

TECHNISCHE UNIVERSITÄT DRESDEN

ZENTRUM FÜR INFORMATIONSDIENSTE
UND HOCHLEISTUNGSRECHNEN
PROF. DR. WOLFGANG E. NAGEL

Master-Arbeit

zur Erlangung des akademischen Grades
Master of Science

**Anwendung neuronaler Netze zur inhaltsbasierten
Bildsuche bei historischen Bilddarstellungen**

Jan Philipp Simon Langen
(Geboren am 26. Dezember 1994 in Münsterlingen, Schweiz)

Hochschullehrer: Prof. Dr. Wolfgang E. Nagel
Betreuer: Dr. Christoph Lehmann & Dr. Taras Lazariv

Dresden, 20. Oktober 2020

Aufgabenstellung der Masterarbeit

Name des Studenten:	Langen, Jan Philipp Simon
Matrikelnummer:	4081562
Studiengang:	Master Informatik TU Dresden
Thema:	Anwendung neuronaler Netze zur inhaltsbasierten Bildersuche bei historischen Bilddarstellungen auf Basis des DELF-Ansatzes (Deep Local Feature)
Zielstellung:	Aufgrund oftmals unvollständiger Metadaten und der großen Heterogenität historischer Bilddarstellungen ist die inhaltsbasierte Bildersuche bei der Nutzung großer Bilddatenbanken ein wichtiges Werkzeug. Gegenüber herkömmlichen Methoden (z.B. RANSAC, SIFT) liefern Ansätze basierend auf neuronalen Netzen vielversprechende Ergebnisse. Dabei ist DELF eine der aktuellen Entwicklungen. In der Masterarbeit werden folgende Punkte behandelt: <ol style="list-style-type: none">1. Anwendung und Dokumentation der verfügbaren Implementierung von DELF speziell zur inhaltsbasierten Suche in historischen Bildern2. Realisation der Implementierung auf PowerPC auf Taurus3. Abgrenzung und Vergleich von DELF mit mindestens einem weiteren Ansatz4. Experimente zur Parametergestaltung von DELF für historische Bilder
Betreuer Hochschullehrer:	Prof. Dr. Wolfgang E. Nagel
Betreuer:	Dr. Christoph Lehmann, Dr. Taras Lazariv
Beginn:	22.06.2020
Abgabe:	23.11.2020

Selbstständigkeitserklärung

Hiermit erkläre ich, dass ich die von mir am heutigen Tag dem Prüfungsausschuss der Fakultät Informatik eingereichte Master-Arbeit zum Thema:

Anwendung neuronaler Netze zur inhaltsbasierten Bildsuche bei historischen Bilddarstellungen

vollkommen selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt sowie Zitate kenntlich gemacht habe.

Dresden, den 20. Oktober 2020

Jan Philipp Simon Langen

Kurzfassung

Abstract

Inhaltsverzeichnis

1 Motivation	2
2 Verwandte Arbeiten	4
3 DELF	8
3.1 ResNet	9
3.2 Trainingsdaten	11
3.3 Fine-Tuning	11
3.4 Attention-Training	12
3.5 Extraktion und Verarbeitung	14
3.5.1 Multi-Skalen-Extraktion	14
3.5.2 Deskriptorlokalisierung	14
3.5.3 Deskriptorselektion	16
3.5.4 Datentransformation und Dimensionsreduktion	17
3.6 Matching	18
3.7 Verfahrensunterschiede im DELF Artikel	20
4 Evaluation	22
4.1 Evaluationsdaten	22
4.2 Retrievalmetriken	23
4.3 Parameteranalyse	25
4.3.1 Hyperparameteroptimierung der Trainingsphasen	26
4.3.2 Variieren der Deskriptorlänge	33
4.3.3 Alternativen zur Bewertung von Descriptormatches	38
4.3.4 Alternative Extraktionspunkte	42
4.4 Verfahrensvergleich	45
5 Fazit und Ausblick	51
5.1 Fazit	51
5.2 Ausblick	52
Literaturverzeichnis	53

1 Motivation

Präzise und effiziente Suchwerkzeuge sind essenziell, um große Datenmengen für einen Nutzer sinnvoll verwertbar zu machen. Dies gilt insbesondere auch im Bereich der Bildersuche. Die klassische Bildersuche basiert auf vom Nutzer formulierten Anfragen, mit deren Hilfe das Suchsystem eine Liste an passenden Bildkandidaten zusammenstellt und zurückgibt. Hierbei nutzt das System eine Reihe von Zusatzinformationen, sogenannte Metadaten wie Tags, Titel, Aufnahmeort oder Datum. Ein alternativer Ansatz der Suche, der in dieser Arbeit behandelt wird, ist die inhaltsbasierte Bildersuche (engl. Content-Based Image Retrieval, kurz CBIR). Hierbei werden vom Nutzer keine Anfragen formuliert. Stattdessen dient ein Bild als Suchanfrage. Ziel ist es, Bilder mit gleichem oder ähnlichem Bildinhalt als Ergebnis zurückzugeben. Ein Vorteil dieser Herangehensweise ist, dass der Nutzer keine Informationen über den Inhalt des als Anfrage verwendeten Bildes (Suchbild) benötigt. Das System arbeitet ausschließlich mit den Pixelinformationen der vorgegebenen Bilder. Ein weiterer Vorteil ist daher, dass weder im Suchbild noch in der Suchdatenbank Metadaten zu den Bildern vorhanden sein müssen, was die Einsatzmöglichkeiten von inhaltsbasierter Suche sehr flexibel gestaltet. Im Folgenden wird die inhaltsbasierte Suche auch als Image Retrieval bezeichnet.

Das Anwendungsgebiet dieser Arbeit ist die Suche in historischen Bildern. Diese weit gefasste Domäne ist besonders herausfordernd, da sie sehr heterogene Daten enthält. Dabei gibt es nicht nur große Unterschiede in den abgebildeten Bildinhalten wie Gebäuden, Naturaufnahmen oder Portraits, sondern auch in den verwendeten Aufnahmeverfahren. Durch die Fortschritte der Aufnahmetechnik können historische Bilder sowohl in Form von Zeichnungen oder Malerei, aber auch als Druck oder in anfänglichen Formen der Photographie vorliegen. Da Metadaten zu historischen Bildern erst bei der Digitalisierung hinzugefügt werden können, sind diese oft gar nicht oder nur lückenhaft vorhanden. Dies macht die inhaltsbasierte Suche für diese Domäne im Vergleich zur klassischen Bildersuche zu einer besonders geeigneten Methode. Mit der Umsetzung einer unterstützenden Suche für das UrbanHistory4D Projekt [1] ergibt sich ein konkreter Anwendungsfall für diese Arbeit. Das UrbanHistory4D Projekt befasst sich mit der Erstellung interaktiver Stadtkarten. An unterschiedlichen Positionen innerhalb der Stadtkarten lassen sich historische Aufnahmen anzeigen, welche Einblicke auf diese Orte aus vergangenen Zeiten ermöglichen. Das Akkumulieren und Zuordnen von historischen Bildern zu diesen Plätzen ist ein wesentlicher Arbeitsanteil bei der Erstellung der Karten. Image Retrieval Systeme können helfen, den Suchaufwand für die Ersteller der Karten signifikant zu reduzieren. Dabei handelt es sich um einen aktiven Forschungsbereich, in dessen Rahmen zurzeit unterschiedliche Suchsysteme analysiert werden.

Das 2016 vorgestellte Image Retrieval Verfahren DELF (attentive DEep Local Features) [2] entwickelt von Noh, Araujo et al., welches in dieser Arbeit untersucht wird, ist ein Deep Learning Ansatz. Durch den raschen Fortschritt im Bereich tiefer neuronaler Netzwerkarchitekturen der letzten Jahre erfreuen sich angelernte Ansätze immer größerer Beliebtheit. DELF erzielt auf bekannten Benchmarkdatensätzen

wie Oxford5k [3] und Paris6k [4] sehr gute Ergebnisse. Besonders gut schneidet DELF im Vergleich auf dem eigens erstellten Google Landmarks Datensatz [5] ab. Dieser enthält mit über 1 Mio. Bildern und 13k unterschiedlichen Motiven eine deutlich heterogeneren Mischung an Objekten als andere Benchmarks. Die gute Performanz auf diesem Datensatz lässt also hoffen, dass sich das DELF-Verfahren auch für die historische Domäne eignet.

2 Verwandte Arbeiten

Bei Information Retrieval handelt es sich um ein Problem aus dem Bereich der Computer Vision, welches bereits seit langem intensiv erforscht wird. In frühen Ansätzen versuchte man vor allem globale Beschreibungen von Bildern zu erstellen, um diese untereinander vergleichen zu können. Diese basierten zum Beispiel auf Farbhistogrammen oder Texturbeschreibungen [6]. Allerdings waren diese Ansätze oft sehr anfällig für Unterschiede in Beleuchtung, Skalierung und anderen Transformationen, wie sie bei unterschiedlichen Aufnahmen des selben Motivs auftreten können.

Ein wesentlicher Durchbruch gelang David G. Lowe 2004 mit der Entwicklung des SIFT-Verfahrens (Scale Invariant Feature Transform) [7]. Hierbei werden mehrere Konzepte vereint, um Bildbeschreibungen zu erzeugen, die robuster gegenüber unterschiedlichen Transformationen sind. So arbeitet der SIFT-Algorithmus beispielsweise nicht direkt auf der direkten Bilddarstellung, sondern im sogenannten Scale Space. Dieser besteht aus unterschiedlich skalierten Versionen des Ursprungsbildes, auf welche wiederum unterschiedlich starke Gauß-Filter angewendet werden. Betrachtet werden schließlich Differenzbilder zwischen benachbarten Stärken der Gauß-Filter-Ergebnisse. Die Verwendung von unterschiedlich skalierten Bildversionen macht die berechneten SIFT-Merkmale deutlich robuster gegen Skalierungsunterschiede. Das SIFT-Verfahren besteht aus zwei Phasen. In der ersten Phase werden über die Suche nach lokalen Extrema bedeutsame Bildpunkte ausgewählt. Für diese werden in der zweiten Phase einzelne Deskriptoren berechnet. Das Bild wird also nicht global beschrieben, sondern über viele lokale Deskriptoren dargestellt. Die lokalen Deskriptoren ergeben sich aus Histogrammen der Gradientrichtungen umliegender Bildpunkte. Diese werden relativ zu der dominanten Gradientrichtung in der Umgebung berechnet, was die Deskriptoren invariant gegenüber Rotationen macht. Lowes Entwicklung bildet den Ursprung für viele abgeleitete Verfahren wie SURF[8], PCA-SIFT[9] und RIFT[10]. Auch in aktueller Forschung werden Image Retrieval Verfahren untersucht, die mit SIFT-Merkmalen arbeiten [11].

Der Trend bei der Entwicklung neuer Image Retrieval Systeme geht aktuell jedoch hauptsächlich in Richtung von angelernten Verfahren. Die Basis dieser Verfahren bilden tiefe CNN-Architekturen (Convolutional Neural Networks). Ein neuronales Netzwerk lässt sich als eine schichtweise Aneinanderreihung nicht-linearer Funktionen auffassen. Convolutional Neural Networks sind ein Sonderfall neuronaler Netze, welche sogenannte Convolutional Layer (faltende Schichten) enthalten. In diesen Schichten werden Faltungs- bzw. Filteroperationen auf die Eingabedaten angewendet, um für das Netzwerk hilfreiche Merkmale in den Daten hervorzuheben. Dies ist durchaus vergleichbar mit den Filteroperationen, die im SIFT-Verfahren verwendet werden. Der Unterschied besteht jedoch darin, dass die Parameter der verwendeten Filtermasken sowie alle anderen Netzparameter nicht per Hand gewählt, sondern in einem Trainingsverfahren für den aktuellen Anwendungsfall optimiert werden. Der Entwickler bestimmt lediglich die grobe Architektur des Netzwerks, also die Anzahl, Größe und Reihenfolge der verwendeten Schichten sowie die Art der Operationen, die in ihnen durchgeführt werden. In Image Retrieval Systemen werden CNNs eingesetzt, um Bilddeskriptoren zu erstellen. Hierfür werden Zwischenergebnisse des Netzwerks, also die Ausgaben einer bestimmten Schicht genutzt. An welcher Stelle im Netzwerk die

Deskriptoren entnommen werden, ist dabei von entscheidender Bedeutung. Zeiler und Fergus haben in ihrer Studie zur Visualisierung von CNNs gezeigt [12], dass die früheren Schichten von CNNs typischerweise einfache Konzepte wie Kanten oder Ecken hervorheben. Mit wachsender Tiefe der betrachteten Netzwerkschicht steigt auch die Komplexität der Konzepte, die von den Ausgaben der Schicht beschrieben werden können.

In dem in [13] beschriebenen Image Retrieval System von Babenko, Slesarev et al. wird als Modell ein CNN bestehend aus fünf faltenden gefolgt von drei voll verbundenen Schichten (im Englischen fully-connected layer) genutzt. Als Deskriptoren werden die Ausgaben der ersten bzw. zweiten voll verbundenen Schicht verwendet. In einer voll verbundenen Schicht hat jeder Wert der Eingabe Einfluss auf jeden Wert in der Ausgabe. Die Ausgaben solcher Schichten werden also von der gesamten Bildeingabe beeinflusst und können daher als globale Deskriptoren verstanden werden. Diese intuitive Herangehensweise erzielt leichte Verbesserung gegenüber den zurzeit der Veröffentlichung gängigen algorithmischen Verfahren.

Razavian, Sullivan et al. stellen in [14] ein System auf Basis der in [15] beschriebenen Netzwerkarchitektur vor. Das Modell besteht ebenfalls aus fünf faltenden und drei voll verbundenen Schichten. Die Deskriptoren stammen aus den Ausgaben der letzten faltenden Schicht. Anders als bei Babenko, Slesarev et al. werden in diesem System mehrere Deskriptoren pro Bild erstellt. Hierfür werden systematisch Teilbilder aus Bildbereichen unterschiedlicher Größe generiert. Anschließend werden die Teilbilder auf eine feste Größe skaliert und als Eingabe in das Netzwerk gegeben. So wird für jeden betrachteten Bildbereich ein eigener lokaler Deskriptor erstellt. In ihren Experimenten stellen die Autoren fest, dass die Verwendung von lokalen Deskriptoren gegenüber einer globalen Betrachtung zu einer signifikanten Verbesserung der Retrievalperformanz führt. Der überwiegende Teil aktueller Retrieval Systeme setzt auf die Erstellung von lokalen Deskriptoren.

Eine interessante Frage bei der Konzeption von Image Retrieval Systemen, die mit lokalen Deskriptoren arbeiten, ist, wie man entscheidet, welche Bildregionen am sinnvollsten zu betrachten sind. Das ONE-Verfahren [16] von Xie, Hong et al. nutzt ein VGG-19 [17] Modell und extrahiert Deskriptoren aus der vorletzten voll verbundenen Schicht. Als Eingaben in das Netzwerk dienen sogenannte Object Proposals. Dabei handelt es sich um Bildausschnitte, welche Regionen umschließen, in denen Objekte vermutet werden. Die Autoren testen sowohl manuell annotierte sowie automatisch extrahierte Object Proposals und erzielen mit beiden Ansätzen ähnlich gute Ergebnisse. Für die automatische Bestimmung von Object Proposals nutzen sie das Selective Search Verfahren [18].

Das DELF-Verfahren [2], welches in der vorliegenden Arbeit untersucht wird, basiert ebenfalls auf lokalen Deskriptoren. Die Deskriptoren werden aus einer faltenden Schicht aus dem hinteren Teil eines ResNet-50 [19] Modells extrahiert. Als Eingabe in das Netzwerk werden Bilder in ihrer Gesamtheit betrachtet. Da bis zur Extraktionsschicht keine voll verbundenen Schichten genutzt werden, kann für jeden extrahierten Wert zurückgerechnet werden, von welchen Bereichen des Ursprungsbildes er beeinflusst wurde. Dies erlaubt es, die Ausgaben der Extraktionsschicht in einzelne lokale Deskriptoren zu unterteilen. Um auszuwählen, welche der lokalen Deskriptoren zur Darstellung eines Bildes genutzt werden sollen, werden die lokalen Deskriptoren in ein weiteres neuronales Netz gegeben. Dieses Netz hat die Aufgabe zu bewerten, wie geeignet die einzelnen Deskriptoren zur Beschreibung des Gesamtbildes sind. Auf Grund dieser Bewertung werden die wichtigsten Deskriptoren zu jedem Bild ausgewählt,

wogegen schlecht bewertete Deskriptoren verworfen werden. Der konzeptionelle Unterschied bei der Auswahl der Deskriptoren im Vergleich zum ONE-Verfahren ist, dass die Auswahl auf Grund der bereits berechneten Deskriptoren geschieht (anstatt auf Grund des Ursprungsbildes). Die Funktionsweise des DELF-Verfahrens wird in Kapitel 3 ab Seite 8 im Detail erklärt.

Bevor neuronale Netze für die Erstellung von Deskriptoren genutzt werden können, müssen ihre Parameter in einem Trainingsverfahren optimiert werden. Während des Trainings muss das Netzwerk eine Aufgabe lösen. Wie erfolgreich das Netzwerk dabei ist, wird mit Hilfe einer Fehlerfunktion dargestellt. Das Netz versucht seine Parameter so anzupassen, dass die Fehlerfunktion minimiert wird. Im Fall der bereits vorgestellten Verfahren wird dabei eine Dummy-Aufgabe, typischerweise die Klassifikation von Bildern, gelöst. In neueren Veröffentlichungen wurden jedoch einige Ansätze vorgestellt, die versuchen neuronale Netze direkt an Image Retrieval Aufgaben zu trainieren. Radenović, Tolias und Chum stellen in [20] einen solchen Ansatz vor. Während des Trainings betrachten sie dabei Bildpaare. Bei diesen Paaren handelt es sich entweder um korrekte Matches, falls die Bilder ähnliche Bildinhalte darstellen, oder um inkorrekte Matches, falls dies nicht der Fall ist. Die Bilder eines Paares durchlaufen identische Netze und erzeugen dabei jeweils eine Ausgabe. Für die Optimierung wird eine spezielle Fehlerfunktion definiert, welche bewertet wie stark sich die Netzwerkausgaben der Paare voneinander unterscheiden. Für korrekte Matches sollten sich die Netzwerkausgaben möglichst ähneln. Wird jedoch ein inkorrekt Match betrachtet, sollten auch die Ausgaben eine Mindestdifferenz aufweisen. Die Autoren testen ihr Verfahren auf unterschiedlichen CNN-Architekturen wie VGG [17] und AlexNet [21] und erzielen damit sehr gute Ergebnisse auf gängigen Retrievalbenchmarks.

Da Image Retrieval Systeme meist auf großen Bilddatenbanken eingesetzt werden und somit für eine Suchanfrage viele Bilder miteinander verglichen werden müssen, ist es sinnvoll, Bildrepräsentationen so kompakt wie möglich zu gestalten, um die Laufzeit der Suche zu verbessern. Insbesondere bei Verfahren, die lokale Deskriptoren erstellen und häufig hunderte oder tausende Merkmale pro Bild extrahieren, kann mit einer guten Kodierung viel Rechenzeit gespart werden. Ein beliebter Ansatz zur Erstellung kompakter Darstellungen aus lokalen Deskriptoren ist das BOVW-Modell (Bag-of-Visual-Words) [22], erstmals vorgestellt im Kontext von Textklassifikation von McCallum und Nigam. Hierbei werden zunächst alle aus einem Datensatz extrahierte Deskriptoren mittels Clusteranalyse (bspw. K-Means-Clustering [23]) in Gruppen eingeteilt. Deskriptoren, die dem gleichen Cluster zugeordnet werden, werden dabei auf das selbe "visuelle Wort" abgebildet. Als Beschreibung des Gesamtbilds dient ein Histogramm über die im Bild enthaltenen visuellen Wörter. Bei diesem Verfahren geht durch Quantisierung ein Teil der Information verloren. Das ebenfalls auf Clustering basierte VLAD-Verfahren [24] von Jégou, Douze et al. versucht diese Information nutzbar zu machen, indem es statt der Vorkommen die Quantisierungsfehler akkumuliert, die beim Abbilden auf die nächsten visuellen Worte entstehen.

Um eine Suchanfrage mit einer Rangliste der ähnlichsten Bilder zum Suchbild beantworten zu können, werden die Deskriptoren der Bilder in der Datenbank mit denen des Suchbildes verglichen. Als Metrik dient hierbei meist die euklidische Distanz. Auf kleinen Datensätzen ist es laufzeittechnisch sinnvoll, alle Kombinationen von Such- und Datenbankbildern zu vergleichen. Häufig werden bei größeren Datensätzen jedoch Methoden der approximierten nächsten Nachbarsuche (ANN) verwendet. Diese garantieren zwar kein optimales Ergebnis, erlauben jedoch eine deutlich schnellere Verarbeitung von Suchanfragen. So gibt es zum Beispiel Ansätze, Deskriptoren mit Hilfe spezieller Hashfunktionen zu vergleichen. Diese

werden so konstruiert, dass ähnliche Deskriptoren auf die gleichen bzw. möglichst ähnliche Hashcodes abgebildet werden, während gleichzeitig die Kollisionswahrscheinlichkeit für sehr unterschiedliche Deskriptoren minimal gehalten wird. Wang et al. beschreiben in ihrer Studie [25] unterschiedliche Konzepte für die Erstellung solcher Hashfunktionen.

3 DELF

Das DELF-Verfahren [2] von Noh, Araujo et al. bildet die Basis für die Experimente, die in dieser Arbeit durchgeführt werden. In dem zu DELF veröffentlichten Artikel werden einige Verfahrensschritte nicht oder nur oberflächlich beschrieben, was als Grundlage für eine detailgetreue Neuimplementierung nicht ausreicht. Die Autoren stellen ebenfalls den Quellcode einer Implementierung des DELF-Verfahrens zur Verfügung¹. Dieser weist jedoch einige konzeptionelle Unterschiede zu dem im Artikel beschriebenen Vorgehen auf. Für den Experimentalteil dieser Arbeit wurde DELF neu implementiert². Diese Implementierung orientiert sich an dem offiziell veröffentlichten Quellcode und wird im folgenden Kapitel schrittweise im Detail erklärt. Die Unterschiede zu den Beschreibungen des veröffentlichten Artikels werden in den letzten Abschnitten dieses Kapitels ab Seite 20 erörtert.

Das DELF-Verfahren lässt sich in vier Phasen einteilen (siehe Abb. 3.1). Zu Beginn steht das sogenannte Fine-Tuning. Hierbei wird ein vortrainiertes Modell, in unserem Fall ein ResNet-50 Netzwerk, auf einem neuen Datensatz weiter trainiert. Die Domäne der Bilder dieses Datensatzes sollte dabei möglichst nahe an der späteren Retrievalaufgabe sein, damit das Modell lernen kann, aussagekräftige Deskriptoren für diese Art von Bildern zu berechnen. In der nächsten Phase wird auf dem Modell aufbauend ein Attention-Netzwerk trainiert, welches die Güte berechneter Deskriptoren bewertet. In der dritten Phase werden für die Bilder der Datenbank, in der gesucht werden soll, Deskriptoren extrahiert. Mit Hilfe des Attention-Netzwerks wird eine Vorauswahl besonders geeigneter Deskriptoren getroffen. In anschließenden Vorverarbeitungsschritten werden die Deskriptoren in den Bildern lokalisiert und in eine kompaktere Form überführt. In der finalen Phase kann DELF aktiv genutzt werden. Es können nun Bilder als Suchanfragen gestellt werden. DELF vergleicht eine Anfrage mit allen Bildern des Datensatzes anhand der vorverarbeiteten Deskriptoren. Potentielle Matches zwischen Deskriptoren werden in einem letzten Schritt geometrisch verifiziert. Das Ergebnis einer Anfrage ist eine Rangliste der ähnlichsten Bilder, sortiert nach der Anzahl verifizierte Deskriptoren-Matches mit dem Anfragebild.

¹<https://github.com/tensorflow/models/tree/master/research/delf>, zuletzt besucht 16.06.20

²<https://gitlab.hrz.tu-chemnitz.de/s5407552--tu-dresden.de/hist4delf>

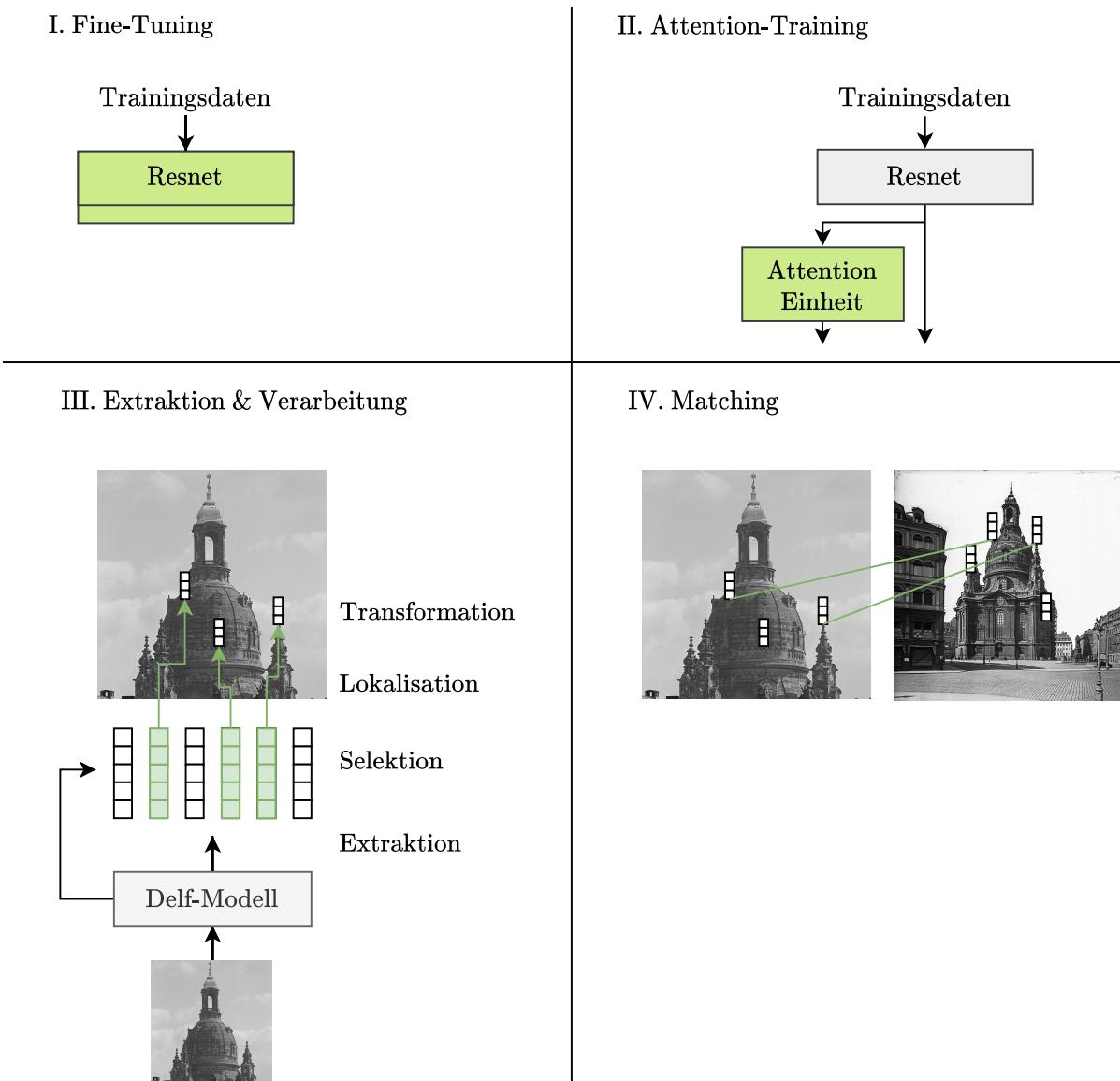


Abbildung 3.1: Die Phasen des DELF-Verfahrens

3.1 ResNet

Das DELF-Verfahren nutzt zur Erstellung von Deskriptoren ein Residuales Netzwerk (kurz ResNet). Bei der im Jahre 2015 vorgestellten ResNet Architektur [19] von He, Zhang et al. handelt es sich um eine der meist genutzten tiefen CNN-Architekturen der aktuellen Forschung. ResNets finden Anwendung in unterschiedlichen Machine Learning Aufgaben, wie Klassifikation, Objektdetektion oder Image Retrieval. Zeiler und Fergus haben gezeigt [12], dass CNNs mit wachsender Netzwerktiefe in der Lage sind, komplexere Merkmale zu detektieren. Es scheint daher intuitiv zur Lösung immer komplexerer Aufgaben zunehmend tiefere Netzwerke zu konstruieren. Allerdings stellt sich heraus, dass ab einem gewissen Punkt keine Verbesserungen mehr mit dem bloßen Aneinanderreihen von immer mehr Schichten erreicht werden können. Werden zu viele Schichten hinzugefügt, kann es sogar passieren, dass die Quali-

tät der Netzwerkvorhersagen abnimmt. Mit dem rasanten Anstieg der Anzahl an Netzwerkparametern wird es immer schwieriger, das Netzwerk zu optimieren. Parameter konvergieren deutlich langsamer zu einem Optimum und es gibt mehr lokale Minima, in denen ein Netzwerk im Optimierungsprozess stecken bleiben kann. ResNets wirken diesem Problem mit der Einführung sogenannter Skip-Verbindungen entgegen. Hierbei werden zusätzliche Direktverbindungen im Netzwerk geschaffen, bei denen einige Schichten übersprungen werden. Fließt eine Eingabe an den Beginn einer Skip-Verbindung, so wird auf dieser die Identität der Eingabe mitgeführt. Parallel durchläuft die Eingabe die übersprungenen Schichten. Am Ausgangspunkt der Verbindung wird schließlich die Ausgabe der übersprungenen Schichten mit der Identität summiert (siehe Abb. 3.2a). Durch die Bereitstellung der Identität hat das Netzwerk eine bessere Grundlage zur Optimierung und einzelne schlecht optimierte Schichten weniger negative Auswirkung auf die Netzwerkausgabe. Die Autoren stellen fest, dass CNNs bei Verwendung von Skip-Verbindung schneller zu einem Optimum konvergieren und dabei bessere Minima gefunden werden.

ResNets können in unterschiedlichen Konfigurationen erstellt werden. Das für DELF verwendete ResNet-50

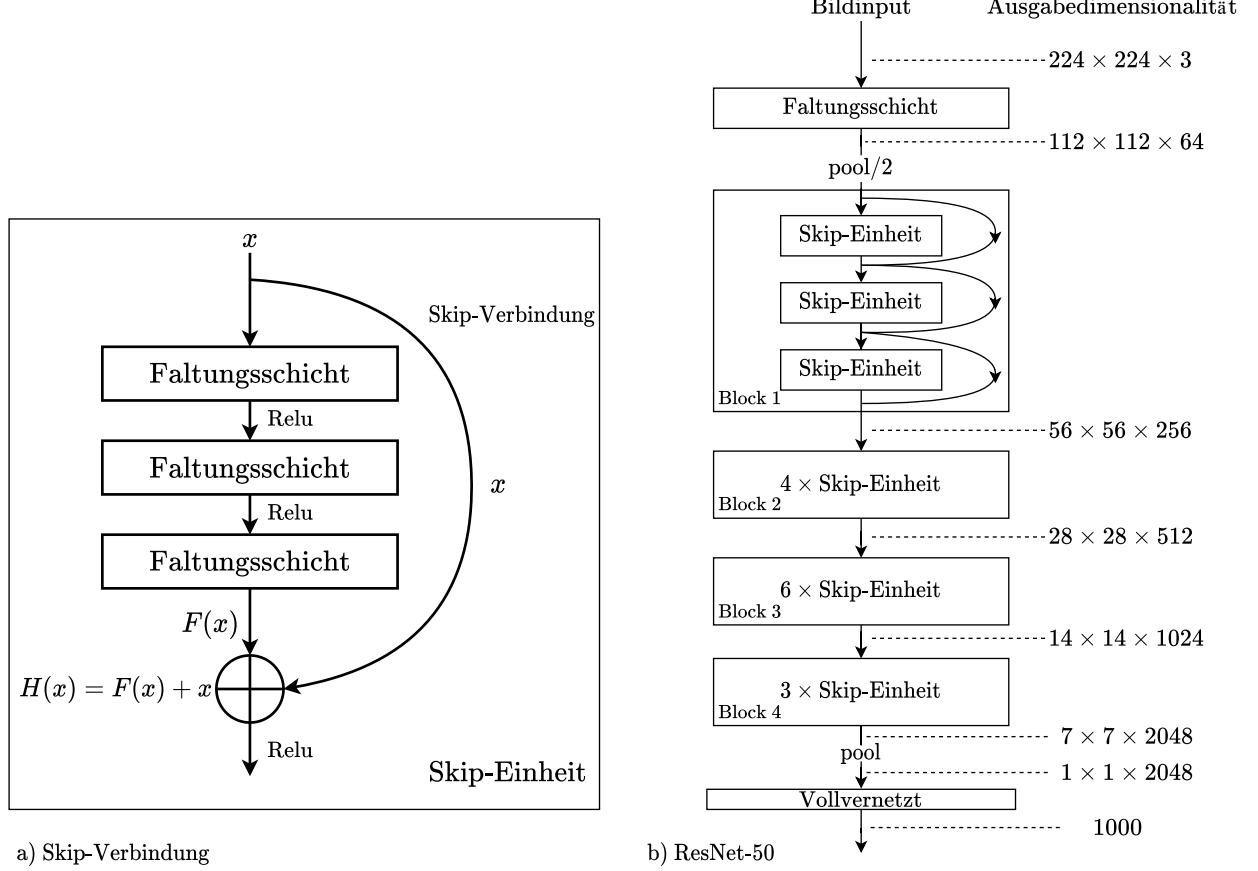


Abbildung 3.2: Aufbau der ResNet-Architektur (vgl. Fig.2, Fig.3 aus [19])

besteht aus 49 faltenden gefolgt von einer vollvernetzen Schicht. Skip-Verbindungen überspringen jeweils drei Schichten. Das Netzwerk kann in vier Blöcke unterteilt werden. Die Größe der einzelnen Featuremaps der Ausgabe verringert sich nach jedem Block um den Faktor vier, wohingegen die Merkmalstiefe bzw. Anzahl der Featuremaps in der Ausgabe steigt (vgl. Abb. 3.2b). In der Implementierung der vorliegenden Arbeit wird die von Torchvision zur Verfügung gestellte ResNet-50 Architektur ge-

nutzt³.

3.2 Trainingsdaten

Um die Modelle für das DELF-Verfahren zu trainieren, wird ein gelabelter Datensatz benötigt. Zum jetzigen Zeitpunkt steht kein solcher Datensatz von historischen Stadtaufnahmen mit ausreichender Größe zur Verfügung. Für die vorliegende Arbeit wird daher alternativ auf die Bilder der Google Landmark Challenge V2 [26] zurückgegriffen. Die Bilder entstammen einer Websuche auf der Wikimedia Datenbank⁴ und zeigen Sehenswürdigkeiten aus der ganzen Welt. Der überwiegende Teil (72%) zeigt dabei menschengemachte Sehenswürdigkeiten wie Kirchen, Museen oder Häuser. Auch wenn historische Aufnahmen keinen wesentlichen Teil der Bilder ausmachen, enthält der Datensatz viele ähnliche Inhalte zu den historischen Datensätzen, die für das Retrieval genutzt werden. Der Trainingssatz der Landmark Challenge ist mit über 4 Millionen Bildern aus über 200k unterschiedlichen Kategorien sehr groß und heterogen. Da bei der Zusammenstellung keine Verifizierung der Bildinhalts durchgeführt wird, kommt es häufig vor, dass Bilder in der falschen Kategorie eingesortiert sind. Für das DELF-Training wird daher ein bereinigtes Subset des Datensatzes verwendet, welches von Yokoo, Ozaki et al. im Rahmen ihrer Arbeit [27] erstellt wurde. Aus dem bereinigten Datensatz werden die 40 häufigsten Kategorien gewählt und so ein Trainingsdatensatz aus 39 790 gelabelten Bildern erstellt.

3.3 Fine-Tuning

Das Ziel der ersten Trainingsphase ist es, das ResNet Modell so zu optimieren, dass das Modell bei der Verarbeitung eines Bildes Zwischenergebnisse erzeugt, die den Bildinhalt aussagekräftig beschreiben. Dies ist die Voraussetzung, um später leistungsstarke Deskriptoren erstellen zu können. Während der Optimierung versucht das Netzwerk die Klassifikationsaufgabe des Trainingsdatensatzes zu lösen. Zu Beginn werden die Netzwerkparameter dabei nicht zufällig initialisiert. Stattdessen wird ein vortrainiertes Modell als Ausgangspunkt genutzt. Dieser Ansatz des Transferlernens wird als Fine-Tuning bezeichnet und ist eine gängige Methode, um den Trainingsprozess zu erleichtern. Auch in anderen Image Retrieval Systemen [14] [20] wird diese Methode für die Netzwerkoptimierung genutzt. Zeiler und Fergus zeigen in ihren Experimenten (vgl. [12] Kapitel 5.2), dass Netzwerke beim Training lernen, allgemein nützliche Merkmale zu extrahieren, die sich auf unterschiedliche Datensätze anwenden lassen. Um ein vortrainiertes Netzwerk auf einen neuen Datensatz anzupassen, sind daher nur kleine Veränderungen der Netzwerkparameter notwendig.

Als Ausgangspunkt für das DELF-Training wird ein auf ImageNet trainiertes ResNet-50 genutzt. Bei ImageNet handelt es sich um einen sehr großen Klassifikationsdatensatz mit 1.4M Bildern aus 1000 sehr unterschiedlichen Kategorien. Durch die Vielfalt an Kategorien eignen sich auf ImageNet trainierte Netzwerke als Ausgangspunkt für viele Klassifikationsaufgaben. Daher stellen die meisten Machine Learning Frameworks auf ImageNet trainierte Netzwerke zur Verfügung⁵.

Während des Trainings erwartet das Netzwerk quadratische Bilder mit einer Seitenlänge von 224 Pixeln

³<https://github.com/pytorch/vision/blob/c2e8a00885e68ae1200eb6440f540e181d9125de/torchvision/models/resnet.py>, zuletzt besucht 16.06.20

⁴<https://commons.wikimedia.org>, zuletzt besucht 18.06.20

⁵<https://pytorch.org/docs/stable/torchvision/models.html>, zuletzt besucht 23.06.20

und 3 Farbkanälen als Eingabe. Hierfür werden die Trainingsdaten zunächst quadratisch zugeschnitten und auf 250×250 Pixel skaliert. Anschließend wird ein zufälliger 224×244 Pixelbereich ausgewählt. Um das vortrainierte Netzwerk auf dem Trainingssatz weiter zu trainieren, muss das Netzwerk so angepasst werden, dass die Ausgaben der letzten Schicht die korrekte Form für die zur Optimierung verwendete Fehlerfunktion hat. Als Fehlerfunktion wird hier die Kreuzentropie berechnet:

$$\text{Kreuzentropie}(Y, \hat{Y}) = - \sum_{c \in C} Y(c) * \log \hat{Y}(c) \quad (3.1)$$

\hat{Y} ist hierbei die Verteilung der Klassenwahrscheinlichkeiten, die das Modell für eine Eingabe vorhergesagt hat. $\hat{Y}(c)$ ist die vom Netzwerk bestimmte Wahrscheinlichkeit, mit der die Eingabe der Klasse c zuzuordnen ist. Y beschreibt die tatsächliche Kategorie der Eingabe. Y ist also eine Verteilung, bei der die Wahrscheinlichkeit für jede bis auf die korrekte Kategorie 0 und für die tatsächliche Klasse 1 ist. Als Ausgabe des Netzwerks wird also ein Vektor der Wahrscheinlichkeitsverteilung erwartet, dessen Dimensionalität der Anzahl der unterschiedlichen Klassen $|C|$ im Datensatz entspricht und dessen Einträge sich auf 1 summieren.

Damit die Ausgaben des Netzwerks die richtige Dimensionalität haben, wird zunächst die letzte voll verbundene Schicht entfernt. An ihre Stelle tritt eine faltende Schicht mit 1×1 Filtermasken, die eine Merkmalstiefe von 2048 erwarten, was der Dimensionalität vorangehenden Schicht entspricht (vgl. Abb. 3.2b). In der faltenden Schicht werden $|C|$ dieser Filtermasken auf die Eingabe angewendet, wodurch die Ausgabe die gewünschte Dimensionalität erhält. Um die Netzwerkausgaben in den richtigen Wertebereich zu überführen, werden diese von einer Softmaxfunktion aktiviert, bevor die Fehlerfunktion berechnet wird:

$$\text{Softmax}(\hat{Y}')_c = \frac{e^{\hat{Y}'_c}}{\sum_{j \in C} e^{\hat{Y}'_j}} \forall c \in C \quad (3.2)$$

Hierbei ist \hat{Y}' ein Ausgabevektor des Netzwerks und \hat{Y}'_c der Eintrag des Vektors, welcher der Klasse c zugeordnet ist. Der resultierende Vektor kann als Wahrscheinlichkeitsverteilung über die unterschiedlichen Klassen in Bezug zur Eingabe interpretiert werden. Mit den vorgenommenen Modifikationen kann das ResNet-Modell auf dem Trainingsdatensatz optimiert werden. Die für das Fine-Tuning und Attention-Training verwendeten Hyperparameter werden im Experimentalteil in Abschnitt 4.3.1 ab Seite 26 erläutert.

3.4 Attention-Training

DELF erzeugt eine große Anzahl an lokalen Deskriptoren, um Bilder zu beschreiben. Da jeder Deskriptor nur einen Ausschnitt des Originalbildes beschreibt, werden auch Deskriptoren für wenig aussagekräftige Bereiche, wie z.B. Teile des Himmels erstellt. Diese Deskriptoren beanspruchen nicht nur zusätzliche Rechenzeit während des Matchingprozesses, sondern können auch zu falsch positiven Matches führen. Ziel der zweiten Trainingsphase ist es daher, auf Basis dieser Deskriptoren ein Netzwerk zu trainieren, welches in der Lage ist, die Qualität der Deskriptoren zu bewerten und so ungeeignete Kandidaten herauszufiltern. Zur Erstellung der Deskriptoren dient das ResNet-50, welches in der ersten Phase trainiert wurde. Als Deskriptoren werden dabei die Ausgaben aus dem dritten ResNet-Block genutzt (vgl. Abb. 3.2b).

Die Ausgaben haben eine Dimensionalität von $w \times h \times 1024$, wobei w und h abhängig von der Breite und Höhe des Eingabebildes sind. Die einzelnen Koordinaten der 1024 Featuremaps lassen sich jeweils auf einen Bildbereich in der Eingabe zurückführen. So kann die Ausgabe in $w \times h$ Deskriptoren der Größe 1024 eingeteilt werden. Wie auch beim Fine-Tuning werden die Bilder für das Training zunächst quadratisch zugeschnitten. Anschließend werden sie auf eine zufällige Seitenlänge zwischen 255 bis 720 Pixel skaliert. Die Skalierung hat Einfluss auf die Beschaffenheit der entstehenden Deskriptoren. Durch das zufällige Skalieren der Trainingsbilder lernt das Attention-Netzwerk mit Deskriptoren aus verschiedenen Skalen umzugehen. Während des Attention-Trainings werden die Parameter des ResNets nicht mehr verändert. Die Schichten nach dem Extraktionspunkt in Block 3 werden nicht mehr benötigt und können verworfen werden.

Aufgabe des Attention-Netzwerks ist es, für eine Eingabe an Deskriptoren der Form $w \times h \times 1024$ eine

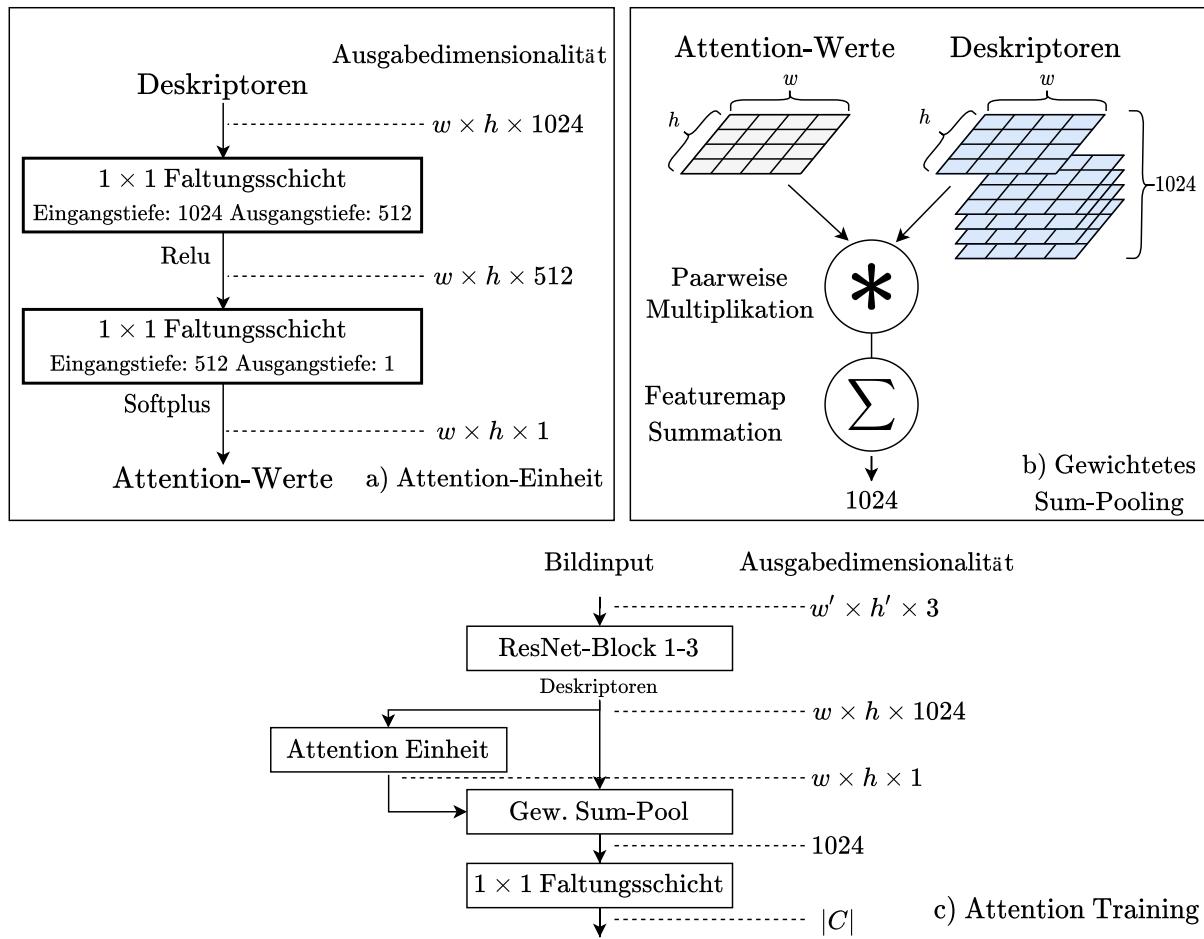


Abbildung 3.3: Architektur des Attention-Trainings

einzelne Featuremap der Größe $w \times h$ zu generieren, dessen Werte jeweils einen Descriptor bewerten. Wichtig ist dabei, dass die Berechnung der einzelnen Attention-Werte nur von den Werten der dazugehörigen Descriptoren abhängen dürfen. Dies kann durch faltende Schichten mit 1×1 -Filtermasken realisiert werden. Die Attention-Einheit besteht aus zwei solcher Schichten, welche die Merkmalstiefe der Descriptoren sukzessive auf 1 reduzieren (vgl. Abb. 3.3a). Um die Parameter der Attention-Einheit zu optimieren, müssen ihre Ausgaben zur Lösung der Trainingsaufgabe beitragen. Da die Attention-Werte

später genutzt werden um zu entscheiden, welche Deskriptoren Einfluss auf die Lösung der Retrievalaufgabe haben, ist es sinnvoll, sie auch beim Training in einer Form zu nutzen, die den Einfluss der Deskriptoren zur Lösung der Klassifikationsaufgabe reguliert. Die Attention-Werte werden zur Gewichtung der Deskriptoren genutzt und dafür elementweise mit den Featuremaps der Deskriptoren multipliziert. Anschließend werden die gewichteten Featuremaps zu jeweils einem Wert summiert. Als Ausgabe ergibt sich ein 1024 dimensionaler Vektor (vgl. Abb. 3.3b). Abschließend muss die Ausgabe in eine passende Form für die Berechnung der Kreuzentropie gebracht werden. Dies geschieht analog wie im Fine-Tuning durch Verwendung einer 1×1 -Faltungsschicht und anschließender Softmaxaktivierung der Ausgabe (vgl. 3.3c).

3.5 Extraktion und Verarbeitung

Nachdem das Training der Modelle abgeschlossen ist, kann mit der Extraktion aller benötigten Informationen über den Retrievaldatensatz begonnen werden. Netzwerkparameter werden ab jetzt nicht mehr modifiziert. Schichten und Operationen nach der Attention-Einheit erfüllen daher keine Zweck mehr und können entfernt werden.

3.5.1 Multi-Skalen-Extraktion

Für jedes Bild des Retrievaldatensatzes werden die lokalen Deskriptoren am Extraktionspunkt nach dem dritten ResNet-Block und die dazugehörigen Attention-Werte nach der Attention-Einheit extrahiert. Da die Skalierung des Bildinhalts Einfluss auf die resultierenden Deskriptoren hat, wird für jedes Bild eine Reihe von unterschiedlich skalierten Versionen betrachtet. Dies ist vergleichbar mit der Verwendung des Scale Spaces im SIFT-Verfahren (vgl. Kap.2 Abs.2) und soll zur Invarianz gegenüber Skalierungsoperationen beitragen. Für jedes Bild werden sechs unterschiedliche Skalen mit Skalierungsfaktoren zwischen 2 und $\frac{1}{4}$ erstellt, wobei sich benachbarte Skalen um den Faktor $\sqrt{2}$ unterscheiden.

3.5.2 Deskriptorlokalisierung

Für den weiteren Verlauf des Verfahrens muss jeder lokale Deskriptor einem Bereich des Eingangsbildes zuordenbar sein. Bei allen verwendeten Schichten des Modells bis zum Extraktionspunkt handelt es sich um Faltungs- oder Poolingschichten. Von welchen Bereichen der Eingabe die Ausgaben dieser Schichten abhängen, lässt sich an drei Parametern festmachen. Die Größe der Filtermasken k bestimmt die Größe des Einflussbereiches einzelner Ausgaben. Die Verschiebung der Filtermasken bzw. die Schrittgröße s bestimmt die Verschiebung zwischen den Einflussbereichen benachbarter Ausgaben. Das Padding p bestimmt die Größe des Pufferbereichs, welcher der Eingabe hinzugefügt wird, und sorgt so für eine initiale Verschiebung der Einflussbereiche. Das Padding ist im Folgenden immer symmetrisch und wird daher an jeder Seite der Eingabe hinzugefügt. Betrachtet man mehrere aufeinanderfolgende Schichten, so lassen

sich die Einflüsse dieser Parameter wie folgt rekursiv berechnen:

$$\hat{k}_n = \hat{k}_{n-1} + ((k_n - 1) * \hat{s}_{n-1}) \quad (3.3)$$

$$\hat{s}_n = \hat{s}_{n-1} * s_n \quad (3.4)$$

$$\hat{p}_n = \hat{p}_{n-1} + (p_n * \hat{s}_{n-1}) \quad (3.5)$$

$$\hat{k}_0 = k_0 \quad (3.6)$$

$$\hat{s}_0 = s_0 \quad (3.7)$$

$$\hat{p}_0 = p_0 \quad (3.8)$$

Wobei \hat{k}_n , \hat{s}_n und \hat{p}_n die Größe der Einflussbereiche, Schrittgröße und Paddinggröße in Bezug zur ursprüngliche Eingabe nach n Schichten repräsentieren. k_n , s_n und p_n zeigen dieselben Größen für Schicht n im Bezug zur Ausgabe der vorangehenden Schicht. In Abbildung 3.4 werden diese Berechnungen exemplarisch erklärt. Berechnet man diese Werte für den Extraktionspunkt nach dem dritten ResNet-

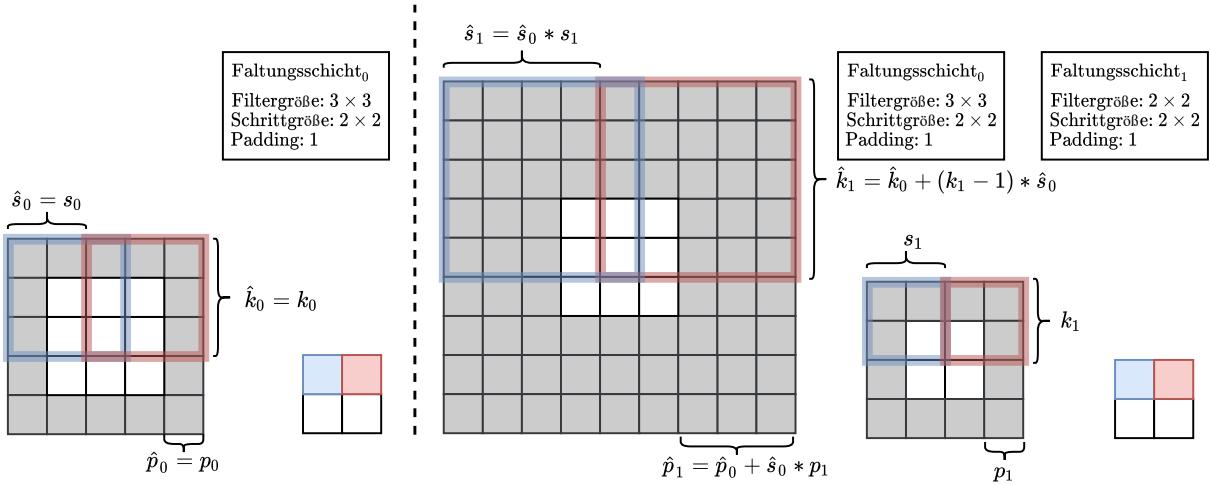


Abbildung 3.4: Berechnung des Einflussbereichs einzelner Ausgaben in der ursprünglichen Eingabe. Links nach einer Faltungsschicht, rechts nach zwei Schichten. Graue Bereiche repräsentiert hinzugefügte Puffer.

Block, ergibt sich für jeden lokalen Deskriptor ein quadratischer Einflussbereich \mathbf{k} mit einer Seitenlänge von 267 Pixeln im Ursprungsbild. Die Verschiebung zwischen Einflussbereichen benachbarter Deskriptoren \mathbf{s} beträgt dabei 16 Pixel. Das Ursprungsbild erhält bis zu dieser Schicht ein effektives Padding \mathbf{p} von 133 Pixeln in jede Richtung. Aufgrund dieser Werte lassen sich die Einflussbereiche für alle $w \times h$ Deskriptoren, die am Extraktionspunkt anfallen, wie folgt berechnen:

$$x_{min}(i, j) = i * \mathbf{s} - \mathbf{p} \quad (3.9)$$

$$x_{max}(i, j) = x_{min}(i, j) + \mathbf{k} \quad (3.10)$$

$$y_{min}(i, j) = j * \mathbf{s} - \mathbf{p} \quad (3.11)$$

$$y_{max}(i, j) = y_{min}(i, j) + \mathbf{k} \quad \text{wobei } 0 \leq i < w \text{ und } 0 \leq j < h \quad (3.12)$$

3.5.3 Deskriptorselektion

Im nächsten Verarbeitungsschritt werden die aussagekräftigsten lokalen Deskriptoren auf Basis der Attention-Werte selektiert. Da während der Deskriptorextraktion unterschiedlich skalierte Bildversionen betrachtet wurden, steht vor der Selektion eine große Anzahl an Deskriptoren zur Auswahl, die teilweise auch stark überlappende Bildbereiche beschreiben. Das kann dazu führen, dass aus einigen Bildbereichen sehr viele Deskriptoren ausgewählt werden, die zwar individuell viel Information über das Bild bereitstellen, in der Summe aber wenig Nutzen haben, da der gleiche Bildbereich vielfach beschrieben wird. Da die Anzahl der selektierten Deskriptoren pro Bild begrenzt ist, kann dies zu einer Verdrängung von Deskriptoren aus anderen wichtigen Bildbereichen führen. Um dem entgegenzuwirken, werden zunächst Deskriptoren, die stark überlappende Bildbereiche beschreiben, mittels Non-Maximum-Suppression ausgesortiert. Als Metrik für die Überlappung wird hierfür die Intersection-over-Union(IoU) zwischen den Einflussbereichen der Deskriptoren berechnet, also das Verhältnis zwischen überlappenden Flächeninhalt zu gemeinsam eingenommenem Flächeninhalt der Einflussbereiche,

$$\text{IoU}(e_1, e_2) = \frac{e_1 \cap e_2}{e_1 \cup e_2} \quad (3.13)$$

wobei e_1 und e_2 Einflussbereiche zweier Deskriptoren sind. Der Non-Maximum-Suppression Algorithmus (siehe Alg. 3.1) erhält als Eingabe eine Liste mit den Einflussbereichen der lokalen Deskriptoren \mathcal{E} , die dazugehörigen Attention-Werte \mathcal{A} , sowie einen Schwellwert T der maximal tolerierten IoU zwischen den Einflussbereichen der Deskriptoren. Zunächst werden die Kandidaten nach Attention-Wert sortiert. Der Einflussbereich mit dem höchsten Attention-Wert wird ausgewählt und von der Kandidatenliste in die Ergebnisliste geschoben. Anschließend wird die IoU zwischen dem ausgewählten Einflussbereich und den übrigen Einflussbereichen der Kandidatenliste berechnet. Kandidaten, deren IoU den Schwellwert T überschreiten, werden aus der Kandidatenliste entfernt. Anschließend wird aus der Kandidatenliste erneut der Einflussbereich mit dem nun höchsten Attention-Wert ausgewählt und der Prozess wiederholt, bis die Kandidatenliste leer ist. Abschließend wird die Ergebnisliste mit den ausgewählten Bereichen und dazugehörigen Attention-Werten zurückgegeben.

Algorithmus 3.1: Non-Maximum-Suppression

```

Input:  $\mathcal{E} \leftarrow \{e_1, \dots, e_n\}$ ,  $\mathcal{A} \leftarrow \{a_1, \dots, a_n\}$ ,  $T$ 
 $\mathcal{S} \leftarrow \emptyset$ 
while  $\mathcal{E} \neq \emptyset$  do
     $x \leftarrow \text{argmax}(\mathcal{A})$  // Wähle bestbewertete Box
     $currBox \leftarrow e_x$ 
     $\mathcal{E} \leftarrow \mathcal{E} \setminus \{e_x\}$  // Verschiebe Box von Kandidaten in Ergebnisliste
     $\mathcal{A} \leftarrow \mathcal{A} \setminus \{a_x\}$ 
     $\mathcal{S} \leftarrow \mathcal{S} \cup \{(e_x, a_x)\}$ 
    foreach  $e_i \in \mathcal{E}$  do
        // Entferne Boxen mit  $\text{IoU} > T$  zu currBox aus Kandidatenliste
        if  $\text{IoU}(e_i, currBox) > T$  then
             $\mathcal{E} \leftarrow \mathcal{E} \setminus \{e_i\}$ 
             $\mathcal{A} \leftarrow \mathcal{A} \setminus \{a_i\}$ 
return  $\mathcal{S}$ 

```

Im DELF-Verfahren wird für die Vorsortierung ein IoU-Schwellwert von 0.8 genutzt. Nach der Vorsortierung findet die finale Selektion der Deskriptoren statt. Für jedes Bild werden hierbei von den verbliebenen Deskriptoren die 1000 Deskriptoren mit den höchsten Attention-Werten ausgewählt.

3.5.4 Datentransformation und Dimensionsreduktion

Ziel des letzten Vorverarbeitungsschrittes ist es, die Daten der lokalen Deskriptoren in eine Form zu bringen, die ein effizientes Vergleichen möglich macht. Obwohl durch den Selektionsprozess bereits die Anzahl an lokalen Deskriptoren je Bild drastisch reduziert wurde, ist die aktuelle Repräsentation der einzelnen Bilder noch sehr groß. Mit 1000 Deskriptoren zu je 1024 Dimensionen umfasst die Beschreibung jedes Bildes über eine Million Werte. Durch die folgenden Transformationen wird die Größe der lokalen Deskriptoren und damit der gesamte Repräsentation stark reduziert. Die für DELF beschriebenen Transformationen bezeichnen die Autoren dabei als "common practice" (vgl. [2] Kap.4.3) und beziehen sich dabei auf das von Jegou und Chum in [28] untersuchte Vorgehen zur Transformation von VLAD und BOW-Deskriptoren. Tatsächlich werden Deskriptoren auch in anderen Image Retrieval-Verfahren [14] [16] auf ähnliche Weise verarbeitet.

Zunächst werden die Deskriptoren der Länge nach normiert. Anschließend wird auf ihnen eine Hauptkomponentenanalyse durchgeführt. Die resultierende Transformationsmatrix soll dabei repräsentativ für die Deskriptoren des gesamten Datensatzes sein, daher werden bei der Analyse Deskriptoren des ganzen Datensatzes oder zumindest eines erheblichen Teils betrachtet. Ziel der Hauptkomponentenanalyse ist es, die Daten anhand neuer Dimensionen, den sogenannten Hauptkomponenten, darzustellen. Die Richtungen der Hauptkomponenten ergeben sich aus einer Linearkombination der bisherigen Dimensionen und sind so gewählt, dass für die transformierten Daten keine Korrelation zwischen den Dimensionen besteht. Eine weitere Eigenschaft dieser Darstellung ist es, dass die einzelnen Hauptkomponenten jeweils den größtmöglichen Anteil der in den Daten vorhandenen Varianz abbilden. D.h. die Hauptkomponente, die den größten Teil der Varianz abbildet, zeigt in die Richtung, entlang welcher die Daten am stärksten variieren. Die Hauptkomponente, die den nächst größeren Anteil der Varianz erklärt, beschreibt den größtmöglichen Anteil der verbleibenden Varianz. Dies führt dazu, dass der wesentliche Teil der Varianz von wenigen Dimensionen erklärt wird. Dadurch ist es möglich, einen großen Teil der Dimensionen zu entfernen, ohne einen starken Informationsverlust zu erleiden.

Sei $X_{n \times d}$ eine Matrix von n zu analysierenden Deskriptoren mit je d Dimensionen. Zunächst werden die Daten zentriert, sodass ihr Mittelwert entlang der d Dimensionen 0 ist:

$$x'_{i,j} = x_{i,j} - \frac{1}{n} \sum_{k=1}^n x_{k,j} \quad \text{für } 0 < i \leq n, 0 < j \leq d \quad (3.14)$$

Häufig werden die Daten anschließend durch die Standardabweichung der einzelnen Dimensionen geteilt, um die Varianz je Dimension auf 1 zu setzen. Je größer die Varianz in den Dimensionen ist, desto größer ist ihr Anteil in den berechneten Hauptkomponenten. Diese Standardisierung wird daher durchgeführt, wenn die Größe der Varianzen der ursprünglichen Dimensionen nicht adäquat ihre Bedeutsamkeit widerspiegeln. Für DELF wird keine Anpassung der Varianz durchgeführt.

Nach der Hauptkomponentenanalyse werden die Daten in den Raum der Hauptkomponenten transformiert. Die neuen Dimensionen sind dabei nach der von ihnen erklärten Varianz sortiert (vgl. Abb. 3.5ab).

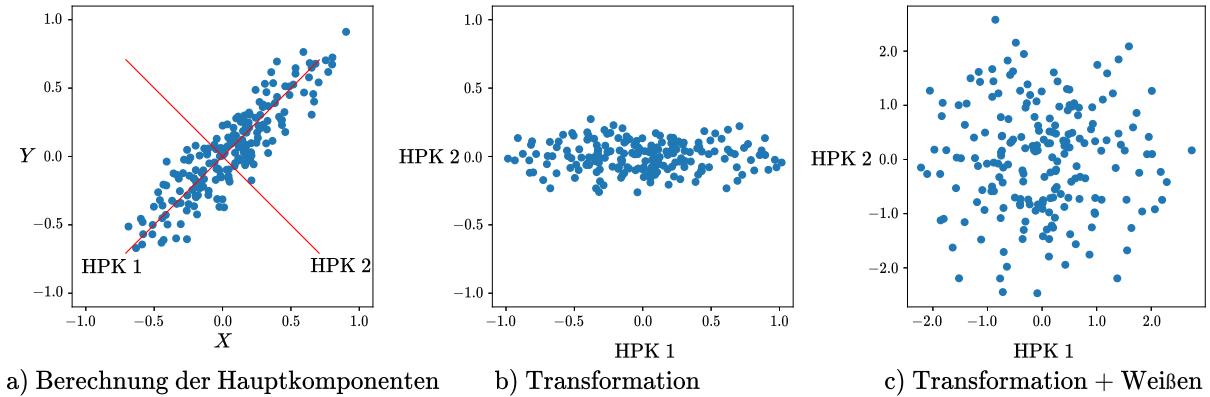


Abbildung 3.5: Beispiel zur Hauptkomponentenanalyse auf 2D-Daten

DELF erhält von den transformierten Deskriptoren nur die ersten 40 Dimensionen und erzeugt damit eine deutlich kompaktere Repräsentation der Bilder. Wie sich die Anzahl erhaltener Dimensionen auf das Retrievalergebnis auswirkt, wird im Rahmen der vorliegenden Arbeit experimentell untersucht (vgl. Kap. 4.3.2, S. 33). Weiterhin sieht das Verfahren vor, die Deskriptorendaten zu weißen, d.h. in eine Form zu bringen, in der zwischen den Dimensionen keine Korrelation herrscht und die Varianz entlang der Dimensionen jeweils 1 beträgt (vgl. Abb. 3.5c). Die erste Voraussetzung ist dabei durch die Transformation in den Raum der Hauptkomponenten automatisch erfüllt. Für die Anpassung der Varianz müssen die Daten durch die Standardabweichungen der neuen Dimensionen geteilt werden. Bevor die Deskriptoren für das Matching genutzt werden können, werden sie erneut der Länge nach normiert.

3.6 Matching

Sobald die Daten der lokalen Deskriptoren des Datensatzes erhoben, verarbeitet und in den Bildern lokalisiert sind, ist das DELF-System in der Lage, Suchanfragen zu bearbeiten. Wird ein Bild als Anfrage an das System gestellt, werden auch für dieses Bild alle notwendigen Informationen berechnet, wie im letzten Abschnitt beschrieben. Um die Anfrage zu beantworten, müssen die Bilder im Suchdatensatz anhand ihrer Ähnlichkeit zum Anfragebild sortiert werden. Hierfür wird jede mögliche Kombination an Bildpaaren aus der Anfrage und dem Suchdatensatz einzeln untersucht. Zunächst wird dabei ein initiales Matching zwischen den Deskriptoren des Anfragebildes D_A und den Deskriptoren des aktuell betrachteten Bildes D_S aus dem Suchdatensatz hergestellt. Für jeden lokalen Deskriptor d_{Ai} wird dabei der ähnlichste Deskriptor in D_S gesucht⁶. Als Metrik wird hierfür die euklidische Distanz berechnet. Ist die Distanz zwischen d_{Ai} und dem ähnlichsten Deskriptor in D_S kleiner als ein Schwellwert T , werden die Deskriptoren als potentielles Match vermerkt (vgl. Alg. 3.2). DELF nutzt für T standardmäßig einen Wert von 0.8.

Anschließend wird überprüft, ob sich die gefundenen Matches auch geometrisch erklären lassen. Die Annahme dabei ist, dass es eine affine Transformation vom Anfragebild auf das betrachtete Bild der

⁶In der offiziellen Implementierung wird für die effiziente Suche der ähnlichsten Deskriptoren ein k-d-Baum [29] über die Deskriptoren des Anfragebildes erstellt. Eine Effizienzsteigerung mit k-d-Bäumen lässt sich jedoch nur erzielen, solange ein angemessenes Verhältnis zwischen der Anzahl der Datenpunkte n und Dimensionalität d der Daten besteht ($n >> 2^d$) [30], was in diesem Anwendungsfäll nicht gegeben ist.

Algorithmus 3.2: Initiales Deskriptor Matching

```

Input:  $D_A \leftarrow \{d_{A1}, \dots, d_{An}\}$ ,  $D_S \leftarrow \{d_{S1}, \dots, d_{Sn}\}$ ,  $T$ 
 $\mathcal{M} \leftarrow \emptyset$ 
foreach  $d_{Ai} \in D_A$  do                                // Für jeden Deskriptor in  $A$ 
     $closestDistance \leftarrow \infty$ 
     $closestDescriptor \leftarrow -1$ 
    foreach  $d_{Sj} \in D_S$  do                  // Finde ähnlichsten Deskriptor in  $S$ 
        if  $\|d_{Ai} - d_{Sj}\|_2 < closestDistance$  then
             $closestDistance \leftarrow \|d_{Ai}, d_{Sj}\|_2$ 
             $closestDescriptor \leftarrow j$ 
    if  $closestDistance < T$  then                // Falls näher als  $T$ 
         $\mathcal{M} \leftarrow \mathcal{M} \cup \{(i, closestDescriptor)\}$  // Füge potentielles Match hinz
    return  $\mathcal{M}$ 

```

Suchdatenbank geben muss, welche die Deskriptorpositionen im Anfragebild auf die Positionen der gefundenen Matches im Suchbild projiziert, falls das betrachtete Bildpaar tatsächlich den gleichen Bildinhalt zeigt.

Für die geometrische Verifikation müssen die Deskriptoren der Matches Koordinaten in ihren dazugehörigen Bildern zugeordnet werden. Hierfür werden die zuvor berechneten Einflussbereiche der Deskriptoren (siehe Kap. 3.5.2 auf Seite 14) genutzt. Die Zentren der Einflussbereiche dienen dabei als Referenzpunkte für die Deskriptoren. Für die Verifikation wird der RANSAC-Algorithmus (RANdom SAmple Consensus) [31] von Fischler und Bolles genutzt. Als Eingabe erhält der Algorithmus eine Liste mit den Positionen der Deskriptoren aus den potentiellen Matches \mathcal{L} , einen Schwellwert T_{in} , der festlegt, wie weit die projizierten Positionen maximal von den tatsächlichen Positionen ihres Matches entfernt sein dürfen, um noch von einer Transformation erklärt zu werden und die Anzahl an Versuchen $numTrials$, die unternommen werden, um eine passende Transformation zu finden.

Zu Beginn werden zufällig Matchpaare aus \mathcal{L} gewählt, mit denen sich eine affine Transformation zwischen den Bildern bestimmen lässt. Für eine affine Ebenentransformation benötigt man drei Koordinatenpaare, um eine Transformationsmatrix bestimmen zu können. Nun wird überprüft, welche Matches sich von dieser Transformation erklären lassen. Sind die Abstände der transformierten Deskriptorpositionen des Anfragebildes zu ihren Matches im aktuellen Suchbild innerhalb des Schwellwerts T_{in} , werden sie als Inlier bezeichnet, d.h. sie können von der Transformation erklärt werden. Dieser Prozess wird mehrfach wiederholt, wobei jedes Mal zufällige Matchpaare für die Berechnung der Transformationsmatrizen genutzt werden. Nach $numTrials$ Versuchen gibt der Algorithmus die Transformation mit den dazugehörigen Inliern zurück, welche die meisten Matches erklären (vgl. Alg. 3.3). Für DELF werden standardmäßig 1000 RANSAC-Versuche durchlaufen, um eine optimale Transformation zu finden. Die Schwellwertdistanz T_{in} beträgt dabei 20.

DELF nutzt die Anzahl erklärter Deskriptormatches als Metrik für die Ähnlichkeit zwischen Bildpaaren. So können die Bilder der Suchdatenbank anhand ihrer Ähnlichkeit sortiert und dem Nutzer zurückgegeben werden. In Kapitel 4.3.3 ab Seite 38 wird die experimentelle Betrachtung von Metriken zur Bestimmung von Ähnlichkeiten zwischen Bildern betrachtet.

Algorithmus 3.3: RANSAC

```

Input:  $\mathcal{L} \leftarrow \{(l_{A1}, l_{S1}), \dots, (l_{Ak}, l_{Sk})\}, T_{in}, numTrials$ 
 $bestModel \leftarrow \text{None}$ 
 $bestInliers \leftarrow \emptyset$ 
for  $i = 0, i < numTrials, i++ \text{ do}$ 
    // Berechne affine Transformation aus zufälligen Matchpaaren
     $inliers \leftarrow \emptyset$ 
     $selectedMatches \leftarrow \text{drawRandomMatches}(\mathcal{L}, pairsNeeded = 3)$ 
     $model \leftarrow \text{calcModel}(selectedMatches)$ 
    foreach  $(l_{Aj}, l_{Sj}) \in \mathcal{L} \text{ do}$ 
        // Sammle von Transformation erklärte Paare
         $transformedLocation \leftarrow \text{transform}(l_{Aj}, model)$ 
        if  $\|l_{Sj} - transformedLocation\|_2 < T_{in}$  then
             $inliers \leftarrow inliers \cup \{(l_{Aj}, l_{Sj})\}$ 
    if  $\text{len}(inliers) > \text{len}(bestInliers)$  then
         $bestInliers \leftarrow inliers$ 
         $bestModel \leftarrow model$ 
return  $bestInliers, bestModel$       // Gib bestes Modell + Inlier zurück

```

3.7 Verfahrensunterschiede im DELF Artikel

Die Verfahrensbeschreibung aus dem zu DELF erschienen Artikel (vgl. [2] Kap.4 und Kap 5.1) weist einige Unterschiede zu dem dazugehörigen veröffentlichten Quellcode auf. Teilweise sind diese Unterschiede der Kürze des Artikels geschuldet. Verfahrensschritte werden meist nur oberflächlich beschrieben, was einen gewissen Interpretationsspielraum für die tatsächliche Umsetzung zulässt. Einige Schritte, wie beispielsweise die Verwendung von Non-Maximum-Suppression zur Vorsortierung bei der Auswahl der lokalen Deskriptoren (siehe Kap. 3.5.3 auf Seite 16), die für das generelle Verständnis des Verfahrens nicht essentiell sind, werden im Artikel nicht erwähnt.

Wesentliche Unterschiede gibt es in der Strukturierung der extrahierten Deskriptoren des Suchdatensatzes sowie in der Verarbeitung von Suchanfragen. Das im Artikel beschriebene Vorgehen nutzt dabei einen invertierten Index in Kombination mit Produkt-Quantisierung [32], um lokale Deskriptoren effizient zu matchen. Dieser Ansatz ist insbesondere für die Suche auf sehr großen Datensätzen besser geeignet, als das Vorgehen im veröffentlichten Quellcode. Für die vergleichsweise überschaubaren historischen Datensätze, mit denen sich die vorliegende Arbeit beschäftigt, stellt dies jedoch kein Problem dar.

Im Artikel werden beim Beantworten einer Anfrage jeweils nicht nur die Deskriptoren eines Bildpaares, sonder alle Deskriptoren des Suchdatensatzes betrachtet. Zunächst wird ein invertierter Index mit einem Codebuch von 8k verschiedenen visuellen Wörtern über die lokalen Deskriptoren erstellt. Das heißt, die lokalen Deskriptoren werden mit Hilfe des K-Means-Algorithmus[23] jeweils einem der 8k Clusterzentren zugeordnet. In der zum Cluster gehörigen Postingliste wird dann ein neuer Eintrag erstellt, der Informationen über die Beschaffenheit des Deskriptors und des Bildes, aus dem er extrahiert wurde, enthält. Hierfür wird das Residuum $r(x)$ zwischen dem betrachteten Deskriptor x und dem ihm zugeordneten Cluster-Zentrum $q(x)$ berechnet.

$$r(x) = x - q(x) \quad (3.15)$$

Der 40-dimensionale Residuenvektor wird anschließend mit Hilfe eines Produkt-Quantisierers auf einen 50-bit-Code abgebildet. Der Vektor wird dafür in m Subvektoren aufgeteilt. Die Subvektoren werden dann mittels K-Means-Algorithmus einem von k Clusterzentren zugeordnet. Der Code ergibt sich aus den konkatenierten Indizes der zugeordneten Clusterzentren. Im Fall von DELF werden die Residuenvektoren in 10 Subvektoren zerlegt, die jeweils einem von 2^5 Clusterzentren zugeordnet werden. Der erzeugte Code bildet die Information über den Deskriptor, welche in den Postinglisten hinterlegt wird. Bei der Verarbeitung einer Anfrage werden die lokalen Deskriptoren der Anfrage zunächst wieder quantisiert und Postinglisten zugeordnet. Anschließend werden ihre Residuenvektoren mit den Einträgen der Postinglisten verglichen. Hierfür werden die 50-bit-Codes wieder über ihre zugehörigen Clusterzentren repräsentiert und die euklidische Distanz der Zentren zum Residuenvektor des Anfragedeskriptors berechnet (vgl. Abb. 3.6). Auf diese Weise findet das System für jeden lokalen Anfragedeskriptor die nächsten 60 Deskriptoren aus dem Suchdatensatz. Diese Matches werden pro Bild akkumuliert. Anschließend findet eine geometrische Verifikation der Matches mit Hilfe des RANSAC-Algorithmus statt. Die Anzahl an verifizierten Matches bildet auch hier die Grundlage für die Bewertung der Ähnlichkeit zwischen Bildern.

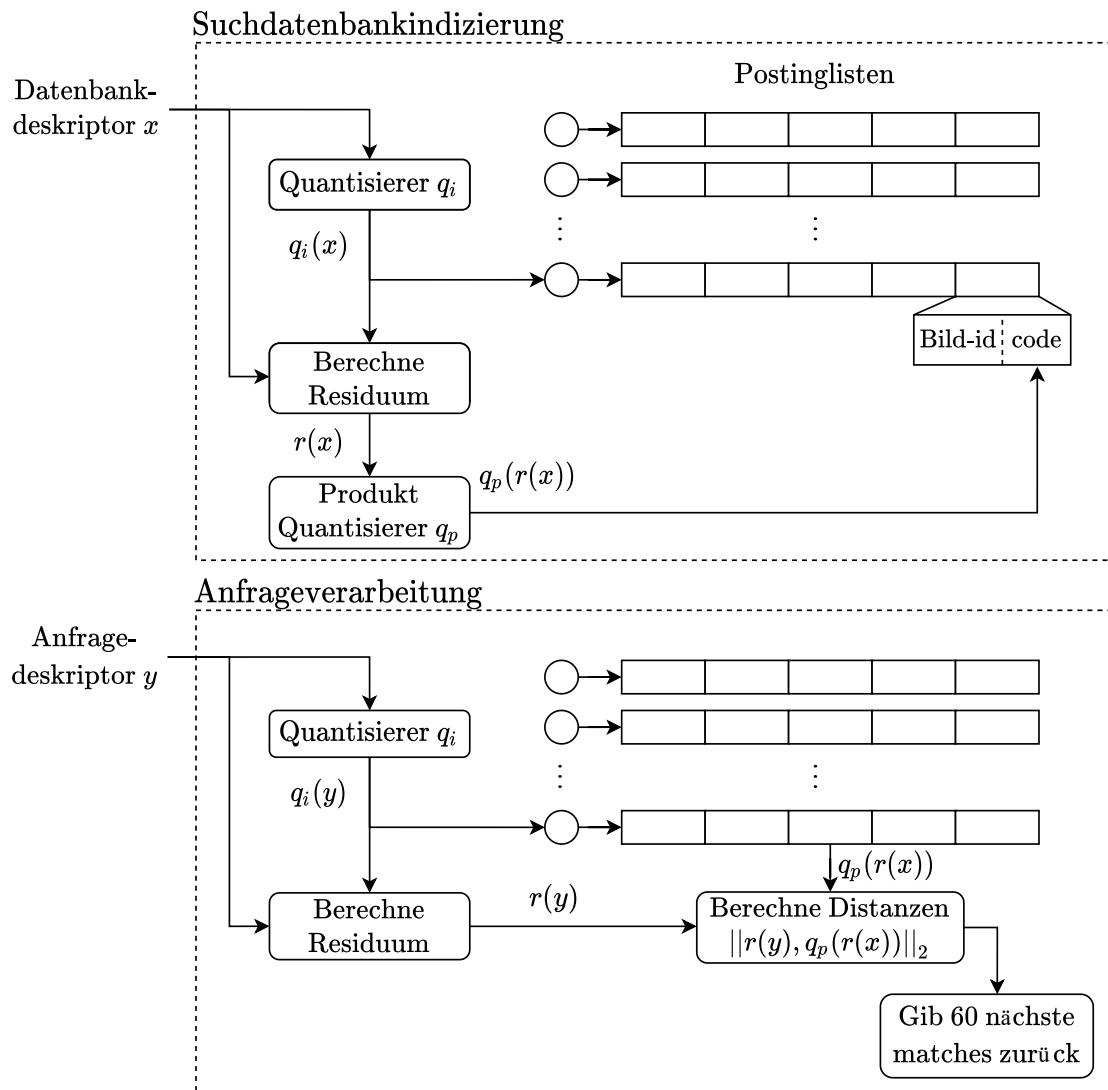


Abbildung 3.6: Invertierter Index und Produkt-Quantisierung zur Anfrageverarbeitung (vgl. [32] Fig.5)

4 Evaluation

Die Untersuchung des DELF-Verfahrens lässt sich in zwei wesentliche Abschnitte unterteilen. Zunächst werden in breit angelegten Experimentalreihen unterschiedliche Parameter innerhalb der DELF-Pipeline variiert, um ihren Einfluss auf die Ergebnisse der gelösten Retrievalaufgaben zu quantifizieren. Ziel ist dabei, sowohl ein besseres Verständnis für die Einflüsse einzelner Parameter zu schaffen, wie auch eine optimale Konfiguration für die Lösung von Retrievalaufgaben speziell für historische Bilder zu finden. Im zweiten Abschnitt der Evaluation werden die Entscheidungen des optimierten DELF-Modells qualitativ untersucht, um ein besseres Verständnis für die Entscheidungsfindung des DELF-Prozesses zu erlangen. Insbesondere wird hier DELF mit anderen Retrievalverfahren verglichen, um Schwächen und Stärken des Verfahrens aufzudecken.

Eine Untersuchung zur Reproduzierbarkeit der Ergebnisse aus dem DELF-Papier [2] findet in der vorliegenden Arbeit nicht statt, da die Autoren für mehreren Parametern innerhalb der DELF-Pipeline keine Angaben zu verwendeten Werten machen. Der von den Autoren verwendete Datensatz für das Training ist außerdem um ein Vielfaches größer, als der in der vorliegenden Arbeit verwendete Trainingssatz. Ein Training auf so vielen Daten ist im Rahmen der vorliegenden Arbeit nicht umsetzbar. Weiterhin fehlen Informationen, wie die Autoren den von ihnen betrachteten Benchmark-Datensatz (Oxford5k) evaluiert haben. Dabei wird nicht erläutert, wie mit den sogenannten Störbildern des Datensatzes umgegangen wird¹, was einen Vergleich von Retrievalergebnissen unmöglich macht.

4.1 Evaluationsdaten

Zur Bewertung von Retrievalsystemen auf historischen Bildern steht ein eigens erstellter Datensatz zur Verfügung (vgl. Tab. 4.1). Die 848 zusammengestellten Bilder umfassen historische Abbildungen von sieben Dresdner Sehenswürdigkeiten und entstammen vorwiegend den Archiven der deutschen Fotothek². Auf Grund der geographischen Nähe einiger Sehenswürdigkeiten sind auf manchen Bildern des Datensatzes mehrere dieser Objekte zu sehen. Bilder sollten von einem Retrievalsystem immer dann zurückgegeben werden, wenn sie mindestens ein Objekt enthalten, welches im Anfragebild zu sehen ist. Bei den Anfragebildern wird darauf geachtet, dass sie immer genau eines der sieben Sehenswürdigkeiten abbilden.

Vorabexperimente haben gezeigt, dass die von DELF durchgeführte Deskriptorselektion nur befriedigende Ergebnisse liefert, wenn die verwendeten Bilder eine Mindestgröße haben. Im Selektionsschritt wird eine feste Anzahl an Deskriptoren vom Attention-Netzwerk ausgewählt. Da die Anzahl an extrahierten Deskriptoren je Bild mit der Bildgröße skaliert kann es bei sehr kleinen Bildern passieren, dass das Attention-Netzwerk gezwungen ist, alle, oder ein Großteil der extrahierten Deskriptoren auszuwählen. In diesem Fall findet keine bedeutsame Auswahl an Deskriptoren statt und der positive Effekt des

¹Mehr Informationen zu Störbildern finden sich im letzten Absatz des folgenden Subkapitels.

²<https://www.slub-dresden.de/sammlungen/deutsche-fotothek/>, zuletzt besucht am 02.09.20

Objekt	Zwinger	Hofkirche	Frauenkirche	Semperoper
Anzahl Objektiressionen	374	216	206	89
Objekt	Sophienkirche	Stallhof	Moritzburg	Total
Anzahl Objektiressionen	66	38	23	1012
Anzahl Anfragen	6	4	4	42
Impressionen pro Bild	0	1	2	5
Anzahl Bilder	0	730	81	1
				Total 848

Tabelle 4.1: Aufbau des historischen Datensatzes

Selektionsprozesses geht verloren (siehe Abb. 4.1). Bei sehr großen Bildern können während des Ex-

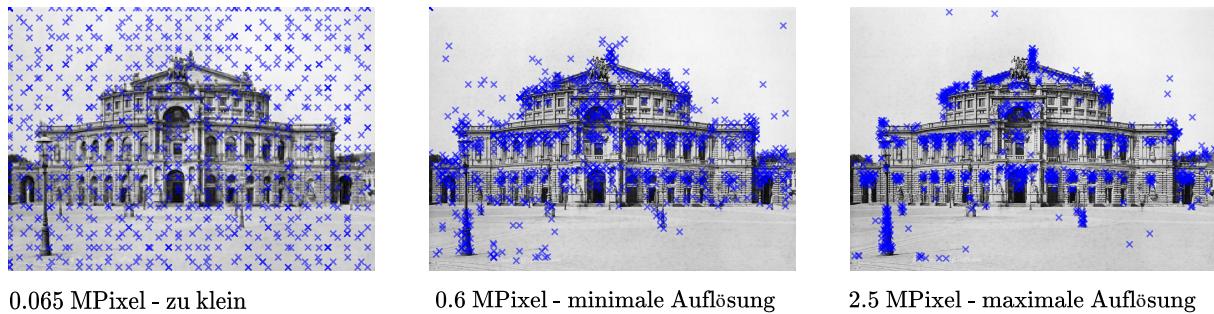


Abbildung 4.1: Referenzpunkte der ausgewählten Deskriptoren bei unterschiedlicher Eingangsauflösung.

traktionsprozesses Speicherprobleme auftauchen. Aus diesem Grund werden nur Bilder verarbeitet, die eine Auflösung zwischen 0.6MPixel und 2.5MPixel aufweisen. Bilder die diese Restriktionen über- bzw. unterschreiten, werden vor der Verarbeitung auf die Maximal- bzw. Mindestgröße skaliert.

Neben historischen Daten wird das DELF-Verfahren zusätzlich auf dem Oxford5k-Datensatz [3] getestet (vgl. Tab. 4.2). Hierbei handelt es sich um einen häufig verwendeten Benchmarkdatensatz, bestehend aus 5063 Bildern. Das Bildmaterial entstammt Suchergebnissen zu 11 unterschiedlichen Sehenswürdigkeiten in und um Oxford, aus der Fotocommunity Flickr³. Häufig finden sich dabei Bilder von Personen oder Aufnahmen von Innenräumen, die keine der gesuchten Sehenswürdigkeiten abbilden. Diese Störbilder können entweder als zusätzliche Herausforderung angesehen oder aber vorab aussortiert werden. In den in der vorliegenden Arbeit durchgeführten Experimenten sind diese Störbilder im Datensatz enthalten.

4.2 Retrievalmetriken

Obwohl sich Klassifikations- und Retrievalaufgaben im Kern ähneln, können viele Metriken mit denen Klassifikationssysteme üblicherweise bewertet werden, wie beispielsweise Genauigkeit, nicht verwendet werden, um die Performanz von Retrievalsystmen zu evaluieren. Generell lassen sich aus der Betrachtung einzelner Bildpaare zwischen Anfragebildern und Bildern des Suchindexes keine Rückschlüsse über die Performanz von Retrievalsystmen ziehen. Grundlage der Bewertung sind stets die gesamten Antworten des Retrievalsystms auf eingehende Suchanfragen. Entscheidend sind hierbei die Rankings, bzw.

³<https://www.flickr.com/>, zuletzt besucht am 03.09.20

Objekt	Radcliffe Camera	Christ Church	All Souls	Magdalen
Anzahl Objektimpressionen	348	133	111	103
Anzahl Anfragen	5	5	5	5
Objekt	Hertford	Ashmolean	Bodleian	Balliol
Anzahl Objektimpressionen	61	31	30	18
Anzahl Anfragen	5	5	5	5
Objekt	Cornmarket	Keble	Pitt Rivers	Total
Anzahl Objektimpressionen	13	11	8	867
Anzahl Anfragen	5	5	5	55
Impressionen pro Bild	0	1	2	Total
Anzahl Bilder	4218	823	22	5063

Tabelle 4.2: Aufbau des Oxford5k Datensatzes

die Reihenfolgen in der die Bilder des Suchindexes auf die Anfragen zurückgegeben werden. Eine Möglichkeit diese Reihenfolgen zu bewerten ist die Erstellung sogenannter ROC-Kurve (Reciever-Operating-Characteristic)(vgl. Abb. 4.2a). Hierfür wird jeweils eine wachsender Anteil der zurückgegebenen Bilder betrachtet und das Verhältnis zwischen Richtig-Positiv-Rate bzw. Recall und Falsch-Positiv-Rate dargestellt. Der Recall gibt an, welcher Anteil an Bildern mit gewünschten Bildinhalt bereits im betrachteten Abschnitt der Rückgabe enthalten war.

$$\text{Recall} = \frac{|\text{Bereits zürückgebene Bilder mit gewünschtem Inhalt}|}{|\text{Im Datensatz enthaltene Bilder mit gewünschtem Inhalt}|} \quad (4.1)$$

Analog beschreibt die Falsch-Positiv-Rate den Anteil der Bilder ohne gewünschten Bildinhalt, der bereits zurückgegeben wurde.

$$\text{Falsch-Positiv-Rate} = \frac{|\text{Bereits zürückgebene Bilder ohne gewünschtem Inhalt}|}{|\text{Im Datensatz enthaltene Bilder ohne gewünschtem Inhalt}|} \quad (4.2)$$

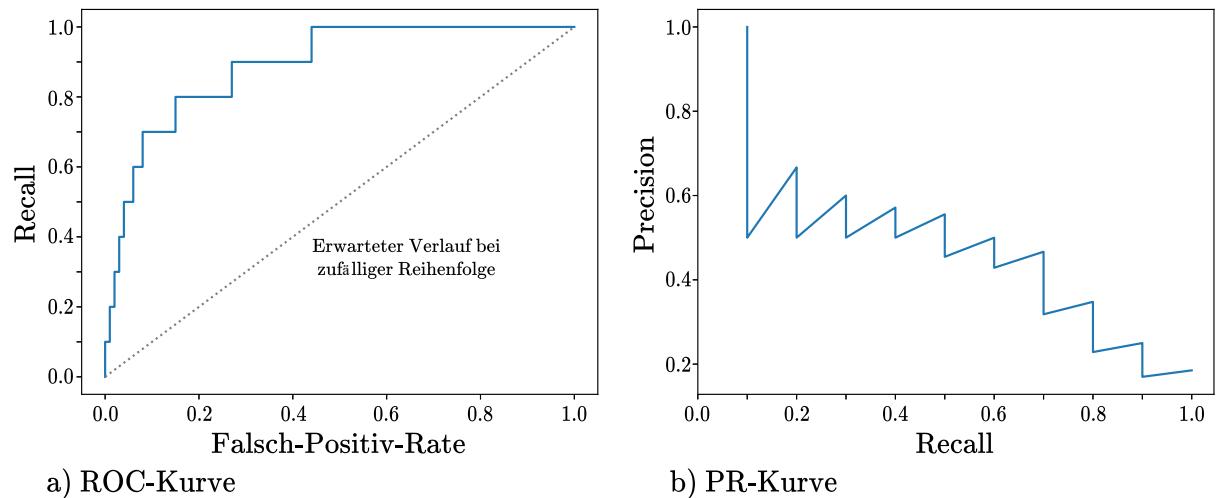


Abbildung 4.2: Vergleich von Kurvenmetriken einem Beispielsranking

ROC-Kurven beginnen stets im Ursprung, wo noch keine Bilder zurückgegeben wurden und enden im Punkt (1, 1) wo der gesamte Datensatz zurückgegeben wurde. Die Gerade zwischen diesen Punkten bildet den erwarteten Verlauf, wenn das System Bilder in einer zufälligen Reihenfolge zurückgibt. Ein Retrievalssystem hat nur dann einen positiven Effekt für den Nutzer, wenn die ROC-Kurven ihrer Anfrageantworten oberhalb dieser Linie verlaufen. Obwohl ROC-Kurven gut in der Lage sind den Unterschied eines Suchsystems gegenüber einer zufälligen Suche aufzuzeigen, vermitteln sie aus der Perspektive eines tatsächlichen Anwenders oft ein zu positives Bild der Ergebnisse. Das liegt daran, dass der Anteil der Bilder, die für eine Anfrage relevant sind meist um ein Vielfaches kleiner ist als der Anteil der irrelevanten Bilder, was an einer ROC-Kurve jedoch nicht ablesbar ist. Angenommen für eine Suchanfrage sind 10% eines durchsuchten Datensatzes tatsächlich relevant und die zur Anfrageantwort erstellte ROC-Kurve zeigt bei einem Recall von 90% eine Falsch-Positiv-Rate von 30%. Obwohl der Verlauf der Kurve ein sehr gutes Ergebnis suggeriert, bedeutet dies für den Nutzer, dass 75% der zurückgegebenen Bilder, die er gesehen hat, bevor ein Recall von 90% erreicht wird, nicht den gesuchten Bildinhalt zeigen. Eine Möglichkeit, die tatsächliche Nutzererfahrung besser abzubilden ist die Erstellung von sogenannten PR-Kurven (Precision-Recall) (vgl. Abb. 4.2b). Hierbei wird die Präzision, also der Anteil der relevanten Bildern in der bisherigen Rückgabe, im Verhältnis zum Recall abgebildet.

$$\text{Precision} = \frac{|\text{Bereits zurückgebene Bilder mit gewünschtem Inhalt}|}{|\text{Bereits zurückgebene Bilder}|} \quad (4.3)$$

So kann direkt abgelesen werden, welchen Anteil an irrelevanten Bildern innerhalb der Rückgabe toleriert werden müssen, um einen bestimmten Anteil der gesuchten Bilder zu finden. Die vorgestellten Kurvenmetriken eignen sich gut um einzelne Anfragen an Retrievalssysteme zu analysieren und zu visualisieren. Um Retrievalssysteme in Gänze, auf Basis mehrerer Anfrage zu bewerten und mit anderen Systemen zu vergleichen, bietet es sich jedoch an ein kompaktere Metriken zu verwenden. Hierfür wird können die Flächen unterhalb der Metrikkurven betrachtet werden. AUC (Area Under Curve) und AP (Average Precision) approximieren jeweils die Flächen unterhalb von ROC bzw. PR-Kurven und können so eine Anfrageantwort mit einer einzelnen Zahl bewerten. Auf Grund der besseren Beschreibung des Nutzererlebnisses werden in der vorliegenden Arbeit PR-Kurven bzw. AP-Werte zur Auswertung genutzt. Für den Vergleich zwischen unterschiedlichen Retrievalssystemen oder Konfigurationen von DELF werden die AP-Werte von mehreren Anfragen an ein System zu einem Mittelwert (mAP) zusammengefasst. So kann die Performanz eines Retrievalssystems in einer einzelnen Zahl dargestellt werden.

4.3 Parameteranalyse

Ein wesentliches Ziel der vorliegenden Arbeit ist es das DELF-Verfahren insbesondere für den Anwendungsfall des Retrievals von historischen Abbildungen zu optimieren. Hierfür werden ein Reihen an Parametern entlang der DELF-Pipeline variiert und ihre Einflüsse, auf die Retrievalergebnisse analysiert. Um belastbare Aussagen über die Einflüsse einzelner Parameter machen zu können, muss eine große Anzahl unterschiedlicher Parameterkonfigurationen betrachtet und Experimente diesbezüglich mehrfach wiederholt werden. Die dafür benötigte Rechenleistung ist sehr groß, weshalb sich die Experimente innerhalb des Zeitrahmens einer Masterarbeit nicht auf einem normalen System durchführen lassen.

Die Experimente werden daher auf dem HPC-DA System⁴ des Zentrums für Informationsdienste und Hochleistungsrechnen (ZIH), der TU Dresden durchgeführt. Die verwendete Partition für maschinelles Lernen besteht aus 32 Knoten, mit jeweils 6 NVIDIA VOLTA V100 GPUs und 2 IBM Power9 CPUs mit je 22 Kernen. Die große Anzahl an leistungsstarken Rechenknoten erlaubt es viele unterschiedliche Konfigurationen parallel und zügig zu testen. In Tab.4.3 findet sich eine Übersicht der rechenintensiven Experimentalreihen, die im Zuge der vorliegenden Arbeit durchgeführt wurden, mit Informationen über verwendete Hardware und benötigter Rechenzeit.

Experiment	#Läufe	#CPU-Kerne	#GPUs	Totale Laufzeit (h)	Totale CPU-Zeit (h)	Totale GPU-Zeit (h)
Hyperparameteroptimierung Finetuning	1	10	10	16	160	160
Hyperparameteroptimierung Attention-Training	2	48	12	40	1920	480
Modelltraining	12	2 – 3	1	78	226	78
Retrievereexperimente	919	30	1	2027	60834	2027
Summe				2325	63140	2745

Tabelle 4.3: Übersicht über rechenintensive Experimente, durchgeführt auf der ML-Partition des HPC-DA Systems. Rechenzeiten sind jeweils über alle Läufe einer Experimentalreihe summiert. CPU-Zeit bezieht sich auf die akkumulierte Rechenzeit der CPU-Kerne.

4.3.1 Hyperparameteroptimierung der Trainingsphasen

Die ersten Experimentalreihen befassen sich mit den Trainingsphasen des DELF-Verfahrens. Um bei späteren Retrievalversuchen gute Ergebnisse erzielen zu können, benötigt man Modelle die in der Lage sind aussagekräftige Bildrepräsentationen zu erstellen. Für die Experimente zum Modelltraining wird dabei angenommen, dass die Güte, der von einem Modell erzeugten Deskriptoren bzw. der von einem Modell getroffenen Auswahl an Deskriptoren positiv mit der Fähigkeit, der Modelle korreliert die beim Training gestellte Klassifikationsaufgaben zu lösen. Um ein Modell zu bewerten wird daher der Fehler, in Form der Kreuzentropie betrachtet, der auftritt wenn das Modell einen Validierungsdatensatz klassifiziert. Vor Beginn jedes Experiments werden zufällig 20% der Trainingsbilder (siehe Kap. 3.2, S. 11) für die Validierung ausgewählt. Die Validierungsdaten stehen dem Modell während des Trainingsprozesses nicht zur Verfügung. Daher können sie genutzt werden, um zu überprüfen, ob ein Modell auch auf ungesuchten Daten vergleichbare Ergebnisse erzielt.

Der Erfolg des Trainingsprozesses hängt wesentlich von der zu trainierenden Architektur und der vorhandenen Datenlage ab. Einfluss haben außerdem Parameter, die den Ablauf des Trainingsprozesses beeinflussen. Die Experiments, die in diesem Abschnitt besprochen werden befassen sich mit der Suche nach optimalen Werten für drei dieser sogenannten Hyperparameter. Betrachtet wird die Anzahl der Trainingsepochen, die Lernrate, die bestimmt wie stark Netzwerkparameter in einem Optimierungsschritt angepasst werden, sowie der Faktor γ mit dem die Lernrate alle 10 Epochen multipliziert wird⁵. Die übrigen Hyperparameter sind für alle Experimente fest gesetzt. Die Modelle werden mit einer Batchgröße von 8 trainiert und mittels Stochastic Gradient Decent (kurz SGD⁶) optimiert. Für die zu untersuchenden Hyperparameter werden jeweils Wertebereiche definiert. Die Länge des Trainings kann zwischen 10 und

⁴<https://tu-dresden.de/zih/hochleistungsrechnen/hpc>

⁵Als weiterer Hyperparameter wird weight decay untersucht, allerdings lässt sich hier kein signifikanter Einfluss auf den Validierungsfehler feststellen. Die Ergebnisse hierzu finden sich im Anhang ab Seite ??

⁶<https://pytorch.org/docs/stable/optim.html#torch.optim.SGD>, zuletzt besucht am 30.07.20

40 Epochen betragen. Die initiale Lernrate wird zwischen 0.01 und 0.001 gewählt und der γ -Faktor liegt im Bereich zwischen 1 und 0.1.

Um den Suchraum effizient erkunden zu können wird das von Norman Koch entwickelte NNOPT-Tool verwendet, um neue Hyperparameterkonfigurationen zu erstellen und auf dem HPC-System zu testen. Das NNOPT-Tool, welches auf dem Hyperopt-Paket [33] von Bergstra, Yamins und Cox basiert, wählt automatisch Werte für die zu untersuchenden Hyperparameter innerhalb der definierten Wertebereiche aus und startet mit diesen Experimenten auf dem Großrechner. Ist ein Experiment abgeschlossen erhält NNOPT zur Bewertung der Konfiguration den erzielten Validierungsfehler. Neue Parameterkonfigurationen werden bevorzugt in der Nähe von bereits getesteten Konfigurationen erstellt, die gute Ergebnisse erzielt haben. Dies erlaubt es schneller optimale Hyperparameterwerte zu finden, als mit einer zufälliger Suche.

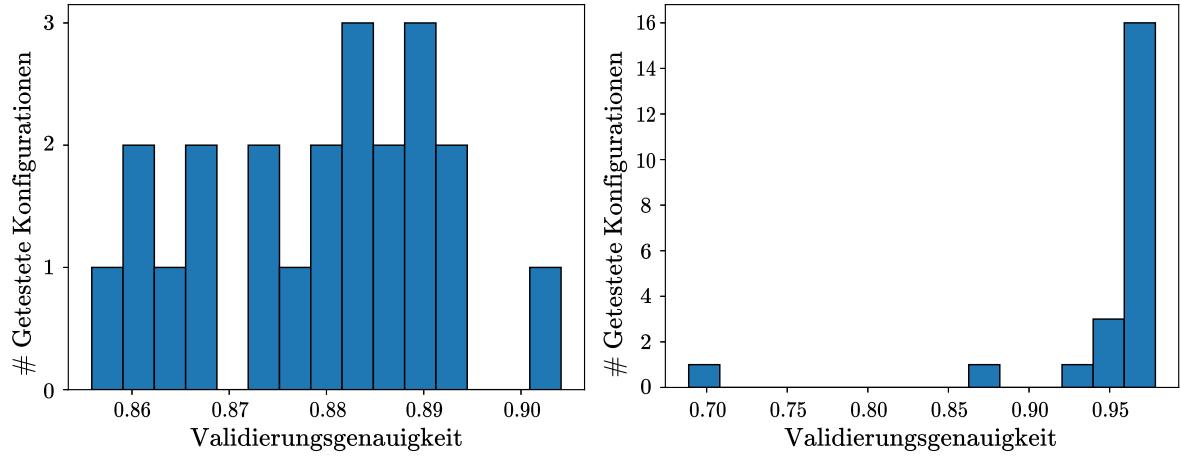
Für das Fine-Tuning wurden auf diese Art 22 unterschiedliche Konfigurationen getestet. Es lässt sich beobachten, dass fast alle getesteten Konfigurationen Modell erzeugen, die in der Lage sind die Trainingsaufgabe sehr gut zu lösen. Im Schnitt erzielen die trainierten Modelle eine Klassifikationsgenauigkeit von 95.13% auf den Validierungsdaten⁷, wobei 20 der 22 Modelle eine Genauigkeit von über 90% erreichen(vgl. Abb.4.3b). Da auf Grund der guten Ergebnisse nur wenig Potential für Verbesserung besteht und die beobachtete Varianz der Ergebnisse unterschiedlicher Konfigurationen gering ist, werden keine weiteren Testläufe zur Hyperparameteroptimierung durchgeführt. Es sei jedoch erwähnt, dass sich auf Grund der im Verhältnis zu Größe der Suchraumes geringen Anzahl an Testläufen keine eindeutigen Abhängigkeiten zwischen Hyperparametern und Klassifikationsperformanz ableiten lassen.

Gut beobachten lässt sich der positive Effekt der Nutzung eines vortrainierten Modells zu Initialisierung der Netzwerkparameter. So erreichen alle getesteten Konfigurationen bereits nach der ersten Trainingsepoke eine Validierungsgenauigkeit von über 85% (vgl. Abb.4.3a). Die initialisierten Parameter müssen nur noch geringfügig angepasst werden, um die neue Trainingsaufgabe zu lösen, weshalb die Modelle von Beginn an gute Ergebnisse erzielen.

Betrachtet man die Ergebnisse der Testläufe in Kombination mit den dazugehörigen Hyperparametern (siehe Abb.4.4), so scheint die Anzahl an Trainingsepochen keinen signifikanten Einfluss auf die erzielte Validierungsgenauigkeit zu haben. Dies deckt sich mit der Annahme, dass sich Netzwerkparameter mittels Fine-Tuning in wenigen Epochen optimieren lassen. Betrachtet man den Trainingsverlauf des Fine-Tunings (vgl. Abb. 4.6), so stellt man fest, das für die meisten Konfigurationen nach 10 – 15 Epochen keine großen Verbesserungen mehr im Hinblick auf die Validierungsgenauigkeit erzielt werden.

Beleuchtet man die in den getesteten Konfigurationen genutzten Lernraten, so stellt sich heraus, dass alle Testläufe mit einer Lernrate unter 0.005 sehr gute Validierungsgenauigkeiten erreichen. Werden höhere Lernraten genutzt, so unterscheiden sich die erzielten Ergebnisse deutlich stärker. Bezieht man die dazugehörigen γ -Faktoren mit ein, lässt sich erkennen, dass Konfigurationen mit hoher Lernrate, aber niedrigem γ -Faktor, also mit starker Reduktion der Lernrate während des Trainingsverlaufes, ebenfalls sehr gute Ergebnisse erzielen. Läufe mit hoher Lernrate, sowie hohem γ -Faktor schneiden dagegen eher schlechter ab (Siehe Abb. 4.5). Analysiert man die Trainingsverläufe der unterschiedlichen Konfigurationen (Siehe Abb.4.6), findet sich eine mögliche Erklärung für diese Verhalten. Bei niedriger Lernrate

⁷Obwohl zum Vergleich der Konfigurationen der Validierungsfehler berechnet wurde, wird die Modellperformanz im Folgenden über die Validierungsgenauigkeit dargestellt, da diese Metrik intuitiver ist.



a) Validierungsgenauigkeit nach erster Trainingsepoke b) Validierungsgenauigkeit nach finaler Trainingsepoke

Abbildung 4.3: Erreichte Klassifikationsgenaugikeit nach der ersten bzw. letzten Epoche des Fine-Tunings, der getesteten Konfigurationen. Achsenskalierung variiert.

nähert sich die Validierungsgenauigkeit ohne starke Einbrüche einem Maximum an. Ist die Lernrate hoch, fluktuiert die Validierungsgenauigkeit jedoch stark. Dies weist darauf hin, dass die Modelle nicht in der Lage sind ein stabiles Optimum für ihre Parameter zu finden. Hohe Lernraten können dazu führen, dass Parameter bei Optimierungsschritten zu stark verändert werden und so ihr Optimum immer wieder überspringen. Nach Abschluss der ersten 10 Epochen wird die Lernrate mit dem γ -Faktor multipliziert. Für Läufe mit hoher initialer Lernrate und niedrigem γ -Faktor lässt ab diesem Punkt ein gleichmäßigerer Lernprozess beobachten.

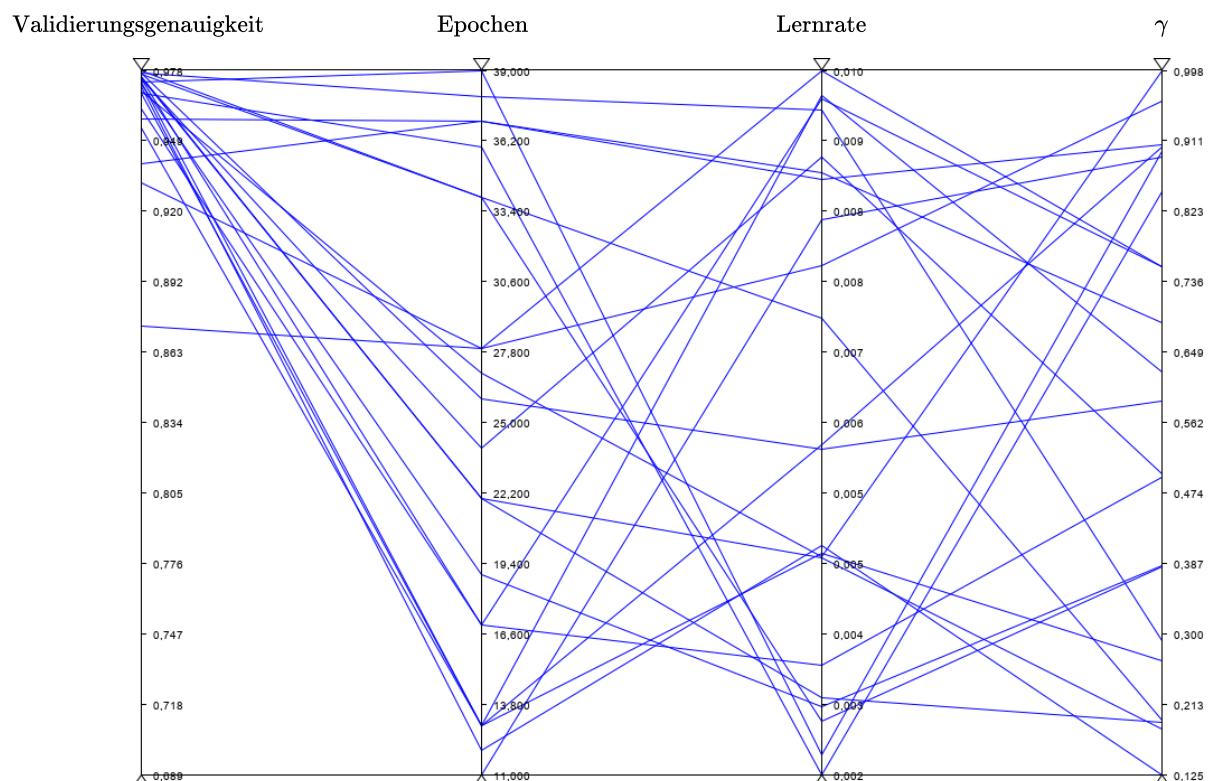


Abbildung 4.4: Parallele Darstellung der getesteten Konfigurationen des Fine-Tunings. Jede Linie repräsentiert einen Konfiguration und die von ihr erzielte Validierungsgenauigkeit.

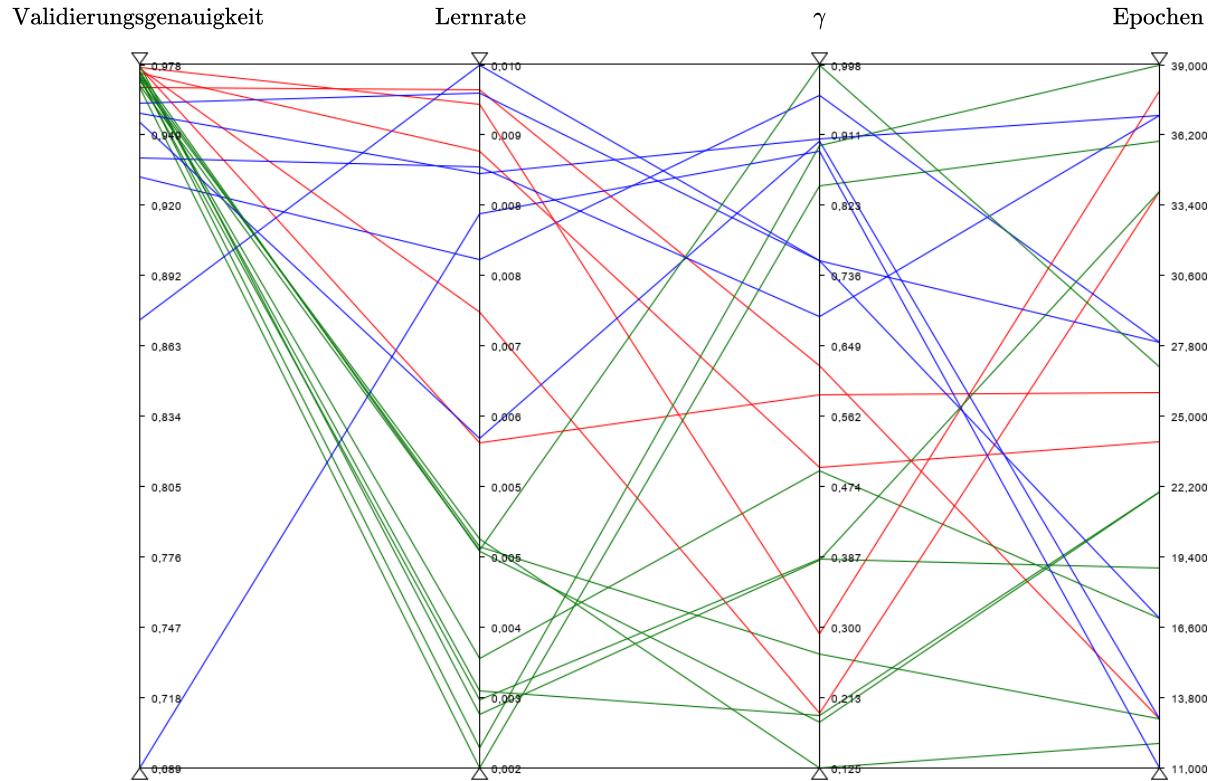


Abbildung 4.5: Betrachtung der genutzten Lernraten und γ -Faktoren. Grün: Konfigurationen mit niedriger Lernrate (< 0.005), Rot: Hohe Lernrate und niedriges γ (< 0.65), Blau: Hohe Lernrate und hohes γ

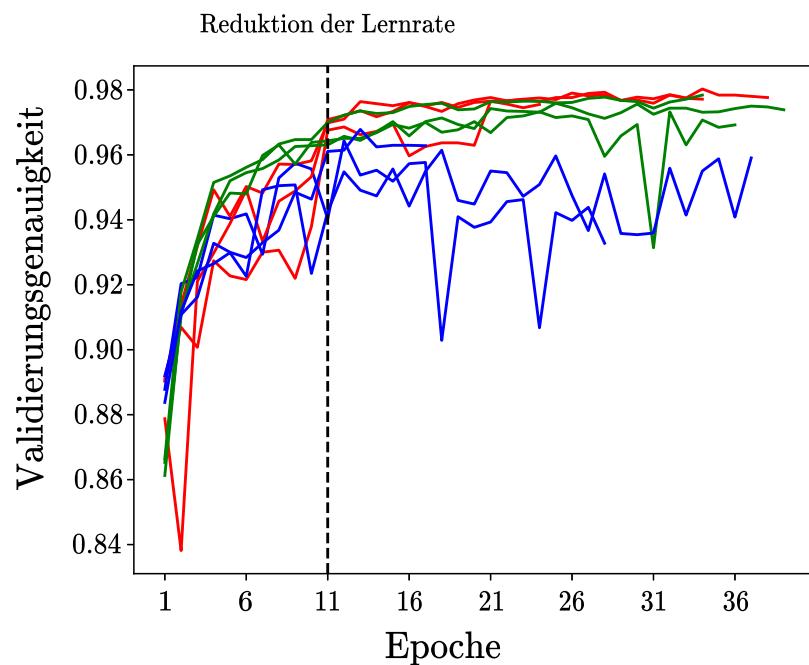
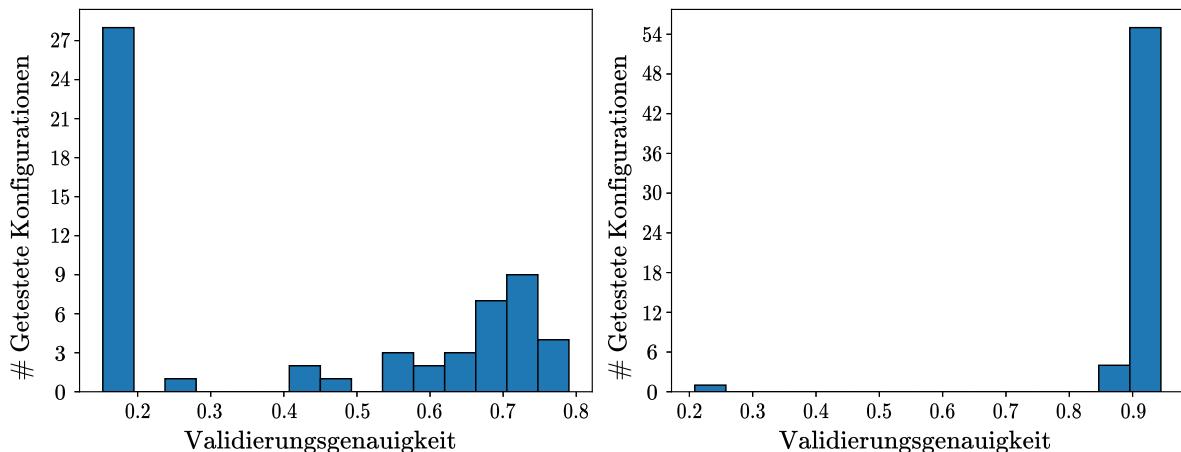


Abbildung 4.6: Trainingsverlauf der Fine-Tunings, bei unterschiedlicher Lernrate und γ . Farbgebung analog zu Abb.4.5

Die Konfiguration mit dem besten Trainingsergebnis nutzt eine Trainingsdauer von 22 Epoche, bei einer initialen Lernrate von 0.0032 und einem γ -Faktor von 0.19 und erzielt nach der letzten Epoche eine Validierungsgenauigkeit von 97.6%.

Das dabei trainierte Modell ist der Ausgangspunkt für die Hyperparameteroptimierung des Attention-Trainings. Analog zu den Experimenten des Fine-Tunings werden mit Hilfe von NNOPT 60 unterschiedliche Konfigurationen für das Attention-Training getestet. Obwohl das Modell für das Attention-Training durch Entfernung des vierten ResNet-Blocks deutlich verkleinert wird und die Ausgaben des ResNets unter Verwendung des gewichteten Sum-Poolings (vgl. Abb.3.3, S. 13) sehr restriktiv genutzt werden, sind die Attention-Modelle in fast allen untersuchten Konfigurationen in der Lage die Klassifikationsaufgaben weiterhin sehr gut zu Lösen. So werden in 59 der 60 untersuchten Konfigurationen eine abschließende Validierungsgenauigkeit von über 89% erreicht (siehe Abb.4.7b). Die durchschnittliche Validierungsgenauigkeit nach der letzten Trainingsepoche beträgt 91.3%, was einer durchschnittlichen Verschlechterung von 6.3% gegenüber dem verwendeten fine-getunten ResNet-Modell entspricht. Im Kontrast zum Fine-Tuning variiert die gemessene Validierungsgenauigkeit nach der ersten Trainingsepoche der Attention-Trainings deutlich stärker und ist allgemein niedriger (vgl. Abb.4.7a). Da die Parameter der zu optimierenden Attention-Einheit nicht vortrainiert sind und daher zufällig initialisiert werden sind größere Unterschiede und schlechter Ergebnisse zu Beginn des Trainings, sowie eine längere benötigte Trainingsdauer zu erwarten.



a) Validierungsgenauigkeit nach erster Trainingsepoche b) Validierungsgenauigkeit nach finaler Trainingsepoche

Abbildung 4.7: Erreichte Klassifikationsgenauigkeit nach der ersten bzw. letzten Epoche des Attention-Trainings, der getesteten Konfigurationen. Achsenkalierung variiert.

Betrachtet man die genutzte Anzahl an Trainingsepochen in Kombination mit der erreichten Validierungsgenauigkeit (siehe Abb.4.8), bestätigt sich diese Annahme. Konfigurationen mit einer Trainingsdauer unter 20 Epochen erzielen in unseren Experimenten durchschnittlich eine geringere Validierungsgenauigkeit. Ab mehr als 20 Epochen schwächt sich der positive Effekt einer längeren Trainingsdauer jedoch deutlich ab.

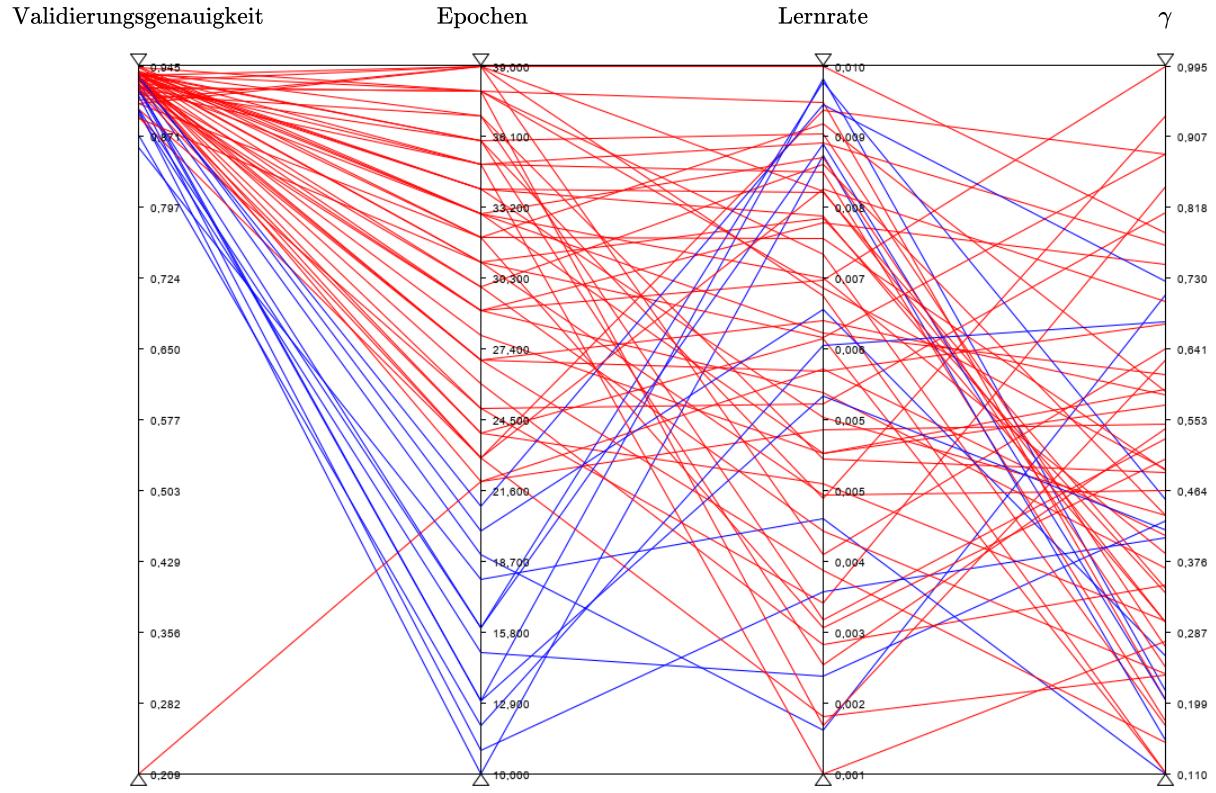


Abbildung 4.8: Betrachtung der gewählten Trainingsdauer. Rot: Über 20 Epochen, Blau: 20 oder weniger Epochen

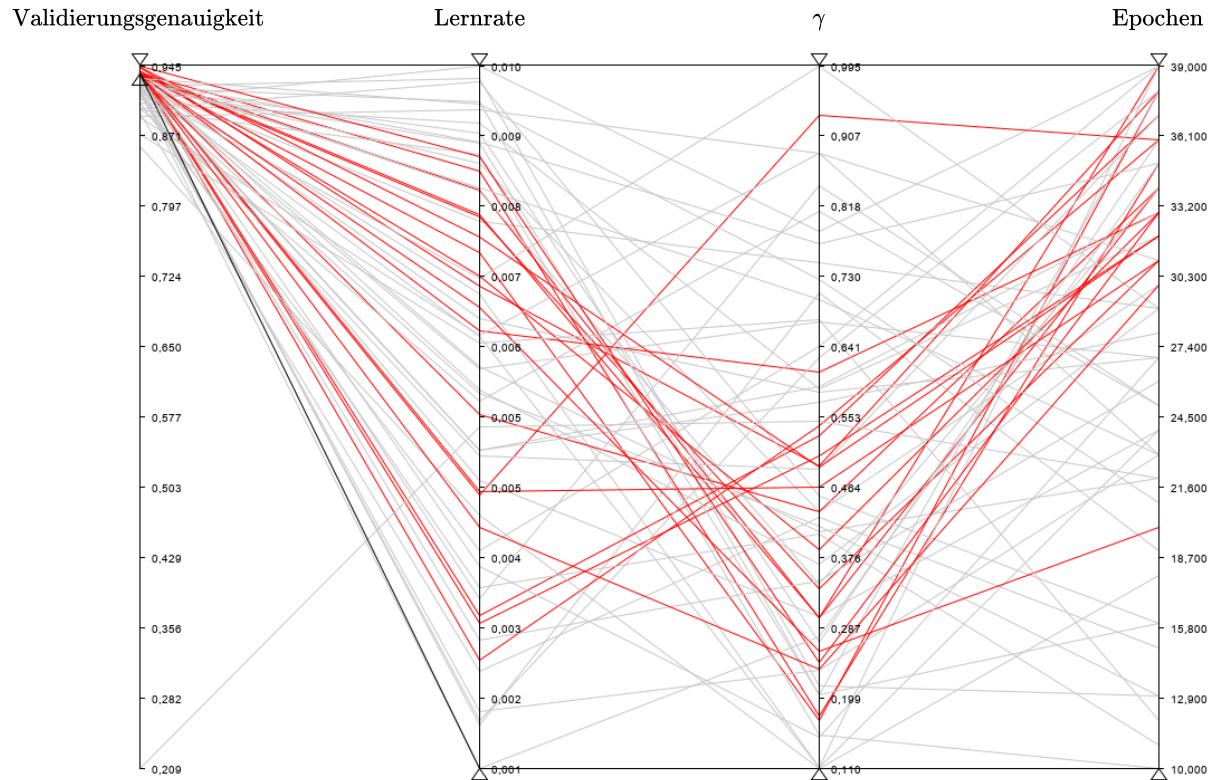


Abbildung 4.9: Darstellung der Validierungsgenauigkeit nach Trainingsende. Rot: Läufe mit Validierungsgenauigkeit $\geq 93.5\%$, Grau: Läufe mit niedrigerer Validierungsgenauigkeit

Für Lernrate und γ -Faktor lassen sich nur schwer Tendenzen erkennen. Über den gesamten Suchraum dieser Parameter lassen sich sowohl Konfigurationen mit hoher, sowie Konfigurationen mit weniger hoher Validierungsgenauigkeiten erreichen finden. Die Konfigurationen, welche die höchsten Validierungsgenauigkeiten erreichen (Siehe Abb.4.9) nutzen jedoch alle eine Lernrate im Bereich zwischen 0.0025 und 0.009. Die γ -Faktoren dieser Konfigurationen liegen meist unter 0.55 und die Trainingsdauer über 30 Epochen. Die beste gefundene Konfiguration erzielt eine Validierungsgenauigkeit von 94.5% und nutzt dabei eine Lernrate von 0.0078 einen γ -Faktor von 0.49 und eine Trainingslänge von 32 Epochen. Mit Hilfe der ermittelten Hyperparameterwerte der besten gefundenen Konfigurationen für Fine-Tuning und Attention-Training werden abschließend 6 Modelle Trainiert, die in den Retrievalexperimenten zur Extraktion der Deskriptoren genutzt werden. Die Trainingsverläufe dieser Modelle sind in Abb.4.10 dargestellt.

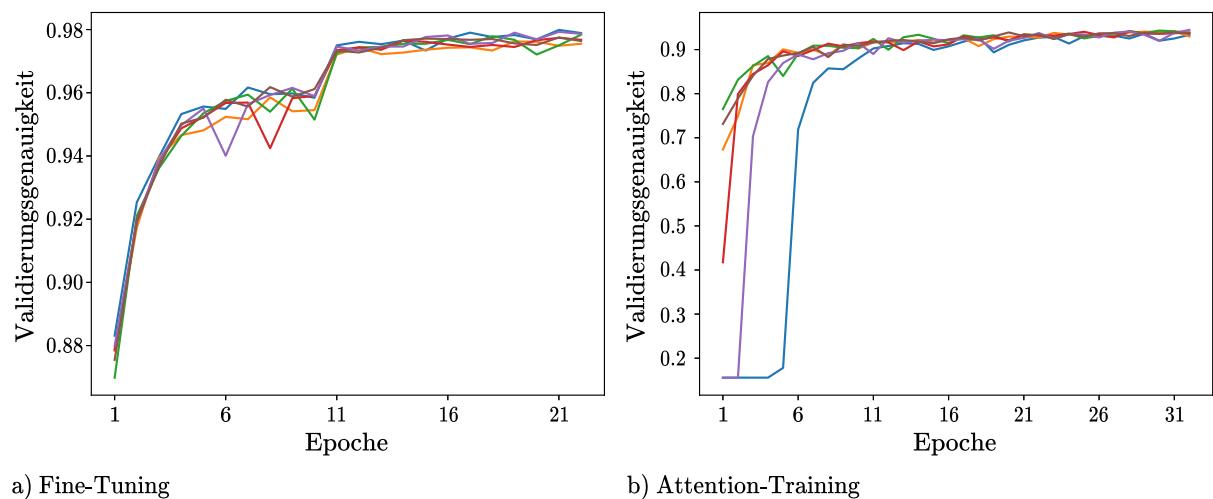


Abbildung 4.10: Trainingsverläufe mit optimierten Hyperparametern. Gleichefarbige Linien gehören zu gemeinsamen Trainingsläufen. Achsenkalierung variiert.

4.3.2 Variieren der Deskriptorlänge

Nach Betrachtung der Trainingsphasen und Erstellung optimierter Modelle, können nun Parameter aus den späteren Phasen des Verfahrens untersucht werden. Ein zentraler Parameter der Extraktions- und Verarbeitungsphase ist die Anzahl an Dimensionen, der Deskriptoren nach der Transformation mittels Hauptkomponentenanalyse (vgl. 3.5.4, S. 17). Da die zu einem Suchdatensatz erstellte Repräsentation fast ausschließlich aus den transformierten Deskriptoren besteht, sind diese für den überwiegenden Teil des Speicherbedarfs verantwortlich, der während der aktiven Verwendung des Suchsystems anfällt⁸. Insbesondere bei der Suche in großen Datenbanken ist es daher notwendig die Deskriptoren stark zu komprimieren. Durch die Komprimierung geht jedoch auch immer ein Teil der ursprünglich in den Deskriptoren enthaltenen Informationen verloren. Es wird erwartet, dass sich der Informationsverlust durch die Deskriptortransformation negativ auf die Qualität des initialen Deskriptormatchings auswirkt. Um

⁸Bei der Suche nach den ähnlichen Deskriptoren während dem Matching hat die Deskriptorlänge auch einen Einfluss auf die Laufzeit. Da der Rechenaufwand für die geometrische Verifizierung von Matches, die nicht von der Deskriptorlänge abhängt, jedoch deutlich größer ist, ist dieser Einfluss vernachlässigbar.

diesen Informationsverlust zu quantifizieren wird der Anteil der ursprünglichen Varianz zwischen den Deskriptoren errechnet, die nach der Transformation erhalten bleibt. Hierbei werden alle Deskriptoren betrachtet, die für die Berechnung der Hauptkomponentenanalyse genutzt werden. Es ist davon auszugehen, dass sich der Informationsverlust auf unbeteiligten Deskriptoren ähnlich verhält. Die Autoren des DELF-Papiers [2] schlagen als Kompromiss zwischen erhaltener Information und Kompaktheit vor, die Deskriptoren auf 40 Dimensionen zu reduzieren. Wie dieser Wert ermittelt wurde wird jedoch nicht beschrieben. Um eine geeigneten Anzahl an Dimensionen zu ermitteln werden neben der von den Autoren empfohlenen Anzahl auch die Hälfte bzw. ein Vielfaches an erhaltenen Dimensionen, bis zu 200 getestet. In Abb. 4.11 wird der Einfluss der Deskriptorlänge auf den Anteil der erklärten ursprünglichen Varianz dargestellt. Die Variation der Deskriptorlänge hat nicht nur Einfluss darauf, welche Deskriptorpaa

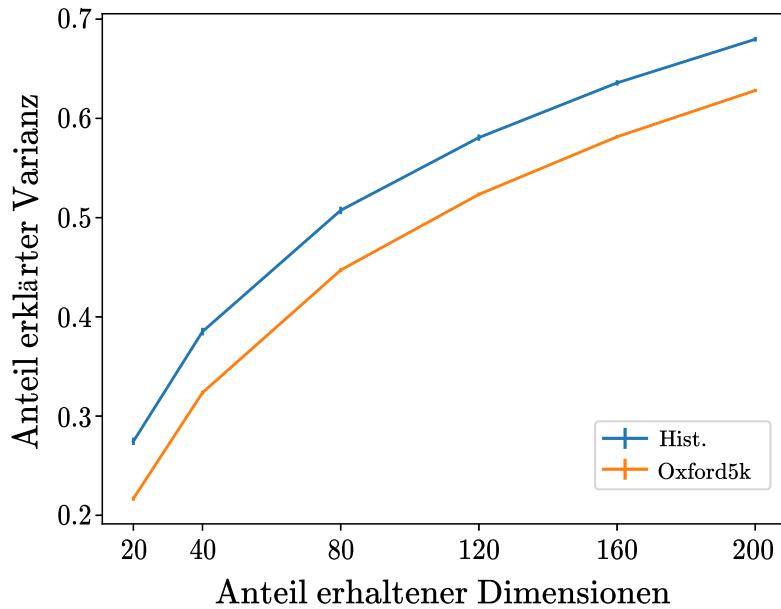


Abbildung 4.11: Erklärter Anteil der Varianz nach Deskriptortransformation durch Hauptkomponentenanalyse auf unterschiedlichen Retrievaldatensätzen, je nach Anzahl erhaltener Dimensionen.

während des initialen Matchings bestimmt werden, sondern auch auf die euklidische Distanz zwischen diesen Paaren ist. Die Deskriptoren werden vor dem Matching zwar Längennormiert, sodass der maximale Abstand zwischen Deskriptoren 2 beträgt. Der mittlere Abstand der bestimmten Deskriptorenpaare ist jedoch von der Varianz zwischen den Deskriptoren abhängig und steigt daher mit wachsender Deskriptorlänge. Während dem initialen Deskriptormatching werden Matchingpaare, deren Abstand über einem Distanzschwellwert liegen verworfen, um inkorrekte Matches auszusortieren (vgl. Alg.3.2, S. 19). In Abb. 4.12 lässt sich beobachten, wie sich die Anzahl an gefundenen Deskriptormatches mit steigender Länge der verwendeten Deskriptoren verringert, wenn der gewählte Distanzschwellwert unverändert bleibt. In Abb. 4.13 ist analog, bei einer konstanten Deskriptorlänge von 80 der Einfluss von unterschiedlichen Schwellwerten dargestellt. Je größer der Schwellwert, desto weniger Deskriptormatches werden aussortiert. Die untersuchten Deskriptorlängen, werden daher auch mit mehreren unterschiedlichen Schwellwerten getestet um geeignete Kombinationen zu finden.

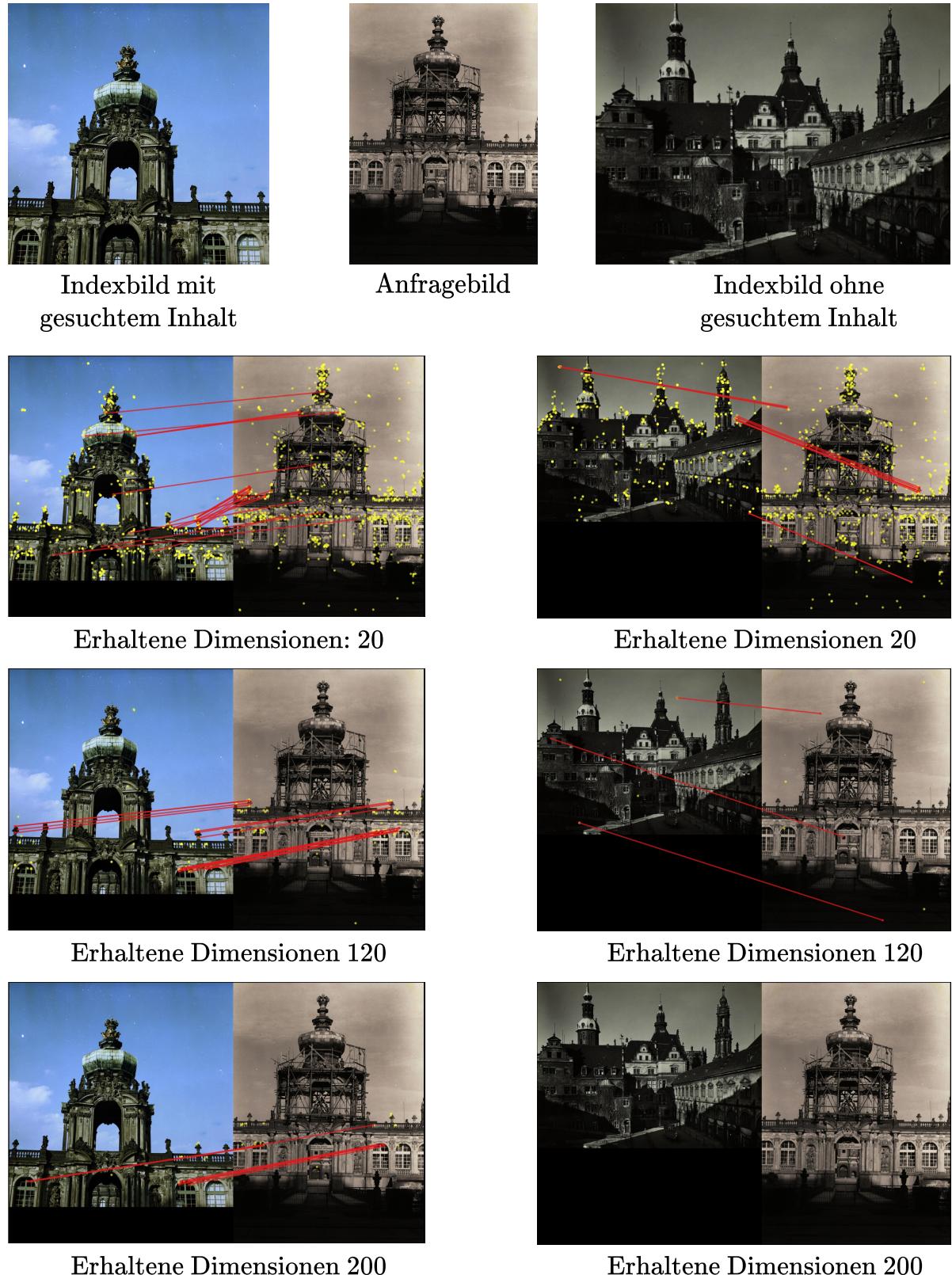


Abbildung 4.12: Einfluss der Deskriptorlänge auf das Deskriptorenmatching zwischen Anfragebild und Bildern mit und ohne gesuchtem Bildinhalt aus dem Suchindex, bei einem Schwellwert für maximale Deskriptordistanz von 0.8. Gelbe Punkte markieren Deskriptoren die im Partnerbild gematched wurden. Rote Linien markieren geometrische verifizierte Matches

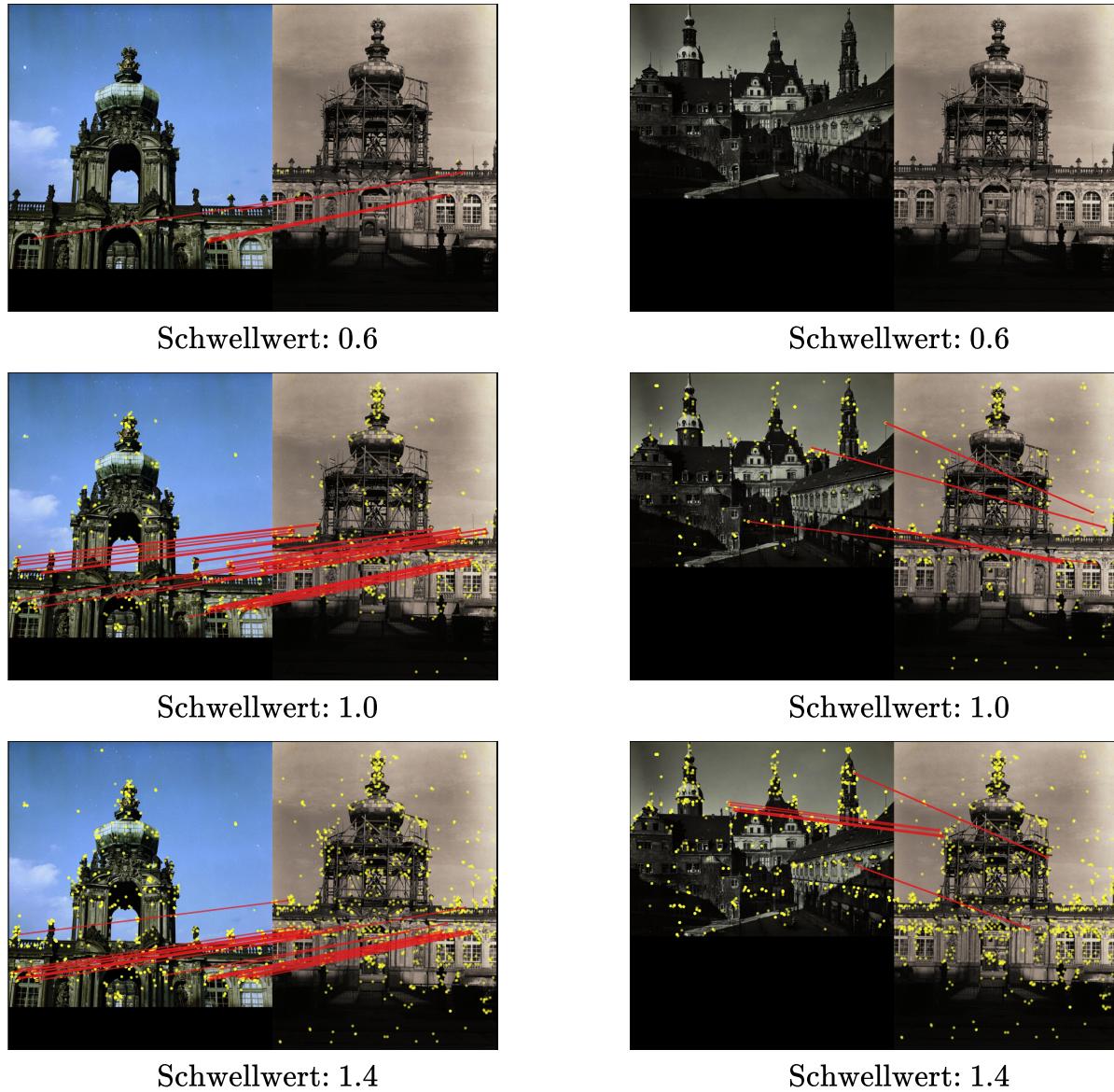


Abbildung 4.13: Einfluss des Distanzschwellwerts auf das Deskriptorenmatching, bei Verwendung von 80 dimensionalen Deskriptoren. Gelbe Punkte markieren Deskriptoren die im Partnerbild gematched wurden. Rote Linien markieren geometrische verifizierte Matches

Die Autoren der DELF-Papiers [2] verwenden für 40-dimensionale Deskriptoren einen Schwellwert von 0.8. Im Folgenden werden neben diesem Wert sowohl kleiner, wie auch größere Werte zwischen 0.6 und 1.4 in Intervallschritten von 0.2 getestet. Alle Parameterkombinationen von Deskriptorlänge und Distanzschwellwert werden auf den 6 trainierten Modellen getestet. Für die Auswertung werden die Ergebnisse jeweils über die Läufe der 6 Modelle gemittelt. Zusätzlich werden Fehlerbereiche im Größer der Standardabweichung in beide Richtungen angegeben. Diese Vorgehen wird auch für alle folgenden Parameteranalysen in der vorliegenden Arbeit angewendet. In Abb. 4.14 ist für alle betrachteten Konfigurationen von Deskriptorlänge und Distanzschwellwert, die auf dem historischen Datensatz sowie auf den Oxford5k-Daten getestet werden, die erreichte mAP dargestellt.

Unabhängig der getesteten Konfiguration erzielt DELF auf dem Oxford5k-Datensatz eine deutlich hö-

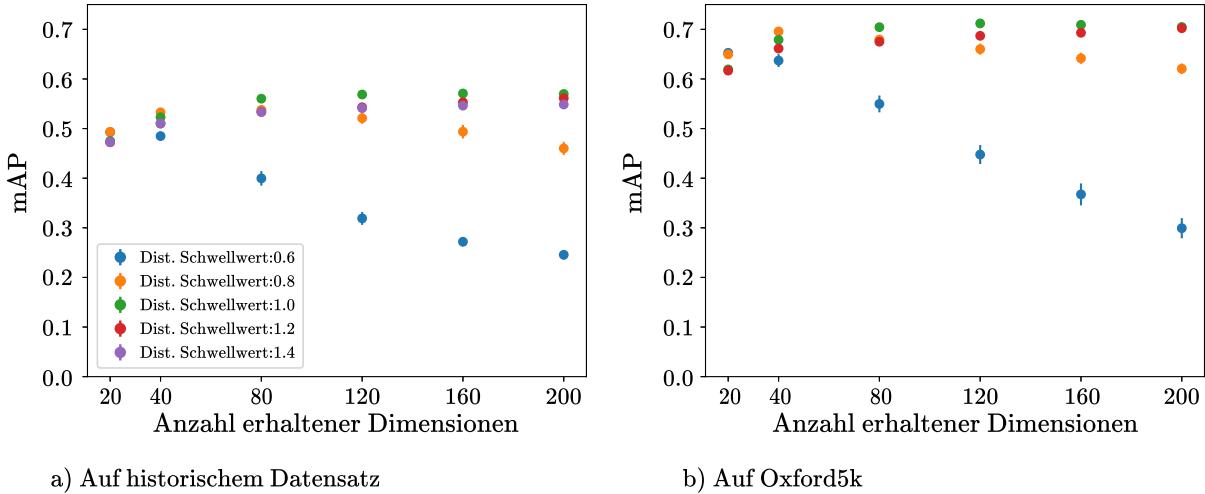


Abbildung 4.14: Erreichte mAP bei unterschiedlicher Deskriptorlänge und Distanzschwellwert.

here mAP, als auf den historischen Daten. Wie erwartet lassen sich mit größerer Deskriptorlänge bei geeignetem Distanzschwellwert tendenziell bessere Ergebnisse erzielen. Die beste gefundene Konfiguration auf den historischen Daten nutzt Deskriptoren der Länge 160 bei einem Schwellwert von 1.0 und erzielt im Mittel eine mAP von 0.57. Auf Oxford5k erreicht die beste Konfiguration im Mittel eine mAP von 0.71 und nutzt dabei eine Deskriptorlänge von 120, mit einem Schwellwert von 1.0. Allerdings lässt sich auf Oxford5k mit der von den DELF-Autoren empfohlenen Deskriptorlänge von 40 und dem empfohlenen Schwellwert von 0.8 mit einer mittleren mAP von 0.70 ein fast gleich gutes Ergebnis erzielen. Lediglich bei Verwendung von noch kleineren Deskriptoren lassen sich signifikante Performanzebüßen feststellen. So erreicht die beste Konfiguration mit Deskriptorlänge 20 im Mittel nur eine mAP von 0.65. Auf den historischen Daten lassen sich mit Deskriptoren mit mehr als 40 Dimensionen noch signifikante Verbesserungen erzielen. So erreicht die beste Konfiguration mit 80 Dimensionen eine im Mittel um 0.03 höhere mAP als die von den Autoren empfohlene Konfiguration. Eine weitere Vergrößerung der Deskriptoren ziehen allerdings auch auf den historischen Daten keine großen Performanzverbesserungen nach sich.

Bei Betrachtung der Distanzschwellwerte, stellt man fest, dass der optimale Schwellwert mit der Länge der verwendeten Deskriptoren steigt. Dies deckt sich mit den Beobachtungen aus Abb. 4.12 und Abb. 4.13, da mit wachsender Deskriptorlänge ein höherer Schwellwert gewählt werden muss, um eine adäquate Anzahl an Deskriptormatches zu erhalten. Bei einem zu klein gewählten Schwellwert werden fast alle potentiellen Deskriptormatches aussortiert, was zu einem drastischen Perfomanzverlust führt. Ein zu groß gewählter Schwellwert führt ebenfalls zu Performanzverlusten, welche jedoch deutlich geringer ausfallen. Das ist der zusätzlichen geometrischen Verifikation mittels RANSAC zu verdanken. Selbst wenn durch den Schwellwert keine unerwünschten Deskriptormatches verworfen werden, wird ein Großteil dieser Matches während der geometrischen Verifikation durch keine Transformation erklärt und somit aussortiert⁹.

Wie bereits erläutert hängt die Wahl eines geeigneten Schwellwerts stark von der Länge der Deskriptoren

⁹In Abbildung 4.16 auf Seite 39 wird deutlich, dass die Performanz stark unter einem zu hohen Schwellwert leidet, wenn keine geometrische Verifikation durchgeführt wird.

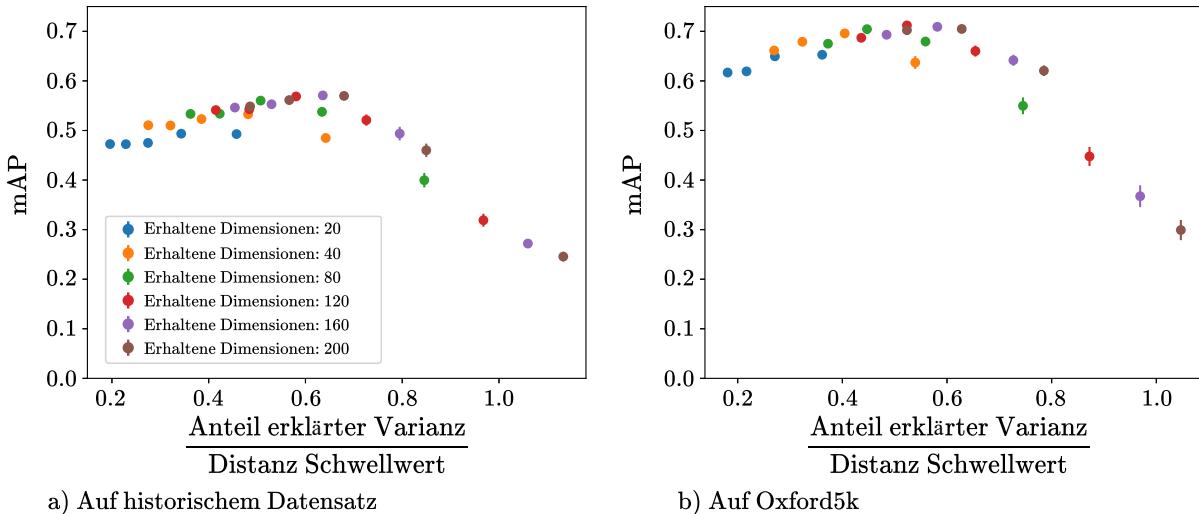


Abbildung 4.15: Erreichte mAP bei unterschiedlichen Verhältnissen zwischen erklärtem Varianzanteil und gewähltem Distanzschwellwert.

bzw. dem Anteil der von ihnen erklärten Varianz ab. Ein mögliche Strategie für die Untersuchung von vielen unterschiedlichen Deskriptorlängen ist daher den Schwellwert direkt abhängig von der erklärten Varianz zu bestimmen und so die Anzahl zu testender Konfigurationen zu reduzieren. In Abb. 4.15 ist die erreichte mAP im Bezug zum Verhältnis zwischen erklärter Varianz und Schwellwert abgebildet. Tatsächlich erreichen fast alle Deskriptorlängen ihr bestes Ergebnis, wenn das Verhältnis zwischen dem Anteil der erklärten Varianz und dem gewählten Schwellwert zwischen 0.5 und 0.7 liegt¹⁰. Für die Untersuchung anderer Deskriptorlängen wäre es daher sinnvoll Schwellwerte zu testen, die innerhalb dieses Bereiches liegen.

4.3.3 Alternativen zur Bewertung von Deskriptormatches

Bei der Überprüfung auf geometrische Plausibilität von Deskriptormatches mittels RANSAC werden durch Betrachtung der Deskriptorpositionen zusätzliche Informationen für das Matching nutzbar gemacht, die idealerweise Suchergebnisse verbessern. Dabei sollte beachtet werden, dass diese geometrische Überprüfung für einen Großteil der für das Matching benötigten Rechenzeit verantwortlich ist. Im Folgenden wird analysiert, ob und wie stark sich Suchergebnisse unter Verwendung von RANSAC gegenüber alternativen Bewertungsmethoden für Deskriptormatches verbessern, um zu bewerten wie sinnvoll dieser zusätzliche Arbeitsschritt ist.

Eine einfache Möglichkeit um ein Matching ohne geometrische Verifikation zu bewerten ist, die Anzahl an Deskriptormatches zu zählen. Je mehr Deskriptorpaares zwischen einem Bildpaar gefunden werden können, desto mehr sollte sich der Inhalt dieser Bilder ähneln. Voraussetzung für eine aussagekräftige Bewertung ist die Verwendung eines geeigneten Distanzschwellwertes. Dieser entscheidet, welche Deskriptorpaares keine ausreichende Ähnlichkeit vorweisen können und daher verworfen werden sollten. Paare, die diesen Schwellwert nicht überschreiten gehen alle mit der gleichen Gewichtung in die Bewertung mit ein.

¹⁰Die wird besonders deutlich, wenn keine geometrische Verifikation stattfindet und zu große Schwellwerte stärkere negative Einflüsse zeigen. Abbildungen hierzu finden sich im Anhang ab Seite ??.

Intuitiv stellen Deskriptorpaare, die sich besonders stark ähneln, jedoch einen besseren Indikator für die Ähnlichkeit zwischen Bildern dar, als Paare, die diesen Schwellwert nur knapp unterschreiten. Daher wird als weitere Bewertungsmethode eine gewichtete Anzahl an Deskriptorpaaren berechnet. Paare, die den Schwellwert unterschreiten, gehen hierbei mit einem Gewicht zwischen 0 und 1 in die Bewertung ein. Ein Paar an identischen Deskriptoren geht dabei mit Gewicht 1 und ein Paar, dessen Distanz genau dem Schwellwert entspricht, mit Gewicht 0 in die Bewertung ein. Die Gewichtung verläuft linear proportional zur Distanz der Deskriptorpaare. Die gewichtete Anzahl bei Distanzschwellwert T ergibt sich aus,

$$\text{gewichtete Anzahl} = \sum_{i=1}^{|D_\Delta|} 1 - \frac{d_{\Delta i}}{T}, \quad (4.4)$$

wobei D_Δ die Distanzen aller Deskriptorpaare $d_{\Delta i}$ enthält, die den Schwellwert nicht überschreiten. In

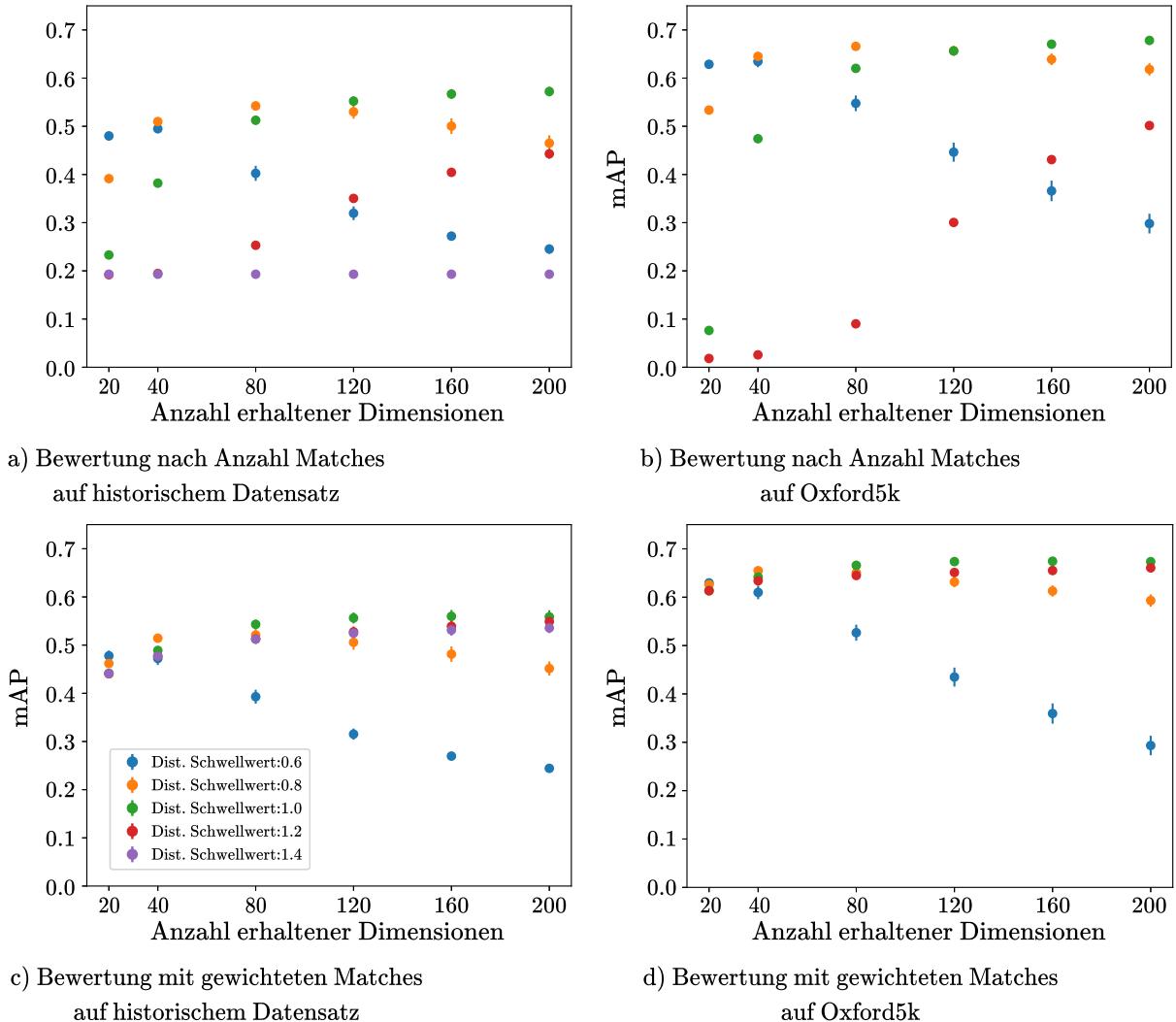


Abbildung 4.16: Erreichte mAP bei unterschiedlicher Deskriptorlänge und Distanzschwellwert unter Verwendung alternativer Bewertungsmethoden.

Abb. 4.16 sind die Ergebnisse der alternativen Bewertungsmethoden, analog wie für RANSAC in Abb.

4.14 dargestellt.

Auch mit alternativen Bewertungsmethoden erzielt DELF auf dem Oxford5k-Datensatz eine deutlich höhere mAP, als auf den historischen Daten. Deskriptoren mit mehr Dimensionen erreichen unter Verwendung eines optimierten Distanzschwellwerts eine tendenziell höhere mAP, wobei die Performanzunterschiede auf den historischen Daten deutlich größer ausfallen, als auf dem Oxford5k-Datensatz. Die größte Verbesserung diesbezüglich wird auf den historischen Daten erzielt, wenn die Anzahl an Descriptormatches zur Bewertung verwendet wird (siehe 4.16 a). Hier verbessert sich die durchschnittliche mAP bei einer Erhöhung der Deskriptorlänge von 20 auf 200 um 0.09.

Die Verwendung eines sehr kleinen Distanzschwellwertes führt, wie bereits unter RANSAC beobachtet (vgl. Abb. 4.14), zu einer drastischen Verschlechterung der Performanz, unabhängig welche Bewertungsmethode genutzt wird. Bei Wahl eines sehr großen Schwellwertes, der unter Verwendung von RANSAC nur zu geringfügigen Verschlechterungen der Performanz führt, werden deutlich schlechtere Ergebnisse erzielt, wenn die Anzahl der Descriptormatches zur Bewertung genutzt wird (siehe 4.16 a, b). Je höher der Schwellwert gewählt ist, desto mehr Deskriptorpaares, dessen Deskriptoren keinen ähnlichen Bildinhalt repräsentieren, werden akzeptiert. Da fälschlicherweise akzeptierte Deskriptorpaares gleichermaßen in die Bewertung eines Matchings eingehen, können auch Bildpaare ohne ähnlichem Bildinhalt hoch bewertet werden. Ab einem gewissen Punkt kann in jedem Bildpaar für jeden Deskriptor ein Partner gefunden werden. Somit erhält jedes Bildpaar die gleiche Bewertung, wodurch Anfragen mit Bildern in einer zufälligen Reihenfolge beantwortet werden. Werden die Deskriptorpaares nach ihrer Ähnlichkeit gewichtet (siehe 4.16 c, d), gehen Deskriptorpaares, die sehr unterschiedliche Bildinhalte repräsentierten, deutlich schwächer in die Bewertung ein, weshalb sich ein zu großer Schwellwert kaum negativ auf die Ergebnisse auswirkt.

Wie sich die Performanz, je nach gewählter Bewertungsmethode und Deskriptorlänge unterscheidet ist

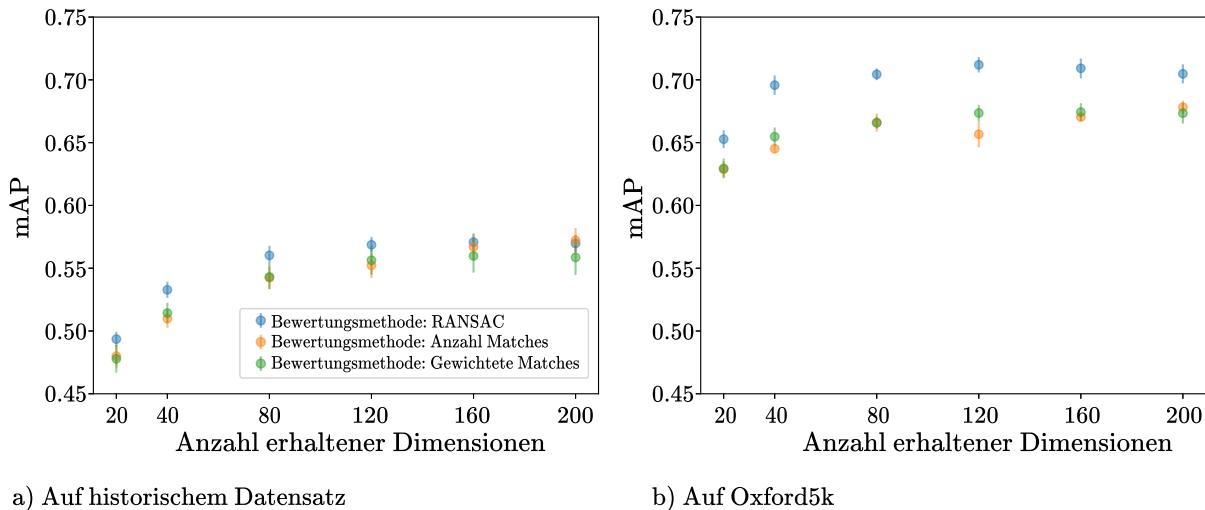


Abbildung 4.17: Erreichte mAP bei unterschiedlicher Deskriptorlänge mit optimiertem Distanzschwellwert unter Verwendung unterschiedlicher Bewertungsmethoden.

in Abb. 4.17 dargestellt. Die hier gezeigten Konfigurationen verwenden jeweils die besten für sie gefundenen Schwellwerte. Auf den Oxford5k-Daten werden mit RANSAC deutlich bessere Ergebnisse, als mit anderen Bewertungsmethoden erzielt. Die beste gefundene Konfiguration erreicht dabei im Mittel

Kategorie	Frauenkirche	Hofkirche	Zwinger	Sophienkirche	Semperoper	Moritzburg	Stallhof
							
ΔmAP	0.076	0.076	0.025	-0.011	-0.034	-0.074	-0.153

Tabelle 4.4: Veränderung der mAP bei Verwendung von RANSAC gegenüber einfachem zählen der Deskriptormatches, je nach Objektkategorie. Der betrachtete Testlauf verwendet eine Deskriptorlänge von 200 und einen Distanzschwellwert von 1.0.

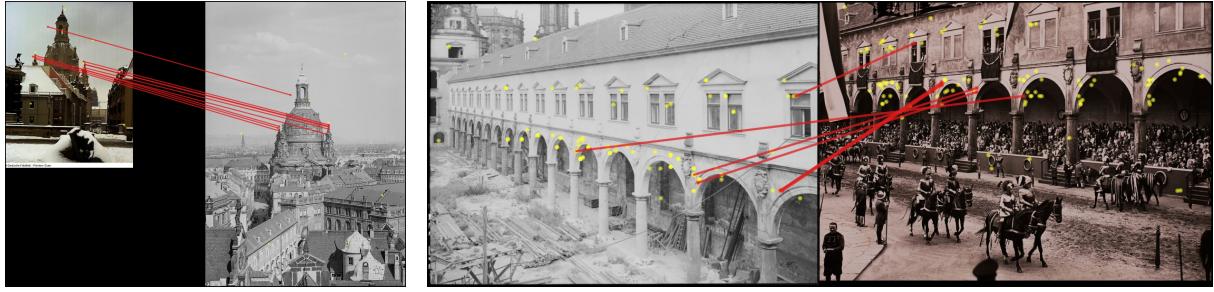
eine mAP von 0.71. Die beste Konfiguration mit einer alternativen Bewertungsmethode nutzt die Anzahl Matches zur Bewertung und erreicht eine mittlere mAP von 0.68. Die Ergebnisse bei Betrachtung von gewichteten Matches unterscheidet sich dabei kaum von den Ergebnissen basierend auf der reinen Matchanzahl.

Auf den historischen Daten fallen die Verbesserungen durch RANSAC deutlich geringer aus. Insbesondere bei Verwendung von hochdimensionalen Deskriptoren werden durch geometrische Verifikation keine besseren Ergebnisse erzielt, als bei Verwendung der Matchanzahl. Die beste gefundene Konfiguration für die historischen Daten nutzt ebenfalls die Anzahl an Matches zur Bewertung und erzielt eine mittlere mAP von 0.57. Die Unterschiede zur besten RANSAC-Konfiguration liegen jedoch innerhalb der ermittelten Fehlerbereiche. Die Betrachtung anhand gewichteter Matches schneidet auch auf den historischen Daten größtenteils ähnlich wie die Betrachtung anhand der Anzahl an Matches ab. Lediglich bei Verwendung sehr großer Deskriptoren werden im Vergleich schlechtere Ergebnisse erzielt. Speziell für den historischen Anwendungsfall scheint der zusätzliche geometrische Verifikationsschritt keinen signifikanten Verbesserungen der Performanz zu erzielen. Um genauer einzugrenzen, auf welchen Daten RANSAC in der Lage ist Retrievalergebnisse zu verbessern, und wo dies nicht funktioniert, ist in Tab. 4.4 für die unterschiedlichen Anfragekategorien aufgeführt, wie sich die mAP bei Verwendung von RANSAC gegenüber der Bewertung mittels Matchanzahl verändert.

Dabei zeigt sich, dass Motive mit gut differenzierbaren markanten Bereichen, wie beispielsweise die Türme der Frauenkirche stark von einer geometrischen Verifizierung profitieren. Objekte die häufig wiederkehrende Elemente enthalten, wie zum Beispiel die Fenster der Moritzburg oder die Arkaden des Stallhofs, verschlechtern sich dagegen bei Verwendung von RANSAC.

Warum die geometrische Überprüfung hier problematisch ist lässt sich gut an dem Beispielmatching zweier Stallhofbilder in Abb. 4.18 b erkennen. Obwohl DELF viele Matchpaare entlang der Arkaden und Fenstern findet, können nur wenige mit einer affinen Transformation erklärt werden. Betrachtet man die verifizierten Matches, so fällt auf, dass DELF Schwierigkeiten hat die korrekten Arkaden miteinander zu matchen. Auf Grund der sich wiederholenden Strukturen, können einzelne Bereiche nicht klar genug unterschieden werden. Das resultierende Matching lässt sich daher nicht einheitlich mit einer Transformation erklären. Im Gegensatz dazu konzentriert sich das Beispielmatching der Frauenkirche (vgl. Abb. 4.18 a) nur auf zwei ihrer seitlichen Türme, wodurch sich eine Transformation finden lässt, die einen Großteil der Deskriptorpaare erklärt. Obwohl hier die Anzahl der Deskriptorpaare vergleichsweise gering ist, kann durch die geometrische Verifikation ein gutes Matching erstellt werden.

Abschließend lässt sich zusammenfassen, dass der zusätzliche Verifikationsschritt, den DELF vorschreibt,



Rückgabeindex mit Ransac: 89
Rückgabeindex bei Anzahl Matches: 395

a) Verbesserung durch RANSAC

Rückgabeindex mit Ransac: 137
Rückgabeindex bei Anzahl Matches: 22

b) Verschlechterung durch RANSAC

Abbildung 4.18: Beispiele für Matchings von Bildpaaren mit gleichem Bildinhalt, die von geometrischer Überprüfung profitieren, bzw. darunter leiden. Der Rückgabeindex gibt an, an wievieler Stelle der Anfrageantwort das entsprechende Bildpaar zurückgegeben wird (niedriger ist besser). Gelbe Punkte repräsentieren Deskriptoren, für die ein Match gefunden wurde. Rote Linien verbinden verifizierte Deskriptorpaare.

nicht uneingeschränkt sinnvoll für das lösen von Retrievalaufgaben ist. Architektonische Besonderheiten und Beschaffenheit der Bildinhalte haben einen signifikanten Einfluss auf Effektivität von RANSAC, somit sollte der Einsatz, je nach Datensatz überdacht werden. Um herauszufinden, ob RANSAC allgemein für die historische Domäne ungeeignet ist, sollten jedoch weiter Experimente auf einem erweiterten Datensatz, insbesondere mit mehr unterschiedlichen Motiven, durchgeführt werden.

4.3.4 Alternative Extraktionspunkte

Das DELF-Verfahren sieht es vor, die Ausgaben des dritten Blocks aus dem zugrundeliegenden ResNet-50 (vgl. Abb. 3.2b, S. 10) als Deskriptoren zu verwenden. Warum genau dieser Punkt zur Extraktion der Deskriptoren verwendet wird, begründen die Autoren des DELF-Papiers [2] jedoch nicht. Als letzte Parameterbetrachtung der vorliegenden Arbeit wird daher ein alternativer Extraktionspunkt untersucht. Dabei werden die Netzwerkausgaben des vierten und letzten ResNet-Blocks betrachtet. Diese Wahl beruht auf den Beobachtungen von Zeiler und Fergus, die in [12] gezeigt haben, dass Netzwerkausgaben aus späteren Schichten eines neuronalen Netzes in der Lage sind komplexere Bildmerkmale detektieren. Deskriptoren aus dieser späten Schicht können daher möglicherweise besser für die Differenzierung komplexer Bildinhalte genutzt werden.

Die bisher verwendeten Extraktionsnetzwerke können für die Deskriptorextraktion aus Block-4 weiter verwendet werden. Da sich die extrahierten Deskriptoren jedoch in Form und Inhalt von Deskriptoren aus Block-3 unterscheiden, müssen für die Deskriptorauswahl neue Attention-Netzwerke trainiert werden. Zunächst findet hierfür eine Optimierung der Hyperparameter analog wie in Kap. 4.3.1 statt. Es zeigt sich dabei, dass alle Testläufe unabhängig ihrer Hyperparameterkonfigurationen, ähnlich gute Testergebnisse erzielen¹¹. Mit den optimierten Hyperparametern werden anschließend 6 neue Attention-Modelle trainiert. Die Trainingsverläufe dieser Modelle (siehe Abb. 4.19) erreichen bereits ab der ersten Epoche sehr hohe Validierungsgenauigkeiten und steigern sich während des Trainings auch nur geringfügig.

¹¹Grafiken zu dieser Hyperparameteroptimierung finden sich im Anhang ab Seite ??.

Im Vergleich dazu zeigen Trainingsverläufe auf Basis von Deskriptoren aus Block-3 eine sehr geringe initiale Validierungsgenauigkeit, die sich während des Trainings deutlich steigert (vgl. Abb. 4.10 b, S. 33). Die Ausgaben aus dem vierten Block gehen während des Fine-Tunings, nach einem finalen Pooling

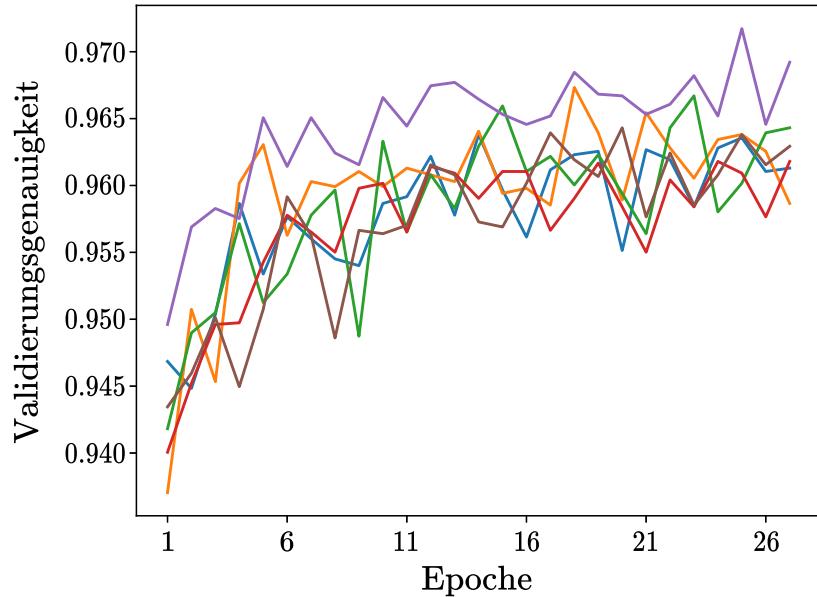


Abbildung 4.19: Verläufe des Attention-Trainings für Deskriptoren aus Block-4, mit optimierten Hyperparametern.

direkt in die Klassifikationsschicht ein. Es ist also nicht verwunderlich, dass diese Ausgaben auch vom Attention-Netzwerk sehr effizient genutzt werden können, um Klassifikationsaufgaben zu lösen. Fraglich ist dabei, ob die vom Attention-Netzwerk gelernte Gewichtung bzw. Auswahl der Deskriptoren für eine erfolgreiche Klassifikation entscheidend ist, oder ob auf Grund der aussagekräftigen Deskriptoren eine beliebige Auswahl an Deskriptoren genügt, um die Klassifikationsaufgabe zu lösen. Falls die Gewichtung der Deskriptoren für das Training nur einen geringen Einfluss hat, könnte sich dies negativ auf die gelernte Selektionsfähigkeit des Attention-Netzwerks auswirken.

In Abb. 4.20 ist das Auswahlverhalten der Attention-Netzwerke bei unterschiedlichen Extraktionspunkten, an dem Beispiel eines Bildes der Semperoper, dargestellt. Es zeigt sich, dass das Attention-Netzwerk auf Basis der Deskriptoren aus Block-4 deutlich mehr Deskriptoren außerhalb der Semperoper selektiert. Ein Aspekt, der die Auswahl von Deskriptoren außerhalb der interessanten Bildbereiche begünstigt ist, dass bei der Verwendung eines späteren Extraktionspunktes, auf Grund der Netzwerkarchitektur, eine geringere Anzahl an Deskriptoren erzeugt wird. Die negativen Auswirkungen einer geringen Auswahl an Deskriptoren wurde bereits in Abbildung 4.1 auf Seite 23 beobachtet¹². Der ungewöhnliche Trainingsverlauf, lässt außerdem vermuten, dass das Attention-Netzwerk in der Lage ist die Trainingsaufgabe zu lösen, ohne ein geeignetes Selektionsverhalten zu erlernen.

In Abb. 4.21 werden die Retrievalergebnisse bei unterschiedlichem Extraktionspunkt verglichen. Wie bei

¹²Bei einer Extraktion nach Block-4 werden ca. 75% weniger Deskriptoren erzeugt, als bei Block-3. Im Vergleich dazu werden nach einer Skalierung von der Originalgröße(2.5 MPixel) auf 0.6 MPixel, wie in Abb. 4.1 dargestellt, ca. 76% weniger Deskriptoren erzeugt. Die Deskriptorauswahl auf der 0.6 MPixel Version scheint jedoch deutlich besser zu funktionieren, als mit Deskriptoren aus Block-4 in Originalgröße. Folglich scheint die Anzahl an erzeugten Deskriptoren nicht alleiniger Grund für das schlechteren Auswahlverhalten zu sein.

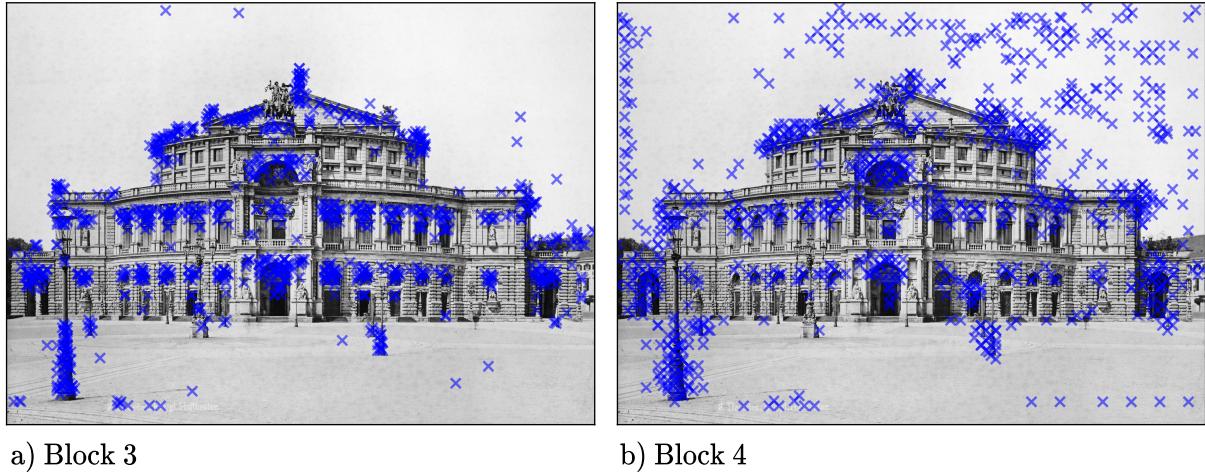


Abbildung 4.20: Unterschiede der Deskriptorselektion bei Verwendung unterschiedlicher Extraktionspunkte.

voran gegangenen Experimenten werden hierbei unterschiedliche Deskriptorlängen und Distanzschwellwerte untersucht. Dabei wird mit allen untersuchten Deskriptorlängen und Distanzschwellwerten, auf beiden untersuchten Datensätzen eine niedrigere mAP erreicht, wenn die verwendeten Deskriptoren aus dem vierten ResNet-Block stammen. Dabei fallen die Performanzunterschiede bei kurzen Deskriptoren besonders stark aus. Bei Verwendung des jeweils besten gefundenen Distanzschwellwerts wird bei einer Deskriptorlänge von 20 eine im Mittel um 0.14 niedrigere mAP auf den historischen Daten erzielt, wenn Deskriptoren aus Block-4 genutzt werden. Bei einer Deskriptorlänge von 200 schrumpft dieser Unterschied auf 0.08. Analog zeigt sich auf Oxford5k bei Deskriptorlänge 20 ein Unterschied der mittleren mAP von 0.20 und bei Länge 200 eine Differenz von 0.05. Die beste gefundene Konfiguration bei Verwendung von Deskriptoren aus Block-4, auf den historischen Daten nutzt eine Deskriptorlänge von 200 und einen Distanzschwellwert von 120. Sie erreicht eine mittlere mAP von 0.49, was um 0.08 niedriger ist, als die beste gefundene Konfiguration mit Deskriptoren aus Block-3. Die beste gefundene Konfiguration für den Oxford5k-Datensatz mit Deskriptoren aus Block-4 nutzt ebenfalls eine Deskriptorlänge von 200 und einen Distanzschwellwert von 1.2 und erreicht eine mittlere mAP von 0.65, was um 0.06 niedriger ist, als die beste gefundene Konfiguration mit Deskriptoren aus Block-3.

Zusammenfassend scheint die Verwendung eines sehr späten Extraktionspunkts zu einer Verschlechterung der Retrievalperformanz zu führen. Es ist naheliegend, dass die Änderungen des Selektionsverhaltens, durch das Training mit Deskriptoren des späten Extraktionspunkts, für einen wesentlichen Teil dieser Verschlechterung verantwortlich sind.

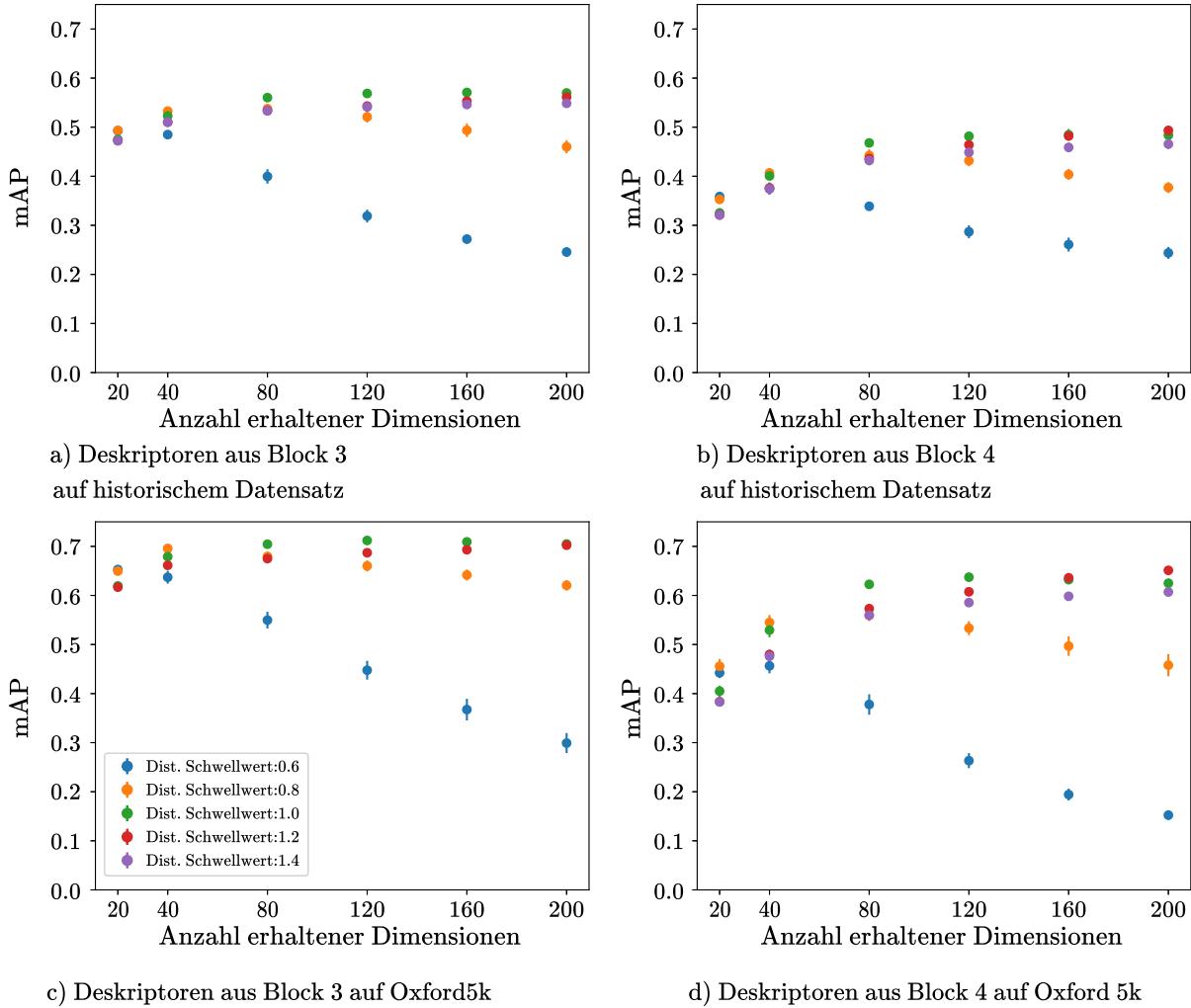


Abbildung 4.21: Erreichte mAP bei unterschiedlicher Deskriptorlänge und Distanzschwellwert unter Verwendung unterschiedlicher Extraktionspunkte. RANSAC wird zur geometrischen Überprüfung von Deskriptormatches eingesetzt.

4.4 Verfahrensvergleich

Mit dem Wissen, um die Einflüsse unterschiedlicher Verfahrensparameter, soll der abschließende Teil der Auswertung dazu dienen, die Ergebnisse des DELF-Verfahrens auf den historischen Daten einzutragen. Hierbei liegt ein besonderes Interesse darin herauszufinden, inwiefern sich das DELF-Verfahren für den Anwendungsfall auf historischen Bildern eignet und welche Stärken und Schwächen es im Vergleich zu anderen Verfahren aufweist. Für den Verfahrensvergleich stehen Ergebnisse des bereits zu Beginn kurz vorgestellten Retrievalsystems von Razavian et. al. [14] (in Folge als ConvNet bezeichnet) zur Verfügung. Das ConvNet-Verfahren nutzt ein tiefes CNN, in diesem Fall eine VGG-16 Architektur, um mehrere lokale Deskriptoren für ein Eingangsbild zu erzeugen. Hierfür werden die Bilder zunächst regelbasiert in 30 unterschiedlich große Teilbereiche zerschnitten. Die Teilbilder durchlaufen das VGG-Netzwerk und aus den Ausgaben der letzten faltenden Schicht wird mit Hilfe von max-pooling für jedes Teilbild ein 2048-dimensionaler Deskriptor erzeugt. Um die Ähnlichkeit zwischen einem Anfragebild

und einem Bild des Suchdatensatzes zu quantifizieren, wird für jeden Deskriptor des Anfragebildes die euklidische Distanz zum ähnlichen Deskriptor des Suchbildes berechnet. Die summierten Distanzen dienen als Ähnlichkeitsmaß für potentielle Bildpaare.

Das DELF-Verfahren nutzt für den Verfahrensvergleich die beste gefundene Konfiguration mit geometrischer Verifizierung, mit einer Deskriptorlänge von 160 und einem Distanzsollwert von 1.0. Da auf den historischen Daten auch ohne geometrische Verifizierung ähnlich gute Ergebnisse erzielt werden, wird außerdem eine Konfiguration mit Deskriptorlänge 200 und Distanzsollwert 1.0 ohne RANSAC betrachtet. Die Deskriptoren entstammen jeweils aus den Ausgaben des dritten ResNet-Blocks. Da sich bei wiederholten Experimenten mit gleicher Konfiguration keine signifikanten Ergebnisunterschiede beobachten lassen, basieren die in Folge präsentierten Ergebnisse jeweils nur auf einzelnen Experimentalläufen.

Ein Aspekt, der für den praktischen Einsatz von Retrievalsystemen bedeutsam ist, in der vorliegenden Arbeit jedoch nur eine untergeordnete Rolle spielt, ist die für eine Suchanfrage benötigte Rechenleistung. Es sei angemerkt, dass das Ziel, der in der vorliegenden Arbeit erstellten Implementierung ist, einen möglichst flexiblen und nachvollziehbaren Prototypen zu erstellen. Daher ist für die praktische Anwendung des Verfahrens noch Optimierungspotential vorhanden. Trotzdem lässt sich die Größenordnung der möglichen Laufzeit mit Hilfe des Prototyps grob einschätzen. Um die Laufzeit der Verfahren zu vergleichen führen alle untersuchten Verfahren Retrielexperimente auf den historischen Daten durch, wobei ihnen je 21 Kerne eines IBM Power9 CPUs und eine NVIDIA VOLTA V100 GPU zur Verfügung stehen. Das DELF-Verfahren unter Verwendung von RANSAC benötigt ca. 54 Sekunden um einen Datensatz von 1000 Bildern mit einem Anfragebild zu vergleichen. Hierbei ist die geometrische Verifikation für den Großteil der benötigten Rechenzeit verantwortlich. Verwendet man DELF ohne geometrische Verifikation dauern 1000 Vergleiche nur ca. 18 Sekunden. Ohne RANSAC ist die Laufzeit von DELF ähnlich, wie die des ConvNet-Verfahrens, welche für 1000 Vergleiche zwischen 14 und 19 Sekunden benötigt.

In Bezug auf die Retrievalergebnisse auf den historischen Daten, erzielen beide Versionen des DELF-

Kategorie	Frauenkirche	Hofkirche	Moritzburg	Semperoper	Sophienkirche	Stallhof	Zwinger	Gesamt
DELF mit RANSAC	0.62	0.74	0.73	0.33	0.34	0.43	0.69	0.56
DELF ohne RANSAC	0.51	0.66	0.86	0.36	0.37	0.61	0.65	0.56
ConvNet	0.53	0.56	0.83	0.22	0.29	0.33	0.78	0.50

Tabelle 4.5: Erreichte mAP der unterschiedlichen Verfahren bzw. Konfigurationen auf den historischen Daten, basierend auf 42 Anfragen. Dargestellt für die einzelnen Kategorien, sowie für alle Anfragen. Achtung, für die einzelnen Kategorien werden unterschiedlich viele Anfragebilder betrachtet (vgl. Tab. 4.1, S. 4.1).

Verfahrens auf den 42 untersuchten Anfrage insgesamt eine um Rund 0.06 höhere mAP als das ConvNet-Verfahren, wobei sich je nach Objektkategorie teilweise starke Unterschiede zeigen (vgl. Tab. 4.5). Hierbei gibt es einige Kategorien, wie die Semperoper oder Sophienkirche, die für alle getesteten Verfahren eine besonders große Herausforderung darstellen. Neben Aspekten wie Auflösung, Aufnahmegerät und Klassenverteilung innerhalb des Datensatzes (vgl. Tab. 4.1) gibt es speziell historische Einflüsse, die die Retrievalaufgabe für diese Kategorien besonders schwierig gestaltet. Betrachtet man die Anfragebilder, welche mit den getesteten Verfahren am schlechtesten beantwortet wurden (vgl. Abb. 4.22), so sieht man, dass sich die hier gezeigten Aufnahmen der Semperoper und des Stallhofs architektonisch deutlich von ihrem heutigen Selbst unterscheiden. Innerhalb des Datensatzes sind unterschiedliche bauliche

Stadien, dieser Sehenswürdigkeiten enthalten, was ein genaues Matching erschwert. Auch zahlreiche Abbildungen von zerstörten Objekten, wie beispielsweise der Sophienkirchen sind ihren intakten Versionen nur schwer zuordenbar.

In Abbildung 4.23 ist der Verlauf einer Suchanfrage der Sophienkirche der unterschiedlichen Verfahren



Abbildung 4.22: Anfragebilder, die unter Verwendung der unterschiedlichen Verfahren die niedrigste AP erreichen. Von links nach rechts, sortiert nach aufsteigender maximal erreichter AP: Semperoper, Sophienkirche, Stallhof, Semperoper.

an Hand von PR-Kurven dargestellt. Bei der Sophienkirche handelt es sich um die Kategorie, welche im Datensatz prozentual am häufigsten (teilweise) zerstört dargestellt ist (vgl. Tab. 4.6). Die Abbildungen der zerstörten Sophienkirchen werden zu einem überwiegenden Teil erst sehr spät innerhalb der Anfrageantworten zurückgegeben, wenn die Precision bereits sehr stark gesunken ist, bzw. schon viele Bilder ohne korrekten Bildinhalt zurückgegeben wurden. Daraus lässt sich schließen, dass Aufnahmen von zerstörten Gebäuden tatsächlich sehr viel schwerer ihren intakten Originalen zugeordnet werden können, als andere unbeschädigte Aufnahmen ihrer selbst.

Zerstörung und architektonischer Wandel sind dabei nicht die einzigen Aspekte, die den historischen

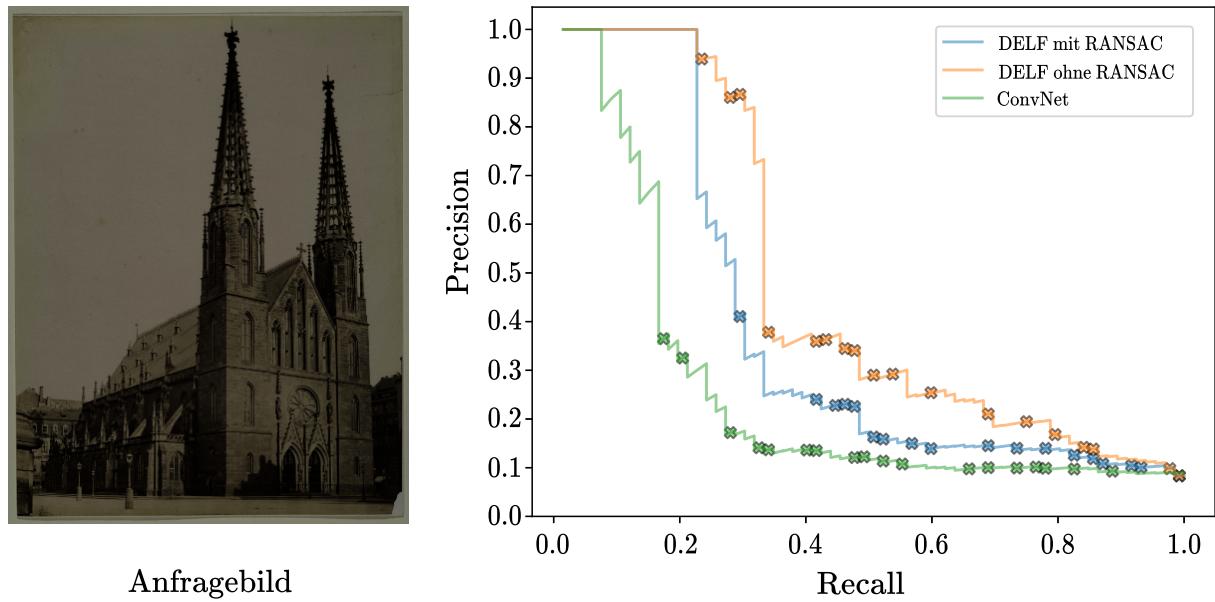


Abbildung 4.23: Anfrageantworten der unterschiedlichen Verfahren auf das Anfragebild, dargestellt als PR-Kurven. Kreuze heben hervor, wann Bilder mit korrektem Bildinhalt zurückgegeben werden, bei denen das Objekt teilweise zerstört dargestellt ist.

Datensatz besonders anspruchsvoll für das Image Retrieval machen. Auch ein großer Anteil an Nacht- und Luftaufnahmen machen die Suche besonders schwierig. Bei in Dunkelheit aufgenommene Bildern,

wie in Abbildung 4.24 a zu sehen, lässt sich durch den geringen Kontrast deutlich schwerer erkennen, welche Teile des Bildes relevante Inhalte darstellen, was die Auswahl geeigneter Deskriptoren erschwert. Auch sind viele Details kaum zu erkennen, die möglicherweise für die Erstellung von aussagekräftigen Deskriptoren genutzt werden könnten. Generell ist davon auszugehen, dass sich Deskriptoren aus Nachbildern deutlich von Deskriptoren aus Aufnahmen bei Tageslicht unterscheiden. In Abbildung 4.25 sind



a) Nachtaufnahme



b) Luftaufnahme

Abbildung 4.24: Beispiele für besonders herausfordernde Aufnahmen im historische Datensatz, in Form von Nacht- und Luftaufnahmen.

PR-Kurven zu einer Suchanfrage nach Bildern des Zwingers dargestellt. Insbesondere bei Verwendung von DELF zeigt sich das Nachtaufnahmen des Zwinger, dargestellt durch Kreuze, fast ausschließlich unter den zuletzt zurückgegebenen Bildern zu finden sind. Interessanterweise scheint das ConvNet-Verfahren deutlich besser in der Lage zu sein Nachtaufnahmen mit der Anfrage zu matchen. Die Nachbilder werden relativ gleichmäßig über den Verlauf der PR-Kurve zurückgegeben, was darauf schließen lässt, dass die Belichtung hier eine deutlich geringere Rolle für den Matchingerfolg spielt. Warum dies so ist, lässt sich aktuell nur schwer feststellen, es erklärt allerdings, warum das ConvNet-Verfahren ausgerechnet für die Kategorie des Zwingers deutlich bessere Ergebnisse, als das DELF-Verfahren erzielt (vgl. 4.5). So findet sich in der Kategorie Zwinger der größte Anteil an Nachtaufnahmen im historischen Datensatz (vgl. Tab. 4.6). Luftaufnahmen, wie in Abbildung 4.24 b zu sehen, stellen aus mehreren

Kategorie	Frauenkirche	Hofkirche	Moritzburg	Semperoper	Sophienkirche	Stallhof	Zwinger
Luftaufnahme	16.5%	18.5%	0.0%	22.5%	7.6%	0.0%	3.7%
Zerstörung	2.9%	9.7%	0.0%	4.5%	27.3%	7.9%	1.1%
Nachtaufnahme	11.7%	5.6%	0.0%	1.1%	4.5%	0.0%	15.0%

Tabelle 4.6: Anteil an Objektdarstellungen im historischen Datensatz, die unterschiedliche erschwerenden Besonderheiten vorweisen.

Gründen eine Herausforderung für Retrievalverfahren dar. Zunächst zeigen sie die abgebildeten Objekte

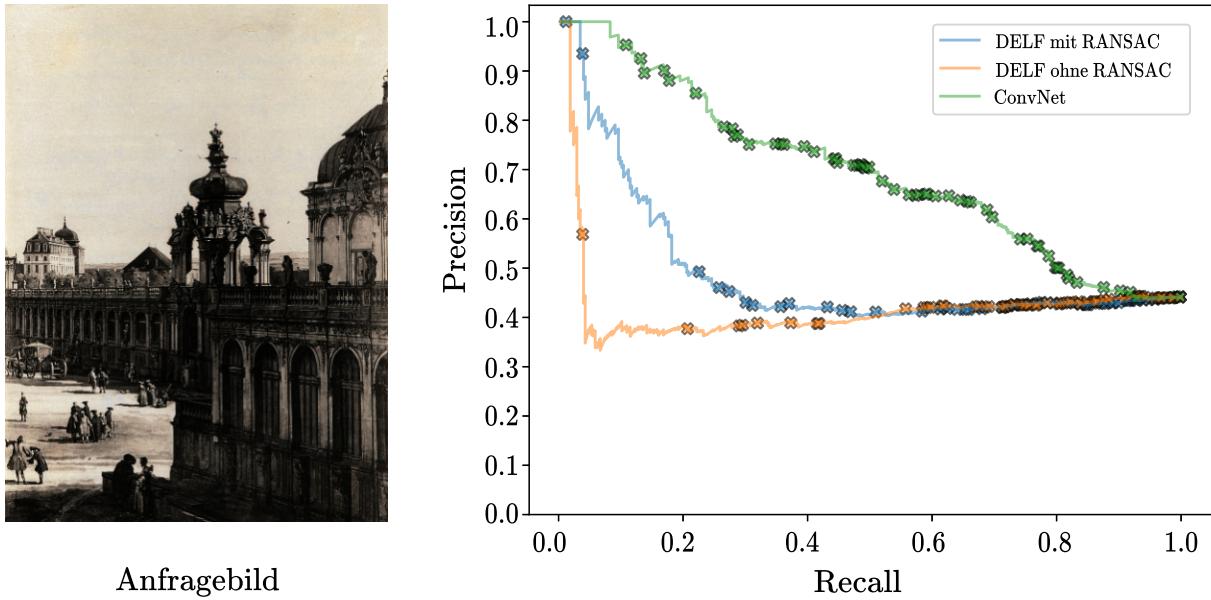


Abbildung 4.25: Anfrageantworten der unterschiedlichen Verfahren auf das Anfragebild, dargestellt als PR-Kurven. Kreuze heben hervor, wann Nachtaufnahmen mit korrektem Bildinhalt zurückgegeben werden.

aus einer ungewöhnlichen Perspektive, sodass Objektteile in den Fokus gerückt werden, die bei einer Frontalaufnahme nicht zu sehen sind. Zugleich werden große Teile der Frontsicht durch Dächer und Türe der Gebäude verdeckt. Luftaufnahmen zeigen auch meist nicht nur das gesuchte Objekt, sondern einen großen Ausschnitt der Umgebung, die häufig viele weitere markante Gebäude und Strukturen enthält. In den Luftaufnahmen des historischen Datensatzes sind auf Grund der Dresdner Stadtstruktur häufig sogar mehrere der untersuchten Kategorien zu sehen. Der jeweils gesuchte Bildinhalt nimmt bei Luftaufnahmen meist nur einen kleinen Teil des Gesamtbildes ein. Dies führt nicht nur dazu, dass weniger Details des Objektes sichtbar sind, sondern auch das ein Großteil der Deskriptoren, die aus einem Luftbild extrahiert werden, nicht das gesuchte Objekt beschreiben. Mit Hilfe des Attention-Netzwerks kann DELF zwar strukturarme Bildbereiche, wie z.B. Abschnitte der Elbe ausgeschlossen werden. Das Selektionsverfahren kann allerdings nicht entscheiden, welche Gebäude für die aktuelle Suchanfrage interessant sind. Das ConvNet-Verfahren, welches mit einer regelbasierten Segmentierung des Bildinhaltes arbeitet, hat keine Möglichkeit Bildbereiche auszuschließen. In Abbildung 4.26 sind PR-Kurven zu einer Suchanfrage der Frauenkirche dargestellt, bei denen zurückgegebene Luftaufnahmen hervorgehoben werden, welche die Frauenkirche enthalten.

Es zeigt sich deutlich, dass Luftaufnahmen für alle untersuchten Verfahren sehr schwierig zu matchen sind. Wie erwartet hat das ConvNet-Verfahren, mit seinem segmentierenden Ansatz jedoch die größten Schwierigkeiten Luftaufnahmen zu matchen. Das DELF-Verfahren profitiert bei der Betrachtung deutlich von der geometrischen Verifikation. Bei der Verwendung von RANSAC ist die Anzahl an verifizierten Matchings auch zwischen Bildern mit gleichem Bildinhalt typischerweise deutlich niedriger, als bei einer reinen Betrachtung der initialen Matchanzahl. Daher kann auch mit einer geringen Anzahl an relevanten Deskriptoren eine hohe RANSAC-Bewertung erzielt werden. Da die relevanten Deskriptoren der betrachteten Luftbilder auf einen kleinen Bildbereich eingegrenzt sind und daher nah beieinander liegen,

ist es außerdem leichter eine geeignete Transformation zu finden, die einen großen Teil der relevanten Deskriptoren innerhalb des RANSAC-Schwellwerts erklären kann. Betrachtet man die erzielte mAP der

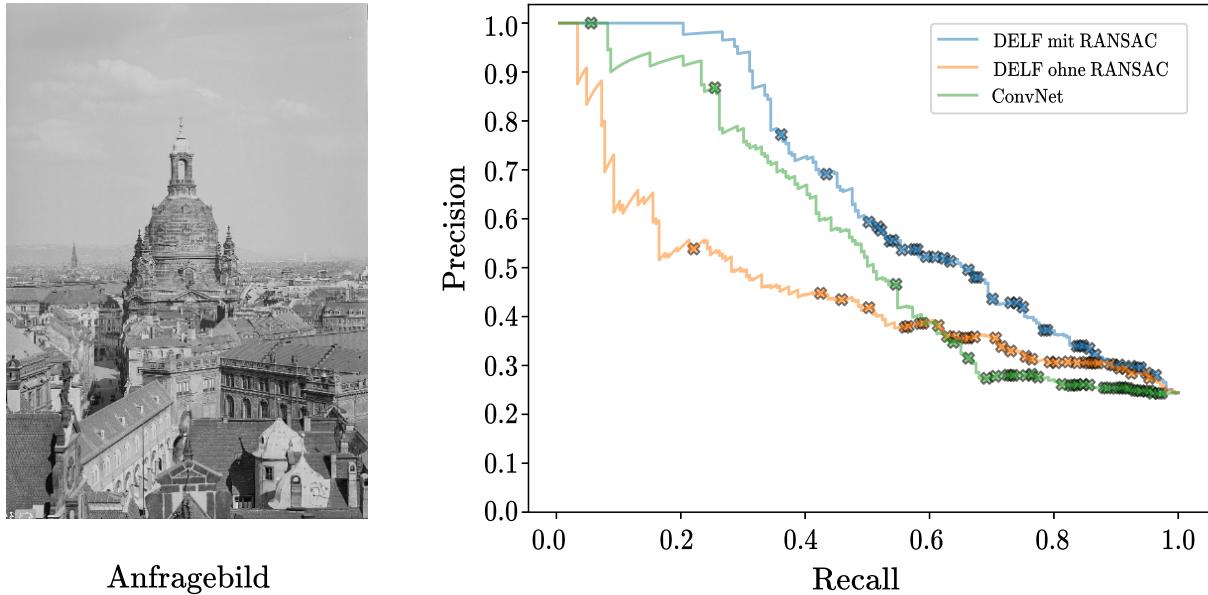


Abbildung 4.26: Anfrageantworten der unterschiedlichen Verfahren auf das Anfragebild, dargestellt als PR-Kurven. Kreuze heben hervor, wann Luftaufnahmen mit korrektem Bildinhalt zurückgegeben werden.

Verfahren in den unterschiedlichen Kategorien (vgl. Tab. 4.5) in Zusammenhang mit den Anteilen an besonders herausfordernden Bildern, je nach Kategorie (vgl. Tab. 4.6), so zeigt sich, dass DELF besonders bei einem hohen Anteil ein Luftaufnahmen (Semperoper, Hofkirche, Frauenkirche) deutlich besser als das ConvNet-Verfahren funktioniert. In der Kategorie der Frauenkirche fällt dieser Effekt, auf Grund des hohen Anteils an Nachtaufnahmen möglicherweise geringer aus.

5 Fazit und Ausblick

Ziel der vorliegenden Arbeit ist es zu untersuchen, ob das Image Retrieval System DELF für die inhaltsbasierte Suche auf historischen Daten geeignet ist und welche Stärken und Schwächen es im Vergleich zu alternativen Verfahren aufweist. Weiterhin sollen die Einflüsse unterschiedlicher Parameter innerhalb der DELF-Pipeline analysiert werden, um diese für die Domäne der historischen Aufnahmen zu optimieren.

5.1 Fazit

Es zeigt sich, dass das DELF-Verfahren bei der Anwendung auf historischen Aufnahmen nicht nur eine deutliche Verbesserung gegenüber einer zufälligen Suche darstellt, sondern in den meisten betrachteten Bilderkategorien auch signifikant bessere Ergebnisse als das Vergleichsverfahren ConvNet erzielt. Trotz des relativ hohen Rechenaufwands, der insbesondere durch den zusätzlichen Schritt der geometrischen Verifikation entsteht, ist DELF damit aktuell ein vielversprechendes Verfahren, für den Einsatz im HistStadt4D-Projekt.

Eine Stärke des DELF-Verfahrens ist sein Umgang mit Luftaufnahmen, welche einen besonders schwierigen Aspekt des historischen Datensatzes bilden. So zeigen sich in Kategorien die häufig in Luftaufnahmen gezeigt werden besonders große Performanzvorteile des DELF-Verfahrens gegenüber ConvNet.

Als problematisch hat sich für das DELF-Verfahren der Umgang mit schlecht belichteten, bzw. NachtAufnahmen erwiesen. So ist DELF kaum in der Lage gleiche Bildinhalte, bei sehr unterschiedlichen Belichtungsverhältnissen zu erkennen. Das ConvNet-Verfahren scheint hingegen kaum von dieser Problematik betroffen zu sein.

Sowohl die DELF, wie auch ConvNet können in Aufnahmen aus unterschiedlichen zeitlichen Perioden, in denen sich die betrachteten Objekte durch Umbauten, oder Zerstörung stark voneinander unterscheiden, kaum gemeinsame Bildinhalte feststellen. Aktuell sind uns keine Möglichkeiten bekannt, mit solchen Veränderungen effektiv umzugehen.

Objekte, die eine Vielzahl von sich ähnelnden Elementen wie Fenster oder Torbögen enthalten, sind für DELF, insbesondere wegen des geometrische Verifikationsschritts mittels RANSAC, problematisch. Diese sich wiederholenden Bildinhalte werden, auf Grund ihrer Ähnlichkeit zueinander, oft falschen Instanzen in dem betrachteten Vergleichsbildern zugeordnet, was zu geometrisch nicht erklärbaren Korrespondenzen führt. In der Parameteranalyse hat sich gezeigt, dass diese Problematik auf dem historischen Datensatz so ausgeprägt ist, dass das DELF-Verfahren hier auch ohne geometrische Verifikation gleich gute Ergebnisse erzielt. Die Vergleichsanalyse auf dem Benchmark-Datensatz Oxford5k zeigt sogleich, dass sich, mittels RANSAC, bei der Betrachtung von Objekten mit weniger Wiederholungen eine signifikante Verbesserung der Retrievalperformanz erreichen lässt.

Die Analyse unterschiedlicher Extraktionspunkte für die Erstellung von Deskriptoren hat ergeben, dass sich die letzten Schichten des ResNets nicht für die Extraktion von DELF-Deskriptoren eignen. Insbesondere das zur Selektion der Deskriptoren trainierte Attention-Netzwerk hat Schwierigkeiten geeignete

Deskriptoren auszuwählen, wenn es auf den Ausgaben des letzten ResNet-Blocks trainiert wurde. Bei einer Extraktion aus dem vorletzten ResNet-Block, wie von den DELF-Autoren empfohlen, werden deutlich bessere Retrievalergebnisse erzielt.

Ein weiterer betrachteter Parameter ist die verwendete Deskriptorlänge. Dabei zeigt sich, dass durch die Erstellung längerer Deskriptoren geringfügige Verbesserungen der Retrievalperformanz erzielt werden können. Dieser Effekt schwächt sich mit wachsender Länge der Deskriptoren deutlich ab. Zusätzlich steigt der benötigte Speicher- und Rechenbedarf mit der Länge der Deskriptoren an. Für einen geeigneten Kompromiss zwischen Retrievalperformanz und Speicherbedarf, bietet sich je nach verwendetem Datensatz eine Deskriptorlänge zwischen 40 und 80 Dimensionen an.

Die zu Beginn durchgeführte Hyperparameteranalyse zur Optimierung der beiden Trainingsphasen des DELF-Verfahrens hat ergeben, dass die hier untersuchten Parameter in einem relativ großen Spektrum gewählt werden können, ohne die Trainingsergebnisse stark zu beeinflussen. Unter fast allen getesteten Konfigurationen, sowohl im Fine-Tuning, wie auch im Attention-Training, konnte das DELF-Netzwerk die gestellte Klassifikationsaufgabe mit sehr hoher Genauigkeit lösen. Hierbei ist es möglich, dass der relative kleine verwendete Trainingsdatensatz bestehend aus nur knapp vierzigtausend Bildern, keine ausreichende Herausforderung für die genutzte Modellarchitektur darstellt, um größere Unterschiede der Trainingsperformanz, auf Grund der Trainingsparameter, aufzuzeigen.

5.2 Ausblick

Literaturverzeichnis

- [1] Ferdinand Maiwald, Jonas Bruschke, Christoph Lehmann, and Florian Niebling. A 4D Information System for the Exploration of Multitemporal Images and Maps using Photogrammetry, Web Technologies and VR/AR. *Virtual Archaeology Review*, 10:1, 07 2019.
- [2] Hyeonwoo Noh, Andre Araujo, Jack Sim, Tobias Weyand, and Bohyung Han. Large-Scale Image Retrieval with Attentive Deep Local Features. pages 3476–3485, 10 2017.
- [3] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2007.
- [4] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. Lost in Quantization: Improving Particular Object Retrieval in Large Scale Image Databases. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.
- [5] Yan-Tao Zheng, Ming Zhao, Yang Song, Hartwig Adam, Ulrich Buddeleier, Alessandro Bissacco, Fernando Brucher, Tat-Seng Chua, Hartmut Neven, and Jay Yagnik. Tour the World: A Technical Demonstration of a Web-Scale Landmark Recognition Engine. In *Proceedings of the 17th ACM International Conference on Multimedia*, MM '09, page 961–962, New York, NY, USA, 2009. Association for Computing Machinery.
- [6] Arnold W. M. Smeulders, Marcel Worring, Simone Santini, Amarnath Gupta, and Ramesh Jain. Content-Based Image Retrieval at the End of the Early Years. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(12):1349–1380, December 2000.
- [7] David Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60:91–, 11 2004.
- [8] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. SURF: Speeded up Robust Features. volume 3951, pages 404–417, 07 2006.
- [9] Yan Ke and Rahul Sukthankar. PCA-SIFT: A more Distinctive Representation for Local Image Descriptors. volume 2, pages II–506, 05 2004.
- [10] Cordelia Schmid and J. Ponce. Semi-Local Affine Parts for Object Recognition. *BMVC04*, 08 2004.
- [11] Miaojing Shi, Yannis Avrithis, and Hervé Jégou. Early Burst Detection for Memory-Efficient Image Retrieval. 06 2015.
- [12] Matthew Zeiler and Rob Fergus. Visualizing and Understanding Convolutional Neural Networks. volume 8689, 11 2013.

- [13] Artem Babenko, Anton Slesarev, Alexandre Chigorin, and Victor Lempitsky. Neural Codes for Image Retrieval. volume 8689, 04 2014.
- [14] Ali S. Razavian, Josephine Sullivan, Stefan Carlsson, and Atsuto Maki. Visual Instance Retrieval with Deep Convolutional Networks. *ITE Transactions on Media Technology and Applications*, 4(3):251–258, 2016.
- [15] Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton. ImageNet Classification with Deep Convolutional Neural Networks. *Neural Information Processing Systems*, 25, 01 2012.
- [16] Lingxi Xie, Richang Hong, Bo Zhang, and Qi Tian. Image classification and retrieval are one. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, ICMR ’15, page 3–10, New York, NY, USA, 2015. Association for Computing Machinery.
- [17] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv 1409.1556*, 09 2014.
- [18] J. R. Uijlings, K. E. Sande, T. Gevers, and A. W. Smeulders. Selective Search for Object Recognition. *Int. J. Comput. Vision*, 104(2):154–171, September 2013.
- [19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. 7, 12 2015.
- [20] Filip Radenović, Giorgos Tolias, and Ondřej Chum. CNN Image Retrieval Learns from BoW: Unsupervised Fine-Tuning with Hard Examples. volume 9905, pages 3–20, 10 2016.
- [21] Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton. ImageNet Classification with Deep Convolutional Neural Networks. *Neural Information Processing Systems*, 25, 01 2012.
- [22] Andrew McCallum and Kamal Nigam. A comparison of event models for naive bayes text classification. In *AAAI 1998*, 1998.
- [23] James B. MacQueen. Some Methods for Classification and Analysis of Multivariate Observations. 1967.
- [24] Hervé Jégou, Matthijs Douze, Cordelia Schmid, and Patrick Perez. Aggregating Local Descriptors into a Compact Image Representation. pages 3304 – 3311, 07 2010.
- [25] Jingdong Wang, Heng Shen, Jingkuan Song, and Jianqiu Ji. Hashing for Similarity Search: A Survey. 08 2014.
- [26] Tobias Weyand, Andre Araujo, Bingyi Cao, and Jack Sim. Google Landmarks Dataset v2 – A Large-Scale Benchmark for Instance-Level Recognition and Retrieval, 2020.
- [27] Shuhei Yokoo, Kohei Ozaki, Edgar Simo-Serra, and Satoshi Iizuka. Two-stage Discriminative Re-ranking for Large-scale Landmark Retrieval, 2020.
- [28] Hervé Jégou and Ondřej Chum. Negative Evidences and Co-occurrences in Image Retrieval: The Benefit of PCA and Whitening. pages 774–787, 10 2012.

- [29] Jerome Friedman, Jon Bentley, and Raphael Finkel. An Algorithm for Finding Best Matches in Logarithmic Expected Time. *ACM Trans. Math. Softw.*, 3:209–226, 09 1977.
- [30] Piotr Indyk. Nearest Neighbors in High-Dimensional Spaces. In Jacob E. Goodman and Joseph O'Rourke, editors, *Handbook of Discrete and Computational Geometry, Second Edition*, pages 877–892. Chapman and Hall/CRC, 2004.
- [31] Martin A. Fischler and Robert C. Bolles. Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. *Commun. ACM*, 24(6):381–395, June 1981.
- [32] Hervé Jégou, Matthijs Douze, and Cordelia Schmid. Product Quantization for Nearest Neighbor Search. *IEEE transactions on pattern analysis and machine intelligence*, 33:117–28, 01 2011.
- [33] J. Bergstra, D. Yamins, and D. D. Cox. Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures. In *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, ICML’13, page I–115–I–123. JMLR.org, 2013.