Data Challenge

Objekt Lokalisation in Münzenbildern vom Corpus-Nummorum Projekt*

*https://www.corpus-nummorum.eu/

Patrick Raphael Melnic, Berna Sen, Ramazan Özdemir

Agenda

- Problemstellung
- Verwendete Modelle:
  - Transformer-Modelle
  - CLIP
  - OWL-ViT
- Herausforderungen & Lösungsansätze
  - 1.Ansatz
  - 2.Ansatz
- Neue Lösung
- Ergebnisse
- Fazit und Ausblick

Problemstellung

Objektdetektion und Lokalisation von Personen und Gegenständen in Münzenbildern

herm

patera

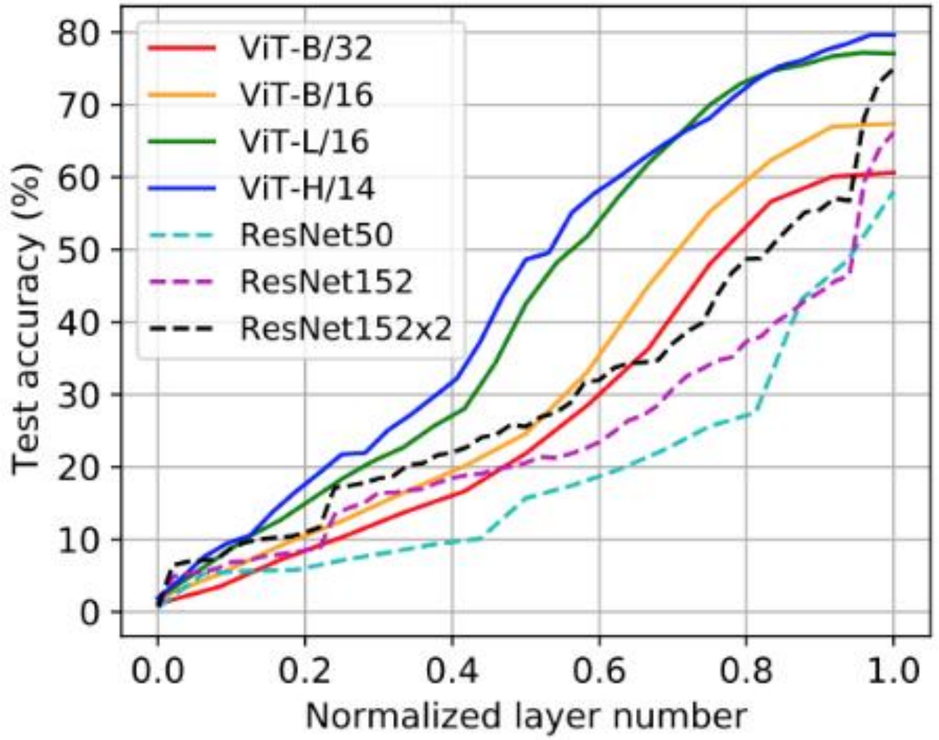
Problemstellung

Wie können wir Transformermodelle nutzen, um die Klassifizierung von Personen und Gegenständen zu verbessern?

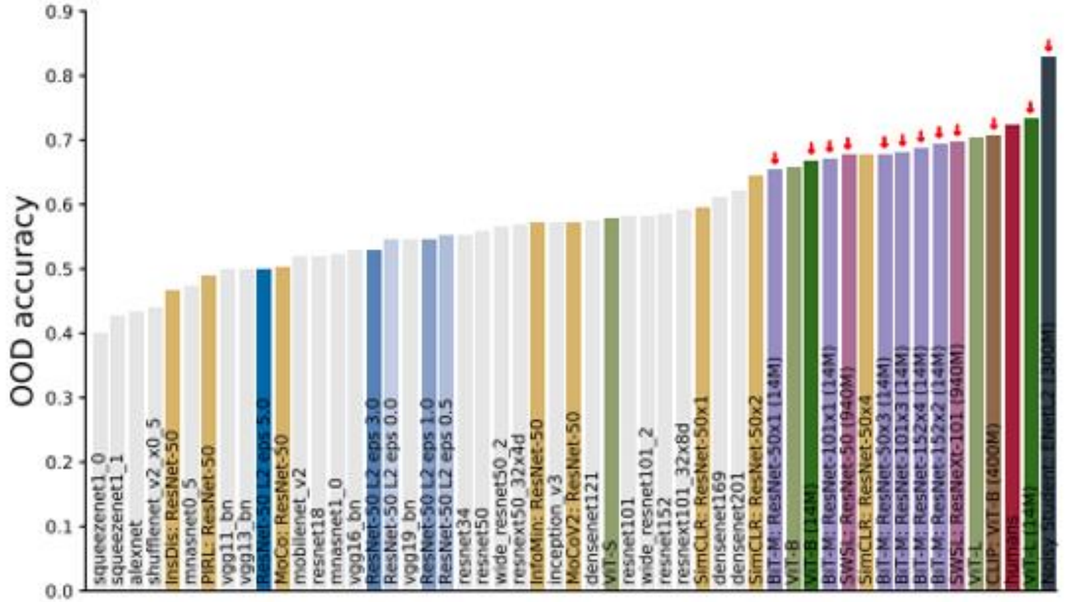
herm

patera

Ansatz: Modell mit Vision Transformer



(b) ViTs vs ResNets



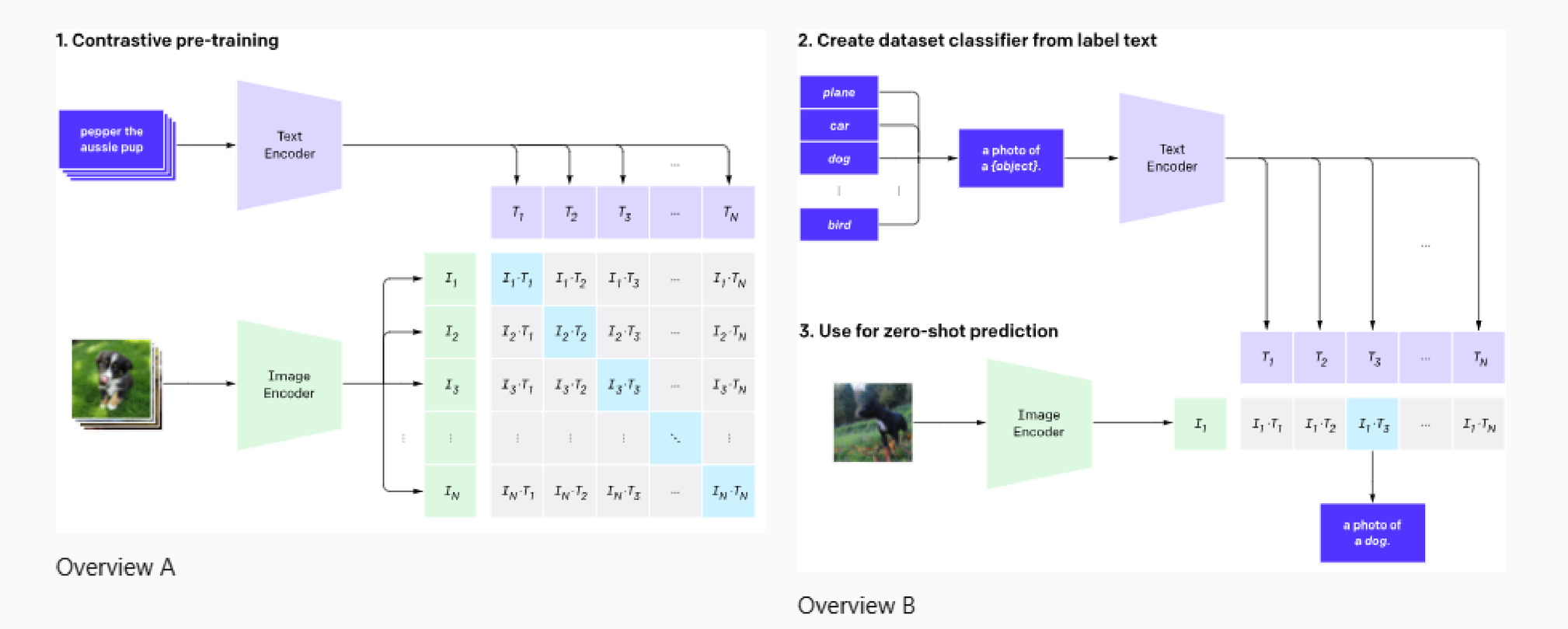
(a) OOD accuracy (higher = better).

Standard supervised models
Self-supervised models
Adversarially trained models (darker: more adv.)
Vision transformers (darker: bigger training set)

Noisy Student
Big Transfer Models (BiT-M)
Semi-weakly supervised learner models (SWSL)
CLIP (with vision transformer backbone)

CLIP

- **CLIP (Contrastive Language-Image Pre-training)**
  - Training mithilfe von Bild-Text-Paaren

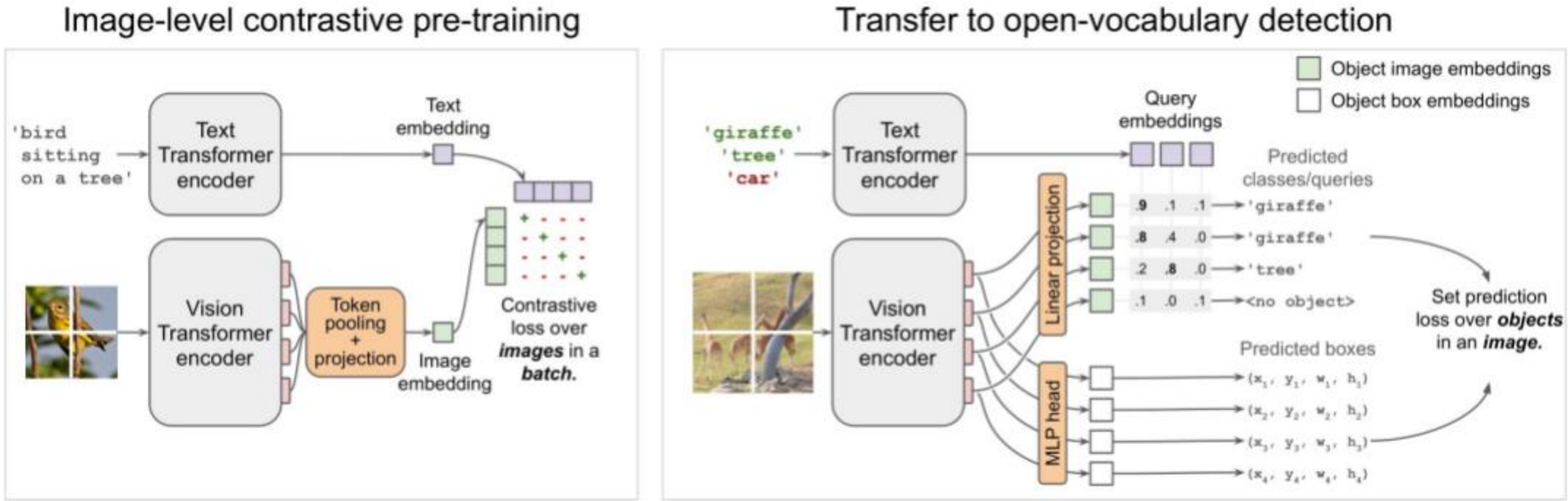


Overview A

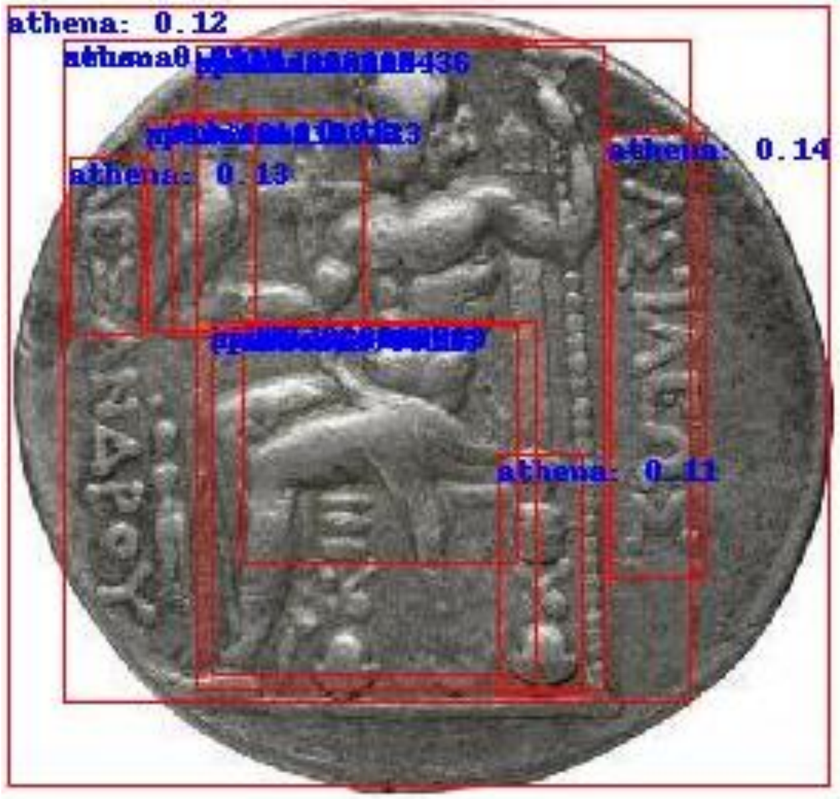
Overview B

OWL-ViT

- **OWL-ViT (Vision Transformer for Open-World Localization)**
  - Modell zur textbasierten Objekterkennung
  - Nutzt CLIP als multimodales Grundgerüst
  - Vision Transformer von CLIP für Labeling und Objektlokalisation

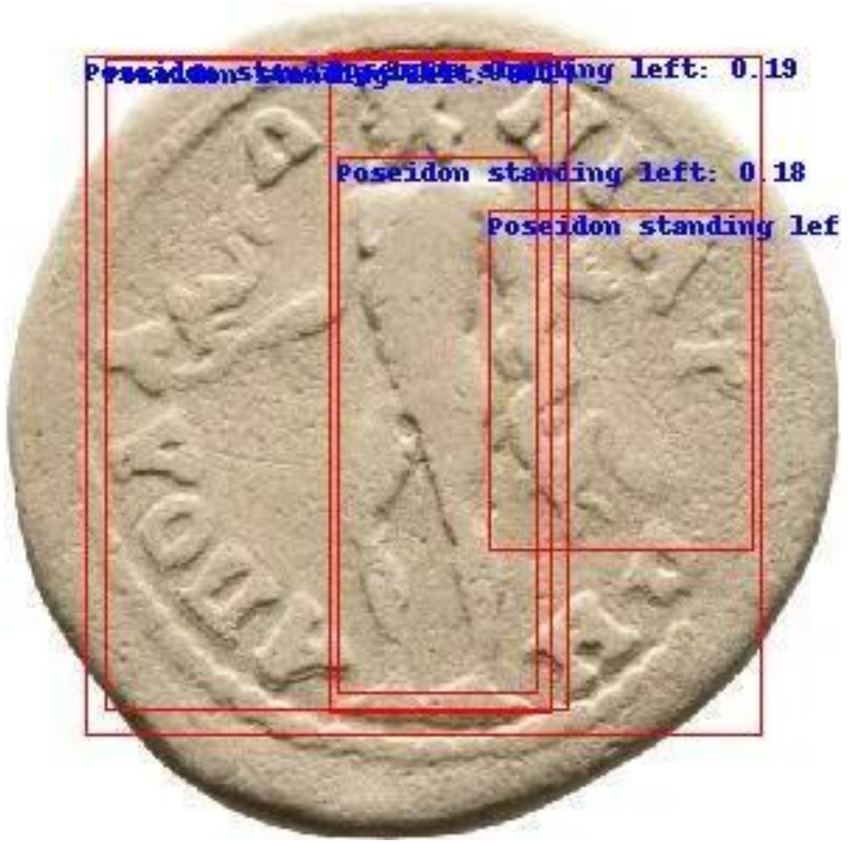


Testfälle mit OWL-ViT



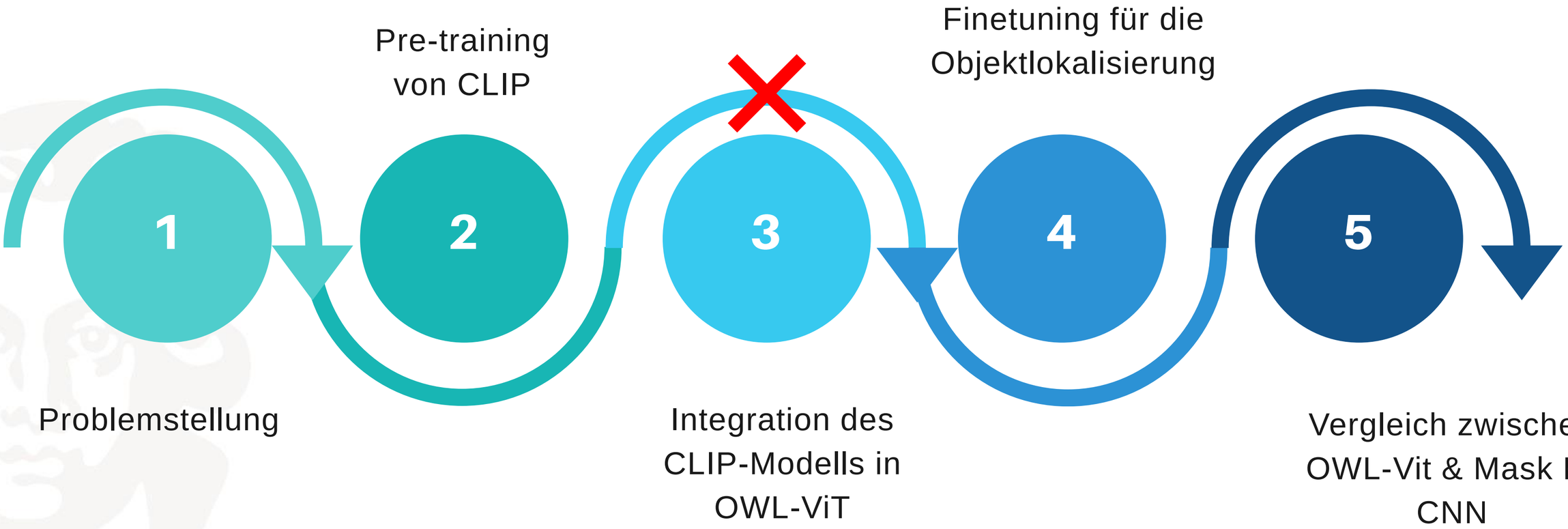
Zeus wurde erwartet, jedoch wurde Athena erkannt

Testfälle mit OWL-ViT



Beschreibung "Poseidon standing left„ wurde falsch interpretiert

Ursprüngliche Vorgehensweise



Herausforderungen & Lösungsansätze

- **CLIP Training**

1. **Ansatz:** CLIP Training mit Skript von Hugging Face

   - Das Skript funktionierte nicht wie erwartet
   - Technische Schwierigkeiten und spezifische Fehler

   ⇨ **2. Ansatz:** Training & Entwicklung eines eigenen Skripts als Alternative

Herausforderungen & Lösungsansätze

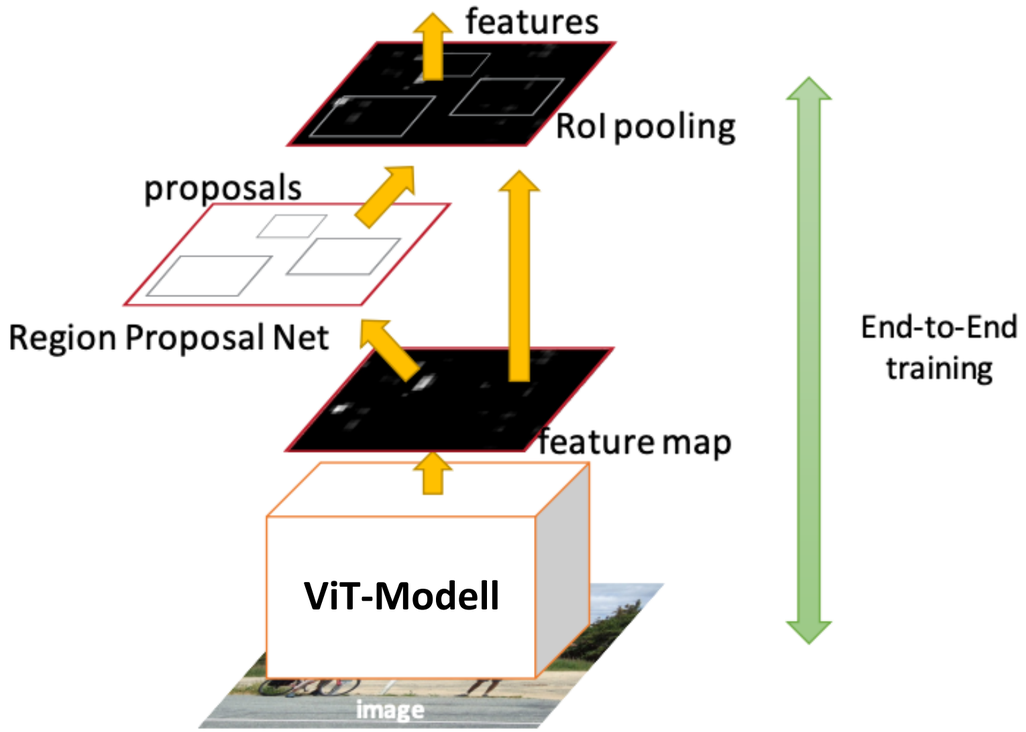
- **OWL-ViT Skript für das Training mit vortrainiertem CLIP-Modell**
  - Schwierigkeiten beim Zusammenführen der Modelle
  - Inkompatibilitäten und technische Limitierungen:
    - Modulinstallationen mit untereinander inkompatiblen Versionen, trotz isolierten virtuellen Environments
    - Dependancy Probleme mit unzureichender Dokumentation
  - Nicht funktionierendes Setup Skript

Neuer Lösungsansatz?

Neuer Lösungsansatz: Multilabel Klassifikation

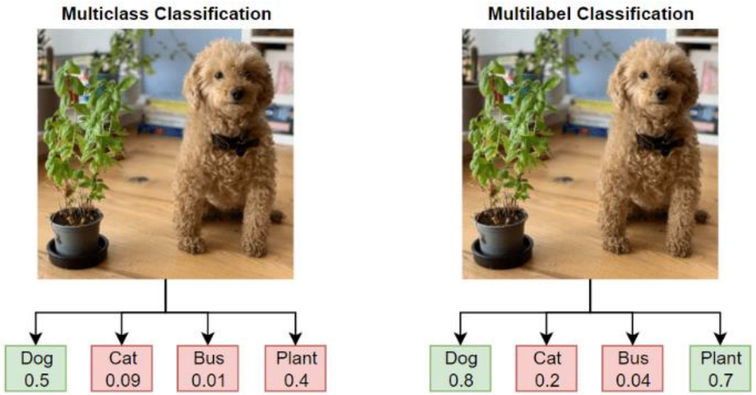
Warum Multilabel Klassifikation?

Ein ViT Modell soll für Multilabeling trainiert werden, um es als Backbone für die Object Detection zu verwenden.



Multilabel Klassifikation

- **Multilabel Klassifikation anhand eines Vision Transformers**
  - Finde alle Klassen, welche in einem Bild vertreten sind
  - Multilabeling hilft dabei, jedes Objekt und jede Person im Bild wiederzuerkennen aber auch welche Objekte und Personen eher auf einer Münze auftreten würden



Multilabel Klassifikation

Dataset:

```
---
Image path: CN_type_2720_BNF_41766282_cn_coin_14816_o_rev
Absolute Path: datasets/CN_dataset_obj_detection_04_23/dataset_obj_detection/hand/CN_type_2720_BNF_41766282_cn_coin_14816_o_rev.jpg
Labels: ['demeter', 'scepter', 'throne', 'corn', 'hand', 'poppy']
Encoded Labels: [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0
, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
---
Image path: CN_type_21272_BNF_41767414_cn_coin_22817_o_obv
Absolute Path: datasets/CN_dataset_obj_detection_04_23/dataset_obj_detection/bull/CN_type_21272_BNF_41767414_cn_coin_22817_o_obv.jpg
Labels: ['bull']
Encoded Labels: [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0
, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
---
```

- **Dataset Split**
  - Train: 80%
  - VAL: 10%
  - Test: 10%

Multilabel Klassifikation

Modellauswahl (Metriken für CN-Dataset)

Modell	Accuracy	F1	Beschreibung
Swin (Shifted Window Transformer)	ca. 65%	Ca. 80	General Purpose Backbone
Vision Transformer	Ca. 50%	Ca. 70	Erster Image Transformer
Vision Transformer Hybrid	Ca. 65%	Ca. 85	Kombination von CNN und Transformer
DeiT (Data-efficient image Transformers)	Ca. 55%	Ca. 75	Effizienz bei kleinen Datensätzen

Trainingsprozesse und Servernutzung im Machine Learning Lab

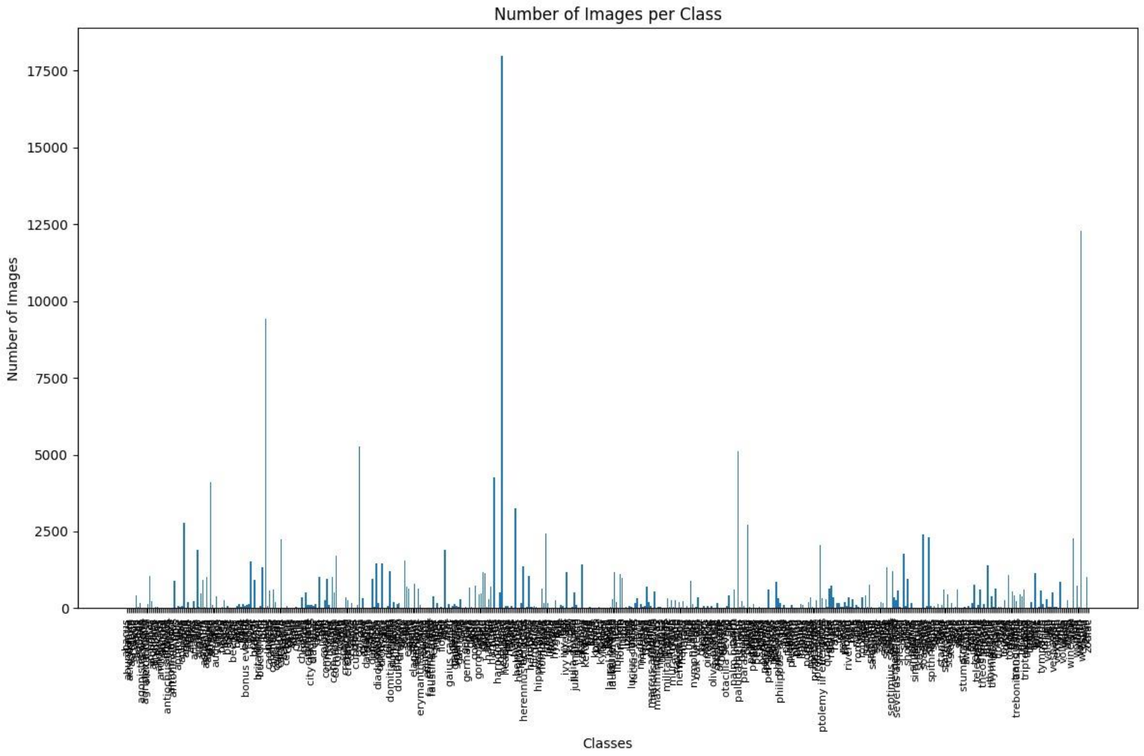
Training auf dem G4 Server des Machine Learning Lab unter Prof. Ramesh

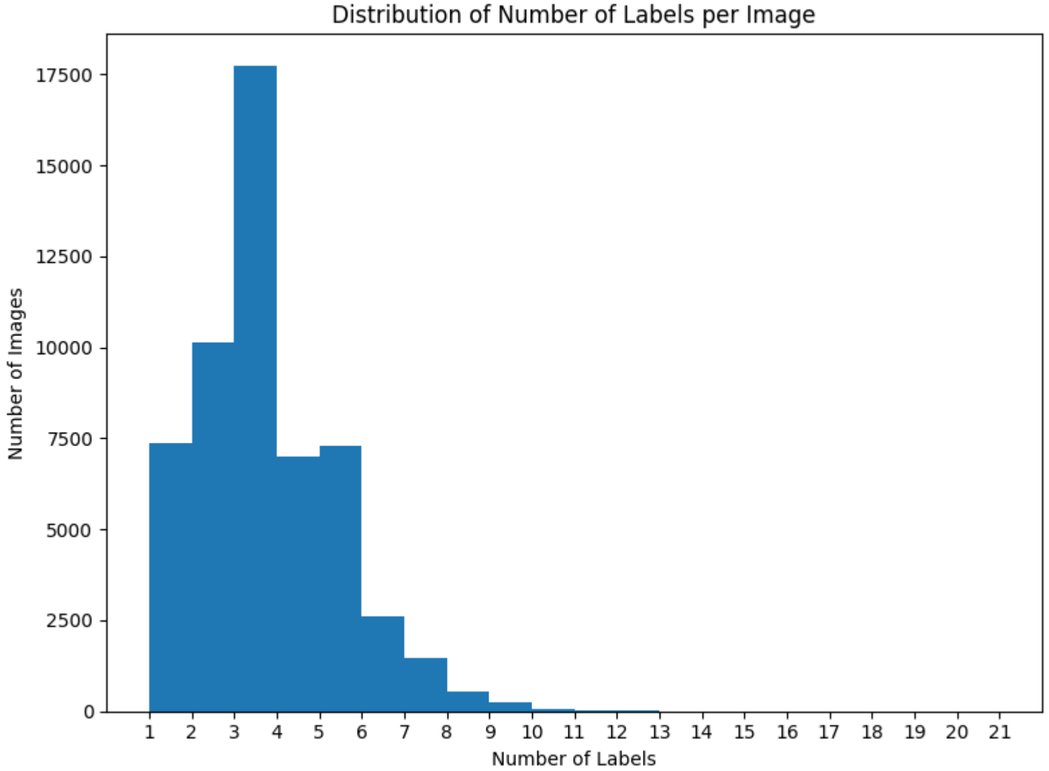
G4 Server	Trainingsdauer
4x Nvidia Tesla V50 GPUs	Eine Epoche dauert ca. 1,5-2 Stunden. 15 Epochen dauern ca. 20-30 Stunden.

100%|| 1360/1360 [18:56<00:00, 1.20it/s]
Epoch 5/25, Train Loss: 0.008875619610840017, Val Loss: 0.012416313676273l2, Val Acc: 0.5344732487589631, Val F1: 0.7835011227457681
100%|| 1360/1360 [18:52<00:00, 1.20it/s]
Epoch 6/25, Train Loss: 0.006880011992111309, Val Loss: 0.012139239208772778, Val Acc: 0.5517558374701231, Val F1: 0.7942775982391829
100%|| 1360/1360 [18:49<00:00, 1.20it/s]
Epoch 7/25, Train Loss: 0.005478392142258064, Val Loss: 0.011763934863676481, Val Acc: 0.5765765765765766, Val F1: 0.8078879429846133
100%|| 1360/1360 [18:50<00:00, 1.20it/s]
Epoch 8/25, Train Loss: 0.004355051430363l, Val Loss: 0.012231500312576l, Val Acc: 0.5822761537047252, Val F1: 0.8121213978941282
100%|| 1360/1360 [18:59<00:00, 1.19it/s]
Epoch 9/25, Train Loss: 0.003524191153915393l, Val Loss: 0.012218222619198702, Val Acc: 0.5903658760801618, Val F1: 0.8167584977324662
100%|| 1360/1360 [18:59<00:00, 1.19it/s]
Epoch 10/25, Train Loss: 0.002955973774677498l, Val Loss: 0.012503787276663762, Val Acc: 0.5912851627137341, Val F1: 0.8169020429414061
100%|| 1360/1360 [18:58<00:00, 1.19it/s]
Epoch 11/25, Train Loss: 0.002559197994945434, Val Loss: 0.012374800009488622, Val Acc: 0.6065453208310351, Val F1: 0.8220726421791059
100%|| 1360/1360 [18:47<00:00, 1.21it/s]
Epoch 12/25, Train Loss: 0.002060739322341225, Val Loss: 0.012857120471787366, Val Acc: 0.6032358889501747, Val F1: 0.8192469631845383
100%|| 1360/1360 [19:09<00:00, 1.18it/s]
Epoch 13/25, Train Loss: 0.001791038354520638l, Val Loss: 0.012990556923015152, Val Acc: 0.6144511858797573, Val F1: 0.8230809828617373
100%|| 1360/1360 [18:46<00:00, 1.21it/s]
Epoch 14/25, Train Loss: 0.001625819335330401, Val Loss: 0.013269396772717728, Val Acc: 0.6236440522154808, Val F1: 0.82924973900522
100%|| 1360/1360 [19:00<00:00, 1.19it/s]
Epoch 15/25, Train Loss: 0.001586614147698342l, Val Loss: 0.0140696534749997, Val Acc: 0.6074646074646075, Val F1: 0.822596358593389
100%|| 1360/1360 [18:47<00:00, 1.21it/s]
Epoch 16/25, Train Loss: 0.001288818722498356l, Val Loss: 0.013693445741527659, Val Acc: 0.6275050560764847, Val F1: 0.834856790481672
100%|| 1360/1360 [18:49<00:00, 1.20it/s]
Epoch 17/25, Train Loss: 0.001242454943049404l, Val Loss: 0.013849894628476571, Val Acc: 0.6280566280566281, Val F1: 0.829794534402167
100%|| 1360/1360 [18:54<00:00, 1.20it/s]
Epoch 18/25, Train Loss: 0.001128962931515536, Val Loss: 0.014318788699422l5, Val Acc: 0.6308144879573451, Val F1: 0.8344421484830279
100%|| 1360/1360 [18:43<00:00, 1.21it/s]
Epoch 19/25, Train Loss: 0.001074279923212837l, Val Loss: 0.014236802302355713, Val Acc: 0.6403750689464975, Val F1: 0.8388414572056448
100%|| 1360/1360 [18:46<00:00, 1.21it/s]
Epoch 20/25, Train Loss: 0.000969232537762068l, Val Loss: 0.014532398491385667, Val Acc: 0.6366979224122081, Val F1: 0.8344027238100656
100%|| 1360/1360 [18:50<00:00, 1.20it/s]
Epoch 21/25, Train Loss: 0.000928324552767341l, Val Loss: 0.014763619656236295, Val Acc: 0.6447876447876448, Val F1: 0.838519283053939
100%|| 1360/1360 [18:39<00:00, 1.21it/s]
Epoch 22/25, Train Loss: 0.000898953979102610l, Val Loss: 0.015243064208120546, Val Acc: 0.6230924802353374, Val F1: 0.827958310738013
100%|| 1360/1360 [19:06<00:00, 1.19it/s]
Epoch 23/25, Train Loss: 0.000850505221719418l, Val Loss: 0.014670340293634902, Val Acc: 0.6501195072623644, Val F1: 0.8425076065595432
100%|| 1360/1360 [18:34<00:00, 1.22it/s]
Epoch 24/25, Train Loss: 0.000886167933774285l, Val Loss: 0.015567314789137419, Val Acc: 0.6359624931053502, Val F1: 0.8354412432570012
100%|| 1360/1360 [19:01<00:00, 1.19it/s]
Epoch 25/25, Train Loss: 0.000741237216184612l, Val Loss: 0.015196186313679551, Val Acc: 0.6565545136973708, Val F1: 0.8431780351284994
Test Loss: 0.015771622021737344, Test Acc: 0.644170650974623, Test F1: 0.8340179011170724

Multilabel Klassifikation

Für alle 179014 Bilder (54489 ohne Duplikate) aus dem CN-Dataset





Multilabel Klassifikation

- **Ziel:** Untersuchung der Auswirkungen der extrem ungleichmäßigen Klassenverteilung auf die Modellleistung.

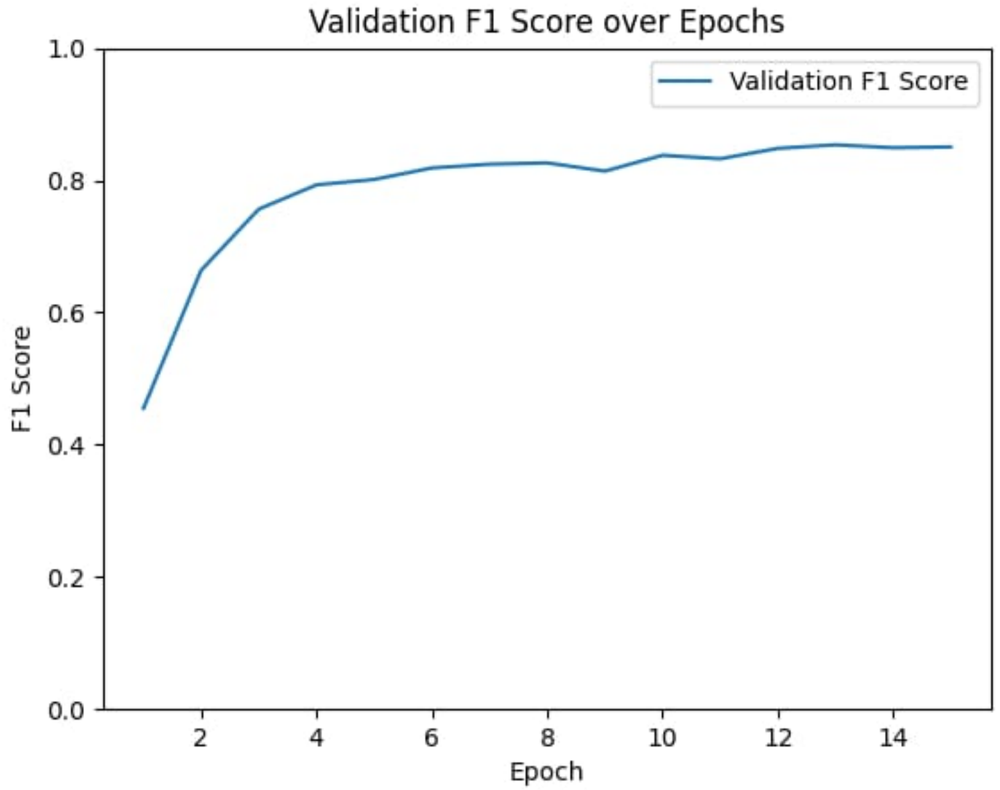
- **Übersicht:** Wie viele Klassen bleiben übrig, wenn Klassen mit weniger als einer bestimmten Anzahl von Bildern ausgeschlossen werden?

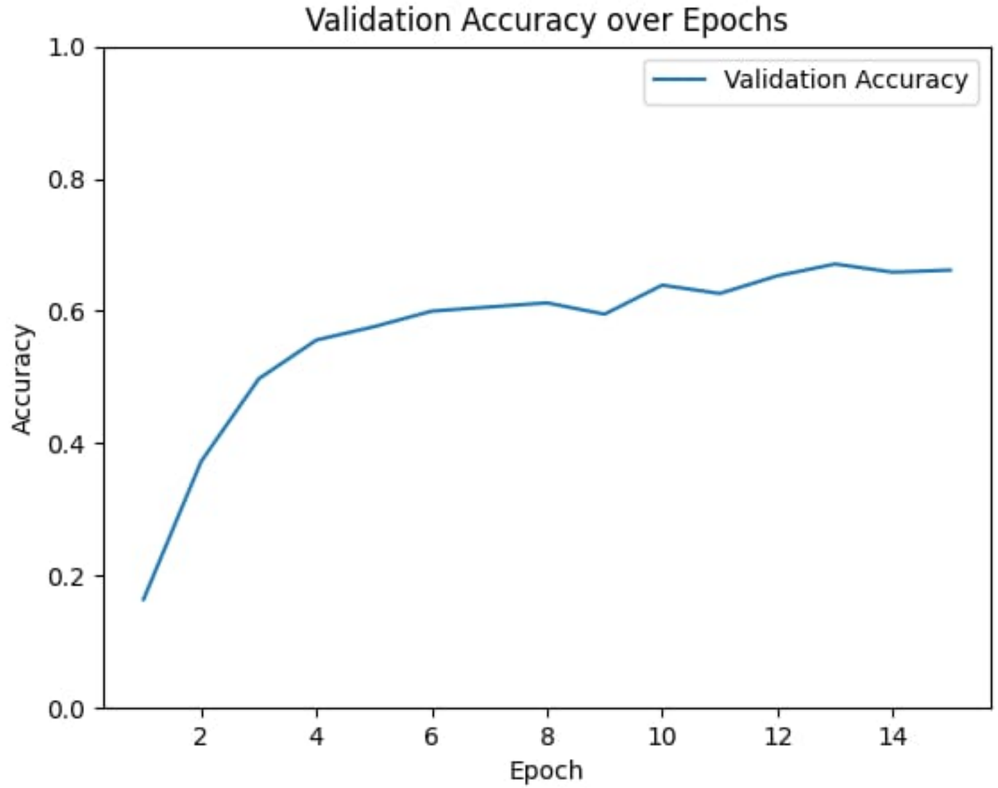
- **Analyse:** Untersuchen, ob das Entfernen von Klassen mit wenigen Bildern die Modellleistung verbessert, für eine bessere Generalisierung und effizienteres Training.

Dataset	0	100	500	1000
Train	43591	43015	41040	38862
Val	5448	5376	5130	4857
Test	5450	5378	5130	4857
Gesamt:	54489	53769	51300	48576
Anzahl Klassen:	506	190	86	44

Ergebnisse

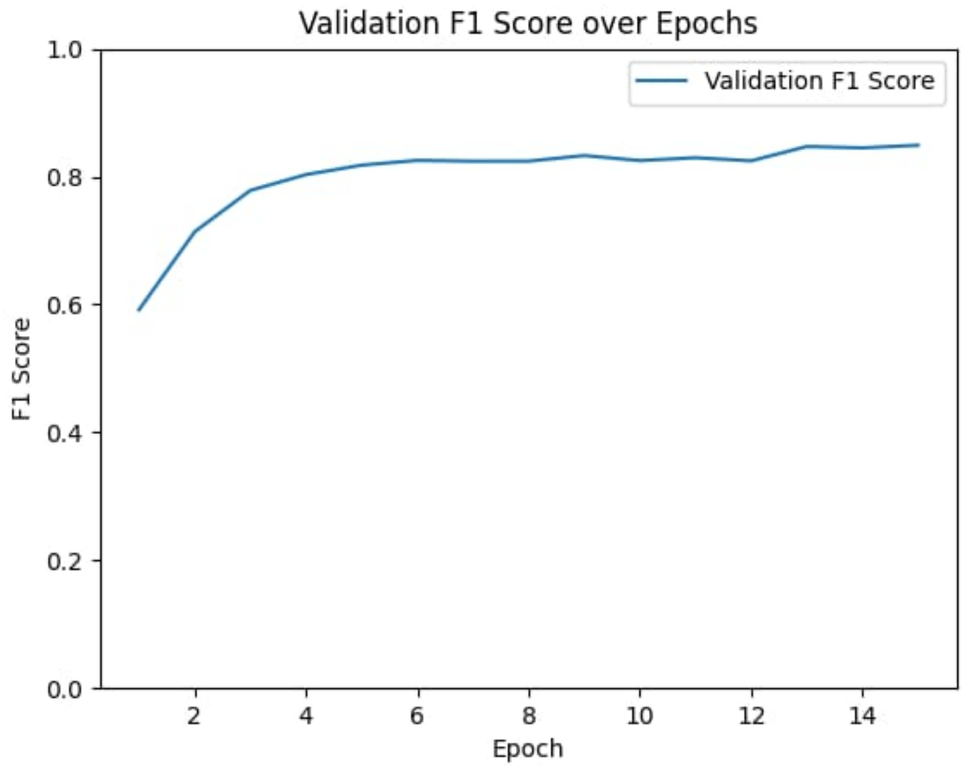
Ergebnisse unserer Analyse mit einem Schwellenwert von 0

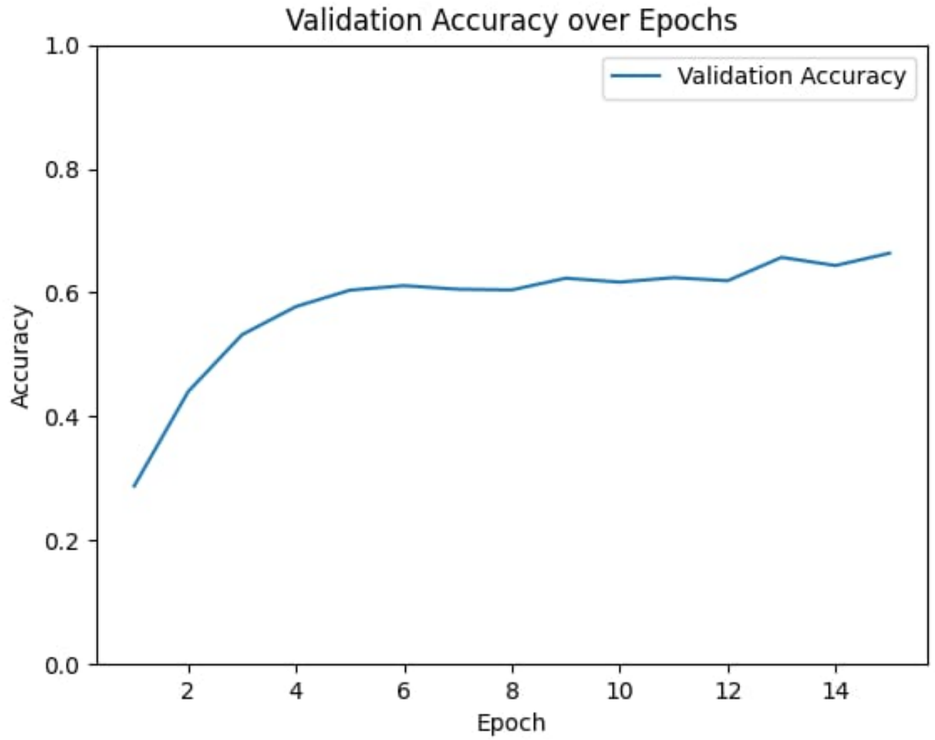




Ergebnisse

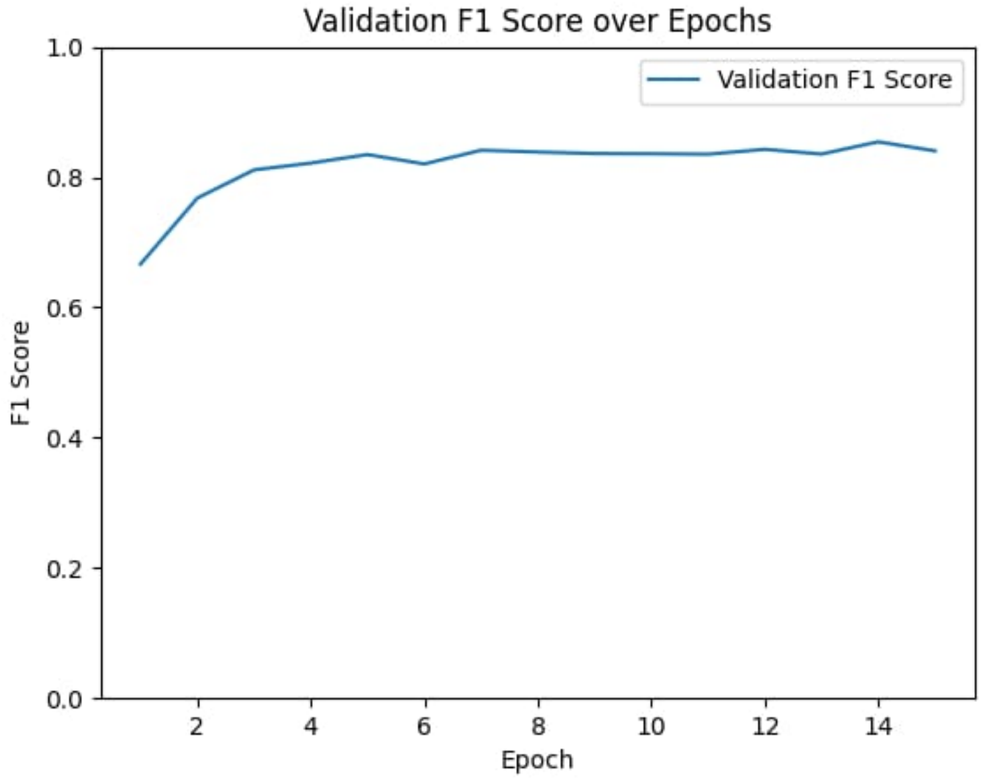
Ergebnisse unserer Analyse mit einem Schwellenwert von 100

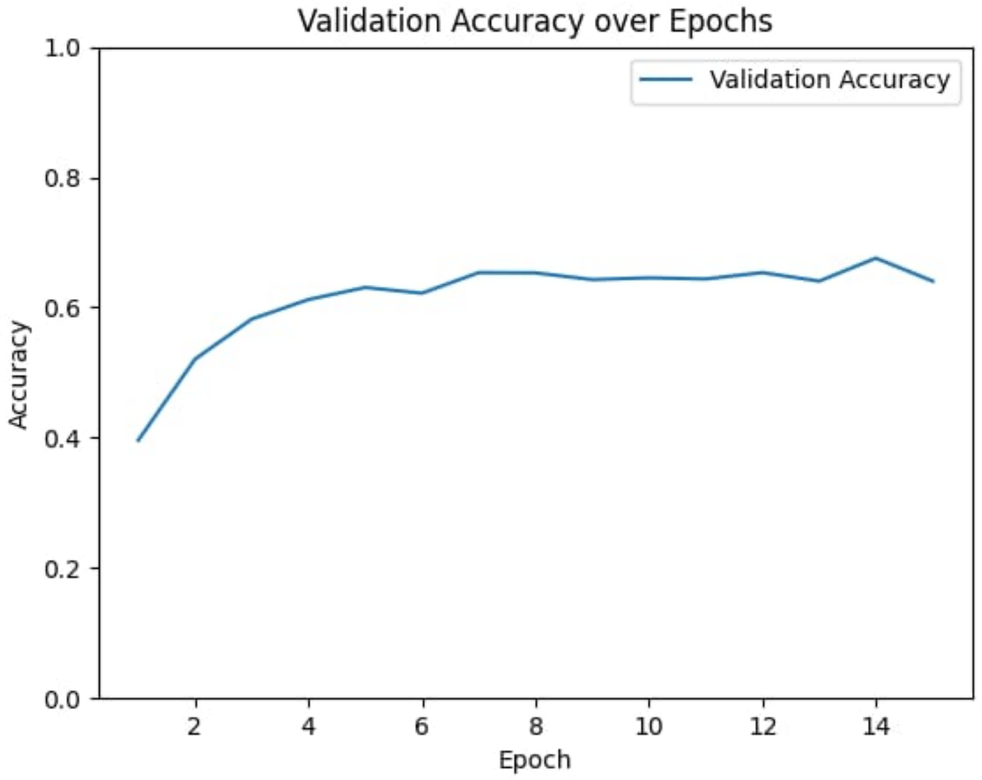




Ergebnisse

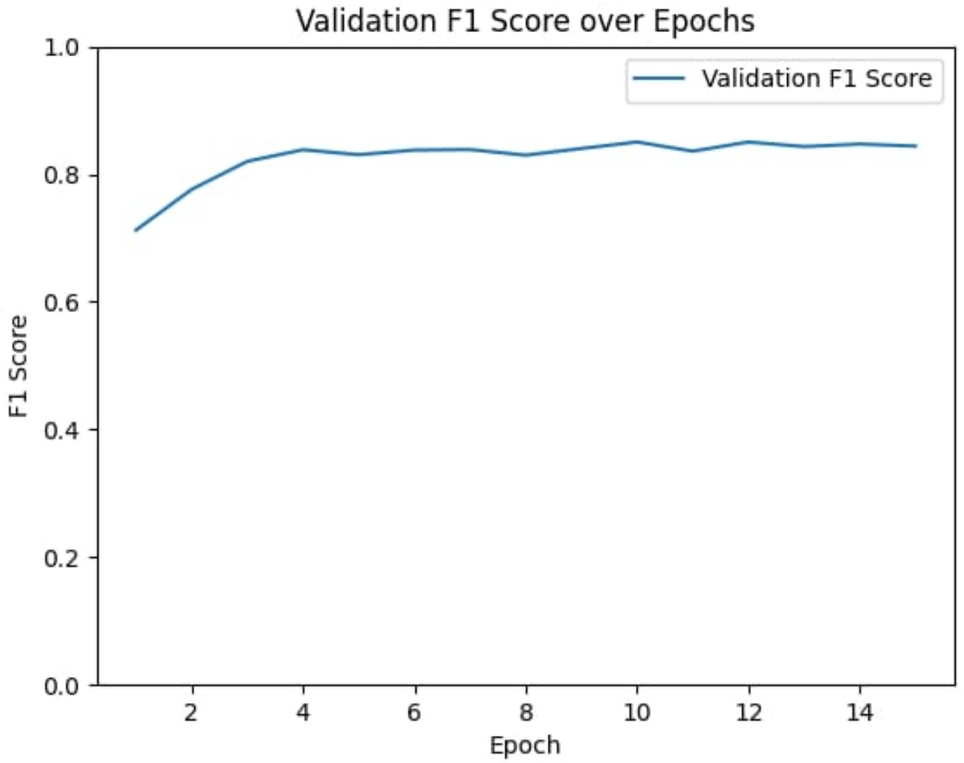
Ergebnisse unserer Analyse mit einem Schwellenwert von 500

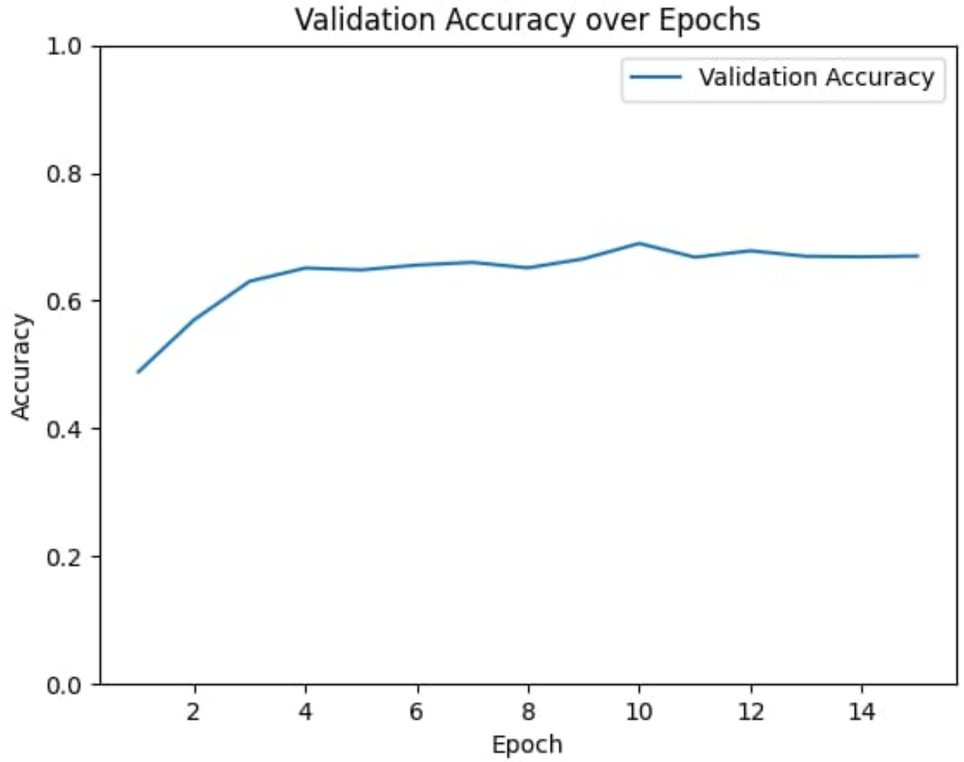




Ergebnisse

Ergebnisse unserer Analyse mit einem Schwellenwert von 1000





Beste und Schlechteste Ergebnisse nach F1-Score für Schwellwert 0

Beste

F1-Score: 0.5

Wahre Labels:
'serpent staff', 'hand', 'asclepius‘

Vorhergesagte Labels:
'asclepius‘

Schlechteste

F1-Score: 0.0

Wahre Labels:
'head', 'tunny‘

Vorhergesagte Labels:
'gorgo', 'laurel branch', 'satrap‘

Beste und Schlechteste Ergebnisse nach F1-Score für Schwellwert 1000

Beste

F1-Score: 0.3556

Wahre Labels:
'bull', 'cornucopia‘, 'eagle', 'foot', 'scepter', 'patera', 'throne', 'zeus‘

Vorhergesagte Labels:
'altar', 'apollo', 'athena', 'bow', 'bull', 'bust', 'caracalla', 'club', 'corn', 'cornucopia', 'cuirass', 'diadem', 'dionysus', 'dolphin', 'eagle', 'foot', 'grape', 'griffin', 'hand', 'head', 'helmet', 'heracles', 'horse', 'ivy wreath', 'kalathos', 'laurel wreath', 'lion', 'patera', 'protome', 'scepter', 'septimius severus', 'shield', 'snake', 'spear', 'throne', 'tunny', 'zeus'

Schlechteste

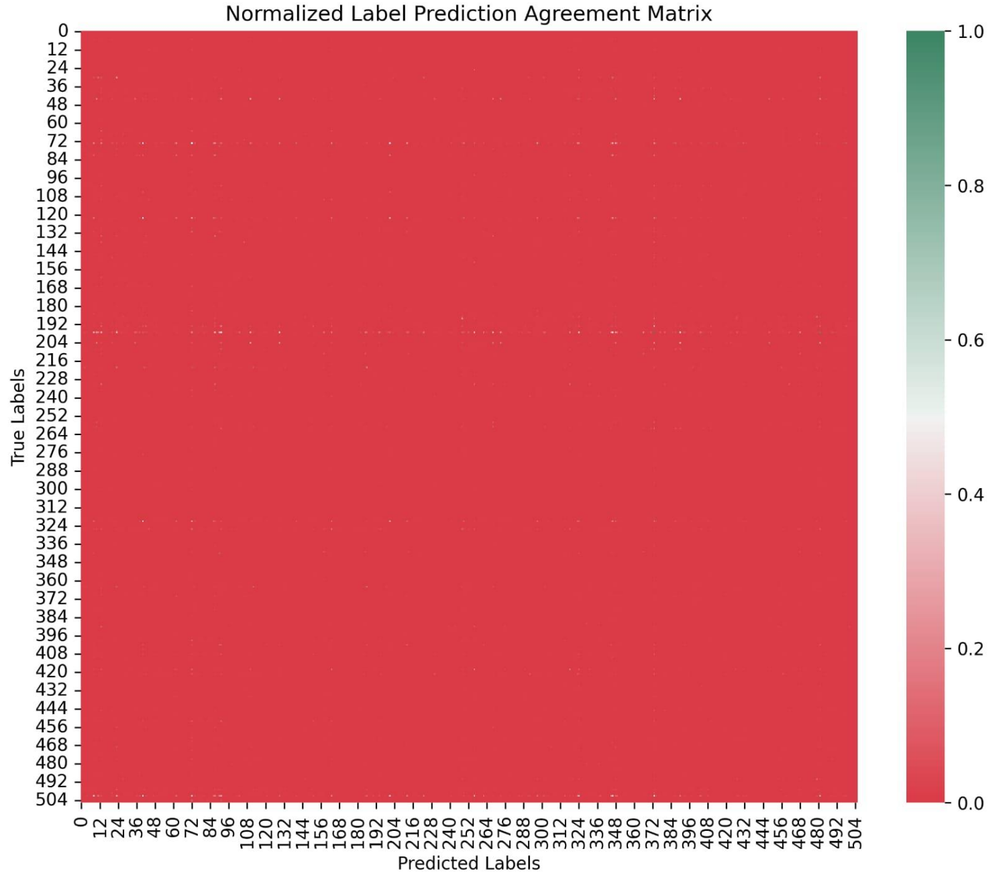
F1-Score: 0.0

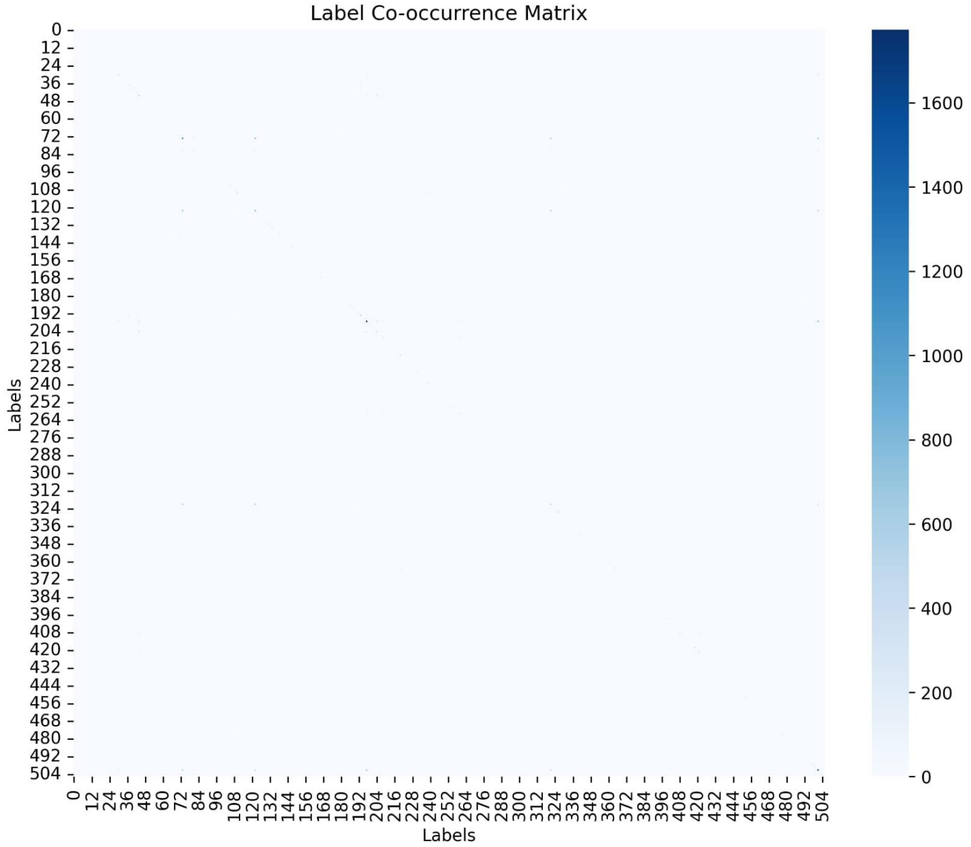
Wahre Labels:
‚wing‘

Vorhergesagte Labels:
'altar', 'arm', 'athena', 'bow', 'bull', 'bust', 'caracalla', 'club', 'corn', 'cuirass', 'diadem', 'dionysus', 'eagle', 'foot', 'grape', 'head', 'helmet', 'heracles', 'hermes', 'horse', 'ivy wreath', 'paludamentum', 'protome', 'scepter', 'septimius severus', 'shield', 'snake', 'throne', 'torch', 'tunny', 'wreath'

Ergebnisse

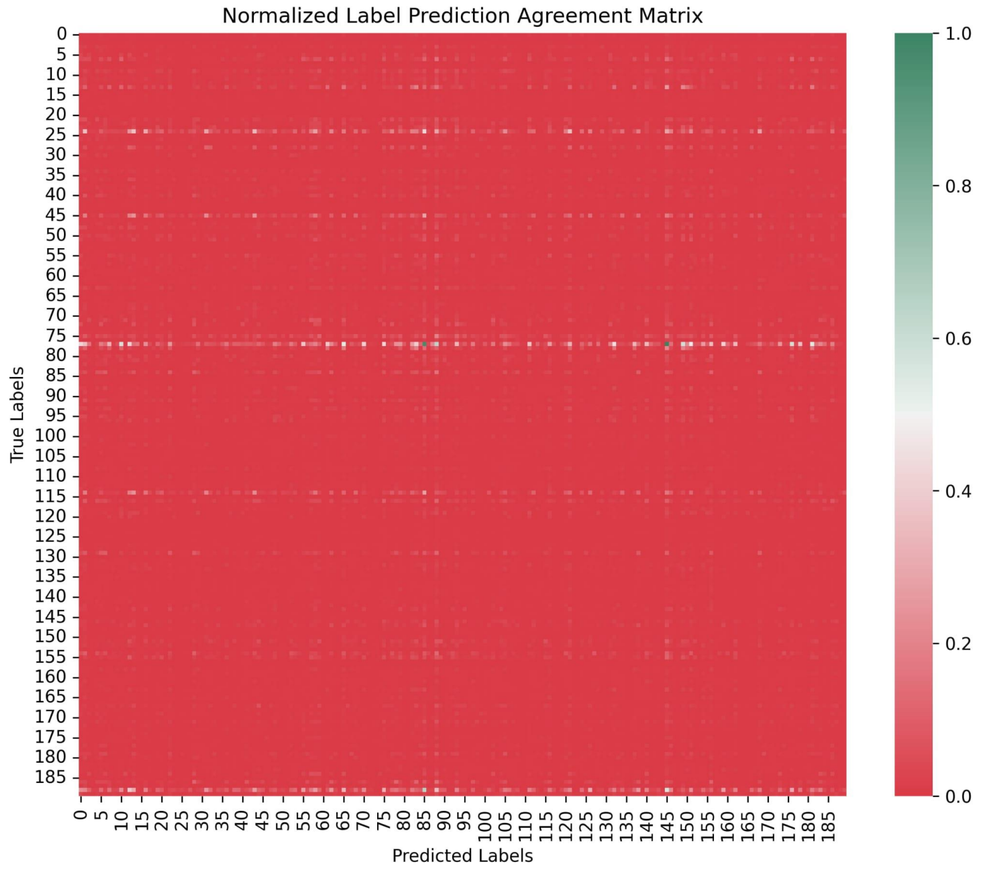
Ergebnisse unserer Analyse mit einem Schwellenwert von 0 (Anzahl Bilder: 54489)

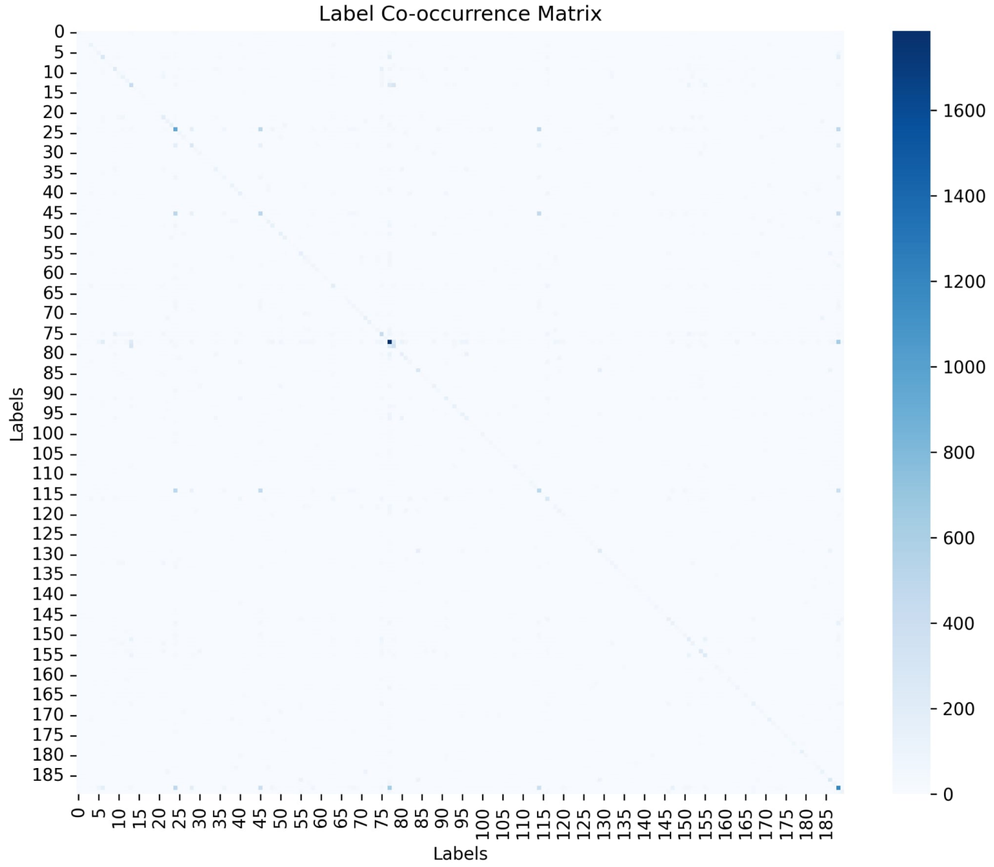




Ergebnisse

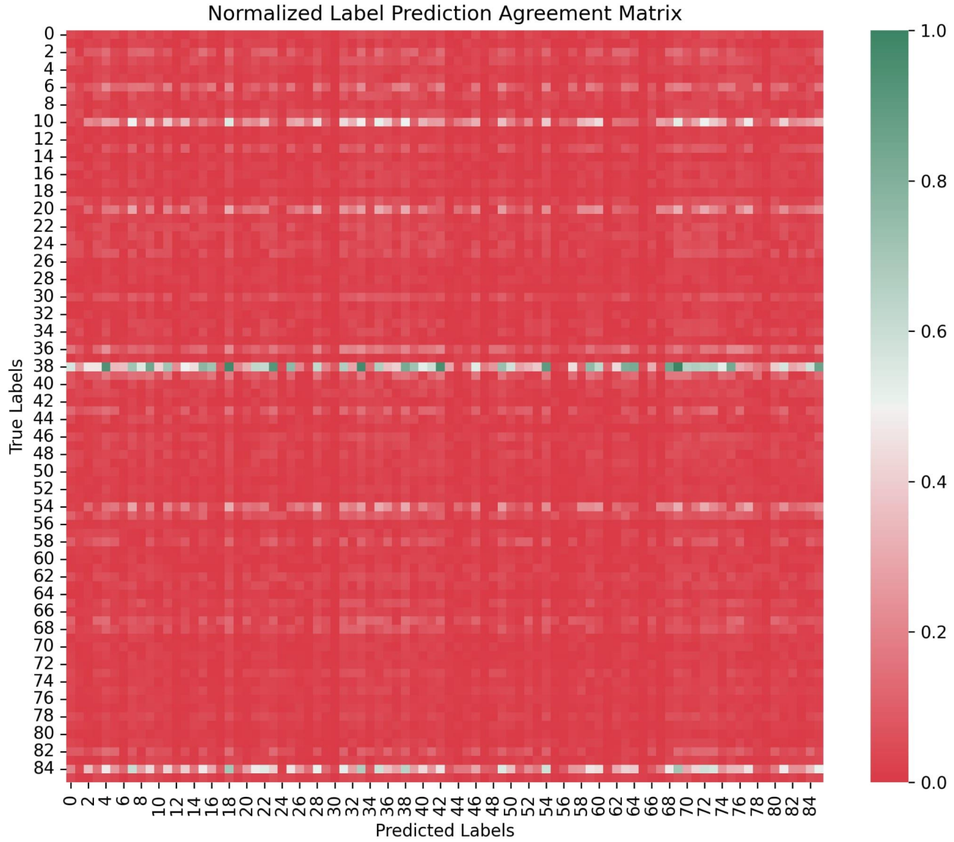
Ergebnisse unserer Analyse mit einem Schwellenwert von 100 (Anzahl Bilder: 53769)

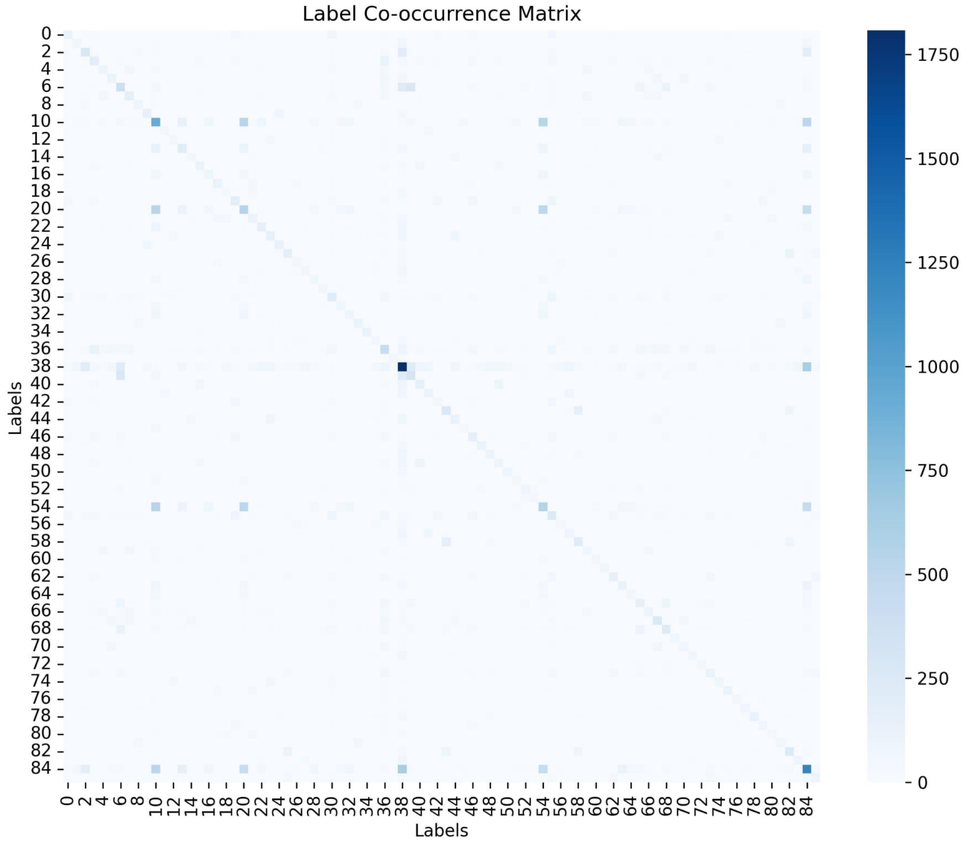




Ergebnisse

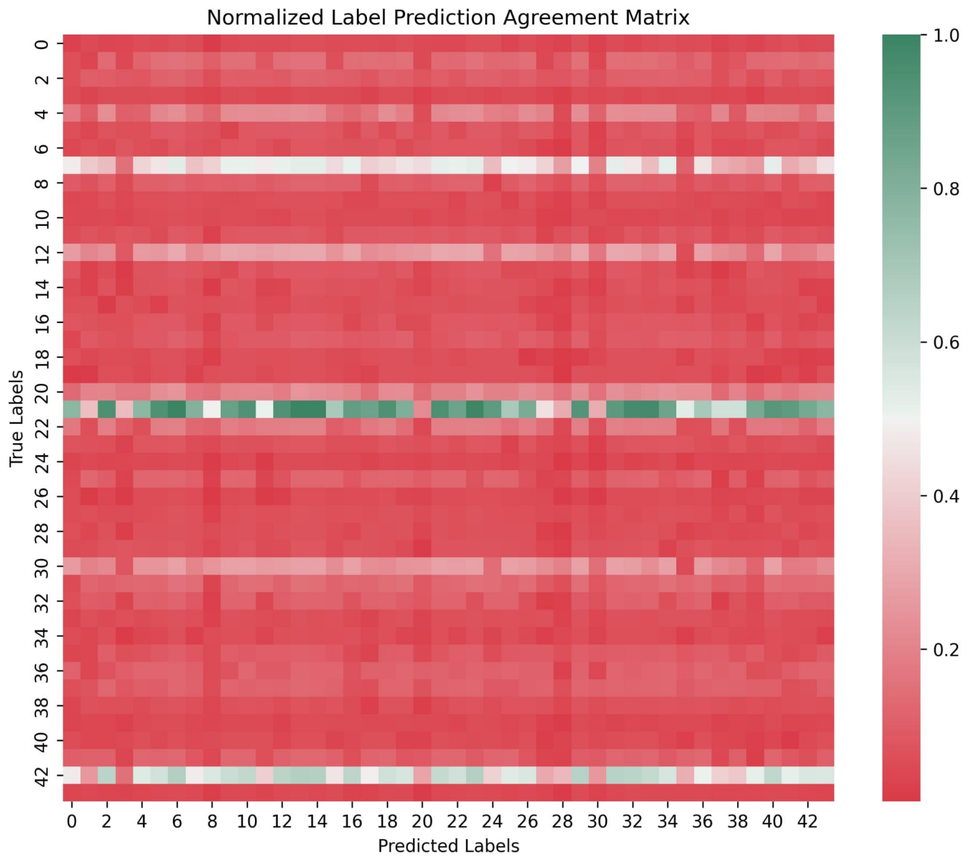
Ergebnisse unserer Analyse mit einem Schwellenwert von 500 (Anzahl Bilder: 51300)

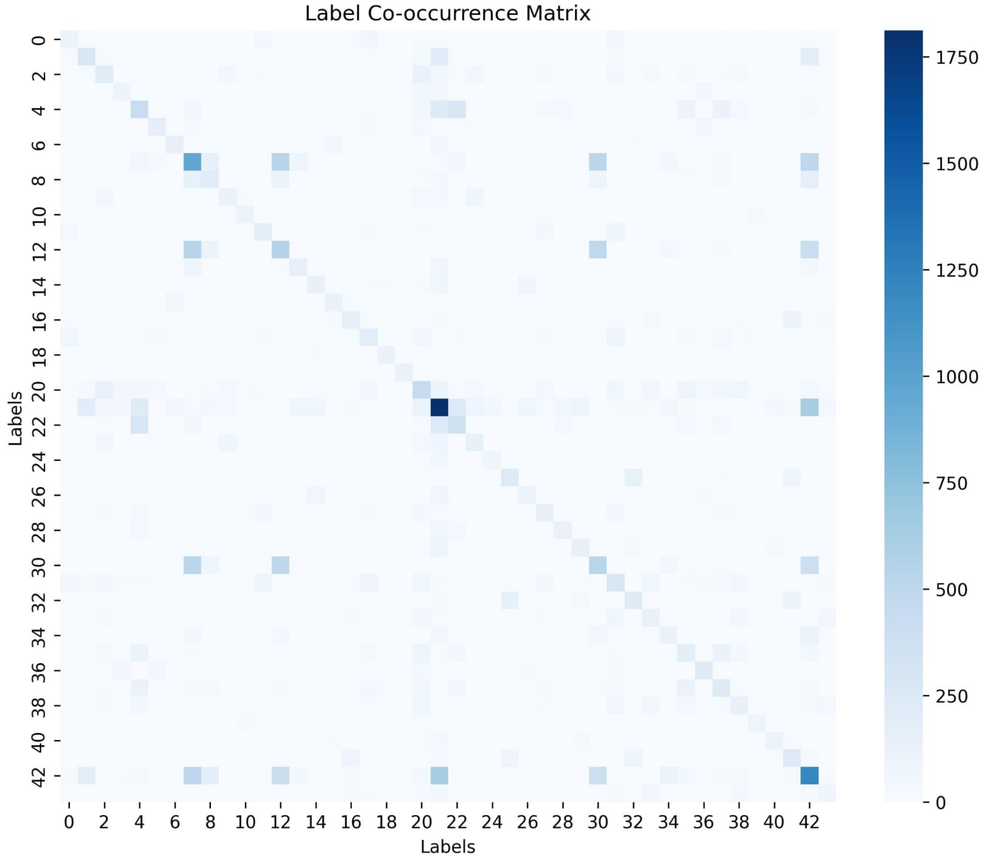




Ergebnisse

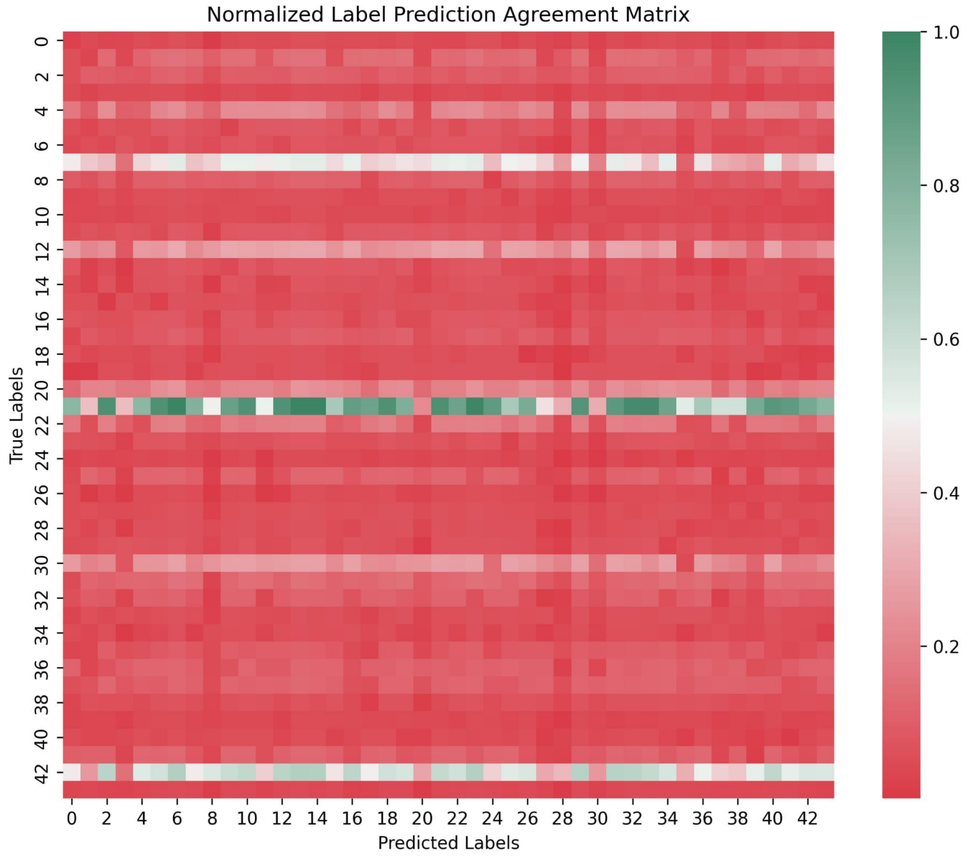
Ergebnisse unserer Analyse mit einem Schwellenwert von 1000 (Anzahl Bilder: 48576)





Ergebnisse

Ergebnisse unserer Analyse mit einem Schwellenwert von 1000 (Anzahl Bilder: 48576)



Welche Klasse werden in der label prediction Agreement Matrix überrepräsentiert?

- Bei Schwellenwert **1000:**
  - Label ID 7: 'bust‘
  - Label ID 21: 'head‘
  - Label ID 42: 'wreath'

.

Fazit

Multi-Label-Ansatz:

- Trotz leichter Leistungssteigerungen Schwierigkeiten, die Klassen korrekt zu klassifizieren, unabhängig vom Schwellenwert
- Gesamtleistung des Modells suboptimal
- Weitere Untersuchungen und Optimierungen sind notwendig

Ausblick

Für den Multi-Label-Ansatz:

- Sollten wir die Anzahl der Labels pro Bild nach oben begrenzen?
- Erweiterung auf Graustufenbilder
- Gewichtung der Klassen

Integration von CLIP in die Pipeline

- Ansatz: Einführung von CLIP in die bestehende Pipeline
- Ziel: Nutzen des Text Encoders zur Labelfindung neben dem Vision Encoder
- Methode:
  - Integration des Vision Encoders
  - Nutzung des Text Encoders zur Labelfindung

Quellen

- Rohit Kundu,The Beginner’s Guide to Contrastive Learning (v7labs.com), https://www.v7labs.com/blog/contrastive-learning-guide#h1, 22. Mai 2022, (letzter Zugriff: 02.05.2024)
- George Lawton, Was ist Transformer-Modell? - Definition von Computer Weekly, https://www.computerweekly.com/de/definition/Transformer-Modell, Januar 2024, (Zugriff: 02.05.2024)
- Strikingloo, Do Vision Transformers See Like Convolutional Neural Networks? (strikingloo.github.io), 07 Sep 2021, (Zugriff: 09.05.2024)
- Robert Geirhos, Kantharaju Narayanappa, Benjamin Mitzkus, Tizian Thieringer, Matthias Bethge, Felix A. Wichmann, Wieland Brendel. Partial success in closing the gap between human and machine vision. https://doi.org/10.48550/arXiv.2106.07411, 14. Juni 2021 (Letzter Zugriff: 09.05.2024)
- Alec Radford, Ilya Sutskever, Jong Wook Kim, Gretchen Krueger, Sandhini Agarwal, CLIP: Connecting text and images | OpenAI, 5. Januar 2021
- Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, Xiao Wang, Xiaohua Zhai, Thomas Kipf, Neil Houlsby. Simple Open-Vocabulary Object Detection with Vision Transformers. https://doi.org/10.48550/arXiv.2205.06230. 12. Mai 2022 (Letzter Zugriff: 09.05.2024)

Quellen

- Pantelis Monogioudis, Fast and Faster RCNN Object Detection, https://pantelis.github.io/artificial-intelligence/aiml-common/lectures/scene-understanding/object-detection/faster-rcnn-object-detection/index.html, 2023 (letzter Zugriff: 09.07.2024)
- MathWorks. "Multilabel Image Classification Using Deep Learning." MathWorks, https://de.mathworks.com/help/deeplearning/ug/multilabel-image-classification-using-deep-learning.html (Letzter Zugriff: 09.07.2024)

Danke für die Aufmerksamkeit!