements[3] (CTE) und Design Patterns[4] (DP) darstellen, und das Zusammenspiel in der gemeinsamen Nutzung von CTE und DP durch die Projektpartner in einer Übersicht zusammengeführt. Die Übersicht zeigt für jeden Anwendungsfall und jede Bedrohung

- welche Technologien auf diese spezifische Bedrohung abzielen und dieselben oder kompatible CTE/DP verwenden sowie
- den Namen derjenigen Konsortialpartner, die auch Technologie für denselben Anwendungsfall/dieses Bedrohungsszenario entwickelt.

Damit werden “direkte Nachbarn” im Projekt, unabhängig davon, an welchem WP sie beteiligt sind, sichtbar. Dieser Ansatz wurde in Form des “Cross-Track”-Tool (CT-Tool) operationalisiert, im Projekt durch den Konsortialleiter NXP ausgerollt und durch alle Projektpartner gefüllt und fortlaufend aktualisiert.

Innerhalb des Konsortiums waren für Merantix die zentralen Ansprechpartner die Projektpartner TNO (Bereitstellung Fahrzeug für Demonstrator, ebenfalls Entwickler eines Anomaliedetektors) und Commsignia (Bereitstellung der On-board Unit für die Fahrzeug-Fahrzeug- bzw. Fahrzeug-Infrastruktur-Kommunikation) mit denen gemeinsam auch die Demonstration der Ergebnisse in WP9 vorbereitet wurden.

[3] CTE sind bestehende industriell erprobte Technologien (ab TRL 7), die verwendet werden können, um neue Sicherheitslösungen in den verschiedenen Bereichen in SECREDAS zu entwickeln. CTE sind domänenunabhängig und können beispielsweise Kryptografiebibliotheken, Hardwareanker für die sichere Schlüsselspeicherung, Kommunikationsnetzwerke und -protokolle oder vorhandene Sicherheitsprodukte wie Firewalls, vertrauenswürdige Ausführungsumgebungen oder Distributed-Ledger-Technologien sein. In SECREDAS wurden 24 CTE definiert.
[4] DP stellen i.S.v. SECREDAS wiederverwendbare Lösungen inklusive Formulierung von Best Practices für Security und Privacy sowie Protokoll- und Architekturspezifikationen. Im Vorhaben wurden in Summe 34 DP definiert.

II. Eingehende Darstellung

1. Verwendung der Zuwendung und des erzielten Ergebnisses im Einzelnen, mit Gegenüberstellung der vorgegebenen Ziele

Der wesentliche Teil der Zuwendung wurde für die anteilige Deckung der Personalkosten verwendet. Diese Aufwendungen spiegeln sich in den Ergebnissen der Arbeitspakete, an denen Merantix beteiligt war, wider. Zu einem geringen Teil fielen Reisekosten - resultierend aus Projekttreffen auf Gesamtprojektebene - sowie Ausgaben für die Nutzung von Cloud Services an.

Merantix war im Vorhaben SECREDAS an den Arbeiten in fünf Arbeitspaketen beteiligt:

WP1 Nutzerszenarien

Ziel von WP1 war die Entwicklung des Designs eines idealtypischen Szenarios im Bereich autonomes Fahren inklusive der ausführlichen Formulierung potentieller Attacken und Gefahren.

In WP1 wirkte Merantix an der Formulierung des Nutzerszenarios des “Sicheren automatisierten Fahrens” mit. Dabei hat sich Merantix insbesondere darauf konzentriert, die Anforderungen für eine robuste Wahrnehmung, die sicheres autonomes Fahren ermöglicht, zu skizzieren. Hierzu wurden potentielle Risiken ungesehener Grenzfälle sowie die Risiken verschiedenartiger feindlicher Angriffe von außen näher beschrieben.

Die Arbeiten an WP1 erfolgten im ersten Projektjahr. Zu Beginn des zweiten Projektjahres wurden in Vorbereitung auf die Demonstration (WP9) die Szenarien und Use Cases weiterentwickelt und verfeinert. Merantix hat in die Diskussion den Use Case 1.5 (Angriffe auf automatisiert fahrendes Fahrzeug an der Kreuzung) eingebracht. Die Ergebnisse wurden in Deliverable D1.2 dokumentiert.

WP2 Referenzarchitekturen und Anforderungen

Ziel von WP2 war die Analyse der Sicherheits- und Datenschutzanforderungen, welche für die Entwicklung, Implementierung und Evaluierung der SECREDAS-Anwendungsfälle erforderlich sind.

Innerhalb von Task T2.3 (Security & Privacy Reference Architecture for Safe Automated Systems) hat Merantix an der Formulierung der Sicherheitsprinzipien für verschiedene Use Cases mitgewirkt (mit dem Fokus auf der Umgebungswahrnehmung der On-board Kameras, um die Robustheit gegenüber Grenzfällen und Angriffen auf das System zu gewährleisten.) und damit zur Definition einer Referenzarchitektur für “Secure and safe automated systems in the context of autonomous driving” beigetragen.

WP4 Sensorik im Automobil

Ziel dieses Work Packages war die Entwicklung einer robusten Bildsegmentierung, die für hochautomatisiertes Fahren eingesetzt werden kann. Die Methode sollte robust gegenüber Grenzfällen (sogenannte Corner Cases) und bösartigen Angriffen sein.

Damit ein Fahrzeug autonom fahren kann, muss die Steuerung der Quer- und Längslenkung automatisiert werden. Hierzu müssen Erfassungsdaten verarbeitet werden, um Informationen über die Fahrumgebung abzuleiten. Mit Hilfe von Objekterkennungs- und Bildsegmentierungsverfahren kann ein solches Verständnis erreicht werden.

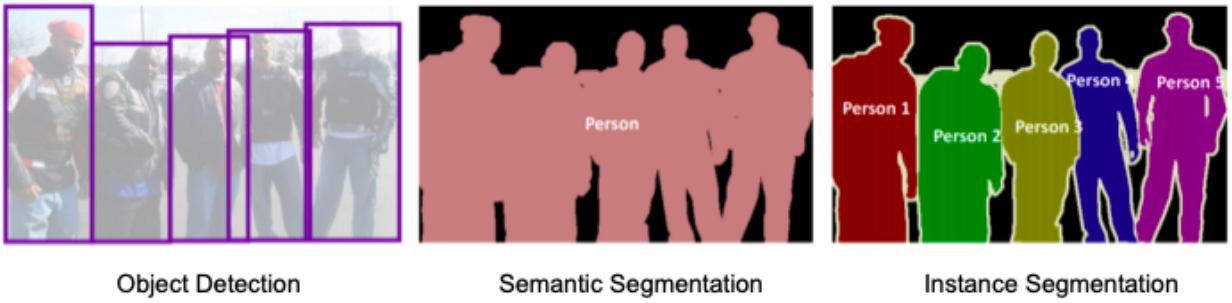


Abbildung 2: Übersicht möglicher Ergebnisse der Bildverarbeitungs- und -erkennungsalgorithmen

Bildsegmentierungsalgorithmen (Semantic Segmentation) liefern ein feingranulares Verständnis des gesamten Bildes, da sie Informationen für jedes einzelne Pixel liefern. Die zugrunde liegenden Algorithmen helfen, die globale Bildszene zu verstehen. Jedes Pixel wird einer Entität zugeordnet, während gleichzeitig Informationen des gesamten Bildes und der Nachbarschaft des Pixels berücksichtigt werden. Dementsprechend kann der Vorgang des Gruppierens von Pixeln, die ein bestimmtes Objekt oder eine bestimmte Kategorie in einer Szene darstellen, als Bildsegmentierung beschrieben werden. Dabei wird das digitale Bild anhand seiner Eigenschaften in verschiedene Teile zerlegt. Die bekanntesten Ansätze zur semantischen Segmentierung beinhalten künstliche neuronale Netze (auch als Deep Learning bezeichnet). Durch die Implementierung eines neuronalen Netzwerks können Eingabedaten wie Video-Feeds von den Bordsensoren des Fahrzeugs verarbeitet werden.

Für das Vorhaben SECREDAS und den hier betrachteten Anwendungsfall wurde eine Bilderkennung auf Basis von Video-Daten durch Merantix in Form von neuronalen Netzen implementiert.

Im Bereich der Bilderkennung hat maschinelles Lernen in den letzten Jahren erstaunliche Ergebnisse gezeigt. Es ist jedoch von größter Bedeutung, die Robustheit solcher intelligenter Verfahren für den Einsatz an realen Systemen und im Produktivbetrieb sicherzustellen. Bei der Verwendung von neuronalen Netzen zur Bilderkennung ist es wichtig, nicht nur die Erkennungsgenauigkeit der Systeme zu berücksichtigen, sondern auch die Robustheit ihrer Ausgabe im Falle von Daten, die im Voraus nicht bekannt waren oder sogar gezielt zur Manipulation des Systems produziert wurden.

Eine Herausforderung beim Einsatz von Deep Learning besteht darin, dass Modelle in der Regel eine gute Leistung auf den Daten, auf denen sie trainiert wurden, zeigen. Oft bleibt jedoch unklar, wie diese sich bei Daten verhalten, die nicht zum Trainieren des Modells verwendet wurden. Hierzu zählen Out-of-Distribution (OOD)-Daten (Daten, die sich von dem Datentyp unterscheiden, mit dem das Modell trainiert wurde) oder Adversarial Attacks, also mutwilligen Angriffen, die eine geeignete Eingabe für ein Modell darstellen (z. B. ein Bild für ein

Bildseg- mentierungsmodell), aber künstlich mit dem Ziel erstellt wurden, das Modell zu täuschen.

Die Robustheit von Bilderkennungsmethoden kann analysiert werden, indem die Rohdaten augmentiert bzw. geändert oder gestört werden, beispielsweise durch die künstliche Veränderung der Lichtverhältnisse, des Kontrastes, der Schärfe, der Verzerrung oder durch Hinzufügen künstlichen Sensorrauschens. Während ein robustes Modell nur wenig Veränderung der Resultate zeigen wird, werden klassische Ansätze sensitiv auf die Veränderung der Rohdaten reagieren. Die relevanten Vorarbeiten hierfür wurden in T4.1 durch Merantix geleistet.

Merantix adressiert die beschriebene Herausforderung innerhalb von SECREDAS primär durch die Entwicklung

- von Modellen, die robust gegenüber OOD-Daten und mutwilligen Angriffen sind,
- eines Algorithmus, der es erlaubt, OOD-Samples und mutwillige Angriffe zu erkennen und somit manipulierte Eingabedaten zu identifizieren und
- eines Testing Framework, welches es erlaubt, die beiden erstgenannten zu evaluieren.

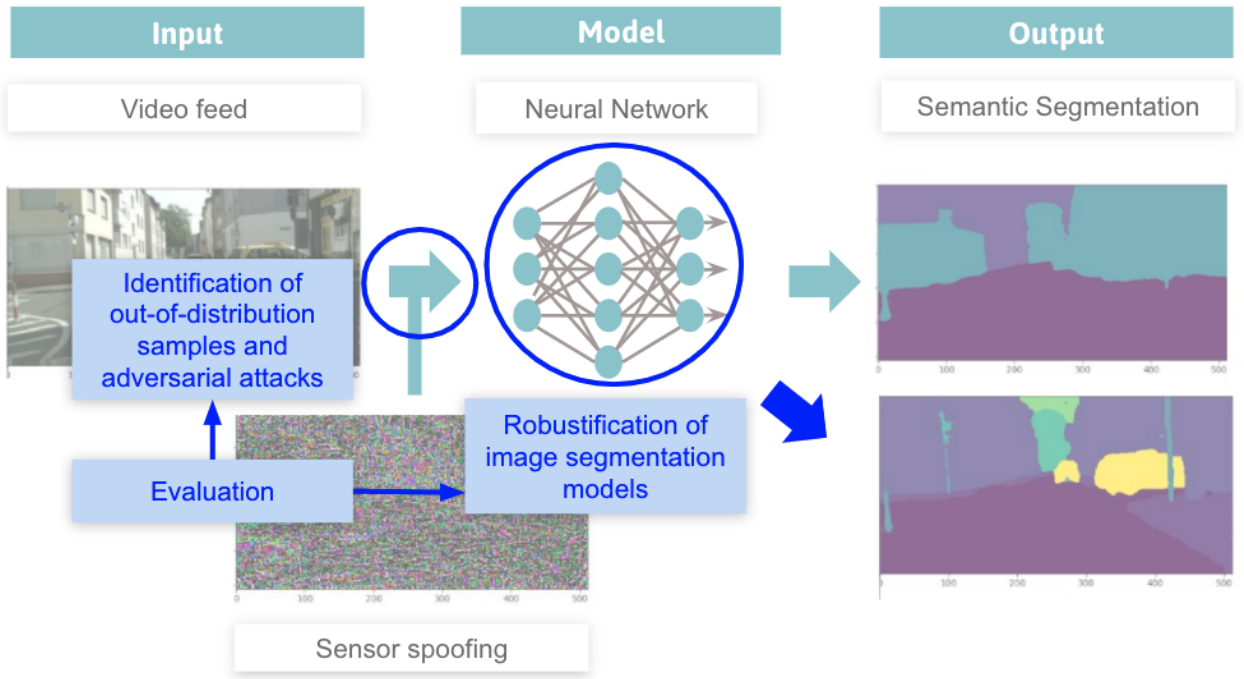
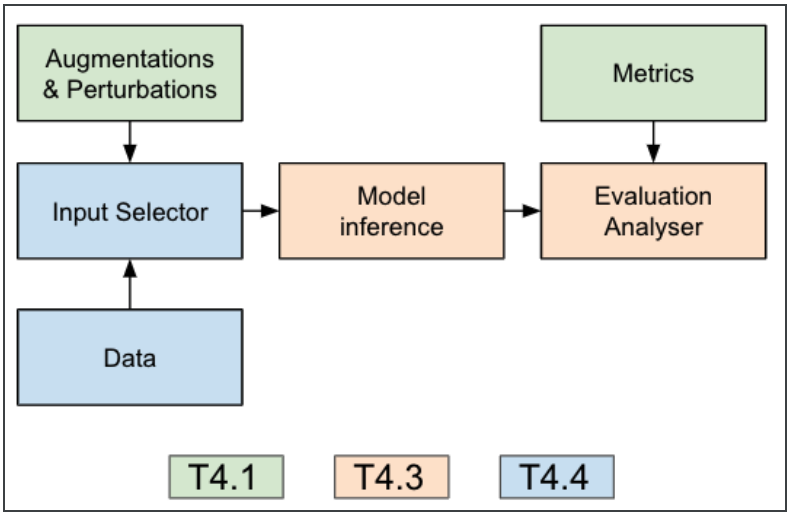


Abbildung 3: Konzeptioneller Ansatz von Merantix in WP4

Die Aufgaben von Merantix innerhalb von WP4 teilen sich auf drei Tasks (T) auf, deren Zusammenspiel in nachfolgender Abbildung dargestellt ist:

- T4.1 Konzeption von Sensor, Komponenten und Sensordatenverarbeitung
- T4.3 Entwicklung Datenverarbeitung
- T4.4 Integration und Validierung von Sensorsystemen



Abbbildung 4: Zusammenspiel der Merantix-Tasks (T) in WP4

T4.1 Konzeption von Sensor, Komponenten und Sensordatenverarbeitung

Für die Evaluation der Performanz und Robustheit von Bildsegmentierungsnetzwerken hat Merantix in SECREDAS ein so genanntes “Testing Framework” implementiert. Dieses Framework wurde verwendet, um die entwickelte Methode zur Erhöhung der Robustheit von Bildsegmentierungsnetzwerken (Model Ensembling, siehe T4.3) zu evaluieren, indem die Leistungsabnahme eines Modells gemessen wird, wenn Erweiterungen auf den Evaluierungsdatensatz angewendet werden. Darüber hinaus wird es zur Generierung der OOD-Daten verwendet, die zur Bewertung des Ansatzes zur OOD-Erkennung verwendet werden. Für das Testing Framework wurden eine Reihe von Funktionalitäten implementiert:

Zu den wesentlichen Ergebnissen zählen die Implementierung von Möglichkeiten zum Einspielen von Bildstörungen (sogenannte “Perturbations” und "Augmentations") - wie z.B. Änderungen der Belichtung, Rotationen, Spiegelungen, Verschiebungen, Unschärfe und Entfernen einzelner Pixel. U.a. wurden auch 15 Augmentierungen der realen Welt (z.B. durch Simulation von Frost, Nebel, unterschiedlicher Belichtungsdauer usw.) auf Basis des Hendrycks-Framework realisiert.

In einem nächsten Schritt wurden vier verschiedene Arten von Attacken implementiert, die einen Kenntnisstand über das Bilderkennungssystem (= das neuronale Netz) (sog. „White-Box-Angriffe“) bzw. keinen Kenntnisstand („Black-Box-Angriffe“) auf Seiten der Angreifer voraussetzen. Zu ersteren gehört die Generierung von Bildern, die gezielt zu einem Versagen der Bilderkennung führen („Adversarial Attacks“). Diese umfassen verschiedene Ausprägungen von Fast Gradient Sign Attacks (FGSM). Für Black-Box-Angriffe wurden auf anderen Modellen berechnete Tests zur Anfälligkeit für mutwillige Angriffe implementiert und die Methode „Generalizable Data-free Objective for Crafting Universal Adversarial Perturbations“ (GD-UAP) zur Generierung von universellen mutwilligen Angriffen in das Framework integriert.

Zur Darstellung der Auswirkungen dieser unterschiedlichen Angriffe wurden Visualisierungen implementiert, die

- den originalen Input (Unperturbed Image)
- das modifizierte Bild (der "Angriff", Perturbed Image)
- die Bilddifferenz zwischen originalem und modifiziertem Bild
- den Output der Bildsegmentierungsmethode (in diesem Fall ESPNet) für Original-Input (Unperturbed Mask), modifiziertem Input (Perturbed Mask) und Differenz zwischen diesen Outputs (Mask Diff)

erzeugen. Nachfolgende Abbildung zeigt eine solche Visualisierung für eine ausgewählte Attacke:

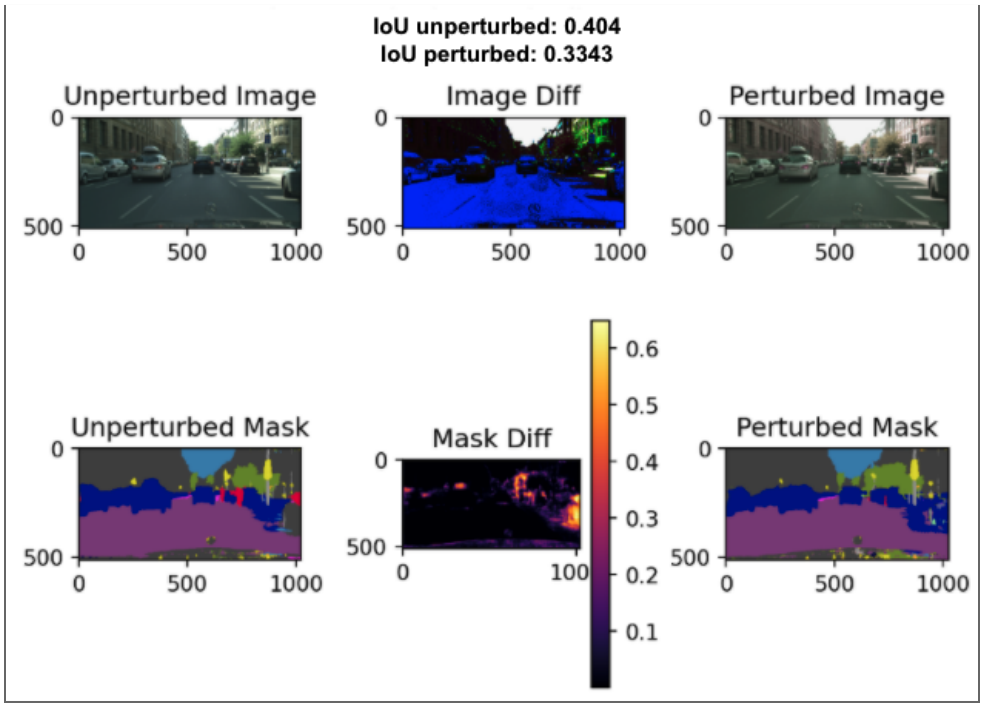


Abbildung 5: Visualisierung der Effekte eines ausgewählten Angriffs ("Colour Jitter"[5]) auf die Bildsegmentierung

Der Output wird bewertet anhand verschiedener Metriken, u.a. zur Quantifizierung der Robustheit und Generalisierungsfähigkeit von neuronalen Netzen zur semantischen Segmentierung, die in T4.1 konzipiert und implementiert wurden.

Für die semantische Segmentierung wurde auf eine bestehende Modellarchitektur (ESPNet) zurückgegriffen und das neuronale Netz auf dem Cityscapes-Datensatz trainiert.

Aufbauend auf diesen Ergebnissen lassen sich Trainingsstrategien entwickeln, die es ermöglichen, das Netzwerk gegen die demonstrierten Angriffe robuster zu machen. Hierzu bedarf es der Einbindung von modifiziertem Input in das Training des Netzwerkes.

Die Ergebnisse wurden umfassend in Deliverable D4.6 dokumentiert. Ferner wurde ein Video generiert, welches die Funktionalitäten des Testing Frameworks aufzeigt.

[5] ermöglicht u.a. die Änderung der Helligkeit, Änderung des Kontrasts, Änderung des Farbspektrums

T4.3 Entwicklung Datenverarbeitung

Innerhalb von T4.3 wurde im Kern mittels Model Ensembling eine Methode entwickelt und implementiert, die eine Robustifizierung von neuronalen Netzwerken ermöglicht.

Um gezielte Angriffe auf den Inputdatenstrom zu erkennen und im Kontext des Vorhabens SECREDAS diese Information auch anderen Verkehrsteilnehmern verfügbar machen zu können, wurde durch Merantix ein Algorithmus entwickelt, der es erlaubt, OOD-Samples und mutwillige Angriffe zu erkennen und somit manipulierte Eingabedaten zu identifizieren. Die Performance des entwickelten Anomaliedetektors ließe sich durch den Einsatz größerer Modelle oder durch die Implementierung und das Training komplexerer Klassifizierer noch deutlich weiter steigern.

In den vergangenen Jahren wurden erhebliche Fortschritte beim Training von Convolutional Neural Networks (CNN) erzielt, die neue Anwendungen in der semantischen Segmentierung ermöglichen. Diese Fortschritte gehen jedoch mit hohen Anforderungen an Rechenressourcen einher. Hochmoderne Modelle haben in der Regel Parameter in der Größenordnung von 10^7, was den Einsatz in Umgebungen mit knappem Rechenbudget, wie autonomen Fahrzeugen oder On-Edge-Geräten, schwierig macht.

Um die Rechenlast bestehender hochmoderner Modelle zu reduzieren und eine hohe Leistung bei allen Aufgaben aufrechtzuerhalten, wird aktiv an Methoden zur effizienten Komprimierung und Beschleunigung neuronaler Netze geforscht. Um perspektivisch die Ergebnisse von der Forschung auch in den Realeinsatz bringen zu können, hat Merantix in SECREDAS die Methode des Pruning als Komprimierungstechnik für Bildverarbeitungsaufgaben weiterentwickelt und von der Aufgabe der Klassifizierung auf jene der semantischen Segmentierung erweitert.

Ensembling Framework

Die Güte eines maschinellen Lernmodells zeigt sich u.a. auch darin, wie gut es mit Daten umgehen kann, die von der Menge der Daten, auf der trainiert wurde, abweicht. Einem Modell mit dieser Qualität wird eine hohe Generalisierungsfähigkeit nachgesagt.

Die Robustheit eines maschinellen Lernmodells steht in direktem Zusammenhang mit seiner Generalisierungsfähigkeit. Es kann als jene Eigenschaft definiert werden, dass sowohl auf Trainingsdaten als auch auf Testdaten - wobei sich die Testdaten von den Trainingsdaten in Bezug auf Datenverteilung und Rauschen unterscheiden und Störungen enthalten können - ähnliche gute Leistungen erzielt werden.

Für die Generalisierung von Modellen bestehen mehrere Ansätze, die von Merantix im Rahmen von Literaturrecherchen näher betrachtet und strukturiert wurden:

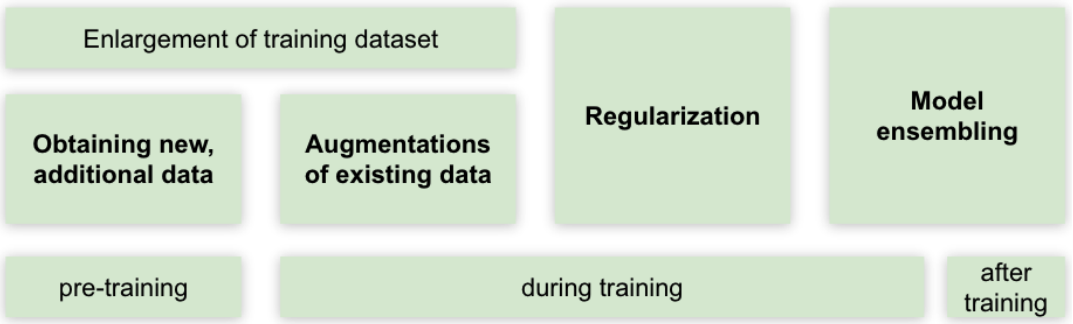


Abbildung 6: Ansätze zur Erhöhung der Robustheit von Machine-Learning-Modellen

Ein Ansatz zur besseren Generalisierung eines Modells besteht darin, die Größe des Datensatzes zu erhöhen, der zum Trainieren des Modells verwendet wird. Die Verwendung von mehr Daten ist jedoch nicht immer eine praktikable Lösung. Zum einen sind in einigen Anwendungsfällen die verfügbaren Daten begrenzt oder das Sammeln zusätzlicher Daten ist aufwändig. Zum anderen erfordert das Trainieren eines Modells mit mehr Daten mehr Zeit und Rechenleistung - je nach Trainingsinfrastruktur und -ressourcen stellt dies eine weitere Herausforderung dar.

Eine Alternative hierzu ist die Erhöhung der effektiven Größe des Trainingsdatensatzes, indem Daten dieses Datensatzes augmentiert werden. Auf diese Weise können aus vorhandenen Mustern weitere Muster generiert werden. Augmentationen können dazu beitragen, dass das Modell robuster gegenüber Transformationen wird. Beispielsweise kann die Verwendung von Drehung und Spiegelung als Erweiterungen von Bildern das Modell gegenüber solchen Transformationen invariant machen. Die Datenerweiterung in Bildverarbeitungsaufgaben ist ein aktives Forschungsgebiet. Neuere Ansätze umfassen die Verwendung von Generative Adversarial Networks (GAN).

Darüber hinaus lässt sich die Robustheit des Modells auch erhöhen, indem man das Modell während des Trainings reguliert. Bei der Regularisierung werden dem Modell zusätzliche implizite oder explizite Einschränkungen auferlegt, sodass die Trainingsdaten nicht einfach überangepasst (so genanntes “Overfitting”). “Early Stopping”, “Weight Decay” und Dropout-Varianten sind verbreitete Alternativen zur Regularisierung im Deep Learning. In den meisten Fällen wird eine Art Regularisierung verwendet, während tiefe neuronale Netze trainiert werden, um zu verhindern, dass sie überangepasst werden und zu robusteren Modellen führen. Daher kann die Regularisierung als eine Möglichkeit zur Robustifizierung von Modellen zur Trainingszeit angesehen werden.

Ein weiterer Weg, um ein robusteres Modell zu erhalten, besteht darin, ein Ensemble von Modellen zu verwenden, d.h. ein Metamodell zu haben, das die Vorhersagen von Teilmodellen verwendet, um die endgültige Vorhersage zu erzeugen. Es wurde festgestellt, dass im Allgemeinen ein Ensemble von Modellen eine höhere Leistung als einzelne Teilmodelle erreicht und robuster ist. Dies ermöglicht das Training schwächerer Submodelle mit begrenzter Zeit und begrenzten Rechenressourcen und erzielt dennoch eine hohe Leistung durch Ensembling. Die Aussage, dass ein Ensemble eine bessere Generalisierungsleistung hat, gilt auch für tiefe

neuronale Netze, die in der Bildverarbeitung verwendet werden. Daher haben wir uns für die Arbeit in SECREDAS für die Verwendung von Ensembles entschieden, um eine robuste Segmentierungsleistung zu erzielen.

In unserer Arbeit testen wir die Robustheit unserer Modelle gegenüber mutwilligen Angriffen und OOD-Samples und validieren, dass ein Ensembling-Ansatz ein praktikabler Weg ist, um eine solche Robustheit zu erreichen. Um zu validieren, ob Ensembling ein robustes Ergebnis erzeugt, haben wir drei Modelle und diese in ihrer Performanz gegen verschiedene mutwillige Angriffe einzeln und auch als Ensemble evaluiert und konnten dabei den Nachweis erbringen, dass Model Ensembling einen vielversprechender Weg zur Robustifizierung von Bildsegmentierungsaufgaben darstellt.

Die Ergebnisse wurden umfassend in Deliverable D4.6 dokumentiert.

Anomaliedetektor

Da Deep-Learning-(DL-)Ansätze immer häufiger in realen Umgebungen eingesetzt werden, ist es von größter Bedeutung, zu verstehen, wann das zugrunde liegende Modell unsicher über seine Vorhersagen ist. Dies gilt umso mehr für Bereiche, in denen falsche Vorhersagen verheerende Folgen haben können und in denen die Bildverarbeitung mittels DL beeindruckende Ergebnisse erzielt hat, wie etwa beim autonomen Fahren oder der Erkennung von Brustkrebs.

Solche Modelle können jedoch auch das Ziel gezielter Angriffe sein, die versuchen, das Modell durch manipulierte Inputdaten zu falschen und möglicherweise gefährlichen Vorhersagen zu führen.

Es wird davon ausgegangen, dass ein Angreifer - z.B. ein Hacker - Zugriff auf den Eingangsbildstrom eines Fahrzeugs hat. Dies stellt das Worst-Case-Szenario im Vergleich dazu dar, den Gegner darauf zu beschränken, Objekte in der realen Welt zu platzieren oder zu modifizieren. Solche Inputdaten stellen Daten dar, die nicht in den Trainingsdaten enthalten waren. Das Modell war daher nicht darauf vorbereitet, mit diesen Daten umzugehen und weiß nicht, wie es darauf richtig reagieren soll, was zu fehlerhaften Ausgaben führt.

Merantix hat daher im Vorhaben SECREDAS eine neuartige Methode mit dem Titel Detection of Anomalous and Adversarial Input using Normalizing Flows (DAAIN) konzipiert und entwickelt, die es ermöglicht, sowohl anomale Daten als auch mutwillige Angriffe zu erkennen, indem Aktivierungen aus den angegriffenen Schichten des Modells beobachtet und diese Aktivierungen mit Hilfe von “Normalizing Flows” auf eine gewünschte Zielverteilung umgewandelt werden. Auf diese Zielverteilung sind einfache Abstandsmetriken anwendbar, die eine Klassifizierung ermöglichen. Im Vorhaben SECREDAS hat Merantix diesen Ansatz auf die Aufgabe der Bildsegmentierung erweitert. Während Ansätze für die Klassifizierung bereits entwickelt wurden, ist eine Anpassung für die Segmentierung nicht trivial.

Der gewählte Ansatz misst zuerst die Aktivierungen, ähnlich wie bei einem EEG beim Menschen, transformiert diese Beobachtungen und verwendet schließlich einen einfachen Klassifikator, um die Ausgabe zu erzeugen – ist es anomal oder nicht?

Die entwickelte Methode wurde hinsichtlich der Fähigkeit, anomale und adverseriale Eingaben zu erkennen, bewertet. Wir konnten den Nachweis erbringen, dass die vorgeschlagene Methode zwei etablierte Verfahren, welche als Baseline für einen Vergleich herangezogen wurden- Maximum-Softmax-Probability (MSP) und Monte-Carlo Dropout (MCD) - deutlich übertrifft.

Die Ergebnisse wurde in Deliverable D4.6 sowie in einer eigenen Publikation dokumentiert.

Modellkomprimierung

Merantix hat im Vorhaben SECREDAS einen neuen Ansatz zur Modellkomprimierung für die semantische Bildsegmentierung mit CNNs erforscht und entwickelt. Auto-Compressing Subset Pruning, kurz ACOSP, basiert auf einer Modifikation des wissenschaftlichen Ansatzes des Differential Subset Pruning und reduziert effizient die Anzahl der Faltungsfilter im so komprimierten Modell, was zu einem dünnen Netzwerk im Vergleich zum nicht komprimierten Modell führt. Während sich die meisten Pruning-Ansätze darauf konzentrieren, die Parameter oder Filter zu finden, die sicher entfernt werden können, konzentriert sich ACOSP auf den Entfernungsprozess selbst.

ACOSP zielt vor allem auf Anwendungsfälle ab, welche von einer sehr hohen Kompressionsrate profitieren und demgegenüber geringe Einbußen in der Erkennungsgenauigkeit tolerieren können und übertrifft in diesem Bereich frühere Ansätze bei den meisten Datensätzen erheblich. So werden Ergebnisse mit ausreichender Qualität selbst dann erreicht, wenn mehr als 93% des Modells verworfen werden.

Die Ergebnisse wurden in Form einer eigenen Publikation dokumentiert.

T4.4 Integration und Validierung von Sensorsystemen

Mit der Beteiligung an T4.4 verfolgte Merantix das Ziel, die entwickelten Konzepte und Methoden in das Gesamttestsystem einzubinden. Relevante Arbeitsschritte hierfür stellten die Einbindung der Daten anderer Partner sowie die Synthese relevanter Testszenarien dar.

Hierzu fanden eine Vielzahl von Abstimmungen insbesondere mit den beiden Projektpartnern TNO und Commsignia statt: TNO stellte das Fahrzeug (inklusive Kamera(daten)), in dem die Algorithmen von Merantix demonstriert werden sollen (siehe WP9). Commsignia stellte die On-board unit (OBU) im TNO-Fahrzeug, über die mögliche Fehlermeldungen - u.a. aufgrund von (für die Demo simulierten) Attacken auf den Kamerasensor - an andere Verkehrsteilnehmer mit Hilfe von V2X-Kommunikation verteilt werden.

WP9 Gemeinsame Demonstratoren

Die im Gesamtprojekt SECREDAS entwickelten Technologien und deren Zusammenspiel wurde in Form gemeinsamer Demonstratoren präsentiert. Für die drei im Vorhaben adressierten Anwendungsdomänen wurden jeweils eigene Demonstratoren (Automotive: “Demo I”, Rail: “Demo II”, Health: “Demo III”) konzipiert und an verschiedenen geographischen Orten umgesetzt. Allen Demonstratoren in SECREDAS gemein war die Vernetzung der beteiligten Einzelsysteme, um auf der einen Seite deren Zusammenspiel aufzeigen zu können, auf der

anderen Seite aber auch sichtbar machen zu können, wie sich vermeiden lässt, dass sich solche Angriffe im Gesamtsystem ausbreiten. Für die Anwendungsdomäne Automotive wurden daher Systeme implementiert, die eine Fahrzeug-zu-Fahrzeug- bzw. Fahrzeug-zu-Infrastruktur-Kommunikation - als V2X-Kommunikation im Folgenden zusammengefasst - ermöglichen.

Gemeinsam mit den anderen an der Entwicklung von Lösungen für das automatisierte Fahren (Anwendungsdomäne Automotive) beteiligten Partnern hat Merantix ein Demonstrationskonzept erarbeitet, welches sowohl die Präsentation einzelner Lösungen als auch deren Zusammenspiel umfasste. Hierfür wurden zunächst intern mehrere (Sub-)Szenarien konzipiert, die - für die Demonstration gegenüber Externen - in eine klare Storyline für mit mehreren, aufeinander aufbauenden und sich logisch fortsetzenden Szenarien, überführt wurde. Vom ersten bis zum letzten Szenario wird dabei der Fokus aus dem einzelnen Fahrzeug heraus auf die Gesamtverkehrssituation gewechselt.

Die Präsentation dieser sogenannten “Demo I” fand am 26.06.2021 auf dem Automotive Campus (TNO) in Helmond (Niederlande) statt.

Szenario	Beschreibung	Interne ID
1	Ein Kamerasensor eines automatisierten Fahrzeugs wird gehackt und eine mutwillige Bildattacke durchgeführt, um die Objekterkennung des Fahrzeugs zu beeinträchtigen.	3C
2	Die Kommunikation eines Infrastruktursensors wird gehackt, um bösartige CPM-Nachrichten zu senden, die falsche Objektdaten über den erkannten Verkehr enthalten.	3B
3	Die Ampel an der Kreuzung wird gehackt und sendet bösartige SPAT-Nachrichten, die darauf hinweisen, dass alle Ampeln grün sind.	3A
4	Ein entführtes automatisiertes Fahrzeug sendet bösartige CAM-Nachrichten, die darauf hinweisen, dass es an einer roten Ampel anhält, während es in Wirklichkeit die Kreuzung überquert und die rote Ampel verletzt.	1C
5	Ein Fußgänger mit einem Ultra-wideband-Tag überquert eine Kreuzung mit einem sich nähernden (entführten) Fahrzeug.	2
6	Eine gehackte oder fehlerhaft funktionierende Road Side Unit (Sensor auf Seiten der Infrastruktur) sendet unvollständige oder falsche Objektdaten	2

Tabelle 1: Demo I Szenarien

Merantix verantwortete die Konzeption und die Umsetzung des Szenario 3C:

Szenario 3C geht von der Bedrohung durch einen Hacker aus, der direkt einen Sensor eines automatisierten Fahrzeugs angreift. Angenommen wird ein feindlicher Angreifer, der die Kamerabilder manipuliert, indem er den Live-Bildstrom im Wesentlichen teilweise korrumpiert.

Diese Kamerabilder werden kontinuierlich von den Objekterkennungs- (und -klassifizierungs-) und Objektverfolgungsalgorithmen des automatisierten Fahrzeugsystems verwendet. Diese Objekte werden mit anderen Messungen und Detektionen des Fahrzeugs fusioniert, um die Umgebung zu modellieren (World Modeling) und somit richtige Entscheidungen zu treffen, um letztendlich sichere Fahrmanöver durchzuführen. Schlechte Informationen über erkannte Objekte können, wenn sie vertrauenswürdig erscheinen, zu falschen Entscheidungen und unsicheren Manövern durch das automatisierte Fahrzeugsystem führen. Obwohl sehr situationsabhängig, besteht hierdurch grundsätzlich eine Gefahr für die Insassen und die umliegenden Verkehrsteilnehmer.

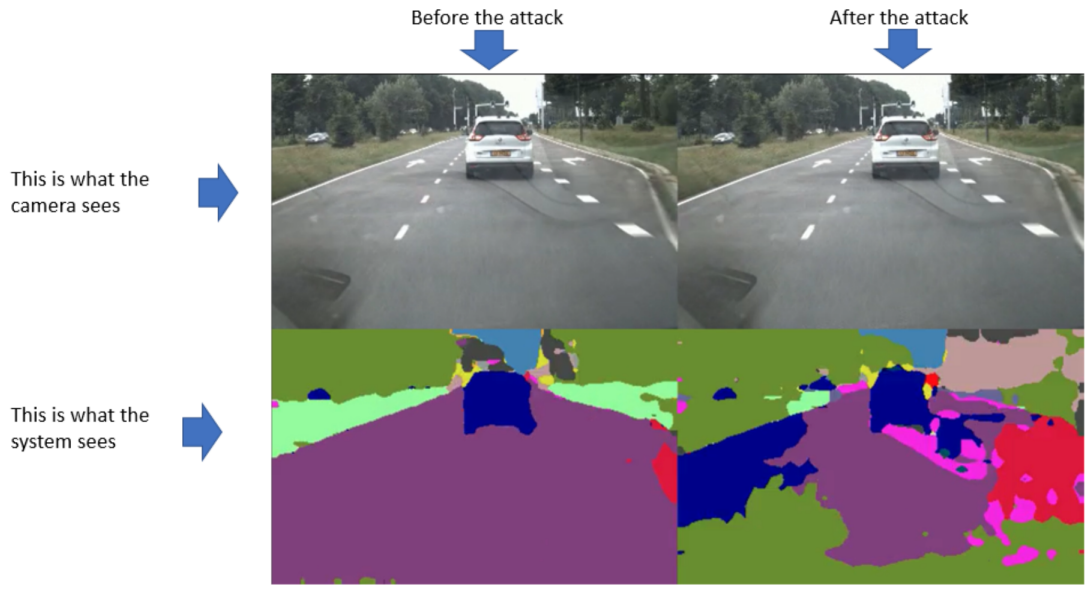


Abbildung 5: Bildsegmentierung auf TNO-Kameradaten

Ein Ausfall des Kamerasensors, der nicht durch menschliches Eingreifen verursacht wurde, kann zu ähnlichen Effekten führen und so könnte eine solche Anomalie-Bedrohung auch durch den gleichen (Art) Algorithmus erkannt werden. Letzteres wurde jedoch nicht für SECREDAS Demo I getestet.

In Szenario 3C wird die Bedrohung durch einen von Merantix entwickelten Algorithmus zur Erkennung von Bildanomalien erkannt. Diese Software ist so konzipiert, dass sie im Fahrzeug läuft und das Fahrzeug auf potenzielle Angriffe auf den Kamerabildstrom aufmerksam macht. Als Eingabe werden die Aktivierungen des Segmentierungsmodells verwendet, welches Objekte aus den Live-Bildern segmentiert (wie Straßen, Vegetation, Gebäude, Fahrzeuge, Fußgänger), um anormale Bilder aufgrund eines mutwilligen Bildangriffs auf Pixelebene zu erkennen. Wird ein solcher Angriff erkannt, löst die Software die Aussendung einer Sicherheitsbenachrichtigung aus. Dann werden in der Implementierung dieses Demonstrators die Schlüssel des Sicherheitszertifikats widerrufen (umgesetzt seitens des Projektpartners Commsignia) - allen anderen V2X-kommunikationsfähigen Fahrzeugen wird damit signalisiert, dass die Kommunikation mit diesem Fahrzeug nicht mehr vertrauenswürdig ist.

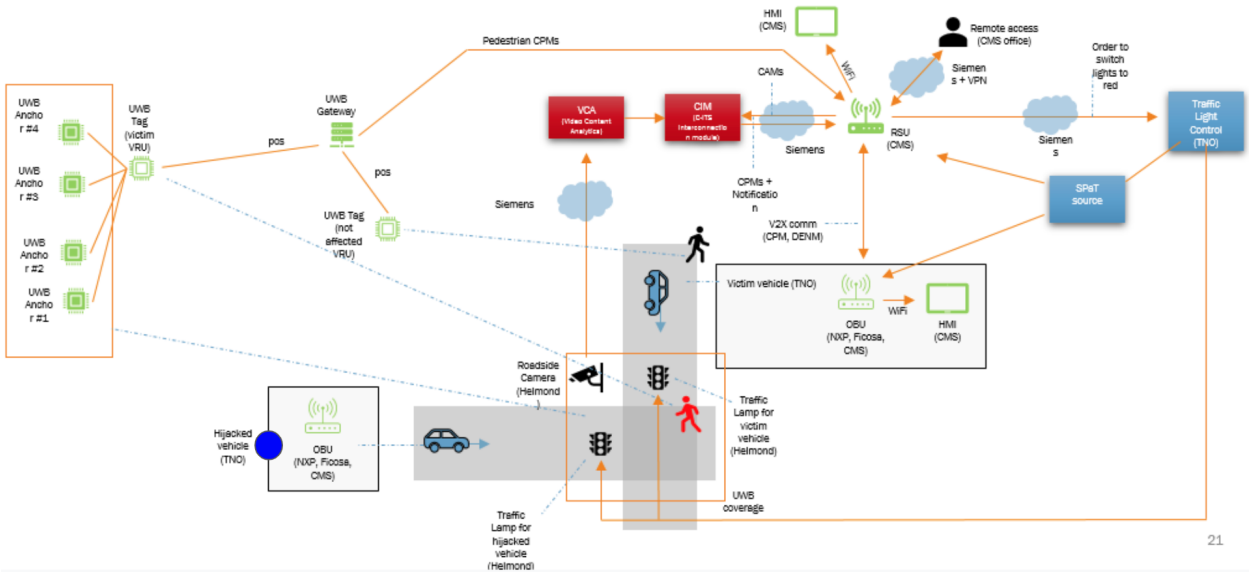


Abbildung 6: Demo I Architektur und Verortung des Anomaliedetektors von Merantix (blauer Punkt)

Im vorliegenden Szenario werden die anderen Fahrzeuge über V2X-Kommunikation gewarnt und ignorieren zukünftig die gemeinsamen Objekte, die vom angegriffenen Fahrzeug mit Hilfe spezifischer Nachrichtenformate (Collective Perception Messages) gesendet werden. Hierdurch werden potenzielle Fehlentscheidungen und unsichere Manöver vermieden.

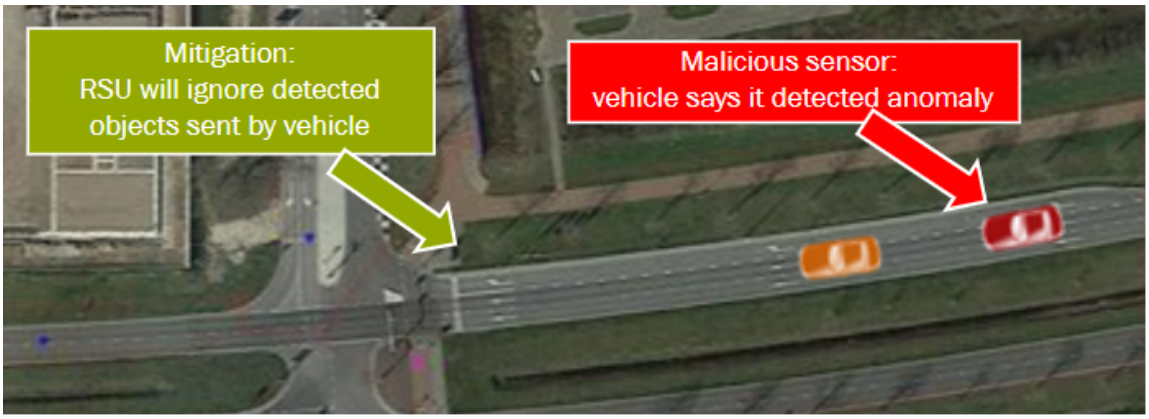


Abbildung 7: Layout von Szenario 3C in Helmond

Während der Präsentation von Demo I (inkl. Szenario 3C) vor Ort wurden Videoaufnahmen von den Einzelsystemen und deren Zusammenspiel erstellt und in Form eines Projektvideos veröffentlicht (siehe WP11).

Management

Merantix oblag im Vorhaben keine eigene Verantwortlichkeit für die Organisation von WPs oder Tasks, an denen Merantix beteiligt war. Lediglich für die Erstellung von Deliverable D4.6 war Merantix alleinig in der Verantwortung.

Während der Projektlaufzeit hat Merantix an allen Projekttreffen auf Gesamtprojektebene (insbesondere die drei Reviews, inklusive der Final Conference zum Projektabschluss) teilgenommen.

Ergebnisverbreitung

Die im Vorhaben SECREDAS erarbeiteten Ergebnisse liegen in verschiedenen Formen (Dokument, Video), die über verschiedene Kommunikationskanäle (z.B. Webseiten, Social Media Posts) - sowohl des Projekts als auch von Merantix - ausgespielt wurden, vor:

- Die Ergebnisse wurden in elf Deliverables dokumentiert (siehe Tabelle 2).
- Es wurden drei Publikationen veröffentlicht (siehe Kapitel 6).
- Es wurde ein Konferenzbeitrag generiert.
- Es wurde ein Projektvideo erstellt.

Beiträge zu Deliverables

An der Konzeption und Erstellung der nachfolgend benannten Deliverables hat Merantix im Vorhaben SECREDAS mitgewirkt:

Deliverable	Datum	Beiträge von Merantix
D1.2 Final reference set of scenarios & use cases	Dezember 2018	Beschreibung von Use Case 1.5 “Resilience of the vehicle’s perception systems against false information about the traffic situation”
D4.1 Conception of sensor, components, and sensor data processing	September 2019	Beschreibung der notwendigen Systemelemente (Kamera, Compute Plattform (GCP)) zur Umsetzung von Use Case 1.5
D4.2 Test concept for sensor, components, and sensor data processing	Februar 2019	Beschreibung des Testkonzepts für die geplante Analyse von Kamerawahrnehmungsmodulen auf ihre Robustheit
D4.6 Robustified Image Segmentation	April 2021	Beschreibung der Kernergebnisse von WP4/T4.3: Model ensembling, Testing framework, Anomaly detector
D9.1 List of requirements and specifications for each system developed	Mai 2020	Beschreibung der “Image Segmentation under attack” als Beitrag für Demo I

D9.2 Single demonstrators working	Juni 2020	Beschreibung des Demonstratorbeitrags der Anomaliedetektion in Kamerasensordaten
D9.5 Document of specifications and requirements for system fusion	August 2020	Beschreibung des Szenarios 1.5 “An automated vehicle under attack”
D9.8 Report on use of DPs and CTEs in the Common Demonstrators	September 2021	Einordnung der im Demonstrator implementierten und gezeigten Lösungen nach Technologiereifegraden (TRL) hinsichtlich der im Projekt definierten Design Patterns (DP)
D9.10 Validation procedures from Demo I, II and III and introduction of Demo IV	September 2021	Beschreibung der geplanten Tests zur Validierung der von Merantix entwickelten Komponenten
D9.11 Report on results of WP9 Demo I	August 2021	Beschreibung der für die Demo eingebrachten Software-Komponenten
D11.7 Dissemination and Exploitation Report	September 2021	Beschreibung der Disseminationsaktivitäten von Merantix

Tabelle 2: Deliverables mit Beiträgen seitens Merantix

Konferenzbeitrag

Auf dem “2nd Workshop for Artificial Intelligence for Small and Medium Enterprises” im Rahmen der "INFORMATIK 2021" hat Merantix die Ergebnisse der Publikation “Chameleon: A Semi-AutoML framework targeting quick and scalable development and deployment of production-ready ML systems for SMEs” Ende September 2021 vorgestellt.

Projektvideo

Um über die Ergebnisdemonstration im Vorhaben selber - im Rahmen der Demo im Sommer 2021 sowie der Final Conference im Herbst 2021 - hinaus die Projektergebnisse interessierten Dritten anschaulich und verfügbar machen zu können wurde innerhalb des Konsortiums der Entschluss gefasst, die Kernergebnisse in einem Video zusammenzustellen. Hierfür wurde ein Konzept entwickelt, welche hierarchisch den roten Faden auf Gesamtprojektebene, auf Ebene der verschiedenen Demos, die die unterschiedlichen Anwendungsdomänen abbilden, sowie der einzelnen Partnerbeiträge (so genannte “Partner Snippet”) zusammenfasst.

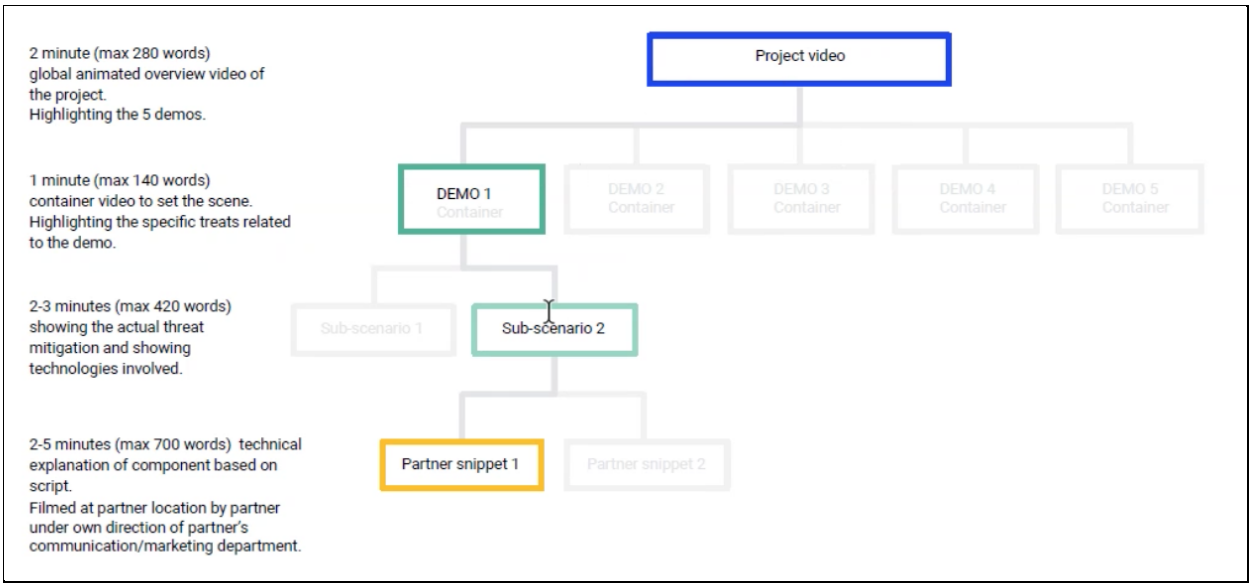


Abbildung 8: Struktureller Aufbau des Projektvideos

Im Ergebnis liegt seit Oktober 2021 ein Projektvideo vor, welches über die SECREDAS Homepage sowie direkt über die Webseite https://secredas.vercel.app/ erreichbar ist. Der Lösungsbeitrag von Merantix ist hier an vorderster Stelle mit dem Beitrag “Hacked Images” (Demo 1 > Scenario 1) sichtbar.

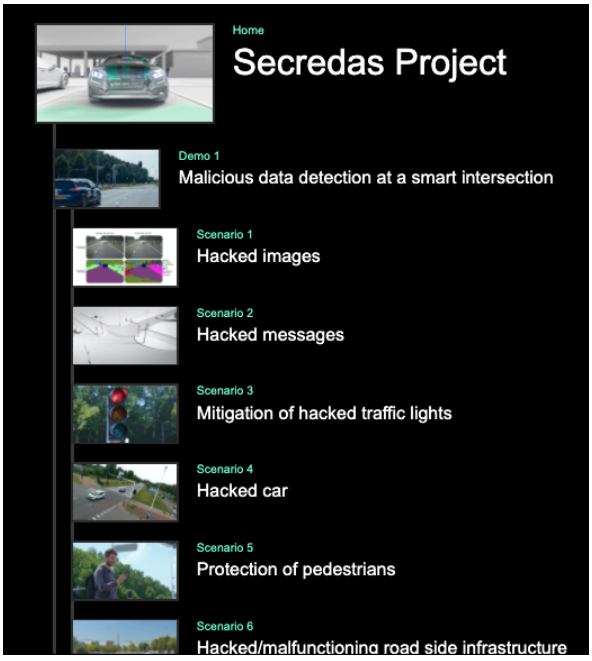


Abbildung 9: Projektvideo (Screenshot)

2. Wichtigste Positionen des zahlenmäßigen Nachweises

Das Gros des Budgets entfiel auf Personalkosten, die 99,2% der Gesamtkosten ausmacht. Die restlichen Aufwände waren bedingt durch Reisekosten (0,4%) sowie Kosten für das Training der neuronalen Netze, die für Datenspeicherung und Rechenleistung, die extern eingekauft wurden.

3. Notwendigkeit und Angemessenheit der geleisteten Arbeit

Die durchgeführten Forschungsarbeiten im Verbundprojekt SECREDAS sowie die dafür aufgewandten Ressourcen waren notwendig und angemessen, da sie der im Projektantrag formulierten Planung entsprachen und alle wesentlichen im Arbeitsplan formulierten Aufgaben erfolgreich bearbeitet wurden.

Die durch Merantix erbrachten Arbeiten und erzielten Ergebnisse hätten ohne die bereitgestellte Förderung nicht durchgeführt werden können. Die erzielten Ergebnisse des Vorhabens und die bei der Projektbearbeitung gewonnenen Methoden- und Domänenkenntnisse bieten Merantix eine Vielzahl von Anknüpfungspunkten für eine weitere Verwertung.

4. Voraussichtlicher Nutzen

Merantix ist darauf spezialisiert, aktuelle Forschungsergebnisse im Bereich des maschinellen Lernens (ML) in industrielle Anwendungen zu überführen. Durch Merantix konnten im Vorhaben SECREDAS die folgenden Projektergebnisse erreicht werden:

(A) Aufbau von Domänenwissen im Bereich Safety und Security im Automotive-Umfeld

(B) Entwicklung eines Testing Frameworks zur Bewertung der Robustheit neuronaler Netze

(C) Aufbau von Methoden-Know-how zur Robustifizierung von neuronalen Netzen

(D) Entwicklung eines Anomaliedetektors für die Bildverarbeitungsaufgabe der semantischen Segmentierung

(E) Entwicklung von Methoden zur Modellkompression

Der Nutzen des Vorhabens SECREDAS für die Merantix AG schlägt sich im Kern in der Weiterverwertung der gewonnenen Erkenntnisse und Ergebnisse im eigentlichen Geschäftsbetrieb und -zweck - der Umsetzung von Machine Learning Lösungen für (Industrie-)Kunden sowie der Gründung neuer Unternehmen auf Basis von Methoden des Maschinellen Lernens - nieder.

Konkret stellen sich die folgenden Verwertungsoptionen dar, die im Erfolgskontrollbericht detaillierter beschrieben werden und für die im Folgenden angegeben ist, welche Ergebnisse auf die Verwertungsoptionen einzahlen:

(1) Beratung von Kunden bei der Ideation, Konzeption und Umsetzung von KI- bzw. ML-Projekten (A)

(2) Technische Realisierung von Proof-of-Concepts (B), (C), (D), (E)

(3) Lizenzierung einzelner Solutions und Module (B), (C), (D), (E)

(4) Bereitstellung von Artefakten als Open Source Software (B), (C), (D)

(5) Ausgründung neuer Unternehmen (A)

5. Während der Durchführung des Vorhabens dem ZE bekannt gewordenen Fortschritts auf dem Gebiet des Vorhabens bei anderen Stellen

Während des Projektzeitraums hat Merantix intensiv den wissenschaftlichen Fortschritt im Bereich des Maschinellen Lernens gemonitort und die identifizierten relevanten Arbeiten in die Konzeption der eigenen Forschungsansätze einfließen lassen. Wesentlich Quellen hierfür waren Fachkonferenz sowie öffentliche Publikationsplattformen.

Im Ergebnis liegt eine umfassende Literaturliste vor, die in Auszügen in Kapitel I.4 dokumentiert worden ist.

6. Erfolgte oder geplante Veröffentlichungen des Ergebnisses

Aus dem Teilvorhaben heraus wurden seitens Merantix drei Publikationen erstellt:

Ditschuneit, Konstantin; Otterbach, Johannes: *Auto-Compressing Subset Pruning for Semantic Image Segmentation*. Veröffentlicht auf arXiv.org am 26.01.2022: https://arxiv.org/abs/2201.11103

Otterbach, Johannes; Wollmann, Thomas: *Chameleon: A Semi-AutoML framework targeting quick and scalable development and deployment of production-ready ML systems for SMEs*. Veröffentlicht auf arXiv.org am 08.05.2021: https://arxiv.org/abs/2105.03669

Von Baußnern, Samuel; Otterbach, Johannes; Loy, Adrian; Salzmann, Mathieu; Wollmann, Thomas: *DAAIN: Detection of Anomalous and Adversarial Input using Normalizing Flows*. Veröffentlicht auf arXiv.org am 30.05.2021: https://arxiv.org/abs/2105.14638