ffk Journal
Film- und Medienwissenschaftliches Kolloquium

Lisa Rein
Weimar

„Basketballplayers", „Cheerleaders" und andere „Persons"

Zu Epistemologie und Visualität der Bilddatenbank ImageNet

Abstract: Ausgehend von der Kategorie „basketballplayer" analysiert der Beitrag, wie in ImageNet, der Pionier-Bilddatenbank der Computer Vision, Bilder von Körpern klassifiziert werden. Dazu zeichne ich kursorisch die Entstehungsgeschichte von ImageNet nach und analysiere im Anschluss die epistemische Struktur der Datenbank anhand der drei Ebenen Bild, Text und Bild-Text-Relation. Ausgehend vom *screenwalking* durch die Datenbank zeige ich, dass Epistemologie und Visualität von ImageNet auf den Prämissen beruhen, dass die visuelle Welt erstens in klar voneinander abgrenzbaren Objekten erfassbar ist, und dass sich Bilder und Begriffe zweitens eindeutig und unmittelbar einander zuordnen lassen. Darüber hinaus wird deutlich, dass ImageNet stark von ontologischen Bedingungen, wie der Verfügbarkeit von Daten, geprägt ist, und damit in der Praxis eine *messy* Assemblage menschlicher und maschineller Informationsverarbeitung bildet.

Lisa Rein (M.A.), wissenschaftliche Mitarbeiterin im DFG-Projekt „Curating the Feed – Interdisziplinäre Perspektiven auf digitale Bilderfeeds und ihre *Curatorial Assemblages*" an der Bauhaus-Universität Weimar an der Professur Digitale Kulturen. Aktuelle Forschungsschwerpunkte: Bilddatenbanken und fotografische Archive, Critical Algorithm Studies, Computer Vision, Fotografie- und Archivtheorie.

1. Circle it, classify it

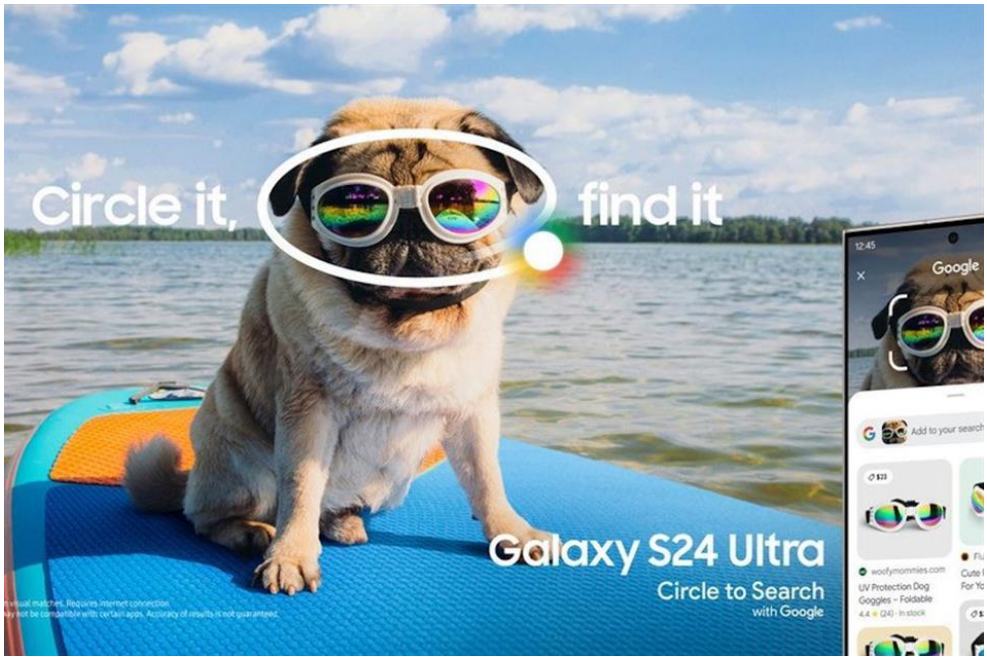

Abb. 1: Werbekampagne „Circle it, find it" für *Samsung Galaxy*, 2024.

Samsung bewirbt an seinem neuesten Galaxy Smartphone vor allem die „KI"-Funktionen, und insbesondere die „Circle to Search"-Funktion: eine Weiterentwicklung von Google Lens, die von Nutzer_innen per Touch eingekreiste Objekte auf Bildern erkennt. Diese können im Anschluss basierend auf verschiedenen Optionen weitergenutzt werden: zum Beispiel bei der Suche nach den hippen Cowboystiefeln aus dem neuesten Instagram-Video, oder bei der Recherche zu dem buddhistischen Tempel, der im letzten Urlaub fotografiert wurde.[1] Auch Apples iPhones können bereits seit 2022 mit dem Tool „Visuelles Nachschlagen" Objekte in Bildern identifizieren und dann z. B. den Namen einer unbekannten Pflanze ausfindig machen oder ein Rezept für das im Restaurant fotografierte Gericht heraussuchen.[2] Eine ähnliche Form der „Objekterkennung" nutzt der iPhone-„Porträtmodus": Um die tiefenscharfe Ästhetik analoger Spiegelreflexkameras zu simulieren, werden Machine-Learning-Modelle der Computer Vision eingesetzt, die „Objekte" (in der Regel Menschen) identifizieren können, welche dann – in der Tradition der klassischen Porträtfotografie – visuell vom Hintergrund abgesetzt werden.[3]

1 Vgl „Introducing a new way to search | Circle to Search".
2 Vgl. Walker-Todd 2023.
3 Vgl. *On-Device Panoptic Segmentation for Camera Using Transformers* 2021.

Die algorithmische Bildklassifizierung hat Fuß gefasst in alltäglichen Medienumgebungen, in denen sie menschliche Körper, Gebäude oder hippe Sonnenbrillen gleichermaßen aufspüren, zuordnen und weiterverarbeiten kann. Dabei sind es gerade diese trivialen Formen der Anwendung, die sich im Alltag von Milliarden von Smartphone-Nutzer_innen einschreiben — wenngleich sie im öffentlichen Diskurs weniger präsent sind als die Gesichtserkennung oder das automatisierte Fahren. All diese Anwendungsbereiche gehören zur sogenannten „Computer Vision", einem Teilbereich der Informatik, der sich seit den 1960er Jahren mit einem automatisierten Verständnis komplexer Bildzusammenhänge beschäftigt.[4] Sie haben gemeinsam, dass sie auf algorithmischen Modellen basieren, die Bilder verrechnen, analysieren und klassifizieren, und dass diese Modelle auf Datenbanken aufbauen, die große Mengen an Bildern mit einer zweiten Dateneinheit verknüpfen. Wenn es sich um Bilder von Körpern handelt, werden dabei also dreidimensionale Körper zu zweidimensionalen Bildern umformatiert, die wiederum in Pixelkonstellationen gerechnet werden. Diese Konstellationen werden schließlich algorithmisch gelesen und semantischen Einheiten – z. B. dem Label „girl", „basketball player" oder „supermom" – als zweiter Dateneinheit zugeordnet.[5] Die Bilddatenbanken bilden somit das Herzstück algorithmischer Bildklassifizierung. Der vorliegende Beitrag widmet sich ImageNet, der *godmother* ebenjener Bilddatenbanken, und fragt nach ihrer epistemischen Ordnung und Visualität, insbesondere in Bezug auf die Repräsentation von Körpern. Dabei soll die Analyse der epistemischen Bedingungen am Beispiel von ImageNet zeigen, dass algorithmische Bildklassifizierung auf einem dinghaften Verständnis der Welt aufbaut, wodurch Körper zu schematischen Objekten strukturiert werden.

2. ImageNet: *Godmother* der Computer Vision

Der Ordner „n09842047_basketballplayer" liegt im Finder meines Macbooks zwischen „n09841696_baserunner" und „n09842288_basketweaver". Rechtsklick auf den Ordner: Darin liegen 1228 „Objekte", so nennt es die Informationsanzeige des Finders, ich klicke auf den Ordner und scrolle durch eine Liste an Dateien, die Vorschau rechts zeigt Bilder an: Überbelichtete Amateurfotografien von Basketballspieler_innen in schlecht ausgeleuchteten Hallen; Schnappschüsse von Kindern und jungen oder mittelalten Erwachsenen auf Basketballplätzen; mal mit Ball, mal ohne, mal in Trikots, mal nicht, mal in Bewegung, mal still, mal mehrere Personen, mal nur ein Körperteil, viele Schwarze Basketballer_innen, einige weiße*, wenige Frauen, viele Kinder; Spieler_innen in Umkleiden und Fotostudios, für Gruppenfotos arrangiert, am Händeschütteln und Pokaleschwenken; Autogrammkarten von Basketballspielern; Pressefotos in Schwarzweiß; dann wieder Bälle, Körbe, Hände.*

[4] Hunger 2021: 2.

[5] Beispiel-Kategorien aus dem Datensatz ImageNet-21k.

Bei diesem Sammelsurium von fotografischen Bildern, deren größte Gemeinsamkeit darin besteht, dass sie in einem Ordner mit dem Namen „n09842047_basketballplayer“ liegen, handelt es sich um einen Ausschnitt aus der Bilddatenbank ImageNet. ImageNet wurde an den US-amerikanischen Universitäten Stanford und Princeton entwickelt, 2009 erstmals veröffentlicht, und beinhaltet in der vollständigen Version vierzehn Millionen gelabelte Bilder, die in über 21.000 Kategorien sortiert sind,[6] von „a“ wie „abacus“ bis „z“ wie „zoo keeper“ – oder „b“ wie „basketballplayer“. Die Bilder in der Datenbank wurden zwischen 2007 und 2010 aus dem Internet heruntergeladen, wobei ein großer Teil von der frühen Fotosharing-Plattform Flickr stammt.[7] Die Kategorien oder „Labels“, denen die Bilder zugeordnet sind, entstammen WordNet, einer in den 1980er Jahren entwickelten Begriffstaxonomie,[8] und die Zuordnungen von Bildern zu Labels wurden von knapp 50.000 über Amazon Mechanical Turk beschäftigten Clickworker_innen ausgeführt.[9]

Das Forschungsprojekt und die Datenbank ImageNet haben über das Feld der Computer Vision hinaus entscheidend zu den wichtigen Transformationen des letzten Jahrzehnts im Bereich der sogenannten Künstlichen Intelligenz, sprich zur Renaissance neuronaler Netze unter der Bezeichnung „deep learning“ und zum Big Data-Paradigma, beigetragen.[10] Auch wenn mittlerweile andere Bilddatenbanken wie LAION, die z. B. für KI-Bild-Generatoren wie Stable Diffusion und Midjourney verwendet wird, im Zentrum aktueller KI-Diskurse stehen, wirkt ImageNet fort: Die Datenbank zählt nach wie vor zu den wichtigsten Trainingsstandards im Bereich Computer Vision und wird bei nahezu allen neueren Modellen und Datenbanken als Kontrollinstanz zu Grunde gelegt. Somit manifestiert sich die epistemische Ordnung von ImageNet stets aufs Neue in Infrastrukturen der Bildklassifizierung.[11]

3. *Screenwalking* ImageNet

Wie kann man vierzehn Millionen Bildern begegnen? ImageNets wohl bekannteste Kritikerin Kate Crawford rechnet uns vor, dass, wenn wir jedes Bild aus der Datenbank für zehn Sekunden anschauen würden, wir viereinhalb Jahre beschäftigt wären.[12] In Anbetracht dieser schieren Massen an Bildern ist es einerseits naheliegend, auch in der Erforschung der Bilddatenbanken auf computerbasierte, quantitative Bildanalysewerkzeuge zurückzugreifen. Andererseits laufen Methoden wie die Data Analytics, die große Mengen von Daten automatisiert

6 Vgl. Russakovsky et al. 2015: 2–3.
7 Vgl. Malevé/Sluis 2023.
8 Vgl. Yang et al. 2020: 3.
9 Vgl. Fei-Fei 2017.
10 Vgl. Denton et al. 2021: 5–6.
11 Vgl. Offert/Bell 2021: 1141.
12 Vgl. Crawford o. J.

auswerten, stets Gefahr, die Komplexitätsreduktion der Datenbanken zu reproduzieren. Zudem steht ImageNet seit mehreren Jahren im Fokus kritischer Forschung zu Bilddatenbanken, und ist dementsprechend bereits hinreichend in seiner Gesamtheit untersucht worden, u.a. mithilfe automatisierter, quantitativer Methoden[13] oder klassischer Diskursanalysen[14]. Daher schlage ich mit dem vorliegenden Beitrag einen anderen Weg ein: Inspiriert von Estelle Blaschkes Beforschung von kommerziellen Massen-Bildsammlungen[15] stelle ich das gezielte Stöbern und Streunen in den Bildermeeren der Datenbank in den Vordergrund der Wissensproduktion. Ich gehe davon aus, dass explorative Mikro-Erkundungen des Bildmaterials Erkenntnisse über Bilddatenbanken, ihre epistemischen Einschreibungen und ihre Visualität erzeugen können, welche großflächige, quantitative Methoden nicht zu generieren vermögen. In Anlehnung an die *walkthrough*-Methode, die häufig für die kultur- und medienwissenschaftliche Analyse von digitalen Bildschirmartefakten verwendet wird,[16] setze ich hier das *screenwalking* als assoziativere Version der *walkthroughs* ein, als Bildschirmspaziergang, der Raum für Umwege und zufällige Entdeckungen bietet und damit ein punktuelles *close reading* von Bilddatenbanken ermöglicht.

4. Körper in ImageNet

Ich scrolle weiter durch die Liste der Bildobjekte in „n09842047_basketballplayer". Mein Blick bleibt an einem fast quadratischen Bild hängen, das wenig in die Kategorie zu passen scheint: Es zeigt eine Cheerleaderin von hinten, ab der Taille aufwärts: Ein Arm mit Pompon in die Höhe gereckt, den anderen abgewinkelt vorm Körper, sichtbar muskulös, sie trägt ein bordeauxrotes Top und eine Schleife im langen braunen Haar und scheint mitten in einem cheer zu sein. Im unscharfen Hintergrund kann ich Figuren in Trikots und einen Basketballkorb erahnen, auch dieses Bild ist in einer Sporthalle geschossen. Ich starre auf den in die Luft gereckten Pompon, die Haare, den Rücken dieser Cheerleaderin, in der leisen Hoffnung, dass sie sich umdreht, aber sie bleibt erstarrt in ihrer Bewegung.

Die ImageNet-Kategorien sind innerhalb einer hierarchischen Taxonomie verortet: Das *subset* „basketballplayer" ist eine Unterkategorie der Hauptkategorie „person", welche wiederum eine der neun Oberkategorien der Datenbank bildet.[17] Für den Annotationsprozess wurden den Kategorien kurze Definitionen und Beispielsätze zur Seite gestellt. Die Definition von „basketballplayer" in der Liste der ImageNet-

13 Vgl. z. B. Prabhu/Birhane 2020, Yang et al. 2020.
14 Vgl. Denton et al. 2021, Hanna et al. 2020.
15 Vgl. Blaschke 2016: 14–15.
16 Vgl. Light et al. 2018.
17 Die anderen acht Oberkategorien sind: „plant", „geologic formation", „natural object", „sport", „artifact", „fungus", „animal", und „miscellaneous"; vgl. Crawford, 2021: 137.

Kategorien lautet „an athlete who plays basketball“[18]. In der vollständigen Version von ImageNet (ImageNet-21k) liegen innerhalb des „person“-*subtrees* 2832 Unterkategorien, darunter unter anderem zutiefst rassistische, misogyne, ableistische, queerfeindliche und anderweitig diskriminierende Kategorien.[19] Nachdem Kate Crawford und Trevor Paglen öffentlichkeitswirksam auf die verletzenden Zuschreibungen eines großen Teils des ImageNet-„person“-*subtrees* aufmerksam gemacht hatten,[20] wurde der *subtree* 2020 von den ImageNet-Entwickler_innen einer gründlichen Revision unterzogen.[21] Dabei bewerteten sie etwa die Hälfte der unter „person“ subsumierten Kategorien als potenziell „offensive“, und 2674 der 2832 Kategorien als nicht ausreichend „imageable“, also nicht vorstellbar.[22] Als Konsequenz dieser Revision beinhalten aktuelle Versionen von ImageNet-21k nur noch die verbleibenden 158 als harmlos und vorstellbar eingestuften „person“-*subsets*, darunter auch „basketballplayer“. Jedoch finden Derivate der ursprünglichen ImageNet-21k-Version inklusive aller ursprünglicher „person“-Unterkategorien nach wie vor rege Verwendung im Feld der sogenannten „Objekterkennung“.[23] Zudem sind, wie oben beschrieben, in den fünfzehn Jahren, in denen ImageNet existiert, bereits unzählige algorithmische Modelle auf ebenjener Version der Datenbank trainiert worden.[24]

In diesem Beitrag soll es jedoch um eine Kategorie gehen, die sowohl als harmlos als auch als vorstellbar kategorisiert wurde, in der englischen Version des Wortes („basketballplayer“) nicht geschlechtlich markiert ist, und zudem vom ImageNet-Team als Beispiel für eine Kategorie mit überproportional hohem Anteil Schwarzer Menschen genannt wird.[25] Damit verschiebt sich der Fokus von den eindeutig diskriminierenden Formen der Körperklassifizierung und den demografischen Missverhältnissen von ImageNet in Richtung einer generellen Betrachtung der epistemischen Wirkweise und Visualität der Datenbank, auf die nicht zuletzt eine Cheerleaderin im „basketballplayer“-*subset* hinweist. Im Folgenden untersuche ich, wie ImageNet als epistemische Struktur durch Auswahl und Zurichtung von Bild- und Textdaten Wissen über die Welt, und, insbesondere anhand des „person“-*subtrees*, Wissen über Körper erzeugt. Dazu betrachte ich drei miteinander verknüpfte Ebenen der Wissensproduktion, die in dieser Reihenfolge ImageNets Produktions-Pipeline entsprechen:[26] erstens die textuelle Ebene, auf der Konzepte

18 Die Liste der imageability scores kann nach Erstellung eines Accounts von der ImageNet-Website heruntergeladen werden: https://image-net.org/data/imageability_scores.csv.
19 Vgl. Crawford/Paglen 2021: 1110.
20 Vgl. ebd.: 1108–1111.
21 Vgl. Yang et al. 2020.
22 Vgl. ebd.: 4–5.
23 Vgl. Birhane et al. 2021: 10.
24 Vgl. Hunger 2024: 31.
25 Vgl. Yang et al. 2020: 9.
26 Vgl. ebd.: 3.

und Begriffe entstehen, zweitens die Ebene der Bilder, und drittens die Ebene der Zuordnung von Begriff und Bild.

Text

Ich scrolle durch eine Excel-Liste mit 2395 Einträgen. Es ist die Liste mit den imageability-scores, *die die ImageNet-Forscher_innen 2020 erstellt haben, links steht die Indexnummer der Kategorie, dann der Name, daneben die Beschreibung, teils ergänzt durch einen Beispielsatz: Direkt über „n09842047 – basketballplayer – an athlete who plays basketball" lautet der vorige Eintrag: „n09992837 – daughter, girl – a female human offspring", mit dem Beispielsatz: „Her daughter cared for her in her old age". Hinter der Excel-Tabelle ragt das Bild der Cheerleaderin hervor. Ob sie schon ihre Mutter pflegen muss? Ich suche in der Liste nach dem Begriff „cheerleader", und finde den Eintrag zweimal, einmal als „n09913455 – cheerleader – someone who leads the cheers by spectators at a sporting event", und einmal als „n09913593 – cheerleader – an enthusiastic and vocal supporter", mit dem Beispielsatz, „he has become a cheerleader for therapeutic cloning".*

Im Gegensatz zu neueren Datenbanken wie LAION beruht ImageNet nicht auf selbstlernenden Algorithmen, die Kategorien proaktiv und automatisiert anlegen, sondern auf einer klassisch enzyklopädischen, von den Macher_innen selbst angelegten Begriffs-Struktur, wie die meisten Bilddatenbanken dieser Generation.[27] Die Beschreibungen und Beispielsätze in den ImageNet-Begriffsdefinitionen, die im Annotationsprozess verwendet wurden, stammen nahezu alle aus WordNet: einer englischsprachigen lexikalischen Datenbank semantischer Verknüpfungen, die Mitte der 1980er Jahren mit finanzieller Unterstützung des US-amerikanischen Verteidigungsministeriums an der Princeton University am Institut für kognitive Psychologie entwickelt wurde, und wiederum auf Enzyklopädien wie dem Brown Corpus aus den 1960er Jahren beruht.[28] Teils wurden die Einträge zudem um Informationen aus Wikipedia ergänzt.[29]

Die Auswahl der Kategorien einer Datenbank soll in der Regel dazu beitragen, eine epistemische Struktur zu schaffen, die aus den vorhandenen Daten bestmöglich nutzbares Wissen generiert.[30] Die konzeptuellen Reduzierungen, die dabei im Fall von ImageNet entstehen, werden an der Taxonomie der Datenbank, die von WordNet übernommen wurde, deutlich: Unter „natural object / body / human body" finden sich die fünf Unterkategorien „male body", „female body", „juvenile body", „adult body" und „person".[31] Neben den normativen, geschlechterbinären Vorstellungen von Körpern wird an diesem Beispiel vor allem sichtbar, dass

27 Vgl. Yang et al. 2020: 10.
28 Vgl. Crawford 2021: 136.
29 Vgl. Yang et al. 2020: 3.
30 Vgl. Bechmann/Bowker 2019: 1.
31 Vgl. Crawford/Paglen 2021: 1109.

Klassifizierungssysteme in Datenbanken zwangsläufig sowohl Lücken durch Auslassungen, als auch Überbetonungen durch Mehrfachnennungen produzieren, und damit mitunter chaotische epistemische Ordnungen hervorbringen. So werden hier einerseits alle Körper, die weder auf einer Geschlechter- noch einer Altersachse eindeutig und anhand eines binären Schemas verortet werden können, in die „person"-Kategorie ausgelagert, andererseits werden Bilder *entweder* „adult body" *oder* „female body" zugeordnet, obwohl sie beide Kategorien repräsentieren.

Deutlich wird also, dass Datenbanken in der Auswahl der verwendeten Begriffe und Kategorien keineswegs eine vorgefundene analoge Welt der Dinge in ihrer ‚natürlichen' Ordnung abbilden, sondern höchstens gesellschaftliche Machtverhältnisse fortschreiben. Entgegen ImageNets Anspruch, die „entire world of objects"[32] zu erfassen, ließe sich die analoge Welt nie vollständig in einzelnen, voneinander abgegrenzten, hierarchisch angeordneten Kategorien darstellen. Dieser unauflösbare Widerspruch resultiert im Fall von ImageNet in einer *messiness*[33] der Kategorien, die sich beispielsweise auch darin manifestiert, dass dieselben oder sehr ähnliche Kategorien an verschiedenen Stellen der Datenbank existieren, wie am Beispiel der beiden „cheerleader"-Kategorien deutlich wird.

Die Wahl der Kategorien folgt allerdings nicht nur der von WordNet vorgelegten, mitunter chaotischen Taxonomie, sondern basiert auch auf pragmatischen Entscheidungen, die bei weitem nicht immer so konzeptuell intendiert sind, wie die ImageNet-Entwickler_innen (und viele ihrer Kritiker_innen) behaupten.[34] Im Gegenteil bestimmen praktische Faktoren maßgeblich mit, welche Kategorien Anwendung finden: Die Verfügbarkeit von genügend qualitativ ausreichendem Bildmaterial zum Zeitpunkt der Datenbank-Produktion und die Wahl von Begriffen, die sich möglichst eindeutig verbildlichen lassen, spielen für ImageNet letztlich eine größere Rolle als enzyklopädische Vollständigkeit,[35] so gibt es beispielsweise keine „art"-Kategorie[36], dafür aber die Kategorie „iPod"[37].

Zudem wird insbesondere am „temporal bias"[38] von ImageNet sichtbar, dass der Universalitätsanspruch der Datenbank weit hinter ihre Abhängigkeit von verfügbaren Daten und deren zeitliche Begrenztheit zurückfällt. Der aus WordNet importierte Beispielsatz zum Eintrag „daughter, girl" zeigt, was die ImageNet-Entwickler_innen in ihrer eigenen kritischen Reflektion als Problem des „stagnierenden Konzeptvokabulars" beschreiben: Sowohl Begriffe selbst, als auch ihre Definitionen, sind semiotisch instabil und verändern sich mit der Zeit.[39] So wirkt

32 Vgl. Gershgorn 2017.
33 Zum Konzept der *messiness* in digitalen Bildkulturen vgl. Moskatova et al. 2022.
34 Vgl. Smits/Wevers 2022: 338, 344.
35 Vgl. ebd.: 338.
36 Vgl. Hunger 2021: 5.
37 Vgl. Smits/Wevers 2022: 340.
38 Vgl. ebd.: 344.
39 Vgl. Yang et al. 2020: 3–4.

die Auswahl der Begriffe, Definitionen und Beispielsätze mindestens veraltet und klischeehaft, wie z. B. die Kategorie „iPod“, oder der Beispielsatz „her daughter cared for her in her old age“. Viele der Bezeichnungen und Beispielsätze verweisen als Relikte spezifischer, historisch gebundener Diskurse auf die mangelnde „historische Tiefe“[40] des ImageNet-Vokabulars, egal, ob es um Geschlechterrollen geht, oder um ‚therapeutisches Klonen‘[41], wie im Beispielsatz der Cheerleader-Kategorie. Vor allem innerhalb des „person“-*subtrees* sind (zumindest aus heutiger Sicht) aber ein Großteil der Begriffe und Beispielsätze auch schlicht extrem stigmatisierend und diskriminierend.[42]

Bild

Ich stelle die Darstellung im Finder auf „Symboldarstellung“ um, statt einer Liste sehe ich jetzt ein Raster von Bildkacheln, mit dem Zeigefinger auf der Maus scrolle ich weiter. Basketballkörbe, trostlose Hallen, Sportplätze; auf einem Bild steht unten auf dem weißen Rand „www.fotosearch.com“. Ich klicke auf das Bild daneben und vergrößere es: ein Bild von einer Spielerin im violetten Trikot, sie bindet sich die Hallensportschuhe, rutscht aus dem rechten Bildrand; ich scrolle weiter, dann erkenne ich die Frau plötzlich auf einem anderen Bild wieder, die Schuhe diesmal zu, sie wirft einen Basketball Richtung Korb. Jetzt halte ich bewusst Ausschau nach ihr, meine Augen scannen die Bildreihen nach dem violetten Trikot ab, und tatsächlich finde ich sie einige Reihen später wieder, es scheint dasselbe Spiel zu sein, sie trägt dasselbe rote Zopfgummi im Haar, die Gegnerinnen in blau-weiß, sie dribbelt. Ob sie wohl gewonnen hat?[43]

Die Bilder in ImageNet wurden mithilfe von mehreren, zu der Zeit verfügbaren Suchmaschinen gesammelt.[44] Das gesamte Bildmaterial stammt also aus dem World Wide Web, so wie es sich Anfang der 2000er Jahre konstituierte, und damit ist auch hier ein eindeutiger temporärer Beschnitt angelegt: Abgesehen von einer kleinen Anzahl historischer Bilder, die offensichtlich gescannt und digitalisiert wurden, sprechen die meisten der „basketballplayer“-Abbildungen eine zeitlich eindeutig zu verortende Sprache: Die ästhetische Sprache der Digitalkameras, die Anfang der 2000er Jahre analoge Kameras abzulösen begannen, Smartphone-Kameras vorausgingen, und eine Flut an schlecht belichteten, niedrig aufgelösten Amateurfotografien ins Internet geschwemmt haben. Dazwischen finden sich

40 Vgl. Taurino 2024: 89.

41 Öffentliche Debatten um das Klonen embryonaler Stammzellen entbrannten 2001 in den USA und Westeuropa und wurden danach noch für einige Jahre unter dem Schlagwort „therapeutisches Klonen“ diskutiert. Vgl. Nerlich/Clarke 2003: 44, 46.

42 Vgl. Yang et al. 2020: 4.

43 Ich verzichte hier aus Gründen des Datenschutzes darauf, das beschriebene Bild zu zeigen, da es sich offensichtlich um ein Privatfoto handelt und die abgebildete Person identifizierbar sein könnte. Zur Kritik am nicht-konsensuellen Umgang mit privaten Bildern durch Datenbanken wie ImageNet vgl. Prabhu/Birhane 2020: 1, 4–5.

44 Vgl. Yang et al. 2020: 3.

professionelle Fotografien, wie Stockfotos und Pressebilder, wobei der Großteil aller Bilder von *Flickr* stammt.[45] Der „temporal bias" von ImageNet zeichnet sich demnach nicht nur in der zeitlichen Verortung der Begriffsauswahl ab, sondern vor allem durch die starke zeitliche Beschränktheit der Bildauswahl.[46]

Analog zum ImageNet-Textkorpus, der aufgrund seiner WordNet-Herkunft hegemoniale Taxonomien und Epistemologien in ImageNet einschreibt, wird die Bildauswahl außerdem durch das Bias der Bildsuchmaschinen strukturiert.[47] So stammen die Bilder fast durchgehend von Seiten mit vielen Aufrufen und einem für Suchmaschinen leicht lesbaren System der Verschlagwortung, wie es z.B. bei Flickr der Fall ist. Dass im Zuge dieser Form der Bild-Akquise Serien von Bildern Teil der Datenbank wurden, die von ein und demselben Basketballspiel stammen, ist in der pragmatischen Verfügbarkeits-Logik von ImageNet also keine Überraschung. Zudem ist bemerkenswert, dass nicht jede Kategorie gleich viele Bilder enthält: Je nach Verfügbarkeit von brauchbarem Bildmaterial beheimaten Kategorien unterschiedlich viele Bilder, wobei die *subsets* „gal", „grandfather", „dad" und „chief executive officer" im „person"-*subtree* mit ca. 1600 Bildern pro Kategorie die Liste anführen[48] („basketballplayer" enthält 1228 Bilder) und damit verdeutlichen, dass sich in ImageNets Epistemologie auch patriarchale Repräsentationsverhältnisse widerspiegeln.

45 Vgl. Smits/Wevers 2022: 332.
46 Vgl. ebd.: 335.
47 Zum Bias der Suchmaschinen vgl. Noble 2018.
48 Vgl. Crawford/Paglen 2021: 1109.

Text – Bild

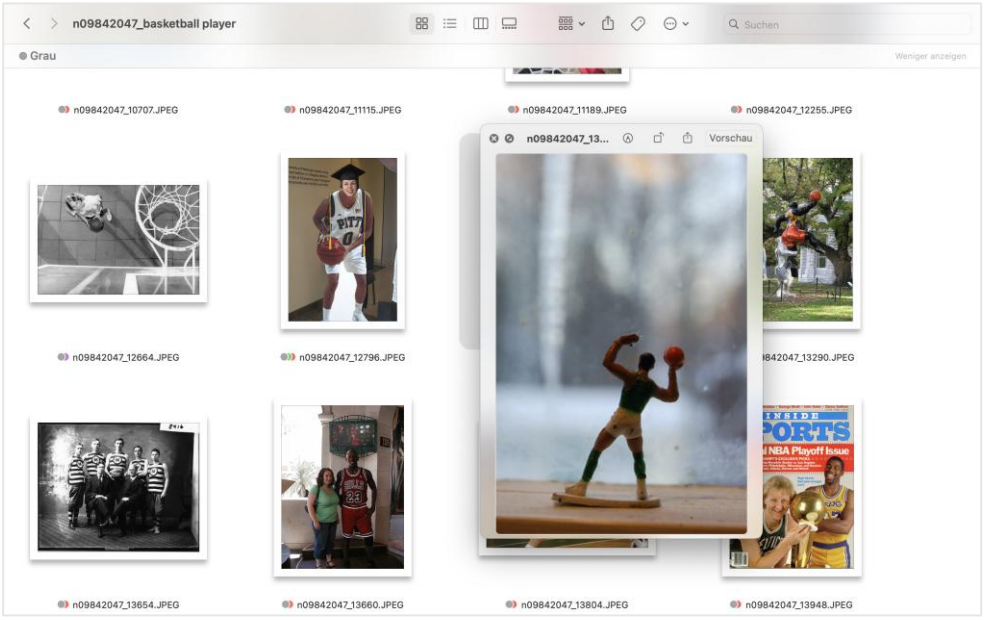

Abb. 2: Ausschnitt aus „basketballplayer" (ImageNet-21k). Eigener Screenshot

Irgendwann fange ich beim Scrollen durch die Kacheln an, per Rechtsklick diejenigen Bilder mit bunten Tags zu versehen, die mir besonders auffallen, weil sie seltsam fehlplatziert wirken: Eine Tribüne voller Pyrotechnik; als Basketballspieler verkleidete Männer mit großen Bierkrügen; ein Basketballspieler-Pappaufsteller, zwei Augen lugen durch das Pappgesicht; eine Basketballspieler-Wachsfigur, daneben zum Größenvergleich ihr treuer Fan; eine Spielfigur mit Laserschwert vor einem Himmel aus Pappkarton; mehrere Bilder von derselben bunten Skulptur, vielleicht von Niki de Saint Phalle?

An der Genese von ImageNet wird deutlich: Das selbsternannte Ziel der Projektbeteiligten war es von Anfang an, Computern das Sehen beizubringen, wobei Projektinitiatorin Fei-Fei Li nicht müde wurde, ihr Verständnis vom Sehen als berechenbaren, körperlosen Vorgang der Objekterkennung zu artikulieren.[49] Die dadurch vom menschlichen Wahrnehmungssinn losgelöste ‚Fähigkeit', Bilder zu interpretieren, verzahnt sich in ImageNets Epistemologie mit einer Vorstellung der visuellen Welt als Ansammlung klar voneinander abzugrenzender Objekte. Das Forschungsprojekt ImageNet – exemplarisch für das Feld der Computer Vision zu Beginn des Jahrtausends – konzipiert Visualität als entkörpertes, objektorientiertes Phänomen, das aus organischer, analoger Materie herausgelöst und, transportiert durch die Kameralinse, in den Digitalcomputer hineinverlagert werden kann.[50] Im Zuge dieser Automatisierung der Visualität wird Sehen enthistorisiert und

49 Vgl. Denton et al. 2021: 6–7.
50 Vgl. Malevé/Sluis 2023.

naturalisiert.[51] Dabei zeigen zahlreiche Beispiele aus ImageNet, dass der vermeintlich universelle und neutrale „way of seeing“[52] der Datenbank von hegemonialen, *weißen,* westlichen Vorstellungen der Welt geprägt ist. Besonders erkennbar wird dies an der Kategorie „black person“, und zwar nicht nur aufgrund der kolonial-rassistischen Logik dieser Kategorisierung, sondern auch anhand der höchst rassistischen Begriffe, die als ergänzende Kategorienbezeichnung verwendet werden,[53] sowie anhand der Tatsache, dass 79 der 1286 Bilder des *subsets* nicht Schwarze Menschen, sondern *weiße* Menschen, die Blackfacing praktizieren, zeigen.[54] Aber auch zunächst unscheinbare Kategorien belegen, dass die visuelle Erfassung der Welt von ImageNet nicht universell, sondern historisch und partikular ist.[55] ImageNets „instrumenteller Realismus“[56] und statistischer Positivismus verkörpern direkt und indirekt semiotische Praktiken und Kulturtechniken, die hegemonialen und normativen Blickregimen entsprechen.[57]

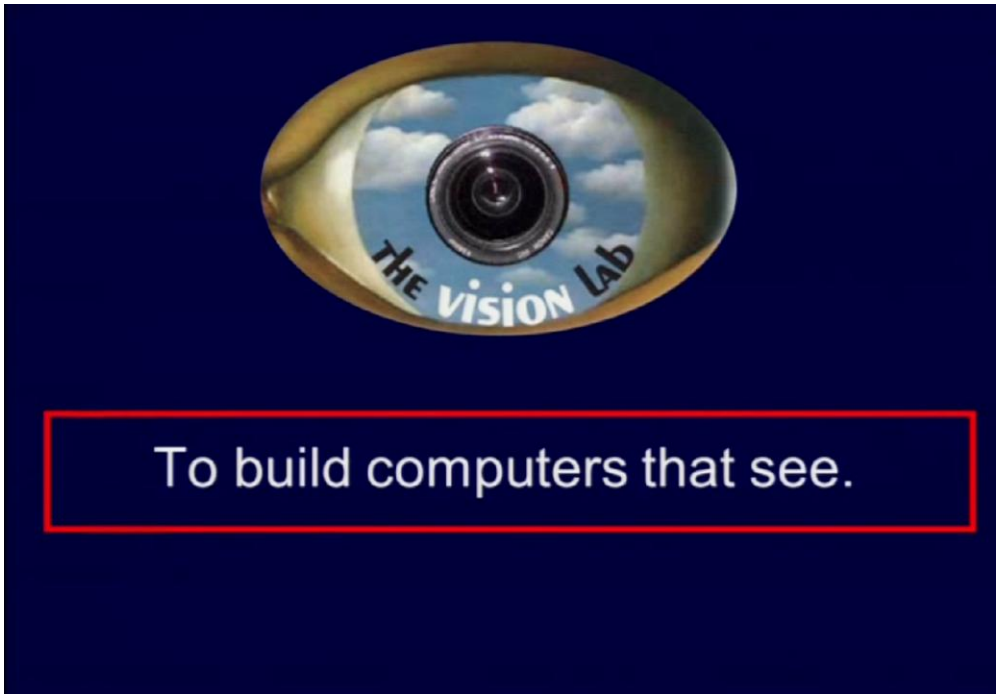

Abb. 3: Slide aus einer ImageNet-Präsentation von Fei-Fei Li, inklusive Logo des Stanford Vision Labs, 2012.

51 Vgl. ebd.
52 Vgl. Berger 2008.
53 Die Original-Bezeichnung der Kategorie beinhaltet zusätzlich zu „Black person“ noch weitere Begriffe, die ich hier aufgrund ihrer rassistischen Aufladungen nicht nenne.
54 Vgl. Monea 2019: 198.
55 Vgl. Malevé/Sluis 2023.
56 Vgl. ebd.
57 Vgl. ebd.

Die Objekt-Bezogenheit von ImageNets Visualität offenbart sich weiterhin darin, dass in der Datenbank ausschließlich Nomen als Kategorien verwendet werden, in der Annahme, dass Nomen in der Regel Begriffe seien, die sich zuverlässig und unkompliziert bildlich darstellen ließen.[58] Blicken wir nochmal zurück zu den beiden „cheerleader"-*subsets*, wird jedoch schnell deutlich, dass nicht alle Nomen gleichermaßen abbildbar sind: Während das erste „cheerleader"-Konzept sich, zumindest in einer klischeehaften Zuspitzung, durch visuelle Merkmale wie Kleidung, Pompons, Pose, und Hintergrund relativ klar bebildern lässt, bleibt bei der zweiten „cheerleader"-Kategorie („an enthusiastic and vocal supporter") rätselhaft, an welchen visuellen Markern sich eine generell unterstützende geistige Haltung festmachen ließe. Die Auswahl der Nomen, die als Begriffe in ImageNet Einzug erhalten haben, bezeugt erneut, dass praktische Entscheidungen Vorrang vor konzeptuellen Überlegungen haben: Begriffe, die sich möglichst objekthaft darstellen lassen und möglichst eindeutig textuell beschrieben werden können, dominieren die Welt von ImageNet. Das führt nicht zuletzt dazu, dass die Bilddatenbank in vielerlei Hinsicht wie ein Warenkatalog anmutet, und ImageNet-Experten Gabriel Pereira und Bruno Moreschi zufolge eine konsumorientierte Darstellung der Welt verkörpert.[59] Zumindest demonstrieren die eingangs beschriebenen aktuellen Anwendungsbeispiele von Google und Apple neben weiteren Computer-Vision-Großprojekten von Meta und Co., dass Bilddatenbanken auch Teil von Wertschöpfungsketten des digitalen Kapitalismus sind.

ImageNet postuliert, dass die eindeutige Zuordnung von Bild und Begriff einer natürlichen, vorgefundenen Ordnung der Welt entspreche, in der die Beziehung zwischen Bild und Konzept eindeutig, ablesbar und unmittelbar sei. Dass diese Setzungen in aktuellen Diskursen der Computer Vision weiter Bestand haben, zeigt beispielsweise der im Bereich *Machine Learning* derzeit hochgehandelte Begriff der „multimodality". Der unproblematische Switch zwischen Bild- und Textmodalität, der vorausgesetzt wird, um sogenannte „large-scale visio-linguistic models" (z. B. OpenAIs CLIP-Modell) zu konzipieren, basiert auf denselben epistemischen Annahmen wie ImageNet: dass sowohl Bilder und Begriffe als auch die Welt und ihre fotografische Abbildung sich mühelos und eindeutig ineinander übersetzen ließen.

5. Epistemologie und Visualität von ImageNet

Zusammengefasst lässt sich folgendes zur Visualität und Epistemologie von ImageNet als Pionierprojekt der Computer Vision feststellen: Das Forschungsprojekt und die daraus entstandene Bilddatenbank verfolg(t)en das Ziel, menschliches Sehen algorithmisch so überzeugend zu imitieren, dass die Ergebnisse

58 Vgl. Crawford/Paglen 2021: 1109.
59 Vgl. Pereira/Moreschi 2021: 1209.

im Rahmen einer stark schematisierten Erfassung der Welt menschliche Akteur_innen von der *Sehstärke* algorithmischer Modelle überzeugen würden. Dazu konzipiert ImageNet Visualität als entkörpertes, universelles, ahistorisches Erfassen einer Welt, die aus einer Akkumulation von zueinander abgegrenzten, eindeutig erfassbaren und benennbaren Objekten besteht. Indem menschliches Sehen als natürliche und objektive Erfassung der Welt und die Computer Vision als ihr technisches Äquivalent formuliert werden, erhebt ImageNet Anspruch auf eine allgemeine Gültigkeit seiner Wissensordnung. Auch jenseits der höchst diskriminierenden Zuschreibungen, die insbesondere bei der Klassifizierung von Menschen aus dieser Ordnung erwachsen, wird anhand einer eher trivialen Kategorie wie „basketballplayer" deutlich, dass ImageNet auf einem Dispositiv der Datafizierung basiert, welches auf die Reduktion von Komplexität, auf starke Stereotypisierung und auf die Objekt-Machung der Welt angewiesen ist.

Zudem ist die epistemische Struktur von ImageNet, entgegen der Idealvorstellung ihrer Entwickler_innen, weitaus weniger intentional und allumfassend als angenommen. Durch den „temporal bias" des Vokabulars und der Bildsammlung, durch die (Nicht-)Verfügbarkeit von bestimmten Bildern und nicht zuletzt durch die subjektiven, menschlichen Text-Bild-Zuordnungen von 49.000 Clickworker_innen ist die Wissensordnung in ihrem Anspruch auf universelle Gültigkeit und lückenlose Erfassung stark eingeschränkt und deutlich *messier* als vermutet. Dabei verschwimmen nicht nur maschinelle und menschliche Informationsverarbeitung, sondern auch fotografische und algorithmische Visualitäten, und ImageNet muss letztlich als mehrschichtige, teilautomatisierte und funktionalistische Infrastruktur des Wissens betrachtet werden. Trotz alledem erfüllt ImageNet in seiner Rolle als Pionier-Datenbank für das Feld der Computer Vision eine normierende Funktion für die visuelle, computerbasierte Erfassung der Welt. Dass diese visuelle Erfassung fester Bestandteil des alltäglichen Mediengebrauchs geworden ist, lässt erkennen, dass Bilddatenbanken wie ImageNet mitsamt ihrer mitunter *messy* Epistemologien als medientechnologische Akteure Sehgewohnheiten und Blickregime nicht nur reproduzieren, sondern längst entscheidend mitkonfigurieren. Dabei bestehen neuere Datenbanken wie LAION aus Bildsammlungen, die mindestens dreihundertmal so groß sind wie *ImageNet,* und überlassen die Zuordnung von Text und Bild nahezu gänzlich vortrainierten Modellen. Die Auseinandersetzung mit Bilddatenbanken der algorithmischen Klassifizierung wird also noch vielschichtiger und komplexer – nichtsdestotrotz bleibt sie angesichts der wachsenden Rolle von Bilddatenbanken und algorithmischen Modellen für visuelle, digitale Medienkulturen umso relevanter.

Literaturverzeichnis

Bechmann, Anja/Bowker, Geoffrey (2019): „Unsupervised by any other name. Hidden layers of knowledge production in artificial intelligence on social media". In: *Big Data & Society* 6.1, S. 1–11.

Bell, Sean et al. (2020): „GrokNet. Unified Computer Vision Model Trunk and Embeddings For Commerce". In: *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, S. 2608–2616.

Berger, John (2008 [1972]): *Ways of Seeing.* London: Penguin.

Birhane, Abeba/Prabhu, Vinay Uday/Kahembwe, Emmanuel (2021): „Multimodal datasets. misogyny, pornography, and malignant stereotypes". *arXiv.* https://arxiv.org/abs/2110.01963 (25.09.2024).

Blaschke, Estelle (2016): *Banking on images. From the Bettmann Archive to Corbis.* Leipzig: Spector books.

Crawford, Kate (2021): *Atlas of AI. Power, Politics, and the Planetary Costs of Artificial Intelligence.* New Haven: Yale University Press.

Crawford, Kate: „9 Ways To See A Dataset. What's at stake in examining datasets?". *Knowing Machines.* https://knowingmachines.org/publications/9-ways-to-see/essays/9-ways-to-see-a-dataset (30.09.2024).

Crawford, Kate/Paglen, Trevor (2021): „Excavating AI. The Politics of Images in Machine Learning Training Sets". In: *AI & SOCIETY* 36.4, S. 1105–1116. Denton, Emily et al. (2021): „On the genealogy of machine learning datasets. A critical history of ImageNet". In: *Big Data & Society* 8.2, S. 1–14.

Dulhanty, Chris/Wong, Alexander (2019): „Auditing ImageNet. Towards a Model-driven Framework for Annotating Demographic Attributes of Large-Scale Image Datasets". *arXiv.* http://arxiv.org/abs/1905.01347 (08.08.2024).

Fei-Fei, Li/Deng, Jia (2017): „ImageNet. Where have we been? Where are we going?". CVPR. Beyond ImageNet. Large Scale Visual Recognition Challenge workshop. https://image-net.org/static_files/files/imagenet_ilsvrc2017_v1.0.pdf (30.09.2024).

Gershgorn, Dave (2017): „The data that transformed AI research—and possibly the world". *Quartz*. https://qz.com/1034972/the-data-that-changed-the-direction-of-ai-research-and-possibly-the-world (01.10.2024).

Hanna, Alex et al. (2020): „Lines of Sight". In: *Logic Magazine* 12: Commons, https://logicmag.io/commons/lines-of-sight/ (25.09.2024).

Hunger, Francis (2021). „Working Paper 2. ‚Why so many windows?' – Wie die Bilddatensammlung ImageNet die automatisierte Bilderkennung historischer Bilder beeinflusst." *Zenodo*. https://zenodo.org/record/4742621 (28.01.2025).

Hunger, Francis (2024): „Im Gespräch mit Gabriel Pereira über ‚Wie die Bilddatensammlung ImageNet Wirklichkeit (re-)konstruiert'". In: Arns, Inke et al. (Hrsg.): *Training the Archive Ludwig Forum für International Kunst, Aachen*. Köln: Walther König, S. 30–31.

Kirillov, Alexander et al. (2023): „Segment Anything". *arXiv.* http://arxiv.org/abs/2304.02643 (28.09.2024).

Light, Ben/Burgess, Jean/Duguay, Stefanie (2018): „The Walkthrough Method. An Approach to the Study of Apps". In: *New Media & Society* 20.3, S. 881–900.

Malevé, Nicolas/Sluis, Katrina (2023): „The Photographic Pipeline of Machine Vision; or, Machine Vision's Latent Photographic Theory". In: *Critical AI* 1.1–2, https://read.dukeupress.edu/critical-ai/article/doi/10.1215/2834703X-10734066/382465/The-Photographic-Pipeline-of-Machine-Vision-or (24.08.2024).

Monea, Alexander (2019): „Race and Computer Vision". In: Sudmann, Andreas (Hrsg.): *The democratization of artificial intelligence. Net politics in the era of learning algorithms*. Bielefeld: transcript, S. 189–208.

Moskatova, Olga/Mücke, Laura Katharina/Tedjasukmana, Chris (2022). „Editorial. Messy Images – Unordnungen vernetzter Bilder" In: *montage av* 31.1, S. 5–18. https://montage-av.de/31-1-2022/ (28.01.2025).

Nerlich, Brigitte/Clarke, David D. (2003). „Anatomy of a Media Event. How Arguments Clashed in the 2001 Human Cloning Debate". In: *New Genetics and Society*, *22*.1, S. 43–59. https://doi.org/10.1080/1463677032000069709 (23.01.2025).

Noble, Safiya Umoja (2018): *Algorithms of Oppression. How Search Engines Reinforce Racism.* New York: New York University Press.

Offert, Fabian/Bell, Peter (2021): „Perceptual Bias and Technical Metapictures. Critical Machine Vision as a Humanities Challenge". In: *AI & SOCIETY* 36.4, S. 1133–1144.

O.V. (2021): „On-device Panoptic Segmentation for Camera Using Transformers". *Apple Machine Learning Research.* https://machinelearning.apple.com/research/panoptic-segmentation (30.09.2024).

Pereira, Gabriel/Moreschi, Bruno (2021): „Artificial Intelligence and institutional Critique 2.0. Unexpected Ways of Seeing with Computer Vision". In: *AI & SOCIETY* 36.4, S. 1201–1223.

Prabhu, Vinay Uday/Birhane, Abeba (2020): „Large Image Datasets. A Pyrrhic Win for Computer Vision?". *arXiv*. http://arxiv.org/abs/2006.16923 (02.09.2024).

Russakovsky, Olga et al. (2015): „ImageNet Large Scale Visual Recognition Challenge". *arXiv*. http://arxiv.org/abs/1409.0575 (23.07.2024).

Smits, Thomas/Wevers, Melvin (2022): „The Agency of Computer Vision Models as Optical Instruments". In: *Visual Communication* 21.2, S. 329–349.

Taurino, Giulia (2024): „Auf der Suche nach Grenzobjekten: Ein taxonomiebasierter Ansatz zur algorithmischen Co-Kuratierung in Archivsammlungen". In: Arns, Inke et al. (Hrsg.): *Training the Archive Ludwig Forum für International Kunst, Aachen*. Köln: Walther König, S. 83–95.

Walker-Todd, Alex (2023): „How to use Visual Look Up to find a food recipe in iOS 17". *iMore*. https://www.imore.com/ios/ios-17/how-to-use-visual-look-up-to-find-a-food-recipe-in-ios-17, (30.09.2024).

Yang, Kaiyu et al. (2020): „Towards Fairer Datasets. Filtering and Balancing the Distribution of the People Subtree in the ImageNet Hierarchy". In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, S. 547–558.

Medienverzeichnis

„Introducing a new way to search | Circle to Search", Google, *YouTube,* (17.01.2024), https://www.youtube.com/watch?v=WdbeqSQjZI8 (23.01.2025).

Abbildungsverzeichnis