



Bewertung der Auswirkungen alternativer Farbräume auf die Bildklassifizierung

Bachelor Thesis

Bachelor of Science in Applied Computer Science

Giresse Ngansop

14.09.2026

Supervisor:

1st: Prof. Dr. Christian Ledig
2nd: Sebastian Dörrich, M.Sc.

Chair of Explainable Machine Learning
Faculty of Information Systems and Applied Computer Sciences
Otto-Friedrich-University Bamberg

Abstract

Almost all computer vision models process images in the RGB color space – a choice determined by acquisition hardware rather than by the task. Its three channels are strongly correlated, since a change in illumination shifts all of them at once. Decorrelated color spaces such as CIELAB separate luminance from chrominance by construction, which suggests that they should make a network more robust to illumination-related corruptions.

This thesis tests that hypothesis. Two architectures – a ResNet-18 and a VGG-11 – are trained from scratch on CIFAR-10 in four input representations (RGB, CIELAB, HSV and grayscale) under otherwise identical conditions, each replicated across five random seeds and evaluated on the fifteen standard corruptions of CIFAR-10-C. Three further experiments complement this comparison: an eigenvalue analysis of the learned first-layer filters, a variation of the normalization layer across three levels, and a targeted manipulation of the input normalization. The replication on the second architecture follows four predictions registered in writing beforehand.

The hypothesis is not supported. CIELAB leads RGB by 0,21 percentage points on illumination-related corruptions, a difference that changes sign across seeds and is smaller than the spread between two runs of the same condition; after correction for multiple testing, no robustness comparison between color spaces remains significant, and the result carries over to the second architecture. The grayscale control, by contrast, loses 6,97 percentage points in every configuration, confirming that the design is sensitive enough to detect a genuine difference. The filter analysis explains why: a network trained on RGB constructs a luminance-chrominance decomposition in its first layer by itself, independently of the input format. This null result holds for the customary per-channel standardization; under a variant that does not stretch the chroma channels, a small advantage in the predicted direction appears at all five seeds without reaching significance.

The aggregate null result nevertheless conceals two opposing effects. Broken down by corruption type, CIELAB loses 9,72 percentage points on noise and gains 4,64 on blur, consistently across all seeds; the two cancel in the mean, and the same pattern reappears on the second architecture with almost twice the amplitude. A targeted experiment identifies the cause: unifying the per-channel scaling of the input normalization – leaving the centering untouched – reduces the amplitude of the trade-off by about sixty percent, while the RGB control condition remains unchanged. The dominant effect, finally, concerns neither the color space nor the architecture as such: whether a normalization layer is present, and which one, changes robustness by roughly six percentage points in both architectures – some ten times the effect of the input format.

The input color space therefore changes little about *what* a network learns, but a great deal about how a corruption appears before the network processes it – and that effect arises mostly in the normalization that follows the color space transformation, a step usually treated as neutral. For robustness, the architecture is the more effective lever.

Zusammenfassung

Nahezu alle Verfahren des maschinellen Sehens verarbeiten Bilder im RGB-Farbraum – eine Wahl, die von der Aufnahmetechnik bestimmt wird und nicht von der Aufgabe. Seine drei Kanäle sind stark korreliert, da eine Beleuchtungsänderung sie gemeinsam verschiebt. Dekorrelierte Farbräume wie CIELAB trennen Helligkeit und Farbe konstruktiv voneinander, was nahelegt, dass sie ein Netz robuster gegenüber beleuchtungsbedingten Störungen machen sollten.

Die vorliegende Arbeit prüft diese Hypothese. Zwei Architekturen – ein ResNet-18 und ein VGG-11 – werden auf CIFAR-10 in vier Eingaberepräsentationen (RGB, CIELAB, HSV und Graustufen) unter sonst identischen Bedingungen von Grund auf trainiert, jeweils über fünf Zufallsstartwerte wiederholt und auf den fünfzehn Standardstörungen von CIFAR-10-C ausgewertet. Drei weitere Experimente ergänzen diesen Vergleich: eine Eigenwertanalyse der gelernten Filter der ersten Faltungsschicht, eine Variation der Normalisierungsschicht über drei Stufen sowie ein gezielter Eingriff in die Eingabenormalisierung. Der Wiederholung auf der zweiten Architektur liegen vier vorab schriftlich festgehaltene Vorhersagen zugrunde.

Die Hypothese bestätigt sich nicht. CIELAB liegt auf den beleuchtungsbedingten Störungen um 0,21 Prozentpunkte vor RGB; dieser Unterschied wechselt zwischen den Startwerten das Vorzeichen und ist kleiner als die Streuung zwischen zwei Läufen derselben Bedingung. Nach Korrektur für multiples Testen bleibt kein Robustheitsvergleich zwischen den Farbräumen signifikant, und der Befund überträgt sich auf die zweite Architektur. Die Graustufen-Kontrolle verliert dagegen in jeder Konfiguration 6,97 Prozentpunkte, was belegt, dass das Versuchsdesign einen tatsächlich vorhandenen Unterschied zu messen vermag. Die Filteranalyse erklärt den Befund: Ein auf RGB trainiertes Netz konstruiert in seiner ersten Schicht selbständig eine Luminanz-Chrominanz-Zerlegung, und zwar unabhängig vom Eingabeformat. Der Nullbefund gilt dabei für die übliche kanalweise Standardisierung; unter einer Variante, welche die Chromakanäle nicht spreizt, zeigt sich bei allen fünf Startwerten ein geringer Vorsprung in der vorhergesagten Richtung, ohne das Signifikanzniveau zu erreichen.

Der Nullbefund im Mittelwert verdeckt gleichwohl zwei gegenläufige Effekte. Nach Störungsart aufgeschlüsselt verliert CIELAB 9,72 Prozentpunkte auf Rauschstörungen und gewinnt 4,64 auf Unschärfestörungen, über alle Startwerte hinweg konsistent; im Mittel heben sich beide auf, und auf der zweiten Architektur tritt dasselbe Muster mit nahezu doppelter Amplitude auf. Ein gezieltes Experiment weist die Ursache nach: Wird die kanalweise Skalierung der Eingabenormalisierung vereinheitlicht – bei unveränderter Zentrierung –, so sinkt die Amplitude des Trade-offs um rund sechzig Prozent, während die Kontrollbedingung RGB unverändert bleibt. Der deutlichste Effekt betrifft schließlich weder den Farbraum noch die Architektur als solche: Ob eine Normalisierungsschicht vorhanden ist und welche, verändert die Robustheit in beiden Architekturen um rund sechs Prozentpunkte – etwa das Zehnfache dessen, was das Eingabeformat bewirkt.

Der Eingabefarbraum verändert somit wenig daran, *was* ein Netz lernt, wohl aber erheblich, wie eine Störung aussieht, bevor das Netz sie verarbeitet – und diese Wirkung entsteht überwiegend erst in der Normalisierung, die auf die Farbraumtransformation folgt und üblicherweise als neutraler Zwischenschritt behandelt wird. Für die Robustheit ist die Architektur der wirksamere Hebel.

Danksagung

Mein Dank gilt zuallererst den Menschen in meinem persönlichen Umfeld – für die anregenden Gespräche und für den Rückhalt während der vergangenen Monate.

Ebenso danke ich Sebastian Dörrich für die Betreuung dieser Arbeit: für die vielen frischen Ideen, für die konstruktive Kritik und für die Geduld, mit der er meine Fragen beantwortet hat. Prof. Dr. Christian Ledig danke ich für die Möglichkeit, diese Arbeit an seinem Lehrstuhl anzufertigen.

Dem Lehrstuhl für Erklärbares Maschinelles Lernen danke ich für die Bereitstellung der Rechenressourcen, ohne die die hier berichteten Experimente nicht möglich gewesen wären.

Der vollständige Quellcode, mit dem die berichteten Experimente durchgeführt wurden, sowie die Konfigurationsdateien, die Rohergebnisse und die Auswertungsskripte sind öffentlich zugänglich unter:

`https://github.com/giressengansop/decol-gen`

Inhaltsverzeichnis

Abbildungsverzeichnis **x**

Tabellenverzeichnis **xi**

Abkürzungsverzeichnis **xii**

Notation **xiii**

1 Einleitung **1**

1.1 Motivation 1

1.2 Forschungsfrage und Hypothese 2

1.3 Aufbau der Arbeit 2

2 Verwandte Arbeiten **3**

2.1 Farbräume in der Bildklassifikation 3

2.2 Robustheit, Normalisierung und gelernte Farbkodierung 3

2.3 Abgrenzung dieser Arbeit 4

3 Beitrag dieser Arbeit **5**

3.1 Aufschlüsselung der Robustheit nach Störungsart 5

3.2 Nachweis eines gegenläufigen Effekts 5

3.3 Direkte Messung der gelernten Farbverarbeitung 5

3.4 Wiederholung anhand vorab festgelegter Vorhersagen 6

4 Grundlagen **6**

4.1 Farbräume und Kanalkorrelation 6

4.1.1 RGB und die Korrelation seiner Kanäle 6

4.1.2 HSV 8

4.1.3 CIELAB 9

4.1.4 Graustufen 9

4.1.5 Was Dekorrelation bedeutet – und was nicht 9

4.2 Convolutional Neural Networks und ResNet-18 10

4.2.1 Die Faltung als gewichtete Kanalmischung 10

4.2.2 ResNet-18 11

4.3 Normalisierungsschichten: BatchNorm und GroupNorm 11

4.3.1 Batch-Normalisierung 12
4.3.2 Gruppen-Normalisierung 13
4.3.3 Warum dieser Vergleich für die Fragestellung geeignet ist . . . 13
4.4 Robustheit gegenüber Bildstörungen 13
4.4.1 Warum Testgenauigkeit nicht genügt 13
4.4.2 CIFAR-10-C 14
4.4.3 Kennzahlen 14
4.5 Statistische Methoden 14
4.5.1 Gepaarte Vergleiche 15
4.5.2 Der p-Wert und seine Grenzen 15
4.5.3 Multiples Testen und die Holm-Korrektur 16
4.5.4 Vorab festgelegte und nachträgliche Analysen 16

5 Methodik **16**
5.1 Datenpipeline und Farbraumtransformationen 16
5.1.1 Aufbau der Pipeline 17
5.1.2 Die vier Farbräume 18
5.2 Normalisierung der Eingaben 18
5.2.1 Vier Normalisierungsschemata 19
5.2.2 Bestimmung der Konstanten 19
5.3 Modellarchitektur und Training 20
5.3.1 ResNet-18 und die Anpassung an 32×32 20
5.3.2 VGG-11 als zweite Architektur 21
5.3.3 Training ohne Vortraining 21
5.3.4 Reproduzierbarkeit 22
5.3.5 Modellauswahl 22
5.4 Evaluationsprotokoll 22
5.4.1 Genauigkeit auf ungestörten Daten 22
5.4.2 Robustheit auf CIFAR-10-C 23
5.4.3 Kennzahlen 23
5.5 Statistisches Protokoll 24
5.5.1 Wiederholung über Zufallsstartwerte 24
5.5.2 Testverfahren 25
5.5.3 Korrektur für multiples Testen 25
5.5.4 Vorab festgelegte und nachträgliche Analysen 25
5.5.5 Systematische Wiederholung der Prüfungen 26

6 Experimente und Ergebnisse **26**

6.1 Genauigkeit auf unveränderten Testdaten 26

6.1.1 Konvergenzverhalten 27

6.2 Die Haupthypothese bestätigt sich nicht: RGB vs. CIELAB 28

6.2.1 Der vorhergesagte Vorteil bleibt aus 29

6.2.2 Der Befund ist stabil 29

6.2.3 Einordnung des Nullbefunds 29

6.3 Der versteckte Trade-off: Rauschen gegen Unschärfe 30

6.3.1 Zwei gegenläufige Effekte 30

6.3.2 Eine mechanistische Erklärung 31

6.3.3 Einordnung 32

6.4 Ursache des Trade-offs: gezielte Variation der Kanalskalierung 33

6.4.1 Vermutung und Versuchsaufbau 33

6.4.2 Die Vermutung bestätigt sich 33

6.4.3 Eine Beobachtung zur Haupthypothese 34

6.5 HSV im Vergleich zu RGB und CIELAB 35

6.6 Test der BatchNorm-Erklärung: GroupNorm 36

6.6.1 Vorhersage und Versuchsaufbau 36

6.6.2 Die Vorhersage bestätigt sich nur teilweise 36

6.6.3 Die Normalisierungsschicht wirkt stärker als der Farbraum . . 37

6.6.4 Zwischenfazit 38

6.7 Analyse der gelernten Eingangsfilter 38

6.7.1 Vorgehen 38

6.7.2 Die Kontrolle ist erfüllt 38

6.7.3 Das Netz entkoppelt die Farbkanäle 39

6.7.4 Das Ergebnis hängt weder vom Eingabeformat noch von der Normalisierungsschicht ab 40

6.7.5 Grenzen dieser Analyse 40

6.8 Wiederholung auf einer zweiten Architektur 41

6.8.1 Vorab festgelegte Vorhersagen 41

6.8.2 Versuchsaufbau 41

6.8.3 V1 bis V3 bestätigen sich 42

6.8.4 V4 bestätigt sich nicht 42

6.8.5 Ein Nebenbefund zur Normalisierungsschicht 43

6.9 Robustheits-Checks . . . 44
6.9.1 Korrektur für multiples Testen . . . 44
6.9.2 Die zurückgehaltenen Validierungsstörungen . . . 44
6.9.3 Variation der Eingabenormalisierung . . . 45
6.9.4 Negativkontrolle . . . 45

7 Diskussion **45**
7.1 Der Farbraum verändert die Störungsgeometrie, nicht das Lernen . . . 45
7.1.1 Warum eine dekorrelierte Eingabe keinen Vorteil bringt . . . 46
7.1.2 Wo der Farbraum dennoch wirkt . . . 46
7.1.3 Prüfung der Erklärung . . . 47
7.1.4 Die Rolle der Normalisierungsschicht . . . 47
7.1.5 Zusammenfassung . . . 48
7.2 Einordnung in die Literatur . . . 49
7.2.1 Der Nullbefund ist keine Anomalie . . . 49
7.2.2 Die Filteranalyse bestätigt einen bekannten Befund und erweitert ihn . . . 49
7.2.3 Eine Lücke: Robustheit wird aggregiert berichtet . . . 50
7.2.4 Der Bezug zum Form-Textur-Bias . . . 50
7.2.5 Die Normalisierungsschicht als der wirksamere Hebel . . . 50
7.3 Limitationen . . . 51
7.3.1 Reichweite des Versuchsaufbaus . . . 51
7.3.2 Hyperparameter und Batchgröße . . . 51
7.3.3 Status der einzelnen Befunde . . . 52
7.3.4 Grenzen der Filteranalyse . . . 52
7.3.5 Modellauswahl auf dem Testdatensatz . . . 52
7.3.6 Störungsrobustheit ist nicht Domänenverallgemeinerung . . . 53

8 Fazit und Ausblick **53**
8.1 Fazit . . . 53
8.1.1 Die Arbeitshypothese bestätigt sich nicht . . . 53
8.1.2 Der Farbraum wirkt dennoch – an anderer Stelle . . . 54
8.1.3 Die Architektur ist der wirksamere Hebel . . . 54
8.2 Ausblick . . . 55
8.2.1 Die vollständige Dosis-Wirkungs-Kurve der Kanalskalierung . . . 55

8.2.2 Die Rolle der bildweisen Normalisierung 55
8.2.3 Eine Verlustfunktion auf dem Helligkeitskanal 55
8.2.4 Übertragung auf reale Domänenverschiebungen 56

A Vollständige Ergebnistabellen **57**

B Vorab festgelegte Vorhersagen **61**

C Software und Konfiguration **62**

D Reproduzierbarkeit **62**

Literaturverzeichnis **64**

Abbildungsverzeichnis

1 Ein CIFAR-10-Bild in vier Repräsentationen 7
2 Korrelation zwischen den Farbkanälen 8
3 Nichtlinearität der CIELAB-Transformation 10
4 Versuchsaufbau im Überblick 17
5 Konvergenzverhalten der vier Bedingungen 28
6 Aufschlüsselung des Farbraumeffekts nach Störungsart 32
7 Trade-off in Abhängigkeit von der Kanalskalierung 34
8 Farbräume unter beiden Normalisierungsschichten 37
9 Genutzte Farbrichtungen in conv1 39
10 HSV-Rückstand über Architekturen und Normalisierungsschichten . . 43

Tabellenverzeichnis

1 Aufbau von ResNet-18 11
2 Normalisierungskonstanten 20
3 Trainingskonfiguration 21
4 Störungsarten in CIFAR-10-C 23
5 Genauigkeit auf unveränderten Testdaten 27
6 Robustheit auf CIFAR-10-C 28
7 Aufschlüsselung nach Störungsart 31
8 Trade-off unter vereinheitlichter Kanalskalierung 33
9 Abstand zwischen den Farbräumen unter beiden Normalisierungsschichten 36
10 Genutzte Farbrichtungen in conv1 39
11 Ergebnisse auf VGG-11 42
12 HSV-Rückstand über Architekturen 43
13 Genauigkeit je Störungsart (BatchNorm) 57
14 Genauigkeit je Störungsart (GroupNorm) 58
15 Zusätzliche Störungsarten 58
16 Genauigkeit je Störungsart (VGG-11) 59
17 Genauigkeit je Störungsart (VGG-11 ohne Normalisierung) 59
18 Genauigkeit je Störungsart (vereinheitlichte Skalierung) 60
19 Verwendete Softwareversionen 62

Abkürzungsverzeichnis

BN	Batch-Normalisierung (Batch Normalization)
CIELAB	Farbraum der CIE mit Helligkeit L* und Farbkanälen a*, b*
CNN	Convolutional Neural Network
GN	Gruppen-Normalisierung (Group Normalization)
HSV	Farbraum: Farbton (Hue), Sättigung (Saturation), Hellwert (Value)
mCA	mittlere Korruptions-Genauigkeit (mean Corruption Accuracy)
$\text{mCA}_{\text{Helligkeit}}$	mCA über die drei helligkeitsbezogenen Störungsarten
mCA_{Rest}	mCA über die zwölf übrigen Störungsarten
RGB	Farbraum: Rot, Grün, Blau

Notation

Farbräume

R, G, B	Rot-, Grün- und Blaukanal, jeweils auf $[0, 1]$ skaliert
H, S, V	Farbton, Sättigung und Hellwert des HSV-Farbraums
L^*	Helligkeitskanal des CIELAB-Farbraums
a^*, b^*	Chromakanäle des CIELAB-Farbraums (grün–rot, blau–gelb)
X, Y, Z	Koordinaten des geräteunabhängigen XYZ-Farbraums
X_n, Y_n, Z_n	Koordinaten des Referenzweißpunktes
$f(t)$	Nichtlinearität der CIELAB-Transformation
δ	Schwellenwert der CIELAB-Nichtlinearität, $\delta = 6/29$

Netzarchitektur und Normalisierung

$w_{c,u,v}$	Gewicht eines Faltungsfilters für Kanal c an Kernposition (u, v)
N, C, H, W	Achsen eines Aktivierungstensors: Batch, Kanal, Höhe, Breite
G	Anzahl der Gruppen bei der Gruppen-Normalisierung
μ, σ^2	Mittelwert und Varianz einer Normalisierungsschicht
γ, β	gelernte Skalierungs- und Verschiebungsparameter
ε	numerische Konstante zur Vermeidung einer Division durch null
$\mathcal{F}$, $\mathcal{H}$	Residualabbildung und Zielabbildung eines Residualblocks

Auswertung

$\mathcal{C}$	Menge der ausgewerteten Störungsarten
$\mathrm{acc}(c, s)$	Genauigkeit bei Störungsart c und Schweregrad s
mCA	mittlere Korruptions-Genauigkeit über $\mathcal{C}$ und alle Schweregrade
M	zweites Moment der Kanalvektoren der ersten Faltungsschicht
$\lambda_1 \geq \lambda_2 \geq \lambda_3$	auf Summe eins normierte Eigenwerte von M

Statistik

s	Zufallsstartwert eines Trainingslaufs
δ_s	paarweise Differenz zweier Bedingungen beim Startwert s
$\bar{\delta}$, s_δ	Mittelwert und Standardabweichung der paarweisen Differenzen
d	Effektstärke nach Cohen für gepaarte Daten
p	p-Wert eines statistischen Tests
α	Signifikanzniveau
m	Anzahl der Tests bei der Korrektur für multiples Testen

1 Einleitung

1.1 Motivation

Nahezu alle Verfahren des maschinellen Sehens verarbeiten Bilder im RGB-Farbraum. Diese Wahl ist selten das Ergebnis einer Abwägung, sondern historisch bedingt: Kameras nehmen in RGB auf, Bildschirme geben in RGB wieder, und die gängigen Bibliotheken liefern Bilder in diesem Format. Das Eingabeformat eines neuronalen Netzes wird damit von der Aufnahmetechnik bestimmt und nicht von der Aufgabe, die das Netz lösen soll.

Aus statistischer Sicht ist RGB für diese Aufgabe allerdings ungünstig angeordnet. Die drei Kanäle sind stark korreliert, weil die Helligkeit eines Bildpunktes in alle drei zugleich eingeht: Wird eine Szene heller beleuchtet, so steigen R, G und B gemeinsam. Auf CIFAR-10 trägt eine einzige Richtung im Farbraum 91,1 Prozent der Gesamtvarianz der Pixelfarben, während auf die beiden übrigen zusammen weniger als neun Prozent entfallen. Ein Netz, dem eine derart angeordnete Eingabe präsentiert wird, muss die Trennung von Helligkeit und Farbe selbst herstellen, bevor es eine Beleuchtungsänderung von einer Farbänderung unterscheiden kann.

Diese Eigenschaft ist mehr als eine Randnotiz, denn sie berührt eine praktische Schwäche neuronaler Netze. Ein Modell, das auf unveränderten Testdaten hohe Genauigkeit erreicht, kann deutlich schlechter abschneiden, sobald sich die Aufnahmebedingungen ändern – bei anderer Beleuchtung, mit anderen Geräten oder unter widrigen Umständen. Besonders deutlich tritt das in Anwendungsfeldern hervor, in denen die Daten aus mehreren Quellen stammen. In der medizinischen Bildgebung etwa unterscheiden sich Gewebeschnitte verschiedener Kliniken systematisch in ihrer Färbung, ohne dass sich der diagnostisch relevante Inhalt änderte; ein Modell, das die Färbung mitlernt, verliert beim Wechsel des Hauses an Zuverlässigkeit.

Vor diesem Hintergrund liegt ein naheliegender Gedanke nahe: Wenn ein Farbraum die Helligkeit von vornherein als eigenen Kanal führt, muss das Netz diese Trennung nicht mehr selbst erlernen, und eine Beleuchtungsänderung betrifft nur einen Kanal statt drei. Farbräume mit dieser Eigenschaft sind seit Jahrzehnten bekannt und standardisiert. CIELAB etwa trennt die Helligkeit L^* konstruktiv von den beiden Farbkanälen a^* und b^*. Die Transformation ist umkehrbar, verlustfrei und mit wenigen Zeilen Code umzusetzen – ein bemerkenswert einfacher Eingriff, sollte er tatsächlich die Robustheit verbessern.

Ob er das tut, ist bislang nicht abschließend geklärt. Die vorliegenden empirischen Vergleiche berichten kleine und uneinheitliche Unterschiede zwischen Farbräumen (Kapitel 2), und sie messen die Robustheit als Mittelwert über sehr verschiedenartige Bildstörungen hinweg. Ob ein Vorteil gerade dort entsteht, wo der vermutete Mechanismus ihn erzeugen sollte – bei Störungen der Helligkeit –, wurde bisher nicht untersucht.

1.2 Forschungsfrage und Hypothese

Die vorliegende Arbeit geht der folgenden Frage nach:

> *Verändert der Eingabefarbraum, wie gut ein Convolutional Neural Network lernt und wie robust es gegenüber Bildstörungen bleibt?*

Untersucht werden vier Bedingungen: RGB als Referenz, die beiden dekorrelierten Farbräume CIELAB und HSV sowie eine Graustufendarstellung, die als Kontrollbedingung ohne chromatische Information dient. Alle vier werden unter identischen Bedingungen trainiert und ausgewertet; einzig das Eingabeformat unterscheidet sich.

Der Untersuchung liegt die folgende Arbeitshypothese zugrunde:

> *Ein Netz, das mit CIELAB-Eingaben trainiert wird, ist robuster gegenüber helligkeitsbezogenen Bildstörungen als ein mit RGB trainiertes Netz, weil die Helligkeit dort als eigener Kanal geführt wird.*

Diese Hypothese ist gerichtet und mechanistisch begründet, was sie prüfbar macht: Sie sagt einen Vorteil nicht allgemein vorher, sondern für eine bestimmte, im Voraus festgelegte Untergruppe von Störungen. Sollte sich ein Vorteil in gleicher Größe auch außerhalb dieser Untergruppe zeigen, wäre er nicht auf den vermuteten Mechanismus zurückzuführen.

Die Untersuchung erfolgt an einem ResNet-18 auf CIFAR-10 sowie dem zugehörigen Robustheits-Benchmark CIFAR-10-C. Diese Wahl erlaubt es, jedes Experiment über mehrere Zufallsstartwerte zu wiederholen und die Ergebnisse statistisch abzusichern – eine Voraussetzung dafür, Unterschiede in der erwarteten Größenordnung von wenigen Prozentpunkten überhaupt beurteilen zu können. Sie begrenzt zugleich die Reichweite der Aussagen, worauf Abschnitt 7.3 eingeht.

1.3 Aufbau der Arbeit

Kapitel 2 ordnet die Arbeit in den Forschungsstand ein und benennt die Lücke, an der sie ansetzt. Kapitel 3 führt die daraus abgeleiteten Beiträge aus. Kapitel 4 stellt die benötigten Grundlagen bereit: die betrachteten Farbräume und ihre mathematischen Eigenschaften, den Aufbau eines ResNet-18, die beiden verwendeten Normalisierungsschichten sowie die statistischen Verfahren.

Kapitel 5 beschreibt die Datenpipeline, das Training und das Evaluationsprotokoll. Kapitel 6 berichtet die Ergebnisse, zunächst auf ungestörten Daten, anschließend auf den Störungen von CIFAR-10-C, aufgeschlüsselt nach Störungsart, und schließlich anhand einer Analyse der gelernten Eingangsfilter. Kapitel 7 deutet diese Befunde, ordnet sie in die Literatur ein und benennt die Grenzen der Arbeit. Kapitel 8 fasst zusammen und benennt Anschlussfragen.

2 Verwandte Arbeiten

Die Frage, ob der Eingabefarbraum die Leistung eines neuronalen Netzes beeinflusst, ist nicht neu. Dieses Kapitel ordnet die vorliegende Arbeit in den bestehenden Forschungsstand ein und benennt die Lücke, an der sie ansetzt.

2.1 Farbräume in der Bildklassifikation

Die umfassendste empirische Untersuchung zu dieser Frage stammt von Gowda and Yuan (2019). Die Autoren trainieren neuronale Netze auf CIFAR-10, CIFAR-100, SVHN und ImageNet in sieben verschiedenen Farbräumen – darunter RGB, CIELAB, HSV und YCbCr – und vergleichen die erreichte Genauigkeit.

Ihr Ergebnis ist in zweifacher Hinsicht aufschlussreich. Einzeln verwendet unterscheiden sich die Farbräume nur um ein bis zwei Prozentpunkte, ohne dass sich eine klare Rangfolge über die Datensätze hinweg abzeichnete. Ein nennenswerter Gewinn entsteht erst, wenn mehrere Farbräume gleichzeitig verarbeitet und ihre Ausgaben zusammengeführt werden. Der namensgebende Beitrag der Arbeit betrifft somit die Fusion, nicht die Wahl eines einzelnen Farbraums – eine Unterscheidung, die beim Zitieren dieser Quelle zu beachten ist.

Einen anderen Weg verfolgt Karargyris (2015): Statt einen festen Farbraum vorzugeben, lässt er das Netz die Farbraumtransformation selbst erlernen. Auch hier bleiben die Gewinne moderat, was ebenfalls dagegen spricht, im Eingabeformat einen wesentlichen Hebel zu vermuten.

Für größere Datensätze weist der Befund eher in die entgegengesetzte Richtung. In einer systematischen Untersuchung von Architektur- und Trainingsentscheidungen auf ImageNet erweist sich RGB als der im Mittel geeignetste Eingaberaum (Mishkin et al., 2017). Ein Vorteil dekorrelierter Farbräume, sofern er auf kleinen Datensätzen auftritt, überträgt sich demnach nicht ohne Weiteres auf größere Maßstäbe.

Insgesamt zeigt die Literatur keinen belastbaren Vorteil eines einzelnen dekorrelierten Farbraums. Die berichteten Unterschiede sind klein, nicht systematisch und hängen vom Datensatz ab.

2.2 Robustheit, Normalisierung und gelernte Farbkodierung

Die Bewertung der Robustheit gegenüber Bildstörungen wurde durch Hendrycks and Dietterich (2019) standardisiert, die mit ImageNet-C und CIFAR-10-C Benchmarks aus systematisch gestörten Testdaten bereitstellen. Diese Arbeit variiert allerdings die Störungen und nimmt RGB als Eingabeformat als gegeben hin; der Farbraum ist dort kein Untersuchungsgegenstand.

Näher an der Fragestellung liegt Chiu et al. (2022). Die Autoren zeigen, dass die Empfindlichkeit neuronaler Netze gegenüber Farbunterscheidbarkeit weder durch

größere Modelle noch durch veränderte Architekturen oder gängige Augmentierungsverfahren nennenswert abnimmt. Untersucht wird dort jedoch die Farbabhängigkeit als solche, nicht der Vergleich verschiedener Eingabeformate.

Eine zweite Gruppe von Arbeiten befasst sich mit der Frage, wodurch die Robustheit eines Netzes überhaupt bestimmt wird. Geirhos et al. (2019) zeigen, dass auf ImageNet trainierte Netze vorrangig anhand von Textur und nicht anhand von Form klassifizieren, und dass ein durch das Training erzwungener Formbias die Robustheit gegenüber einem breiten Spektrum von Störungen verbessert. Kamann et al. (2020) erreichen einen vergleichbaren Effekt, indem sie die Verlässlichkeit der Farbinformation im Training gezielt zerstören. Beide Ansätze greifen in das Trainingsziel ein und verändern damit, worauf ein Netz sich stützt.

Eine dritte Gruppe betrifft die Verarbeitung von Farbe innerhalb des Netzes. Ioffe and Szegedy (2015) motivieren die Batch-Normalisierung ausdrücklich mit dem Gedanken eines Whitenings der Eingaben, führen aber aus Aufwandsgründen eine kanalweise Standardisierung ohne Dekorrelation ein. Huang et al. (2018) greifen diesen Punkt auf und schlagen ein tatsächlich dekorrelierendes Verfahren vor. Wu and He (2018) entwickeln mit der Gruppen-Normalisierung eine Alternative, die ohne Batch-Statistiken auskommt. Harris et al. (2020) schließlich zeigen, dass trainierte Netze in ihren frühen Schichten spontan eine gegenfarbartige Kodierung entwickeln, wie sie CIELAB von Konstruktion wegen bereitstellt.

2.3 Abgrenzung dieser Arbeit

Aus diesem Forschungsstand ergeben sich drei offene Punkte.

Erstens berichten die vorliegenden Farbraumvergleiche die Robustheit ausschließlich als aggregierte Kennzahl. In der gesichteten Literatur findet sich keine Arbeit, die den Vergleich zwischen Farbräumen nach Störungsarten aufschlüsselt. Ein Verfahren, das auf verschiedene Störungsarten gegenläufig wirkt, bliebe in einem solchen Mittelwert unsichtbar.

Zweitens wird die Untergruppe der helligkeitsbezogenen Störungen nirgends gesondert betrachtet – obwohl gerade sie den Mechanismus prüft, der einen Vorteil von CIELAB begründen soll. Die theoretische Begründung dekorrelierter Farbräume verweist auf die Trennung von Helligkeit und Farbe; die empirische Prüfung erfolgt dennoch über Kennzahlen, die diese Trennung nicht abbilden.

Drittens wird die zugrunde liegende Annahme – ein Netz profitiere von einer bereits dekorrelierten Eingabe – ausschließlich indirekt über die Klassifikationsleistung geprüft. Harris et al. (2020) zeigen zwar, dass Netze eine gegenfarbartige Kodierung erlernen, untersuchen jedoch nicht, ob diese Fähigkeit vom Eingabeformat abhängt. Ob eine dekorrelierte Eingabe dem Netz überhaupt Arbeit abnimmt, bleibt damit ungeklärt.

An diesen drei Punkten setzt die vorliegende Arbeit an; Kapitel 3 führt die daraus abgeleiteten Beiträge aus.

3 Beitrag dieser Arbeit

Aus der in Abschnitt 2.3 festgestellten Lücke ergeben sich die vier Beiträge der vorliegenden Arbeit.

3.1 Aufschlüsselung der Robustheit nach Störungsart

Bestehende Vergleiche zwischen Farbräumen berichten die Robustheit als aggregierte Kennzahl über sämtliche Störungsarten hinweg. Diese Arbeit schlüsselt den Vergleich stattdessen nach einzelnen Störungsarten auf und wertet jede Kombination aus Farbraum und Störung getrennt aus.

Die Aufschlüsselung ist keine bloße Verfeinerung der Darstellung. Sie legt offen, dass ein Mittelwert über Störungsarten implizit voraussetzt, ein Verfahren wirke auf alle Störungen in dieselbe Richtung – eine Annahme, die sich hier als unzutreffend erweist.

3.2 Nachweis eines gegenläufigen Effekts

Auf dieser Grundlage weist die Arbeit nach, dass CIELAB gegenüber RGB auf Rauschstörungen deutlich verliert und auf Unschärfestörungen ebenso deutlich gewinnt. Beide Effekte sind über alle Zufallsstartwerte hinweg konsistent und überstehen die Korrektur für multiples Testen; im Mittelwert heben sie einander nahezu vollständig auf.

Für diesen Zusammenhang wird eine Erklärung vorgeschlagen und in einem eigenen Experiment geprüft. Die Vermutung lautet, dass nicht die Farbraumtransformation selbst den Trade-off erzeugt, sondern die kanalweise Standardisierung, die auf sie folgt: Sie verstärkt die Chromakanäle von CIELAB um ein Vielfaches dessen, was ein RGB-Kanal erfährt, und wirkt dadurch auf hochfrequente und niederfrequente Störungen gegenläufig. Geprüft wird dies durch eine Normalisierungsvariante, die ausschließlich die Skalierung verändert und die Zentrierung unberührt lässt – eine Manipulation an einer einzigen Größe.

Damit rückt ein Verarbeitungsschritt in den Blick, der in Farbraumvergleichen üblicherweise als neutraler Zwischenschritt behandelt wird.

3.3 Direkte Messung der gelernten Farbverarbeitung

Drittens wird die Frage, ob ein Netz eine dekorrelierte Eingabe überhaupt benötigt, nicht nur über die Klassifikationsleistung beantwortet, sondern unmittelbar an den trainierten Gewichten gemessen. Eine Eigenwertzerlegung der Kanalgewichte der ersten Faltungsschicht zeigt, in welchem Ausmaß ein Netz die drei Farbrichtungen tatsächlich nutzt und wie die gelernten Achsen ausgerichtet sind.

Ergänzt wird dies durch einen Vergleich von Normalisierungsschichten unter sonst identischen Bedingungen, der die Wirkung des Eingabeformats gegen die Wirkung einer Architekturentscheidung stellt.

3.4 Wiederholung anhand vorab festgelegter Vorhersagen

Viertens werden die zentralen Befunde auf einer zweiten Architektur wiederholt. Vor dem Start der zugehörigen Rechenläufe wurden vier Vorhersagen schriftlich festgehalten, jede mit ihrer mechanistischen Begründung und so formuliert, dass sie scheitern kann.

Dieses Vorgehen unterscheidet sich von einer bloßen Wiederholung. Eine Vorhersage, die vor der Betrachtung der Daten feststand, lässt sich weder nachträglich anpassen noch selektiv berichten; ihr Scheitern ist ebenso aussagekräftig wie ihr Zutreffen. Von den vier Vorhersagen bestätigen sich drei, während die vierte widerlegt wird und die in dieser Arbeit vorgeschlagene Deutung der Normalisierungsschicht korrigiert.

4 Grundlagen

4.1 Farbräume und Kanalkorrelation

Ein Farbraum ist eine Vorschrift, die eine Farbe durch Zahlen beschreibt. Dieselbe Farbe lässt sich auf verschiedene Weisen kodieren, ohne dass sich der Informationsgehalt ändert. Was sich ändert, ist die Anordnung dieser Information auf die Kanäle – und genau darauf zielt die Fragestellung dieser Arbeit. Abbildung 1 zeigt dieselbe Aufnahme in den vier Repräsentationen, die in dieser Arbeit verglichen werden, jeweils zerlegt in ihre drei Kanäle.

4.1.1 RGB und die Korrelation seiner Kanäle

Der RGB-Farbraum beschreibt eine Farbe als additive Mischung dreier Grundfarben. Er entspricht der Funktionsweise von Kameras und Bildschirmen und ist deshalb das übliche Eingabeformat für neuronale Netze.

Für die vorliegende Fragestellung ist eine statistische Eigenschaft dieses Formats maßgeblich: Die drei Kanäle sind stark korreliert. Der Grund ist anschaulich. Die Helligkeit eines Bildpunktes geht in alle drei Kanäle gleichermaßen ein; wird eine Szene heller beleuchtet, so steigen R, G und B gemeinsam. Ein erheblicher Teil der Variation zwischen Bildpunkten ist daher keine Farbvariation, sondern eine Helligkeitsvariation, die sich über alle Kanäle erstreckt.

Wie ausgeprägt dieser Effekt ist, lässt sich an CIFAR-10 beziffern. Eine Eigenwertzerlegung der Pixelfarben ergibt, dass eine einzige Richtung 91,1 Prozent der Gesamtvarianz trägt, während auf die beiden übrigen zusammen weniger als neun Pro-

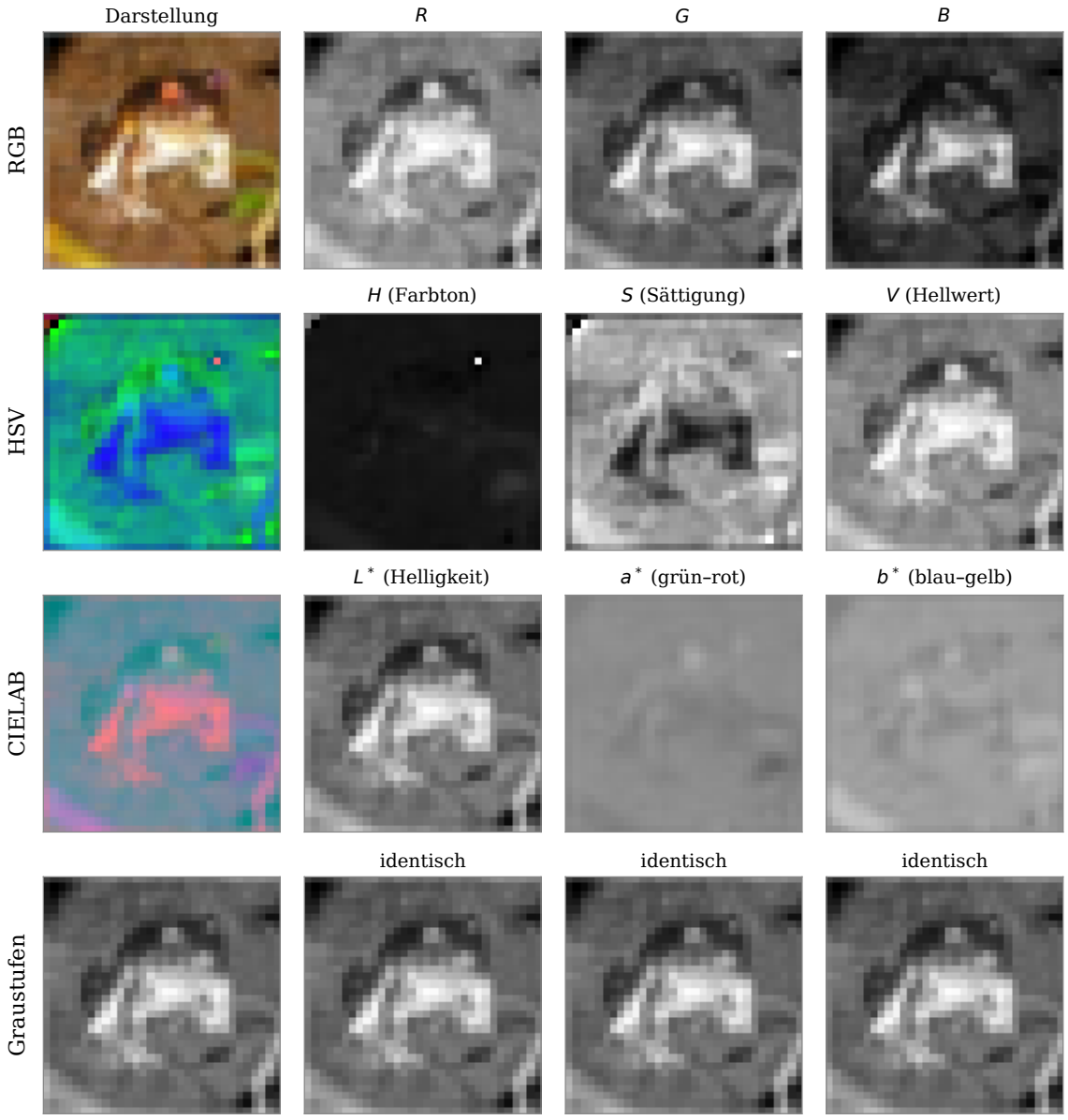


Abbildung 1: Dasselbe CIFAR-10-Bild in den vier verglichenen Repräsentationen. Die linke Spalte zeigt die Darstellung als Ganzes, die drei rechten Spalten die einzelnen Kanäle, einheitlich auf das Intervall $[0, 1]$ abgebildet. Die drei RGB-Kanäle enthalten sichtbar dieselbe Struktur; in HSV und CIELAB konzentriert sich diese Struktur auf den Helligkeitskanal, während die Farbkanäle deutlich weniger Kontrast aufweisen. Die geringe Streuung von a^* und b^* ist keine Eigenschaft dieses Bildes, sondern des Farbraums (Tabelle 2); sie wird in Abschnitt 6.3 als Ursache eines der zentralen Ergebnisse wieder aufgegriffen. Der Farbtonkanal H erscheint hier nahezu schwarz, weil die Farbtöne dieses Bildes fast alle im unteren Bereich der zyklischen Skala liegen; das ist kein Darstellungsfehler. Die Graustufenvariante wiederholt denselben Kanal dreifach und dient als Negativkontrolle.

zent entfallen (Abschnitt 6.7). Diese dominante Richtung ist die Helligkeitsrichtung $(1, 1, 1)$.

Dieselbe Eigenschaft zeigt sich unmittelbar an den paarweisen Korrelationen der Kanäle (Abbildung 2). In RGB liegen sie zwischen 0,78 und 0,91; in CIELAB und HSV sind sie dem Betrag nach kleiner als 0,3. Die Kanäle der beiden zuletzt genannten Farbräume tragen also weitgehend voneinander unabhängige Information, während sich die drei RGB-Kanäle in weiten Teilen wiederholen.

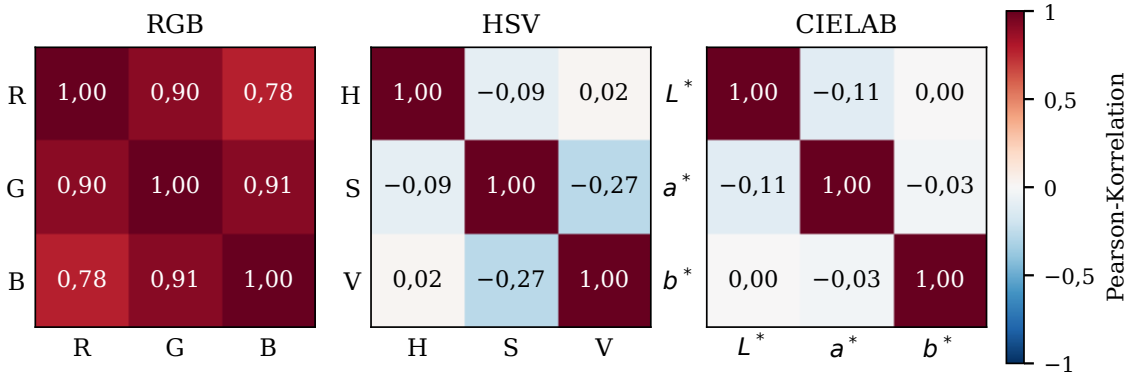


Abbildung 2: Korrelation zwischen den Farbkanälen, berechnet über die Bildpunkte von 1000 Trainingsbildern aus CIFAR-10. Die drei RGB-Kanäle sind stark positiv korreliert, da die Helligkeit eines Bildpunktes in alle drei zugleich eingeht. In CIELAB und HSV ist diese Kopplung konstruktionsbedingt weitgehend aufgehoben.

Die Überlegung, die dieser Arbeit zugrunde liegt, knüpft hier an: Ein Netz, dem eine derart anisotrope Eingabe präsentiert wird, verfügt über keine Möglichkeit, eine Helligkeitsänderung von einer Farbänderung zu unterscheiden, ohne diese Trennung selbst zu erlernen.

4.1.2 HSV

Der HSV-Farbraum beschreibt eine Farbe durch Farbton (H), Sättigung (S) und Hellwert (V). Mit $M = \max(R, G, B)$ und $m = \min(R, G, B)$ gilt

$$V = M, \qquad S = \begin{cases} (M - m)/M & \text{für } M > 0 \\ 0 & \text{sonst,} \end{cases} \tag{1}$$

während der Farbton H stückweise danach bestimmt wird, welcher der drei Kanäle das Maximum stellt.

Zwei Eigenschaften dieser Definition werden in späteren Kapiteln benötigt. Erstens ist V ein Maximum und damit keine Mittelung: Ein Maximum reagiert auf eine Glättung des Bildes anders als ein Mittelwert. Zweitens ist H eine zyklische Größe – der Farbton 0 und der Farbton 1 bezeichnen dieselbe Farbe. Für ein Netz, das seine Eingabe als reelle Zahl verarbeitet, ist diese Zyklizität nicht erkennbar.

4.1.3 CIELAB

Der CIELAB-Farbraum wurde entwickelt, um wahrnehmungsbezogene Abstände abzubilden: Gleiche numerische Abstände sollen möglichst gleichen wahrgenommenen Farbunterschieden entsprechen. Er trennt die Helligkeit L^* konstruktiv von den beiden Farbkanälen a^* (grün–rot) und b^* (blau–gelb).

Die Umrechnung erfolgt über den geräteunabhängigen XYZ-Raum:

$$L^* = 116\, f\left(\tfrac{Y}{Y_n}\right) - 16, \tag{2}$$

$$a^* = 500\left[f\left(\tfrac{X}{X_n}\right) - f\left(\tfrac{Y}{Y_n}\right)\right], \qquad b^* = 200\left[f\left(\tfrac{Y}{Y_n}\right) - f\left(\tfrac{Z}{Z_n}\right)\right], \tag{3}$$

wobei X_n, Y_n, Z_n die Koordinaten des Referenzweißpunktes bezeichnen und

$$f(t) = \begin{cases} t^{1/3} & \text{für } t > \delta^3 \\ \dfrac{t}{3\delta^2} + \dfrac{4}{29} & \text{sonst,} \end{cases} \qquad \delta = \tfrac{6}{29} \;\; (\delta^3 \approx 0{,}0089). \tag{4}$$

Für diese Arbeit ist die Nichtlinearität in Gleichung 4 entscheidend. Die Kubikwurzel besitzt bei kleinen Argumenten eine sehr steile Ableitung, $\frac{\mathrm{d}}{\mathrm{d}t}t^{1/3} = \frac{1}{3}t^{-2/3}$: Änderungen in dunklen Bildbereichen werden dadurch stärker abgebildet als gleich große Änderungen in hellen. Abbildung 3 zeigt beides – die Kennlinie und den zugehörigen Verstärkungsfaktor. Bei einer relativen Luminanz von 0,02 beträgt dieser 4,52, bei 0,5 nur noch 0,53; eine Änderung im dunklen Bildbereich erscheint in CIELAB also um etwa das Achtfache dessen verstärkt, was dieselbe Änderung im mittleren Helligkeitsbereich bewirkt. Die lineare Fortsetzung unterhalb von δ^3 begrenzt diese Verstärkung, greift jedoch erst bei sehr kleinen Werten.

4.1.4 Graustufen

Die Graustufendarstellung fasst die drei Kanäle zu einem luminanzgewichteten Mittel zusammen und verwirft damit die chromatische Information vollständig. Sie ist kein alternativer Farbraum im eigentlichen Sinne, sondern eine Reduktion. In dieser Arbeit dient sie als Negativkontrolle: Eine Bedingung, bei der Information nachweislich fehlt, muss messbar schlechter abschneiden, wenn das Versuchsdesign empfindlich genug ist.

4.1.5 Was Dekorrelation bedeutet – und was nicht

Von den drei Farbräumen führt allein CIELAB die Helligkeit als eigenen Kanal; in RGB ist sie über alle drei Kanäle verteilt, in HSV liegt sie mit V zwar in einem Kanal, dieser ist jedoch als Maximum definiert und nicht als gewichtetes Mittel.

Zu betonen ist, dass eine solche Transformation den Informationsgehalt nicht verändert. Sie ist umkehrbar und fügt nichts hinzu. Was sie verändert, ist die Verteilung dieser

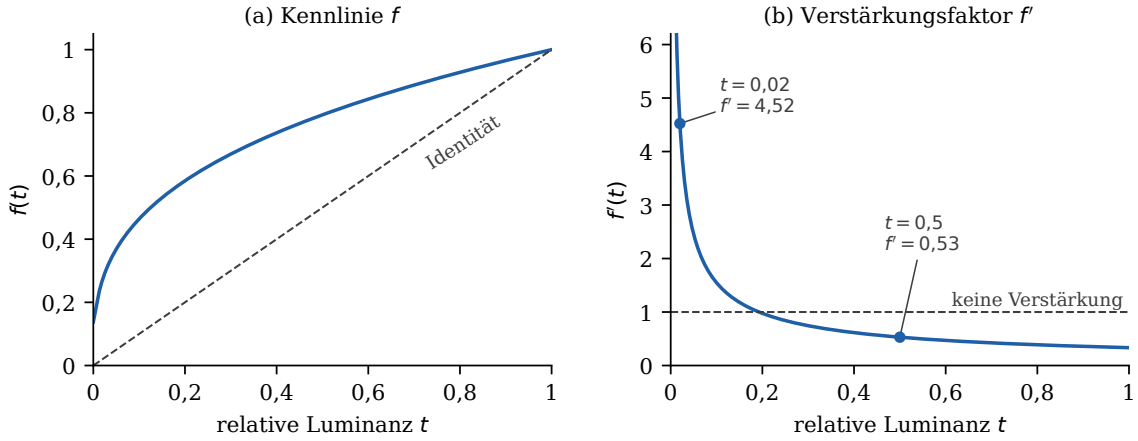


Abbildung 3: Nichtlinearität der CIELAB-Transformation. **(a)** Die Kennlinie f aus Gleichung 4 im Vergleich zur Identität. **(b)** Ihre Ableitung f', die angibt, um welchen Faktor eine kleine Änderung der relativen Luminanz in CIELAB verstärkt wird. Oberhalb der gestrichelten Linie wird verstärkt, unterhalb gedämpft. Additive Störungen, die im RGB-Raum über das ganze Bild gleich groß sind, treffen dunkle Bildbereiche in CIELAB damit deutlich stärker – ein Zusammenhang, auf den Abschnitt 6.3 zurückgreift.

Information auf die Kanäle – und damit, welche Eingangsgrößen ein Netz miteinander verrechnen muss, um eine bestimmte Eigenschaft zu erkennen. Ob diese Umverteilung dem Lernvorgang zugutekommt, ist die empirische Frage, der diese Arbeit nachgeht.

4.2 Convolutional Neural Networks und ResNet-18

4.2.1 Die Faltung als gewichtete Kanalmischung

Ein Convolutional Neural Network verarbeitet Bilder, indem es kleine Gewichtsmatrizen – die Filter – über das Bild schiebt und an jeder Position eine gewichtete Summe bildet (LeCun et al., 2015; Goodfellow et al., 2016). Ein Filter der ersten Schicht mit einem Kern der Größe 3×3 und drei Eingangskanälen besitzt $3 \cdot 3 \cdot 3 = 27$ Gewichte; sein Ausgabewert an einer Bildposition ist

$$y = \sum_{c=1}^{3} \sum_{u,v} w_{c,u,v}\, x_{c,\, i+u,\, j+v}, \tag{5}$$

wobei c über die Eingangskanäle und (u, v) über die Kernpositionen läuft.

Für diese Arbeit ist die Summe über c der wesentliche Punkt. Sie besagt, dass die erste Schicht die Eingangskanäle *linear mischt*, und die Gewichte $w_{c,u,v}$ legen fest, in welchem Verhältnis. Ein Filter, dessen drei Kanalgewichte annähernd gleich sind, bildet ein Helligkeitsmittel; ein Filter mit entgegengesetzten Vorzeichen bildet eine

Farbdifferenz. Welche Kanalmischungen ein Netz erlernt, ist damit unmittelbar aus seinen Gewichten ablesbar – die Grundlage der Analyse in Abschnitt 6.7.

In tieferen Schichten geht diese unmittelbare Deutbarkeit verloren, da die Eingangskanäle dort keine Farbkanäle mehr sind, sondern die Ausgaben vorangegangener Filter.

4.2.2 ResNet-18

Mit zunehmender Tiefe wird das Training vorwärtsgerichteter Netze zunehmend schwierig. He et al. (2016) schlagen deshalb Residualverbindungen vor: Ein Block berechnet nicht die Abbildung $\mathcal{H}(x)$ selbst, sondern die Abweichung $\mathcal{F}(x) = \mathcal{H}(x) - x$ und addiert die Eingabe anschließend hinzu,

$$y = \mathcal{F}(x) + x. \tag{6}$$

Die zusätzliche Verbindung leitet das Gradientensignal unverändert an frühere Schichten weiter und macht damit deutlich tiefere Netze trainierbar.

ResNet-18 besteht aus einer Eingangsfaltung, vier Stufen mit je zwei Residualblöcken und einem abschließenden Klassifikationskopf. Zwischen den Stufen verdoppelt sich die Kanalzahl bei gleichzeitiger Halbierung der räumlichen Auflösung.

Tabelle 1: Aufbau von ResNet-18 in der für diese Arbeit verwendeten, an 32×32 angepassten Fassung (vgl. Abschnitt 5.3).

Stufe	Aufbau	Ausgabe
Eingang	Faltung 3×3, Schrittweite 1	$64 \times 32 \times 32$
Stufe 1	2× Residualblock	$64 \times 32 \times 32$
Stufe 2	2× Residualblock	$128 \times 16 \times 16$
Stufe 3	2× Residualblock	$256 \times 8 \times 8$
Stufe 4	2× Residualblock	$512 \times 4 \times 4$
Kopf	globales Mittel, lineare Schicht	10

Die Eingangsfaltung besitzt damit die Gewichtsform $[64, 3, 3, 3]$: 64 Filter, je drei Eingangskanäle und neun Kernpositionen. Die zweite Achse dieser Gewichte entspricht unmittelbar den drei Kanälen des gewählten Farbraums – und ist damit die einzige Stelle im Netz, an der das Eingabeformat explizit sichtbar ist.

Jeder Residualblock enthält zwei Faltungen, denen jeweils eine Normalisierungsschicht und eine ReLU-Aktivierung folgen. Welche Normalisierungsschicht dabei verwendet wird, ist Gegenstand des folgenden Abschnitts.

4.3 Normalisierungsschichten: BatchNorm und GroupNorm

Normalisierungsschichten standardisieren die Aktivierungen im Inneren eines Netzes. Sie unterscheiden sich darin, über welche Achsen des Aktivierungstensors Mit-

telwert und Streuung gebildet werden – und gerade dieser Unterschied ist für die vorliegende Arbeit von Bedeutung.

Ein Aktivierungstensor besitzt vier Achsen: Beispiel im Batch (N), Kanal (C) sowie Höhe und Breite (H, W). Beide im Folgenden beschriebenen Verfahren normalisieren nach demselben Schema

$$\hat{x} = \frac{x - \mu}{\sqrt{\sigma^2 + \varepsilon}}, \qquad y = \gamma\,\hat{x} + \beta, \tag{7}$$

mit gelernten Parametern γ und β. Sie unterscheiden sich ausschließlich darin, über welche Achsen μ und σ^2 berechnet werden.

4.3.1 Batch-Normalisierung

Die Batch-Normalisierung (Ioffe and Szegedy, 2015) bildet Mittelwert und Streuung *je Kanal* über alle Beispiele des Batches und alle Bildpositionen, also über die Achsen N, H und W:

$$\mu_c = \frac{1}{NHW} \sum_{n,h,w} x_{n,c,h,w}. \tag{8}$$

Jeder Kanal wird damit auf denselben Wertebereich gebracht, und zwar anhand einer Statistik, die über viele Beispiele gemittelt ist und daher die Verteilung des gesamten Datensatzes annähert.

Zur Testzeit steht kein Batch zur Verfügung, dessen Statistik sinnvoll wäre. Stattdessen werden während des Trainings gleitende Mittelwerte von μ_c und σ_c^2 mitgeführt und zur Vorhersage als feste Größen verwendet. Das Verhalten eines Netzes mit Batch-Normalisierung unterscheidet sich somit zwischen Training und Anwendung.

Zwei Eigenschaften sind für die spätere Diskussion wesentlich. Erstens erfolgt die Normalisierung *kanalweise und datensatzweit*: Systematische Unterschiede im Wertebereich einzelner Kanäle werden dadurch eingeebnet. Zweitens sind die verwendeten Statistiken zur Testzeit *fest*; sie stammen aus den Trainingsdaten und passen sich einer veränderten Eingabeverteilung nicht an.

Bereits Ioffe and Szegedy (2015) motivieren das Verfahren mit dem Gedanken, dass Netze schneller lernen, wenn ihre Eingaben *weiß* sind, also mittelwertfrei, auf Einheitsvarianz gebracht und dekorreliert. Ein vollständiges Whitening ist jedoch rechenaufwendig, weshalb die Batch-Normalisierung jeden Kanal einzeln standardisiert. Sie leistet damit die ersten beiden Teile, nicht aber die Dekorrelation – Huang et al. (2018) weisen ausdrücklich darauf hin und schlagen deshalb ein tatsächlich dekorrelierendes Verfahren vor.

4.3.2 Gruppen-Normalisierung

Die Gruppen-Normalisierung (Wu and He, 2018) teilt die C Kanäle in G Gruppen und normalisiert *je Beispiel und Gruppe* über die Kanäle der Gruppe sowie über die Bildpositionen:

$$\mu_{n,g} = \frac{1}{(C/G)HW} \sum_{c \in \mathcal{G}_g} \sum_{h,w} x_{n,c,h,w}. \tag{9}$$

In dieser Arbeit wird $G = 32$ verwendet; in einer Schicht mit 64 Kanälen umfasst eine Gruppe folglich zwei Kanäle.

Der entscheidende Unterschied liegt in der Achse N: Sie geht in Gleichung 9 nicht ein. Die Normalisierung eines Bildes hängt damit weder von den übrigen Bildern des Batches noch von der Batchgröße ab, und es werden keine gleitenden Statistiken benötigt. Training und Anwendung verhalten sich identisch, und jedes Bild wird anhand seiner eigenen Statistik normalisiert – auch dann, wenn es gestört ist.

Umgekehrt entfällt damit die datensatzweite Vereinheitlichung der Kanäle: Eine Gruppe umfasst hier lediglich zwei Kanäle eines einzelnen Bildes, sodass systematische Unterschiede zwischen Kanälen nicht in gleichem Maße eingeebnet werden wie bei der Batch-Normalisierung.

4.3.3 Warum dieser Vergleich für die Fragestellung geeignet ist

Beide Schichten besitzen dieselbe Schnittstelle und dieselbe Zahl gelernter Parameter. Der Austausch der einen gegen die andere verändert genau eine Eigenschaft des Netzes – die Achsen, über die normalisiert wird – und lässt Tiefe, Kanalzahl, Filtergrößen und sämtliche Hyperparameter unberührt. Ein beobachteter Unterschied lässt sich deshalb der Normalisierung zuschreiben und nicht einem Architekturwechsel. Von dieser Eigenschaft macht das Experiment in Abschnitt 6.6 Gebrauch.

4.4 Robustheit gegenüber Bildstörungen

4.4.1 Warum Testgenauigkeit nicht genügt

Die übliche Bewertung eines Klassifikators setzt voraus, dass Trainings- und Testdaten derselben Verteilung entstammen. In der Anwendung ist diese Annahme regelmäßig verletzt: Aufnahmen entstehen bei anderer Beleuchtung, mit anderen Geräten oder unter widrigen Bedingungen. Ein Modell, das auf unveränderten Testdaten hohe Genauigkeit erreicht, kann unter solchen Verschiebungen deutlich schlechter abschneiden.

Man unterscheidet dabei verschiedene Formen der Verteilungsverschiebung. Diese Arbeit befasst sich mit der *Störungsrobustheit*: Die Testbilder stammen aus derselben Quelle wie die Trainingsbilder, werden jedoch durch eine definierte Operation verändert – Rauschen, Unschärfe, veränderte Helligkeit. Davon zu unterscheiden ist die *Domänenverallgemeinerung* im engeren Sinne, bei der Trainings- und Testdaten

aus unterschiedlichen Quellen stammen, etwa aus verschiedenen Kliniken mit je eigenen Färbeprotokollen. Die Störungsrobustheit ist die kontrolliertere und leichter messbare der beiden Formen; sie erlaubt es, eine einzelne Ursache gezielt zu variieren. Sie ist jedoch nicht mit der allgemeineren Form gleichzusetzen – ein Punkt, auf den Abschnitt 7.3 zurückkommt.

4.4.2 CIFAR-10-C

Hendrycks and Dietterich (2019) stellen mit CIFAR-10-C einen Benchmark bereit, der diese Bewertung standardisiert. Auf den Testdatensatz von CIFAR-10 werden Störungen angewandt, die typischen realen Beeinträchtigungen nachempfunden sind und sich vier Gruppen zuordnen lassen: Rauschen, Unschärfe, Wettereinflüsse und digitale Artefakte. Jede Störung liegt in fünf Schweregraden vor.

Zwei Eigenschaften dieses Aufbaus sind für die vorliegende Arbeit wesentlich. Erstens treten die Störungen ausschließlich zur Testzeit auf; ein Modell wird unverändert übernommen und lediglich auf veränderten Daten ausgewertet. Zweitens ist die Störung nach Art und Stärke bekannt, sodass sich die Leistung nicht nur insgesamt, sondern getrennt nach Störungsart beurteilen lässt. Von dieser zweiten Möglichkeit macht Abschnitt 6.3 Gebrauch.

4.4.3 Kennzahlen

Hendrycks and Dietterich (2019) berichten ihre Ergebnisse als *mean Corruption Error*, bei dem die Fehlerraten vor der Mittelung an denen eines Referenzmodells normiert werden. Diese Normierung dient dazu, unterschiedlich schwierige Störungen vergleichbar zu gewichten.

Die vorliegende Arbeit verwendet stattdessen die ungewichtete mittlere Genauigkeit über alle Störungen und Schweregrade, im Folgenden als mCA bezeichnet (Gleichung 12). Der Grund ist, dass hier ausschließlich Modelle desselben Aufbaus miteinander verglichen werden: Eine Normierung an einem Referenzmodell würde alle verglichenen Größen mit demselben Faktor versehen und die Differenzen nicht verändern. Für einen Vergleich mit absoluten Werten aus der Literatur ist der Unterschied der Kennzahlen allerdings zu beachten.

Da ein Mittelwert über Störungsarten voraussetzt, dass ein Verfahren auf allen Störungen in dieselbe Richtung wirkt – eine Annahme, die sich in dieser Arbeit als unzutreffend erweist –, wird die Genauigkeit zusätzlich für jede Störungsart einzeln ausgewertet. Das genaue Vorgehen beschreibt Abschnitt 5.4.

4.5 Statistische Methoden

Das Training eines neuronalen Netzes ist kein deterministischer Vorgang. Die Initialisierung der Gewichte, die Reihenfolge der Trainingsbeispiele und die zufälligen

Augmentierungen führen dazu, dass zwei Läufe derselben Konfiguration zu unterschiedlichen Modellen führen. Ein einzelnes Trainingsergebnis ist insofern eine Zufallsgröße, und ein Vergleich zweier Konfigurationen anhand je eines Laufes lässt nicht erkennen, ob ein beobachteter Unterschied auf die Konfiguration oder auf diesen Zufall zurückgeht.

Der Ausweg besteht darin, jede Konfiguration mehrfach mit verschiedenen Zufallsstartwerten zu trainieren. Die Streuung der Ergebnisse über die Startwerte hinweg beziffert dann den Anteil des Zufalls und liefert den Maßstab, an dem ein gemessener Unterschied zu beurteilen ist.

4.5.1 Gepaarte Vergleiche

Verwenden zwei Konfigurationen dieselben Startwerte, so lassen sich ihre Läufe einander zuordnen: Der Lauf mit Startwert s der einen Konfiguration und der Lauf mit Startwert s der anderen teilen dieselbe Initialisierungssequenz. Statt zweier unabhängiger Stichproben betrachtet man dann die Differenzen $\delta_s = a_s - b_s$ der zugeordneten Paare.

Der Vorteil dieses Vorgehens liegt darin, dass ein Teil der Zufallsvariation in beiden Läufen gleichermaßen auftritt und sich in der Differenz aufhebt. Der gepaarte t-Test prüft anschließend, ob der Mittelwert $\bar{\delta}$ mit der Annahme vereinbar ist, dass die zugrunde liegende mittlere Differenz null beträgt.

4.5.2 Der p-Wert und seine Grenzen

Der p-Wert eines solchen Tests gibt an, wie wahrscheinlich es wäre, unter der Annahme *keines* tatsächlichen Unterschieds eine mindestens so große Differenz zu beobachten wie die gemessene. Ein kleiner p-Wert besagt, dass das beobachtete Ergebnis unter dieser Annahme ungewöhnlich wäre.

Zwei verbreitete Fehldeutungen sind auszuschließen. Der p-Wert ist weder die Wahrscheinlichkeit, dass kein Unterschied besteht, noch ein Maß für die Größe eines Unterschieds. Insbesondere erlaubt ein großer p-Wert nicht den Schluss, dass kein Unterschied vorliegt – er besagt lediglich, dass die Daten mit dieser Annahme vereinbar sind. Bei kleinen Stichproben ist das häufig auch dann der Fall, wenn ein Unterschied besteht.

Aus diesem Grund werden in dieser Arbeit stets zwei weitere Größen mit angegeben. Die *Effektstärke* nach Cohen für gepaarte Daten,

$$d = \frac{\bar{\delta}}{s_\delta}, \tag{10}$$

setzt den mittleren Unterschied ins Verhältnis zu seiner eigenen Streuung und ist von der Stichprobengröße unabhängig. Die *Konsistenz* gibt an, bei wie vielen Startwerten die Differenz dasselbe Vorzeichen aufweist; ein Effekt, der bei allen Startwerten

in dieselbe Richtung weist, ist anders zu bewerten als einer, der das Vorzeichen wechselt, selbst bei gleichem p-Wert.

Da der t-Test eine annähernd normalverteilte Differenz voraussetzt und sich diese Annahme bei fünf Beobachtungen nicht prüfen lässt, wird ergänzend der verteilungsfreie Wilcoxon-Vorzeichen-Rang-Test berechnet.

4.5.3 Multiples Testen und die Holm-Korrektur

Wird eine Untersuchung nicht auf einen, sondern auf mehrere Vergleiche gestützt, so steigt die Wahrscheinlichkeit, allein durch Zufall mindestens einen p-Wert unterhalb der gewählten Schwelle zu finden. Bei m voneinander unabhängigen Tests und einer Schwelle von $\alpha = 0{,}05$ beträgt diese Wahrscheinlichkeit $1 - (1 - \alpha)^m$ – bei neun Tests bereits rund 37 Prozent.

Das Verfahren von Holm begegnet dem, indem es die m p-Werte aufsteigend ordnet und den k-kleinsten mit der verschärften Schwelle $\alpha/(m - k + 1)$ vergleicht; sobald ein Vergleich scheitert, gelten auch alle folgenden als nicht signifikant. Es ist damit strenger als ein einzelner Test, aber weniger konservativ als die Bonferroni-Korrektur, die sämtliche Schwellen gleichermaßen durch m teilt.

4.5.4 Vorab festgelegte und nachträgliche Analysen

Die Aussagekraft eines p-Wertes setzt voraus, dass die geprüfte Hypothese vor der Betrachtung der Daten feststand. Wird eine Auffälligkeit erst in den Daten entdeckt und anschließend an denselben Daten geprüft, so ist der resultierende p-Wert nicht mehr als Wahrscheinlichkeit im obigen Sinne zu lesen: Die Auswahl der Hypothese hat die Daten bereits verwendet.

Diese Arbeit kennzeichnet daher durchgängig, welche Befunde auf vorab festgelegten Fragestellungen beruhen und welche sich erst bei der Auswertung ergeben haben. Nachträgliche Befunde werden nicht verworfen – sie sind ein legitimer Bestandteil empirischer Arbeit –, aber als das ausgewiesen, was sie sind: Beobachtungen, die einer eigenständigen Prüfung bedürfen.

5 Methodik

5.1 Datenpipeline und Farbraumtransformationen

Alle Experimente dieser Arbeit basieren auf CIFAR-10 (Krizhevsky, 2009), einem Datensatz aus 60 000 Farbbildern der Größe 32×32 Pixel, aufgeteilt in 50 000 Trainings- und 10 000 Testbilder über zehn Objektklassen. Die geringe Auflösung erlaubt es, jedes Experiment über mehrere Zufallsstartwerte zu wiederholen – eine Voraussetzung für die statistische Absicherung, die in Abschnitt 5.5 beschrieben wird.

5.1.1 Aufbau der Pipeline

Der zentrale methodische Grundsatz dieser Arbeit lautet: *zwischen den vier Versuchsbedingungen darf sich ausschließlich der Eingabe-Farbraum unterscheiden.* Modellarchitektur, Optimierer, Augmentierung, Anzahl der Epochen und Evaluationsprotokoll bleiben identisch. Nur so lässt sich ein beobachteter Unterschied dem Farbraum zuschreiben und nicht einer anderen Änderung.

Die Verarbeitung eines Bildes durchläuft vier Schritte, die Abbildung 4 zusammen mit dem Evaluationsteil des Aufbaus darstellt:

1. **Geometrische Augmentierung** (nur im Training): `RandomCrop` mit vier Pixeln Randauffüllung und `RandomHorizontalFlip`.
2. **Farbraumkonvertierung** von RGB in den Zielfarbraum.
3. **Umwandlung in einen Tensor** der Form $3 \times 32 \times 32$.
4. **Kanalweise Normalisierung** (siehe Abschnitt 5.2).

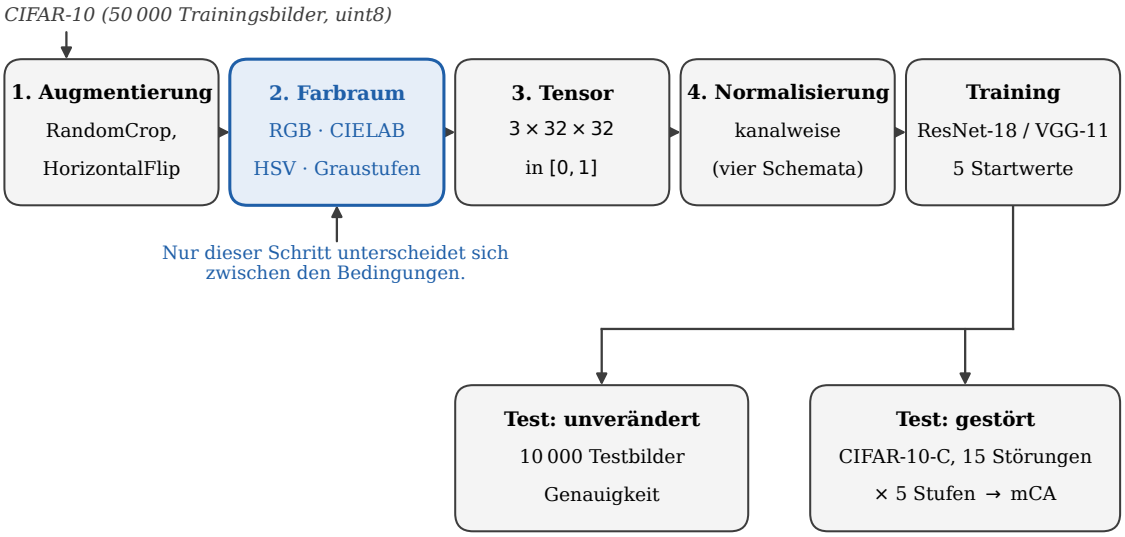


Abbildung 4: Versuchsaufbau im Überblick. Ein Bild durchläuft vier Vorverarbeitungsschritte, bevor es das Netz erreicht; das trainierte Netz wird anschließend zweimal ausgewertet – auf den unveränderten Testdaten und auf den gestörten Daten von CIFAR-10-C. Zwischen den Versuchsbedingungen unterscheidet sich ausschließlich der zweite Schritt.

Die Reihenfolge der ersten beiden Schritte ist bewusst gewählt: Zuschnitt und Spiegelung sind rein geometrische Operationen, die jedes Pixel unverändert an eine andere Position verschieben. Sie kommutieren daher mit jeder punktweisen Farbraumtransformation, sodass die Reihenfolge das Ergebnis nicht verändert. Die Augmentierung vor der Konvertierung auszuführen ist lediglich effizienter, da sie auf dem kompakten `uint8`-Format arbeitet.

Die Konvertierung erfolgt *on the fly*, also für jedes Bild beim Laden, statt einmalig einen konvertierten Datensatz auf die Festplatte zu schreiben. Damit bleibt sichergestellt, dass alle Bedingungen von identischen Rohdaten ausgehen und keine Zwischenspeicherung eine zusätzliche Fehlerquelle einführt.

5.1.2 Die vier Farbräume

Alle Konvertierungen verwenden `scikit-image` und liefern drei Kanäle im Intervall $[0, 1]$. Diese einheitliche Skalierung ist notwendig, damit die anschließende Normalisierung für alle Farbräume dieselbe Bedeutung hat.

RGB dient als Referenzbedingung. Die `uint8`-Werte werden lediglich durch 255 geteilt.

HSV wird mit `rgb2hsv` berechnet; Farbton, Sättigung und Hellwert liegen bereits im Intervall $[0, 1]$.

CIELAB wird mit `rgb2lab` berechnet und liefert $L^* \in [0, 100]$ sowie $a^*, b^* \in [-128, 127]$. Um die wahrnehmungsbezogene Bedeutung der Kanäle zu erhalten, werden diese affin auf $[0, 1]$ abgebildet:

$$L' = \frac{L^*}{100}, \qquad a' = \frac{a^* + 128}{255}, \qquad b' = \frac{b^* + 128}{255}. \tag{11}$$

Graustufen entstehen durch die luminanzgewichtete Summe der RGB-Kanäle. Als einzige der vier Konvertierungen verwendet sie `OpenCV` statt `scikit-image` und damit die Gewichte der Norm ITU-R BT.601 ($0{,}299\,R + 0{,}587\,G + 0{,}114\,B$) anstelle von BT.709. Auf natürlichen Bildern unterscheiden sich die beiden Normen im Mittel um etwa 1,2 % des Wertebereichs. Für die Rolle dieser Bedingung ist der Unterschied unerheblich: Beide Normen entfernen die chromatische Information vollständig, und der gemessene Leistungsabfall der Graustufen-Bedingung (rund sieben Prozentpunkte, siehe Kapitel 6) übersteigt diese Abweichung um mehr als eine Größenordnung. Entscheidend ist, dass die zugehörigen Normalisierungskonstanten mit demselben Konverter berechnet wurden und die Bedingung damit in sich konsistent bleibt. Der resultierende Einkanal wird dreifach dupliziert, sodass die Eingabe weiterhin die Form $3 \times 32 \times 32$ besitzt. Diese Verdopplung fügt keine Information hinzu; sie stellt lediglich sicher, dass in allen vier Bedingungen *dieselbe* Netzarchitektur mit derselben Parameterzahl trainiert wird. Ohne diesen Schritt würde die Graustufen-Bedingung ein anderes Netz vergleichen und wäre als Kontrollbedingung wertlos. Die Graustufen-Bedingung dient in dieser Arbeit als Negativkontrolle: Da sie den Farbanteil vollständig entfernt, muss ein funktionierendes Versuchsdesign hier einen deutlichen Leistungsabfall messen (vgl. Abschnitt 6.9).

5.2 Normalisierung der Eingaben

Nach der Konvertierung liegen alle Kanäle im Intervall $[0, 1]$, besitzen jedoch je nach Farbraum sehr unterschiedliche Verteilungen. Der Farbton in HSV ist nahezu

gleichverteilt, während die Chromakanäle a^* und b^* in CIELAB stark um ihren Mittelwert konzentriert sind. Da die Gewichte der ersten Faltungsschicht für alle Kanäle aus derselben Verteilung initialisiert werden, würde ein Kanal mit geringer Streuung von Beginn an ein schwächeres Signal liefern als seine Nachbarn. Die Normalisierung stellt daher sicher, dass jeder Kanal dem Netz mit vergleichbarer Skala präsentiert wird.

5.2.1 Vier Normalisierungsschemata

Um auszuschließen, dass die Ergebnisse von dieser Vorverarbeitung abhängen, wurden vier Schemata implementiert und miteinander verglichen:

- **z-Score** (Referenz): kanalweise Standardisierung auf Mittelwert 0 und Standardabweichung 1 mit den in Tabelle 2 angegebenen Konstanten, $x' = (x - \mu_c)/\sigma_c$.
- **Min-Max**: keine zusätzliche Transformation; die Kanäle bleiben im Intervall $[0, 1]$, wie sie der Konverter liefert.
- **Zentriert**: lineare Abbildung von $[0, 1]$ auf $[-1, 1]$, also $\mu = \sigma = 0{,}5$ für alle Kanäle.
- **Vereinheitlichte Skalierung**: kanalweise Zentrierung wie beim z-Score, jedoch Division sämtlicher Kanäle durch die Streuung des ersten Kanals, $x' = (x - \mu_c)/\sigma_1$ (vgl. Gleichung 13).

Die ersten drei Schemata unterscheiden sich darin, wie viel farbraumspezifisches Vorwissen sie einbringen: Der z-Score verwendet für jeden Farbraum eigene, aus den Daten geschätzte Konstanten, während Min-Max und die zentrierte Variante für alle Farbräume identisch sind. Findet sich derselbe Befund unter allen dreien, so kann er kein Artefakt einer bestimmten Vorverarbeitung sein.

Das vierte Schema verfolgt einen anderen Zweck. Es unterscheidet sich vom z-Score in genau einer Hinsicht – der Skalierung –, während die Zentrierung unverändert bleibt. Min-Max und zentrierte Normalisierung verändern dagegen beides zugleich. Diese Variante erlaubt es daher, den Einfluss der kanalweisen Skalierung isoliert zu bestimmen; sie wird in Abschnitt 6.4 verwendet.

5.2.2 Bestimmung der Konstanten

Die z-Score-Konstanten wurden über den vollständigen Trainingsdatensatz von 50 000 Bildern berechnet, und zwar mit *demselben* Konverter, der auch im Training verwendet wird (`compute_cifar10_stats.py`). Diese Konstruktion ist wesentlich: Würden die Statistiken mit einer anderen Implementierung geschätzt als der, die die Bilder erzeugt, könnten Mittelwert und Streuung systematisch von den tatsächlich beim Training auftretenden Werten abweichen.

Tabelle 2: Kanalweise Mittelwerte und Standardabweichungen, berechnet über die 50 000 Trainingsbilder von CIFAR-10 nach der jeweiligen Farbraumkonvertierung.

Farbraum	Kanal 1	Kanal 2	Kanal 3
RGB	$0{,}4914 \pm 0{,}2470$	$0{,}4822 \pm 0{,}2435$	$0{,}4465 \pm 0{,}2616$
HSV	$0{,}3254 \pm 0{,}2765$	$0{,}2738 \pm 0{,}2184$	$0{,}5393 \pm 0{,}2472$
CIELAB	$0{,}5085 \pm 0{,}2426$	$0{,}5035 \pm 0{,}0398$	$0{,}5244 \pm 0{,}0631$
Graustufen	$0{,}4809 \pm 0{,}2392$	(identisch, dupliziert)	

Bemerkenswert ist die geringe Streuung der Chromakanäle in CIELAB (0,0398 und 0,0631 gegenüber 0,2426 für die Helligkeit): Nach der Skalierung aus Gleichung 11 nutzen a^* und b^* nur einen schmalen Teil des verfügbaren Wertebereichs aus. Die z-Score-Normalisierung teilt jeden Kanal durch seine eigene Streuung und verstärkt diese beiden Kanäle daher um ein Vielfaches dessen, was ein RGB-Kanal erfährt. Diese Eigenschaft wird in Abschnitt 7.1 zur Erklärung eines zentralen Ergebnisses herangezogen.

5.3 Modellarchitektur und Training

5.3.1 ResNet-18 und die Anpassung an 32×32

Als Modell dient ein ResNet-18 (He et al., 2016) aus der Bibliothek `timm`. Die Wahl fiel auf diese Architektur, weil sie klein genug ist, um jede Versuchsbedingung über fünf Zufallsstartwerte hinweg in vertretbarer Rechenzeit zu wiederholen, und zugleich etabliert genug, dass die Ergebnisse vergleichbar bleiben.

Die Standardvariante ist für Eingaben der Größe 224×224 ausgelegt und reduziert die räumliche Auflösung bereits am Netzeingang zweifach: durch eine Faltung mit 7×7-Kern und Schrittweite 2 sowie durch eine anschließende Max-Pooling-Schicht. Auf CIFAR-10 angewendet würde ein Bild dadurch unmittelbar von 32×32 auf 8×8 schrumpfen, noch bevor der erste Residualblock erreicht wird. Die für kleine Auflösungen übliche Anpassung ersetzt deshalb die erste Faltung durch einen 3×3-Kern mit Schrittweite 1 und entfernt das Max-Pooling. Alle übrigen Schichten bleiben unverändert.

Die Normalisierungsschicht ist parametrisierbar. In der Referenzkonfiguration wird die Voreinstellung von `timm` verwendet, also Batch-Normalisierung (Ioffe and Szegedy, 2015). Für das in Abschnitt 6.6 beschriebene Experiment wird sie durch Gruppen-Normalisierung mit 32 Gruppen (Wu and He, 2018) ersetzt. Da beide Schichten dieselbe Schnittstelle besitzen, betrifft dieser Austausch ausschließlich die Normalisierung; Tiefe, Kanalzahl, Parameterzahl und alle übrigen Hyperparameter bleiben identisch. Dies ist die Voraussetzung dafür, den beobachteten Unterschied der Normalisierungsschicht zuschreiben zu können und nicht einem Architekturwechsel.

5.3.2 VGG-11 als zweite Architektur

Um zu prüfen, ob die Befunde Eigenschaften des Eingabeformats oder Eigenheiten des ResNet-18 beschreiben, wurde die Untersuchung auf einem VGG-11 (Simonyan and Zisserman, 2015) wiederholt (Abschnitt 6.8). Diese Architektur eignet sich dafür aus zwei Gründen. Sie unterscheidet sich vom ResNet-18 grundlegend, da sie keine Residualverbindungen enthält, und sie lässt sich *ohne jede Normalisierungsschicht* trainieren – was beim ResNet-18 wegen seiner Tiefe nicht ohne Weiteres möglich ist.

Trainiert wurden daher zwei Konfigurationen: eine mit Batch-Normalisierung und eine gänzlich ohne Normalisierungsschicht. Zusammen mit den beiden ResNet-Konfigurationen ergibt sich eine Abstufung über vier Bedingungen, von der datensatzweiten Statistik über die bildweise Statistik bis zum vollständigen Verzicht.

Die erste Faltungsschicht des VGG-11 besitzt wie die des ResNet-18 die Form $[64, 3, 3, 3]$, sodass die Analyse der Eingangsfilter aus Abschnitt 6.7 unverändert übertragbar wäre. Sämtliche Hyperparameter aus Tabelle 3 bleiben unverändert.

5.3.3 Training ohne Vortraining

Alle Modelle werden von zufälliger Initialisierung an trainiert. Auf vortrainierte Gewichte wurde bewusst verzichtet: Öffentlich verfügbare Gewichte sind auf ImageNet in RGB gelernt worden. Ein damit initialisiertes Netz brächte für die RGB-Bedingung bereits passende Filter der ersten Schichten mit, während CIELAB und HSV zunächst gegen diese Voreinstellung anarbeiten müssten. Der Vergleich wäre systematisch zugunsten von RGB verzerrt. Von zufälliger Initialisierung aus starten dagegen alle vier Bedingungen unter identischen Voraussetzungen.

Tabelle 3: Trainingskonfiguration. Alle Werte sind für die vier Farbräume und beide Normalisierungsschichten identisch.

Parameter	Wert
Epochen	50
Batchgröße	32
Optimierer	Adam
Lernrate	$1 \cdot 10^{-3}$
Gewichtsabnahme	0
Lernratenplan	StepLR, Faktor 0,1 nach 25 Epochen
Verlustfunktion	Kreuzentropie
Zufallsstartwerte	0, 1, 2, 3, 42

Auf eine Gewichtsabnahme wurde verzichtet, obwohl für ResNets auf CIFAR-10 üblicherweise ein Wert von $5 \cdot 10^{-4}$ verwendet wird. Dieser Verzicht hat für die vorliegende Fragestellung einen Vorteil: Vor einer Normalisierungsschicht ist die Ausgabe des Netzes invariant gegenüber der Skala der vorangehenden Gewichte. Eine L_2-Strafe wirkt dort weniger als Regularisierung denn als Veränderung der effektiven

Lernrate (van Laarhoven, 2017), und dieser Effekt fällt unter Batch- und Gruppen-Normalisierung unterschiedlich aus. Bei einer von null verschiedenen Gewichtsabnahme würde der Vergleich beider Normalisierungsschichten (Abschnitt 6.6) somit zwei Änderungen zugleich enthalten. Mit dem Wert null entfällt dieser Störfaktor.

5.3.4 Reproduzierbarkeit

Vor jedem Lauf werden die Zufallsgeneratoren von `random`, `NumPy` und `PyTorch` (einschließlich CUDA) auf denselben Wert gesetzt. Zusätzlich wird cuDNN in den deterministischen Modus geschaltet und die automatische Algorithmenauswahl deaktiviert. Ein Lauf ist damit bei gleichem Startwert und gleicher Hardware wiederholbar. Die vollständige Konfiguration wird zu Beginn jedes Laufs in das Ergebnisverzeichnis geschrieben, sodass jede berichtete Zahl auf die Einstellungen zurückgeführt werden kann, mit denen sie erzeugt wurde.

5.3.5 Modellauswahl

Nach jeder Epoche wird das Modell auf dem Testdatensatz ausgewertet; die Gewichte der Epoche mit der höchsten Genauigkeit werden gespeichert und für alle weiteren Auswertungen verwendet. Diese Vorgehensweise verwendet den Testdatensatz zur Modellauswahl und liefert daher leicht optimistische Absolutwerte. Für die Fragestellung dieser Arbeit ist das unproblematisch, da derselbe Auswahlmechanismus auf alle Bedingungen in gleicher Weise angewendet wird und die berichteten Größen Differenzen zwischen Bedingungen sind. Der Punkt wird in Abschnitt 7.3 erneut aufgegriffen.

5.4 Evaluationsprotokoll

Die Bewertung erfolgt auf zwei Ebenen. Die erste misst, wie gut ein Modell die Aufgabe unter denselben Bedingungen löst, unter denen es trainiert wurde; die zweite misst, was von dieser Leistung übrig bleibt, wenn die Eingabebilder gestört werden. Nur die zweite Ebene beantwortet die Forschungsfrage dieser Arbeit, denn hohe Genauigkeit auf ungestörten Daten sagt wenig über die Robustheit eines Modells aus.

5.4.1 Genauigkeit auf ungestörten Daten

Als erste Kennzahl dient die Klassifikationsgenauigkeit auf den 10 000 Testbildern von CIFAR-10. Sie wird für jede Epoche berechnet; berichtet wird der Wert der besten Epoche (siehe Abschnitt 5.3.5).

5.4.2 Robustheit auf CIFAR-10-C

Die Robustheit wird auf CIFAR-10-C (Hendrycks and Dietterich, 2019) gemessen. Dieser Benchmark enthält den Testdatensatz von CIFAR-10 in gestörten Fassungen: Jede Störungsart liegt in fünf Schweregraden vor, wobei jeder Schweregrad sämtliche 10 000 Testbilder umfasst.

Die Veröffentlichung unterscheidet dabei zwei Gruppen. Fünfzehn Störungsarten bilden den eigentlichen Benchmark und definieren die üblicherweise berichteten Kennzahlen; vier weitere – *speckle noise*, *gaussian blur*, *spatter* und *saturate* – werden als zusätzliche Störungen bereitgestellt und gehen in der Literatur regelmäßig nicht in die Kennzahl ein. Als Hauptkennzahl dieser Arbeit dienen daher die fünfzehn Standardstörungen; dies stellt die Vergleichbarkeit mit bestehenden Arbeiten sicher. Die vier zusätzlichen Störungen werden getrennt ausgewertet und dienen als unabhängige Prüfmenge (Abschnitt 6.9).

Tabelle 4: Die fünfzehn Standardstörungen von CIFAR-10-C, gruppiert nach Herkunft. Die drei mit $*$ markierten bilden die Untergruppe der helligkeitsbezogenen Störungen, auf die sich die Arbeitshypothese bezieht. Die vier zurückgehaltenen Validierungsstörungen sind nicht aufgeführt.

Gruppe	Störungsarten
Rauschen	gaussian noise, shot noise, impulse noise
Unschärfe	defocus blur, glass blur, motion blur, zoom blur
Wetter	snow, frost, fog*, brightness*
Digital	contrast*, elastic transform, pixelate, jpeg compression

Ein bereits trainiertes Modell wird ohne jede Anpassung auf allen $15 \times 5 = 75$ gestörten Testmengen ausgewertet. Die Störungen treten ausschließlich zur Testzeit auf und sind dem Modell während des Trainings nie begegnet. Wichtig ist dabei, dass die gestörten Bilder dieselbe Verarbeitung durchlaufen wie die Trainingsbilder: Farbraumkonvertierung und Normalisierung werden mit identischen Konstanten angewandt. Die Störung wirkt also im RGB-Raum, bevor die Konvertierung stattfindet – genau wie eine reale Bildstörung, die eine Kamera aufzeichnet, bevor eine Software das Bild umrechnet.

5.4.3 Kennzahlen

Aus diesen 75 Messungen werden drei Kennzahlen gebildet. Sei $\mathrm{acc}(c, s)$ die Genauigkeit bei Störungsart c und Schweregrad s sowie $\mathcal{C}$ die Menge der fünfzehn Standardstörungen:

$$\mathrm{mCA} = \frac{1}{|\mathcal{C}| \cdot 5} \sum_{c \in \mathcal{C}} \sum_{s=1}^{5} \mathrm{acc}(c, s). \tag{12}$$

Analog wird $\mathrm{mCA}_{\mathrm{Helligkeit}}$ über die Teilmenge $\{\textit{brightness}, \textit{fog}, \textit{contrast}\}$ und $\mathrm{mCA}_{\mathrm{Rest}}$ über die verbleibenden zwölf Störungsarten berechnet.

Die Trennung in diese beiden Teilmengen ist keine nachträgliche Auswertung, sondern folgt unmittelbar aus der Arbeitshypothese: Wenn ein Farbraum wie CIELAB deshalb hilft, weil er die Helligkeit als eigenen Kanal führt, dann muss sich sein Vorteil vor allem dort zeigen, wo die Störung die Helligkeit verändert. Die Untergruppe wurde vor der Auswertung festgelegt. Die Kennzahl $\mathrm{mCA}_{\mathrm{Rest}}$ dient als Gegenprobe: Zeigte sich ein vermeintlicher Helligkeitsvorteil in gleicher Größe auch dort, wäre er nicht auf den vermuteten Mechanismus zurückzuführen.

Der Mittelwert über alle Störungsarten ist die in der Literatur übliche Kennzahl und erlaubt den Anschluss an bestehende Arbeiten. Er hat jedoch eine Schwäche, die in Abschnitt 6.3 zentral wird: Gegenläufige Effekte auf verschiedenen Störungsarten heben sich in ihm auf. Ein Farbraum, der auf Rauschen deutlich verliert und auf Unschärfe deutlich gewinnt, erscheint im Mittel unauffällig. Der Wert von mCA hängt dann davon ab, wie viele Störungen welcher Art in der Menge $\mathcal{C}$ enthalten sind – eine Eigenschaft der Kennzahl, nicht des Modells. Aus diesem Grund wird die Genauigkeit zusätzlich für jede Störungsart einzeln gespeichert und ausgewertet; die aggregierten Kennzahlen werden nie ohne diese Aufschlüsselung interpretiert.

5.5 Statistisches Protokoll

Die in dieser Arbeit erwarteten Unterschiede zwischen Farbräumen liegen in der Größenordnung von ein bis zwei Prozentpunkten. Ein einzelner Trainingslauf kann diese Frage nicht beantworten, denn seine Genauigkeit hängt auch von der zufälligen Initialisierung der Gewichte und der zufälligen Reihenfolge der Trainingsbeispiele ab. Ohne ein Maß für diese Schwankung ist nicht entscheidbar, ob ein beobachteter Unterschied auf den Farbraum oder auf den Zufall zurückgeht.

5.5.1 Wiederholung über Zufallsstartwerte

Jede Bedingung wurde daher mit fünf verschiedenen Zufallsstartwerten (0, 1, 2, 3 und 42) vollständig trainiert und ausgewertet – über vier Farbräume, zwei Architekturen, drei Normalisierungsschichten und vier Normalisierungsschemata hinweg. Die Streuung über die Startwerte liefert die Vergleichsgröße, an der jeder gemessene Unterschied zu beurteilen ist: Ein Unterschied, der kleiner ist als die Schwankung zwischen zwei Läufen derselben Bedingung, kann nicht dem Farbraum zugeschrieben werden. Eine Ausnahme bilden die Min-Max- und die zentrierte Normalisierung, die als Kontrollbedingungen dienen und mit drei Startwerten (0, 1 und 42) ausgewertet wurden; die betreffenden Angaben sind jeweils gekennzeichnet.

Wesentlich ist, dass alle Bedingungen *dieselben* fünf Startwerte verwenden. Dadurch lassen sich die Läufe paarweise zuordnen: Der RGB-Lauf mit Startwert 0 und der CIELAB-Lauf mit Startwert 0 teilen dieselbe Initialisierungssequenz und dieselbe

Datenreihenfolge. Der Vergleich kontrolliert damit einen Teil der Zufallsvariation, statt sie als Rauschen hinnehmen zu müssen.

5.5.2 Testverfahren

Verglichen wird jeweils paarweise mit einem gepaarten t-Test über die fünf Startwerte. Da fünf Beobachtungen für eine Prüfung der Normalverteilungsannahme zu wenig sind, wird zusätzlich der Wilcoxon-Vorzeichen-Rang-Test als verteilungsfreie Alternative berechnet; berichtet werden Abweichungen zwischen beiden Verfahren. Dabei ist eine Eigenschaft dieses Tests zu beachten: Bei fünf Paaren beträgt der kleinste erreichbare zweiseitige p-Wert $2/2^5 = 0{,}0625$. Er kann das übliche Signifikanzniveau von 0,05 also grundsätzlich nicht unterschreiten und dient hier ausschließlich der Absicherung gegen Verteilungsannahmen, nicht als eigenständiges Entscheidungskriterium.

Der p-Wert allein ist bei fünf Beobachtungen wenig aussagekräftig. Er wird daher stets zusammen mit zwei weiteren Größen angegeben:

- der **Effektstärke** nach Cohen für gepaarte Daten, $d = \bar{\delta}/s_\delta$, wobei $\bar{\delta}$ und s_δ Mittelwert und Standardabweichung der paarweisen Differenzen sind. Sie beziffert den Unterschied in Einheiten seiner eigenen Streuung und ist unabhängig von der Stichprobengröße;
- der **Konsistenz über die Startwerte**, also der Anzahl der Startwerte, bei denen die Differenz dasselbe Vorzeichen hat. Ein Effekt, der bei fünf von fünf Startwerten in dieselbe Richtung weist, ist anders zu bewerten als einer, der bei drei von fünf auftritt, selbst bei gleichem p-Wert.

5.5.3 Korrektur für multiples Testen

Diese Arbeit vergleicht mehrere Farbräume auf mehreren Kennzahlen. Mit jedem zusätzlichen Test steigt die Wahrscheinlichkeit, allein durch Zufall mindestens einen Wert unterhalb der üblichen Schwelle von 0,05 zu finden. Über alle berichteten Vergleiche wird daher die Holm-Bonferroni-Korrektur angewandt und neben dem unkorrigierten stets der korrigierte p-Wert angegeben.

5.5.4 Vorab festgelegte und nachträgliche Analysen

Ein p-Wert lässt sich nur dann als Wahrscheinlichkeit lesen, wenn die Hypothese vor der Betrachtung der Daten feststand. Diese Arbeit unterscheidet deshalb konsequent zwischen zwei Arten von Befunden:

- **Vorab festgelegte Analysen** – die Haupthypothese (CIELAB gegenüber RGB auf helligkeitsbezogenen Störungen), die Untergruppenbildung aus Abschnitt 5.4, die Vorhersage des GroupNorm-Experiments sowie die vier Vorher-

sagen zur Wiederholung auf der zweiten Architektur (Abschnitt 6.8); letztere wurden vor dem Start der zugehörigen Rechenläufe schriftlich festgehalten.

- **Nachträgliche Beobachtungen** – alle Befunde, die sich erst bei der Auswertung ergaben, insbesondere die Aufschlüsselung nach einzelnen Störungsarten. Sie werden als solche gekennzeichnet und nicht als bestätigte Hypothesen dargestellt.

5.5.5 Systematische Wiederholung der Prüfungen

Über die statistischen Tests hinaus wurde jeder zentrale Befund unter veränderten Bedingungen wiederholt: über fünf Zufallsstartwerte, über die drei Normalisierungsschemata aus Abschnitt 5.2, mit neu berechneten Normalisierungskonstanten sowie auf den zurückgehaltenen Validierungsstörungen. Ein Befund gilt in dieser Arbeit erst dann als belastbar, wenn er alle diese Prüfungen übersteht. Die Ergebnisse dieser Wiederholungen werden in Abschnitt 6.9 berichtet.

6 Experimente und Ergebnisse

Dieses Kapitel berichtet die Ergebnisse in der Reihenfolge, in der sich die Fragestellung entwickelt hat. Abschnitt 6.1 prüft zunächst, ob der Farbraum die Genauigkeit auf ungestörten Daten und die Konvergenzgeschwindigkeit beeinflusst. Abschnitt 6.2 wendet sich der eigentlichen Arbeitshypothese zu und zeigt, dass sie sich nicht bestätigt. Die folgenden Abschnitte gehen der Frage nach, was hinter diesem Nullbefund steht: Abschnitt 6.3 deckt zwei gegenläufige Effekte auf, die sich im Mittelwert aufheben, Abschnitt 6.4 klärt deren Ursache in einem gezielten Experiment, Abschnitt 6.5 ordnet den Befund zu HSV ein, und die Abschnitte 6.6 und 6.7 prüfen die vorgeschlagene Erklärung auf zwei unabhängigen Wegen. Abschnitt 6.8 wiederholt die Untersuchung anhand vorab festgelegter Vorhersagen auf einer zweiten Architektur, und Abschnitt 6.9 fasst schließlich zusammen, welchen Prüfungen die berichteten Befunde standhalten.

Sofern nicht anders angegeben, beziehen sich alle Werte auf fünf Zufallsstartwerte, die fünfzehn Standardstörungen von CIFAR-10-C und die z-Score-Normalisierung nach Abschnitt 5.2. Angegebene p-Werte sind unkorrigiert; die Korrektur für multiples Testen wird in Abschnitt 6.9 berichtet.

6.1 Genauigkeit auf unveränderten Testdaten

Tabelle 5 zeigt die Genauigkeit auf dem ungestörten Testdatensatz. Die drei Farbräume liegen innerhalb von 0,36 Prozentpunkten beieinander, während die Graustufen-Bedingung um 1,72 Prozentpunkte zurückbleibt.

Für die Arbeitshypothese ist der Vergleich zwischen CIELAB und RGB maßgeblich. CIELAB liegt um 0,18 Prozentpunkte vorn, was bei einem gepaarten t-Test über

Tabelle 5: Genauigkeit auf den unveränderten Testdaten von CIFAR-10 (Mittelwert ± Standardabweichung über fünf Zufallsstartwerte, BatchNorm-Modelle).

Farbraum	Genauigkeit (%)
CIELAB	$92{,}76 \pm 0{,}14$
RGB	$92{,}58 \pm 0{,}13$
HSV	$92{,}40 \pm 0{,}08$
Graustufen	$90{,}95 \pm 0{,}22$

die fünf Startwerte nicht signifikant ist ($p = 0{,}14$; CIELAB gewinnt bei vier von fünf Startwerten). Der Unterschied liegt in derselben Größenordnung wie die Streuung zwischen zwei Läufen derselben Bedingung und ist damit nicht vom Zufall der Initialisierung zu trennen.

Für HSV zeigt sich ein kleiner, aber durchgängiger Rückstand: Der Farbraum liegt bei allen fünf Startwerten hinter RGB ($-0{,}18$ Prozentpunkte, $p = 0{,}034$, $d = 1{,}41$) und hinter CIELAB ($-0{,}36$ Prozentpunkte, $p = 0{,}004$, $d = 2{,}75$). Der Effekt ist konsistent, in seiner absoluten Größe jedoch gering.

Deutlich abgesetzt ist allein die Graustufen-Bedingung: Sie verliert gegenüber RGB 1,72 Prozentpunkte bei allen fünf Startwerten ($p < 0{,}001$). Dieser Befund ist inhaltlich nicht überraschend – die Bedingung entfernt die Farbinformation vollständig –, erfüllt aber die ihm zugedachte Funktion als Negativkontrolle: Das Versuchsdesign ist in der Lage, einen tatsächlich vorhandenen Informationsverlust zu messen. Dass es zwischen den drei Farbräumen keinen vergleichbaren Unterschied findet, lässt sich damit nicht auf eine mangelnde Empfindlichkeit des Verfahrens zurückführen.

6.1.1 Konvergenzverhalten

Die Endgenauigkeit sagt nichts darüber aus, wie schnell sie erreicht wird. Da CIELAB die Trennung von Helligkeit und Farbe bereits mitbringt, liegt die Vermutung nahe, dass ein damit trainiertes Netz früher ein gegebenes Genauigkeitsniveau erreicht. Abbildung 5 stellt den Verlauf der Testgenauigkeit über die fünfzig Epochen dar.

Ein Vorteil von CIELAB zeigt sich nicht. Die Schwelle von 85 Prozent erreichen beide Bedingungen bei exakt denselben Epochen (im Mittel $7{,}80 \pm 0{,}84$), die Schwelle von 80 Prozent ebenfalls ($4{,}60 \pm 0{,}55$). Bei 90 Prozent liegt CIELAB um 0,80 Epochen vorn (18,60 gegenüber 19,40), was bei zwei von fünf Startwerten zutrifft und nicht signifikant ist ($p = 0{,}41$). HSV benötigt durchgängig etwas länger, ohne dass der Unterschied gesichert wäre. Lediglich in der ersten Epoche liegt RGB um 1,76 Prozentpunkte vor CIELAB; dieser Vorsprung ist ab Epoche fünf verschwunden.

Die Negativkontrolle greift auch hier. Die Graustufen-Bedingung benötigt bis zur 85-Prozent-Schwelle $10{,}00 \pm 1{,}00$ statt 7,80 Epochen ($p = 0{,}029$, bei allen fünf Startwerten langsamer) und erreicht 90 Prozent erst nach der Absenkung der Lernrate.

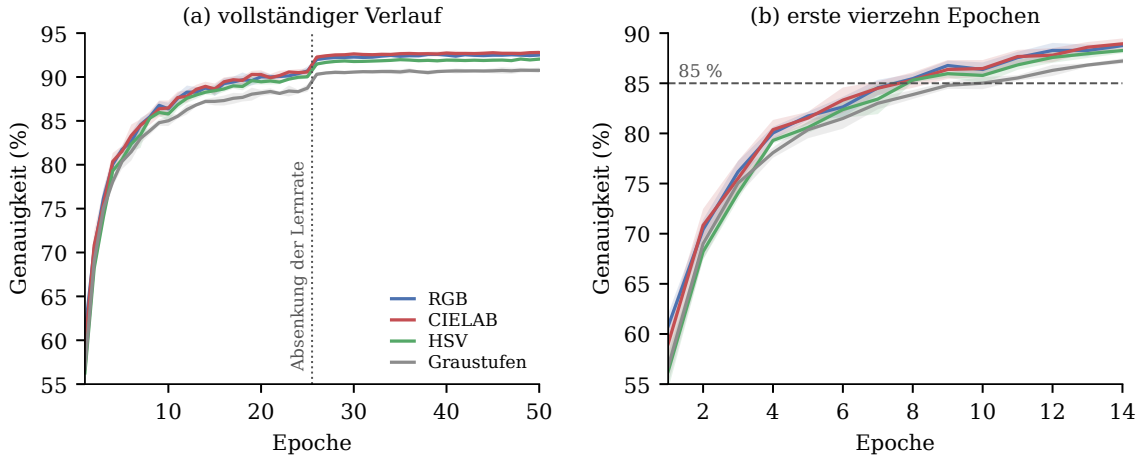


Abbildung 5: Genauigkeit auf den unveränderten Testdaten im Verlauf des Trainings (Mittelwert über fünf Zufallsstartwerte, schattierter Bereich: Standardabweichung). **(a)** Der vollständige Verlauf; die gepunktete Linie markiert die Absenkung der Lernrate nach Epoche 25. **(b)** Ausschnitt der ersten vierzehn Epochen, in denen sich Unterschiede im Lerntempo zeigen müssten. Die drei Farbräume verlaufen nahezu deckungsgleich; allein die Graustufen-Bedingung bleibt zurück.

Die Messung ist also in der Lage, einen Unterschied im Lerntempo aufzudecken, wenn einer vorliegt – zwischen den drei Farbräumen findet sie keinen. Zu beachten ist dabei, dass der Sprung aller Kurven bei Epoche 26 nicht vom Farbraum herrührt, sondern von der Lernratenabsenkung nach Epoche 25 (Tabelle 3).

Festzuhalten bleibt, dass sich die Genauigkeit auf ungestörten Daten und das Konvergenzverhalten kaum zwischen den Farbräumen unterscheiden. Beides ist jedoch nicht die Kennzahl, auf die sich die Arbeitshypothese bezieht; diese betrifft die Robustheit gegenüber Störungen und wird im folgenden Abschnitt geprüft.

6.2 Die Haupthypothese bestätigt sich nicht: RGB vs. CIELAB

Tabelle 6: Mittlere Korruptions-Genauigkeit auf den fünfzehn Standardstörungen von CIFAR-10-C (Mittelwert ± Standardabweichung über fünf Zufallsstartwerte, BatchNorm-Modelle).

Farbraum	mCA (%)	mCA Helligkeit (%)	mCA Rest (%)
CIELAB	$71{,}79 \pm 0{,}68$	$82{,}08 \pm 0{,}51$	$69{,}22 \pm 0{,}74$
HSV	$71{,}73 \pm 0{,}68$	$81{,}40 \pm 0{,}36$	$69{,}32 \pm 0{,}77$
RGB	$71{,}66 \pm 0{,}52$	$81{,}87 \pm 0{,}36$	$69{,}11 \pm 0{,}58$
Graustufen	$64{,}59 \pm 0{,}34$	$80{,}66 \pm 0{,}20$	$60{,}57 \pm 0{,}41$

Die Arbeitshypothese sagt vorher, dass CIELAB robuster gegenüber helligkeitsbezogenen Störungen ist als RGB, weil es die Helligkeit als eigenen Kanal führt. Tabelle 6 enthält die dafür maßgeblichen Werte.

6.2.1 Der vorhergesagte Vorteil bleibt aus

Auf der Untergruppe der helligkeitsbezogenen Störungen liegt CIELAB um 0,21 Prozentpunkte vor RGB. Der gepaarte t-Test über die fünf Startwerte ergibt $p = 0{,}31$, der Wilcoxon-Test $p = 0{,}63$; die Effektstärke beträgt $d = 0{,}52$. CIELAB gewinnt bei drei von fünf Startwerten.

Aufschlussreicher als diese Kennzahlen sind die Einzelwerte. Die Differenz CIELAB − RGB beträgt je nach Startwert +0,17, +0,56, −0,18, −0,18 und +0,67 Prozentpunkte: Sie wechselt das Vorzeichen. Zum Vergleich beträgt die Standardabweichung zwischen fünf Läufen *derselben* Bedingung 0,36 Prozentpunkte – also mehr als der gemessene Unterschied zwischen den beiden Bedingungen. Der Effekt ist damit kleiner als das Rauschen, aus dem er gewonnen wurde.

Dasselbe Bild ergibt sich auf den beiden übrigen Kennzahlen: über alle fünfzehn Störungen +0,13 Prozentpunkte ($p = 0{,}41$) und auf den zwölf nicht helligkeitsbezogenen Störungen +0,11 Prozentpunkte ($p = 0{,}56$), jeweils mit drei Erfolgen von fünf. Auch die Gegenprobe fällt damit negativ aus: Der minimale Vorsprung auf den Helligkeitsstörungen findet sich in gleicher Größe auf allen übrigen Störungen wieder. Es gibt also keinen Hinweis auf einen Vorteil, der gerade dort entstünde, wo der vermutete Mechanismus ihn erzeugen sollte.

6.2.2 Der Befund ist stabil

Damit dieser Nullbefund nicht selbst ein Artefakt ist, wurde er unter veränderten Bedingungen wiederholt. Unter der zentrierten Normalisierung beträgt der Unterschied auf den Helligkeitsstörungen +0,21 Prozentpunkte ($p = 0{,}38$), unter der Min-Max-Normalisierung −0,39 Prozentpunkte ($p = 0{,}54$) – das Vorzeichen kehrt sich um. Auch diese beiden Schemata, die die Kanäle nicht einzeln skalieren und CIELAB damit anders präsentieren als die z-Score-Normalisierung, lassen keinen Vorteil hervortreten. Ein viertes Schema, das die Spreizung der Chromakanäle vollständig aufhebt, wird in Abschnitt 6.4 behandelt; es führt als einziges zu einem Ergebnis in der vorhergesagten Richtung.

6.2.3 Einordnung des Nullbefunds

Ein nicht signifikantes Ergebnis belegt für sich genommen nicht die Abwesenheit eines Effekts. Drei Beobachtungen zusammen rechtfertigen hier gleichwohl die Aussage, dass die Hypothese in der untersuchten Konfiguration nicht zutrifft.

Erstens sagt die Hypothese keinen beliebig kleinen, sondern einen mechanistisch begründeten und daher sichtbaren Effekt vorher: Wenn eine Helligkeitsstörung in

CIELAB nur einen Kanal verändert statt drei, sollte sich das messen lassen. Gemessen wird stattdessen ein Unterschied, der kleiner ist als die Schwankung zwischen zwei identischen Läufen.

Zweitens überdauert der Befund jede angewandte Prüfung – fünf Startwerte und drei Normalisierungsschemata – ohne sich in einer einzigen Konfiguration umzukehren.

Drittens, und am wichtigsten: Dasselbe Verfahren misst sehr wohl Effekte, wo welche vorhanden sind. Die Graustufen-Bedingung verliert gegenüber RGB 6,97 Prozentpunkte auf denselben fünfzehn Störungen, bei allen fünf Startwerten und mit $p < 0{,}001$ (beide Werte aus derselben Trainingsserie, siehe Abschnitt 6.9). Das Versuchsdesign ist also nicht blind; es findet an der Stelle, an der die Hypothese einen Effekt vorhersagt, keinen.

Die Hypothese bestätigt sich damit für die untersuchte Konfiguration nicht – ResNet-18, CIFAR-10, Bildgröße 32×32, z-Score-Standardisierung. Ein nicht signifikanter Unterschied belegt allerdings nicht die Abwesenheit eines Effekts; er zeigt, dass ein Effekt der vorhergesagten Größe mit diesem Versuchsaufbau nicht nachweisbar ist. Eine allgemeine Aussage über andere Architekturen oder Auflösungen ist damit nicht verbunden; die Grenzen dieser Aussage werden in Abschnitt 7.3 behandelt.

Die Nennung der Normalisierung in dieser Aufzählung ist bewusst gewählt; Abschnitt 6.4 zeigt, dass sie für diesen Vergleich nicht gleichgültig ist.

Der Nullbefund im Mittelwert bedeutet zudem nicht, dass der Farbraum wirkungslos bliebe. Der folgende Abschnitt zeigt, dass der Mittelwert zwei gegenläufige Effekte verdeckt.

6.3 Der versteckte Trade-off: Rauschen gegen Unschärfe

Der vorangegangene Abschnitt hat gezeigt, dass sich CIELAB und RGB in den aggregierten Kennzahlen nicht unterscheiden. Diese Kennzahlen mitteln jedoch über fünfzehn sehr verschiedene Störungsarten. Der folgende Abschnitt schlüsselt den Vergleich nach einzelnen Störungsarten auf.

6.3.1 Zwei gegenläufige Effekte

Tabelle 7 und Abbildung 6 zeigen ein Bild, das der aggregierte Nullbefund nicht erwarten ließ. Die Unterschiede sind keineswegs klein: Sie reichen von $-12{,}4$ bis $+7{,}3$ Prozentpunkten. Sie verteilen sich jedoch nicht zufällig, sondern folgen der Art der Störung.

Auf allen drei Rauschstörungen verliert CIELAB deutlich, und zwar bei *keinem einzigen* der fünf Startwerte anders. Zusammengefasst beträgt der Rückstand 9,72 Prozentpunkte ($p < 0{,}001$, $d = -6{,}7$, 0/5). Auf allen vier Unschärfestörungen gewinnt CIELAB ebenso ausnahmslos, im Mittel um 4,64 Prozentpunkte ($p < 0{,}001$, $d = +12{,}9$, 5/5). Effektstärken dieser Größenordnung sind nicht als außergewöhnlich

Tabelle 7: Differenz CIELAB − RGB je Störungsart (Prozentpunkte, Mittelwert über fünf Startwerte). Die Spalte „Konsistenz“ gibt an, bei wie vielen der fünf Startwerte CIELAB vor RGB liegt.

Störungsart	Gruppe	CIELAB − RGB	p	Konsistenz
gaussian noise	Rauschen	−12,37	$< 0{,}001$	0/5
shot noise	Rauschen	−11,58	$< 0{,}001$	0/5
impulse noise	Rauschen	−5,20	0,013	0/5
jpeg compression	Digital	−0,47	0,055	1/5
contrast	Digital	−0,39	0,489	2/5
brightness	Wetter	+0,02	0,839	3/5
fog	Wetter	+0,98	0,001	5/5
elastic transform	Digital	+1,83	$< 0{,}001$	5/5
snow	Wetter	+2,67	$< 0{,}001$	5/5
motion blur	Unschärfe	+3,43	0,005	5/5
defocus blur	Unschärfe	+3,50	0,001	5/5
frost	Wetter	+3,92	$< 0{,}001$	5/5
pixelate	Digital	+4,00	0,010	5/5
zoom blur	Unschärfe	+4,33	0,002	5/5
glass blur	Unschärfe	+7,29	0,002	5/5

großer Effekt zu lesen: Sie entstehen, weil die gepaarte Versuchsanordnung die Streuung zwischen den Startwerten weitgehend herausrechnet, sodass der Nenner sehr klein wird. Aussagekräftig sind hier die Differenzen in Prozentpunkten und die Konsistenz über die Startwerte, nicht der Absolutwert von d.

Zum Vergleich: Der in Abschnitt 6.2 berichtete Unterschied im Mittel über alle Störungen betrug 0,13 Prozentpunkte. Die hier gemessenen Einzeleffekte sind also rund fünfzig- bis hundertmal größer als der Wert, der sich nach der Mittelung ergibt. Der Nullbefund entsteht nicht dadurch, dass der Farbraum keine Wirkung hätte, sondern dadurch, dass zwei gegenläufige Wirkungen sich in der Summe nahezu aufheben.

6.3.2 Eine mechanistische Erklärung

Für dieses Muster bietet die Definition des CIELAB-Farbraums eine Erklärung. Die Transformation enthält eine Kubikwurzel, die auf die relativen Luminanzwerte angewandt wird. Ihre Ableitung $\frac{\mathrm{d}}{\mathrm{d}t}\, t^{1/3} = \frac{1}{3} t^{-2/3}$ wächst für kleine Argumente stark an: Bei einer relativen Luminanz von 0,02 beträgt sie 4,52, bei 0,5 nur 0,53 – ein Faktor von rund 8,5.

Eine additive Rauschstörung verändert einzelne Pixelwerte um einen etwa gleich großen Betrag, unabhängig von der lokalen Helligkeit. In dunklen Bildbereichen wird diese Änderung durch die Kubikwurzel um ein Vielfaches verstärkt, bevor das Netz

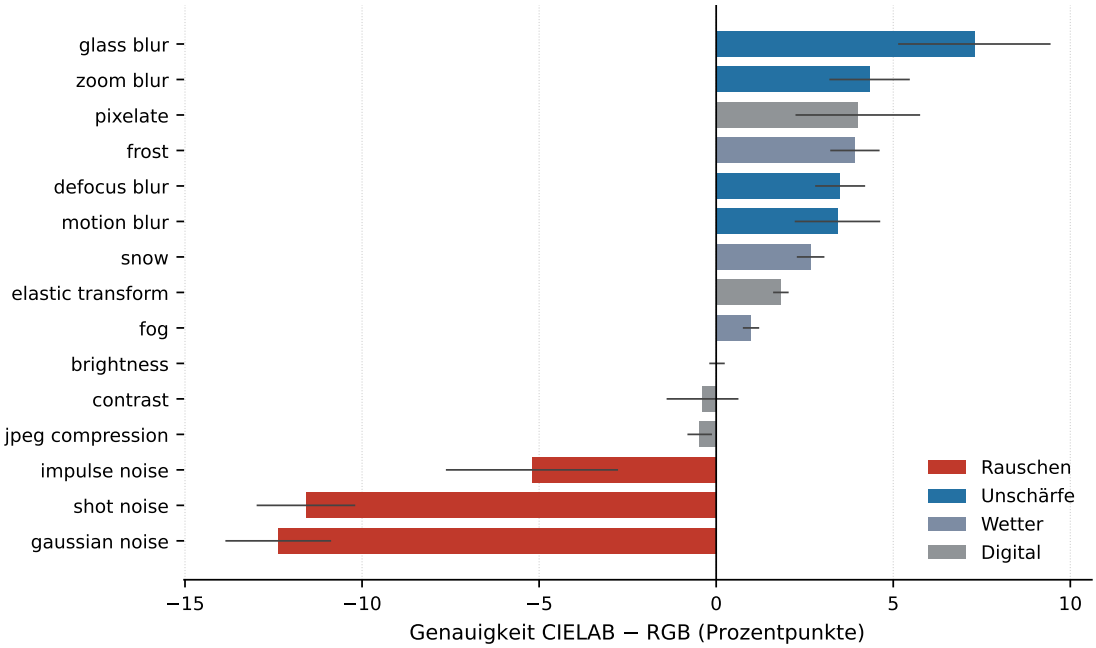


Abbildung 6: Differenz der Genauigkeit zwischen CIELAB und RGB, aufgeschlüsselt nach Störungsart (Mittelwert über fünf Zufallsstartwerte, Fehlerbalken: Standardabweichung). Die Störungsarten sind nach der Größe der Differenz geordnet. Rauschstörungen treten ausnahmslos am unteren, Unschärfestörungen am oberen Ende auf.

sie überhaupt zu sehen bekommt. Dieselbe Störung erscheint dem Netz in CIELAB also größer als in RGB.

Eine Unschärfestörung wirkt anders: Sie ersetzt jeden Pixelwert durch einen lokalen Mittelwert. Die Luminanz eines Bildbereichs bleibt dabei weitgehend erhalten, sodass keine vergleichbare Verstärkung auftritt. Warum CIELAB in diesem Fall sogar gewinnt, wird in Abschnitt 7.1 diskutiert.

6.3.3 Einordnung

Dieser Befund ist eine nachträgliche Beobachtung: Die Aufschlüsselung nach Störungsart war nicht Teil der ursprünglichen Fragestellung, sondern ergab sich bei der Auswertung. Er ist daher nicht als bestätigte Hypothese zu lesen. Drei Eigenschaften sprechen gleichwohl gegen einen Zufallsbefund: die Größe der Effekte, ihre vollständige Konsistenz über alle Startwerte (0/5 beziehungsweise 5/5 in beiden Richtungen), und der Umstand, dass die Gruppierung nicht nachträglich gewählt wurde, sondern der von Hendrycks and Dietterich (2019) vorgegebenen Kategorisierung folgt.

Für die Bewertung aggregierter Robustheitskennzahlen folgt daraus eine allgemeinere Beobachtung: Ein Mittelwert über Störungsarten setzt implizit voraus, dass ein Verfahren auf allen Störungen in dieselbe Richtung wirkt. Trifft das nicht zu,

hängt das Ergebnis davon ab, wie viele Störungen welcher Art in die Mittelung eingehen. Abschnitt 6.9 zeigt an den zurückgehaltenen Validierungsstörungen, dass dieser Einfluss messbar ist.

6.4 Ursache des Trade-offs: gezielte Variation der Kanalskalierung

Der im vorangegangenen Abschnitt berichtete Trade-off wirft die Frage nach seiner Ursache auf. Eine naheliegende Vermutung betrifft nicht die Farbraumtransformation selbst, sondern die anschließende Normalisierung.

6.4.1 Vermutung und Versuchsaufbau

Die z-Score-Standardisierung teilt jeden Kanal durch seine eigene Streuung. Für CIELAB beträgt diese bei den Chromakanälen lediglich 0,0398 und 0,0631 gegenüber 0,2426 für die Helligkeit (Tabelle 2); a^* und b^* werden dadurch um das Vier- bis Sechsfache dessen verstärkt, was ein RGB-Kanal erfährt. Die Vermutung lautet, dass diese Spreizung den Trade-off verursacht: Additives Rauschen ist hochfrequent und landet in den verstärkten Kanälen, während Unschärfe als Tiefpass die niederfrequente Farbstruktur erhält, die dort ebenfalls verstärkt ankommt.

Zur Prüfung wurde eine Normalisierungsvariante implementiert, die *ausschließlich* die Skalierung verändert: Die kanalweise Zentrierung bleibt unverändert, sämtliche Kanäle werden jedoch durch die Streuung des ersten Kanals geteilt,

$$x'_c = \frac{x_c - \mu_c}{\sigma_1}, \qquad c \in \{1, 2, 3\}, \tag{13}$$

mit $\sigma_1 = \sigma_{L^*}$ für CIELAB. Die relative Spreizung der Chromakanäle beträgt damit 1,00 statt 6,10. RGB dient als Kontrollbedingung: Da seine drei Streuungen mit 0,2470, 0,2435 und 0,2616 nahezu gleich sind, verändert der Eingriff dort kaum etwas. Trainiert wurde über dieselben fünf Zufallsstartwerte.

Tabelle 8: Differenz CIELAB − RGB auf Rausch- und Unschärfestörungen unter z-Score-Standardisierung und unter vereinheitlichter Skalierung (fünf Zufallsstartwerte). Die Amplitude bezeichnet die Differenz zwischen beiden Werten.

Normalisierung	Spreizung a^*	Rauschen	Unschärfe	Amplitude
z-Score	6,10×	−9,72	+4,64	14,35
vereinheitlicht	1,00×	−3,73	+1,94	**5,67**

6.4.2 Die Vermutung bestätigt sich

Die Amplitude des Trade-offs sinkt von 14,35 auf 5,67 Prozentpunkte und damit um rund sechzig Prozent. Abbildung 7 stellt diesen Befund zusammen mit den beiden

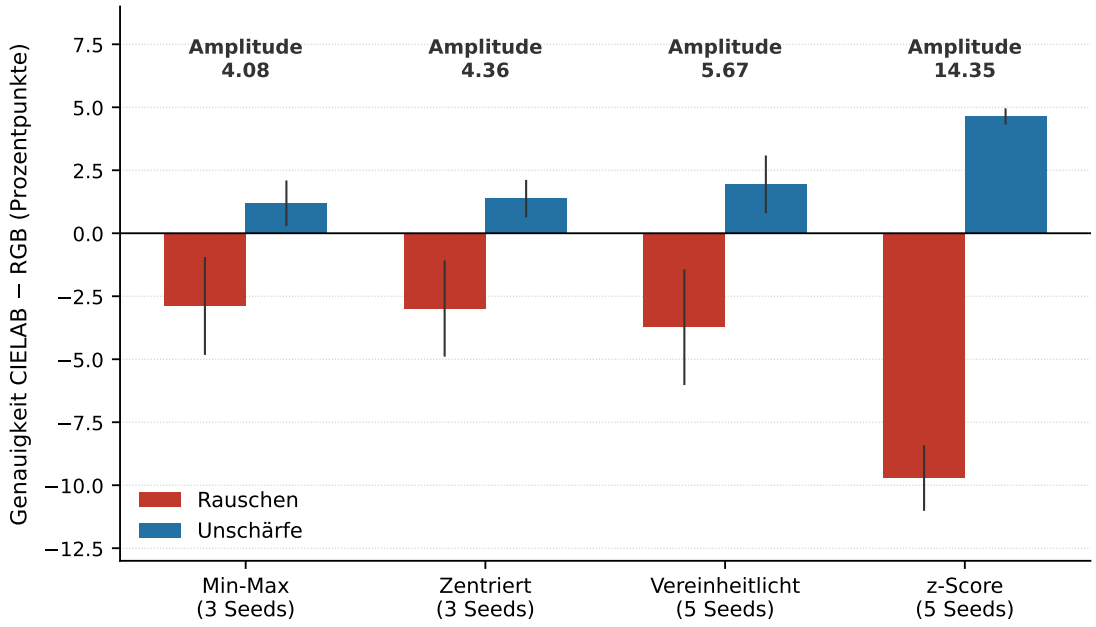


Abbildung 7: Differenz CIELAB − RGB auf Rausch- und Unschärfestörungen unter den vier Normalisierungsschemata (Fehlerbalken: Standardabweichung über die Startwerte). Allein das z-Score-Schema skaliert die Kanäle einzeln; nur dort erreicht die Amplitude des Trade-offs mehr als sechs Prozentpunkte.

übrigen Normalisierungsschemata dar, die die Kanäle ebenfalls nicht einzeln skalieren und dieselbe Größenordnung erreichen. Beide Einzeleffekte schrumpfen dabei in gleichem Maße: der Verlust auf Rauschstörungen von 9,72 auf 3,73 Prozentpunkte, der Gewinn auf Unschärfestörungen von 4,64 auf 1,94. Ihre Richtung und ihre Konsistenz über die Startwerte bleiben unverändert (0/5 beziehungsweise 5/5, $p < 0{,}05$).

Die Kontrollbedingung bestätigt, dass der Eingriff wirkt wie beabsichtigt: Für RGB verändert sich die Robustheit um lediglich $-0{,}54$ Prozentpunkte. Der Eingriff betrifft also im Wesentlichen denjenigen Farbraum, dessen Kanäle ungleich gestreut sind.

Damit ist die Ursache des Trade-offs weitgehend geklärt. Er entsteht nicht allein aus der Farbraumtransformation, sondern überwiegend aus der kanalweisen Standardisierung, die auf diese Transformation folgt. Der verbleibende Anteil von rund 5,7 Prozentpunkten ist vermutlich der Kubikwurzel der CIELAB-Transformation zuzuschreiben (Abschnitt 4.1), die bereits vor jeder Normalisierung wirkt und daher in beiden Bedingungen gleichermaßen vorhanden ist.

6.4.3 Eine Beobachtung zur Haupthypothese

Ein Nebenergebnis dieses Experiments betrifft die Arbeitshypothese selbst. Unter vereinheitlichter Skalierung liegt CIELAB auf den helligkeitsbezogenen Störungen um 0,91 Prozentpunkte vor RGB, und zwar bei allen fünf Startwerten ($d = 1{,}09$;

Einzelwerte +0,43, +0,29, +2,12, +1,46 und +0,27). Unter z-Score-Standardisierung beträgt derselbe Abstand lediglich 0,21 Prozentpunkte bei drei von fünf Startwerten.

Dieser Wert verfehlt die Signifikanzschwelle bereits ohne Korrektur für multiples Testen ($p = 0{,}071$) und wird daher nicht als Ergebnis, sondern als Beobachtung berichtet. Er ist gleichwohl bemerkenswert, weil er in der Richtung liegt, die die Arbeitshypothese vorhersagt, und weil er über alle Startwerte hinweg dasselbe Vorzeichen aufweist.

Eine mögliche Deutung lautet, dass die übermäßige Spreizung der Chromakanäle CIELAB benachteiligt und einen geringen tatsächlichen Vorteil auf helligkeitsbezogenen Störungen verdeckt. Der in Abschnitt 6.2 berichtete Nullbefund gilt demnach zunächst für die dort verwendete, in der Literatur übliche Standardisierung; ob er unabhängig davon Bestand hat, lässt sich mit fünf Startwerten nicht entscheiden. Abschnitt 8.2 greift diesen Punkt auf.

6.5 HSV im Vergleich zu RGB und CIELAB

Neben der Haupthypothese wurde HSV als dritter Farbraum untersucht. Anders als CIELAB trennt HSV Helligkeit und Farbe nicht durch eine luminanzgewichtete Summe, sondern über das Maximum der drei RGB-Werte (Gleichung 1). Ob diese abweichende Konstruktion zu einem anderen Robustheitsverhalten führt, ist eine naheliegende Anschlussfrage.

Die Ergebnisse aus Tabelle 6 geben dazu keinen Anlass. Über die fünfzehn Standardstörungen liegt HSV um 0,07 Prozentpunkte vor RGB ($p = 0{,}90$) und um 0,06 Prozentpunkte hinter CIELAB ($p = 0{,}93$). Beide Unterschiede sind kleiner als die Streuung zwischen zwei Läufen derselben Bedingung und damit nicht interpretierbar. Der Befund entspricht dem in Abschnitt 6.2 für CIELAB berichteten: Auch dieser dritte Farbraum unterscheidet sich in der aggregierten Robustheit nicht von der Referenz.

Ein durchgängiges, wenn auch kleines Muster zeigt sich auf den helligkeitsbezogenen Störungen. Dort liegt HSV bei allen fünf Startwerten hinter CIELAB ($-0{,}67$ Prozentpunkte, $p = 0{,}155$, $d = -0{,}78$) und bei zwei von fünf Startwerten vor RGB ($-0{,}47$ Prozentpunkte, $p = 0{,}208$). Die Signifikanzschwelle wird bereits ohne Korrektur für multiples Testen verfehlt; angesichts der vollständigen Konsistenz gegenüber CIELAB wird der Rückstand gleichwohl als Beobachtung festgehalten. Er gewinnt in Abschnitt 6.6 an Bedeutung, wo sich derselbe Rückstand unter veränderter Normalisierungsschicht deutlich verstärkt.

Auf ungestörten Daten liegt HSV hinter beiden übrigen Farbräumen (Abschnitt 6.1). Dieser Rückstand ist mit 0,36 Prozentpunkten gegenüber CIELAB der einzige Farbraum-Unterschied dieser Arbeit, der die Korrektur für multiples Testen übersteht (Abschnitt 6.9). Eine mögliche Ursache liegt in der zyklischen Natur des Farbtonkanals: Die Werte 0 und 1 bezeichnen dieselbe Farbe, was ein Netz, das seine Eingabe als reelle Zahl verarbeitet, nicht erkennen kann. Diese Deutung wurde nicht geprüft und bleibt eine Vermutung.

6.6 Test der BatchNorm-Erklärung: GroupNorm

Die bisherigen Abschnitte belegen, dass der Eingabefarbraum die Robustheit kaum beeinflusst. Eine naheliegende Erklärung lautet, dass die Batch-Normalisierung die Kanäle fortlaufend neu ausrichtet und die Unterschiede zwischen den Eingabeformaten dadurch verdeckt. Diese Erklärung ist prüfbar.

6.6.1 Vorhersage und Versuchsaufbau

Vor Beginn der Rechenläufe wurde folgende Vorhersage schriftlich festgehalten: *Wenn die Batch-Normalisierung die Unterschiede zwischen den Farbräumen verdeckt, muss sich der Abstand zwischen ihnen vergrößern, sobald sie entfernt wird.* Getestet wird sie, indem in allen Modellen die Batch-Normalisierung durch Gruppen-Normalisierung ersetzt und das Training unter sonst identischen Bedingungen wiederholt wird (vgl. Abschnitt 5.3). Als Kennzahl dient der Abstand zwischen dem besten und dem schlechtesten der drei Farbräume, je Startwert berechnet und anschließend gemittelt.

Tabelle 9: Abstand zwischen bestem und schlechtestem Farbraum (RGB, CIELAB, HSV) in Prozentpunkten, je Startwert berechnet und über fünf Startwerte gemittelt. Da der Abstand zwischen Extremwerten je Startwert gebildet wird, ist er konstruktionsbedingt positiv und wächst mit der Streuung zwischen den Läufen; er liegt daher über dem Abstand der Mittelwerte aus Tabelle 6. Interpretierbar ist das Verhältnis der beiden Spalten, nicht der Absolutwert.

Kennzahl	BatchNorm	GroupNorm	Faktor
mCA gesamt	1,01	0,85	0,84
mCA Helligkeit	0,81	**2,49**	**3,07**

6.6.2 Die Vorhersage bestätigt sich nur teilweise

Abbildung 8 stellt beide Kennzahlen einander gegenüber. Über alle fünfzehn Störungen bleibt der Abstand geschlossen; er verringert sich sogar leicht von 1,01 auf 0,85 Prozentpunkte ($p = 0{,}78$). In dieser Hinsicht trifft die Vorhersage nicht zu: Ohne Batch-Normalisierung trennen sich die Farbräume in der Gesamtrobustheit nicht.

Auf den helligkeitsbezogenen Störungen – also dort, wo die ursprüngliche Arbeitshypothese einen Unterschied erwartete – verdreifacht sich der Abstand dagegen von 0,81 auf 2,49 Prozentpunkte. Der Anstieg tritt bei allen fünf Startwerten auf ($p = 0{,}035$).

Aufschlussreich ist, wodurch dieser Anstieg zustande kommt. Er beruht nicht darauf, dass CIELAB seinen vorhergesagten Vorteil endlich zeigen würde: Unter GroupNorm liegt RGB auf den Helligkeitsstörungen um 0,57 Prozentpunkte *vor* CIELAB,

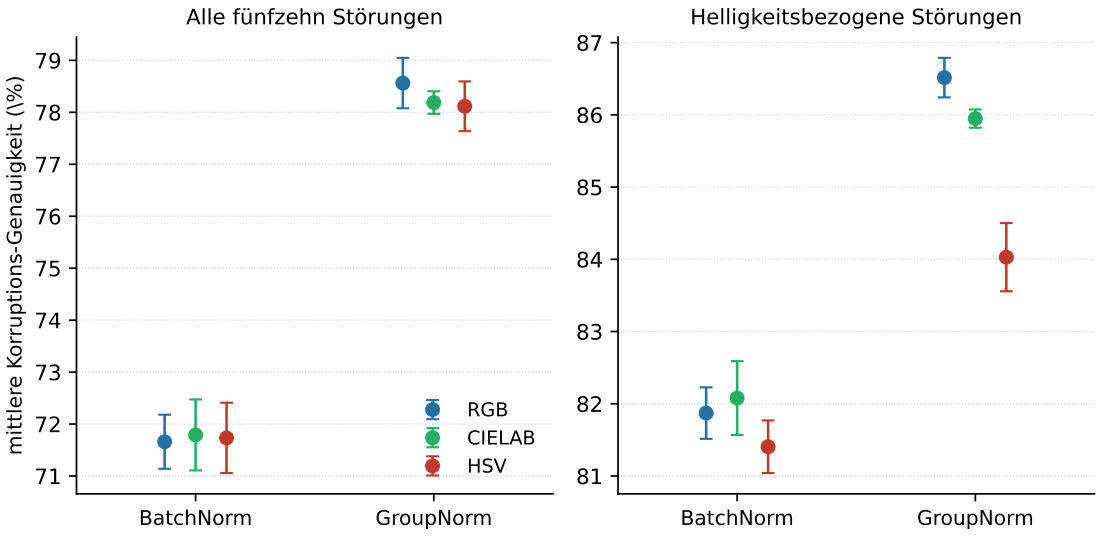


Abbildung 8: Mittlere Korruptions-Genauigkeit der drei Farbräume unter Batch- und Gruppen-Normalisierung (Mittelwert über fünf Zufallsstartwerte, Fehlerbalken: Standardabweichung). Links über alle fünfzehn Störungen, rechts über die helligkeitsbezogenen. Der Abstand zwischen den Farbräumen bleibt links geschlossen und öffnet sich rechts, wobei allein HSV zurückfällt.

bei allen fünf Startwerten ($p = 0{,}018$) – also in der der Hypothese entgegengesetzten Richtung. Der Abstand vergrößert sich vielmehr, weil HSV einbricht: um 2,49 Prozentpunkte hinter RGB und 1,92 hinter CIELAB, jeweils bei null von fünf Startwerten in die andere Richtung ($p \leq 0{,}001$).

6.6.3 Die Normalisierungsschicht wirkt stärker als der Farbraum

Unabhängig vom Vergleich zwischen den Farbräumen zeigt das Experiment einen Effekt, der alle bisher berichteten Unterschiede in den Schatten stellt: Der Wechsel von Batch- zu Gruppen-Normalisierung erhöht die Robustheit *jedes* Farbraums um mehr als sechs Prozentpunkte – RGB um 6,90, CIELAB um 6,40 und HSV um 6,38 Prozentpunkte, jeweils bei allen fünf Startwerten und mit $p < 0{,}001$.

Der Unterschied zwischen den Farbräumen beträgt im selben Aufbau weniger als einen Prozentpunkt. Eine einzige Änderung an der Architektur bewirkt hier also etwa das Zehnfache dessen, was die Wahl des Eingabeformats bewirkt. Dieser Effekt ist in der Literatur zur Gruppen-Normalisierung angelegt (Wu and He, 2018); seine Größenordnung im vorliegenden Vergleich war nicht vorhergesagt.

Der Gewinn ist nicht kostenlos: Auf ungestörten Daten verlieren die GroupNorm-Modelle zwischen 2,28 und 2,92 Prozentpunkte an Genauigkeit. Der Zugewinn an Robustheit übersteigt diesen Verlust jedoch um mehr als das Doppelte.

6.6.4 Zwischenfazit

Die geprüfte Erklärung ist damit weder bestätigt noch vollständig widerlegt. Die Batch-Normalisierung verdeckt einen Unterschied – aber nicht den vorhergesagten: Ohne sie tritt kein CIELAB-Vorteil hervor, sondern eine Schwäche von HSV. Ob die Batch-Normalisierung tatsächlich die vermutete Farbentkopplung leistet, lässt sich an den trainierten Gewichten direkt messen; der folgende Abschnitt tut dies.

6.7 Analyse der gelernten Eingangsfilter

Der vorangegangene Abschnitt hat die Erklärung geprüft, dass die Batch-Normalisierung die Farbräume angleicht, indem ihre Wirkung auf das Ergebnis untersucht wurde. Die zugrunde liegende Annahme – das Netz erlerne eine eigene Farbentkopplung – lässt sich jedoch auch unmittelbar an den trainierten Gewichten überprüfen. Dafür ist kein zusätzliches Training erforderlich.

6.7.1 Vorgehen

Die erste Faltungsschicht hat die Form $[64, 3, 3, 3]$; die zweite Achse entspricht den drei Eingangskanälen. Für jeden der 64 Filter und jede der neun Kernpositionen liegt somit ein Vektor aus $\mathbb{R}^3$ vor, der beschreibt, wie dieser Filter die Eingangskanäle mischt. Über alle 576 Vektoren x_i hinweg wird das zweite Moment

$$M = \frac{1}{N} \sum_i x_i x_i^\top \in \mathbb{R}^{3\times 3} \tag{14}$$

gebildet und in seine Eigenvektoren zerlegt. Diese Formulierung ist vorzeicheninvariant, was notwendig ist, da das Vorzeichen eines Filters ohne Bedeutung ist. Die auf die Summe eins normierten Eigenwerte $\lambda_1 \geq \lambda_2 \geq \lambda_3$ geben an, wie stark das Netz die jeweilige Farbrichtung nutzt. Als zusammenfassendes Maß dient der effektive Rang $1/\sum_k \lambda_k^2$: Er beträgt 3,00, wenn alle drei Farbrichtungen gleich stark genutzt werden, und 1,00, wenn nur eine einzige genutzt wird. Gemittelt wird über die fünf Startwerte.

Als Kontrolle dient erneut die Graustufen-Bedingung. Sie erhält drei identische Eingangskanäle und kann daher höchstens eine Farbrichtung tatsächlich nutzen. Ihr effektiver Rang muss deutlich unter dem der Farbnetze liegen; andernfalls misst das Verfahren nicht das Beabsichtigte.

6.7.2 Die Kontrolle ist erfüllt

Das Graustufen-Netz erreicht einen effektiven Rang von 1,37 und nutzt damit im Wesentlichen eine einzige Farbrichtung, während alle Farbnetze zwischen 2,6 und 2,8 liegen. Aufschlussreich ist dabei nicht nur der niedrige Rang, sondern die Struktur der beiden verbleibenden Eigenwerte: Mit 0,080 und 0,072 sind sie nahezu gleich

Tabelle 10: Genutzte Farbrichtungen in der ersten Faltungsschicht (BatchNorm-Modelle, gemittelt über fünf Startwerte). Zum Vergleich sind in der ersten Zeile die Hauptachsen der Pixelfarben von CIFAR-10 selbst angegeben.

Quelle	λ_1	λ_2	λ_3	effektiver Rang
CIFAR-10, Pixelfarben (Referenz)	0,911	0,074	0,015	1,20
Graustufen-Netz (Kontrolle)	0,848	0,080	0,072	1,37
RGB-Netz	0,501	0,297	0,201	2,63
CIELAB-Netz	0,469	0,290	0,241	2,76
HSV-Netz	0,440	0,333	0,228	2,81

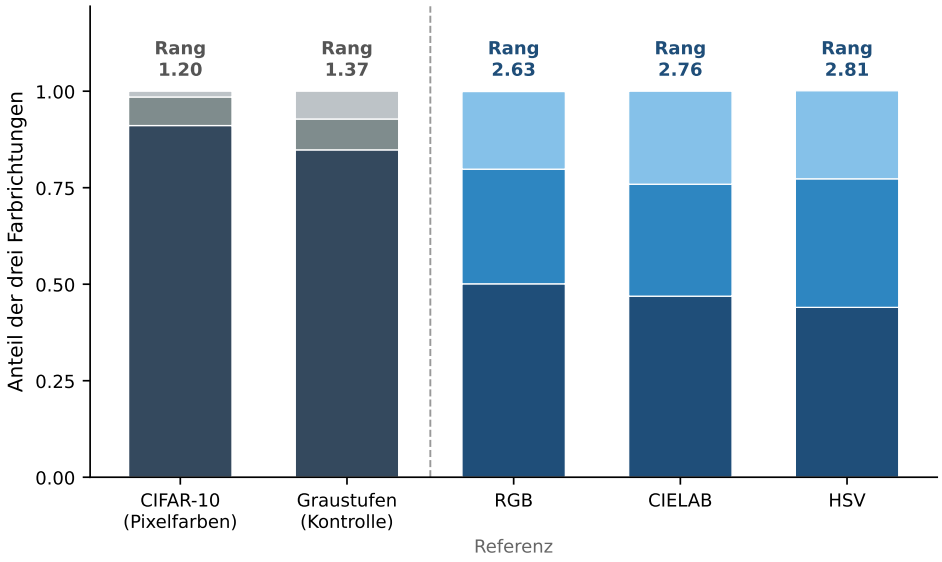


Abbildung 9: Anteil der drei Farbrichtungen an der gesamten Kanalgewichtung, gemittelt über fünf Zufallsstartwerte. Links die beiden Referenzgrößen – die Pixelfarben von CIFAR-10 und das Graustufen-Netz –, rechts die drei Farbnetze. Der über den Balken angegebene effektive Rang beträgt 3,00, wenn alle drei Farbrichtungen gleich stark genutzt werden, und 1,00, wenn nur eine genutzt wird.

groß. Diese Quasi-Entartung bedeutet, dass innerhalb der zur Luminanz orthogonalen Ebene keine Richtung bevorzugt wird – genau das ist zu erwarten, wenn die Eingabe überhaupt keine Farbinformation enthält. Das Verfahren erkennt somit nicht nur die fehlende Farbdimension, sondern gibt auch die erwartete Symmetriestruktur korrekt wieder.

6.7.3 Das Netz entkoppelt die Farbkanäle

Abbildung 9 macht den Kern des Befundes sichtbar. Die Pixelfarben von CIFAR-10 sind stark anisotrop: Die erste Hauptachse trägt 91,1 Prozent, der effektive Rang

beträgt lediglich 1,20. Die gelernten Filter des RGB-Netzes verteilen sich dagegen wesentlich gleichmäßiger auf die drei Farbrichtungen (0,50 / 0,30 / 0,20) und erreichen einen effektiven Rang von 2,63. Das Netz übernimmt die Hauptachsen der Daten also nicht, sondern gleicht sie an: Es verstärkt die chromatischen Richtungen relativ zu ihrem Anteil in den Daten. Genau das leistet eine dekorrelierende Transformation wie CIELAB von Konstruktion wegen.

Die gelernte Zerlegung entspricht zudem inhaltlich der Trennung von Luminanz und Chrominanz. Die erste gelernte Achse des RGB-Netzes lautet [0,58 0,65 0,50] und liegt mit $|\cos| = 0{,}994$ nahezu exakt auf der Luminanzrichtung $(1, 1, 1)$. Die beiden übrigen Achsen stehen mit $|\cos| = 0{,}035$ und 0,104 praktisch senkrecht dazu und spannen damit die chromatische Ebene auf. Ihre genaue Ausrichtung innerhalb dieser Ebene ist nicht bestimmt, da λ_2 und λ_3 nahe beieinander liegen; maßgeblich ist die Trennung zwischen Luminanz und Chrominanz, nicht die Drehung innerhalb der chromatischen Ebene.

6.7.4 Das Ergebnis hängt weder vom Eingabeformat noch von der Normalisierungsschicht ab

Alle drei Farbräume erreichen einen effektiven Rang derselben Größenordnung (2,6 bis 2,8). Ob das Netz eine bereits dekorrelierte Eingabe erhält (CIELAB) oder eine stark korrelierte (RGB), verändert nicht, wie stark es die Farbdimensionen am Ende nutzt.

Unter Gruppen-Normalisierung ergeben sich nahezu identische Werte – RGB 2,64, CIELAB 2,80, HSV 2,76, Graustufen 1,42 – und dieselbe Achsenausrichtung ($|\cos| = 0{,}998$ zur Luminanzrichtung). Die Entkopplung entsteht somit in der gelernten Faltungsschicht selbst und nicht erst durch die Batch-Statistik.

Dieser Befund präzisiert das Ergebnis des vorangegangenen Abschnitts. Die dort geprüfte Erklärung schrieb der Batch-Normalisierung die Farbentkopplung zu; die Messung zeigt, dass diese Zuschreibung nicht zutrifft. Dass ihre Entfernung dennoch den Abstand auf den Helligkeitsstörungen vergrößert, muss folglich einen anderen Grund haben als die Kanalentkopplung. Diese Frage wird in Abschnitt 7.1 aufgegriffen.

6.7.5 Grenzen dieser Analyse

Die Untersuchung betrachtet ausschließlich die erste Faltungsschicht; über tiefere Schichten sind damit keine Aussagen möglich. Das zweite Moment aus Gleichung 14 beschreibt zudem nur die lineare Kanalmischung – nichtlineare Wechselwirkungen, die erst durch die Aktivierungsfunktionen entstehen, werden nicht erfasst. Bei CIELAB und HSV liegen die Achsen im jeweiligen Eingaberaum und sind daher nicht unmittelbar als Luminanz- beziehungsweise Gegenfarbachsen im Sinne von RGB zu lesen; vergleichbar ist zwischen den Farbräumen allein der effektive Rang.

Schließlich entstehen die berichteten Werte durch Mittelung der Matrizen M über die fünf Startwerte, sodass keine Streuung je Startwert vorliegt. Die Unterschiede zwischen den drei Farbräumen (2,63, 2,76 und 2,81) werden daher nicht als Rangfolge interpretiert; belastbar ist allein der Größenordnungsvergleich zwischen Farbnetzen (rund 2,6 bis 2,8), Graustufen-Kontrolle (1,4) und Rohdaten (1,2).

6.8 Wiederholung auf einer zweiten Architektur

Alle bisher berichteten Befunde beruhen auf einer einzigen Architektur. Ob sie Eigenschaften des Eingabeformats oder Eigenheiten des ResNet-18 beschreiben, lässt sich daran nicht entscheiden. Der folgende Abschnitt wiederholt die Untersuchung deshalb auf einem VGG-11.

6.8.1 Vorab festgelegte Vorhersagen

Vor dem Start der Rechenläufe wurden vier Vorhersagen schriftlich festgehalten. Sie sind so formuliert, dass sie scheitern können:

V1 Der Nullbefund bleibt bestehen – CIELAB und RGB unterscheiden sich auf helligkeitsbezogenen Störungen nicht bedeutsam. Begründung: Die Filteranalyse aus Abschnitt 6.7 schreibt die Farbentkopplung der gelernten Faltungsschicht zu, nicht dem ResNet.

V2 Der Trade-off bleibt bestehen, in derselben Richtung und Größenordnung. Begründung: Er entsteht in der Farbraumtransformation und in der Eingabenormalisierung, also *vor* dem Netz, und darf daher nicht von der Architektur abhängen.

V3 Die Negativkontrolle greift weiterhin – Graustufen liegt um mehr als vier Prozentpunkte hinter RGB. Ohne diese Bestätigung ist V1 nicht interpretierbar.

V4 Der Abstand zwischen den Farbräumen auf helligkeitsbezogenen Störungen ist ohne Normalisierungsschicht größer als unter Batch-Normalisierung. Begründung: Ohne sie entfällt die in Abschnitt 6.6 beschriebene Angleichung der Kanalwertebereiche vollständig.

6.8.2 Versuchsaufbau

Trainiert wurde ein VGG-11 in zwei Konfigurationen – einmal mit Batch-Normalisierung und einmal gänzlich ohne Normalisierungsschicht – in allen vier Farbräumen und über dieselben fünf Zufallsstartwerte. Sämtliche übrigen Einstellungen entsprechen Tabelle 3. Die erste Faltungsschicht besitzt in beiden Architekturen die Form $[64, 3, 3, 3]$, sodass die Auswertung unverändert übertragbar ist.

Tabelle 11: Mittlere Korruptions-Genauigkeit auf VGG-11, mit und ohne Normalisierungsschicht (fünfzehn Standardstörungen, fünf Zufallsstartwerte). Die entsprechenden ResNet-18-Werte enthält Tabelle 6.

	VGG-11 mit BatchNorm		VGG-11 ohne Normalisierung	
Farbraum	mCA (%)	mCA Helligkeit (%)	mCA (%)	mCA Helligkeit (%)
RGB	$73{,}95 \pm 0{,}48$	$76{,}50 \pm 0{,}52$	$68{,}20 \pm 0{,}44$	$70{,}00 \pm 0{,}35$
CIELAB	$72{,}03 \pm 0{,}59$	$76{,}80 \pm 0{,}26$	$65{,}91 \pm 0{,}42$	$70{,}04 \pm 0{,}36$
HSV	$73{,}90 \pm 0{,}34$	$75{,}12 \pm 0{,}51$	$67{,}82 \pm 0{,}36$	$68{,}31 \pm 0{,}31$
Graustufen	$67{,}98 \pm 0{,}22$	$74{,}70 \pm 0{,}50$	$60{,}56 \pm 0{,}60$	$67{,}25 \pm 0{,}41$

6.8.3 V1 bis V3 bestätigen sich

Der Nullbefund überträgt sich unverändert. Auf den helligkeitsbezogenen Störungen liegt CIELAB in beiden Konfigurationen praktisch gleichauf mit RGB: um 0,30 Prozentpunkte vorn mit Batch-Normalisierung ($p = 0{,}36$) und um 0,04 Prozentpunkte ohne sie ($p = 0{,}75$). Auf einer zweiten Architektur, mit und ohne Normalisierungsschicht, zeigt sich damit ebenso wenig ein Vorteil wie auf der ersten.

Der Trade-off überträgt sich ebenfalls – und fällt deutlich größer aus. Auf Rauschstörungen verliert CIELAB gegenüber RGB 20,03 Prozentpunkte mit Batch-Normalisierung und 22,48 ohne sie, jeweils bei null von fünf Startwerten in die andere Richtung ($p \leq 0{,}0002$). Auf Unschärfestörungen gewinnt es 4,70 beziehungsweise 5,03 Prozentpunkte, jeweils bei allen fünf Startwerten. Die Amplitude des Trade-offs beträgt damit 24,73 und 27,51 Prozentpunkte gegenüber 14,35 auf dem ResNet-18. Richtung und Konsistenz sind identisch, das Ausmaß ist es nicht: Die Architektur bestimmt mit, wie stark sich der Trade-off auswirkt, nicht aber, ob er auftritt.

Die Negativkontrolle greift in beiden Konfigurationen. Graustufen liegt um 5,97 beziehungsweise 7,64 Prozentpunkte hinter RGB, jeweils bei allen fünf Startwerten und mit $p < 0{,}001$.

6.8.4 V4 bestätigt sich nicht

Die vierte Vorhersage trifft nicht zu. Der Abstand zwischen den drei Farbräumen auf helligkeitsbezogenen Störungen beträgt ohne Normalisierungsschicht 1,81 Prozentpunkte und mit Batch-Normalisierung 1,79 Prozentpunkte – die beiden Werte sind nicht unterscheidbar. Das vollständige Entfernen der Normalisierungsschicht öffnet den Abstand also nicht.

Damit ist die in Abschnitt 6.6 geprüfte Erklärung in ihrer allgemeinen Form nicht haltbar. Wäre die Angleichung der Kanalwertebereiche ursächlich, müsste ihr vollständiges Entfallen den größten Effekt zeigen; es zeigt gar keinen. Der in Abschnitt 6.6 berichtete Anstieg von 0,81 auf 2,49 Prozentpunkte ist demnach nicht dem *Fehlen* einer datensatzweiten Angleichung zuzuschreiben, sondern einer Eigenschaft der Gruppen-Normalisierung selbst.

Aufschlussreich ist dabei der Verlauf des HSV-Rückstands, der den Abstand in allen Konfigurationen bestimmt:

Tabelle 12: Rückstand von HSV gegenüber RGB auf helligkeitsbezogenen Störungen in allen vier untersuchten Konfigurationen (Prozentpunkte, fünf Zufallsstartwerte).

Konfiguration	HSV − RGB	p
ResNet-18, BatchNorm	−0,47	0,208
VGG-11, BatchNorm	−1,38	0,029
VGG-11, ohne Normalisierung	−1,69	0,001
ResNet-18, GroupNorm	**−2,49**	0,001

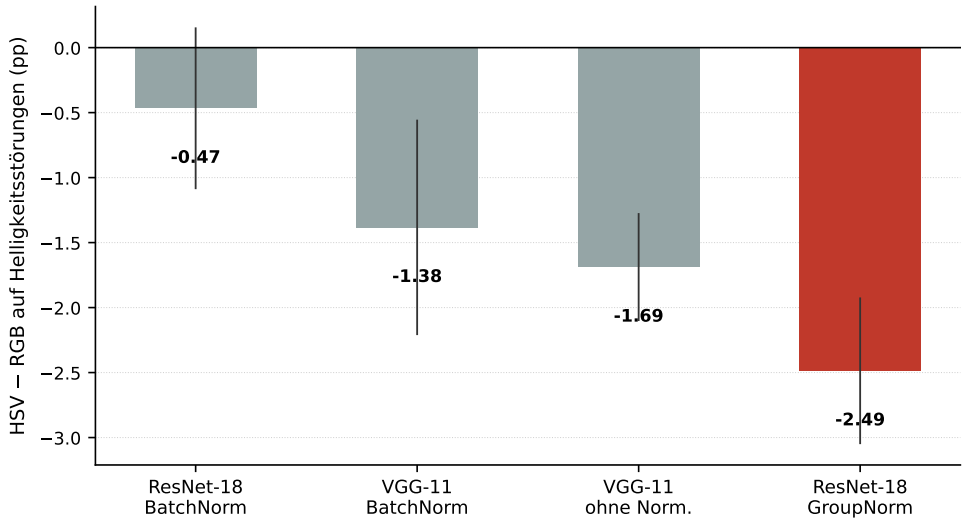


Abbildung 10: Rückstand von HSV gegenüber RGB auf helligkeitsbezogenen Störungen in den vier untersuchten Konfigurationen (fünf Zufallsstartwerte, Fehlerbalken: Standardabweichung). Der Rückstand besteht überall; hervorgehoben ist die Konfiguration mit bildweiser Normalisierung, in der er am größten ausfällt.

Der Rückstand besteht in allen vier Konfigurationen und ist damit weder an eine Architektur noch an eine bestimmte Normalisierungsschicht gebunden. Am stärksten fällt er jedoch unter Gruppen-Normalisierung aus – also gerade dort, wo jedes Bild anhand seiner eigenen Statistik normalisiert wird. Diese Beobachtung führt zu einer anderen Deutung als der ursprünglich geprüften; sie wird in Abschnitt 7.1 aufgegriffen.

6.8.5 Ein Nebenbefund zur Normalisierungsschicht

Unabhängig vom Farbraumvergleich bestätigt sich der in Abschnitt 6.6 berichtete Größenvergleich auf der zweiten Architektur. Der Verzicht auf jede Normalisierungsschicht kostet das VGG-11 zwischen 5,75 und 6,12 Prozentpunkte Robustheit,

je nach Farbraum und mit $p < 0{,}001$ – eine Größenordnung, die dem Gewinn von 6,4 bis 6,9 Prozentpunkten beim Wechsel von Batch- zu Gruppen-Normalisierung auf dem ResNet-18 entspricht. Ob eine Normalisierungsschicht vorhanden ist und welche, wirkt sich damit über beide Architekturen hinweg um etwa das Zehnfache stärker aus als die Wahl des Eingabefarbraums.

6.9 Robustheits-Checks

Dieser Abschnitt fasst zusammen, welchen Prüfungen die berichteten Befunde standhalten. Wie in Abschnitt 5.5 festgelegt, gilt ein Befund erst dann als belastbar, wenn er alle diese Prüfungen übersteht.

6.9.1 Korrektur für multiples Testen

Die Arbeit vergleicht drei Farbräume paarweise auf drei Kennzahlen, was neun Tests ergibt. Nach Holm-Bonferroni-Korrektur über diese neun Vergleiche überdauert ein einziger die Signifikanzschwelle: der Rückstand von HSV gegenüber CIELAB auf ungestörten Daten ($-0{,}36$ Prozentpunkte, $p_{\text{Holm}} = 0{,}032$). Alle übrigen acht Vergleiche erreichen einen korrigierten p-Wert von 1,000 – darunter sämtliche Vergleiche zur Robustheit und insbesondere der für die Arbeitshypothese maßgebliche zwischen CIELAB und RGB auf den helligkeitsbezogenen Störungen.

Für den in Abschnitt 6.3 berichteten Trade-off wurde die Korrektur getrennt über die fünfzehn Vergleiche je Störungsart durchgeführt. Hier überstehen zehn der fünfzehn die Korrektur, darunter beide Extreme: *gaussian noise* ($-12{,}37$ Prozentpunkte, $p_{\text{Holm}} = 0{,}0011$) und *glass blur* ($+7{,}29$ Prozentpunkte, $p_{\text{Holm}} = 0{,}017$) sowie alle vier Unschärfestörungen. Nicht signifikant bleiben *impulse noise*, *pixelate*, *brightness*, *contrast* und *jpeg compression*.

Das Ergebnis der Korrektur ist damit doppelt aufschlussreich: Sie entkräftet sämtliche aggregierten Farbraum-Unterschiede und lässt zugleich den Trade-off nach Störungsart nahezu unangetastet bestehen.

6.9.2 Die zurückgehaltenen Validierungsstörungen

Die vier in Abschnitt 5.4 genannten zusätzlichen Störungsarten wurden getrennt ausgewertet. Sie bestätigen den Trade-off auf Daten, die nicht in die Hauptkennzahl eingehen: *speckle noise* folgt mit $-10{,}38$ Prozentpunkten dem Muster der Rauschstörungen, *gaussian blur* mit $+5{,}68$ Prozentpunkten dem der Unschärfestörungen.

Bemerkenswert ist die Folge für die aggregierte Kennzahl. Über diese vier Störungen gemittelt liegt CIELAB um 0,74 Prozentpunkte *hinter* RGB, und dieser Unterschied ist signifikant ($p = 0{,}0011$); über die fünfzehn Standardstörungen gemittelt liegt CIELAB um 0,13 Prozentpunkte *vorn* und der Unterschied ist es nicht ($p = 0{,}41$). Die Rangfolge kehrt sich also um, je nachdem welche Störungen in die Mittelung eingehen.

Bei einem gleichmäßig wirkenden Farbraum wäre das nicht möglich. Die Empfindlichkeit des Aggregats gegenüber seiner Zusammensetzung ist damit ein dritter, von der Aufschlüsselung in Abschnitt 6.3 unabhängiger Beleg für den Trade-off. Sie unterstreicht zugleich, warum eine aggregierte Robustheitskennzahl ohne Aufschlüsselung nach Störungsart irreführend sein kann.

6.9.3 Variation der Eingabenormalisierung

Der Vergleich zwischen CIELAB und RGB wurde unter allen vier Normalisierungsschemata aus Abschnitt 5.2 wiederholt. Auf den helligkeitsbezogenen Störungen ergibt sich unter z-Score-Normalisierung +0,21 Prozentpunkte ($p = 0{,}31$), unter zentrierter Normalisierung +0,21 Prozentpunkte ($p = 0{,}38$) und unter Min-Max-Normalisierung −0,39 Prozentpunkte ($p = 0{,}54$). Unter keinem dieser drei Schemata zeigt sich der vorhergesagte Vorteil; unter einem von ihnen kehrt sich das Vorzeichen sogar um.

Das vierte Schema, die vereinheitlichte Skalierung, bildet die Ausnahme: Dort liegt CIELAB bei allen fünf Startwerten vor RGB, ohne das Signifikanzniveau zu erreichen (Abschnitt 6.4). Der Nullbefund gilt somit für die kanalweise Standardisierung und die ihr nahestehenden Schemata, nicht notwendigerweise für jede denkbare Vorverarbeitung.

6.9.4 Negativkontrolle

Die Graustufen-Bedingung dient in allen Auswertungen als Negativkontrolle. Sie verliert gegenüber RGB 1,72 Prozentpunkte auf ungestörten Daten und 6,97 Prozentpunkte auf den fünfzehn Standardstörungen, jeweils bei allen fünf Startwerten und mit $p < 0{,}001$. In der Analyse der Eingangsfilter (Abschnitt 6.7) erreicht sie zudem den erwarteten niedrigen effektiven Rang von 1,37 bei quasi-entarteter chromatischer Ebene.

Diese Kontrolle ist für die Bewertung der Nullbefunde entscheidend. Sie belegt, dass das Versuchsdesign in der Lage ist, einen tatsächlich vorhandenen Unterschied im Farbgehalt der Eingabe zu messen – und zwar in einer Größenordnung, die die zwischen den Farbräumen gemessenen Unterschiede um mehr als das Fünfzigfache übersteigt. Das Ausbleiben eines Effekts zwischen RGB, CIELAB und HSV lässt sich folglich nicht auf eine mangelnde Empfindlichkeit des Verfahrens zurückführen.

7 Diskussion

7.1 Der Farbraum verändert die Störungsgeometrie, nicht das Lernen

Die Ergebnisse ergeben zunächst ein widersprüchlich erscheinendes Bild. Der Eingabefarbraum ist für die aggregierte Leistung praktisch bedeutungslos (Abschnitte 6.1

und 6.2), erzeugt aber je nach Störungsart Unterschiede von bis zu zwölf Prozentpunkten (Abschnitt 6.3). Dieser Abschnitt schlägt eine Erklärung vor, die beide Beobachtungen zusammenführt.

7.1.1 Warum eine dekorrelierte Eingabe keinen Vorteil bringt

Die Arbeitshypothese ging davon aus, dass ein Netz von einer bereits dekorrelierten Eingabe profitiert, weil es die Trennung von Helligkeit und Farbe nicht selbst herstellen muss. Die Analyse der Eingangsfilter (Abschnitt 6.7) zeigt, dass diese Annahme die Ausgangslage falsch beschreibt: Das Netz stellt die Trennung ohnehin her.

Die Pixelfarben von CIFAR-10 sind stark anisotrop – eine einzige Richtung trägt 91 Prozent der Varianz, der effektive Rang beträgt 1,20. Die gelernten Filter eines aus RGB trainierten Netzes nutzen dagegen alle drei Farbrichtungen weitgehend gleichmäßig (effektiver Rang 2,63), und ihre Hauptachse liegt mit $|\cos| = 0{,}994$ auf der Luminanzrichtung. Das Netz konstruiert in seiner ersten Schicht also genau jene Luminanz-Chrominanz-Zerlegung, die CIELAB von vornherein bereitstellt.

Damit lässt sich der Nullbefund erklären, ohne einen Mangel des Verfahrens annehmen zu müssen. Eine dekorrelierte Eingabe erspart dem Netz eine Leistung, die es ohne erkennbare Kosten selbst erbringt. Der theoretische Vorteil existiert, er ist jedoch bereits eingelöst, bevor er wirksam werden könnte. Diese Deutung steht im Einklang mit Gowda and Yuan (2019), wo einzelne Farbräume ebenfalls kaum voneinander abweichen und ein Gewinn erst durch die gleichzeitige Verwendung mehrerer Farbräume entsteht.

7.1.2 Wo der Farbraum dennoch wirkt

Ein Farbraum bestimmt nicht nur, wie ein Bild kodiert ist, sondern auch, wie eine *Störung* dieses Bildes aussieht. Und diese zweite Wirkung kann das Netz nicht ausgleichen: Sie tritt ein, bevor die erste Faltungsschicht das Bild erreicht.

Für die Rauschstörungen wurde dieser Zusammenhang bereits in Abschnitt 6.3 beschrieben. Die Kubikwurzel der CIELAB-Transformation besitzt bei kleinen Argumenten eine steile Ableitung – 4,52 bei einer relativen Luminanz von 0,02 gegenüber 0,53 bei 0,5. Eine additive Störung, die in RGB überall gleich groß ist, erscheint in CIELAB in dunklen Bildbereichen um ein Vielfaches verstärkt.

Hinzu kommt ein zweiter, in derselben Richtung wirkender Umstand. Die Chromakanäle von CIELAB besitzen eine sehr geringe Streuung (0,0398 und 0,0631 gegenüber 0,2426 für die Helligkeit, siehe Tabelle 2). Die kanalweise Standardisierung teilt jeden Kanal durch seine Standardabweichung und verstärkt die Chromakanäle dadurch um das etwa Vier- bis Sechsfache dessen, was ein RGB-Kanal erfährt. Was in diesen Kanälen ankommt – Signal wie Störung – erreicht das Netz entsprechend verstärkt.

Diese Verstärkung erklärt beide Richtungen des Trade-offs. Additives Rauschen ist hochfrequent und verteilt sich auf alle Kanäle; in den gespreizten Chromakanälen

wird es überproportional groß. Unschärfe wirkt dagegen als Tiefpass: Sie zerstört die hochfrequente Textur, lässt aber die niederfrequente Farbstruktur weitgehend bestehen. Ein Netz, dem die chromatische Information verstärkt präsentiert wird, verfügt bei zerstörter Textur noch über verwertbares Signal. Dieselbe Eigenschaft, die CIELAB gegenüber Rauschen anfällig macht, kommt ihm bei Unschärfe zugute.

Diese Aufteilung nach Frequenzbereichen ist für CIFAR-10-C unabhängig belegt. Yin et al. (2019) zeigen, dass sich die Störungen dieses Benchmarks im Fourier-Spektrum trennen lassen – Rauschen liegt überwiegend im hohen, Unschärfe- und Wetterstörungen im niedrigen Frequenzbereich – und dass Eingriffe, die die Robustheit im einen Bereich erhöhen, sie im anderen verringern. Der hier berichtete Trade-off ist somit kein Sonderfall von CIELAB, sondern eine Ausprägung dieses allgemeineren Musters. Neu ist der Mechanismus: Er entsteht nicht durch eine Änderung des Trainingsverfahrens, sondern durch die kanalweise Skalierung der Eingabenormalisierung.

7.1.3 Prüfung der Erklärung

Diese Deutung ist prüfbar, und Abschnitt 6.4 berichtet die zugehörige Messung. Wird ausschließlich die Spreizung der Chromakanäle aufgehoben – bei unveränderter Zentrierung –, so verliert der Trade-off rund sechzig Prozent seiner Amplitude, während die Kontrollbedingung RGB nahezu unberührt bleibt. Die beiden übrigen Normalisierungsschemata, die die Kanäle ebenfalls nicht einzeln skalieren, bestätigen denselben Befund unabhängig davon.

Für die Beweisführung ist dabei ein Detail entscheidend: Min-Max- und zentrierte Normalisierung unterscheiden sich untereinander im Zentrieren, stimmen im Ergebnis jedoch überein. Die Verschiebung der Kanäle scheidet damit als Ursache aus, und es bleibt allein ihre Skalierung.

Die vorgeschlagene Erklärung trägt demnach den größeren Teil des Trade-offs. Der verbleibende Anteil ist vermutlich der Kubikwurzel der CIELAB-Transformation zuzuschreiben (Gleichung 4), die bereits vor jeder Normalisierung wirkt und daher in allen Bedingungen gleichermaßen vorhanden ist. Beide Mechanismen wirken in dieselbe Richtung, unterscheiden sich jedoch in ihrem Gewicht.

Bemerkenswert ist die Stelle, an der dieser Effekt entsteht. Nicht die Farbraumtransformation erzeugt ihn überwiegend, sondern die Normalisierung, die auf sie folgt – ein Verarbeitungsschritt, den Farbraumvergleiche üblicherweise als neutralen Zwischenschritt behandeln.

7.1.4 Die Rolle der Normalisierungsschicht

Aus dem bisher Gesagten ergibt sich eine Frage, die Abschnitt 6.7 offengelassen hat. Wenn die Farbentkopplung in der Faltungsschicht entsteht und nicht in der Batch-Statistik – warum vergrößert dann der Ersatz der Batch- durch eine Gruppen-

Normalisierung den Abstand zwischen den Farbräumen auf den Helligkeitsstörungen um das Dreifache?

Die naheliegende Erklärung wäre, dass die Batch-Normalisierung die Wertebereiche der Kanäle datensatzweit angleicht und dadurch verdeckt, was ein Farbraum an ihnen verändert. Diese Erklärung wurde in Abschnitt 6.8 geprüft und trifft nicht zu. Wäre die Angleichung ursächlich, müsste ihr vollständiges Entfallen den größten Effekt zeigen; tatsächlich ist der Abstand in einem Netz ohne jede Normalisierungsschicht mit 1,81 Prozentpunkten praktisch derselbe wie mit Batch-Normalisierung (1,79).

Die Ursache liegt demnach nicht im *Fehlen* einer Angleichung, sondern in einer Eigenschaft der Gruppen-Normalisierung selbst. Aufschlussreich ist dabei, dass der Abstand in allen Konfigurationen von derselben Bedingung getragen wird: dem Rückstand von HSV auf helligkeitsbezogenen Störungen. Er beträgt $-0{,}47$ Prozentpunkte unter Batch-Normalisierung im ResNet-18, $-1{,}38$ und $-1{,}69$ in den beiden VGG-Konfigurationen und $-2{,}49$ unter Gruppen-Normalisierung (Tabelle 12).

Der Rückstand besteht somit unabhängig von Architektur und Normalisierungsschicht; verstärkt wird er allein dort, wo jedes Bild anhand seiner *eigenen* Statistik normalisiert wird. Dafür bietet sich folgende Deutung an. Der V-Kanal von HSV ist als Maximum der drei RGB-Werte definiert und verschiebt sich unter einer Helligkeitsänderung unmittelbar und gleichgerichtet über das gesamte Bild. Eine bildweise Normalisierung nimmt diese Verschiebung in ihre Statistik auf und rechnet sie mit heraus – und entfernt damit zugleich einen Teil des Signals, den ein anderer Farbraum erhalten hätte. Diese Deutung wurde nicht eigens geprüft und bleibt eine Vermutung; ein möglicher Test ist in Abschnitt 8.2 skizziert.

Festzuhalten bleibt der belastbare Teil: Die Batch-Normalisierung dekorreliert die Kanäle nicht, wie Huang et al. (2018) festhalten und die Filteranalyse bestätigt. Sie gleicht deren Wertebereiche an – was jedoch, wie das Experiment ohne Normalisierungsschicht zeigt, für den Abstand zwischen den Farbräumen nicht maßgeblich ist.

7.1.5 Zusammenfassung

Die Befunde lassen sich in einem Satz zusammenfassen: Der Eingabefarbraum verändert kaum, was ein Netz lernt, wohl aber, wie eine Störung aussieht, bevor das Netz sie verarbeitet. Was gelernt wird, stellt das Netz selbst her und ist deshalb gegenüber dem Eingabeformat weitgehend unempfindlich. Wie eine Störung im Eingaberaum geometrisch wirkt, liegt dagegen außerhalb seines Zugriffs – und genau dort entstehen die gemessenen Unterschiede.

Diese Wirkung entsteht dabei überwiegend nicht in der Farbraumtransformation selbst, sondern in der Normalisierung, die auf sie folgt. Das ist insofern bemerkenswert, als die Vorverarbeitung in Farbraumvergleichen üblicherweise als neutraler Zwischenschritt behandelt wird.

Für die Praxis folgt daraus, dass die Wahl des Eingabefarbraums kein wirksamer Hebel für Robustheit ist. Ob eine Normalisierungsschicht vorhanden ist und welche, bewirkt in beiden untersuchten Architekturen mehr als das Zehnfache (Abschnitte 6.6 und 6.8).

7.2 Einordnung in die Literatur

7.2.1 Der Nullbefund ist keine Anomalie

Dass sich einzelne Farbräume kaum unterscheiden, steht im Einklang mit den vorliegenden empirischen Vergleichen. Gowda and Yuan (2019) berichten für CIFAR-10 Unterschiede von ein bis zwei Prozentpunkten zwischen sieben Farbräumen; ein nennenswerter Gewinn entsteht dort erst, wenn mehrere Farbräume gleichzeitig verwendet und ihre Ausgaben zusammengeführt werden. Die vorliegende Arbeit untersucht ausschließlich einzeln verwendete Farbräume und findet in derselben Größenordnung ebenfalls keinen belastbaren Unterschied.

Dieser Punkt verdient Beachtung, weil der Titel von Gowda and Yuan (2019) zunächst das Gegenteil nahelegt. Die dort berichteten Gewinne stammen aus der Fusion, nicht aus der Wahl eines einzelnen Farbraums – die Ergebnisse widersprechen einander also nicht, sondern betreffen unterschiedliche Fragestellungen.

In dieselbe Richtung weist Chiu et al. (2022): Die Farbabhängigkeit neuronaler Netze lässt sich weder durch Modellgröße noch durch Architektur oder gängige Augmentierungsverfahren nennenswert beheben. Das Eingabeformat als Stellschraube für farbbezogene Robustheit erscheint vor diesem Hintergrund von vornherein wenig aussichtsreich.

7.2.2 Die Filteranalyse bestätigt einen bekannten Befund und erweitert ihn

Der in Abschnitt 6.7 berichtete Befund – ein aus RGB trainiertes Netz konstruiert in seiner ersten Schicht eine Luminanz-Chrominanz-Zerlegung – ist der Literatur nicht fremd. Harris et al. (2020) zeigen, dass trainierte CNN in ihren frühen Schichten spontan eine gegenfarbartige Kodierung entwickeln, wie sie CIELAB von Konstruktion wegen bereitstellt.

Die vorliegende Arbeit ergänzt diesen Befund um den für die Fragestellung entscheidenden Vergleich: Der effektive Rang der genutzten Farbrichtungen ist nahezu gleich, unabhängig davon, ob die Eingabe bereits dekorreliert ist (CIELAB) oder nicht (RGB). Damit ist nicht nur gezeigt, *dass* ein Netz eine solche Zerlegung erlernt, sondern auch, dass diese Fähigkeit vom Eingabeformat unabhängig ist – was den Nullbefund unmittelbar erklärt.

Bemerkenswert ist zudem, dass die Entkopplung nicht der Normalisierungsschicht zuzuschreiben ist. Zwar wurde die Batch-Normalisierung von Ioffe and Szegedy

(2015) ausdrücklich als recheneffiziente Näherung eines Whitening-Verfahrens eingeführt; Huang et al. (2018) weisen jedoch darauf hin, dass sie die Kanäle lediglich einzeln zentriert und skaliert, ohne sie zu dekorrelieren, und schlagen aus diesem Grund ein tatsächlich whitenendes Verfahren vor. Die vorliegende Messung fügt sich in diese Einschätzung: Der effektive Rang bleibt unter Gruppen-Normalisierung nahezu unverändert (2,63 gegenüber 2,64). Die Dekorrelation entsteht in der gelernten Faltungsschicht, nicht in der Normalisierungsstatistik.

7.2.3 Eine Lücke: Robustheit wird aggregiert berichtet

Die aufgeschlüsselte Betrachtung nach Störungsart (Abschnitt 6.3) findet in der gesichteten Literatur keine Entsprechung. Farbraumvergleiche werden dort über aggregierte Kennzahlen geführt. Angesichts eines Trade-offs von rund fünfzehn Prozentpunkten Amplitude, der sich im Mittelwert auf 0,13 Prozentpunkte reduziert, ist das eine relevante Einschränkung: Ein systematischer Effekt kann in aggregierten Vergleichen unsichtbar bleiben, ohne dass dies bemerkt wird. Dass eine Aufschlüsselung nach Frequenzcharakteristik der Störung aufschlussreich ist, zeigen Yin et al. (2019) für andere Eingriffe; für den Farbraum als Eingriff steht sie bislang aus.

Dass die Rangfolge zweier Farbräume davon abhängt, welche Störungsarten in die Mittelung eingehen (Abschnitt 6.9), betrifft dabei nicht nur die vorliegende Arbeit, sondern jede Studie, die Farbräume anhand einer einzelnen aggregierten Robustheitskennzahl vergleicht.

7.2.4 Der Bezug zum Form-Textur-Bias

Die Frage, warum die Wahl des Eingabeformats so wenig bewirkt, lässt sich mit Geirhos et al. (2019) weiter einordnen. Dort wird gezeigt, dass auf ImageNet trainierte CNN vorrangig anhand von Textur und nicht anhand von Form klassifizieren, und dass ein durch Training erzwungener Formbias die Robustheit gegenüber einem breiten Spektrum von Bildstörungen verbessert. Kamann et al. (2020) erreichen einen vergleichbaren Effekt, indem sie die Verlässlichkeit der Farbinformation gezielt zerstören.

Beiden Ansätzen ist gemeinsam, dass sie in das Trainingsziel eingreifen und das Netz zwingen, sich auf andere Merkmale zu stützen. Eine Farbraumtransformation tut dies gerade nicht: Sie kodiert dieselbe Information anders, lässt aber unberührt, worauf das Netz sich stützt. Vor diesem Hintergrund ist der Nullbefund der vorliegenden Arbeit erwartbar – die Transformation setzt an einer Stelle an, die für den ausschlaggebenden Mechanismus nicht maßgeblich ist.

7.2.5 Die Normalisierungsschicht als der wirksamere Hebel

Der in Abschnitt 6.6 berichtete Zugewinn von sechs bis sieben Prozentpunkten durch Gruppen-Normalisierung ist der Sache nach nicht neu: Wu and He (2018) führen das

Verfahren gerade wegen seiner Unabhängigkeit von Batch-Statistiken ein. Neu ist der Größenvergleich innerhalb eines identischen Aufbaus. Dieselben Modelle, dieselben Daten, dasselbe Evaluationsprotokoll ergeben für die Wahl des Farbraums weniger als einen Prozentpunkt und für die Wahl der Normalisierungsschicht mehr als sechs.

Für die Praxis lässt sich daraus eine klare Priorisierung ableiten: Wer die Robustheit eines Modells gegenüber Bildstörungen erhöhen will, sollte an der Architektur ansetzen und nicht am Eingabeformat.

7.3 Limitationen

Die folgenden Einschränkungen begrenzen die Reichweite der berichteten Ergebnisse. Sie werden jeweils mit der Angabe verbunden, welche Aussagen davon betroffen sind und welche nicht.

7.3.1 Reichweite des Versuchsaufbaus

Sämtliche Ergebnisse beruhen auf einer Bildgröße (32×32) und einem Datensatz (CIFAR-10). Ob die Farbraumunterschiede bei höherer Auflösung, bei Architekturen ohne Faltung oder bei natürlicheren Bildern anders ausfallen, lässt sich aus diesen Daten nicht ableiten. Die Aussage, die Arbeitshypothese bestätige sich nicht, bezieht sich ausdrücklich auf die untersuchte Konfiguration.

Die Beschränkung auf eine einzige Architektur wurde durch die Wiederholung auf einem VGG-11 aufgehoben (Abschnitt 6.8). Beide Architekturen sind jedoch faltungsbasiert und in ihrem Aufbau verwandt; über Modelle ohne Faltung, etwa Vision Transformer, ist damit nichts ausgesagt.

Die verbleibende Einschränkung wiegt für den Nullbefund schwerer als für den Trade-off. Letzterer beruht auf Eigenschaften der Verarbeitung *vor* dem Netz – der Kubikwurzel der CIELAB-Transformation und der Spreizung der Chromakanäle durch die Normalisierung –, was seine Übertragbarkeit auf beide untersuchten Architekturen erklärt. Der Nullbefund dagegen setzt voraus, dass das Netz die Farbentkopplung selbst erlernt; ob das bei grundlegend anderen Architekturen ebenso gilt, bleibt offen.

7.3.2 Hyperparameter und Batchgröße

Die Hyperparameter aus Tabelle 3 wurden einmal festgelegt und für alle Bedingungen unverändert übernommen. Das sichert die Vergleichbarkeit, schließt aber nicht aus, dass einzelne Farbräume unter anderen Einstellungen besser abschnitten. Eine getrennte Optimierung je Farbraum hätte umgekehrt den Vergleich verzerrt und wurde deshalb nicht vorgenommen.

Für den Vergleich der Normalisierungsschichten ist zusätzlich die Batchgröße von 32 zu beachten. Der Vorteil der Gruppen-Normalisierung gegenüber der Batch-Normalisierung ist bekanntermaßen von der Batchgröße abhängig (Wu and He,

2018). Der berichtete Zugewinn von sechs bis sieben Prozentpunkten gilt daher für diese Batchgröße und ist nicht ohne Weiteres auf andere übertragbar.

7.3.3 Status der einzelnen Befunde

Die Befunde dieser Arbeit haben unterschiedliches Gewicht. Der Nullbefund zur Haupthypothese und das GroupNorm-Experiment beruhen auf vorab festgelegten Fragestellungen. Die Aufschlüsselung nach Störungsart und die Analyse der Eingangsfilter ergaben sich dagegen erst bei der Auswertung; sie sind nachträgliche Beobachtungen, auch wenn sie über alle Startwerte konsistent sind und die Korrektur für multiples Testen überstehen.

Die in Abschnitt 7.1 vorgeschlagene Erklärung des Trade-offs beruht auf einem eigens dafür angelegten Experiment: Die vereinheitlichte Skalierung verändert ausschließlich die Spreizung der Kanäle, lässt die Zentrierung unberührt, umfasst fünf Startwerte und führt RGB als Kontrollbedingung mit (Abschnitt 6.4). Ihre Grenze liegt anderswo: Gemessen wurden nur zwei Punkte der Skalierung – der Faktor 6,10 und der Faktor 1,00 –, sodass der Verlauf zwischen ihnen offenbleibt. Eine vollständige Dosis-Wirkungs-Kurve steht aus; Abschnitt 8.2 skizziert sie.

7.3.4 Grenzen der Filteranalyse

Die Analyse in Abschnitt 6.7 betrachtet allein die erste Faltungsschicht und beschreibt dort nur die lineare Kanalmischung. Nichtlineare Wechselwirkungen und die Verarbeitung in tieferen Schichten bleiben unberücksichtigt. Die Aussage lautet daher, dass die Eingangsschicht eine Luminanz-Chrominanz-Zerlegung konstruiert – nicht, dass das gesamte Netz farbraumunabhängig arbeitet. Da die Werte über die Startwerte gemittelt wurden, liegt zudem keine Streuung vor, weshalb die Unterschiede zwischen den Farbräumen nicht als Rangfolge interpretiert werden.

7.3.5 Modellauswahl auf dem Testdatensatz

Wie in Abschnitt 5.3.5 beschrieben, wird die Epoche mit der höchsten Testgenauigkeit ausgewählt. Die absoluten Genauigkeiten sind dadurch leicht optimistisch. Für die berichteten Differenzen ist dies unerheblich, da das Verfahren auf alle Bedingungen in gleicher Weise angewandt wird; für einen Vergleich mit absoluten Werten aus der Literatur ist es zu berücksichtigen.

Der Vorbehalt reicht weiter, als es zunächst scheint: CIFAR-10-C entsteht durch Störung genau derjenigen 10 000 Testbilder, auf denen auch die Epoche ausgewählt wird. Die Robustheitswerte sind damit ebenfalls leicht optimistisch und nicht vollständig unabhängig von der Modellauswahl. Ein eigener Validierungsteil des Trainingsdatensatzes hätte diese Kopplung vermieden; er wurde hier zugunsten eines identischen Protokolls über alle Bedingungen nicht abgetrennt. Da die Kopplung alle Bedingungen in gleichem Maße betrifft, bleiben die berichteten Differenzen davon unberührt.

7.3.6 Störungsrobustheit ist nicht Domänenverallgemeinerung

CIFAR-10-C misst die Robustheit gegenüber synthetisch erzeugten Störungen desselben Datensatzes. Das ist eine kontrollierte, aber enge Form der Verteilungsverschiebung. Eine Verschiebung, wie sie etwa zwischen Aufnahmegeräten, Erhebungsorten oder – im medizinischen Kontext – zwischen Färbeprotokollen verschiedener Kliniken auftritt, ist damit nicht abgebildet. Die Ergebnisse dieser Arbeit sind auf solche Szenarien nicht ohne Weiteres übertragbar; entsprechende Untersuchungen bleiben dem in Abschnitt 8.2 skizzierten Ausblick vorbehalten.

8 Fazit und Ausblick

8.1 Fazit

Diese Arbeit ist der Frage nachgegangen, ob der Eingabefarbraum verändert, wie gut ein Convolutional Neural Network lernt und wie robust es gegenüber Bildstörungen bleibt. Vier Farbräume wurden unter identischen Bedingungen trainiert und über fünf Zufallsstartwerte auf CIFAR-10 sowie den fünfzehn Standardstörungen von CIFAR-10-C ausgewertet.

8.1.1 Die Arbeitshypothese bestätigt sich nicht

Die Hypothese, ein mit CIELAB trainiertes Netz sei robuster gegenüber helligkeitsbezogenen Störungen, hat sich nicht bestätigt. Der gemessene Unterschied gegenüber RGB beträgt 0,21 Prozentpunkte, wechselt zwischen den Startwerten das Vorzeichen und ist kleiner als die Streuung zwischen zwei Läufen derselben Bedingung. Er überdauert keine der angewandten Prüfungen: Er zeigt sich unter keinem der Normalisierungsschemata, nach Korrektur für multiples Testen erreicht kein Robustheitsvergleich zwischen den Farbräumen die Signifikanzschwelle, und auf einer zweiten Architektur – einem VGG-11 mit und ohne Normalisierungsschicht – fällt er mit 0,30 beziehungsweise 0,04 Prozentpunkten ebenso aus.

Dass dieser Nullbefund nicht auf ein unempfindliches Versuchsdesign zurückgeht, belegt die Graustufen-Kontrolle: Dieselbe Auswertung misst dort einen Rückstand von 6,97 Prozentpunkten bei allen fünf Startwerten, und zwar in jeder untersuchten Konfiguration.

Eine Erklärung dafür liefert die Analyse der gelernten Eingangsfilter. Ein aus RGB trainiertes Netz konstruiert in seiner ersten Faltungsschicht selbständig eine Luminanz-Chrominanz-Zerlegung: Seine Hauptachse liegt mit $|\cos| = 0{,}994$ auf der Luminanzrichtung, und der effektive Rang der genutzten Farbrichtungen ist mit 2,63 nahezu identisch mit dem eines aus CIELAB trainierten Netzes (2,76). Eine bereits dekorrelierte Eingabe erspart dem Netz somit eine Leistung, die es ohne erkennbare Kosten selbst erbringt.

Eine Einschränkung ist dabei zu benennen. Die übliche kanalweise Standardisierung benachteiligt CIELAB, indem sie dessen Chromakanäle übermäßig spreizt. Unter einer Normalisierungsvariante, die diese Spreizung aufhebt, liegt CIELAB auf den helligkeitsbezogenen Störungen bei allen fünf Startwerten vor RGB – ohne Signifikanz, aber durchgängig. Der Nullbefund betrifft demnach den Farbraum in Verbindung mit der in der Literatur üblichen Vorverarbeitung; ob er unabhängig davon Bestand hat, ist mit dieser Stichprobengröße nicht zu entscheiden.

8.1.2 Der Farbraum wirkt dennoch – an anderer Stelle

Aus dem Nullbefund im Mittelwert folgt nicht, dass der Farbraum ohne Wirkung bliebe. Die Aufschlüsselung nach Störungsart zeigt zwei gegenläufige Effekte von erheblicher Größe: Gegenüber RGB verliert CIELAB auf Rauschstörungen 9,72 Prozentpunkte und gewinnt auf Unschärfestörungen 4,64 Prozentpunkte, jeweils über alle Startwerte hinweg konsistent. Im Mittelwert heben sich beide nahezu vollständig auf. Auf der zweiten Architektur tritt derselbe Trade-off auf, dort sogar mit annähernd doppelter Amplitude.

Der Nullbefund entsteht demnach nicht durch Wirkungslosigkeit, sondern durch Kompensation. Für die Bewertung aggregierter Robustheitskennzahlen folgt daraus eine allgemeinere Beobachtung: Ein Mittelwert über Störungsarten setzt voraus, dass ein Verfahren auf alle Störungen in dieselbe Richtung wirkt. Trifft das nicht zu, hängt das Ergebnis davon ab, welche Störungen in die Mittelung eingehen – eine Eigenschaft der Kennzahl, nicht des Modells.

Als überwiegende Ursache dieses Trade-offs erweist sich die kanalweise Standardisierung. Wird ausschließlich sie verändert – alle Kanäle mit demselben Faktor skaliert, die Zentrierung unangetastet –, so sinkt die Amplitude des Trade-offs um rund sechzig Prozent, während die Kontrollbedingung RGB nahezu unverändert bleibt. Der Farbraum verändert somit nicht, *was* ein Netz lernt, wohl aber, wie eine Störung aussieht, bevor das Netz sie verarbeitet – und diese Wirkung entsteht überwiegend erst in der Normalisierung, die auf die Farbraumtransformation folgt.

8.1.3 Die Architektur ist der wirksamere Hebel

Der deutlichste Effekt dieser Arbeit betrifft nicht den Farbraum. Der Wechsel von Batch- zu Gruppen-Normalisierung erhöht die Robustheit jedes untersuchten Farbraums um 6,4 bis 6,9 Prozentpunkte; der vollständige Verzicht auf eine Normalisierungsschicht kostet das VGG-11 zwischen 5,75 und 6,12 Prozentpunkte. In beiden Architekturen liegen die Unterschiede zwischen den Farbräumen im selben Aufbau unterhalb eines Prozentpunktes.

Beide Befunde sind unterschiedlich abgesichert. Ob eine Normalisierungsschicht überhaupt vorhanden ist, wirkt sich unabhängig von der gewählten Batchgröße aus; dieser Teil der Aussage steht ohne Vorbehalt. Der Vorsprung der Gruppen- gegenüber der Batch-Normalisierung ist dagegen bekanntermaßen von der Batchgröße abhängig (Wu and He, 2018) und gilt hier für die verwendete Größe von 32 (Abschnitt 7.3).

Eine einzige Architekturentscheidung bewirkt damit etwa das Zehnfache dessen, was die Wahl des Eingabeformats bewirkt – und dieser Größenvergleich gilt über beide untersuchten Architekturen hinweg. Wer die Robustheit eines Modells gegenüber Bildstörungen erhöhen will, sollte an der Architektur ansetzen und nicht am Eingabeformat.

8.2 Ausblick

Aus den Ergebnissen ergeben sich vier Anschlussfragen, die im Folgenden mit dem jeweils vorgeschlagenen Vorgehen und der zugehörigen Vorhersage skizziert werden.

8.2.1 Die vollständige Dosis-Wirkungs-Kurve der Kanalskalierung

Der Einfluss der kanalweisen Skalierung wurde an zwei Punkten bestimmt: bei einer Spreizung der Chromakanäle um den Faktor 6,10 und um den Faktor 1,00 (Abschnitt 6.4). Eine systematische Variation – etwa mit dem 0,25-, 0,5-, 2- und 4-fachen der jeweiligen Streuung – ergäbe eine vollständige Kurve statt zweier Stützpunkte.

Zwei Fragen ließen sich damit beantworten. Erstens, ob der Zusammenhang monoton verläuft, wie es die vorgeschlagene Erklärung nahelegt. Zweitens, wo die Kurve den Wert null erreicht: Der dort verbleibende Trade-off wäre derjenige Anteil, der allein auf die Kubikwurzel der CIELAB-Transformation entfällt und sich durch Normalisierung nicht beeinflussen lässt.

8.2.2 Die Rolle der bildweisen Normalisierung

Der Abstand zwischen den Farbräumen auf helligkeitsbezogenen Störungen wird in allen untersuchten Konfigurationen vom Rückstand des HSV-Farbraums getragen und fällt unter Gruppen-Normalisierung am größten aus (Tabelle 12). Die in Abschnitt 7.1 vorgeschlagene Deutung – eine bildweise Normalisierung rechnet die gleichgerichtete Verschiebung des V-Kanals mit heraus – wurde nicht geprüft.

Ein geeigneter Test bestünde darin, die Gruppengröße der Normalisierung zu variieren. Bei einer Gruppe je Kanal entspricht das Verfahren einer Instanz-Normalisierung, bei einer einzigen Gruppe einer Schicht-Normalisierung über alle Kanäle. Die Vorhersage lautet, dass der HSV-Rückstand mit zunehmender Trennung der Kanäle wächst, da der V-Kanal dann für sich normalisiert wird.

8.2.3 Eine Verlustfunktion auf dem Helligkeitskanal

Eine Farbraumtransformation verändert die Kodierung der Eingabe, lässt aber unberührt, worauf das Netz sich stützt. Arbeiten, die die Robustheit tatsächlich erhöhen, greifen dagegen in das Trainingsziel ein (Geirhos et al., 2019; Kamann et al., 2020). Ein entsprechender Ansatz für den vorliegenden Fall wäre ein zusätzlicher Verlustterm, der die Vorhersagen für ein Bild und für dieselbe Aufnahme mit veränderter

Helligkeit aneinander bindet. Die Vorhersage lautet, dass ein solcher Zwang die Helligkeitsrobustheit stärker verbessert als jede Wahl des Eingabeformats – und dass der Effekt dann nicht mehr vom Farbraum abhängt.

8.2.4 Übertragung auf reale Domänenverschiebungen

CIFAR-10-C bildet synthetisch erzeugte Störungen desselben Datensatzes ab. Die in der Motivation genannte Anwendung – Aufnahmen aus verschiedenen Kliniken mit je eigenen Färbeprotokollen – stellt eine andersartige und komplexere Verschiebung dar. Ein histopathologischer Datensatz mit mehreren Erhebungsorten, etwa Camelyon17 (Bandi et al., 2019), wäre der natürliche Prüffall.

Ob der hier gemessene Trade-off dort in vergleichbarer Form auftritt, ist offen: Eine Färbungsverschiebung ist weder hochfrequent wie Rauschen noch tiefpassartig wie Unschärfe, sondern eine systematische Verschiebung im Farbraum selbst. Die vorliegenden Ergebnisse lassen für diesen Fall keine Vorhersage zu.

A Vollständige Ergebnistabellen

Die folgenden Tabellen enthalten die Genauigkeit je Störungsart, gemittelt über die fünf Schweregrade und über fünf Zufallsstartwerte. Angegeben sind Mittelwert und Standardabweichung in Prozent. Die aggregierten Kennzahlen aus Kapitel 6 ergeben sich als Mittelwerte über die jeweiligen Zeilen dieser Tabellen.

Tabelle 13: Genauigkeit je Störungsart auf CIFAR-10-C, BatchNorm-Modelle (fünf Zufallsstartwerte).

Störungsart	RGB	CIELAB	HSV	Graustufen
gaussian noise	$49{,}77 \pm 2{,}41$	$37{,}40 \pm 2{,}02$	$46{,}97 \pm 2{,}34$	$41{,}08 \pm 0{,}86$
shot noise	$60{,}69 \pm 1{,}88$	$49{,}11 \pm 2{,}00$	$55{,}97 \pm 2{,}05$	$52{,}07 \pm 1{,}02$
impulse noise	$59{,}02 \pm 1{,}00$	$53{,}82 \pm 2{,}92$	$54{,}56 \pm 1{,}37$	$57{,}36 \pm 0{,}65$
defocus blur	$77{,}71 \pm 0{,}63$	$81{,}21 \pm 0{,}69$	$78{,}81 \pm 0{,}86$	$68{,}84 \pm 1{,}15$
glass blur	$51{,}47 \pm 2{,}01$	$58{,}76 \pm 1{,}58$	$55{,}77 \pm 1{,}43$	$40{,}21 \pm 1{,}58$
motion blur	$71{,}42 \pm 1{,}02$	$74{,}85 \pm 0{,}54$	$72{,}39 \pm 1{,}06$	$60{,}20 \pm 1{,}10$
zoom blur	$72{,}15 \pm 0{,}68$	$76{,}48 \pm 0{,}78$	$73{,}50 \pm 0{,}90$	$58{,}61 \pm 2{,}24$
snow	$77{,}65 \pm 0{,}71$	$80{,}31 \pm 0{,}46$	$78{,}71 \pm 0{,}39$	$69{,}68 \pm 0{,}88$
frost	$74{,}50 \pm 1{,}05$	$78{,}42 \pm 1{,}02$	$76{,}40 \pm 0{,}76$	$66{,}36 \pm 0{,}91$
fog	$83{,}73 \pm 0{,}20$	$84{,}72 \pm 0{,}35$	$84{,}17 \pm 0{,}42$	$79{,}71 \pm 0{,}23$
brightness	$90{,}34 \pm 0{,}10$	$90{,}36 \pm 0{,}14$	$89{,}94 \pm 0{,}17$	$88{,}27 \pm 0{,}17$
contrast	$71{,}55 \pm 0{,}98$	$71{,}16 \pm 1{,}13$	$70{,}10 \pm 0{,}90$	$74{,}00 \pm 0{,}60$
elastic transform	$81{,}71 \pm 0{,}21$	$83{,}54 \pm 0{,}13$	$82{,}16 \pm 0{,}22$	$73{,}61 \pm 0{,}75$
pixelate	$73{,}36 \pm 0{,}87$	$77{,}36 \pm 0{,}96$	$75{,}78 \pm 0{,}51$	$62{,}85 \pm 1{,}42$
jpeg compression	$79{,}82 \pm 0{,}28$	$79{,}35 \pm 0{,}28$	$80{,}79 \pm 0{,}26$	$75{,}99 \pm 0{,}31$

Tabelle 14: Genauigkeit je Störungsart auf CIFAR-10-C, GroupNorm-Modelle (fünf Zufallsstartwerte).

Störungsart	RGB	CIELAB	HSV	Graustufen
gaussian noise	67,43 ± 1,87	55,25 ± 1,61	64,37 ± 1,12	48,85 ± 1,74
shot noise	72,44 ± 1,46	63,42 ± 1,38	69,28 ± 0,78	57,36 ± 1,30
impulse noise	70,28 ± 1,19	66,54 ± 0,76	66,55 ± 2,02	65,67 ± 0,61
defocus blur	83,12 ± 0,24	84,99 ± 0,35	82,80 ± 0,61	78,26 ± 0,52
glass blur	63,15 ± 0,96	70,67 ± 0,54	69,85 ± 1,08	45,86 ± 0,77
motion blur	79,49 ± 0,59	81,08 ± 0,35	79,22 ± 0,87	71,56 ± 0,46
zoom blur	79,69 ± 0,45	82,60 ± 0,38	79,93 ± 0,86	72,30 ± 0,77
snow	80,61 ± 0,58	81,92 ± 0,23	80,87 ± 0,49	73,17 ± 0,47
frost	79,60 ± 0,92	81,31 ± 0,61	80,17 ± 0,35	71,00 ± 0,61
fog	84,11 ± 0,35	83,92 ± 0,19	81,05 ± 0,52	79,26 ± 0,30
brightness	88,59 ± 0,14	88,00 ± 0,20	87,26 ± 0,27	86,96 ± 0,34
contrast	86,85 ± 0,50	85,92 ± 0,24	83,78 ± 0,80	87,27 ± 0,33
elastic transform	82,40 ± 0,44	83,88 ± 0,12	82,21 ± 0,50	76,01 ± 0,39
pixelate	78,90 ± 1,15	82,97 ± 0,91	82,80 ± 0,95	67,05 ± 0,52
jpeg compression	81,76 ± 0,32	80,34 ± 0,41	81,59 ± 0,34	76,83 ± 0,16

Tabelle 15: Genauigkeit auf den vier zusätzlichen Störungsarten von CIFAR-10-C, BatchNorm-Modelle (fünf Zufallsstartwerte). Diese Störungen gehen nicht in die Hauptkennzahl ein; ihre Auswertung wird in Abschnitt 6.9 berichtet.

Störungsart	RGB	CIELAB	HSV
speckle noise	63,33 ± 1,64	52,96 ± 1,99	58,98 ± 1,95
gaussian blur	66,83 ± 1,08	72,51 ± 1,04	69,24 ± 1,16
spatter	82,52 ± 0,47	84,45 ± 0,41	83,99 ± 0,52
saturate	87,58 ± 0,27	87,39 ± 0,22	89,01 ± 0,20

Tabelle 16: Genauigkeit je Störungsart auf CIFAR-10-C, VGG-11 mit Batch-Normalisierung (fünf Zufallsstartwerte).

Störungsart	RGB	CIELAB	HSV	Graustufen
gaussian noise	$57{,}08 \pm 2{,}01$	$33{,}70 \pm 3{,}31$	$53{,}87 \pm 2{,}82$	$49{,}00 \pm 0{,}76$
shot noise	$66{,}62 \pm 1{,}47$	$44{,}51 \pm 2{,}98$	$61{,}13 \pm 1{,}70$	$59{,}18 \pm 0{,}32$
impulse noise	$56{,}67 \pm 1{,}62$	$42{,}08 \pm 1{,}82$	$52{,}85 \pm 1{,}51$	$55{,}35 \pm 0{,}63$
defocus blur	$79{,}90 \pm 0{,}77$	$83{,}18 \pm 0{,}17$	$81{,}74 \pm 0{,}62$	$73{,}39 \pm 0{,}62$
glass blur	$61{,}74 \pm 1{,}14$	$71{,}47 \pm 0{,}68$	$68{,}77 \pm 1{,}10$	$48{,}34 \pm 1{,}38$
motion blur	$75{,}06 \pm 0{,}87$	$77{,}32 \pm 0{,}34$	$76{,}09 \pm 0{,}61$	$65{,}87 \pm 0{,}68$
zoom blur	$76{,}74 \pm 0{,}79$	$80{,}27 \pm 0{,}27$	$78{,}62 \pm 0{,}65$	$68{,}22 \pm 0{,}83$
snow	$79{,}29 \pm 0{,}51$	$82{,}87 \pm 0{,}36$	$80{,}63 \pm 0{,}32$	$72{,}50 \pm 0{,}23$
frost	$77{,}81 \pm 0{,}55$	$82{,}42 \pm 0{,}23$	$79{,}33 \pm 0{,}52$	$71{,}47 \pm 0{,}59$
fog	$78{,}48 \pm 0{,}41$	$79{,}71 \pm 0{,}56$	$77{,}52 \pm 0{,}65$	$75{,}13 \pm 0{,}51$
brightness	$88{,}45 \pm 0{,}35$	$88{,}16 \pm 0{,}22$	$87{,}56 \pm 0{,}32$	$85{,}87 \pm 0{,}17$
contrast	$62{,}58 \pm 1{,}20$	$62{,}53 \pm 0{,}26$	$60{,}27 \pm 0{,}86$	$63{,}11 \pm 0{,}99$
elastic transform	$81{,}58 \pm 0{,}29$	$83{,}12 \pm 0{,}13$	$81{,}88 \pm 0{,}33$	$75{,}37 \pm 0{,}43$
pixelate	$83{,}78 \pm 0{,}21$	$86{,}96 \pm 0{,}24$	$85{,}67 \pm 0{,}47$	$75{,}11 \pm 0{,}29$
jpeg compression	$83{,}50 \pm 0{,}16$	$82{,}17 \pm 0{,}35$	$82{,}62 \pm 0{,}24$	$81{,}82 \pm 0{,}17$

Tabelle 17: Genauigkeit je Störungsart auf CIFAR-10-C, VGG-11 ohne Normalisierungsschicht (fünf Zufallsstartwerte).

Störungsart	RGB	CIELAB	HSV	Graustufen
gaussian noise	$54{,}24 \pm 2{,}56$	$28{,}36 \pm 3{,}38$	$48{,}13 \pm 1{,}19$	$38{,}70 \pm 1{,}95$
shot noise	$62{,}30 \pm 2{,}29$	$37{,}78 \pm 2{,}59$	$55{,}56 \pm 0{,}79$	$48{,}56 \pm 1{,}52$
impulse noise	$54{,}41 \pm 0{,}88$	$37{,}36 \pm 1{,}80$	$47{,}87 \pm 1{,}70$	$48{,}34 \pm 1{,}65$
defocus blur	$73{,}87 \pm 0{,}64$	$76{,}76 \pm 0{,}06$	$75{,}72 \pm 0{,}69$	$66{,}18 \pm 0{,}72$
glass blur	$54{,}64 \pm 1{,}85$	$65{,}60 \pm 0{,}66$	$62{,}45 \pm 1{,}03$	$42{,}73 \pm 1{,}08$
motion blur	$68{,}14 \pm 0{,}93$	$70{,}39 \pm 0{,}24$	$69{,}07 \pm 0{,}80$	$58{,}52 \pm 0{,}97$
zoom blur	$69{,}49 \pm 0{,}90$	$73{,}50 \pm 0{,}08$	$71{,}87 \pm 0{,}91$	$59{,}83 \pm 0{,}98$
snow	$72{,}06 \pm 0{,}31$	$76{,}56 \pm 0{,}15$	$73{,}98 \pm 0{,}52$	$64{,}78 \pm 0{,}75$
frost	$72{,}45 \pm 0{,}48$	$76{,}70 \pm 0{,}28$	$73{,}92 \pm 0{,}50$	$64{,}86 \pm 0{,}99$
fog	$72{,}12 \pm 0{,}53$	$72{,}74 \pm 0{,}56$	$70{,}55 \pm 0{,}59$	$67{,}32 \pm 0{,}39$
brightness	$83{,}11 \pm 0{,}24$	$82{,}74 \pm 0{,}19$	$82{,}58 \pm 0{,}22$	$80{,}69 \pm 0{,}34$
contrast	$54{,}76 \pm 0{,}61$	$54{,}64 \pm 0{,}57$	$51{,}80 \pm 0{,}72$	$53{,}75 \pm 1{,}11$
elastic transform	$75{,}46 \pm 0{,}36$	$77{,}29 \pm 0{,}20$	$75{,}89 \pm 0{,}34$	$68{,}69 \pm 0{,}46$
pixelate	$76{,}63 \pm 0{,}84$	$80{,}73 \pm 0{,}28$	$79{,}49 \pm 0{,}88$	$68{,}41 \pm 1{,}06$
jpeg compression	$79{,}33 \pm 0{,}16$	$77{,}49 \pm 0{,}26$	$78{,}43 \pm 0{,}21$	$77{,}03 \pm 0{,}33$

Tabelle 18: Genauigkeit je Störungsart auf CIFAR-10-C unter vereinheitlichter Kanalskalierung, ResNet-18 mit Batch-Normalisierung (fünf Zufallsstartwerte).

Störungsart	RGB	CIELAB
gaussian noise	48,61 ± 2,19	40,99 ± 1,41
shot noise	59,86 ± 2,05	53,12 ± 1,04
impulse noise	58,36 ± 1,03	61,54 ± 1,12
defocus blur	77,16 ± 0,79	79,17 ± 1,49
glass blur	52,42 ± 1,83	52,77 ± 1,64
motion blur	70,66 ± 0,67	73,02 ± 1,24
zoom blur	70,99 ± 1,19	74,04 ± 1,79
snow	77,41 ± 0,49	78,31 ± 0,50
frost	73,44 ± 1,38	75,25 ± 0,78
fog	83,23 ± 0,51	84,22 ± 0,58
brightness	90,23 ± 0,04	90,63 ± 0,13
contrast	71,02 ± 0,74	72,37 ± 1,16
elastic transform	81,08 ± 0,55	82,20 ± 0,40
pixelate	72,83 ± 1,27	74,19 ± 0,87
jpeg compression	79,57 ± 0,36	79,16 ± 0,60

B Vorab festgelegte Vorhersagen

Die folgenden Vorhersagen wurden vor dem Start der Rechenläufe zur zweiten Architektur schriftlich festgehalten. Die zugehörigen Ergebnisse berichtet Abschnitt 6.8.

V1 – Der Nullbefund bleibt bestehen.
CIELAB und RGB unterscheiden sich auf $\mathrm{mCA}_{\mathrm{Helligkeit}}$ nicht bedeutsam. Begründung: Die Filteranalyse zeigt, dass das Netz die Luminanz-Chrominanz-Zerlegung selbst konstruiert; das ist keine Eigenschaft des ResNet.

V2 – Der Trade-off bleibt bestehen.
CIELAB verliert auf Rauschstörungen und gewinnt auf Unschärfestörungen, in derselben Richtung und derselben Größenordnung. Begründung: Der Trade-off entsteht in der Farbraumtransformation und in der Eingabenormalisierung, also vor dem Netz. Er darf nicht von der Architektur abhängen; verschwindet er, ist die Erklärung aus Abschnitt 7.1 widerlegt.

V3 – Die Negativkontrolle greift weiterhin.
Graustufen liegt um mehr als vier Prozentpunkte hinter RGB auf mCA. Ohne diese Bestätigung ist V1 nicht interpretierbar.

V4 – Der Abstand vergrößert sich ohne Normalisierungsschicht.
Der Abstand zwischen den Farbräumen auf $\mathrm{mCA}_{\mathrm{Helligkeit}}$ ist größer als unter Batch-Normalisierung. Begründung: Ohne Normalisierungsschicht entfällt die Angleichung der Kanalwertebereiche vollständig. Bleibt der Abstand geschlossen, ist die Angleichung als Erklärung ausgeschlossen.

C Software und Konfiguration

Sämtliche Experimente wurden auf einer NVIDIA RTX A5000 mit 24 GB Grafikspeicher durchgeführt. Eine Trainingsepoche benötigt im Mittel 34 Sekunden; ein vollständiger Lauf über 50 Epochen dauert somit rund 29 Minuten. Diese Angabe beruht auf einer Stichprobe von 550 protokollierten Epochen, wobei die Einzelwerte je nach Auslastung des Rechenknotens zwischen 17 und 36 Sekunden schwanken.

Tabelle 19: Versionen der verwendeten Bibliotheken.

Bibliothek	Version
Python	3.11.15
PyTorch	2.6.0
torchvision	0.21.0
timm	1.0.27
scikit-image	0.26.0
OpenCV	4.13.0
NumPy	2.4.3
SciPy	1.17.1

Die vollständige Trainingskonfiguration ist in Tabelle 3 angegeben und für alle Bedingungen identisch – über beide Architekturen, alle Normalisierungsschichten und alle Normalisierungsschemata hinweg. Die verwendeten Normalisierungskonstanten finden sich in Tabelle 2.

Insgesamt wurden die folgenden Konfigurationen trainiert, sofern nicht anders angegeben über fünf Zufallsstartwerte: ResNet-18 mit Batch-Normalisierung und mit Gruppen-Normalisierung in vier Farbräumen, ResNet-18 unter der vereinheitlichten Skalierung in zwei Farbräumen, sowie VGG-11 mit Batch-Normalisierung und ohne Normalisierungsschicht in vier Farbräumen. Die Min-Max- und die zentrierte Normalisierung wurden als Kontrollbedingungen mit drei Startwerten (0, 1 und 42) trainiert (Abschnitt 5.5).

D Reproduzierbarkeit

Vor jedem Lauf werden die Zufallsgeneratoren von `random`, `NumPy` und `PyTorch` einschließlich CUDA auf denselben Wert gesetzt; cuDNN wird in den deterministischen Modus geschaltet und die automatische Algorithmenauswahl deaktiviert. Verwendet wurden die Zufallsstartwerte 0, 1, 2, 3 und 42.

Zu Beginn jedes Laufs wird die vollständige Konfiguration in das Ergebnisverzeichnis geschrieben, sodass jede berichtete Zahl auf die Einstellungen zurückgeführt werden kann, mit denen sie erzeugt wurde. Die Auswertung auf CIFAR-10-C ist deterministisch und ergibt bei wiederholter Ausführung identische Werte.

Ein einzelner Trainingslauf wird über folgenden Aufruf gestartet:

```
python -m scripts.train_model \
    --config configs/exp_lab.yaml --seed 0
```

Die Auswertung auf CIFAR-10-C erfolgt anschließend über `scripts/eval_cifar10c.py`, die Analyse der Eingangsfilter über `scripts/analyze_conv1.py`.

Digitaler Anhang

Der gesamte für diese Arbeit entwickelte Code ist unter der MIT-Lizenz öffentlich zugänglich:

`https://github.com/giressengansop/decol-gen`

Das Repositorium enthält:

- `src/color_dg/` – die Farbraumkonvertierungen, die Datenpipeline und die beiden Modellarchitekturen;
- `configs/` – je eine YAML-Datei pro Versuchsbedingung, aus der sich jeder berichtete Lauf unverändert reproduzieren lässt;
- `scripts/` – Training, Evaluation auf CIFAR-10-C, statistische Auswertung und Filteranalyse;
- `runs/` – die Rohergebnisse aller berichteten Läufe, je Bedingung und Startwert getrennt abgelegt;
- `logs/` – die Trainingsprotokolle, auf denen Abbildung 5 beruht;
- `tests/` – Testfälle für die Farbraumkonvertierungen;
- `PREREGISTRATION_vgg11.md` – die in Anhang B wiedergegebenen Vorhersagen in ihrer ursprünglichen Fassung.

Die trainierten Modellgewichte sind aufgrund ihres Umfangs nicht Teil des Repositoriums; sie lassen sich mit den angegebenen Konfigurationen und Startwerten neu erzeugen. CIFAR-10 wird beim ersten Aufruf automatisch geladen, CIFAR-10-C ist separat zu beziehen; der Ablauf ist in der Datei `README.md` beschrieben.

Literaturverzeichnis

Peter Bandi et al. From detection of individual metastases to classification of lymph node status at the patient level: The camelyon17 challenge. In *IEEE Transactions on Medical Imaging*, volume 38, pages 550–560, 2019.

Ming-Chang Chiu, Yingfei Wang, Derrick Kim, Pin-Yu Chen, and Xuezhe Ma. Colorsense: A study on color vision in machine visual recognition. *arXiv preprint arXiv:2212.08650*, 2022.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations (ICLR)*, 2019.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

Shreyank N. Gowda and Chun Yuan. Colornet: Investigating the importance of color spaces for image classification. In *Computer Vision – ACCV 2018*, volume 11364 of *Lecture Notes in Computer Science*, pages 581–596. Springer, 2019.

Ethan Harris, Daniela Mihai, and Jonathon Hare. How convolutional neural network architecture biases learned opponency and colour tuning. *arXiv preprint arXiv:2010.02634*, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations (ICLR)*, 2019.

Lei Huang, Dawei Yang, Bo Lang, and Jia Deng. Decorrelated batch normalization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 791–800, 2018.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning (ICML)*, pages 448–456, 2015.

Christoph Kamann, Burkhard Güssefeld, Robin Hutmacher, Jan Hendrik Metzen, and Carsten Rother. Increasing the robustness of semantic segmentation models with painting-by-numbers. In *European Conference on Computer Vision (ECCV)*, 2020.

Alexandros Karargyris. Color space transformation network. *arXiv preprint arXiv:1511.01064*, 2015.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521 (7553):436–444, 2015.

Dmytro Mishkin, Nikolay Sergievskiy, and Jiri Matas. Systematic evaluation of convolution neural network advances on the imagenet. *Computer Vision and Image Understanding*, 161:11–19, 2017.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*, 2015.

Twan van Laarhoven. L2 regularization versus batch and weight normalization. *arXiv preprint arXiv:1706.05350*, 2017.

Yuxin Wu and Kaiming He. Group normalization. In *European Conference on Computer Vision (ECCV)*, pages 3–19, 2018.

Dong Yin, Raphael Gontijo Lopes, Jonathon Shlens, Ekin Dogus Cubuk, and Justin Gilmer. A Fourier perspective on model robustness in computer vision. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

Declaration of Authorship

Ich erkläre hiermit gemäß §9 Abs. 12 APO, dass ich die vorstehende Abschlussarbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe. Des Weiteren erkläre ich, dass die digitale Fassung der gedruckten Ausfertigung der Abschlussarbeit ausnahmslos in Inhalt und Wortlaut entspricht und zur Kenntnis genommen wurde, dass diese digitale Fassung einer durch Software unterstützten, anonymisierten Prüfung auf Plagiate unterzogen werden kann.

Place, Date

Signature