ZfdG

Zeitschrift für digitale Geisteswissenschaften

Fachartikel aus:
Hartmut Beyer / Thomas Mandl (Hg.): Bildähnlichkeit und Bildsuche: Geistes- und informationswissenschaftliche Zugänge zu historischem Material (= Zeitschrift für digitale Geisteswissenschaften / Sonderbände, 8). 2026. DOI: 10.17175/sb008

Titel:
Ähnlichkeiten erklären. Explainable Artificial Intelligence für die multimodale Bildsuche und -analyse in der Kunstgeschichte

Autor*in:
Julian Stalter

Kontakt: julian.stalter@kunstgeschichte.uni-muenchen.de
Institution: Ludwig-Maximilians-Universität München
GND: 1269024078 ORCID: 0000-0003-1149-1688
Contribution (CRediT): Conceptualization | Data curation | Funding acquisition | Investigation | Methodology | Project administration | Writing – original draft

Autor*in:
Matthias Springstein

Kontakt: matthias.springstein@tib.eu
Institution: Technische Informationsbibliothek (TIB) Hannover
GND: 1374806765 ORCID: 0000-0002-6509-8534
Contribution (CRediT): Investigation | Methodology | Software | Visualization | Writing – original draft

Autor*in:
Stefanie Schneider

Kontakt: stefanie.schneider@itg.uni-muenchen.de
Institution: Ludwig-Maximilians-Universität München
GND: 1220379301 ORCID: 0000-0003-4915-6949
Contribution (CRediT): Conceptualization | Data curation | Funding acquisition | Methodology | Software | Supervision | Visualization | Writing – review & editing

DOI des Beitrags:
10.17175/sb008_003

Nachweis im OPAC der Herzog August Bibliothek:
1933638664

Erstveröffentlichung:
21.05.2026

Letzte Überprüfung aller Verweise:
23.01.2026

Format:
PDF ohne Paginierung, Lesefassung

GND-Verschlagwortung:
Kunstgeschichte | Erklärbare künstliche Intelligenz | Ähnlichkeit | Bildanalyse

Empfohlene Zitierweise:
Julian Stalter / Matthias Springstein / Stefanie Schneider: Ähnlichkeiten erklären. Explainable Artificial Intelligence für die multimodale Bildsuche und -analyse in der Kunstgeschichte. In: Hartmut Beyer / Thomas Mandl (Hg.): Bildähnlichkeit und Bildsuche: Geistes- und informationswissenschaftliche Zugänge zu historischem Material (= Zeitschrift für digitale Geisteswissenschaften / Sonderbände, 8). Wolfenbüttel 2026. 21.05.2026. HTML / XML / PDF. DOI: 10.17175/sb008_003

Julian Stalter / Matthias Springstein / Stefanie Schneider

Ähnlichkeiten erklären. Explainable Artificial Intelligence für die multimodale Bildsuche und -analyse in der Kunstgeschichte

Abstract

Der Beitrag untersucht das Konzept der Ähnlichkeit im Rahmen der *Explainable Artificial Intelligence (XAI)* in Modellen des maschinellen Lernens für die kunsthistorische Bildsuche und -analyse. Anhand des bildorientierten Forschungswerkzeugs *iART* wird gezeigt, welche Parameter in künstlichen neuronalen Netzen die Ergebnisse der Bildsuche bestimmen und wie. Dabei werden zwei entscheidende Faktoren hervorgehoben: die Architektur der neuronalen Netze und die verwendeten Trainingsdaten. Durch die Anwendung von Methoden der *XAI* können diese Prozesse transparenter gemacht werden, um ein kritisches Verständnis ihrer Anwendung in der kunsthistorischen Forschung zu fördern. Dieser interdisziplinäre Ansatz unterstreicht die Notwendigkeit von Transparenz und methodischer Reflexion beim Einsatz von Technologien des maschinellen Lernens in den Geisteswissenschaften.

This article examines the concept of similarity in the context of *Explainable Artificial Intelligence (XAI)* in machine learning models for art-historical image search and analysis. Using the image-oriented research tool *iART*, it is shown which parameters in artificial neural networks determine the results of image search and how. Two decisive factors are highlighted: the architecture of the neural networks and the training data used. By applying *XAI* methods, these processes can be made more transparent in order to promote a critical understanding of their application in art-historical research. This interdisciplinary approach emphasizes the necessity for transparency and methodological reflection when using machine learning technologies in the humanities.

1. Die Rolle(n) der Ähnlichkeit in algorithmischen Prozessen der Bildsuche und -analyse

[1] Geben Nutzer*innen der Bildersuchmaschine *iART* [1] den Begriff ›creation‹[2] in den Suchschlitz ein, erscheint zunächst ein *Studieblad Met Vier Handen* (1710–1777) eines unbekannten Künstlers (Abbildung 1). An dritter Stelle folgt wieder eine Studie von Händen, diesmal die eines Armbrustschützen. Sie halten einen Pfeil, dessen Spitze jedoch nicht zu erkennen ist, sondern eher an einen Stift oder Pinsel erinnert. Für Kunsthistoriker*innen ein interessantes Ergebnis: Beziehen sich diese Ergebnisse vielleicht auf die schöpferisch tätige Hand Gottes oder sogar der Künstler*innen, die aus Lehm, Erde oder mit dem Pinsel die Welt erschaffen und ihr Form geben? Greift hier die Suchmaschine, die auf Basis eines künstlichen neuronalen Netzes die Ergebnisse selektiert, kunsthistorische Topoi auf und liefert kontextuell anspruchsvolle Suchergebnisse? Erscheint »der Künstler als Griffel der Gottheit«,[3] wie es Ernst Kris und Otto Kurz formuliert haben? Oder liegt eine andere, prosaischere Erklärung näher?

[1] Springstein et al. 2021; Schneider et al. 2022. *iART*, kurz für *Interaktives Analyse- und Retrieval-Tool*, wurde im Rahmen eines von der DFG geförderten Projekts von 2019 bis 2021 entwickelt vom Lehrstuhl für Mittlere und Neuere Kunstgeschichte der Ludwig-Maximilians-Universität München, der Forschungsgruppe Visual Analytics der Technischen Informationsbibliothek (TIB) Hannover und der Fachgruppe Intelligente Systeme und Maschinelles Lernen des Heinz Nixdorf Instituts der Universität Paderborn. Es handelt sich um eine für kunsthistorische Bildinhalte optimierte Suchmaschine, die verschiedene *Deep-Learning*-Methoden zur automatischen Klassifizierung und Ähnlichkeitsbestimmung von Bildern einsetzt.

[2] Da das in *iART* verwendete Modell größtenteils mit Daten in englischer Sprache trainiert wurde, wird ebenso die Suche auf Englisch durchgeführt. Die deutsche Entsprechung der Suchanfrage wäre ›Erschaffung‹ oder ›Schöpfung‹.

[3] Kris / Kurz 2010, S. 74.

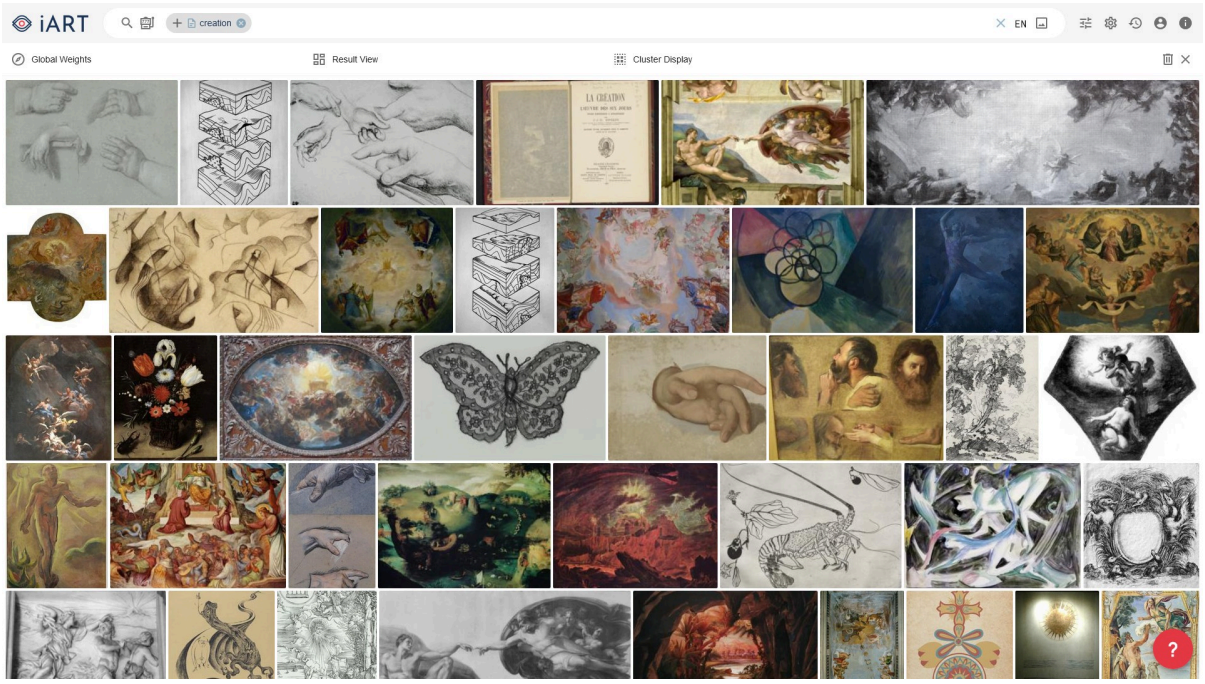

Abb. 1: Suchergebnisse im Forschungswerkzeug *iART* für den Begriff ›creation‹. [Bildquelle: *iART*, Suchbegriff creation]

Auch in Michelangelos berühmter Darstellung der Erschaffung Adams (*The Creation of Adam*, 1508–1512), die auf Platz fünf der Suchergebnisse rangiert, stehen zwei Hände im Mittelpunkt: Gott schwebt auf einer Wolke heran und richtet seinen Zeigefinger auf Adam, der wiederum seine linke Hand, auf sein Knie gestützt, Gott entgegenstreckt. Könnten die vielen Hände in den Suchergebnissen hier auf eine Ähnlichkeit hindeuten und ›creation‹ direkt mit Michelangelos Gemälde in Verbindung bringen? Eine Möglichkeit, diese Hypothese zu überprüfen, besteht darin, mit der Eingabe von ›creation of adam‹ gezielt nach Michelangelos *Erschaffung Adams* zu suchen. Und tatsächlich finden sich in den Suchergebnissen weitere Studien von Händen; außerdem tauchen Werke wie das Relief *Die Vertreibung aus dem Paradies* (1649) eines unbekannten Künstlers auf, in dem der Engel eine ähnliche Zeigegeste aufweist wie Michelangelos Adam (Abbildung 2). Diese Ergebnisse deuten darauf hin, dass gerade dieser Aspekt – die Darstellung der Hände – mit dem Suchbegriff ›creation‹ und insbesondere mit Michelangelos Werk assoziiert wird. Es stellt sich jedoch die Frage, warum der Algorithmus gerade dieses Merkmal hervorhebt und nicht etwa die liegende Figur Adams, den sich auf der Wolke nähernden Gott oder andere in der Kunstgeschichte vertretene Schöpfungsszenen. Welche Faktoren beeinflussen also die Auswahl der Bilder, und welche Rolle spielt Ähnlichkeit in diesem algorithmischen Auswahlprozess? [2]

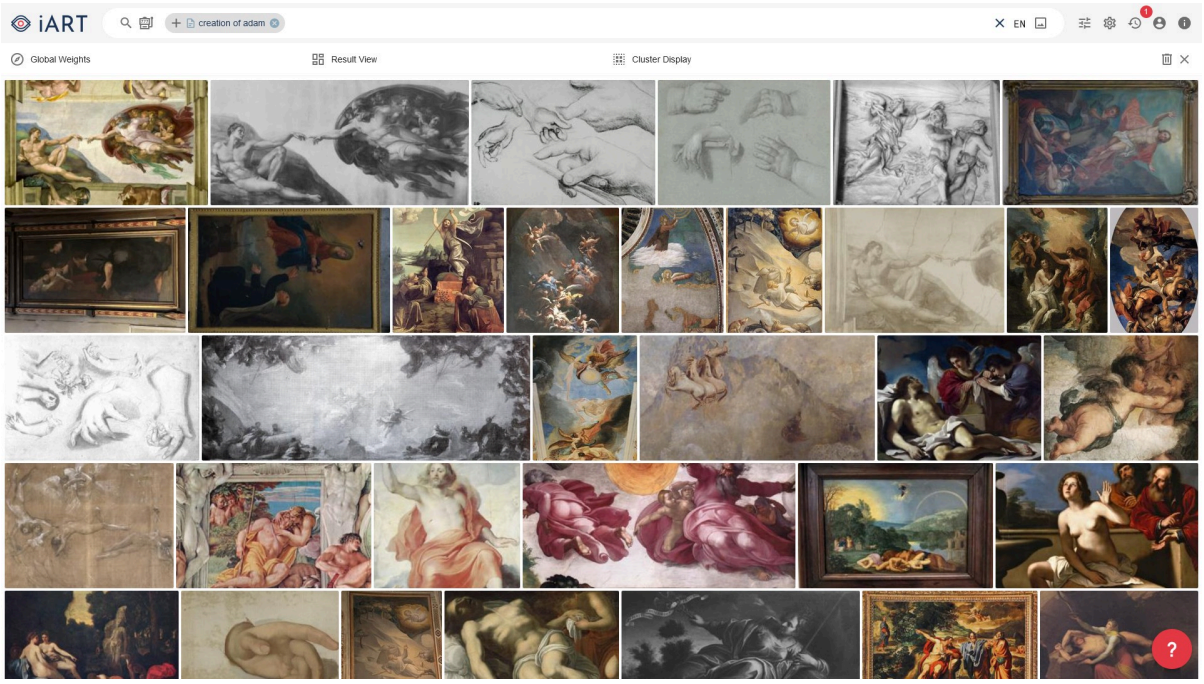

Abb. 2: Suchergebnisse im Forschungswerkzeug *iART* für den Begriff ›creation of adam‹. [Bildquelle: *iART*, Suchbegriff creation of adam]

[3] Im Folgenden untersuchen wir zwei Faktoren, die die Suchergebnisse in bildbasierten Forschungswerkzeugen wie *iART* beeinflussen: (1) die Architektur neuronaler Netze und (2) die Daten, die in sie zum Training eingespeist werden. Dabei konzentrieren wir uns auf die *Dimensionen* des Konzepts der Ähnlichkeit, und wie diese die algorithmisch generierten Ergebnisse bestimmen. Eingebettet in den Rahmen der *Explainable Artificial Intelligence (XAI)* zielt der Beitrag darauf ab, die Entscheidungsprozesse künstlicher neuronaler Netze transparent zu machen,[4] um komplexe Mechanismen in einer für Forscher*innen verständlichen Sprache abzubilden.[5] Im Kontext von *iART* – und ähnlichen geisteswissenschaftlichen Forschungswerkzeugen, die maschinelle Lernverfahren integrieren[6] – können im Sinne des *Tool Criticism* so auch Spezifika der jeweiligen Anwendung thematisiert werden.[7] Auf diese Weise soll eine vertiefte Diskussion über den Gebrauch und die Auswirkungen von maschinellen Lernverfahren in der Kunstgeschichte angeregt werden. Durch die konstante Bezugnahme auf die eingangs vorgestellte Suche nach ›creation‹ wird die methodologische Herangehensweise illustriert und die Diskussion entlang eines spezifischen Anwendungsszenarios entfaltet. Dabei beleuchtet der Beitrag sowohl die mathematischen Grundlagen der Ähnlichkeitsbestimmung als auch deren Verortung im bildwissenschaftlichen Kontext. Dieser Bestandsaufnahme folgt ein Ausblick auf Methoden zur Verbesserung der Transparenz und Nachvollziehbarkeit von neuronalen Netzen, die für die Vorhersage der Netze entscheidende Bereiche visualisieren um die sogenannte *Black Box* der Modelle zu ›entmystifizieren‹.

[4] Diese Arbeiten sind Teil des Projekts Reflexionsbasierte künstliche Intelligenz in der Kunstgeschichte, das im Rahmen des DFG-Schwerpunktprogramms »Das digitale Bild« seit 2022 gefördert wird. Die Kooperation zwischen der Technischen Informationsbibliothek (TIB) der Leibniz Universität Hannover und dem Lehrstuhl für Mittlere und Neuere Kunstgeschichte der Ludwig-Maximilians-Universität München befasst sich interdisziplinär mit den Herausforderungen, die sich aus dem Einsatz künstlicher neuronaler Netze in der kunsthistorischen Bildsuche und -analyse ergeben. Ziel des Projektes ist es, die spezifischen Anforderungen

[4] Vgl. Molnar 2020.
[5] Vgl. Doshi-Velez / Kim, S. 2.
[6] Vgl. Ohm et al. 2023; Offert / Bell 2023; Ufer et al. 2021.
[7] Vgl. Herrmann et al. 2023.

der Kunstgeschichte an den reflexiven Einsatz künstlicher neuronaler Netze im Forschungsprozess zu untersuchen. Dies beinhaltet die Erstellung eines kunsthistorischen Textkorpus, das mittels eines automatisierten Extraktionsprozesses in einen Wissensgraphen überführt wird; dieser Wissensgraph dient als Grundlage für das Training domänenspezifischer Modelle. Durch den Einsatz von Merkmalsvisualisierungen und sogenannten ›Szenengraphen‹ wird zudem eine verbesserte Nachvollziehbarkeit der Klassifikations- und Retrieval-Entscheidungen künstlicher neuronaler Netze angestrebt, um die Ergebnisfindung transparenter zu gestalten.[8] Der vorliegende Beitrag schließt an diesen letzten Aspekt an. Unserem interdisziplinären Ansatz folgend, möchten wir zunächst den Begriff der Ähnlichkeit in den jeweiligen Domänen verorten.

2. Vorgehensweisen zur bildorientierten Ähnlichkeitsbestimmung

[5] Der Begriff der Ähnlichkeit ist, wie vielfach diskutiert und kritisiert wurde, allgegenwärtig und doch nicht greifbar. Ähnlichkeit sei keine inhärente Eigenschaft einer Entität, sondern ein von außen bestimmtes Attribut, führt Nelson Goodman aus, und so »relative, variable, [and] culture-dependent«.[9] Ihre Bestimmung erfordere die Interpretation einer beobachtenden Instanz und sei daher von Natur aus von menschlichem Ermessen geprägt, so Michel Foucault.[10]

2.1 Das Konzept der Ähnlichkeit aus kunsthistorischer Perspektive

[6] In der Kunstgeschichte wird Ähnlichkeit methodisch unter den Begriffen des ›vergleichenden Sehens‹ und des ›vergleichenden Blicks‹ diskutiert. Dabei werden mindestens zwei Objekte anhand verschiedener, zu definierender Merkmale miteinander in Beziehung gesetzt und verglichen. Disziplingeschichtlich geht dieser Vorgang bis ins 19. Jahrhundert auf Anton Springer zurück, der die Kunstgeschichte als Formwissenschaft verstand: Ohne auf vorgefasste Theorien zurückzugreifen, wollte er über Vergleiche zu grundlegenden Individual-, Orts- und Zeitstilen gelangen.[11]

[7] Diese vergleichende Betrachtung differenzierte sich später in zwei Varianten aus: in die Analyse der Verschiedenheit der Erscheinungen und in der Analyse von Gemeinsamkeiten, Analogien und Übergängen.[12] Die Verschiedenheit der Erscheinungen wird methodisch bei Heinrich Wölfflin angewandt, der mit »Kontrasteindrücken«[13] die Unterschiede in Gegensatzpaaren zu erfassen sucht. Analogien hingegen verfolgt Aby Warburgs Bilderatlas *Mnemosyne*. Dieser versammelt fotografische Reproduktionen auf Tafeln, um einen Überblick über die Summe der Bilder zu geben, die durch die Übernahme von »Pathosformeln« in einem Verwandtschaftsverhältnis zu einander stehen.[14] Pathosformeln, etwa als prägnante Ausdruckshaltungen, zeugen für Warburg von einer jahrtausendelangen Tradierung bestimmter Gesten.[15] Die Tafeln stehen für Warburg damit in einem Ähnlichkeitsverhältnis und können vergleichend betrachtet werden. Diese und andere Formen der Ähnlichkeitsbestimmung bieten letztlich, wie George Kubler es formuliert, eine Möglichkeit, das Universum zu verstehen, indem wir es durch Identitäten – wie Klassen, Typen und Kategorien – vereinfachen. Auf diese Weise kann die unendliche Folge nichtidentischer Ereignisse in ein endliches System von Ähnlichkeiten überführt werden.[16] Die so gebildeten Identitäten dienen auch in der

[8] Vgl. Stalter et al. 2024.
[9] Goodman 1972, S. 437.
[10] Foucault 1966, S. 41.
[11] Vgl. Pfisterer 2020, S. 154.
[12] Vgl. Geimer 2010, S. 47.
[13] Wölfflin 1915.
[14] Vgl. Thürlemann 2013, S. 109.
[15] Vgl. Ubl 2011, S. 430.
[16] Kubler 2008, S. 61.

Kunstgeschichte als Grundlage für die vergleichende Betrachtung: Die Fähigkeit zur Kategorienbildung setzt, nach Felix Thürlemann, voraus, »das Gemeinsame und das jeweils Eigene der […] zusammengestellten Bilder – sei es auf inhaltlich-ikonografischer, sei es auf formal-stilistischer Ebene, begrifflich zu fassen«.[17]

2.2 Das Konzept der Ähnlichkeit aus informatischer Perspektive

[8] Die algorithmische Mustererkennung hingegen beruht darauf, dass Ähnlichkeiten als Nähe- und Abstandsverhältnisse in einem Vergleichsraum modelliert werden – die Quantifizierung von Ähnlichkeit schafft erst deren Operationalisierung. In diesem Raum stellt jedes Bild einen Vektor in einem hochdimensionalen Koordinatenraum dar. Die relative Lage dieser Vektoren zueinander offenbart ihre relationalen Zusammenhänge, die durch den Vergleich der extrahierten numerischen Merkmale konkretisiert werden. Ziel ist es, die Ähnlichkeit als Wert zwischen 0 und 1 zu bestimmen, wobei Metriken wie der euklidische Abstand oder die Kosinusähnlichkeit verwendet werden (Abbildung 3).

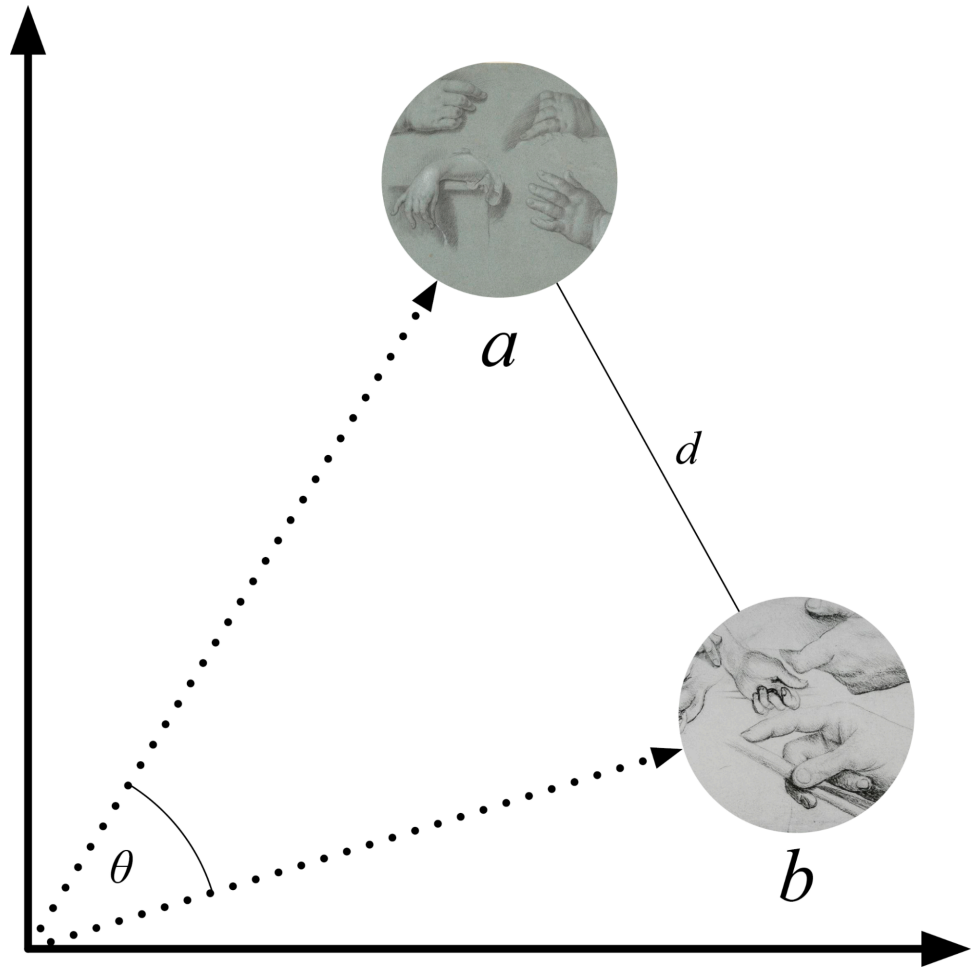

Abb. 3: Visualisierung des euklidischen Abstands d und der Kosinusähnlichkeit *cos*(θ) zweier Merkmalsvektoren a und b, die die Kunstwerke *Studieblad Met Vier Handen* (1710–1777) und *Studie für die Hände eines Armbrustschützen* (1512–1516) repräsentieren. [Grafik: Stefanie Schneider / Matthias Springstein 2024]

[9] In *Convolutional Neural Networks (CNNs)* wird ein bildrepräsentierender Merkmalsvektor zum Beispiel erzeugt, indem die Pixel des Eingabebildes sequentiell durch mehrere Schichten von Neuronen geleitet werden. Diese Neuronen sind darauf ausgelegt, Bildmerkmale zu identifizieren, von einfachen Farbgradienten in den ersten

[17] Thürlemann 2005, S. 167.

Schichten bis hin zu semantisch interpretierbaren Mustern und Objektformen in höheren Schichten. Jede dieser Schichten besteht aus einer Reihe von Filtern, die zunächst willkürlich konfiguriert sind, aber während des Trainings zunehmend auf die Erkennung bestimmter Merkmale eingestellt werden. Die *Outputs* dieser Filter werden auf *Feature Maps* abgebildet, die die Lokalisation und Relevanz der erkannten Merkmale im Bild anzeigen. Die Feature Maps der tieferen Schichten schließlich werden zu einem Merkmalsvektor, dem sogenannten *Embedding*, zusammengeführt, der das Bild in einem hochdimensionalen Raum repräsentiert;[18] der euklidische Abstand, oder die Kosinusähnlichkeit, bestimmt somit die zu quantifizierende Ähnlichkeit in diesem Merkmalsraum. Sie bleibt jedoch, auch bei mathematischer Quantifizierung, ein von menschlichen Urteilen geprägtes Konzept: Wenn bestimmte kulturelle oder ästhetische Perspektiven in einem Datensatz überrepräsentiert sind, können diese Präferenzen in das Modell eingeschrieben werden. Dies wiederum beeinflusst die Art und Weise, wie das Modell Bilder wahrnimmt und welche Merkmale es als signifikant für die Ähnlichkeitsbewertung ansieht – worauf im Folgenden zurückzukommen sein wird.

Aufgrund der vielfältigen in den Ähnlichkeitsraum eingeschriebenen Merkmale ist auch die erfassbare Ähnlichkeit in Forschungswerkzeugen wie *iART* variabel, sofern sie visuell fixiert werden kann: Bei einer Suche nach ›creation‹ kann sie stilistischer oder ikonographischer Natur sein, aber auch formal bedingt, etwa durch die Körperhaltung der dargestellten Figuren oder die Farbgebung des Bildes. Eine Ähnlichkeit, die über das rein Visuelle hinausgeht und auf der Historizität der Werke beruht, erfordert jedoch eine tiefer gehende, kontextuelle Betrachtung, die beispielsweise von CNNs nicht explizit geleistet wird. [10]

3. Ähnlichkeit als multimodales Konstrukt

Das »Gemeinsame und das jeweils Eigene […] begrifflich zu fassen«,[19] wie Thürlemann es formuliert, und damit Kategorien zu bilden, ist für die kunsthistorische wie für die informatische Ähnlichkeitsbestimmung aus zwei Gründen zentral: Einerseits definiert sie ein oder mehrere Merkmale, die aus einer Auswahl eine definierte Teilmenge erzeugen – und damit Mengen, die als Kategorien definiert werden können; andererseits führt sie die Subsumtion dieser Teilmenge unter einen natürlichsprachlichen Begriff ein. Insbesondere der letztgenannte Aspekt muss auch in *iART* berücksichtigt werden: In unserem Beispiel suchen wir mit Sprache nach visuellen Konzepten. Daraus ergibt sich zwangsläufig eine weitere Form der Ähnlichkeit: die des sprachlichen Begriffs zu einem visuellen Konzept. Ähnlichkeit ist hier multimodal zu verstehen – als Zusammenspiel von Bild und Text. [11]

3.1 Multimodale Klassifizierung

Wie viele andere neuronale Netze werden CNNs in der Regel überwacht trainiert: Jedem Eingabebild wird eine textuelle Klasse zugeordnet. Diese Zuordnung ermöglicht es dem Modell, zwischen vordefinierten Klassen (wie ›creation‹) zu differenzieren, indem ein Merkmalsraum erzeugt wird, der die visuellen Eigenschaften der jeweiligen Klasse repräsentiert. Jeder Klasse wird dabei – als Maß für die Ähnlichkeit – eine Wahrscheinlichkeit zugeordnet, die die Zugehörigkeit eines Bildes zu dieser Klasse angibt; ein Schwellenwert schränkt die Auswahl auf die Klassen ein, deren Wahrscheinlichkeit hinreichend groß ist. Das Modell kann demnach keine Klassen erkennen, die nicht im Trainingsdatensatz enthalten sind, auch wenn sie ähnliche Merkmale aufweisen, oder semantische Beziehungen zu bereits vordefinierten Kategorien (wie ›creation of adam‹) besitzen – jede unbekannte Klasse muss zunächst durch Trainingsdaten visuell spezifiziert werden. [12]

[18] Vgl. Goodfellow et al. 2016.
[19] Thürlemann 2005, S. 167.

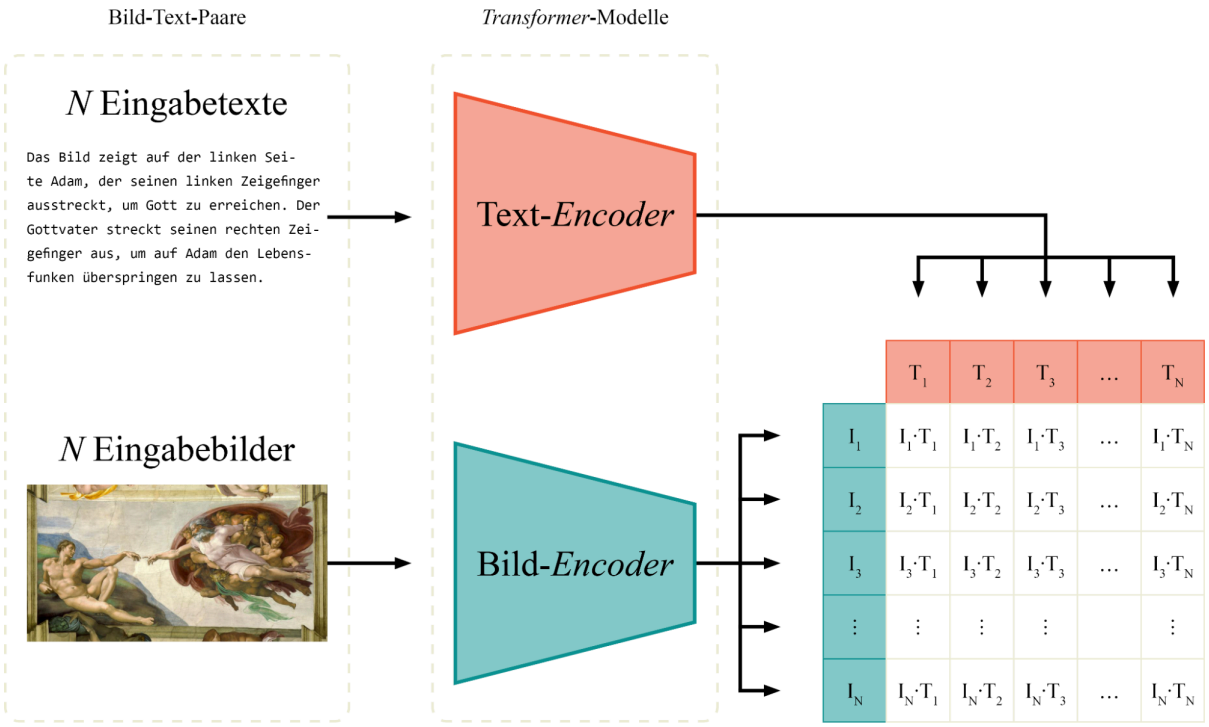

Abb. 4: Schematische Darstellung des Trainingsprozesses mit *CLIP* anhand eines Bild-Text-Paares zu Michelangelos *The Creation of Adam* (1508–1512). [Grafik: Stefanie Schneider / Matthias Springstein 2024]

[13] Modelle wie *CLIP* (Contrastive Language-Image Pre-Training)[20] erweitern diese Grenzen, indem sie Bilder selbstüberwacht mit natürlichsprachlichen Beschreibungen für das Training verknüpfen. Sie überführen Bild- und Textdaten mit modalitätsspezifischen Encodern, die auf sogenannten Transformer-Modellen basieren,[21] in einen gemeinsamen Merkmalsraum, um die Korrespondenz zwischen sprachlichem Begriff und visuellem Konzept zu erzeugen (Abbildung 4). Wesentlich für die *CLIP*-Architektur ist es, die aus Bild und zugehörigem Text generierten Embeddings so anzupassen, dass eine maximale Ähnlichkeit erreicht wird und so ein räumliches Netzwerk ähnlicher sprachlicher Begriffe und visueller Konzepte entsteht. Dieser Ansatz ermöglicht es, den realweltlichen Kontext eines Bildes zu integrieren und die dem Bild innewohnende Komplexität adäquater als bisherige Ansätze zu modellieren; nicht nur einfache, auf einen Begriff reduzierte Klassen können auf diese Weise mit *CLIP* erkannt werden, sondern auch komplexe Szenarien, die in der kunsthistorischen Forschung beispielsweise durch das alphanumerische Klassifikationssystem *Iconclass*[22] abgebildet werden können. Wie in Abbildung 5 dargestellt, können so für Michelangelos Adam ikonographisch bedeutsame Notationen wie 93A211 (»assemblies of the gods in the air, possibly on the clouds«) automatisch hinterlegt werden. Dadurch wird ein semantischer Rahmen geschaffen, der die Generalisierbarkeit und die Anwendbarkeit der Modelle auf eine Vielzahl von Domänen erhöht. Diese Fähigkeit, auch mit nicht zum Training verwendeten Klassen umgehen zu können, wird in der Informatik als *Zero-Shot-Lernen* bezeichnet.[23] Die für jedes Bild gefundenen Klassen können in *iART* zum Beispiel zusammen mit den von der jeweiligen Institution manuell bereitgestellten Metadaten verwendet werden, um die Ergebnisse zu facettieren.

[20] Radford et al. 2021.
[21] Vgl. Vaswani et al. 2017.
[22] Van de Waal 1973–1985.
[23] Vgl. Xian et al. 2019.

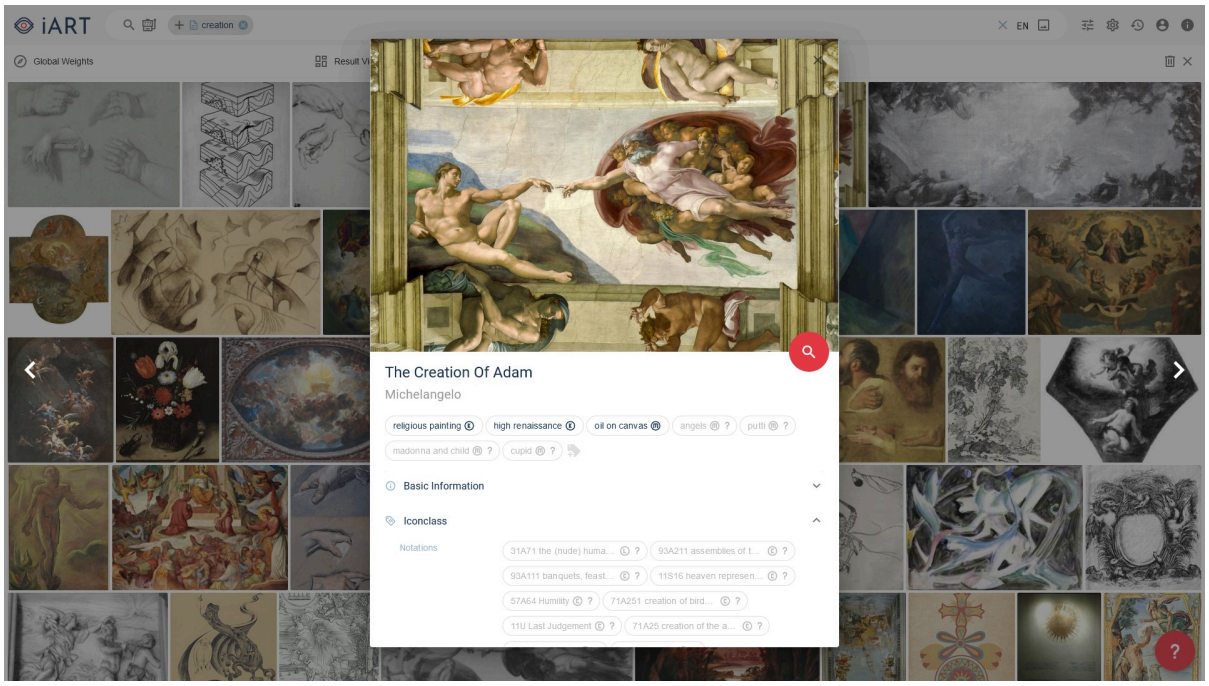

Abb. 5: Einzelobjektansicht von Michelangelos *The Creation of Adam* (1508–1512) im Forschungswerkzeug *iART* mit den für das Bild gefundenen Iconclass-Notationen. [Bilquelle: *iART*, Suchbegriff creation]

3.2 Multimodales Retrieval

*CLIP*s Zero-Shot-Fähigkeiten prädestinieren es sowohl für Klassifikations- als auch für Retrieval-Aufgaben – den computergestützten Prozess des Wiederfindens, hier von Bildern, die für die Nutzer*innen gemäß ihrem Informationsbedarf relevant sein könnten. Bei einer Bild-zu-Bild-Suche wird der Bild-Encoder von *CLIP* im Retrieval zunächst dazu verwendet, um die in einer Datenbank gespeicherten Bilder in Embeddings umzuwandeln, wie in Abbildung 6 zu sehen ist, während eine (textuelle) Suchanfrage, ›creation of adam‹, über den Text-Encoder verarbeitet wird. Das Embedding der Anfrage wird dann mit den Embeddings der Bilder verglichen, und die Ergebnisse dieses Vergleichs nach der Wahrscheinlichkeit sortiert, mit der das Bild das über den textuellen Begriff gesuchte visuelle Konzept enthält; die Bilder mit der höchsten Übereinstimmung werden zurückgegeben. Beispielsweise wird Michelangelos *The Creation of Adam* in der Nähe des Begriffs ›creation‹ positioniert und mit hoher Wahrscheinlichkeit in *iART* angezeigt, wenn eine entsprechende Suchanfrage gestellt wird. Da *CLIP* als sogenanntes Foundation Model mit einer großen Anzahl von Bild-Text-Paaren aus unterschiedlichen Domänen trainiert wurde, verfügt es über ein grundlegendes »Weltwissen«[24] für nachgelagerte Aufgaben – Downstream Tasks wie ein kunsthistorisches Retrieval. Diese Verknüpfung von Bild und Text ermöglicht es, den durch unimodale Methoden entstehenden Semantic Gap bei der Anwendung digitaler Methoden zu überwinden. Vor allem visuelle Forschungsdaten können durch die Einbettung in einen natürlichsprachigen Beschreibungskontext zielgerichteter untersucht werden und Semantiken komplexer und facettenreicher wiedergeben. Die Abkehr von den starren Kategorisierungen unimodaler Modelle hin zu intuitiven Abfragen multimodaler Modelle birgt das Potenzial für einen Multimodal Turn in den digitalen Geisteswissenschaften.[25] [14]

[24] Vgl. Bommasani et al. 2021.
[25] Vgl. Smits / Wevers 2023.

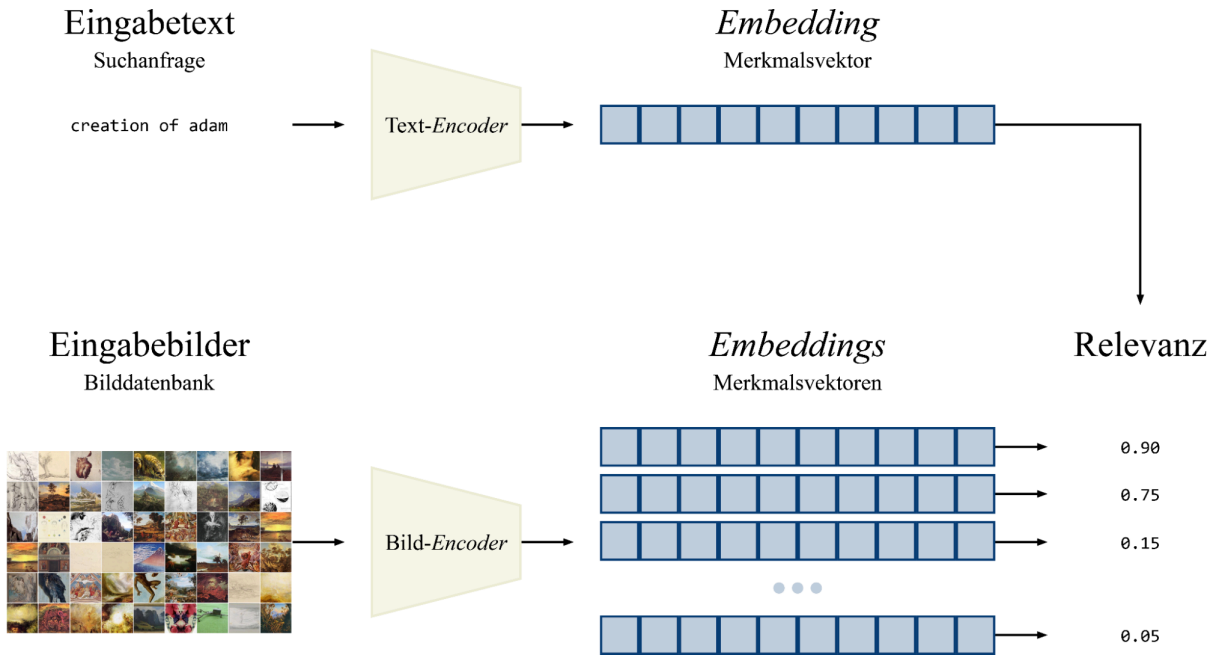

Abb. 6: Schematische Darstellung des Retrieval-Prozesses mit *CLIP* für die Suchanfrage ›creation of adam‹. [Grafik: Stefanie Schneider / Matthias Springstein 2024]

4. Die *Black Box* künstlicher neuronaler Netze

[15] Beide in den vorangegangenen Abschnitten vorgestellten Modelle, CNNs und Transformer, sind zwar prinzipiell in der Lage, Ähnlichkeiten durch den Vergleich von Merkmalsvektoren zu quantifizieren, jedoch sind die Prozesse, durch die diese Vektoren während des Trainings generiert werden, nicht transparent; warum ein Merkmal als relevant erachtet wird und ein anderes nicht, bleibt unklar. Die Vielzahl der während des Trainings erlernten Parameter führt dazu, dass die Vorhersagen – und somit die Basis der Klassenzuordnungen – für die Nutzer*innen von Retrieval-Werkzeugen wie *iART* einer sogenannten *Black Box* gleichen.[26] In Transformer-Modellen erschwert der integrierte Aufmerksamkeitsmechanismus die Interpretation der Parameter zusätzlich, weil er keinen Einblick in die Merkmale der Trainingsbilder erlaubt, die die Entscheidungsfindung der Modelle tatsächlich beeinflussen.[27] So ist nicht nachvollziehbar, ob die in *iART* häufig in den Suchergebnissen vertretenen Hände für die Zuordnung zum Begriff ›creation‹ bedeutsam sind. Gerade aus der Perspektive der Geisteswissenschaften, die traditionell hermeneutisch das einzelne Objekt kontextualisieren, ist es entscheidend, der Opazität dieser Prozesse zu begegnen: »[R]ather than providing descriptions purely from the domains of a formal, technical and causal model of explanation [...], these technologies would benefit from critical approaches that take account of understanding, more common in the humanities and social sciences«.[28] Im Bereich der *XAI* werden daher Techniken entwickelt, die nicht nur die Prozesse erklären, sondern auch die den neuronalen Netzen zugrundeliegenden Konzepte verständlich machen.[29] Neben Erklärbarkeit ist der Begriff der Interpretierbarkeit essentiell: Zielt Erklärbarkeit darauf ab, für den Menschen verständliche Erklärungen der Modellvorhersagen zu liefern, konzentriert sich Interpretierbarkeit darauf, wie diese Vorhersagen in den Merkmalsräumen der neuronalen Netze repräsentiert werden.[30]

[26] Vgl. Kuang 2017.
[27] Vgl. Vaswani et al. 2017.
[28] Berry 2023.
[29] Vgl. Guidotti et al. 2019.
[30] Vgl. Ries et al. 2024, S. 3.

[16] Bereits 2015 haben Google-Forscher unter der Leitung von Alexander Mordvintsev mit *DeepDream* eine Technik zur Visualisierung eingeführt, um die Klassifizierungsprozesse von CNNs nachvollziehbar zu machen. *DeepDream* sollte nicht nur veranschaulichen, was ein Netz während des Trainings ›lernt‹, sondern auch zeigen, wie visuelle Konzepte kombiniert werden können und so Einblicke in die Ursprünge des kreativen Prozesses geben. Der Ansatz ermöglicht es, die ›Essenz‹ eines Konzepts zu visualisieren, indem seine für die Klassifizierung entscheidenden Merkmale dargestellt werden.[31] In der Kunstgeschichte können diese Merkmale so durch hermeneutische Betrachtung in den Forschungsprozess integriert werden.[32] Auch diese ›Sichtbarmachung‹ algorithmischer Prozesse ist jedoch nicht vollständig nachvollziehbar und sollte kritisch hinterfragt werden, da Visualisierungen – dem Begriff des Metabildes von W.J.T. Mitchell folgend – als ›Bilder von Bildern‹ betrachtet werden müssen, die ihre eigene Bedingtheit reflektieren.[33] Ergänzend dazu bieten Aufmerksamkeitskarten, wie sie durch *Grad-CAM* (Gradient-weighted Class Activation Mapping) ermöglicht werden,[34] eine Methode zur Hervorhebung wichtiger Bildbereiche, bei der der Gradient zum Eingabebild anhand eines Suchbegriffs berechnet wird. Aus kunsthistorischer Sicht können zum Beispiel die Regionen eines Bildes identifiziert werden, die für die Klassenzuordnung durch künstliche neuronale Netze entscheidend sind.[35] Attribute, oder auch Körperhaltungen, die auf bestimmte Symbole, Berufe oder Stände hinweisen, können so effizient hervorgehoben und in ihrer Bedeutung validiert werden. Indem die Stärke der Assoziation zwischen einem sprachlichen Begriff und einem visuellen Konzept gemessen und als Heatmap dargestellt wird, können solche Techniken auch in multimodalen Architekturen verwendet werden, um sogenannte »Mental Images«[36] von visuellen Konzepten zu erforschen, wie später gezeigt wird.

5. Aspekte der Erklärbarkeit maschinengenerierter Vorhersagen

[17] Für die Erklärbarkeit und die Interpretierbarkeit künstlicher neuronaler Netze ist die kritische Reflexion von Verzerrungen entscheidend: Um möglichst unverzerrte algorithmische Entscheidungsprozesse zu gewährleisten, muss jede Form der Voreingenommenheit gegenüber Individuen, Gruppen oder – auch – Objekten aufgrund inhärenter oder erworbener Merkmale vermieden werden.[37] Verzerrungen, die bestimmte Gruppen entweder bevorzugen oder benachteiligen, werden unter dem Terminus der Diskriminierung subsumiert. Der Begriff, der sich vom lateinischen ›discriminare‹ – dem Unterscheiden oder Trennen – ableitet, wird in sozialwissenschaftlichen wie informatischen Kontexten verwendet, allerdings mit unterschiedlichen Konnotationen: Während in der Informatik darunter vor allem das selektive Filtern und Ordnen von Daten verstanden wird, fokussiert der Begriff in den Sozialwissenschaften auf die ungerechte Behandlung von Individuen aufgrund sozialer Kategorien (etwa Geschlecht, Sexualität und Alter).[38] Verzerrungen in künstlichen neuronalen Netzen können aber ebenso im letzteren Sinne diskriminierend wirken. So lassen sich in maschinellen Lernverfahren bis zu sieben Arten von Verzerrungen unterscheiden,[39] die wir im Folgenden auf drei Aspekte der Erklärbarkeit reduzieren: (1) historische Verzerrungen, (2) Verzerrungen des Datensatzes und (3) algorithmische Verzerrungen.[40] Diese werden im Rahmen der Suchergebnisse zu ›creation‹ in *iART* eingehend analysiert und diskutiert, wobei Lösungsansätze zur Minimierung – oder zumindest ›Sichtbarmachung‹ – der Verzerrungen mit sogenannten Aufmerksamkeitskarten aufgezeigt werden.

[31] Vgl. Mordvintsev et al. 2015.
[32] Vgl. Offert 2019.
[33] Vgl. Offert / Bell 2021.
[34] Selvaraju et al. 2017.
[35] Vgl. Bell / Offert 2021.
[36] Impett / Offert 2023.
[37] Vgl. Mehrabi et al. 2021.
[38] Vgl. Apprich et al. 2018, S. 9.
[39] Vgl. Suresh / Guttag 2021.
[40] Vgl. Pasquinelli / Joler 2021, S. 1265.

Wie oben erwähnt, können diese Untersuchungen auch im Rahmen des *Tool Criticism* betrachtet werden, bei dem epistemologische und methodologische Aspekte kritisch untersucht werden. Ziel ist es, implizite Eigenschaften – hier: Verzerrungen – aufzudecken, die in den verwendeten Werkzeugen verankert sind und derer sich weder Informatiker*innen noch Geisteswissenschaftler*innen in interdisziplinären Konstellationen zwangsläufig bewusst sind.[41] [18]

5.1 Historische Verzerrung

Historische Verzerrungen sind in der Kunstgeschichte das Ergebnis sich wandelnder gesellschaftlicher Bedingungen, die sich in der Über- und Unterrepräsentation bestimmter Gruppen, in der Wahl der Sujets und in etablierten Konventionen niederschlagen. Entscheidungen über die Repräsentation in Kunstwerken – von Menschen und ihren Funktionen – können daher exkludierend sein. Sie sind eng mit den politischen und sozialen Dynamiken ihrer Zeit verbunden.[42] Künstlerische Praktiken reflektieren immer auch gesellschaftliche Auseinandersetzungen um Klasse, Geschlecht und andere soziale Kategorien.[43] [19]

Wird Gott in den Ergebnissen der Suche nach ›creation of adam‹ in *iART* also stets als (alter) weißer Mann dargestellt, so entspricht dies den Konventionen der jeweiligen zeitgenössischen Kunstproduktion und -praxis – die Verzerrung ist hier Teil des Forschungskontextes und kein Defizit der Suchmaschine. Der Versuch, solche Formen historischer Verzerrung auszugleichen, ist durchaus problematisch: So kann die historische Wahrheit als solche verzerrt werden. Die erzwungene Integration von Diversität in generative Modelle wie Googles Gemini AI führte beispielsweise zu historisch unzutreffenden Darstellungen, etwa von schwarzen Wehrmachtssoldaten.[44] Solche Ergebnisse, die möglicherweise auf die Praxis des *Shadow Prompting* zurückzuführen sind – eine Methode, mit der die Ergebnisse absichtlich diversifiziert werden sollen, indem die Eingabeaufforderung um Wörter wie ›black‹ erweitert wird –,[45] sind für präzise wissenschaftliche Methoden letztlich unbrauchbar, da sie falsche und ungenaue Resultate begünstigen. *Data Balancing*, auf das in Abschnitt 5.3 Algorithmische Verzerrung näher eingegangen wird, wäre eine sinnvollere Möglichkeit, diese Verzerrung auszugleichen. [20]

5.2 Verzerrung des Datensatzes

Verzerrungen des Datensatzes ergeben sich aus der Annotation und Kuration der Trainingsdaten. Diese Verzerrungen wirken sich je nach Modell und Aufgabenstellung unterschiedlich auf die Ergebnisse aus: Für überwacht trainierte Modelle wie CNNs sind mit sprachlichen Begriffen annotierte Trainingsdaten grundlegend, wie sie etwa in Datenbanken wie ImageNet vorliegen.[46] Die Verwendung veralteter Taxonomien wie WordNet und der darin enthaltenen diskriminierenden und simplifizierenden Begriffe führt jedoch zu problematischen Klassifikationen.[47] Darüber hinaus erschwert die Mehrdeutigkeit sowohl der sprachlichen Begriffe als auch der visuellen Konzepte die Annotation.[48] Für Modelle, deren Training selbstüberwacht ohne vordefinierte Klassen erfolgt, ist dagegen die Konstitution der Bild-Text-Paare entscheidend: Für den Trainingsdatensatz von *CLIP* wurden beispielsweise auf Grundlage von Wikipedia-Schlagwörtern 500.000 Suchanfragen durchgeführt, die jeweils bis zu 20.000 Bild-Text-Paare lieferten – insgesamt also 400 Millionen Paare.[49] Bei dieser Datenmenge ist es nahezu unmöglich, alle potenziell fragwürdigen Inhalte [21]

[41] Vgl. Herrmann et al. 2023.
[42] Vgl. Held / Schneider 1993, S. 10–11.
[43] Vgl. Pollock 1988, S. 9–10.
[44] Vgl. Grant 2024.
[45] Vgl. Salvaggio 2023.
[46] Vgl. Deng et al. 2009.
[47] Vgl. Crawford / Paglen 2019.
[48] Vgl. Orr / Crawford 2023.
[49] Vgl. Radford et al. 2021, S. 3.

herauszufiltern und damit Verzerrungen grundsätzlich zu vermeiden. So geben die Autor*innen des Modells zu bedenken: »[O]ur system […] disproportionately attached labels to do with hair and appearance in general to women more than men. […] Additionally, [it] attached some labels that described high status occupations disproportionately more often to men such as ›executive‹ and ›doctor‹. This […] points to historical gendered differences«.[50] Diese – teilweise auch historisch begingte – Verzerrung in den Trainingsdaten führte dazu, dass prestigeträchtige Berufe in den Beschreibungstexten (der Abbildungen) von Männern überrepräsentiert waren, während in denen von Frauen eher äußerliche Merkmale im Vordergrund standen. Solche Formen der Über- und Unterrepräsentation spiegeln sich auch in den Ergebnissen der Modelle wider.

Verzerrungen des Datensatzes müssen auch im Hinblick auf die Anwendung der Modelle – hier in der Kunstgeschichte – diskutiert werden: Die Art und Weise, wie Kunstwerke online dargestellt und beschrieben werden, beeinflusst implizit die Ergebnisse der Ähnlichkeitsbestimmung. Aber nicht nur hier entstehen Fehldarstellungen: Da vor allem nicht-kunsthistorische Daten für das Training von computationalen Modellen verwendet werden, stammt die Mehrzahl der Klassen und Bild-Text-Paare nicht aus der Kunstgeschichte selbst, sondern aus Datenbanken für Stock-Fotografie, von Internet-Shops oder *Wikimedia Commons* – um nur einige der für unseren Kontext relevanten Kategorien zu nennen.[51] Die Überprüfung der Konstitution der Trainingsdaten von *CLIP* ist dabei nur durch eigene stichprobenartige Recherchen möglich, da die Daten nicht frei verfügbar sind. Wir führen daher exemplarisch zwei Recherchen durch: eine für ›creation‹ auf Google, gefiltert nach Bildern mit Creative-Commons-Lizenz (Abbildung 7) und eine für ›creation of adam‹ auf der E-Commerce-Website Etsy (Abbildung 8). Die bereits in *iART* beobachtete Assoziation des Begriffs ›creation‹ mit Händen bestätigt sich in der Google-Suche: Auch hier finden sich zahlreiche Hände, ob als Emoji – in Anlehnung an Michelangelos Werk – oder als fotorealistische Darstellung (eines Mystikers mit Turban). Auch die Suche nach ›creation of adam‹ auf Etsy ergibt, dass häufig nur die ikonische Handgeste des Kunstwerks reproduziert wird; der (vollständige) Titel des Kunstwerks wird jedoch trotzdem in den Beschreibungstexten angegeben. [22]

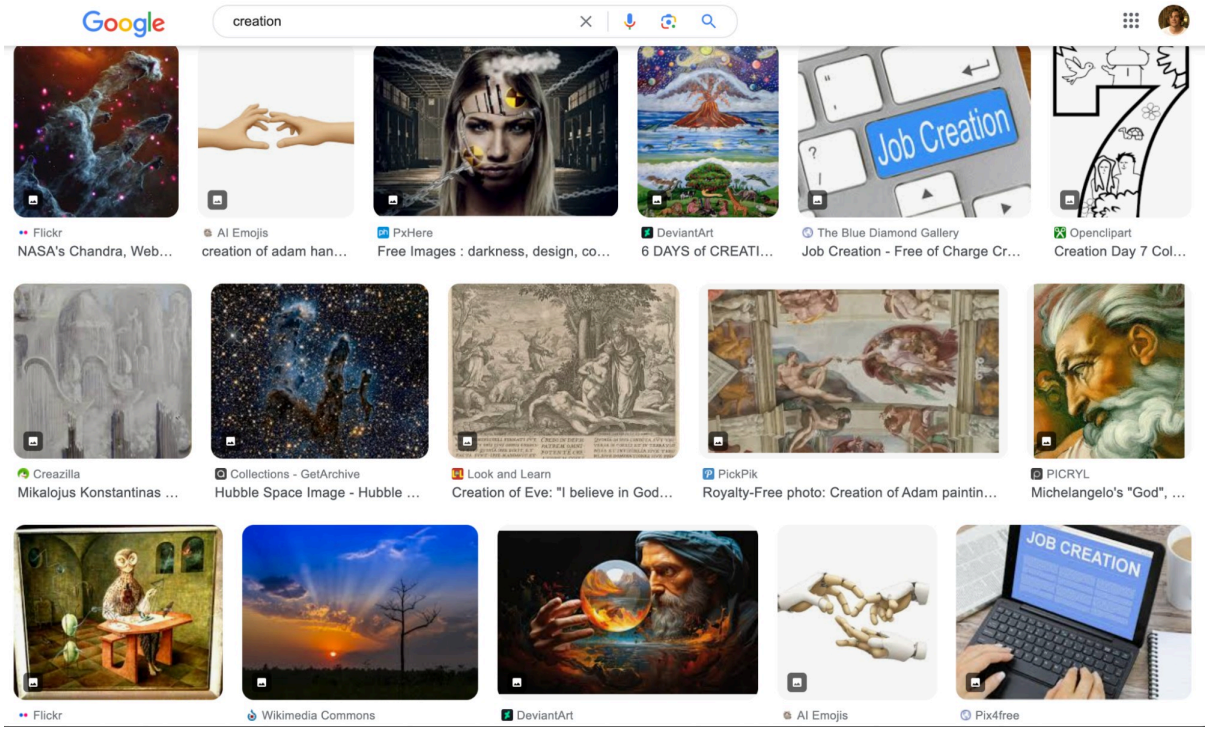

Abb. 7: Suchergebnisse für den Begriff ›creation‹ auf Google, gefiltert nach Bildern mit Creative-Commons-Lizenz. [Bildquelle: Google, Suchbegriff creation / Filter nach CC-Lizenz]

[50] Radford et al. 2021, S. 23.
[51] Vgl. Buschek / Thorp 2024.

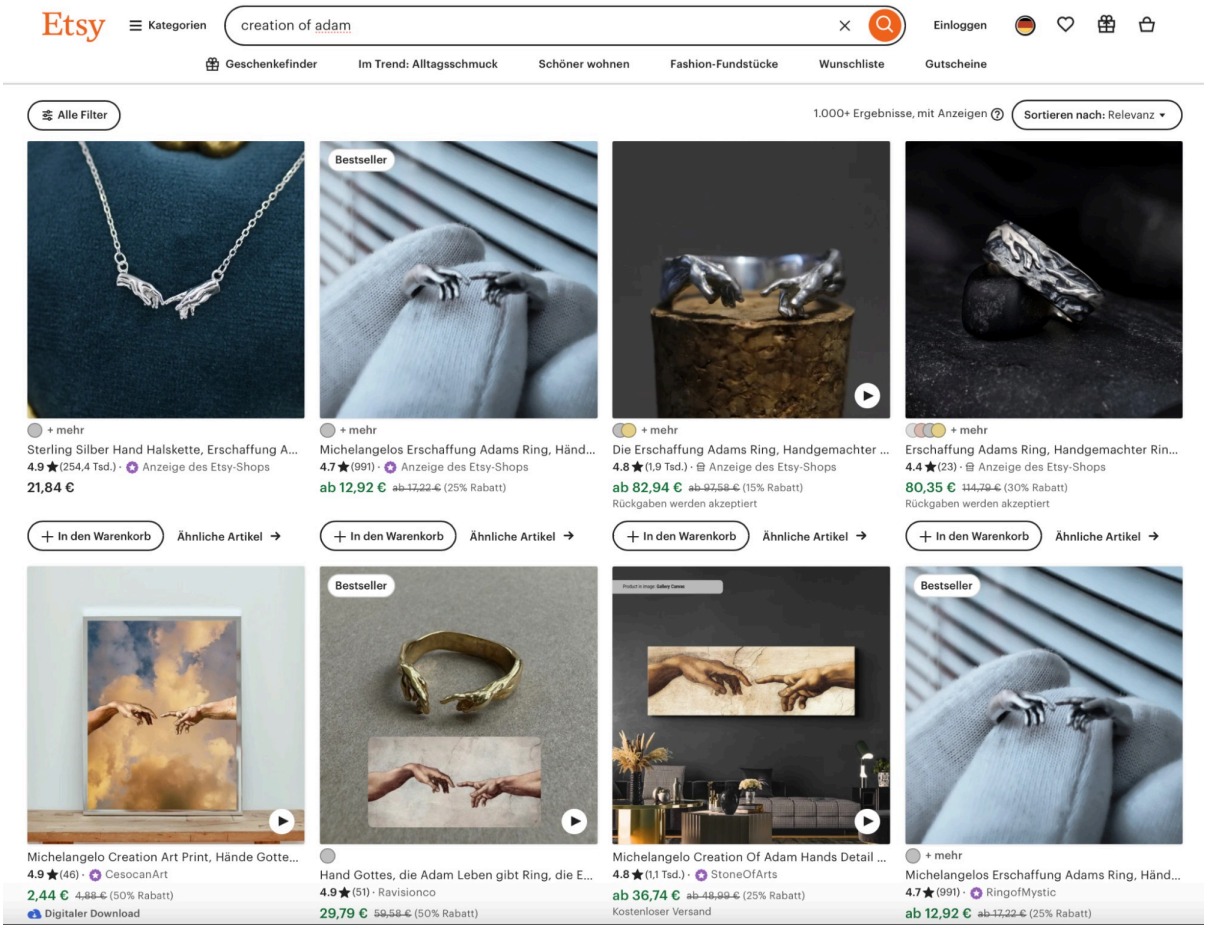

Abb. 8: Suchergebnisse auf der E-Commerce-Website Etsy für den Begriff ›creation of adam‹. [Bildquelle: Etsy, Suchbegriff creation of adam]

Erhält das Modell nun im Training für den sprachlichen Begriff ›creation‹ viele dieser Abbildungen von Händen als Bild-Text-Paare, so weist auch der sich bildende Merkmalsraum eine größere Ähnlichkeit des Begriffs ›creation‹ mit Händen als mit Adam oder der Darstellung Gottes auf. Die hierdurch resultierende Verzerrung der Trainingsdaten kann also dazu führen, dass das Modell nur das visuelle Konzept der Hände als entscheidend für den Begriff ›creation‹ abspeichert und bei zukünftigen Suchanfragen bevorzugt ausgibt. [23]

5.3 Algorithmische Verzerrung

Algorithmische Verzerrungen entstehen durch Wechselwirkungen zwischen den beiden zuvor genannten Verzerrungen – der historischen Verzerrung und der Verzerrung des Datensatzes – und können sich gegenseitig verstärken. Die in künstlichen neuronalen Netzen ablaufenden Prozesse komprimieren eben diese Trainingsdaten, wobei Informationen verloren gehen oder sich in ihrer Gewichtung ändern können.[52] Um algorithmische Verzerrungen einzuschränken, ist es entscheidend, klar definierte Fairnesskriterien festzulegen.[53] Verzerrungen in den Trainingsdaten sollten nicht als unveränderlich betrachtet, sondern durch den Algorithmus behandelt werden.[54] Dazu gibt es verschiedene Ansätze: Attribute wie Ethnie oder Geschlecht können etwa beim Training abgeschwächt werden;[55] auch durch Methoden wie *Data Balancing* [24]

[52] Vgl. Pasquinelli / Joler 2021, S. 1265.
[53] Vgl. Corbett-Davies et al. 2017.
[54] Vgl. Narayanan 2019.
[55] Vgl. Hardt et al. 2016.

kann eine geringere Diskriminierung erreicht werden.[56] Insbesondere Zero-Shot-Modelle wie *CLIP* tendieren zum sogenannten *Association Bias*, bei dem soziale Stereotype nicht nur durch Repräsentation, sondern auch durch Assoziation reproduziert werden.[57] So könnte der Begriff ›creation‹ eher männlich konnotiert sein, also mit Gott und Adam assoziiert werden, aber weniger mit Frauen, obwohl auch hier Assoziationen denkbar wären.

[25] Erschwerend für die Auseinandersetzung mit algorithmischen Verzerrungen kommt hinzu, dass beim Training künstlicher neuronaler Netze die Eingaben – Texte wie Bilder – in undurchsichtige Strukturen überführt werden, die es den Nutzer*innen erschweren, die konkreten Prozesse der Ähnlichkeitsbestimmung nachzuvollziehen – und damit auch die in den Eingaben angelegten Verzerrungen. Interne Prozesse der jeweiligen Trainingsverfahren beeinflussen zudem die Wissensrepräsentation im Merkmalsraum. Die daraus resultierende Intransparenz verhindert, dass Nutzer*innen dynamische Prozesse, die zur Entscheidungsfindung beitragen, effektiv analysieren können und erschwert die Bewertung der Modellergebnisse erheblich.

5.4 Aufmerksamkeitskarten

[26] Eine leicht generalisierbare Methode, diese Verzerrungen zu identifizieren und kritisch in den Prozess der Ähnlichkeitsbestimmung einzubeziehen, sind die bereits erwähnten Aufmerksamkeitskarten. Diese Karten geben ein visuelles Feedback darüber, welche Teile einer Eingabe den größten Einfluss auf die Modellausgabe haben und visualisieren somit, welche Merkmale für die Entscheidungsfindung des Modells besonders relevant sind. Da Aufmerksamkeitskarten mit verschiedenen Modalitäten kompatibel sind, können zum Beispiel für die Klassifikation visueller Konzepte relevante Bildbereiche oder bei Textübersetzungen bestimmte Wortgruppen hervorgehoben werden. Etabliert haben sich verschiedene Berechnungsverfahren – wie *Grad-CAM*[58] –, von denen die meisten darauf beruhen, dass ein Suchbegriff vorgegeben und daraus der Gradient zum Eingabebild berechnet wird. In moderneren Architekturen werden jedoch zunehmend sogenannte Aufmerksamkeitsschichten eingesetzt, die es dem Netz ermöglichen, automatisch zu lernen, welche Bild- oder Textregion für eine Vorhersage relevant ist – wobei die Verschachtelung mehrerer solcher Schichten die Darstellung weiter verkomplizieren kann.[59] Um die internen Prozesse von *CLIP* zu visualisieren, können wir die Zero-Shot-Fähigkeit des Modells nutzen: Sie erlaubt uns, kleinere Bildregionen auszuwählen und jede dieser Regionen mit dem Merkmalsvektor eines Eingabewortes zu vergleichen. Auf diese Weise können Ähnlichkeiten zwischen der visualisierten Region und dem konkreten sprachlichen Begriff in einem Raster dargestellt werden.

[27] So werden in Abbildung 9 einzelne Bildregionen von vier Kunstwerken mit *CLIPSeg*[60] hervorgehoben und mit visuellen Konzepten assoziiert. Betrachten wir zunächst Michelangelos Gemälde selbst: Die Aufmerksamkeitskarte zur Suchanfrage ›creation of adam‹ fokussiert nicht alle Bereiche des Gemäldes gleichermaßen, sondern vor allem die Figur Adams unten links – obwohl die Anfrage dem Bildtitel entspricht (Abbildung 9b). Bei der Suche nach ›creation‹ hingegen werden alle Bereiche hervorgehoben, vor allem Adam und Gott scheinen für das Konzept ausschlaggebend zu sein (Abbildung 9c). Schließlich zeigt die Suche nach ›hand‹, wie *CLIPSeg* spezifische Begriffe erkennt und nur den jeweils für das Konzept relevanten Bereich markiert (Abbildung 9d). Auch die Studie für die Hände eines Armbrustschützen (1512–1516), die in *iART* an dritter Stelle bei ›creation‹ zurückgegeben wird, zeigt Interessantes: So scheinen alle dargestellten Hände stark auf die Begriffe ›creation‹ und ›creation of adam‹ zu reagieren (Abbildungen 9j und 9k). In Annibale Carraccis Gemälde *Pan und Diana* (1597–1602) hingegen aktiviert die Suche nach ›creation of adam‹ besonders den Kopf- und Schulterbereich des Fauns (Abbildung 9f). Hier ist eine Ähnlichkeit mit

[56] Vgl. Alabdulmohsin et al. 2024.
[57] Vgl. Alabdulmohsin et al. 2024.
[58] Selvaraju et al. 2017.
[59] Vgl. Vaswani et al. 2017.
[60] Lüddecke / Ecker 2022.

Michelangelos liegender Adam-Figur zu vermuten. Bei ›creation‹ erweitert sich dieser Bereich und fokussiert stärker auf die Interaktion – erweitert sich also um die dargestellte Diana; beide Figuren sind durch die Geste der Hand miteinander verbunden (Abbildung 9g). Auch im Relief *Die Vertreibung aus dem Paradies* (1649) sehen wir dieses kompositorische Verhältnis betont: Bei ›creation‹ zeigt sich der Bereich, der die beiden Figurengruppen verbindet, hervorgehoben, während bei ›creation of adam‹ vor allem die männliche Figur rechts unten – tatsächlich Adam – betont wird (Abbildungen 9n und 9o).

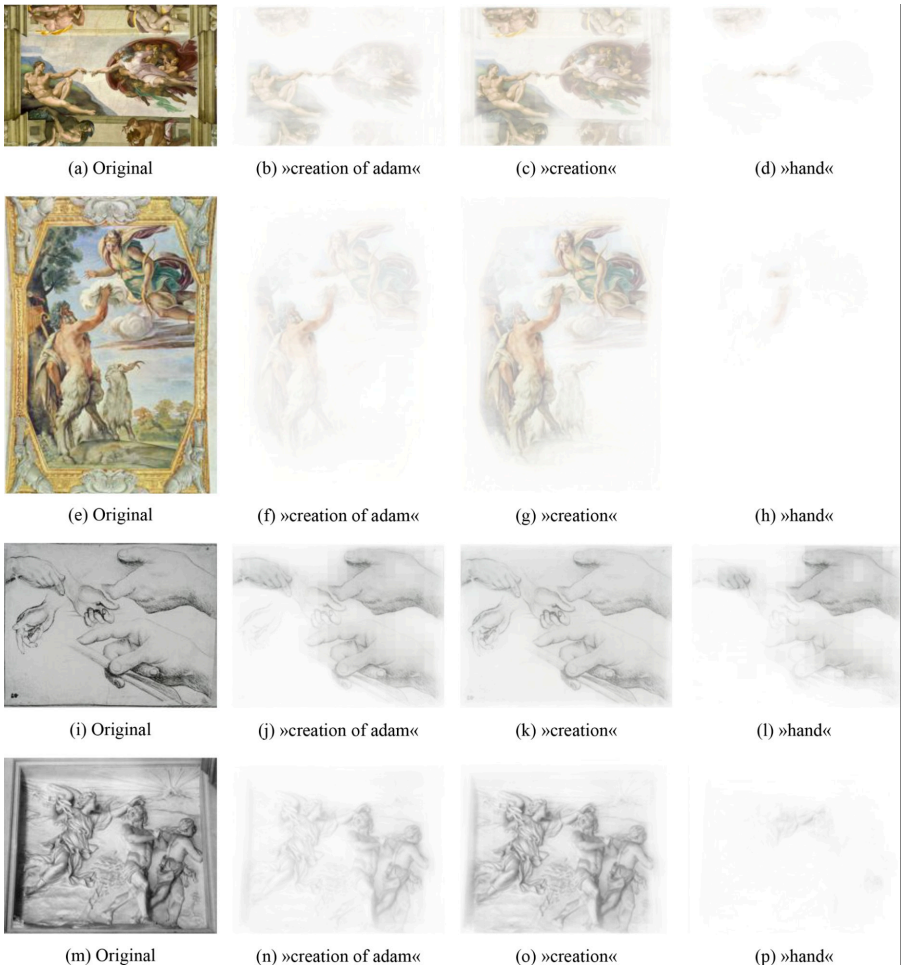

Abb. 9: Aufmerksamkeitskarten für vier Bilder, die im Forschungswerkzeug *iART* für den Begriff ›creation‹ zurückgegeben werden: Michelangelos Erschaffung Adams (1508–1512; a–d), Annibale Carraccis Pan und Diana (1597–1602; e–h), Studie für die Hände eines Armbrustschützen (1512–1516; i–l) und Die Vertreibung aus dem Paradies (1649; m–p). Die Karten wurden mit *CLIP*Seg (Lüddecke / Ecker 2022) und den Begriffen ›creation of adam‹, ›creation‹ und ›hand‹ erstellt. [Grafik: Matthias Springstein / Stefanie Schneider 2024]

Was können uns diese Visualisierungen nun über den Prozess der Ähnlichkeitsbestimmung und die gewonnenen Ähnlichkeitsverhältnisse sagen? Während bei ›creation of adam‹ insbesondere Bildregionen aktiviert werden, die männliche Aktdarstellungen zeigen, liegt der Schwerpunkt bei der Suche nach ›creation‹ eher auf der Bildkomposition und der Interaktion zwischen den Figuren: So werden hier Figuren und ihre jeweilige durch eine Handgeste dominierte Verbindung bevorzugt ausgegeben. Die Darstellung der Hände wiederum scheint auf alle drei Begriffe anzusprechen. Das Modell findet also Ähnlichkeiten zwischen sprachlichem Begriff und visuellem Konzept nicht nur bei den dargestellten Figuren und Figurenkonstellationen, sondern auch bei der Zuordnung abstrakter Begriffe zu einer bestimmten Bildregion, die sich in den Kunstwerken wiederfindet. Dies unterstützt die Hypothese, dass die als Trainingsdaten verwendeten Bild-Text-Paare die Hände stark mit dem Begriff der ›creation‹ assoziieren. [28]

6. Fazit und Ausblick

[29] Ausgehend von zwei kunsthistorischen Fallstudien wurden in diesem Beitrag Dimensionen des Konzepts der Ähnlichkeit vorgestellt und – damit einhergehend – Faktoren untersucht, die die Ähnlichkeitsbestimmung in künstlichen neuronalen Netzen beeinflussen. Darüber hinaus wurde mit den sogenannten Aufmerksamkeitskarten eine Methode dargelegt, die eine transparentere Nachvollziehbarkeit der Ähnlichkeitsbestimmung im automatisierten Bild-Retrieval gewährleisten kann.

[30] Aus kunsthistorischer Sicht konnte Ähnlichkeit als inhaltliche oder formale Gemeinsamkeit beschrieben werden, die sich unter bestimmten Identitäten zusammenfassen lässt. In der Informatik hingegen wird Ähnlichkeit als Nähe- oder Abstandsverhältnis in einem Vergleichsraum definiert, das als Zahl zwischen 0 und 1 dargestellt werden kann. Zentral für den Beitrag war die Integration multimodaler Ansätze zur Operationalisierung der Ähnlichkeitsbestimmung, beispielsweise in der Text-zu-Bild-Suche. In diesem Szenario wird die Dimension der Ähnlichkeit eines visuellen Konzepts um die Ähnlichkeit eines sprachlichen Begriffs erweitert. Im Sinne einer erklärbaren und interpretierbaren Anwendung künstlicher neuronaler Netze in der Kunstgeschichte wurden weiterhin mögliche Parameter untersucht, die einen Einfluss oder sogar eine Verzerrung der Ergebnisse bedingen können. So wurde herausgearbeitet, dass neben den algorithmischen Prozessen auch die Zusammensetzung der Trainingsdaten bei der Ähnlichkeitssuche eine Rolle für die Qualität der Suchergebnisse spielt.

[31] Deutlich wurde, wie durch interdisziplinäre Zusammenarbeit verschiedene Aspekte dieser komplexen Thematik untersucht und mit jeweils eigenen Methoden bearbeitet werden können. Nur durch Anwendung, Untersuchung und kritische Reflexion sowohl von informatischer als auch von kunsthistorischer Seite kann die Funktionalität und der domänenspezifische Einsatz der Werkzeuge in einem Forschungsprozess gewährleistet werden. Eine möglichst enge Zusammenarbeit bei der Auswahl und Aufbereitung der Trainingsdaten sowie bei der Erprobung und Evaluierung der Methoden für transparente und erklärbare Ergebnisse ist dabei zielführend.

[32] Perspektivisch sind weitere Aspekte der Zusammenarbeit beider Disziplinen zu nennen, die die Ähnlichkeitsbestimmung, aber auch generell den domänenspezifischen Einsatz künstlicher neuronaler Netze verbessern und transparenter machen. Die Integration von domänenspezifischen Informationen in die Trainingsdaten könnte durch die Verwendung von strukturiertem Wissen in Form von Wissensgraphen angereichert werden. Dies ermöglicht die Einbindung entsprechender Datenrepositorien in den Trainingsprozess, was nicht nur die Qualität der Ergebnisse für spezifische Anwendungsbereiche verbessert, sondern auch die Transparenz der Trainingsdaten erhöht. Das Sammeln, Zusammenführen und Aufbereiten von Daten in maschinenlesbarer Form ist ein interdisziplinärer Prozess, der Expertise aus informatischer und geisteswissenschaftlicher Warte erfordert.

[33] Durch multimodale Modelle wird die Kontextualisierung von Bild und Text ermöglicht – und damit die Interaktion mit den Modellen in natürlicher Sprache. Es ist nicht nur möglich, mit Text nach Bildern zu suchen, sondern die Modelle können auch Beschreibungen generieren oder anhand von Bildern Fragen beantworten. Dieser Prozess bietet Forscher*innen die Möglichkeit, Einblicke in die Repräsentation des Wissens im Merkmalsraum zu erhalten und spezifische Aspekte und Konzepte zu erfragen. Solche Methoden der direkten Interaktion mit dem Modell könnten die reflexive Nutzung künstlicher neuronaler Netze verbessern und eine intuitive Auseinandersetzung mit dem im Merkmalsraum organisierten Wissen ermöglichen.

Bibliografie

Ibrahim Alabdulmohsin / Xiao Wang / Andreas Steiner / Priya Goyal / Alexander D'Amour / Xiaohua Zhai: CLIP the Bias. How Useful is Balancing Data in Multimodal Learning? arXiv. 07.03.2024. PDF. DOI: 10.48550/arXiv.2403.04547

Clemens Apprich / Wendy Hui Kyong Chun / Florian Cramer / Hito Steyerl: Pattern Discrimination. In Search of Media. Minneapolis u. a. 2018. DOI: 10.14619/1457

Peter Bell / Fabian Offert: Reflections on Connoisseurship and Computer Vision. In: Journal of Art Historiography 24 (2021). PDF. [online]

David Berry: The Explainability Turn. In: Digital Humanities Quarterly 17 (2023), H. 2. HTML. [online]

Rishi Bommasani / Drew A. Hudson / Ehsan Adeli / Russ Altman / Simran Arora / Sydney von Arx / Michael S. Bernstein / Jeannette Bohg / Antoine Bosselut / Emma Brunskill / Erik Brynjolfsson / Shyamal Buch / Dallas Card / Rodrigo Castellon / Niladri Chatterji / Annie Chen / Kathleen Creel / Jared Quincy Davis / Dora Demszky / Chris Donahue / Moussa Doumbouya / Esin Durmus / Stefano Ermon / John Etchemendy / Kawin Ethayarajh / Li Fei-Fei / Chelsea Finn / Trevor Gale / Lauren Gillespie / Karan Goel / Noah Goodman / Shelby Grossman / Neel Guha / Tatsunori Hashimoto / Peter Henderson / John Hewitt / Daniel E. Ho / Jenny Hong / Kyle Hsu / Jing Huang / Thomas Icard / Saahil Jain / Dan Jurafsky / Pratyusha Kalluri / Siddharth Karamcheti / Geoff Keeling / Fereshte Khani / Omar Khattab / Pang Wei Koh / Mark Krass / Ranjay Krishna / Rohith Kuditipudi / Ananya Kumar / Faisal Ladhak / Mina Lee / Tony Lee / Jure Leskovec / Isabelle Levent / Xiang Lisa Li / Xuechen Li / Tengyu Ma / Ali Malik / Christopher D. Manning / Suvir Mirchandani / Eric Mitchell / Zanele Munyikwa / Suraj Nair / Avanika Narayan / Deepak Narayanan / Ben Newman / Allen Nie / Juan Carlos Niebles / Hamed Nilforoshan / Julian Nyarko / Giray Ogut / Laurel Orr / Isabel Papadimitriou / Joon Sung Park / Chris Piech / Eva Portelance / Christopher Potts / Aditi Raghunathan / Rob Reich / Hongyu Ren / Frieda Rong / Yusuf Roohani / Camilo Ruiz / Jack Ryan / Christopher Ré / Dorsa Sadigh / Shiori Sagawa / Keshav Santhanam / Andy Shih / Krishnan Srinivasan / Alex Tamkin / Rohan Taori / Armin W. Thomas / Florian Tramèr / Rose E. Wang / William Wang / Bohan Wu / Jiajun Wu / Yuhuai Wu / Sang Michael Xie / Michihiro Yasunaga /, Jiaxuan You / Matei Zaharia / Michael Zhang / Tianyi Zhang / Xikun Zhang / Yuhui Zhang / Lucia Zheng / Kaitlyn Zhou / Percy Liang : On the Opportunities and Risks of Foundation Models. arXiv. 16.08.2021. Version 3 vom 12.07.2022. PDF. DOI: 10.48550/arXiv.2108.07258

Christo Buschek / Jer Thorp: Models All The Way Down. 2024. [online]

Kate Crawford / Trevor Paglen: Excavating AI. The Politics of Images in Machine Learning Training Sets. 19.09.2019. HTML. [online]

Sam Corbett-Davies / Emma Pierson / Avi Feller / Sharad Goel / Aziz Huq: Algorithmic Decision Making and the Cost of Fairness. In: Stan Matwin / Shipeng Yu / Faisal Farooq (Hg.): Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (Halifax, CA-NS, 13.–17.08.2017), S. 797–806. PDF. DOI: 10.1145/3097983.3098095

Jia Deng / Wei Dong / Richard Socher / Li-Jia Li / Li Fei-Fei: ImageNet. A Large-Scale Hierarchical Image Database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2009, Miami, 20.–25.06.2009), S. 248–255. PDF. DOI: 10.1109/CVPR.2009.5206848

Finale Doshi-Velez / Been Kim: Towards A Rigorous Science of Interpretable Machine Learning. arXiv. 28.02.2017. Version 2 vom 02.03.2017. PDF. DOI: 10.48550/arXiv.1702.08608

Michel Foucault: Les Mots et les Choses. Une archéologie des sciences humaines. Paris 1966. [Nachweis im GVK]

Peter Geimer: Vergleichendes Sehen oder Gleichheit aus Versehen. In: Lena Bader / Martin Gaier / Falk Wolf (Hg.): Vergleichendes Sehen (= Eikones). München 2010, S. 45–69. [Nachweis im GVK]

Ian Goodfellow / Jean Pouget-Abadie / Mehdi Mirza / Bing Xu / David Warde-Farley / Sherjil Ozair / Aaron Courville / Yoshua Bengio: Generative Adversarial Nets. In: Zoubin Ghahramani / Max Welling / Corinna Cortes / Neil D. Lawrence / Kilian Q. Weinberger (Hg.): Advances in Neural Information Processing Systems 27. Annual Conference on Neural Information Processing Systems (Montreal, 08.–13.12.2014), S. 2672–2680. PDF. [online]

Nelson Goodman: Seven Strictures on Similarity. In: Nelson Goodman (Hg.): Problems and Projects. Indianapolis 1972, S. 437–447. [Nachweis im GVK]

Nico Grant: Google Chatbot's A.I. Images Put People of Color in Nazi-Era Uniforms. In: The New York Times vom 22.02.2024. [online]

Riccardo Guidotti / Anna Monreale / Salvatore Ruggiere / Franco Turini / Fosca Gianotti / Dino Pedreschi: A Survey of Methods for Explaining Black Box Models. In: Sartaj Sahni (Hg.): ACM Computing Surveys 51 (2019), H. 5, S. 1–42. PDF. DOI: 10.1145/3236009

Moritz Hardt / Eric Price / Nati Srebro: Equality of Opportunity in Supervised Learning. In: Daniel D. Lee / Ulrike von Luxburg / Roman Garnett / Masashi Sugiyama / Isabelle Guyon (Hg.): Advances in Neural Information Processing Systems 29 (Barcelona, 05.–10.12.2016). PDF. [online]

Jutta Held / Norbert Schneider: Sozialgeschichte der Malerei. Vom Spätmittelalter bis ins 20. Jahrhundert. Köln 1993. [Nachweis im GVK]

Berenike Herrmann / Anne-Sophie Bories / Francesca Frontini / Clèmence Jacquot / Steffen Pielström / Simone Rebora / Geoffrey Rockwell / Stéfan Sinclair: Tool Criticism in Practice. On Methods, Tools and Aims of Computational Literary Studies. In: Digital Humanities Quarterly 17 (2023), H. 2. HTML. [online]

Leonardo Impett / Fabian Offert: There Is a Digital Art History. arXiv. 14.08.2023. PDF. DOI: 10.48550/arXiv.2308.07464

Ernst Kris / Otto Kurz: Die Legende vom Künstler. Frankfurt / Main 2010. [Nachweis im GVK]

Cliff Kuang: Can A.I. Be Taught to Explain Itself? In: The New York Times vom 21.11.2017. [online]

George Kubler: The Shape of Time. Remarks on the History of Things. New Haven 2008. [Nachweis im GVK]

Junnan Li / Dongxu Li / Silvio Savarese / Steven Hoi: BLIP-2. Bootstrapping Language-Image Pre-Training with Frozen Image Encoders and Large Language Models. In: Andreas Krause / Emma Brunskill / Kyunghyun Cho / Barbara Engelhardt / Sivan Sabato / Jonathan Scarlett (Hg.): International Conference on Machine Learning. Proceedings of Machine Learning Research (ICML 2023, Honolulu, 23.–29.07.2023), S. 19730–19742. [online]

Timo Lüddecke / Alexander Ecker: Image Segmentation Using Text and Image Prompts. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2022, New Orleans, 19.–24.06.2022), S. 7086–7096. PDF. DOI: 10.1109/CVPR52688.2022.00695

Ninareh Mehrabi / Fred Morstatter / Nripsuta Saxena / Kristina Lerman / Aram Galstyan: In: A Survey on Bias and Fairness in Machine Learning. In: ACM Computing Surveys 54 (2021), H. 6, S. 1–35. [online]

Christoph Molnar: Interpretable Machine Learning. A Guide for Making Black Box Models Explainable. München 2020. [Nachweis im GVK]

Alexander Mordvintsev / Christopher Olah / Mike Tyka: Inceptionism. Going Deeper into Neural Networks. In: Google Research Blog, 18.06.2015. HTML. [online]

Arvind Narayanan: TL;DS. 21 Fairness Definition and Their Politics. In: Dora's World. Blog. 19.07.2019. HTML. [online]

Fabian Offert / Peter Bell: Perceptual Bias and Technical Metapictures. Critical Machine Vision as a Humanities Challenge. In: AI & SOCIETY 36 (2021), H. 4, S. 1133–1144. PDF. DOI: 10.1007/s00146-020-01058-z

Fabian Offert: Images of Image Machines. Visual Interpretability in Computer Vision for Art. In: Laura Leal-Taixé / Stefan Roth (Hg.): Computer Vision – ECCV 2018 Workshops (München, 08.–14.09.2018), S. 710–715. PDF. DOI: 10.1007/978-3-030-11012-3_54

Tillmann Ohm / Mar Canet Sola / Andres Karjus / Maximilian Schich: Collection Space Navigator. An Interactive Visualization Interface for Multidimensional Datasets. In: Roger Malina / Kang Zhang / Wei Zeng / Günter Wallner (Hg.): VINCI 2023. 16th International Symposium on Visual Information Communication and Interaction. Konferenzproceedings (Guangzhou, 22.–24.09.2023). New York 2023. PDF. DOI: 10.1145/3615522.3615546

Will Orr / Kate Crawford: The Social Construction of Datasets. On the Practices, Processes and Challenges of Dataset Creation for Machine Learning. SocArxiv. 07.11.2023. PDF. DOI: 10.31235/osf.io/8c9uh

Matteo Pasquinelli / Vladan Joler: The Nooscope Manifested. AI as Instrument of Knowledge Extractivism. In: AI & SOCIETY 36 (2021), H. 4, S. 1263–1280. PDF. DOI: 10.1007/s00146-020-01097-6

Ulrich Pfisterer: Kunstgeschichte. Zur Einführung. Hamburg 2020. [Nachweis im GVK]

Griselda Pollock: Vision and Difference. Feminism, Femininity and the Histories of Art. London u. a. 1988. [Nachweis im GVK]

Alec Radford / Jong Wook Kim / Chris Hallacy / Aditya Ramesh / Gabriel Goh / Sandhini Agarwal / Girish Sastry / Amanda Askell / Pamela Mishkin / Jack Clark / Gretchen Krueger / Ilya Sutskever: Learning Transferable Visual Models From Natural Language Supervision. In: Marina Meila / Tong Zhang (Hg.): 38th International Conference on Machine Learning. Conference Proceedings (Online, 18.–24.07.2021). 2021, S. 8748–8763. HTML. [online]

Thorsten Ries / Karina van Dalen-Oskam / Fabian Offert: Reproducibility and Explainability in Digital Humanities. In: International Journal of Digital Humanities 6 (2024), H.1, S. 1–7. PDF. DOI: 10.1007/s42803-023-00083-w

Eryk Salvaggio: Shining a Light on »Shadow Prompting«. In: Tech Policy Press. 19.10.2023. HTML. [online]

Stefanie Schneider / Matthias Springstein / Javad Rahnama / Hubertus Kohle / Ralph Ewerth / Eyke Hüllermeier: iART. Eine Suchmaschine zur Unterstützung von bildorientierten Forschungsprozessen. In: Michaela Geierhos (Hg.): 8. Tagung des Verbands Digital Humanities im deutschsprachigen Raum e. V. (DHd 2022, online, 07.–11.03.2022), S. 142–147. PDF. DOI: 10.5281/zenodo.6328175

Ramprasaath R. Selvaraju / Michael Cogswell / Abhishek Das / Ramakrishna Vedantam / Devi Parikh / Dhruv Batra: Grad-CAM. Visual Explanations from Deep Networks via Gradient-Based Localization. In: IEEE International Conference on Computer Vision (ICCV 2017, Venedig, 22.–29.10.2017), S. 618–626. PDF. DOI: 10.1109/ICCV.2017.74

Thomas Smits / Melvin Wevers: A Multimodal Turn in Digital Humanities. Using Contrastive Machine Learning Models to Explore, Enrich, and Analyze Digital Visual Historical Collections. In: Digital Scholarship in the Humanities 38 (2023), H. 3, S. 1267–1280. PDF. DOI: 10.1093/llc/fqad008

Matthias Springstein / Stefanie Schneider / Javad Rahnama / Eyke Hüllermeier / Hubertus Kohle / Ralph Ewerth: iART. A Search Engine for Art-Historical Images to Support Research in the Humanities. In: Heng Tao Shen / Yueting Zhuang / John R. Smith / Yang Yang / Pablo Cesar / Florian Metze / Balakrishnan Prabhakaran (Hg.): MM '21. ACM Multimedia (Chengdu / online, 20.–24.10.2021), S. 2801–2803. PDF. DOI: 10.1145/3474085.3478564

Julian Stalter / Matthias Springstein / Maximilian Kristen / Stefanie Schneider / Eric Müller-Budack / Ralph Ewerth / Hubertus Kohle: ReflectAI: Reflexionsbasierte künstliche Intelligenz in der Kunstgeschichte. In: Joëlle Weis / Thomas Haider / Estelle Bunout (Hg.): Book of Abstracts DHd2024. Quo vadis DH (Passau, 26.02.–01.03.2024). Passau 2024, S. 414–417. PDF. DOI: 10.5281/zenodo.10686565

Mukund Sundararajan / Ankur Taly / Qiqi Yan: Axiomatic Attribution for Deep Networks. In: Doina Precup / Yee Whye Teh (Hg.): Proceedings of the 34th International Conference on Machine Learning (ICML 2017, Sydney, 06.–11.08.2017), S. 3319–3328. PDF. [online]

Harini Suresh / John V. Guttag: A Framework for Understanding Sources of Harm Throughout the Machine Learning Life Cycle. In: Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO 2021, New York, 05.–09.10.2021). PDF. DOI: 10.1145/3465416.3483305

Felix Thürlemann: Bild gegen Bild. In: Aleida Assmann / Ulrich Gaier / Gisela Trommsdorff (Hg.): Zwischen Literatur und Anthropologie. Diskurse, Medien, Performanzen. Tübingen 2005, S. 163–174. [Nachweis im GVK]

Felix Thürlemann: Mehr als ein Bild. Für eine Kunstgeschichte des hyperimage. München 2013. [Nachweis im GVK]

Ralph Ubl: Symbol. In: Ulrich Pfisterer (Hg.): Metzler Lexikon Kunstgeschichte. Stuttgart 2011, S. 426–433. [Nachweis im GVK]

Nikolai Ufer / Max Simon / Sabine Lang / Björn Ommer: Large-Scale Interactive Retrieval in Art Collections Using Multi-Style Feature Aggregation. In: PLoS ONE 16 (2021), H. 11. DOI: 10.1371/journal.pone.0259718

Henri van de Waal: Iconclass. An Iconographic Classification System. Completed and Edited by L. D. Couprie with R. H. Fuchs. Amsterdam 1973–1985. [Nachweis im GVK]

Ashish Vaswani / Noam Shazeer / Niki Parmar / Jakob Uszkoreit / Llion Jones / Aidan N. Gomez / Lukasz Kaiser / Illia Polosukhin: Attention is All you Need. In: Isabelle Guyon / Ulrike von Luxburg / Samy Bengio / Hanna M. Wallach / Rob Fergus / S. V. N. Vishwanathan / Roman Garnett (Hg.): Advances in Neural Information Processing Systems 30. Annual Conference on Neural Information Processing Systems 2017 (Long Beach, US-CA, 04.–09.12.2017), S. 5998–6008. PDF. [online]

Heinrich Wölfflin: Kunstgeschichtliche Grundbegriffe. Das Problem der Stilentwickelung in der neueren Kunst. München 1915. [Nachweis im GVK]

Yongqin Xian / Christoph H. Lampert / Bernt Schiele / Zeynep Akata: Zero-Shot Learning. A Comprehensive Evaluation of the Good, the Bad and the Ugly. In: IEEE Transactions on Pattern Analysis and Machine Intelligence 41 (2019), H. 9, S. 2251–2265. PDF. DOI: 10.1109/TPAMI.2018.2857768

Abbildungsverzeichnis

Abb. 8: Suchergebnisse auf der E-Commerce-Website Etsy für den Begriff ›creation of adam‹. [Bildquelle: Etsy, Suchbegriff creation of adam]

Abb. 9: Aufmerksamkeitskarten für vier Bilder, die im Forschungswerkzeug *iART* für den Begriff ›creation‹ zurückgegeben werden: Michelangelos Erschaffung Adams (1508–1512; a–d), Annibale Carraccis Pan und Diana (1597–1602; e–h), Studie für die Hände eines Armbrustschützen (1512–1516; i–l) und Die Vertreibung aus dem Paradies (1649; m–p). Die Karten wurden mit *CLIP*Seg (Lüddecke / Ecker 2022) und den Begriffen ›creation of adam‹, ›creation‹ und ›hand‹ erstellt. [Grafik: Matthias Springstein / Stefanie Schneider 2024]