| | |
|---|---|
| Vorhaben | **ExplAINN**<br>**Explainable AI and Neural Networks** |
| Titel | **Abschlussbericht** |
| Förderkennzeichen | 01IS19074 |
| Zuwendungsempfänger | Deutsches Forschungszentrum für Künstliche Intelligenz GmbH<br>Trippstadter Straße 122, D-67663 Kaiserslautern |
| Projektleiter | Sebastian Palacio |
| Bewilligungszeitraum | 01.10.2019 – 30.09.2022 |
| Autoren | Sebastian Palacio, Joachim Folz, Federico Raue, Tushar Karayil, Fatemeh Azimi, Andrey Gushov, Stanislav Frolov, Dayananda Herurkar, Brian Moser, Tobias Nauen |
| Erstellungsdatum | 2. März 2023 |


GEFÖRDERT DURCH


AUFGRUND EINES BESCHLUSSES
DES DEUTSCHEN BUDESTAGES

# Teil I

# Kurzbericht

ExplAINN — 01IS19074

# Kurzbericht

## Aufgabenstellung

Erklärbare KI (Explainable AI, XAI) hat sich zu einem wichtigen Forschungsbereich im Bereich des maschinellen Lernens entwickelt, wobei der Schwerpunkt darauf liegt, KI-Modelle für menschliche Nutzer transparent, interpretierbar und verständlich zu machen. Die zunehmende Komplexität von Black-Box-Modellen hat Bedenken hinsichtlich der Robustheit und Zuverlässigkeit dieser Modelle aufkommen lassen, was XAI zu einem entscheidenden Bestandteil der KI-Forschung macht. In diesem Projekt untersuchten wir die Relevanz von XAI, einschließlich der Bedeutung der Analyse bestehender Modelle, der Entwicklung neuartiger Methoden, die von vornherein interpretierbar sind, und des Verständnisses des Ausmaßes, mit dem die Robustheit eines Modells dessen Leistung beeinflussen kann.

Die Bedeutung der Analyse bestehender Modelle kann nicht hoch genug eingeschätzt werden, da sie wertvolle Einblicke in ihre Robustheit und ihr Verbesserungspotenzial liefert. XAI-Techniken können dazu beitragen, Schwachstellen und Bias in Black-Box-Modellen aufzudecken, was zu einem besseren Verständnis, mehr Vertrauen und letztlich zu einem verantwortungsvollen Einsatz von KI-Systemen führt.

Die Entwicklung von neuartigen, durch ihre Konstruktionsweise erklärbaren Methoden, ist ebenfalls von entscheidender Bedeutung, da sie sicherstellen, dass die Modelle von Grund auf transparent und interpretierbar sind. Dies ist besonders wichtig in Bereichen, in denen auf der Grundlage von KI-Modellen Entscheidungen von signifikanter Tragweite getroffen werden, z. B. im Gesundheitswesen, im Finanzwesen und in der Strafjustiz. Durch die Entwicklung von Modellen, die von vornherein erklärbar sind, kann XAI dazu beitragen, dass KI-Systeme fairer, vertrauenswürdiger und zuverlässiger sind.

Neben der Schaffung von Transparenz und Interpretierbarkeit ist die Entwicklung von Visualisierungen für Computer-Vision-Probleme ein effektives Mittel, um Erklärungen zu Modellvorhersagen zu liefern. Visualisierungen können dabei helfen, komplexe Informationen auf einfache und intuitive Weise zu vermitteln, sodass es für menschliche Benutzer einfacher wird, die Funktionsweise eines komplexen KI-Modells nachzuvollziehen. Durch die visuelle Darstellung des Modellverhaltens und der Beziehungen zwischen Eingaben und Ausgaben können Visualisierungen dazu beitragen, Vertrauen in das Modell aufzubauen und seine Leistung zu überprüfen.

Im Rahmen dieses Projekts wurden verschiedene Techniken und Ansätze für XAI entwickelt und in Anwendungen für verschiedene Bereiche evaluiert, wobei die Robustheit von Black-Box-Modellen untersucht und ihre Entscheidungsgrenzen analysiert wurden, um die nächste Generation interpretierbarer maschineller Lernverfahren zu definieren.

## Ablauf des Vorhabens

Im Allgemeinen wurden alle Ziele gemäß dem Projektantrag erreicht und es wurden mehrere wichtige Beiträge auf dem Gebiet XAI geleistet.

Während wir anfangs mehrere Arbeitspakete als separate Blöcke mit einem gewissen Grad an Überschneidungen identifiziert hatten, stellten wir bald fest, dass die Zusammenhänge zwischen unseren Arbeiten deutlich stärker waren als zunächst angenommen. Diese Erkenntnis führte zu frühen Ergebnissen für die Arbeitspakete 3 und 4, während wir noch Experimente für die ersten beiden Arbeitspakete durchführten, die sich auf XAI für bestehende Modelle, bzw. Robustheitsanalysen konzentrierten.

Eine Literaturrecherche auf dem Gebiet der XAI zeigte auch, dass es noch keine Einigkeit über die Bedeutung von Kernprinzipien wie „Interpretation“ und „Erklärung“ gab. Als Lösungsansatz haben wir einen allgemeinen Rahmen für den Vergleich von Beiträgen in der XAI entwickelt, der als Vorlage für unsere Arbeit in den beiden Arbeitspaketen 1 und 4 diente. Dieses Rahmenwerk wurde in der Community gut aufgenommen, und wir wurden eingeladen, es in mehreren Vorträgen und auf Konferenzen vorzustellen.

Dank der großen Resonanz wurde das Deutsche Institut für Normung (DIN) auf unsere Projektergebnisse im Bereich XAI aufmerksam, was schließlich in unserer Beteiligung an der Erstellung der DIN SPEC 92001-3 „Künstliche Intelligenz - Life Cycle Prozesse und Qualitätsanforderungen - Teil 3: Erklärbarkeit“ mündete. Dies ist eine bedeutende Anerkennung der Wirkung und Relevanz unserer Arbeit und unterstreicht die Bedeutung von XAI für die Entwicklung und den Einsatz von KI-Systemen.

Ein Bereich, in dem wir bessere Ergebnisse erwartet hatten, waren Adversarial Attacks. Trotz mehrerer Experimentreihen, wie sie im Rahmen der Arbeiten für WP3 und WP4 durchgeführt wurden, haben wir nicht den Grad an Robustheit erreicht, den wir uns zu Beginn vorgestellt hatten. Der Bereich der Adversarial Attacks bleibt unter realistischen Bedingungen eine Herausforderung, und es sind weitere Arbeiten erforderlich, um die Robustheit von KI-Modellen gegenüber solchen Angriffen zu verbessern. Wir haben zwar Fortschritte beim Verständnis der Auswirkungen solcher Angriffe auf die Modellleistung erreicht, aber unsere Ergebnisse zeigen, dass selbst die robustesten Modelle für diese Art von Angriffen anfällig sein können. Dies unterstreicht die Bedeutung der Entwicklung neuartiger, konstruktionsbedingt robuster Methoden, und den Bedarf an weiterer Forschung in diesem Bereich.

## Ergebnisse und Zusammenarbeit mit anderen Stellen

Während der Laufzeit dieses Projekts haben wir eine Reihe bedeutender Fortschritte auf dem Gebiet der erklärbaren KI und der Robustheit gegen Adversarial Attacks erzielt. Zu diesen Ergebnissen gehören die Veröffentlichung 26 wissenschaftlicher Artikel auf internationalen Konferenzen, 3 Doktorarbeiten sowie 8 Bachelor- und Masterarbeiten und 2 eingeladene Vorträge. Unsere Arbeit wurde auch in zwei hochkarätigen Workshops vorgestellt, wodurch die Bedeutung und Relevanz unserer Forschung unterstrichen wird.

Zusätzlich zu diesen akademischen Ergebnissen haben wir auch den Code für unsere Forschung in öffentlich zugänglichen Repositories wie GitHub veröffentlicht. Dies erhöht nicht nur die Sichtbarkeit unserer Arbeit, sondern macht es auch anderen Forschenden leichter darauf aufzubauen und das Feld der XAI voranzubringen.

Wir sind besonders stolz auf die laufende Zusammenarbeit mit dem Deutschen Institut für Normung zur Erstellung des Standards 92001-3 für erklärbare KI. Diese Zusammenarbeit ist ein Beleg für die Bedeutung unserer Arbeit und unseres Engagements, das Feld voranzubringen.

Forschende, die an diesem Projekt gearbeitet haben, hatten auch die Möglichkeit, Praktika bei führenden Unternehmen wie Amazon, Adobe, Meta und AlgoLux zu absolvieren. Dies verschaffte unseren Forschern nicht nur wertvolle Erfahrungen, sondern trägt auch dazu bei, die Kluft zwischen Wissenschaft und Industrie zu überbrücken und sicherzustellen, dass unsere Arbeit auch in der Praxis Wirkung zeigt.

# Teil II

# Eingehende Darstellung

ExplAINN — 01IS19074

# Darstellung der durchgeführten Arbeiten

Die Definition der Arbeitspakete dieses Projekts solle die wichtigsten Herausforderungen im Bereich der erklärbaren KI und der Robustheit gegen Adversarial Attacks widerspiegeln. Diese Arbeitspakete waren:

1. **Erklärung bestehender Modelle:** Entwicklung von Methoden zur Erklärung der Entscheidungen von Black-Box-Modellen, mit dem Ziel, diese Modelle transparenter und interpretierbarer zu machen.
2. **Robustheitsanalysen:** Analyse der Robustheit von Modellen des maschinellen Lernens gegenüber Adversarial Perturbations und anderen Rauschquellen. Dies war ein entscheidender Aspekt unserer Forschung, da er uns half zu verstehen, inwieweit diese Faktoren die Modellleistung beeinflussen können.
3. **Entscheidungsgrenzenanalyse:** Untersuchung der Entscheidungsgrenzen von Modellen des maschinellen Lernens, mit dem Ziel zu verstehen, wie diese Modelle ihre Entscheidungen treffen.
4. **Erstellen von besser erklärbaren Modellen:** Entwicklung neuer Modelle für maschinelles Lernen, die durch ihren Aufbau interpretierbar sind, mit dem Ziel, es Anwendern zu erleichtern, diese Modelle zu verstehen und ihnen zu vertrauen.
5. **Erklärbarkeit mittels Visualisierung:** Erstellung von Visualisierungen als Mittel zur Erklärung zu Modellvorhersagen. Wir haben erkannt, dass Visualisierungen ein wirksames Mittel sein können, um komplexe Modelle transparenter und verständlicher zu machen, und haben daher Visualisierungen entwickelt, um den Vorhersageprozess besser zu interpretieren.

Jedes dieser Arbeitspakete spielte eine wichtige Rolle bei der Vertiefung unseres Verständnisses von erklärbarer KI und Modellrobustheit, und die Ergebnisse jedes Arbeitspakets haben dazu beigetragen, die Forschung in den anderen Arbeitspaketen zu informieren und anzuleiten.

## AP1: Erklärung bestehender Modelle

Das erste Arbeitspaket des Projekts konzentrierte sich auf Erklärungen für bestehende ML-Modelle. Wir sind dieses Thema systematisch angegangen: Zunächst haben wir die Literatur im Bereich XAI ausgewertet, um eine gemeinsame Sprache für den Vergleich von Beiträgen in diesem Bereich zu finden. Außerdem haben wir eine Literaturübersicht über zwei Bereiche

des generativen maschinellen Lernens erstellt, nämlich über die Generierung von Text zu Bildern und die Superauflösung von Bildern. Ziel dieser Übersichten ist es, die Defizite bei der Interpretierbarkeit aktueller Techniken und Modellarchitekturen zu ermitteln, die im aktuellen Stand der Forschung bestehen.

Wir verfolgten die Verarbeitungsschritte in den bestehenden Modellen und untersuchten den Einfluss von Normalisierung auf Zwischenergebnisse und verglichen sie mit alternativen Modellen ohne Normalisierungsschichten. Darüber hinaus wurden verschiedene Metriken zur Bestimmung der Interaktion zwischen Eingabedaten und der Trainingsphase von Modellen analysiert. Dazu gehörte die Untersuchung der Auswirkungen der Datensatzgröße auf Suchalgorithmen für neuronale Netzwerkarchitekturen und die Bewertung der Zuverlässigkeit von Feature-Attributionen (auch Merkmalszuordnungskarten oder Feature Attribution Maps), zur Vermittlung räumlicher Informationen mit faltenden neuronalen Netzen (Convolutional Neural Networks, CNNs). Schließlich wurde das Problem der Überanpassung erforscht, indem neuartige Architekturen untersucht wurden, die eine konkurrenzfähige Leistung zu CNNs gezeigt haben, insbesondere MLP-Mixer [Tol+21b]. Diese Modelle wurden durch Variation ihrer Größe innerhalb eines geringen Rechenbudgets ausgewertet, um die Faktoren zu verstehen, die ihre Tendenz zur Überanpassung beeinflussen.

## Literaturrecherche

Bevor wir uns mit den Merkmalen bestehender Modelle befassten, suchten wir zunächst in den bestehenden Definitionen nach Kernprinzipien, die uns helfen könnten zu verstehen, was im Kontext von XAI eine „Erklärung“ und was eine „Interpretation“ ist. Überraschenderweise haben wir festgestellt, dass es in der Literatur keinen Konsens darüber gibt, worauf sich diese Begriffe beziehen sollten. Infolgedessen definiert ein Großteil der Forschung ihre eigenen Ziele, was die Situation noch verschlimmert, die als „die Insassen leiten die Anstalt“ bekannt ist [MHS17]. Um eine einheitliche Definition für Kernprinzipien wie Erklärung und Interpretation zu finden, haben wir eine umfangreiche Literaturrecherche durchgeführt und die häufigsten Gemeinsamkeiten der bestehenden Definitionen identifiziert. Diese Eigenschaften wurden im Hinblick auf den Anwendungsbereich kontextualisiert, in dem XAI verankert ist: maschinelle Lernmodelle. Anschließend fassten wir diese Gemeinsamkeiten in zwei prägnanten Definitionen zusammen und integrierten sie in die Pipeline für maschinelles Lernen. Auf diese Weise konnten wir einen gemeinsamen Rahmen zur Beschreibung bestehender und zukünftiger Beiträge im Bereich der XAI schaffen, sodass sie fair miteinander verglichen werden können.

Es wurde argumentiert, dass nicht alles im Bereich des maschinellen Lernens erklärungsbedürftig sei [DK17]. Um kritische Aspekte des maschinellen Lernens zu identifizieren, die erklärungsbedürftige Methoden erfordern, haben wir Literaturrecherchen für zwei Bereiche des generativen maschinellen Lernens durchgeführt.

Der erste ist bekannt als Adversarial Text-zu-Bild-Generierung (auch T2I genannt). Bei T2I besteht die Aufgabe darin, Bilder zu erzeugen, welche die in natürlicher Sprache eingegebenen

Textbeschreibungen korrekt wiedergeben. Die Verwendung von Textbeschreibungen als Bedingung für die Bilderzeugung ist eine wesentlich flexiblere und besser interpretierbare Schnittstelle für den Menschen im Vergleich zu einzelnen Stichwörtern. Zunächst wurde festgestellt, dass Mechanismen entwickelt werden müssen, um **aussagekräftigere Textbeschreibungen** als Voraussetzung für den Bildgenerator zu kodieren. Dies kann teilweise mit mehr Daten (siehe AP2) oder besseren Texteinbettungen erreicht werden, wie sie von großen Sprachmodellen (LLMs) wie BERT [Dev+19] stammen. Im Vergleich zu früheren Methoden wie *GloVe* [PSM14] und *Word2Vec* [PSM14] verwendet BERT In-Kontext-Einbettungen, die es ermöglichen, zwischen den verschiedenen Bedeutungen zu unterscheiden, die ein einzelnes Wort haben kann (z. B. kann sich *Fliegen* auf den Akt des Fliegens oder auf die Insekten beziehen). Ein weiterer Aspekt, den diese Untersuchung aufgedeckt hat, ist die Tendenz einiger Methoden, die numerischen und Positionsinformationen in Bildunterschriften zu ignorieren, was die Erklärung des Generierungsprozesses erschwert. Glücklicherweise gibt es inzwischen mehrere Benchmarks, die die kompositorischen Fähigkeiten von T2I-Modellen testen [Par+21]. Da in der Beschriftung eines Bildes beliebige feingranulare Details spezifiziert werden können, wird der Abgleich zwischen dem generierten Objekt und seiner Eingabebeschreibung viel interpretierbarer. Daher wurden neuartige Architekturen benötigt, um diese umfassenden Textvorgaben zu kodieren. Dieser Herausforderung widmeten wir uns in AP4.

Der zweite Bereich der generativen ML den wir betrachteten ist die Superauflösung. Unter Superauflösung (Super-Resolution, SR) versteht man die Hochskalierung eines Bildes, häufig von niedriger auf hohe Auflösung, wobei versucht wird, den Inhalt und die Details so weit wie möglich zu erhalten, bzw. fehlende Details hinzuzufügen. Für diese Aufgabe wurde in der Literaturrecherche eine Lücke zwischen der menschlichen Wahrnehmung der Bildqualität und den für SR verwendeten Metriken festgestellt. Aktuelle Metriken wie das Spitzen-Signal-Rausch-Verhältnis (Peak Signal-to-Noise Ratio, PSNR) und die strukturelle Ähnlichkeit (Structural Similarity, SSIM) sind nur begrenzt in der Lage, die subjektiv wahrgenommene Qualität zu erfassen. Diese Metriken spiegeln die menschliche Wahrnehmung der Bildqualität nicht vollständig wider, sodass die Ergebnisse des Trainings mit diesen Metriken nicht direkt als Qualitätsmaßstab interpretiert werden können. Das Problem bei der Entwicklung einer objektiven Metrik, die die Wahrnehmung der Bildqualität erfasst, besteht darin, dass es keine allgemeine Übereinkunft darüber gibt, was eine gute Bildqualität ausmacht. Verschiedene Personen können unterschiedliche Meinungen darüber haben, was gut aussieht, und die menschliche Wahrnehmung wird auch von Faktoren wie den Sichtbedingungen und dem Kontext beeinflusst. Diese Uneinigkeit macht die Interpretation von SR-Ergebnissen sehr viel schwieriger und bleibt ein aktives Forschungsgebiet.

### Transformationen und Bedeutung der Parameter

Wir vertiefen die von uns vorgelegte Analyse, um zu erklären, wie sich die Zwischenergebnisse in einem Modell für maschinelles Lernen verändern, aber auch um den Einfluss der Modellparameter und ihrer Operationen zu ermitteln. Zu diesem Zweck haben wir eine Analyse des Einflusses durchgeführt, den die Parameter in Normalisierungsschichten beim Training von

Deep-Learning-Modellen haben. Die Batch-Normalisierung (BN) [IS15] ist die beliebteste Methode, die beim Deep Learning verwendet wird, um die Eingaben einer Schicht während des Trainingsprozesses für jeden Mini-Batch zu standardisieren. Sie wurde eingeführt, um das Problem der internen Kovarianzverschiebung zu beheben, das auftritt, wenn sich die Verteilung der Eingaben einer Schicht während des Lernprozesses des Netzwerks ändert, wodurch das Training verlangsamt oder sogar instabil wird. Durch die Normalisierung der Eingaben soll BN das Netz stabiler machen und seine Lerngeschwindigkeit und Leistung verbessern. Diese Interpretationen wurden jedoch infrage gestellt, indem gezeigt wurde, dass das Hinzufügen großer Mengen kovarianten Rauschens zu normalisierten Aktivierungen keine übermäßigen negativen Auswirkungen auf die Leistung von Modellen mit BN hat. Wir haben jedoch beobachtet, dass die Magnitude und Standardabweichung der Gradienten signifikant niedriger ist, wenn BN eingeführt wird. Wir modifizierten eine ResNet-Architektur [He+16], indem wir BN durch eine Kombination aus Gewichtsnormalisierung [SK16], Gradientenbeschränkung [Zha+20] und Dropout [Sri+14] ersetzten. Wir konnten hierbei zeigen, dass dieses Netzwerk die Leistung eines regulären ResNet-Modells mit BN erreicht, was die Erklärung, die ursprünglich für die Batch-Normalisierung gegeben wurde, weiter widerlegt.

### Überanpassung und Interaktion zwischen Daten und Training

Im folgenden Abschnitt werden verschiedene Aspekte im Zusammenhang mit der Überanpassung von Modellen, Beschneidungsmethoden und der Interaktion zwischen Daten und Training in Deep-Learning-Modellen behandelt. Insbesondere konzentrierte sich unsere Forschung auf diese Aspekte im Zusammenhang mit der Suche nach neuronalen Netzwerkarchitekturen, Feature-Attributionen und effizienten Implementierungen von mehrschichtigen Perzeptron-Netze (Multi-Layer Perceptrons, MLPs) für das Sehen.

Wir haben die Hauptaspekte untersucht, die sich auf die Überanpassung in einer neuen Modellfamilie, den MLP-Mixern, auswirken. MLP-Mixer [Tol+21a] sind eine Art von neuronaler Netzwerkarchitektur, die vor kurzem als Alternative zu faltenden neuronalen Netzwerken (Convolutional Neural Networks, CNNs) für Bildklassifizierungsaufgaben eingeführt wurde. MLP-Mixer basiert auf der Idee, eine Reihe von MLPs einzusetzen, um Bildbereiche über räumliche Dimensionen hinweg zu mischen und zu transformieren. In Anbetracht des großen Rechenaufwands für das Training solcher Modelle haben wir untersucht, wie anfällig MLP-Mixer für eine Überanpassung bei begrenztem Rechenaufwand ist. Wir haben zwei Erkenntnisse gewonnen. Erstens sind kleine Varianten der MLP-Mixer weniger anfällig für eine Überanpassung, während große Varianten eine stärkere Regularisierung erfordern, um die Tendenz zur Überanpassung zu mindern. Zweitens benötigen MLP-Mixer im Vergleich zu CNNs und Transformer-Modellen im Allgemeinen viel mehr Daten, um verallgemeinerbare Ergebnisse zu erzielen.

Üblicherweise ebenfalls besonders datenintensiv sind sog. Architektursuchverfahren. Unsere Untersuchungen in diesem Bereich hatten ein besseres Verständnis für die Beziehung zwischen Daten und Modellen zum Ziel, insb. im Hinblick auf die Leistung von Black-Box-Methoden bei reduzierter Datenmenge, denn bereits ML-Modelle mit festgelegter Modellarchitektur erfordern ein umfangreiches Training. Neuronale Netzwerkarchitektursuche (Neural Architecture

Search, NAS) [ZL17] ist eine Technik, die dazu dient, automatisch die beste neuronale Netzwerkarchitektur für eine bestimmte Aufgabe zu finden. NAS-Methoden erfordern in der Regel größere Trainingsdatensätze als bekannte Deep-Learning-Architekturen. Dies liegt daran, dass NAS eine rechenintensivere Methode ist, die das Trainieren und Bewerten vieler verschiedener Modelle erfordert, um die beste Architektur für eine bestimmte Aufgabe zu finden. Darüber hinaus wird bei NAS häufig ein größerer Raum möglicher Architekturen durchsucht, was mehr Daten für eine angemessene Untersuchung erfordern kann. Die Undurchsichtigkeit des Verhaltens von NAS und die Art und Weise, wie Architekturen erstellt werden, macht es schwierig, ihren Ergebnissen zu vertrauen. Wir untersuchten Methoden zur Verringerung der Datenmenge, die für die Suche nach einer Architektur verwendet wird, als eine vielversprechende Möglichkeit, erklärbare Modelle zu ermöglichen. Ein kleinerer Datensatz ermöglicht den Einsatz erfolgreicher Erklärungsstrategien, die auf dem Prinzip „Erklärung durch Beispiele“ beruhen [KKK16]. Darüber hinaus wird durch die Verwendung eines kleineren Datensatzes der Suchraum für mögliche Architekturen reduziert, wodurch es einfacher wird zu verstehen, warum eine bestimmte Architektur ausgewählt wurde. Wir haben eine neuartige Methode entwickelt, um die Größe eines Datensatzes erheblich zu reduzieren, ohne dass die Qualität der Architektur darunter leidet. Unsere Experimente zeigen, dass nur ein kleiner Teil eines Datensatzes für gängige NAS-Ansätze erforderlich ist und große Datensätze nur für das Training der endgültigen Architektur nützlich zu sein scheinen. Darüber hinaus haben wir festgestellt, dass es für NAS-Algorithmen wesentlich ist, nur solche Beispiele zu erhalten, welche leichter zu klassifizieren sind (basierend auf dem erzielten Fehlerwert), um eine leistungsfähige Architektur abzuleiten.

Um zu erklären, warum bestimmte architektonische Entscheidungen bei der Verwendung verschiedener NAS-Ansätze getroffen werden, haben wir eine umfassende empirische Analyse der Module durchgeführt, die bei den einzelnen NAS-Methoden am ehesten zustande kommen. Wir fanden heraus, dass sowohl Darts (V1 & V2) [LSY18] als auch ENAS [Pha+18] dazu neigen, Skip-Verbindungen zu erzeugen, was zu einer mittelmäßigen Leistung oder einem lokalen optimalen Zelldesign führen kann. Darüber hinaus haben wir festgestellt, dass GDAS [DY19] plausiblere und dennoch vielfältigere Architekturen ableitet, indem es mehr stochastische Variabilität in seinen Suchzyklus einbezieht.

Eines der am weitesten verbreiteten Interpretationswerkzeuge für Computer Vision sind sog. Feature-Attributionen (auch Merkmalszuordnungskarten oder Feature Attribution Maps), die am häufigsten als „Heatmaps“ bezeichnet werden. Diese Methoden sind beliebt, weil sie versprechen, die Beziehung zwischen dem Modell und dem Teil der Eingabedaten, der eine Vorhersage auslöst, zu vermitteln. Es gibt zahlreiche Forschungsarbeiten [Bar+20; Sam+21], in denen Methoden entwickelt wurden, die die Bedeutung der Eingangspixel für eine bestimmte Vorhersage abbilden können. Jüngste Arbeiten wie ROAR [Hoo+19] haben jedoch gezeigt, dass diese Methoden der Merkmalszuordnung nicht so zuverlässig sind wie ursprünglich angenommen. Die Herausforderung wird durch einen Versuchsaufbau unterstützt, bei dem die Eingabe auf der Grundlage einer anfänglichen Schätzung der Wichtigkeit schrittweise verdeckt wird. Wir erweiterten die Analyse von ROAR, indem wir einen iterativen Prozess zur Schätzung der Wichtigkeit eines gegebenen Beispiels vorschlugen, um zu zeigen, dass Methoden zur Merkmalsschätzung (ohne die richtige Interpretation) eine zu plumpe Schätzung der Wichtigkeit liefern. Wir schlugen ein Evaluierungsprotokoll vor, mit dem gründlich getestet

wird, wie viele Informationen eine Heatmap liefert, die mithilfe von Methoden zur Zuordnung der Bedeutung von Merkmalen erstellt wurde. Dabei stellten wir stellten fest, dass diese Feature-Attributionen häufig in Bereiche fallen, die keine Informationen vermitteln, was der Ad-hoc-Interpretation widerspricht, die diesen Zuordnungsstrategien gegeben wurde.

## AP2: Robustheitsanalysen

Dieser Abschnitt beschreibt unsere Arbeiten, welche sich mit der Robustheit von Modellen des maschinellen Lernens befassten. Adversarial Examples sind Eingaben, die speziell darauf ausgelegt sind, ein maschinelles Lernmodell in die Irre zu führen. Diese Beispiele können durch Hinzufügen unmerklicher Störungen, sog. Adversarial Perturbations, zu den Eingaben erstellt werden, die das Modell zu falschen Vorhersagen veranlassen können. Dies ist ein ernsthaftes Problem bei realen Anwendungen des maschinellen Lernens, da böswillige Akteure solche Beispiele verwenden können, um Sicherheitsmaßnahmen zu umgehen oder Ergebnisse zu manipulieren. Wir haben verschiedene Techniken zur Erfassung der Auswirkungen von Adversarial Perturbations evaluiert, darunter unterschiedliche Arten von Rauschen, die ein Modell zu Fehleinschätzungen verleiten können. Darüber hinaus untersuchten wir, wie sich die Veränderung von Redundanzen innerhalb der Modelle auf die Robustheit auswirken kann, und erforschten Verteidigungsmechanismen, die eingesetzt werden können, um die Modelle robuster gegenüber Störungen durch Angreifer zu machen.

### Modellredundanzen

Unter Anomalieerkennung versteht man die Aufgabe, Datenpunkte zu identifizieren, die von der Norm oder dem erwarteten Verhalten eines Systems abweichen. In der Praxis gibt es viele Anwendungen, die von der Erkennung von Betrug bei Finanztransaktionen bis zur Fehlererkennung in industriellen Systemen reichen. Ein Ansatz zur Erkennung von Anomalien ist die Verwendung von Autoencodern, d. h. neuronalen Netzen, die darauf trainiert sind, ihre Eingabedaten zu rekonstruieren. Die Idee dahinter ist, dass der Rekonstruktionsfehler für anomale Daten höher ist als für normale Daten. Standard-Autoencoder sind jedoch nicht unbedingt robust gegenüber Angriffen oder verrauschten Daten, was bei der Erkennung von Anomalien zu falsch-positiven und falsch-negativen Ergebnissen führt. Darüber hinaus sind ihre Vorhersagen in der Regel undurchsichtig, da die Anomalien auf der Grundlage des Fehlers pro Beispiel markiert werden, was keinen Raum lässt, um zu bestimmen, was genau an einem Beispiel anomal ist. Mit anderen Worten: Anomalievorhersagen von regulären Autoencodern gehen nicht auf die Frage ein, warum ein Beispiel anomal sein könnte. Um beide Probleme anzugehen, schlugen wir eine Modifikation des traditionellen Aufbaus für die Erkennung von Anomalien bei tabellarischen Daten vor. Insbesondere verwendeten wir einen entrauschenden Autoencoder und eine zellenbasierte Verlustfunktion (Tabellenzellen), um die Rekonstruktion einzelner Beispiele zu bewerten. Dieser Aufbau ist nicht nur robuster gegenüber kleinen Änderungen im Eingaberaum, sondern ermöglicht auch eine differenzierte Identifizierung von Anomalien, die auf Merkmalsebene (und nicht nur auf Beispielebene) auftreten. Zusammenfassend lässt sich

sagen, dass die zellenweise Erkennung von Anomalien nicht nur die Frage beantwortet, welche Beispiele anomal sind, sondern auch die Frage, warum es sich um eine Anomalie handelt.

Im Bereich der Suche nach neuronalen Architekturen haben unsere Experimente auch einen Nachteil der Verwendung von Proxy-Datensätzen aufgedeckt, die kleiner sind als die Originaldaten. Konkret bestätigten wir, dass DARTS und ENAS bei der Generierung von Zellarchitekturen weniger robust sind als GDAS und eher dazu neigen, bei ihren Entwürfen in lokalen Optima stecken zu bleiben. Im Gegensatz dazu war GDAS beim Entwurf von hochleistungsfähigen Zellen robuster und konnte dank seiner stochastischen Variabilität während der Suche zu vielfältigeren Architekturen konvergieren.

Bei ML-Modellen ist mangelnde Robustheit oft auf eine ungleichmäßige Auswahl der Trainingsdaten zurückzuführen. Unausgewogene Datensätze (mit deutlich mehr Beispielen für eine Klasse als für die anderen) führen dazu, dass Modelle lernen, die häufigsten Klassen vorherzusagen, auch bekannt als „Shortcut Learning“ [Gei+20]. Ein besonders extremes Beispiel hierfür fanden wir im Bereich Visual Question Answering (VQA) einer der Goldstandards zur Bewertung von Modellen der VQA2.0-Datensatz [Goy+17]. Wir untersuchten die Auswirkungen eines auffälligen Ungleichgewichts in diesem Datensatz, bei dem polare (P) Fragen, d. h. solche, die mit „Ja“ oder „Nein“ beantwortet werden können, im Vergleich zu den übrigen nicht-polaren (NP) Klassen um das 950-fache überrepräsentiert sind (0,02 % aller Beispiele pro NP-Klasse gegenüber 19 % pro P-Klasse). Unsere umfangreichen empirischen Auswertungen bestätigen, dass ein modernes VQA-Modell unter idealen Bedingungen, d. h. wenn polare und nicht-polare Fragen getrennt verarbeitet werden, weniger als 1 Prozentpunkt Unterschied aufweist. Diese Ergebnisse zeigen, dass VQA-Modelle robust gegenüber einer Überrepräsentation polarer Beispiele sind. In AP3 beschreiben wir eine Erweiterung dieser Arbeit, die den Merkmalsraum und die Trennung zwischen polaren und nicht-polaren Beispielen analysiert.

Während der Arbeiten an Arbeitspaket 4 fanden wir eine wirksame Methode zur Schaffung neuartiger Architekturen, die durch eine Neuinterpretation von Daten auf redundante Weise und ihr Design interpretierbar sind. Konkret schlagen wir die Definition von Hilfsaufgaben vor, die Informationen aus dem ursprünglichen Datensatz wiederverwenden, aber die Beziehung zwischen jedem Datenpunkt und seiner ursprünglichen Annotation ändern. Die Hilfsaufgaben werden durch kleine Teilmodelle gelöst, die mit einer Hauptarchitektur verbunden sind (z. B. ResNet50 oder DenseNet). Wir haben festgestellt, dass diese Redundanzen zu kürzeren Trainingsphasen und konsistenteren Leistungsmetriken führen (z. B. Fehlerwerte mit geringerer Variabilität zwischen den Läufen). Daher ist es wichtig zu messen, wie viel robuster diese Modelle dank der Zusatzaufgaben werden. Im nächsten Abschnitt berichten wir über die Ergebnisse einer Bewertung dieser Modelle gegen Adversarial Attacks.

## Adversarial Perturbations

Adversarial Perturbations sind nicht wahrnehmbare Veränderungen im Eingaberaum, die ein Modell zu einem Fehler verleiten. Diese Störungen können von böswilligen Akteuren ausgenutzt werden, um ein automatisches Vorhersagemodell absichtlich zu täuschen. Auch andere Arten von Störungen wie Rauschen, Zoom, Unschärfe und affine Transformationen werden

häufig verwendet und wirken sich nachweislich negativ auf die Leistung von ML-Modellen aus. Das Verständnis der Funktionsweise von Adversarial Attacks ist daher von entscheidender Bedeutung für den Bereich ML, um die richtigen Methoden zur Abschwächung (oder vollständigen Vermeidung) der Auswirkungen dieser Störungen zu entwickeln.

Für dieses Projekt haben wir frühere Arbeiten zu Bildtranskodierern erweitert, die sich darauf konzentrieren, nur die Informationen darzustellen, die zur Lösung der Aufgabe benötigt werden. Bei einem Farbbild der Größe $256 \times 256$ ist es zum Beispiel unwahrscheinlich, dass ein Objektklassifikator alle Informationen in diesem Bild benötigt, um eine zuverlässige Aussage zu treffen. Ausgehend von einem vortrainierten Autoencoder und einem Bildklassifikator wird der Decoderteil auf der Grundlage der Gradienten des Klassifikators aktualisiert. Wir evaluierten einen Aufbau, bei dem ein solcher fein abgestimmter Autoencoder vor einem Klassifikator als Verteidigungsmechanismus gegen feindliche Angriffe eingesetzt wird. Unsere Analyse zeigte, dass Adversarial Attacks die Gradienten dieses Ensembles nicht effektiv ausnutzen können. Dies liegt an der Natur der Gradienten selbst, die eher strukturelle als semantische Informationen vermitteln. Während ein Klassifikator beispielsweise Informationen über die Farbe und die Textur einer kreisförmigen Form benötigt, um festzustellen, dass es sich um einen Tennisball handelt, erzeugt der Autoencoder (und insbesondere der Decoder) starke Gradienten für harte Kanten, die die Kreisform beschreiben, aber nicht auf der Textur oder der Farbe. Wir haben die Robustheit dieses Ensembles für verschiedene neuronale Netzwerkarchitekturen wie ResNet50 und Inception-v3 untersucht. Die Ergebnisse von mehreren gradientenbasierten Angriffen zeigten, dass diese Ensembles die Modelle viel robuster gegenüber Störungen machen, die durch verschiedene Angriffe wie FGSM, BIM und C&W verursacht werden.

Zusätzlich zu den traditionellen Klassifikatoren haben wir auch eine der von uns vorgeschlagenen interpretierbaren Methoden aus AP4 evaluiert. Unsere allgemeine Lösung für besser interpretierbare Modelle, wie sie in diesem Projekt definiert wurde, erfordert die Einbeziehung zusätzlicher Strukturen in das Modell, die den impliziten Einschränkungen der ursprünglichen Aufgabe entsprechen (z. B. Rotationsinvarianz für die Objekterkennung oder Auflösungskovarianz). Die resultierenden Architekturen lösen daher ein Optimierungsproblem mit mehr Einschränkungen. Es ist daher plausibel, dass diese Beschränkungen zu Modellen führen, die robuster gegen Angriffe sind. Aus diesem Grund bezogen wir eine Bewertung unserer vorgeschlagenen interpretierbaren Modelle aus AP4 gegen diese Angriffe ein. Insbesondere evaluieren wir ein selbstüberwachtes Hilfsmodell (Self-Supervised Auxiliary, SSAL), das auf einer ResNet-Architektur basiert, gegen einstufige FGSM- und iterative (PGD-) Angriffe. Die Ergebnisse zeigten, dass Angriffe gegen diese Modelle immer noch erfolgreich sind. Trotz der zusätzlichen Robustheit von SSAL-Modellen konnten wir schlussfolgern, dass Adversarial Attacks Schwachstellen des Modells ausnutzen, die nicht unbedingt mit interpretierbaren Einschränkungen übereinstimmen. Für die Zukunft planen wir die Erforschung von Einschränkungen, die die Mechanismen der menschlichen Wahrnehmung berücksichtigen. Wir stellen die Hypothese auf, dass Einschränkungen aus diesem Bereich die Effektivität Adversarial Attacks drastisch reduzieren können.

Neben Adversarial Attacks gibt es noch andere Arten von Störungen, die sich auf die Leistung von Modellen des maschinellen Lernens auswirken können, wie z. B. Unordnung, Gaußsches

Rauschen, Verteilungsverschiebung, Objektskalierung oder unterschiedliche Textausdrücke. Diese Störungen können die Lernfähigkeit von Modellen stark beeinträchtigen, was zu einer schlechten Leistung auf Testdaten führt. Darüber hinaus führt die Vernachlässigung dieser Faktoren häufig zu falschen Vorhersagen, die mit herkömmlichen XAI-Methoden (z. B. durch die Verwendung von Merkmalszuweisungskarten) nur unzureichend verstanden werden können. Um diese Herausforderungen zu überwinden, wurden mehrere Schutzmechanismen vorgeschlagen, darunter räumliche Transformatorennetzwerke für Unordnung, Korrespondenzabgleich für Skalierung, Feinabstimmung für Daten-zu-Text-Generatoren und neuartige differenzierbare Audiodarstellung für Gaußsches Rauschen. Wir untersuchten die Interpretierbarkeit dieser Methoden, um den genannten Störungen entgegenzuwirken, und ermittelten, welche davon zu robusteren Modellen führen.

Die Klassifizierung von Bildern in komplexen Szenen ist eine schwierige Aufgabe in der Bildverarbeitung, bei der die Bilder nicht nur das relevante Objekt, sondern auch ablenkende Unordnung enthalten, was zu einer schlechten Genauigkeit führen kann. Die Fähigkeit, sich selektiv auf das gesuchte Objekt zu konzentrieren und die Umgebung zu unterdrücken, ist in diesem Fall entscheidend für eine genaue Klassifizierung. In diesem Projekt wurden neuartige Architekturen für die Bildklassifizierung mit Mechanismen zur Erkennung von Unordnung und zur Fokussierung auf das relevante Objekt ausgestattet. Diese Ansätze können den Modellen helfen, besser zwischen dem relevanten Objekt und den umgebenden Störfaktoren zu unterscheiden, was zu robusteren und besser interpretierbaren Klassifizierungsergebnissen führt. Wir haben verschiedene Varianten von CNNs kleiner und mittlerer Größe auf der Basis von LeNet- und ResNet-Architekturen auf Datensätzen mit kontrollierbarem synthetischem Rauschen (Clutter MNIST und Clutter FashionMNIST [MHG+14]), aber auch auf Datensätzen mit komplexen Szenen (Pascal VOC [Eve+10]) evaluiert. Unsere Experimente zeigten, dass es möglich ist, verglichen mit einem Black-Box-Klassifikator, eine robustere Leistung zu erzielen, wenn man sich auf diese Prinzipien verlässt. Bei den synthetischen Datensätzen erreichten wir eine Verbesserung der Modellgenauigkeit um bis zu 41 Prozentpunkte und bei den komplexen Szenen des Pascal VOC-Datensatzes um bis zu 3,6 Prozentpunkte.

Andere natürlich vorkommende Arten von Störungen sind Größe und Verdeckung. Objekten, die weit von der Kamera entfernt oder aus anderen Gründen klein sind (z. B. ein Flugzeug am Himmel vom Boden aus gesehen), mangelt es an strukturellen Details, die ein Modell lernt, wenn es mit großformatigen Versionen derselben Objekte trainiert wird (z. B. ein Flugzeug am Boden). Dieses Problem stellt eine besondere Herausforderung bei der Verfolgung von Objekten in einer Videosequenz dar. Moderne Verfolgungsmodelle stoßen bei der Verfolgung von Objekten über lange Zeiträume an ihre Grenzen, und Verdeckungen, die zu einem späteren Zeitpunkt in der Videosequenz auftreten, sind oft schwer zu erkennen. Wir haben ein neuartiges interpretierbares Prinzip, welches auf Referenzvergleichen basiert, vorgeschlagen, um die Auswirkungen solcher kleinen verdeckten Objekte abzuschwächen, was zu einer robusteren Verfolgung führt (weitere Details in AP4). Experimente mit den anspruchsvollen YouTube-VOS [Xu+18]- und DAVIS-Datensätzen [Per+16] zeigen eine Verbesserung von 0,7 bzw. 1,3 Punkten des mittleren $\mathcal{J}$- & $\mathcal{F}$-Scores im Vergleich zum Stand der Technik.

Bildklassifikatoren sind dafür bekannt, dass sie in kontrollierten Umgebungen gut funktionieren, ihre Leistung kann sich jedoch schnell verschlechtern, wenn sie mit Daten konfrontiert

werden, die sich stark von der beim Training beobachteten Datenverteilung unterscheiden. Dies wird als Verteilungsverschiebung bezeichnet und kann durch Veränderungen in der Umgebung, im Datenerfassungsprozess oder durch andere Faktoren verursacht werden. In diesem Projekt haben wir verschiedene Ansätze zur Verbesserung der Robustheit von Videoklassifikatoren durch Testzeitanpassung (Test Time Adaptation, TTA) von Batch-Normalisierungsschichten oder einer zusätzlichen selbstüberwachten Aufgabe bewertet. Durch die Anpassung dieser Schichten zur Testzeit ist es möglich, die Robustheit des Modells gegenüber Verteilungsverschiebungen zu verbessern und seine Gesamtleistung zu steigern, ohne dass die Kosten für ein erneutes Training des Modells anfallen. Zu den Methoden zur Durchführung von TTA gehören das Ersetzen der Statistiken der Batch-Normalisierungsschichten durch die des Eingangsvideos zum Testzeitpunkt, die Aktualisierung sowohl der Statistiken als auch der Skalierungsparameter aller Normalisierungsschichten oder die Verwendung einer zusätzlichen selbstüberwachten Aufgabe (z. B. Rotationsschätzung) zur Feinabstimmung des gemeinsamen Merkmalsextraktors, bevor dessen Ausgabe an den Klassifikationskopf weitergeleitet wird. Wir haben gezeigt, dass es möglich, wenn auch nicht trivial ist, die Robustheit von Videoklassifikatoren gegen verschiedene Arten von Störungen wie Gaußsches Rauschen, Bewegungsunschärfe, Schnee- und Nebelartefakte zu verbessern. Wir haben festgestellt, dass alle drei Strategien besser sind als der Verzicht auf jegliche Art von Anpassung, jedoch ist kein einzelner Ansatz ist den anderen für alle Arten von Störungen überlegen.

Bei vielen Aufgaben der Verarbeitung natürlicher Sprache, z. B. bei der Umwandlung von Daten in Text und von Text in Bilder, ist die Fähigkeit, konsistente Ausgaben für dieselbe Eingabeanfrage zu erzeugen, von entscheidender Bedeutung. Dies kann jedoch eine Herausforderung sein, da verschiedene Formulierungen derselben Anfrage unterschiedliche Nuancen und Ausdrücke haben können, was zu uninterpretierbaren Abweichungen in der gewünschten Ausgabe führt. Daher ist es wichtig, dass Modelle robust gegenüber solchen Variationen sind, um genaue und vertrauenswürdige Ergebnisse zu produzieren. Wir schlugen zwei Wege vor, um dieses Problem anzugehen, einen pro Anwendungsbereich. Für Daten-zu-Text (d. h. für die Erstellung von Bildunterschriften auf der Grundlage von Tabellendaten) schlugen wir einen strukturierten Such- und Lernansatz vor, bei dem vorab trainierte Sprachmodelle verwendet werden, die später durch das Einfügen fehlender Werte in die generierten Bildunterschriften feinabgestimmt werden, um die Textqualität zu verbessern und gleichzeitig die Effizienz der Inferenz beizubehalten. Experimente zeigten, dass das von uns vorgeschlagene Modell eine hohe Leistung auf E2E- und WikiBio-Datensätzen erzielt, 98,35 % der Eingabeslots auf E2E abdeckt und das Problem der geringen Abdeckung effektiv entschärft. Für die Text-zu-Bild-Generierung erweitern wir den MS-COCO-Datensatz (der üblicherweise zum Trainieren von Text-zu-Bild-Modellen verwendet wird) um einen VQA-Datensatz. Diese Erweiterung erfordert eine Änderung der Modellarchitektur, um eine einheitliche Architektur zu schaffen, die Merkmale aus beiden Aufgaben erzeugt, die für die Konditionierung des Bilderzeugungsmoduls nützlicher sind. Die von uns vorgeschlagene Methode senkte den FID von 27,84 auf 25,38 und erhöht die R-Präzision um fast 1 Prozentpunkt im Vergleich zur Ausgangslösung, was darauf hindeutet, dass die T2I-Synthese durch die Nutzung von Daten (und Architekturen) aus dem Bereich der VQA erfolgreich verbessert werden kann.

Die Klassifizierung von Umgebungsgeräuschen (Environmental Sound Classification, ESC) ist eine anspruchsvolle Aufgabe, die eine korrekte Unterscheidung zwischen Geräuschklassen vor-

aussetzt, die im täglichen Leben vorkommen (z. B. „Niesen“, „Flugzeug“, „Presslufthammer“, „Katze“, „laufender Motor“, „Zähneputzen“, „Straßenmusik“). Um dieses Problem zu lösen, wurde eine Vielzahl von leistungsstarken ML-Modellen entwickelt. Die überwiegende Mehrheit dieser Modelle transformiert jedoch das Eingangssignal vor der eigentlichen Verarbeitung in eine zweidimensionale, bildähnliche Darstellung (Spektrogramme) unter Verwendung der Kurzzeit-Fourier-Transformation (Short-Time Fourier Transform, STFT). Wenn diese Modelle „in freier Wildbahn“ angewendet werden, neigen sie zu geringer Robustheit gegenüber additivem weißem Gaußschen Rauschen (Additive White Gaussian Noise, AWGN), welches eine recht häufige Art der Signalverzerrung ist. Um dieses Problem zu lösen, wurde eine adaptive Zeit-Frequenz-Transformation auf der Grundlage Frequenz-B-Spline (FBSP) Wavelets im komplexen Zahlenraum vorgeschlagen. Eine solche Darstellung hat eine inhärente Robustheit gegenüber AWGN, was auch für die endgültige Architektur gilt. Während des Trainings lernt die FBSP-Schicht Filterparameter, die in Bezug auf die Eingabedaten optimal sind, wodurch das Ergebnis des Modells für die Zielklassen maximiert und die Reaktion auf Rauschkomponenten abgeschwächt wird. Wir zeigten, dass dieser Ansatz die Robustheit eines auf ESC-50 [Pic15] und US8K [SJB14] trainierten Modells erhöht, in unseren Experimenten um bis zu 6 Prozentpunkte Genauigkeit bei Signal-Rausch-Verhältnissen zwischen 0 und -5 dB.

## AP3: Entscheidungsgrenzenanalyse

In diesem Abschnitt erörtern wir unsere Beiträge zur Erklärbarkeit, die durch die Analyse von Entscheidungsgrenzen in Modellen des maschinellen Lernens gewonnen wurden. Entscheidungsgrenzen sind die Regionen im Merkmalsraum, die verschiedene Klassen voneinander trennen. Das Verständnis der Entscheidungsgrenzen eines Modells ist entscheidend für die Interpretation seines Verhaltens, die Erkennung von Schwächen und die Verbesserung seiner Leistung. In diesem Zusammenhang haben wir untersucht, wie Modelle mit verschiedenen Methoden ähnliche Lösungen finden können, wie man feststellt, welche Bereiche des Merkmalsraums durch Trainingsdaten gut abgedeckt sind, und wie man Methoden entwickelt, um die Entscheidungsgrenzen durch Änderung einer minimalen Anzahl von Parametern zu modifizieren. Darüber hinaus haben wir untersucht, wie sich feststellen lässt, welche Grenzen robust und welche instabil sind, um Modelle zur Erkennung unbekannter Klassen zu erweitern.

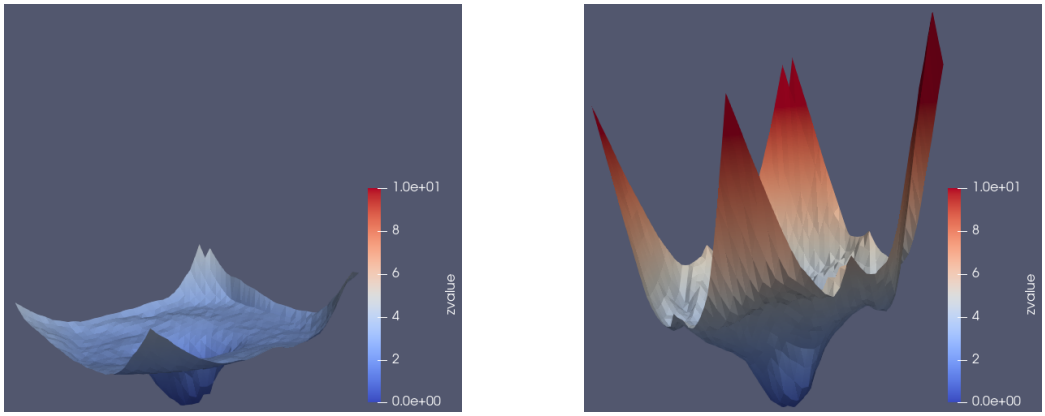
### Ähnliche Lösungen mit verschiedenen Methoden

Für dieses Arbeitspaket haben wir neue Modelle entwickelt, die die Entscheidungsgrenzen verändern. Wir haben drei Ansätze für verschiedene Aufgaben vorgestellt. Der erste Ansatz nutzt Entscheidungsgrenzen, die aus dem visuellen Bereich extrahiert und auf den Audiobereich angewendet werden. Der zweite Ansatz modifiziert die Entscheidungsgrenzen der Verlustfunktion bei Videoobjektsegmentierung (Video Object Segmentation, VOS). Der letzte Ansatz schlägt eine Alternative zu Batch-Normalisierung für die Normalisierung neuronaler Netze bei der Objektklassifizierung vor.

Erstens haben wir ein CNN-Modell für die Klassifizierung von städtischen Umgebungsgeräuschen entwickelt, das dem Stand von Wissenschaft und Technik entspricht und robust gegenüber Rauschen ist (wie in AP2 bewertet). Wir haben eine FBSP-Schicht vorgeschlagen, deren Parameter auf den Filtern (Zentralfrequenz, Bandbreite, Phasenverschiebung und Bias) basieren, um die Abdeckung der Entscheidungsgrenzen zu maximieren. Man beachte, dass die neue Schicht nur vier Parameter hat, was den Freiheitsgrad reduziert. Daher führen starke klassenspezifische Signale zu erheblichen Aktualisierungen der Parameter während der Trainingsphase. Darüber hinaus wirkt sich ein großer Bilddatensatz wie Imagenet, aufgrund der großen Vielfalt an Formen und visueller Konzepte, positiv auf unser Modell zur Parameterinitialisierung aus, da es eine große Vielfalt an zweidimensionalen Formen visueller Konzepte gibt. Im Gegensatz dazu verfügt das Training nur mit dem Audiospektrogramm nicht über genügend differenzierende Merkmale, um feingranulare Unterschiede zwischen einigen Klassen in den Entscheidungsgrenzen zu erlernen, und führt möglicherweise zum Zusammenbruch der entsprechenden Entscheidungsgrenzen zu einzelnen Punkten.

Zweitens haben wir die Entscheidungsgrenzen eines VOS-Modells von einem binären Klassifizierungsfehler zu einem Multi-Klassen-Grenzfehler geändert. Die binäre Klassifizierungsfehlerfunktion verwendet zwei Klassen (Vordergrund und Hintergrund), wobei keine zusätzlichen Informationen, wie z. B. die Pixelposition, genutzt werden. Im Gegensatz dazu haben wir die Grenzklassifizierungsfehlerfunktion erweitert, indem wir die Objektgrenze auf der Grundlage eines Clustering-Algorithmus in mehrere Klassen aufteilen. Die zusätzlichen Positionsinformationen helfen dem VOS-Modell, Objektgrenzen zu lernen, die schwieriger zu segmentieren sind. Durch das Erlernen der relativen Position der Pixel wird das Modell außerdem implizit in die Lage versetzt, Pixeln um die Objektgrenze herum eine höhere Gewichtung zuzuweisen, da sie schwieriger zu segmentieren sind.

Drittens sind Batch-Normalisierungsschichten ein Standardblock für eine breite Palette von CNN-basierten Modellen, welche die Stabilität und Trainingsgeschwindigkeit bei gradientenbasierten Lernansätzen verbessern. Wie bereits in AP1 beschrieben, bestand ihr Zweck ursprünglich darin, die interne Kovarianzverschiebung zu verringern, die während des Trainings eines neuronalen Netzes auftritt und dessen endgültige Leistung negativ beeinflussen kann. Batch-Normalisierung erfordert jedoch einen ausreichend großen Mini-Batch, um ordnungsgemäß zu funktionieren, und verstößt gegen die Annahme der Unabhängigkeit der Beispiele innerhalb des Mini-Batches, die Grundlage für den Multi-GPU-Betrieb ist. Daher haben wir Alternativen zu Batch-Normalisierung untersucht, die diese Nachteile nicht aufweisen und gleichzeitig das gleiche Maß an Stabilität und Leistung wie bisher bieten. In dieser Arbeit haben wir eine normalisierungsfreie Architektur auf der Grundlage von ResNet vorgeschlagen, die bessere Ergebnisse als das Original erzielt. Das vorgeschlagene Modell nutzt Gewichtsnormalisierung, Gradientenbeschneidung und Dropout, um die Batch-Normalisierung zu ersetzen. Unsere Analyse der beiden Modelle zeigte, dass bei der Verwendung von Batch-Normalisierung die Magnitude und die Standardabweichung der Gradienten deutlich reduziert werden. Schließlich zeigt Abbildung 1 die Entscheidungsgrenzen beider Modelle in Form ihrer Fehleroberfläche. Unser normalisierungsfreies ResNet weist größere Spitzen zu den Rändern der Entscheidungslandschaft hin auf, was sich jedoch nicht auf seine Genauigkeit auswirkt.




(a) (b)

**Abb. 1** Fehlerlandschaft eines ResNet 34, mit (a) und ohne (b) Batch-Normalisierung. Trotz deutlich höherer Spitzen zu den Rändern hin erreicht das Modell ohne Batch-Normalisierung eine höhere Genauigkeit.

## Abdeckung durch Trainingsdaten

Die Entscheidungsgrenzen werden durch die Lösung eines komplexen Optimierungsprozesses bestimmt, der weitgehend von den Beispielen abhängt, die zum Trainieren des Modells verwendet werden. Beispiele mit homogenen Merkmalen sind im Merkmalsraum zusammengeballt und weit von jeder Entscheidungsgrenze entfernt. Im Gegensatz dazu liegen Beispiele, die schwieriger zu klassifizieren oder mehrdeutig sind oder sich auf andere Weise von anderen innerhalb ihrer Klasse unterscheiden, mit größerer Wahrscheinlichkeit in der Nähe von Entscheidungsgrenzen. Die Qualität und die Abdeckung der Beispiele in Kombination mit ML-Modellen sind immer noch eine offene Herausforderung. Dieser Abschnitt beschreibt unsere Arbeiten, die sich mit der Messung der Qualität der Beispiele für VQA-Aufgaben befassen. Außerdem haben wir die Modalitätsabdeckung von Contrastive Language-Image Pretraining (CLIP), einem Modell zum Erlernen von Assoziationen zwischen Bildern und Texten, erweitert. Schließlich haben wir die qualitativen Vorteile der Kombination von Adjektiven und Substantiven bei der Erstellung von Einbettungen untersucht.

Bei VQA ist die Qualität der Beispiele angesichts der Art und Weise, wie sie gekennzeichnet sind, eine besondere Herausforderung. Darüber hinaus hilft die Identifizierung der Art der Fragen den VQA-Modellen, das Training mit mehrdeutigen Beispielen zu vermeiden, welche die Leistung beeinträchtigen und zu Bias führen können. Wir haben eine Metrik namens EAsE entwickelt, die in hohem Maße mit der menschlichen Einschätzung der Schwierigkeit korreliert. Die von uns vorgeschlagene Metrik verwendet die inverse Entropie von Einbettungsgruppen, die durch Clustering-Algorithmen und Worteinbettungsmodelle gebildet werden.

In der Umweltgeräuschklassifikation (Environmental Sound Classification, ESC) haben wir das CLIP Verfahren erweitert. Es handelt sich dabei um ein bimodales (Bild und Text) Modell, welches die Korrespondenz zwischen verschiedenen Darstellungen (Modalitäten) desselben Konzepts durch die Verwendung eines gemeinsamen Einbettungsraums erlernt. CLIP deckt eine

breite Palette von Konzepten ab, wodurch ein Kollabieren der Entscheidungsgrenzen verhindert wird. Darüber hinaus werden zwei Teilnetze (visuell und textuell) zum Erlernen eines multimodalen Einbettungsraums verwendet. Auf diese Weise kann jede unterstützte Modalität als Ziel oder Anfrage verwendet werden (z. B. der Name einer Klasse von Interesse), während der andere Teil des Modells (in diesem Fall der visuelle) Eingabeeinbettungen erzeugt, die anhand der Ähnlichkeit der Anfrage angeordnet werden können. Das beschriebene Anfrageverfahren bezieht sich auf eine reguläre Bildklassifizierung, ist aber im Gegensatz zur herkömmlichen Methodik nicht auf eine vordefinierte Menge von Klassen beschränkt. Vor diesem Hintergrund haben wir zu CLIP eine zusätzliche Modalität (z. B. Audio) hinzugefügt, wodurch das Modell mehr Daten abdeckt. In diesem Fall basiert das Training des Audiokopfes auf Wissensdestillation, wobei Text- und Bildkopf als Lehrernetzwerke für den Schülerkopf (Audio) dienen. Letzteres profitierte von gut etablierten, vortrainierten Entscheidungsgrenzen, was sich in der besseren Leistung im Vergleich zu dem in einem eigenständigen Setup trainierten Audiokopf widerspiegelte.

Das Extrahieren von Informationen aus visuellen Inhalten umfasst das Erkennen von Objekten und das Verstehen von Emotionen und Gefühlen, die in den Daten enthalten sind. In der Regel werden Objekte durch Substantive dargestellt, und Adjektive stehen für Gefühle. Daher werden Adjektiv-Nomen-Paare (ANPs) verwendet, um visuelle Merkmale auf niedriger Ebene mit einer durch visuelle Inhalte ausgedrückten Stimmung auf hoher Ebene zu verknüpfen. Infolgedessen werden ANPs häufig in der visuellen affektiven Datenverarbeitung und der visuellen Stimmungsanalyse eingesetzt. In der Computerlinguistik (Natural Language Processing, NLP) ist das Word2Vec-Modell ein gängiger Ansatz für die Berechnung von Worteinbettungen und wird auch für die Ermittlung von ANP-Einbettungen verwendet. Mithilfe eines vortrainierten Word2Vec-Modells werden die einzelnen Einbettungen eines Adjektivs und eines Substantivs abgerufen und kombiniert, um ANP-Einbettungen zu erhalten. Word2Vec behandelt jedoch Adjektive und Substantive nicht als unterschiedlich und legt sie in denselben Einbettungsraum. Daher war es eine Herausforderung, Adjektiv- und Substantiveinbettungen von Word2Vec zu trennen und zu erklären. Um diese Lücke zu schließen, haben wir ein ANP-W2V-Modell vorgeschlagen, das Adjektive und Substantive unabhängig voneinander behandelt und sie verschiedenen Einbettungsräumen zuordnet. Unser Modell erzeugt getrennte Einbettungen für Adjektive und Substantive als Ausgabe der verdeckten Schicht und verwendet eine Fusionsoperation, um während des Trainings ANP-Einbettungen zu erhalten. Einbettungsmodelle, die mit unserem Ansatz trainiert werden, können bei allgemeinen Aufgaben zur Erkennung von Phrasenähnlichkeit effektiv besser abschneiden. Darüber hinaus haben wir sechs Fusionsmethoden für ANP-Einbettungen analysiert (Vektoraddition, Vektormultiplikation, Verkettung, elementweises Maximum, Tensorprodukt und Dilatation). Die optimale Leistung einer Fusionsmethode hängt von der Art der Anwendung ab, da eine einfache Fusionsmethode wie die Konkatenation am besten für die Phrasenähnlichkeit geeignet ist, während komplexe Fusionsoperationen wie das Tensorprodukt gut für das Verständnis der lokalen Beziehungen innerhalb der ANP geeignet sind.

## Ändern eines Minimalen Parametersatzes

Entscheidungsgrenzen werden nach dem Training durch Modellparameter definiert. Jüngste Ergebnisse zeigen, dass schon die alleinige Anpassung der Batch-Normalisierungsschichten bei gleichzeitiger Fixierung der anderen Modellparameter zur Verbesserung der Ergebnisse insb. zur Wiederverwendung für andere Aufgaben von Vorteil ist. In dieser Arbeit haben wir analysiert, wie sich die Entscheidungsgrenzen durch die Änderung einer kleinen Anzahl von Parametern verändern. Es wurden zwei Szenarien betrachtet: dichtes Tracking (offline und online) und Bildklassifizierung.

Jüngste Studien zeigen, dass die Domänenverschiebung eines Modells zu einer Vermischung zwischen den Clustern verschiedener Klassenmerkmale führt, wodurch sich die Klassifizierungsleistung verschlechtert. Die sog. Testzeitanpassung (Test Time Adaptation) kann dieses Problem lindern, indem der Merkmalsextraktor an die Zieldomäne angepasst wird, um die einzelnen Klassencluster wiederherzustellen, sodass der Klassifikator eine geeignete Entscheidungsgrenze zwischen verschiedenen Klassen finden kann. Wir haben zwei selbstlernende Ansätze, VideoWalk [JOE20] und MAST [LLX20], für dichtes Tracking in Videos in Kombination mit Prediction-Time Batch-Normalisierung [Nad+20; Sch+20] für Testzeitanpassung analysiert. Hierbei werden die zuvor gesammelten Statistiken der Batch-Normalisierungsschichten zur Testzeit aktualisiert, ohne die anderen Parameter zu verändern. VideoWalk und MAST profitieren von der Anwendung von Prediction-Time Batch-Normalisierung bei dichtem Tracking sowohl offline als auch online.

Lu *et al*. [Lu+21] schlagen vor, dass vortrainierte Transformer auf Text für verschiedene Aufgaben wie Bit-Operationen, Bildklassifikation und Remote Homology Detection wiederverwendet werden können. Wir untersuchten, wie GPT-2, das auf einem großen Textdatensatz trainiert wurde, für die Bildklassifikation umfunktioniert und verbessert werden kann. Eine weitere Motivation ein vortrainiertes GPT-2 wiederzuverwenden sind die sonst benötigten Trainingsdaten von ca. 8 Millionen Webseiten und Kosten von 256 $ pro Stunde für Cloud-Computing. Das neue Modell fügt eine lineare Schicht zwischen der Eingabe und GPT-2 ein. Das Hinzufügen der linearen Schicht ist eine minimale Änderung im Vergleich zu den ca. 1,5 Milliarden Parametern des Modells. Wir haben festgestellt, dass der Einbettungsraum zwischen den Trainingsbeispielen und GPT-2 die Genauigkeit des Modells beeinflusst. Das ursprüngliche Modell erreicht 72 % Genauigkeit mit einer linearen Projektion, während wir herausgefunden haben, dass das Modell 83 % Genauigkeit erreicht, wenn es mit einem LSTM kombiniert wird. Diese Ergebnisse wurden erzielt, indem nur die Eingabeschicht und die Normalisierungsschichten aktualisiert wurden. Wir haben außerdem festgestellt, dass alle Blöcke in GPT-2 einen positiven Einfluss auf die Bilderkennungsaufgaben haben.

## Klassifizierung bei Unbekannten Klassen

Wir haben bewiesen, dass VQA-Modelle robust gegenüber einer Überrepräsentation von polaren Fragen (d. h. Fragen, deren Antwort entweder „ja“ oder „nein“ lautet) in Trainingsdatensätzen sind. Wir haben unsere Arbeit erweitert, um die Entscheidungsgrenzen zwischen polaren

und nicht-polaren Fragen zu erklären. Zu diesem Zweck wurde der Versuchsaufbau so definiert, dass nicht-polare Fragen mithilfe eines Modells ausgewertet werden, das zuvor mit polaren Fragen trainiert wurde. Das Ergebnis ist eine starke Korrelation zwischen der Genauigkeit der nicht-polaren Antworten und der Anzahl der polaren Fragen zu diesen nicht-polaren Antworten, die für das Training des Modells verwendet wurden. Daraus schließen wir, dass sich die Entscheidungsgrenzen des durch polare Fragen geschaffenen Merkmalsraums mit dem Merkmalsraum der nicht-polaren Fragen überschneiden. Die Beziehung zwischen polaren Fragen und nicht-polaren Klassen führt dazu, dass das Modell zu einem einheitlichen Merkmalsraum konvergiert, der sowohl für die Beantwortung polarer als auch nicht-polarer Fragen verwendet werden kann. Eine interessante Möglichkeit, die sich aus unseren Ergebnissen ergibt, ist, dass VQA-Modelle beliebig viele Klassen lernen können, ohne dass teure, feingranulare Annotationen erforderlich sind, da bereits mit polaren Fragen zu diesen Klassen sehr gute Ergebnisse erzielt werden.

## AP4: Erstellen von Besser Erklärbaren Modellen

In Arbeitspaket 1, 2 und 3 haben wir bereits verschiedene Limitierungen von Modellen des maschinellen Lernens erörtert, wie etwa ihre Undurchsichtigkeit und mangelnde Robustheit. In diesem Arbeitspaket konzentrierten wir uns nun darauf, neue interpretierbare Architekturen zu entwickeln, die diese Probleme angehen. Diese Architekturen wurden von den Limitierungen der Standardmodelle inspiriert, die zuvor identifiziert wurden. Wir diskutieren die Hauptmerkmale dieser Modelle und wie sie zur Interpretierbarkeit und Robustheit beitragen, indem sie modifizierte Zielfunktionen verwenden, sich auf reichhaltigere Beschriftungsstrukturen stützen, die ursprüngliche Aufgabe aufteilen oder interpretierbare Operationen verwenden.

### XAI-Handbook als Blaupause für die Erstellung interpretierbarer Modelle

Wie zu Beginn in AP1 erwähnt, definiert das „XAI Handbook“ ein Vokabular, das alle Beiträge auf dem Gebiet der XAI umfasst. Ein Teil dieses Vokabulars beinhaltet den Begriff der „nicht-funktionalen Anforderungen“ für eine bestimmte Aufgabe. In diesem Zusammenhang ist eine nicht-funktionale Anforderung eine Einschränkung oder Eigenschaft der Aufgabe, auf die sich das Modell stützen sollte, um Vorhersagen zu treffen. Sobald eine solche Anforderung identifiziert wurde, ist es die Aufgabe des ML-Experten, eine „Erklärung“ zu entwerfen, d. h. eine Abbildung bestehender formaler Elemente der ML-Pipeline in ein Muster, das einen Beleg für das Vorhandensein oder Nichtvorhandensein einer nicht-funktionalen Anforderung liefert. Der Ausgabe der Erklärung werden dann explizite Regeln zugewiesen, wie sie zu lesen ist, oder mit anderen Worten, wie sie zu „interpretieren“ ist. Dank des XAI Handbooks waren wir in der Lage, einen allgemeinen Prozess zu identifizieren, mit dem wir Modelle mit einem von vornherein interpretierbaren Design konstruieren können. Zunächst ermitteln wir eine nicht-funktionale Anforderung an die Aufgabe, die das ML-Modell erfüllen soll, z. B. Bildklassifizierung, Segmentierung von Videoobjekten, Beantwortung visueller Fragen, usw. Diese Anforderung muss dann

formal als Einschränkung dargestellt werden, die in die Konzeption des ML-Modells einbezogen wird. Durch die Verwendung der begleitenden Interpretation zur Ausgabe der Erklärung ist ein Modell dank der zusätzlichen Struktur, die mehr Informationen über den Vorhersageprozess des gesamten Modells vermittelt, besser interpretierbar. In den nächsten Abschnitten beschreiben wir die wichtigsten Beiträge zur Konstruktion solcher interpretierbarer Modelle.

## Self-Supervised Auxiliary Learning (SSAL)

Traditionelle ML-Klassifikatoren können nicht auf unterstützende Signale zurückgreifen, die ihre Leistung und Interpretierbarkeit verbessern könnten. Wir führen das selbstüberwachte Hilfslernen (Self-Supervised Auxiliary Learning, SSAL) als einen Ansatz ein, um genau dies zu erreichen. SSAL-Modelle verwenden einen oder mehrere zusätzliche Klassifizierungsköpfe, die von den Zielen der ursprünglichen überwachten Klassifizierungsaufgabe abgeleitet sind, und folgen dabei den Prinzipien des Multi-Task-Lernens. Insbesondere werden die Beschriftungen der ursprünglichen Klassifizierungsaufgabe auf der Grundlage der visuellen Ähnlichkeit ihrer Beispiele gruppiert. Diese Gruppen sind als **zusätzliche Beschriftungen, welche die Ko-Domäne einer surjektiven (hierarchischen Gruppierungs-) Beziehung darstellen**, mit den ursprünglichen Beschriftungen strukturiert. SSAL-Zweige legen dem Optimierungsprozess auf niedriger Ebene Voraussetzungen auf, und ihre Verwendung während der Inferenz ermöglicht eine schnellere Konvergenz der Modelle bei gleichzeitiger Konzentration auf einen reichhaltigeren Satz klassenrelevanter Merkmale. Das Training ist mit denselben Methoden wie bei traditionellen CNN-Architekturen möglich, aber es stützt sich auf eine strukturierte Fehlerfunktion, die sowohl Terme für die ursprüngliche Aufgabe als auch für die Hilfsaufgabe enthält. SSAL-Modelle übertreffen durchweg den Stand von Wissenschaft und Technik und liefern strukturierte Vorhersagen, die den Nutzern helfen können, die zugrunde liegenden Merkmale, die das Modell für die Vorhersage verwendet, einzugrenzen. Vor allem aber liefern sie eine zusätzliche Struktur (d. h. die Ausgaben der Gruppenzweige), für die Heatmaps, z. B. die Klassenaktivierungskarte (Class-Activation Map, CAM), eine nützliche Interpretation bieten: Stark aktivierte Bereiche entsprechen Merkmalen, die allen Beschriftungen der Gruppenklasse gemeinsam sind. Wie in Abbildung 2 gezeigt, umfassen hoch aktivierte Bereiche für die vorhergesagte Klasse LINEAL die Art von gleichmäßigen Markierungen und kleinen Zahlen, die in Beispielen der vorhergesagten Hilfsklasse gefunden werden können: {BABYFLASCHE, PILLENFLASCHE, FERNBEDIENUNG, RECHENSCHIEBER, LINEAL}.

## Hybride Video-Objekt-Segmentierung mit Referenzabgleich

Die One-Shot-Segmentierung von Videoobjekten (Video Object Segmentation, VOS) ist eine Aufgabe, die häufig in Bereichen wie autonome Systeme und Robotik vorkommt, wo ein Objekt von Interesse pixelweise in einer Videosequenz segmentiert werden muss. Bei dieser Aufgabe wird eine anfängliche Maske, welche die Segmentierung eines Objektes definiert, bereitgestellt, und die Aufgabe des Modells besteht darin, dieses spezifische Objekt während des Rests der Sequenz zu verfolgen. Diese Aufgabe ist aufgrund der realen Anforderungen, wie z. B. uneingeschränkte Objektdarstellung, Kamerabewegung, Verdeckung, schnelle Bewegung

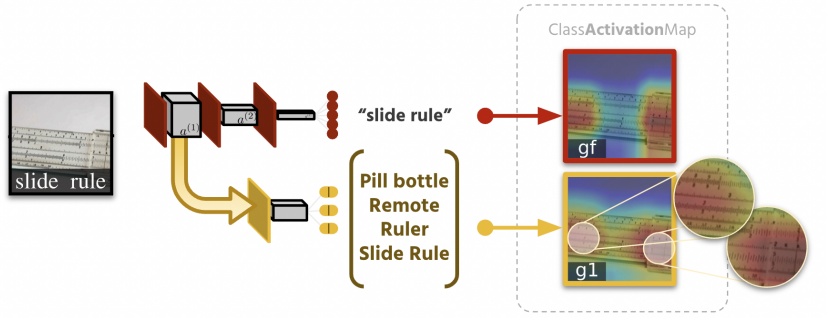



**Abb. 2** SSAL-Modelle sind dank der Hilfsaufgabe, die visuell ähnliche Klassen aus der ursprünglichen Aufgabe gruppiert, besser interpretierbar. Die Interpretation der Vorhersage ist mit Heatmaps möglich, da die Bereiche mit hoher Aktivierung den Merkmalen entsprechen, die allen Beschriftungen in der Gruppenklasse gemein sind.

und Bewegungsunschärfe, eine Herausforderung. In letzter Zeit haben sich Methoden, die auf rekurrenten neuronalen Netzen basieren, als erfolgreich erwiesen, da sie in der Lage sind, sich das Zielobjekt zu merken und räumlich-zeitliche Merkmale zu berechnen, die für sequenzielle Daten nützlich sind. Sie weisen jedoch Einschränkungen auf, wie z. B. eine geringere Segmentierungsgenauigkeit bei längeren Sequenzen und verdeckten Objekten aufgrund ihres begrenzten Speichers und der Fehlerausbreitung, die den meisten rekurrenten Modellen inhärent ist. Es ist offensichtlich, dass einige der vorherrschenden Einschränkungen durch die Speicherdarstellung der rekurrenten Modelle bedingt sind. Anstatt sich also vollständig auf die interne, undurchsichtige Speicherdarstellung dieser Modelle zu verlassen, können wir das Problem auf eine für den Menschen besser interpretierbare Weise darstellen. Insbesondere nutzen wir die anfängliche Segmentierung, die für das erste Bild verfügbar ist, weiter aus und erlauben dem Modell, explizit darauf zurückzugreifen, bevor es die Segmentierungsmaske für jedes neue Bild im Video vorhersagt. Dies ähnelt der Strategie, die ein Mensch bei der Lösung der Aufgabe verfolgt, z. B. wenn er sich ein Referenzbild eines Objekts ansieht, bevor er das gleiche Objekt in einem anderen Kontext identifiziert.

Darüber hinaus haben wir auch einen Schwachpunkt der Fehlerfunktion festgestellt, die zur Vorhersage von Segmentierungsmasken verwendet wird. Der Standardfehlerfunktion für das Training eines VOS-Modells beruht auf einer binären Maske, bei der jedes Pixel entweder als Teil des Vordergrunds oder des Hintergrunds gekennzeichnet wird. Auf diese Weise erhält das Modell keine zusätzlichen Informationen über die Beschaffenheit der (meisten) festen Objekte, z. B. dass konvexe Objekte keine Löcher innerhalb ihrer Umrandung haben. Aus diesem Grund haben wir das Ziel mit einem entfernungsbasierten räumlichen Fehler ergänzt, wie er von Bischke *et al.* [Bis+19] im Bereich der Erdbeobachtung vorgeschlagen wurde. Bei dieser Fehlerfunktion werden Pixeln eine von mehreren Unterklassen entsprechend ihres Abstandes zum Objektrand zugewiesen. Diese Fehlerfunktion ist nicht nur empirisch vorteilhaft, sondern auch besser interpretierbar, da er eine allmähliche Abnahme der Bedeutung von Pixeln darstellt, je weiter sie vom Objekt entfernt sind, was eine natürliche Eigenschaft von Objekten in der realen Welt ist, die jedoch nie explizit modelliert wird (siehe Abbildung 3). Durch das Erlernen der relativen Position der Pixel wird das Modell außerdem implizit in die Lage versetzt,

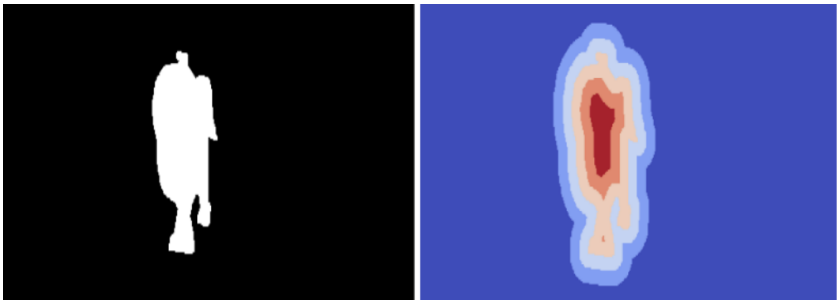


**Abb. 3** Anstelle einer binären Maske für den Fehler verwenden wir eine strukturierte, abstandsbasierte Maske, die zusätzliche Informationen über die Eigenschaften der Objekte vermittelt.

Pixeln in der Nähe der Objektränder eine höhere Gewichtung zuzuweisen, da sie schwieriger zu segmentieren sind. Darüber hinaus haben die Visualisierungen der Pixelmasken eine direkte, für Laien verständliche Interpretation: „Ausgangspixel für einen bestimmten Kanal entsprechen der geschätzten Position der verfolgten Instanz." Wie in AP2 berichtet, statteten wir nicht nur ein VOS-Modell mit interpretierbaren Primitiven aus, sondern verbesserten auch die Segmentierungsergebnisse, insbesondere für kleinere Objekte, längere Videos und verdeckte Szenen.

## Reichhaltigere textuelle Verankerung für die Adversarial Text-zu-Bild-Erzeugung

Bei Text-zu-Bild-Modellen (Text to Image, T2I) entsteht eine interpretierbare Schnittstelle zum Verständnis der Ausgabe des Modells, wenn umfangreichere Textbeschreibungen als Eingabe zulässig sind. Aus Sicht des Benutzers wird der Generierungsprozess viel kontrollierbarer, wenn er nicht nur ein Objekt, sondern auch dessen Eigenschaften spezifizieren kann. Herkömmliche T2I-Systeme verfügen jedoch lediglich über einfache Mechanismen zur Klassenkonditionierung, z. B. eine endliche Menge fester Einbettungen, die den Klassenbeschriftungen entsprechen, und fügen den Zwischendarstellungen Rauschen hinzu, um eine Vielfalt von Ergebnissen zu erzielen (z. B. jedes Mal unterschiedliche Bilder der Klasse KATZE). Aufgrund dieser Einschränkung bei der Darstellung komplexerer Textinformationen haben wir zwei neuartige Architekturen für T2I vorgeschlagen, die die Fähigkeiten dieser Modelle zur Verarbeitung umfangreicherer Texteingaben erweitern.

Wir begannen damit, dass wir sowohl für Objekte als auch für Eigenschaften dieser Objekte, auch bekannt als Adjektiv-Nomen-Paare (ANPs), getrennte Repräsentationen zulassen. Während Positions- und Objekteinbettungen nach etablierten Methoden erzeugt werden, schlugen wir den **Zusatz eines parallelen Moduls zur Einbettung der Adjektive** vor. Während des Adversarial Trainings minimiert der Diskriminator nun **drei Adversarial Hinge Fehlerfunktionen bei Bildmerkmalen, Objekt-Beschriftungsmerkmalen und Objekt-Attributmerkmalen**. Unsere umfangreiche Evaluierung zeigte, dass unser Modell nicht nur in der Lage ist, qualitativ hochwertige Ergebnisse zu produzieren, sondern auch, dass die Bilder durch die zusätzlichen Attribute weiter verankert werden, was durch die Abnahme der Diversität belegt wird (siehe Tabelle 1).

| Method | IS ↑ | SceneIS ↑ | FID ↓ | SceneFID ↓ | DS (↑) | CAS ↑ | Attr-F1 ↑ |
|---|---|---|---|---|---|---|---|
| Real Images | 23.50 ± 0.71 | 13.43 ± 0.33 | 11.93 | 2.46 | - | 46.22 | 15.77 |
| LostGANv1 [SW19] | 10.30 ± 0.19 | 9.07 ± 0.12 | 35.20 | 11.06 | **0.47 ± 0.09** | 31.04 | 11.25 |
| LostGANv2 [SW21] | 10.25 ± 0.20 | 9.15 ± 0.22 | 34.77 | 15.25 | 0.42 ± 0.09 | 30.97 | 11.38 |
| Ke Ma *et al.* [MZS20] | 9.57 ± 0.18 | 8.17 ± 0.13 | 43.26 | 16.16 | 0.30 ± 0.11 | **33.09** | 12.62 |
| *AttrLostGANv1* | 10.68 ± 0.43 | 9.24 ± 0.12 | 32.93 | 8.71 | 0.40 ± 0.11 | 32.11 | 13.64 |
| *AttrLostGANv2* | **10.81 ± 0.22** | **9.46 ± 0.13** | **31.57** | **7.78** | 0.28 ± 0.10 | 32.90 | **14.61** |

**Tab. 1** Ergebnisse für generierte Bilder, die durch visuelle Attribute eingeschränkt sind. Die von uns vorgeschlagene Architektur (kursiv) erzielt bei fast allen Wahrnehmungsmetriken die besten Ergebnisse. Bemerkenswert ist, dass die Metrik, die sich verschlechtert, *Diversität* (DS) ist, was zeigt, dass das Modell wie beabsichtigt durch die zusätzlichen Attribute eingeschränkt wird.

Wir haben außerdem eine Erweiterung dieses Verankerungsprinzips vorgeschlagen, bei der ein Adversarial T2I-Generator beliebige Beschreibungen für ein Objekt eingeben kann. Reichhaltigere Beschreibungen überbrücken die Kluft der Interpretierbarkeit zwischen dem Eingabecodierer und den generierten Bildern, da der Benutzer nicht durch eine feste Anzahl von Elementen oder Attributen beschränkt ist, sondern Objekte in beliebigem Detail beschreiben kann. Jüngste Fortschritte im Bereich der Texteinbettung erlauben es, Wörter und ihren Kontext originalgetreu darzustellen und dabei die Semantik langer Zeichenketten zu erhalten. Für dieses Modell haben wir auch eine neuartige Methode zur Erzeugung von Bildern aus semantisch reichhaltigen Regionsbeschreibungen sowie einer multimodalen Fehlerfunktion beim Abgleich von Regionsmerkmalen eingeführt, um einen zuverlässigen Bild-Text-Abgleich zu ermöglichen. Experimentelle Ergebnisse zeigen weitere Verbesserungen in verschiedenen Qualitätsmetriken wie FID, SceneIS und SceneFID.

## Feingranulare Erklärungen für Anomalien in Tabellendaten

Die Erkennung von Anomalien wird häufig durch das Training von Autoencodern auf einem Datensatz, der aus normalen (nicht anomalen) Beispielen besteht, vorgenommen. Ziel ist es, eine komprimierte Darstellung der Eingabedaten zu erlernen, die in der Lage ist, die normalen Beispiele mit hoher Genauigkeit zu rekonstruieren. Sobald der Autoencoder trainiert ist, kann er verwendet werden, um Anomalien in neuen, bisher ungesehenen Daten zu erkennen, indem der Rekonstruktionsfehler zwischen den Originaldaten und ihrer Rekonstruktion berechnet wird. Anomalien sind in der Regel durch höhere Rekonstruktionsfehler gekennzeichnet als normale Beispiele. Das Problem bei diesem Ansatz ist, dass die resultierende Anomalie als solche markiert wird, unabhängig von den zugrunde liegenden Gründen, die diese Beispiele anomal machen. Die Berechnung des Rekonstruktionsfehlers pro Merkmal hilft dabei, herauszufinden, welche Merkmale zur Anomalie beigetragen haben. Die Feststellung, welcher Teil des Beispiels oder welche Zelle einer tabellarischen Darstellung anomal ist, wird als zellenweise Anomalieerkennung bezeichnet. Die zellenweise Erkennung von Anomalien beantwortet nicht nur die Frage „Welche Beispiele sind anomal?“, sondern auch die Frage „Warum ist es eine Anomalie?“. Um einen solchen interpretierbaren Anomalie-Detektor zu entwickeln, trainierten

wir einen entrauschenden Autoencoder (Denoising Autoencoder, DAE), dessen Ausgabe pro Merkmal (Zelle) und nicht über das gesamte Beispiel ausgewertet wird. Darüber hinaus stellten wir zellenbasierte kontrafaktische Werte bereit, die anomale Beispiele in nicht-anomale transformieren würden. Wir führten eine Bewertung unseres vorgeschlagenen Ansatzes auf drei öffentlich zugänglichen Datensätzen zur Erkennung von Anomalien mit gemischten Attributen durch. Die empirischen Ergebnisse zeigen, dass unser Ansatz in seiner Leistung mit undurchsichtigen Verfahren zur Erkennung von Anomalien vergleichbar ist und gleichzeitig zusätzliche Informationen über die Ursache (d. h. das Eingangsmerkmal) der Anomalie und ein plausibles kontrafaktisches Ergebnis liefert.

### Räumliche Transformer-Netzwerke als interpretierbarer Aufmerksamkeitsmechanismus

Aus der Analyse in AP2 zur Klassifizierung unübersichtlicher Bilder haben wir ein einfaches Prinzip herausgearbeitet, mit dem Modelle durch ein interpretierbares Verfahren der vom Menschen inspirierten Aufmerksamkeit robuster gegen diese Unordnung werden. Im Wesentlichen wird ein Modell mit einem Mechanismus ausgestattet, der es ermöglicht, sich auf interessante Objekte zu konzentrieren und dabei das gesamte Umgebungsrauschen auszublenden. Dieses Prinzip ähnelt dem des foveatistischen Sehens, bei dem die wichtigsten Informationen aus der Mitte des Signals extrahiert werden. Unser Modell ist in der Lage, auffällige Objekte zu finden und das Bild so zu verändern, dass das Objekt zentriert und entsprechend vergrößert wird. Konkret setzen wir diesen Aufmerksamkeitsmechanismus mit einem räumlichen Transformationsnetzwerk (Spatial Transformer Network, STN) um, das eine affine Transformation des ursprünglichen Eingangsbildes erlernt. Durch Verstärkungslernen lernt das STN die notwendigen Transformationen, um das Bild so zu verändern, dass der nachfolgende Klassifikator die Wahrscheinlichkeit einer korrekten Vorhersage maximiert. Da die Transformation selbst in einem einzigen Schritt zu drastischen Veränderungen führen kann, haben wir auch einen Mechanismus zur allmählichen Anpassung mit einem rekurrenten Netzwerk entwickelt, der eine Abfolge von kleinen Transformationen vornimmt, bei denen nur eine konkret interpretierbare Operation pro Zeitschritt durchgeführt wird. Zu den Operationen gehören Translation entlang beider Raumachsen, Rotation und Skalierung. Das Ensemble aus STN- und Klassifizierungsmodulen folgt einer sequentiellen Pipeline, die eine einfache Überprüfung der modifizierten Eingaben ermöglicht, die in den Klassifizierer eingespeist werden. Darüber hinaus kann die Eingabe, die den Klassifikator erreicht, weiter in die Abfolge der Operationen pro Zeitschritt aufgeschlüsselt werden, die die endgültige Transformation ergeben haben. Diese Informationen sind leicht zu interpretieren und erklären mögliche Fälle, in denen der Klassifikator keine korrekte Vorhersage machen kann.

Neben dem Aufbau selbst haben wir auch eine orthogonale Eigenschaft dieser Transformationen untersucht. Wenn wir die transformierten Beispiele mit den ursprünglichen, unübersichtlicheren Versionen vergleichen, können wir die transformierte Menge als eine „einfachere“ Version der ursprünglichen Menge interpretieren. Vor diesem Hintergrund haben wir einen Trainingsplan für die Bildklassifikation anhand stetig steigender Schwierigkeit, so wie dies bei einem Lehrplan (Curriculum) der Fall ist, evaluiert. In diesem Fall besteht das Curriculum

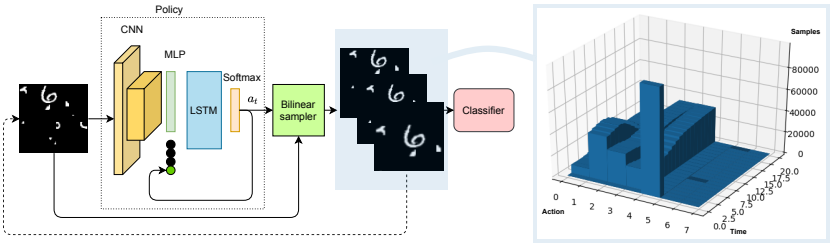



**Abb. 4** Links: Vorgeschlagenes sequentielles STN mit interpretierbaren Operationen, die zu Beispielen führen, die für den Klassifikator durch weniger Unordnung einfacher zu verarbeiten sind. Rechts: Die Zusammenfassung der einzelnen Schritte über den gesamten Datensatz gibt Aufschluss über die häufigste Operation, die für den Klassifikator nützlich ist.

zunächst aus Beispielen, die einfach zu klassifizieren sind. Diese Beispiele werden aus den ursprünglichen Bildern mit so vielen Zeitschritten wie nötig transformiert. Durch die schrittweise Reduktion der Transformationsschritte wird die Schwierigkeit erhöht, bis schließlich die ursprünglichen, unübersichtlichen Bilder verwendet werden. Die Verwendung eines auf inkrementeller Schwierigkeit basierenden Curriculums auf den unübersichtlichen Versionen von FashionMNIST und MNIST zeigt eine Leistungssteigerung von über 1 Prozentpunkt im Vergleich zu konventionellen Lernverfahren. Die Ergebnisse in diesem Bereich zeigen Vorteile in Bezug auf Transparenz und Interpretierbarkeit sowohl für den Vorhersageprozess (über das sequentielle STN) als auch für den Trainingsprozess (basierend auf dem Curriculum-Lernen mit inkrementeller Schwierigkeit).

## AP5: Erklärbarkeit mittels Visualisierung

Der komplexe Entscheidungsprozess von ML-Modellen ist eines der Haupthindernisse für das Verständnis ihrer Vorhersagen. Erklärungsmethoden können diese Komplexität teilweise bewältigen, indem sie sich auf bestimmte globale oder lokale Muster konzentrieren, die für den Entscheidungsprozess relevant sind. Wenn diese Muster jedoch zu komplex oder hochdimensional sind, können Menschen immer noch Schwierigkeiten haben, die Bedeutung der Ergebnisse zu verstehen. In solchen Fällen können wir auf Visualisierungsmethoden zurückgreifen, die komplexe Informationen auf eine intuitivere Weise vermitteln können.

Dieses Arbeitspaket umfasste alle Bemühungen während der gesamten Projektlaufzeit, visuelle Darstellungen der Phänomene zu erzeugen, die in ML-Modellen stattfinden. Wir betrachteten Visualisierungen wie Heatmaps, Grenzen des Eingaberaums und andere visuelle Darstellungen komplexer Strukturen, die aus den Interaktionen zwischen künstlichen Neuronen innerhalb einer ML-Architektur entstehen.

Wir haben uns auf Visualisierungen gestützt, um wichtige Muster des Verhaltens von ML-Modellen zu verstehen. Eine Analyse von Heatmaps für Bildklassifizierer (Abbildung 5) ergab, dass ihre Interpretation nicht dem traditionellen Verständnis von „den Eingangsmerkmalen, die den größten Einfluss auf die Vorhersage haben“ entspricht. Für unsere interpretierbaren

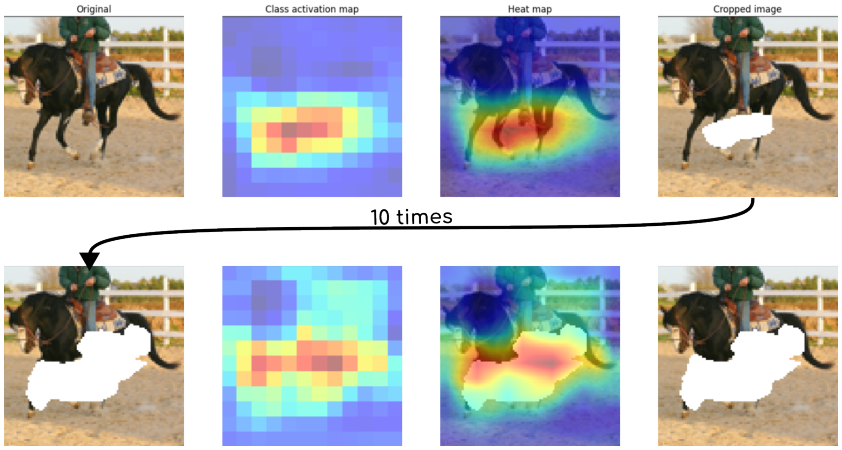



**Abb. 5** Auswertung von Feature-Attributionen. Unsere Methode zeigt, dass gängige Methoden zur Erstellung von Heatmaps nicht mit ihrer Ad-hoc-Interpretation übereinstimmen, die wichtigsten Eingangsmerkmale zu finden.

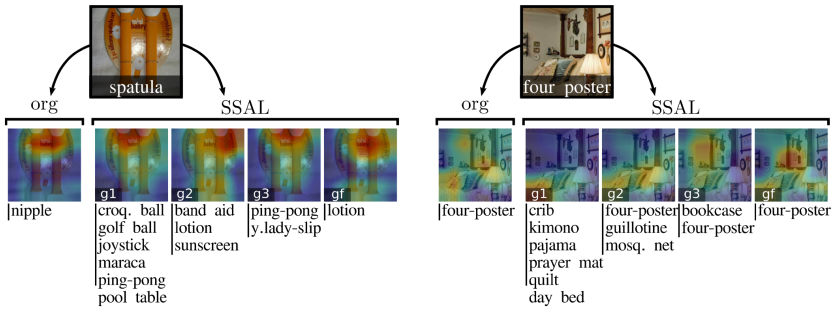



**Abb. 6** Hilfsklassifikatoren mit ähnlichen Gruppen können Bereiche mit interpretierbaren Merkmalen ausfindig machen, d. h. Bildbereiche mit Merkmalen, die den Beschriftungen in der Gruppenklasse gemeinsam sind.

Modelle, die auf Multi-Tasking-Gruppierungszielen basieren, können wir dank der Feature-Attributionen (Abbildung 6) Bereiche mit Merkmalen finden, die allen Beschriftungen in der Gruppenklasse gemeinsam sind. Um das Curriculum für einen Bildklassifikator zu bestimmen, können wir die Bilder wiedergeben, die zu Beginn der Trainingssequenz in einen Klassifikator eingehen, d. h. die einfacheren Beispiele (Abbildung 7). Schließlich können wir die Ergebnisse der Aufgaben zur Segmentierung von visuellen Objekten und zur Umwandlung von Text in Bilder dank ihrer natürlichen Interpretierbarkeit visuell verifizieren. VOS kann verifiziert werden, indem die Ausgabe der Segmentierungsmaske auf das entsprechende Bild gelegt wird (Abbildung 8), während T2I eine textuelle Korrespondenz zu den Umrissen (Bounding Boxes, Layout Cues) und der endgültigen Ausgabe (Abbildung 9) bietet.

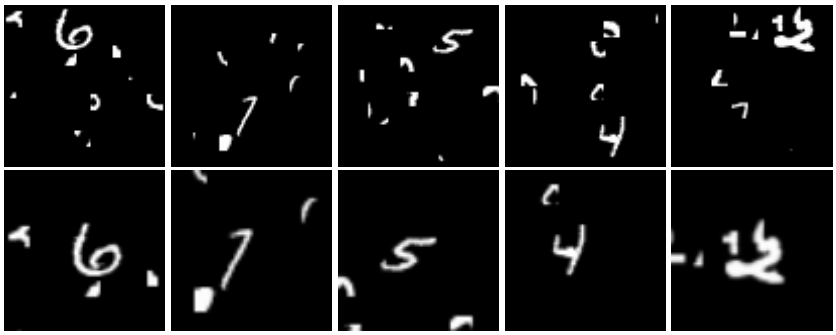


**Abb. 7** Datenbeispiele aus dem unübersichtlichen MNIST-Datensatz in der oberen Reihe und die mit SSTN transformierten Daten in der zweiten Reihe.

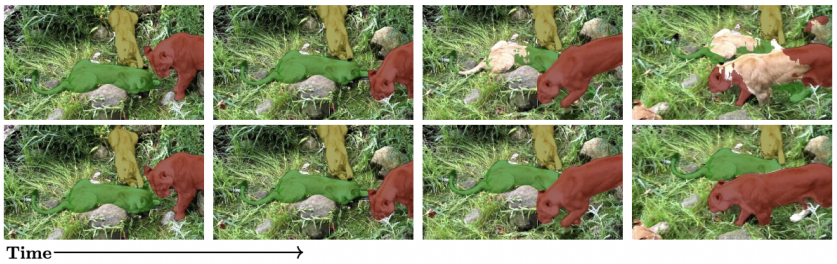



**Abb. 8** Visuelle Objektsegmentierung mit Korrespondenzabgleich hat eine natürliche Visualisierung, wenn die Segmentierungsausgabe mit dem entsprechenden Bild überlagert wird.

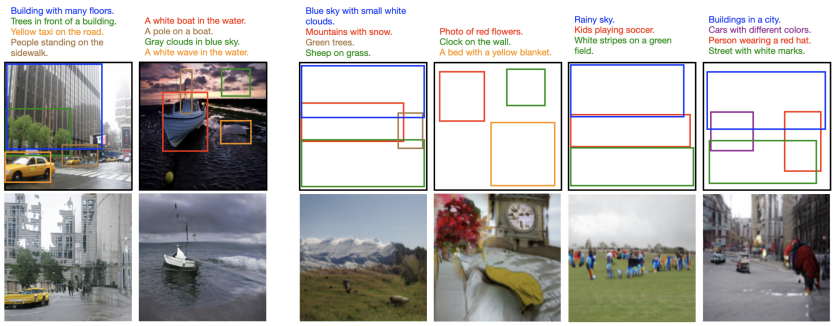



**Abb. 9** Attribute für T2I können iterativ angereichert werden, was eine schrittweise Verfolgung des Ausgangsbildes ermöglicht.

# Wichtigste Positionen des zahlenmäßigen Nachweises

Die wichtigsten Positionen des zahlenmäßigen Nachweises werden dem Bericht als Anhang „Zahlenmäßiger Nachweis“ beigefügt.

# Notwendigkeit der geleisteten Projektarbeiten

Die Notwendigkeit und Angemessenheit der öffentlichen Förderung für die in ExplAINN geleisteten Arbeiten begründet sich einerseits in der geleisteten, anwendungsorientierten Grundlagenforschung und andererseits in der hohen Allgemeinrelevanz der bearbeiteten Fragestellungen. Die Arbeiten folgten dem im Projektantrag festgelegten Arbeitsplan und waren zur Erreichung der Ziele notwendig, in ihrem Umfang angemessen und konnten mit den beantragten Ressourcen durchgeführt werden.

Die hohe Allgemeinrelevanz der geleisteten Arbeiten ergibt sich aus der Undurchsichtigkeit der Funktionsweise von Deep Learning Technologien, welche zunehmenden in sog. Risikoszenarien wie Verkehrskontrolle, Arzneimittelentwicklung, sowie medizinischen, rechtlichen und finanziellen Entscheidungen eingesetzt werden. Hier besteht von allen Seiten ein Bedürfnis für erhöhte Interpretierbarkeit. Hersteller wünschen sich Rechtssicherheit beim Vertrieb von Produkten, die Deep Learning Komponenten enthalten. Bürger wünschen Begründungen für von KI getroffenen Entscheidungen, die sie betreffen. Die Legislative benötigt Einsichten in die Funktionsweise von KI-Modellen, um einen verträglichen Rechtsrahmen festzulegen.

Insofern ist die hohe Aktivität um das Forschungsfeld der erklärbaren KI verständlich. ExplAINN leistete hier wichtige Beiträge, die den Stand von Wissenschaft und Forschung erheblich erweiterten. Neben diversen wissenschaftlichen Publikationen, welche die Interpretation bestehender Modelle zum Ziel haben, wurde das „XAI Handbook“ veröffentlicht. Es definiert Verfahrensweisen und Werkzeuge, mit deren Hilfe besonders interpretierbare Modelle erstellt werden können. Unsere Arbeiten rund um Interpretierbarkeit mündeten dabei in schließlich in der Einladung in das Konsortium, welches die DIN SPEC 92001-3 („Künstliche Intelligenz - Life Cycle Prozesse und Qualitätsanforderungen - Teil 3: Erklärbarkeit“) verfasst. Hiermit werden die Grundlagen von Industrie- und Verwaltungsrahmen für die Rechenschaftspflicht von KI-Systemen geschaffen.

Während diese grundlagenorientierten Forschungs- und Entwicklungsarbeiten erhebliche Aufwände verlangten, bestand – wie immer, wenn wissenschaftliches Neuland beschritten wird – ein durchaus relevantes Risiko, dass die vorgesehenen Lösungsansätze unerwartete Ergebnisse produzieren. Die Durchführung solcher Forschungsarbeiten mit wissenschaftlicher Bedeutung und hoher Anwendungsrelevanz, aber ohne unmittelbaren wirtschaftlichen Ertrag steht in Einklang mit der Charakteristik des DFKI als gemeinnütziges Forschungsunternehmen.

Die erzielten Ergebnisse wurden gemäß dem Verwertungsplan durch die Publikation auf angesehenen wissenschaftlichen Konferenzen, der Teilnahme an Workshops, sowie eingeladenen Vorträgen, diskriminierungsfrei einem breiten Publikum zur Verfügung gestellt und versprechen durch Verbesserungen bei der Interpretierbarkeit von Deep Learning Modellen weiterhin eine signifikante Bereicherung der interessierten deutschen Wirtschaft.

Aufgrund des Zusammenwirkens von Faktoren wie dem nationalen Interesse, den erzielten Ergebnissen, dem Potenzial für langfristige wirtschaftliche Verwertung, dem Forschungsrisiko und der Bedeutung einer öffentlich sichtbaren Behandlung des Themas, war die öffentliche Förderung von ExplAINN angemessen und notwendig.

# Nutzen & Verwertbarkeit des Ergebnisses

Die wichtigsten Beiträge des ExplAINN Projektes bestehen in der Veröffentlichung von 26 wissenschaftlichen Publikationen, dem XAI Handbook, sowie der Teilnahme am Konsortium zur Verfassung der DIN SPEC 92001-3 („Künstliche Intelligenz - Life Cycle Prozesse und Qualitätsanforderungen - Teil 3: Erklärbarkeit“). Somit haben viele der in ExplAINN durchgeführten Arbeiten bereits großes Interesse auch extern bei Forschern und Industriepartnern geweckt. Wir sehen eine Vielzahl von Forschungsprojekten und Anträgen (z. B. XAINES, SensAI, SimLearn, SustainML, ecoKI, AI4EO, CurAIVasc), sowie Kooperationen mit Industriekunden (z. B. enviaM, Sartorius), die von den Ergebnissen profitieren.

## Wirtschaftliche Erfolgsaussichten

Als grundlagenorientiertes Forschungs- und Entwicklungsprojekt zielte ExplAINN auf innovative Projektergebnisse, die nur mittelbar einer direkten wirtschaftlichen Verwertung zugeführt werden können. Die entwickelten Verfahrensweisen und Werkzeuge können jedoch einen wichtigen Beitrag zur Erstellung interpretierbarer KI-Modelle leisten und somit ihren Einsatz in den zuvor erwähnten Risikoszenarien (Verkehrskontrolle, Arzneimittelentwicklung, medizinische, juristische, finanzielle Entscheidungen) ermöglichen. Somit kann aus den Ergebnissen trotz der Ausrichtung des Projekts in anwendungsorientierter Grundlagenforschung mittelbar ein wirtschaftlicher Vorteil generiert werden. Darüber hinaus ermöglichte die Qualität unserer Arbeit den Aufbau von Kooperationsbeziehungen mit technologisch führenden Unternehmen der Branche wie Adobe, AlgoLux, Amazon und Meta im Rahmen von Praktikantenprogrammen.

## Wissenschaftliche Erfolgsaussichten

Wie bereits geschildert erweiterten die im ExplAINN Projekt veröffentlichten wissenschaftlichen Publikationen den Stand von Wissenschaft und Technik im Bereich erklärbare KI erheblich. Insgesamt gab es 26 Veröffentlichungen auf international renommierten Konferenzen und Workshops wie NeurIPS, ICCV, ICPR, TPAMI, und IJCNN, sowie zwei eingeladene Vorträge, die das ebenfalls veröffentlichte XAI Handbook vorstellten[1]. Zudem schlossen innerhalb der

[1] 22.02.2022: ML2R Veranstaltung zu Erklärbarkeit von Künstlicher Intelligenz stärkt Zusammenarbeit der deutschen KI-Zentren (https://www.ml2r.de/en/event-on-explainability-of-artificial-intelligence-strengthens-cooperation-of-german-ai-centers/) ; 30.06.2022: AI Academy

Projektdauer 4 der Kernmitglieder ihre Dissertationen ab und der initiale Projektleiter folgte einem Ruf auf eine Professur. Schließlich reichten 3 Kernmitglieder ihre Doktorarbeit ein und betreuten 8 Abschlussarbeiten an der Technischen Universität Kaiserslautern.

## Wirtschaftliche und Wissenschaftliche Anschlussfähigkeit

Mit dem XAI Handbook und unserem Beitrag zur DIN SPEC 92001-3 wurden wichtige Grundsteine für zukünftige Entwicklungen auf dem Gebiet der erklärbaren KI gelegt. Das Feld kann damit allerdings keineswegs als gelöst betrachtet werden und die Forschungsaktivität ist insb. international betrachtet weiterhin sehr hoch. Wir sehen daher großes Potenzial, die veröffentlichten Verfahrensweisen und Werkzeuge in fortführenden Projekten kontinuierlich weiterzuentwickeln. Initiale Beispiele hierfür sind aktuell laufende Projekte, bzw. Anträge des DFKI wie XAINES, SustainML und SenpAI. Neben den unmittelbaren Ergebnissen, die aus ExplAINN resultierten, werden diese anschließenden Aktivitäten dazu beitragen, das wissenschaftliche Profil und den nationalen und internationalen Ruf des DFKI weiter zu verbessern und zu stärken. Gleichzeitig stellen sie eine wichtige Plattform für den Transfer von Know-how in die Industrie dar.

# Fortschritt auf dem Gebiet des Vorhabens

Als Forschungsbereich mit international besonderes hohem Interesse gab es während der Projektlaufzeit zahlreiche Veröffentlichungen zum Thema erklärbare KI, welche vom ExplAINN-Team kontinuierlich beobachtet wurden. Eine Auflistung würde den Rahmen dieses Berichts sprengen und wäre wenig zielführend. Allgemein lässt sich jedoch ein Trend weg von der Analyse der inneren Abläufe in Modellen, hin zu einem datengetriebenen Ansatz erkennen. Der Fokus liegt dabei darauf, wie Daten verarbeitet werden und vorhandene Bias zu identifizieren. Weiterhin stellen neuartige generative Modelle und sog. „Foundation Models“ wie GPT, Bert und Stable Diffusion eine Herausforderung in Sachen Interpretierbarkeit dar. Verfahren wie etwa „Zero-Shot Learning“, welche es z. B. ermöglichen auf Textdaten trainierte Modelle, ohne weiteres Training für die Bildanalyse zu verwenden, stellen unser bisheriges Verständnis von maschinellen Lernverfahren infrage. Weitere Forschung auf diesem Gebiet ist daher unabdingbar.

Darüber hinaus gibt es eine wachsende Zahl von Forschungsarbeiten, die sich auf theoretische Zusammenhänge zwischen Robustheit und Erklärbarkeit konzentrieren [INM19; Boo+20]. Wir heben die tiefgreifenden Auswirkungen der Arbeit von Jones *et al*. hervor, in der bewiesen wurde, dass Modelle, welche robust gegen Adversarial Attacks sind, zu ähnlichen Merkmalsrepräsentationen konvergieren, unabhängig von der Modellkapazität oder -architektur [Jon+22].

# Veröffentlichungen

Insgesamt wurden im Projekt 26 wissenschaftliche Veröffentlichungen publiziert, 2 eingeladene Vorträge gehalten und 3 Doktorarbeiten sowie 8 studentische Abschlussarbeiten betreut. Das Team wurde außerdem in das Konsortium zur Verfassung der DIN SPEC 92001-3 („Künstliche Intelligenz - Life Cycle Prozesse und Qualitätsanforderungen - Teil 3: Erklärbarkeit“) eingeladen.

Es folgt eine Auflistung der veröffentlichten wissenschaftlichen Arbeiten:

[Azi+19] Fatemeh Azimi u. a. „A reinforcement learning approach for sequential spatial transformer networks“. In: *Artificial Neural Networks and Machine Learning–ICANN 2019: Theoretical Neural Computation: 28th International Conference on Artificial Neural Networks, Munich, Germany, September 17–19, 2019, Proceedings, Part I 28*. Springer. 2019, S. 585–597.

[Fol+20] Joachim Folz u. a. „Adversarial defense based on structure-to-signal autoencoders“. In: *IEEE Winter Conference on Applications of Computer Vision*. IEEE. 2020, S. 3568–3577.

[Fro+20] Stanislav Frolov u. a. „Leveraging visual question answering to improve text-to-image synthesis“. In: *Proceedings of the Second Workshop on Beyond Vision and LANguage: inTEgrating Real-world kNowledge (LANTERN)*. 2020, S. 17–22.

[Mos+20] Brian B Moser u. a. „Dartsrenet: Exploring new rnn cells in renet architectures“. In: *Artificial Neural Networks and Machine Learning–ICANN 2020: 29th International Conference on Artificial Neural Networks, Bratislava, Slovakia, September 15–18, 2020, Proceedings, Part I 29*. Springer. 2020, S. 850–861.

[Azi+21a] Fatemeh Azimi u. a. „Hybrid-S2S: Video Object Segmentation with Recurrent Networks and Correspondence Matching“. In: *VISAPP*. arXiv:2010.05069. 2021, S. 182–192.

[Azi+21b] Fatemeh Azimi u. a. „Revisiting sequence-to-sequence video object segmentation with multi-task loss and skip-memory“. In: *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE. 2021, S. 5376–5383.

[Azi+21c] Fatemeh Azimi u. a. „Spatial transformer networks for curriculum learning“. In: *arXiv preprint arXiv:2108.09696* (2021).

[Fro+21] Stanislav Frolov u. a. „Adversarial text-to-image synthesis: A review“. In: *Neural Networks* 144 (2021), S. 187–209.

[Guz+21a] Andrey Guzhov u. a. „Esresne(x)t-fbsp: Learning robust time-frequency transformation of audio“. In: *2021 International Joint Conference on Neural Networks (IJCNN)*. IEEE. 2021, S. 1–8.

[Guz+21b] Andrey Guzhov u. a. „Esresnet: Environmental sound classification based on visual domain models“. In: *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE. 2021, S. 4933–4940.

[Her+21] Dayananda Herurkar u. a. „ANP-W2V: Effects of Composition Methods for Embedding Adjective-Noun Pairs“. In: *2021 International Joint Conference on Neural Networks (IJCNN)*. IEEE. 2021, S. 1–8.

[JPN21] Shailza Jolly, Sandro Pezzelle und Moin Nabi. „EaSe: A Diagnostic Tool for VQA Based on Answer Diversity“. In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2021, S. 2407–2414.

[Pal+21a] Sebastian Palacio u. a. „Contextual Classification Using Self-Supervised Auxiliary Models for Deep Neural Networks“. In: *International Conference on Pattern Recognition*. IEEE. 2021, S. 8937–8944.

[Pal+21b] Sebastian Palacio u. a. „IteROAR: Quantifying the Interpretation of Feature Importance Methods“. In: *NeurIPS WS: Preregister Science* (2021).

[Pal+21c] Sebastian Palacio u. a. „P≈NP, at least in Visual Question Answering“. In: *International Conference on Pattern Recognition*. IEEE. 2021, S. 2748–2754.

[Pal+21d] Sebastian Palacio u. a. „XAI Handbook: Towards a Unified Framework for Explainable AI“. In: *IEEE International Conference on Computer Vision* (2021).

[Azi+22] Fatemeh Azimi u. a. „Self-supervised test-time adaptation on video data“. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2022, S. 3439–3448.

[Fro+22a] Stanislav Frolov u. a. „Attrlostgan: Attribute controlled image synthesis from reconfigurable layout and style“. In: *Pattern Recognition: 43rd DAGM German Conference, DAGM GCPR 2021, Bonn, Germany, September 28–October 1, 2021, Proceedings*. Springer. 2022, S. 361–375.

[Fro+22b] Stanislav Frolov u. a. „Dt2i: Dense text-to-image generation from region descriptions“. In: *Artificial Neural Networks and Machine Learning–ICANN 2022: 31st International Conference on Artificial Neural Networks, Bristol, UK, September 6–9, 2022, Proceedings, Part II*. Springer. 2022, S. 395–406.

[Guz+22] Andrey Guzhov u. a. „Audioclip: Extending clip to image, text and audio“. In: *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2022, S. 976–980.

[Jol+22] Shailza Jolly u. a. „Search and learn: improving semantic coverage for data-to-text generation“. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Bd. 36. 10. 2022, S. 10858–10866.

[Mos+22a] Brian Moser u. a. „Hitchhiker’s Guide to Super-Resolution: Introduction and Recent Advances“. In: *arXiv preprint arXiv:2209.13131* (2022).

[Mos+22b] Brian Moser u. a. „Less is More: Proxy Datasets in NAS approaches“. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, S. 1953–1961.

[SHH22] Timur Sattarov, Dayananda Herurkar und Jörn Hees. „Explaining Anomalies using Denoising Autoencoders for Financial Tabular Data“. In: *arXiv preprint arXiv:2209.10658* (2022).

[Azi+23a] Fatemeh Azimi u. a. „Rethinking RNN-Based Video Object Segmentation“. In: *Computer Vision, Imaging and Computer Graphics Theory and Applications: 16th International Joint Conference, VISIGRAPP 2021, Virtual Event, February 8–10, 2021, Revised Selected Papers*. Springer International Publishing Cham. 2023, S. 348–365.

[Azi+23b] Fatemeh Azimi u. a. „Sequential Spatial Transformer Networks for Salient Object Classification“. In: *12th International Conference on Pattern Recognition Applications and Methods (ICPRAM)*. SCITEPRESS, 2023.

# Literatur

[Eve+10] M. Everingham u. a. „The Pascal Visual Object Classes (VOC) Challenge“. In: *International Journal of Computer Vision* 88.2 (Juni 2010), S. 303–338.

[Mik+13] Tomas Mikolov u. a. „Efficient estimation of word representations in vector space“. In: *arXiv preprint arXiv:1301.3781* (2013).

[MHG+14] Volodymyr Mnih, Nicolas Heess, Alex Graves u. a. „Recurrent models of visual attention“. In: *Advances in neural information processing systems* 27 (2014).

[PSM14] Jeffrey Pennington, Richard Socher und Christopher D Manning. „Glove: Global vectors for word representation“. In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014, S. 1532–1543.

[SJB14] Justin Salamon, Connor Jacoby und Juan Pablo Bello. „Unsupervised feature learning for urban sound classification“. In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2014), S. 5178–5182.

[Sri+14] Nitish Srivastava u. a. „Dropout: a simple way to prevent neural networks from overfitting“. In: *The journal of machine learning research* 15.1 (2014), S. 1929–1958.

[IS15] Sergey Ioffe und Christian Szegedy. „Batch normalization: Accelerating deep network training by reducing internal covariate shift“. In: *International Conference on Machine Learning*. PMLR. 2015, S. 448–456.

[JSZ+15] Max Jaderberg, Karen Simonyan, Andrew Zisserman u. a. „Spatial transformer networks“. In: *Advances in neural information processing systems* 28 (2015).

[Pic15] Karol J Piczak. „ESC: Dataset for environmental sound classification“. In: *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)* 2015 (2015), S. 2635–2639.

[He+16] Kaiming He u. a. „Deep residual learning for image recognition“. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, S. 770–778.

[KKK16] Been Kim, Rajiv Khanna und Oluwasanmi O Koyejo. „Examples are not enough, learn to criticize! criticism for interpretability“. In: *Advances in neural information processing systems* 29 (2016).

[Per+16] F. Perazzi u. a. „A Benchmark Dataset and Evaluation Methodology for Video Object Segmentation“. In: *Computer Vision and Pattern Recognition*. 2016.

[SK16] Tim Salimans und Durk P Kingma. „Weight normalization: A simple reparameterization to accelerate training of deep neural networks“. In: *Advances in neural information processing systems* 29 (2016).

[Zho+16] Bolei Zhou u. a. „Learning deep features for discriminative localization“. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, S. 2921–2929.

[DK17] Finale Doshi-Velez und Been Kim. „Towards a rigorous science of interpretable machine learning“. In: *arXiv preprint arXiv:1702.08608* (2017).

[Goy+17] Yash Goyal u. a. „Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering“. In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017.

[ZL17] Barret Zoph und Quoc Le. „Neural Architecture Search with Reinforcement Learning“. In: *International Conference on Learning Representations*. 2017.

[LSY18] Hanxiao Liu, Karen Simonyan und Yiming Yang. „DARTS: Differentiable Architecture Search“. In: *International Conference on Learning Representations*. 2018.

[Pha+18] Hieu Pham u. a. „Efficient neural architecture search via parameters sharing“. In: *International conference on machine learning*. PMLR. 2018, S. 4095–4104.

[Xu+18] Ning Xu u. a. „Youtube-vos: Sequence-to-sequence video object segmentation“. In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, S. 585–601.

[Bis+19] Benjamin Bischke u. a. „Multi-task learning for segmentation of building footprints with deep neural networks“. In: *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE. 2019, S. 1480–1484.

[Dev+19] Jacob Devlin u. a. „BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding“. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 2019, S. 4171–4186.

[DY19] Xuanyi Dong und Yi Yang. „Searching for a robust neural architecture in four gpu hours“. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, S. 1761–1770.

[Hoo+19] Sara Hooker u. a. „A benchmark for interpretability methods in deep neural networks“. In: *Advances in neural information processing systems* 32 (2019).

[Rud19] Cynthia Rudin. „Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead“. In: *Nature machine intelligence* 1.5 (2019), S. 206–215.

[SW19] Wei Sun und Tianfu Wu. „Image synthesis from reconfigurable layout and style“. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, S. 10531–10540.

[Bar+20] Alejandro Barredo Arrieta u. a. „Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI“. In: (2020).

[Gei+20] Robert Geirhos u. a. „Shortcut learning in deep neural networks“. In: *Nature Machine Intelligence* 2.11 (2020), S. 665–673.

[JOE20] Allan Jabri, Andrew Owens und Alexei A Efros. „Space-Time Correspondence as a Contrastive Random Walk“. In: *Advances in Neural Information Processing Systems* (2020).

[LLX20] Zihang Lai, Erika Lu und Weidi Xie. „MAST: A Memory-Augmented Self-Supervised Tracker“. In: *IEEE Conference on Computer Vision and Pattern Recognition* (2020).

[MZS20] Ke Ma, Bo Zhao und Leonid Sigal. „Attribute-guided image generation from layout“. In: *arXiv preprint arXiv:2008.11932* (2020).

[Nad+20] Zachary Nado u. a. „Evaluating prediction-time batch normalization for robustness under covariate shift“. In: 2020.

[Sch+20] Steffen Schneider u. a. „Removing covariate shift improves robustness against common corruptions“. In: *CoRR* abs/2006.16971 (2020).

[Zha+20] Jingzhao Zhang u. a. „Why Gradient Clipping Accelerates Training: A Theoretical Justification for Adaptivity“. In: *International Conference on Learning Representations*. 2020.

[Lu+21] Kevin Lu u. a. „Pretrained transformers as universal computation engines“. In: *arXiv preprint arXiv:2103.05247* 1 (2021).

[Par+21] Dong Huk Park u. a. „Benchmark for compositional text-to-image synthesis“. In: *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*. 2021.

[Sam+21] Wojciech Samek u. a. „Explaining deep neural networks and beyond: A review of methods and applications“. In: *Proceedings of the IEEE* 109.3 (2021), S. 247–278.

[SW21] Wei Sun und Tianfu Wu. „Learning layout and style reconfigurable gans for controllable image synthesis“. In: *IEEE transactions on pattern analysis and machine intelligence* 44.9 (2021), S. 5070–5087.

[Tol+21a] Ilya Tolstikhin u. a. „MLP-Mixer: An all-MLP Architecture for Vision“. In: *arXiv preprint arXiv:2105.01601* (2021).

[Tol+21b] Ilya O Tolstikhin u. a. „Mlp-mixer: An all-mlp architecture for vision“. In: *Advances in neural information processing systems* 34 (2021), S. 24261–24272.