



AutoML für Segmentierung

PROJEKTSEMINAR
zum Thema
AUTOML

Westfälische Wilhelms-Universität Münster
Institut für Informatik

Eingereicht von:
Milan Blunk (418650)
Pia Nümann (454700)
Matthias Wolff (458766)

Münster, März 2021

Inhaltsverzeichnis

1 Einführung	2
2 NAS-Net	4
2.1 Funktionsweise / Theorie	4
2.2 Unsere Arbeit / Praxis	7
2.3 Ergebnisse	7
2.4 Fazit	8
3 Auto-DeepLab	11
3.1 Funktionsweise / Theorie	11
3.2 Unsere Arbeit / Praxis	14
3.3 Ergebnisse	15
3.4 Fazit	15
4 nnU-Net	16
4.1 Funktionsweise / Theorie	16
4.2 Unsere Arbeit / Praxis	20
4.2.1 Datensätze aus dem Paper	22
4.2.2 Larven-Datensatz	23
4.2.3 Retina 2D-Datensatz	28
4.2.4 CT-Datensatz	32
4.2.5 Pascal VOC2012	38
4.2.6 Retina 3D-Datensatz	42
4.3 Fazit	44
5 Gesamtfazit	46

1 | Einführung

Diese Ausarbeitung ist Gegenstand des Projektseminars „AutoML“, das im Wintersemester 2020/2021 an der Westfälischen Wilhelms Universität unter der Leitung von Prof. Xiaoyi Jiang und Christof Duhme durchgeführt wurde.

In dem Projektseminar haben sich neun Teilnehmer zu drei Dreiergruppen zusammengefunden. Die Themen der drei Gruppen waren: „AutoML Frameworks“, „AutoML für Segmentierung“ und „NAS ohne Training“.

Wir haben uns mit Thema zwei „AutoML für Segmentierung“ beschäftigt und beschreiben im Folgenden unsere Vorgehensweise, Probleme, die wir hatten, Erfolge, die wir erzielen konnten, und gehen zu jedem Framework, das wir genutzt haben, auf die Theorie zu der Architektursuche ein. Insgesamt haben wir uns mit drei Frameworks beschäftigt: NAS-Unet, Auto-DeepLab und nnU-Net. Die Arbeit ist so aufgebaut, dass wir jedem Framework ein Kapitel widmen, in dem wir erst jeweils die Theorie erklären, anschließend von unserer Vorgehensweise mit dem Framework berichten, dann die Ergebnisse einordnen und bewerten und abschließend ein Fazit ziehen. Im Anschluss an diese Einführung werden wir so also in Kapitel 2 über NAS-Unet berichten, Kapitel 3 befasst sich mit Auto-DeepLab, anschließend gehen wir auf das nnU-Net ein, mit welchem wir die mit Abstand besten Erfolge erzielen konnten, wie wir in Kapitel 4 sehen werden, ehe wir in Kapitel 5 ein Gesamtfazit ziehen.

Zu jedem Framework haben wir das Ziel verfolgt, die Theorie zu verstehen und das Framework anhand verschiedener Datensätze zu testen und zu bewerten. Die Kapitel spiegeln in ihrer Reihenfolge auch die Reihenfolge wieder, in der wir uns mit den verschiedenen Programmen beschäftigt haben. So haben wir mit NAS-Unet angefangen. Zunächst haben wir versucht, auf unseren eigenen Geräten zu arbeiten. Wir mussten jedoch schnell feststellen, dass NAS-Unet auf Windows nicht unterstützt wird und auf MacOS fehlte uns eine Grafikkarte für die Ausführung. Deswegen haben wir an unseren Windows-Geräten eine Linux Distribution installiert. Doch auch so konnten wir nicht zufriedenstellend arbeiten, da die Architektursuche zu aufwendig ist für unsere Geräte.

So haben wir glücklicherweise die Möglichkeit bekommen auf den Universitäts-Servern zu arbeiten. All unser Vorgehen haben wir so auf PALMA II (kurz für „Paralleles Linux-System für Münsteraner Anwender“ [1]) ausgeführt. PALMA II ist ein „Computercluster mit über 18.000 Prozessorkernen, deren Zusammenwirken auch die Lösung komplexester Aufgaben aus Wissenschaft und Forschung ermöglicht.“ [1]. So hatten wir leistungsstarke Hardware zur Verfügung, auf denen wir die komple-

xen Berechnungen ausführen konnten.

Nachdem die technischen Rahmenbedingungen nun auch erklärt sind, sei noch zu erwähnen, dass dieser Bericht ein (fundierte) Grundwissen im Bereich Machine Learning voraussetzt.

2 | NAS-Unet

2.1 Funktionsweise / Theorie

NAS-UNet ist einer der ersten Versuche NAS auf medizinische Bildsegmentierung anzuwenden. Die folgenden Ausführungen beruhen auf dem Paper [2] von Yu Weng et al.

Es sollten MRT-, CT- und Ultraschallbilder segmentiert werden. Die Architektur von NAS-UNet wurde auf Pascal VOC2012 [3] gesucht und diese dann auf den unterschiedlichen medizinischen Datensätzen trainiert. Als medizinische Datensätze werden für die MRT-Bilder der Promise12 Datensatz [4], für die CT-Bilder der Chaos Datensatz [5] und für die Ultraschallbilder der NERVE Datensatz [6] verwendet.

Das vorrangige Ziel von NAS-Unet ist das automatische Finden einer geeigneten Zwei-Zell Architektur. Dabei wird parallel nach der Upsampling und nach der Downsampling Schicht gesucht, die beiden Schichten werden gleichzeitig upgedated. Dabei wird also immer eine Upsampling Zelle gleichzeitig mit der ihr gegenüberliegenden Downsampling Zelle aktualisiert. Die Architektur von NAS-Unet ist streng symmetrisch und es gibt keine zusätzliche Convolutionschicht in der Mitte (siehe Abbildung 2.1).

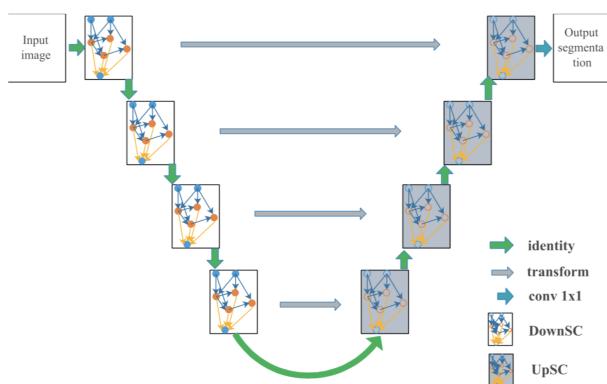


Abbildung 2.1: Zellbasierte Netzarchitektur von NAS-Unet [2]

Der Suchraum, in dem die Architektur gesucht werden soll, enthält die möglichen Architekturen, die prinzipiell verwendet werden können, sowie eine Auswahl von primitiven Operationen. Die möglichen Architekturen sind populäre Unet-Architekturen, von denen nur die mittlere Convolutionschicht entfernt wurde. NAS-Unet verwendet einen zell-basierten Architektursuchraum. Die zell-basierte Architektur (siehe Abbildung 2.1) soll die Generierungsmethode beschränken und so das Problem lösen, dass der Suchraum zu groß wird. Nachdem die beste Zellarchitektur (siehe Abbildung 2.2) gefunden wurde, wird sie im ganzen Netzwerk genutzt und im Rückrad des Netzes gestapelt. Dabei sind nicht nur die Convolutionschichten in die Zellen verlegt, sondern auch alle Up- und Downsampling Operationen. Die Inputknoten einer Schicht sind definiert als die Outputknoten der vorherigen zwei Schichten (siehe Abbildung 2.2). Bei der Auswahl der primitiven Operatoren wurde zum einen darauf geachtet Redundanz zu vermeiden und zum anderen darauf, möglichst wenige Parameter zu haben, um möglichst wenig Memory zu verbrauchen.

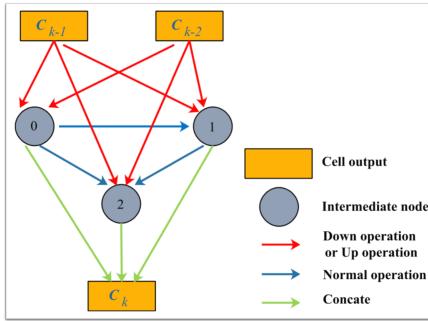


Abbildung 2.2: Zellarchitektur von NAS-Unet [2]

Die Suchstrategie teilt sich in mehrere Schritte auf. Zunächst wird ein überparametrisiertes Netzwerk erstellt (Siehe Abbildung 2.3).

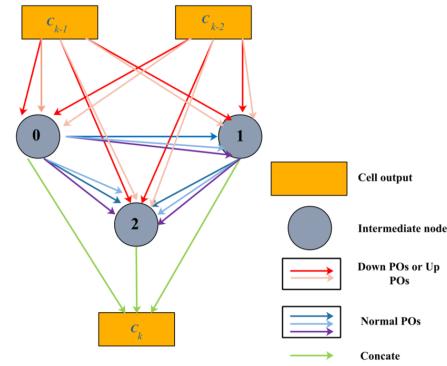


FIGURE 4. An example of Over-Parameterized Cell Architecture, each edge associate with N candidate operations from different primitive operation sets.

Abbildung 2.3: überparametrisierte Zellarchitektur von NAS-Unet [2]

In diesem überparametrisierten Netzwerk lässt sich der Output einer Kante aus der Kombinationen der unterschiedlichen primitiven Operatoren folgendermaßen als Formel darstellen:

$$MixO(x) = \sum_{i=1}^N w_i o_i(x)$$

Dabei ist $o(x)$ die Primitive Operation, w ist das zu der Operation gehörende Gewicht und N ist die Anzahl an primitiven Operationen.

Um die Parameter zu aktualisieren, wird eine effizientere Parameter-Update Strategie verwendet, damit GPU-Memory gespart wird. Da die Output-Feature-Maps nur berechnet werden können, wenn alle Operationen gleichzeitig im GPU-Memory sind, wird das N -fache an GPU-Memory benötigt, als wenn man ein kompaktes Modell trainieren würde. Daher wird hier ein binärer Ansatz verwendet, das heißt anstatt bei jedem Schritt alle Architekturparameter mit dem Gradientenabstiegsverfahren zu aktualisieren, wird immer nur ein Parameter aktualisiert (siehe Abbildung 2.4). Dadurch werden aber mehr Iterationen des Updatens benötigt.

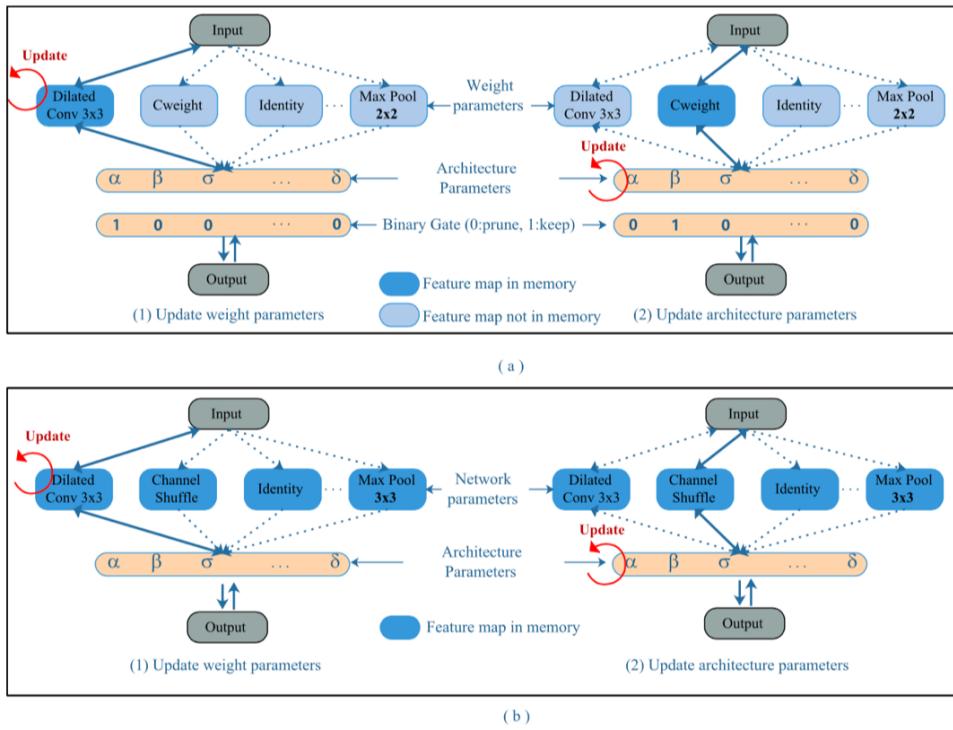


Abbildung 2.4: Vergleich der binären Suchstrategie von NAS-Unet mit dem gleichzeitigen Updaten aller Parameter [2]

Zu den Implementierungsdetails gehört, dass es immer jeweils 4 Up- und Downsampling Zellen gibt. Die Bilder werden zufällig zur einen Hälfte in Trainingsbilder und zur anderen Hälfte in

Testbilder eingeteilt. Das Paper konzentriert sich vorrangig darauf einen effizienten Suchraum zu konstruieren. Die Suchstrategie ist für die Autoren weniger wichtig, da laut ihrer Aussage (Paper, II. B. [2] und Page 44253, Satz2, [2]) jede differentielle Suchstrategie auf dem Suchraum funktionieren würde. In diesem konkreten Fall wird die DARTS-Update Strategie verwendet.

2.2 Unsere Arbeit / Praxis

Bei der praktischen Arbeit mit NAS-Unet sind wir auf verschiedene Schwierigkeiten und Hinderisse gestoßen. Diese belaufen sich vorrangig auf die Problematik, dass es keine Anleitung oder Einführung für Nas-Unet gibt und auch nahezu keine Dokumentation vorliegt. Zunächst einmal haben wir versucht NAS-Unet auf verschiedenen Datensätzen zum Laufen zu bekommen. Dies waren der Datensatz Pascal VOC, der Chaos Datensatz und der Promise Datensatz, welche alle auch im Paper verwendet werden. Bereits hier hatten wie einige Schwierigkeiten, die wir zum Teil auch nicht überwinden konnten.

Als erstes haben wir durch Fehlermeldungen und Suchen im Code herausgefunden, dass NAS-Unet den Datensatz in einem ganz bestimmten Ordner an einem ganz bestimmten Pfad erwartet. Da wir den Pfad auf Palma nicht einrichten konnten, da dieser schon im Wurzelverzeichnis beginnt, mussten wir die Stelle im Code entsprechend anpassen. Die Tatsache, dass der Code fast gar nicht kommentiert wurde, hat uns die Suche und Anpassung erheblich erschwert. Auch die Ordnernamen und die Struktur des Datensatzes mussten angepasst werden. Auch hierzu gab es keinerlei Hinweise oder Dokumentation.

Auf dem Chaos Datensatz, welcher auch im Paper verwendet wird, hatten wir das Problem, dass das Framework im Datensatz nach Bildern sucht, die in keiner öffentlich verfügbaren Version des Datensatzes [5] existieren. Wie das NAS-Unet auf diese Bilder kommt oder wie man verhindert, dass es nach diesen sucht, ist uns unklar geblieben.

Während wir den Promise Datensatz nicht öffentlich zugänglich finden konnten, ließ sich das Framework auf dem Pascal VOC Datensatz erfolgreich ausführen.

2.3 Ergebnisse

Erfolgreich zum Laufen bringen konnten wir das Netz lediglich auf dem Datensatz Pascal VOC2012. Unsere erzielten Ergebnisse waren jedoch leider sehr schlecht. Unsere mIoU auf dem Pascal Datensatz auf den Testbildern war <0.05 (siehe Abbildung 2.5).

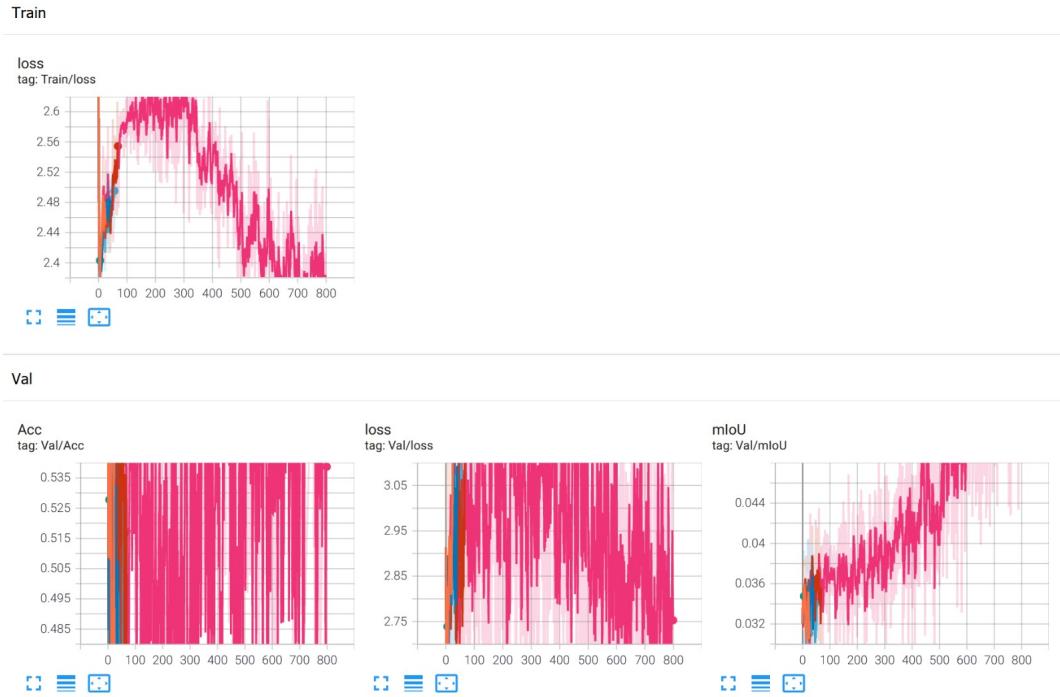


Abbildung 2.5: überparametrisierte Zellarchitektur von NAS-Unet [2]

Wie man sieht, ist der Graphenverlauf sehr schwankend und ergibt nur stark gesmoothed eine sichtbare Tendenz. Die Werte sind schlecht bis sehr schlecht, und verbessern sich auch nur sehr langsam.

Auf dem Chaos-Datensatz haben wir NAS-Unet nicht zum Laufen bekommen, da NAS-Unet nach einem Bild sucht, welches in keiner der öffentlichen Versionen des Datensatzes ([5] unter Download) existiert. Da es uns nicht gelungen ist, dieses Problem zu lösen, haben wir das Framework leider nicht auf dem Chaos-Datensatz zum Laufen bekommen und können daher auch keine Ergebnisse vorstellen. Uns ist auch nach langer Fehlersuche nicht klar geworden, warum das Framework überhaupt nach einem Bild sucht, welches es eigentlich nie eingegeben bekommen hat.

Da wir den Promise Datensatz nicht öffentlich finden konnten, können wir hier leider auch über keine Ergebnisse berichten.

2.4 Fazit

Durch die fehlende Dokumentation des Codes und die fehlende Anleitung zur Nutzung des Frameworks ist es extrem schwierig und zeitaufwändig das Framework zu nutzen. Es war uns leider

auch nicht möglich, den Code vollständig nachzuvollziehen. Wir konnten daher leider nicht nachvollziehen, was genau gemacht wurde. Auch im Internet, zum Beispiel auf Github, konnten wir leider niemanden finden, der den Code nachvollziehen konnte. Man findet auch hier leider nur viele Fragen. Ein Beispiel ist dieser Versuch den Code nachzuvollziehen (siehe Abbildung 2.6) Hier sieht man, dass sich auch die Frage der verschiedenen Metriken auftut. Auch darauf haben wir keine Antwort finden können.

```

@@ -23,6 +23,12 @@ from models import get_segmentation_model
import models.geno_searched as geno_types
from tensorboardX import SummaryWriter

+ '''Questions about the implementation.
+ 1. Why authors asked to run this file to train the network and search the architecture, when nowhere we
+    are using searching?
+ 2. what is the dimension of the nasunet output? whats aux_pred in preds vector?
+ 3. self.test_queue is not even defined anywhere but used in test function
+ 4. Just wondering why are we using two metrics, pixacc andd mIoU?'''

```

Abbildung 2.6: Beispielversuch den Code nachzuvollziehen [7]

Ein weiteres Problem ist das Verhalten von NAS-Unet auf dem Chaos Datensatz. Die Frage, warum es nach einem Bild sucht, welches im Datensatz nicht vorkommt, bleibt offen und damit auch die Frage, wie man NAS-Unet auf diesem Datensatz zum Laufen bekommen könnte. Auffällig ist dies vor allem daher, dass NAS-Unet im Paper auch angeblich auf dem Chaos Datensatz angewendet wird. Hier ist wiederum auffällig, dass das Paper zu NAS-Unet am 04.04.2019 veröffentlicht wurde, während der Datensatz erst am 11.04.2019 veröffentlicht wurde. Möglicherweise hatte die Autoren eine leicht andere Vorversion. Trotzdem erklärt dies nicht, warum das Netz nach mehr Bildern sucht, als ihm eingegeben werden.

Zu den oben genannten Problemen bei der Arbeit mit dem Framework kommt hinzu, dass unsere Ergebnisse auf dem Datensatz Pascal VOC2012 sehr schlecht waren. Besonders auffällig ist dies auf dem Datensatz von Pascal VOC2012, da dieser auch im Paper genutzt wurde und es darauf spezialisiert ist. Leider wurden unserer Recherche nach nie fertig trainierte Modelle von NAS-Unet, welche Ergebnisse wie im Paper angegeben erzielen, veröffentlicht. Daher war es uns leider weder möglich die Ergebnisse zu reproduzieren noch sie nachzuvollziehen.

Im folgenden Github Issue: Issue11 [8] haben wir herausgefunden, dass man die trainierte Netzstruktur anscheinend händisch in den Code zum Trainieren hineinkopieren muss. Per default verwendet NAS-Unet aber die auf Pascal VOC2012 ausgesuchte Netzstruktur. Daher sollten unsere Ergebnisse auf Pascal VOC 2012 eigentlich davon nicht negativ beeinflusst werden.

Auch beim Durchsuchen der Github Issues auf der zugehörigen Github Seite [9] sind wir mehrfach darauf gestoßen, dass die Ergebnisse, die im Paper angegeben wurden, nicht reproduziert werden konnten (zum Beispiel Issue 31 [10]). Da wir auch nicht genau nachvollziehen können, wie diese Ergebnisse entstehen, haben wir uns, auch auf Grund der oben genannten Probleme, dazu entschlossen, nicht länger mit diesem Framework zu arbeiten. Anstelle von NAS-Unet haben wir uns selbstständig ein neues Framework rausgesucht und mit ihm weitergearbeitet (siehe nnU-Net, Kapitel 4).

3 | Auto-DeepLab

3.1 Funktionsweise / Theorie

Auto-DeepLab ist ein NAS Programm, welches im Jahr 2019 zur Segmentierung von Bildern entwickelt wurde. Die folgenden Ausführungen zur Funktionsweise des Programms beruhen auf dem Paper [11] von Chenxi Liu et al.

Auto-DeepLab verfolgt den Ansatz, die Netzstruktur sowie die Zellstruktur des Convolutional Netzes zu suchen. Die Architektur kann mittels Gradient Descent in 3 GPU Tagen gesucht werden. Anders als bei nicht-differenzierbaren Suchtechniken arbeitet Auto-DeepLab somit sehr effektiv: Es wird nicht in einer diskreten Menge von Kandidaten nach der besten Architektur gesucht. Statt dessen wird mit Hilfe von Gradient Descent (also mittels Differenzierbarkeit, also stetig) gearbeitet. Die Optimierung geschieht in Hinblick auf die Validierungsperformance. Diese Effektivität ist auch notwendig, da Bildsegmentierung auf hochauflösenden Bildern funktionieren muss und auch das segmentierte Ausgabebild im besten Fall die gleiche Auflösung wie das zu segmentierende Eingabebild haben sollte.

Zum Zeitpunkt der Veröffentlichung von Auto-DeepLab war der Stand der Entwicklung der NAS Programme so, dass oft nur das Innere der Zellen automatisiert gesucht wurde und dann anhand dieses Resultates das Netz bestimmt wurde. Hier unterscheidet sich Auto-DeepLab, da sowohl Netz als auch die Zellstruktur automatisiert ermittelt werden. Im Folgenden wird die Suche im hierarchischen Suchraum beschrieben. Wir beginnen mit dem Suchraum bezüglich des Zellinneren.

Zunächst einmal definieren wir, was eine Zelle ist:

Eine Zelle ist ein gerichteter, azyklischer Graph, der aus einer geordneten Sequenz aus n Knoten besteht. Jeder Knoten ist hierbei ein sogenannter Block. Mehrere Zellen miteinander verkettet bilden dann das gesamte neuronale Netz.

Die Blöcke in den Zellen sind Strukturen, die zwei Tensoren als Input entgegennehmen und einen Output-Tensor ausgeben. Block i in Zelle l kann mit einem 5-Tupel charakterisiert werden: (I_1, I_2, O_1, O_2, C) mit $I_1, I_2 \in \mathcal{I}_i^l$, wobei \mathcal{I}_i^l die Menge aller möglichen Input-Tensoren darstellt: \mathcal{I}_i^l besteht aus dem Output der vorherigen Zelle, H^{l-1} , dem Output der Zelle zwei Zellen vor der aktuellen Zelle l , H^{l-2} , und jeweils dem Output der Blöcke der aktuellen Zelle, die sich in dem gerichteten, azyklischen Graphen vor dem aktuellen Block i befinden: H_1^l, \dots, H_{i-1}^l . Dadurch haben

die Blöcke einer Zelle, die im gerichteten Graphen weiter hinten sind, mehr mögliche Inputs. Zurück zu dem charakterisierenden 5-Tupel bilden O_1 und O_2 die Layer-Typen jeweils für die beiden Inputs I_1 und I_2 mit $O_1, O_2 \in \mathcal{O}$. Die Menge \mathcal{O} besteht wiederum aus 8 möglichen Operatoren: 3 x 3 depthwise-separable conv, 5 x 5 depthwise-separable conv, 3 x 3 atrous conv with rate 2, 5 x 5 atrous conv with rate 2, 3 x 3 average pooling, 3 x 3 max pooling, skip connection, no connection (zero). Das C stellt den Operator dar, mit dem die beiden auf die Layers angewandten Inputs zu einem Output gemacht werden: $O_1(I_1)$ und $O_2(I_2)$ werden immer einfach elementweise addiert. Damit könnte man die Charakterisierung eigentlich auch auf ein 4-Tupel bestehend aus I_1, I_2, O_1, O_2 beschränken.

Der Output Tensor der Zelle i , H^l , ist einfach die Konkatenation von H_1^l, \dots, H_n^l . D.h. die Outputs der Blöcke der Zelle i werden konkateniert in der Reihenfolge des Auftreten im Graphen.

Nachdem die Struktur der Zellen nun erklärt ist, fahren wir fort mit der Beschreibung, wie sich die Architektur des Netzes zusammenstellt. Dies ist am besten graphisch mit Abbildung 3.1 möglich. Links in Abbildung 3.1 ist ein Gitter dargestellt. Die blauen Punkte des Gitters repräsentieren die Zellen. Die Zweierpotenzen links stehen für die Anzahl an Downsamplings. D.h. eine hohe Zahl (z.B. die 32) steht für eine starke Reduzierung der Bildauflösung. Die Zahlen oben stehen für die Anzahl an Schichten im Netz. Bei Auto-DeepLab gilt $L = 12$, d.h. nach den initialen Zellen zum Anfang des Netzes (grau gefärbt) gibt es 12 weitere Zellen im Netz. Ein Pfad durch das Gitter stellt nun eine mögliche Architektur dar. Von allen Möglichkeiten sucht Auto-DeepLab den Pfad, also die Architektur, die auf den Validierungsdaten am besten performt. Als Verlustfunktion verwendet Auto-DeepLab Cross-Entropy. Ein nächster Schritt in dem Gitter ist entweder immer ein Schritt schräg nach oben (die Auflösung verdoppelt sich), ein Schritt waagerecht (gleiche Auflösung) oder ein Schritt schräg nach unten (die Auflösung halbiert sich). Mit jeder Halbierung der Auflösung verdoppelt sich die Anzahl an Feature Maps.

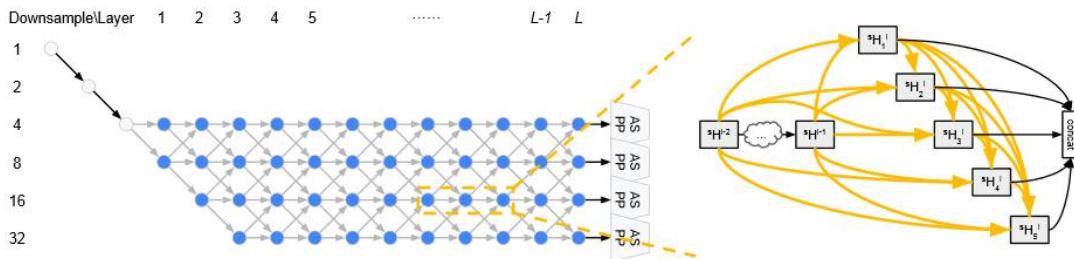


Abbildung 3.1: Auto-DeepLab Architektur [11]

Die drei Schritte (schräg nach oben, waagerecht, schräg nach unten), die durch die hellgrauen Pfeile dargestellt sind, können als Übergangswahrscheinlichkeiten von einem Zustand (einer Zelle) in

den nächsten (nächste Zelle) interpretiert werden. Die drei Skalare sind somit alle nicht-negativ und bilden in der Summe einen Wert von 1. Als beste Architektur wird der Pfad bestimmt, der die Übergangswahrscheinlichkeiten maximiert.

Rechts in Abbildung 3.1 sind insgesamt drei (blaue) Zellen dargestellt: Zelle $l - 2$ (${}^s H^{l-2}$), Zelle $l - 1$ (${}^s H^{l-1}$) und Zelle l (${}^s H^l$). Zelle l ist anders dargestellt als $l - 1$ und $l - 2$: Hier wird das Innere der Zelle, also die Blöcke, dargestellt. Bei Auto-DeepLab besteht jede Zelle aus genau 5 Blöcken. Wie bereits oben beschrieben, bestehen die möglichen Inputs eines Blocks aus den Outputs der zwei Vorgängerzellen und den Outputs der Vorgängerblöcke in der Zelle l . Das hochgestellte s steht für die Stufe an Downsamplings, also $s \in \{4, 8, 16, 32\}$ (es wird bei 4 angefangen, da initial bereits zwei Downsamplings durchgeführt werden).

Zwei beispielhafte, gefundene Architekturen sind in Abbildung 3.2 dargestellt.

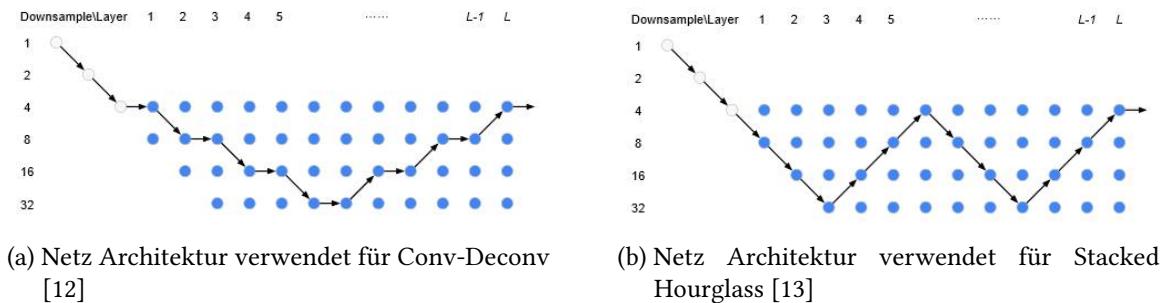


Abbildung 3.2: Architekturen ermittelt durch Auto-DeepLab für zwei verschiedene Bilddatensätze [11]

In Abbildung 3.2a sieht man eine fast U-förmige Architektur. Diese Form werden wir spezifischer und - anders als hier - bewusst forciert bzw. angewandt in Kapitel 4 zum nnU-Net genauer betrachten. Beide hier dargestellten Architekturen enden auf Stufe 4 der Downsamplings (was allerdings keineswegs immer der Fall ist). Damit die Auflösung des Eingabebildes zurückgewonnen werden kann, folgen auf jeder Downsampling Stufe nach der L-ten Schicht die sogenannten Atrous Spatial Pyramid Pooling Module (in Abbildung 3.1 als ASPP abgekürzt). Mit Hilfe dieser Module wird durch Upsampling die ursprüngliche Auflösung wieder gewonnen.

Für die Suche nach der perfekten Architektur nutzt Auto-DeepLab 40 Epochen. Die Batch-Größe beträgt 2. Wie bereits erwähnt, wird mit Gradient Descent gearbeitet. Genauer wird Stochastic Gradient Descent angewandt, um schneller Ergebnisse erzielen zu können. Gestartet wird mit einer Lernrate von 0.025, diese reduziert sich allerdings im Verlauf der Suche immer weiter bis schließlich auf einen Wert von 0.001.

Auffällig ist, dass in den ersten 75% der Layer (also in den ersten 9 Layern) eher Downsampling

dominiert und in den letzten 25% (also letzten 3 Layern) eher Upsampling stattfindet. Zudem ist zu bemerken, dass atrous und depthwise convolution häufig gewählt werden als Operatoren aus O .

3.2 Unsere Arbeit / Praxis

Nachdem wir uns mit der Theorie von Auto-DeepLab auseinander gesetzt hatten, haben wir versucht, das Programm zum Laufen zu bringen. Der Datensatz, den wir zur Verfügung hatten und mit Hilfe von Auto-DeepLab segmentieren wollten, ist ein Datensatz [14], welcher Larven darstellt: Er beinhaltet Graustufen-Bilder von Larven auf einer Glasscheibe, die mittels Frustrated Total Internal Reflection abgelichtet wurden. Ziel war es, die Larven zu segmentieren und die Verschmutzungen um die Larven herum dabei zu ignorieren. Die Larven im Bild sind hell zu sehen. Der Hintergrund ist schwarz und nimmt einen sehr großen Teil der Bilder ein (über 90%). Somit ist der Anteil der zu segmentierenden Objekte, also hier die Larven, sehr klein. Um das Programm auszuführen, haben wir den Link in dem Paper [11] auf Seite 1 unten genutzt, der zu einem GitHub Repository führt. Allerdings ist uns aufgefallen, dass der Link zu dem Repository zu *DeepLab* führt [15], und somit nicht zu einer Veröffentlichung von *Auto-DeepLab*, was uns sehr verwundert hat, nachdem uns das aufgefallen ist. So haben wir also mit DeepLab und nicht, mit Auto-DeepLab gearbeitet. Nach anfänglichen Schwierigkeiten haben wir es letztlich geschafft, das Programm auf dem Larven-Datensatz zum Laufen zu bringen mit Hilfe des Checkpoints, der auf der GitHub Seite zu finden ist. Allerdings wurde jedes Bild ausschließlich als schwarz segmentiert. Da der Hintergrund der Bilder einen sehr großen Teil der Bilder einnimmt, waren die Ergebnisse was den Score angeht nicht gänzlich schlecht. Aber das wirklich brauchbare Ergebnis war natürlich sehr schlecht: Wenn jedes Bild als schwarzes Bild segmentiert wird, existiert kein Mehrwert.

Da aber die Objekte nur einen sehr kleinen Anteil der Bilder einnehmen, haben wir den Ansatz verfolgt, die Bilder in viele kleine Bilder zu zerschneiden (siehe Abbildung 3.3), sodass sich der Anteil der zu segmentierenden Objekte vergrößert. Wir hatten die Hoffnung, dass das Programm die Larven dann erkennt und so in der Lage ist, die Larven zu segmentieren. So haben wir die Bilder zerschnitten zu vielen kleinen Bildern.

Leider hat das aber keine Wirkung gezeigt: Noch immer wurden ausschließlich schwarze Bilder ausgegeben. Die Larven konnten also nicht segmentiert werden.

In den GitHub Issues konnten wir leider keine Lösung unseres Problems finden. Im Gegenteil haben wir dort mehrere Beiträge gefunden, die zeigen, dass scheinbar viele andere Nutzer ebenfalls (ähnliche) Probleme wie wir hatten.

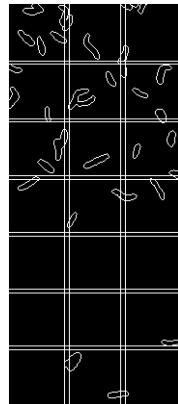


Abbildung 3.3: Zerschnittene Larvenbilder in der Übersicht

Als weiteren Versuch, mit Hilfe von Auto-DeepLab, die Larven zu segmentieren, sind wir auf die GitHub Seite von Noam Rosenberg [16] gestoßen. Hier scheint die vielleicht sogar einzige Implementierung von Auto-DeepLab zu existieren. So haben wir uns mit dieser Umsetzung beschäftigt. Da es uns aber viel zu aufwendig erschien, sich dort weiter hineinzuarbeiten, haben wir den Versuch, mit Hilfe von Auto-DeepLab Bilder zu segmentieren, an dieser Stelle beendet.

3.3 Ergebnisse

Wie bereits in dem vorherigen Unterabschnitt zu dem Praxisteil zu Auto-DeepLab beschrieben, waren unsere Ergebnisse nicht zu gebrauchen. Da jedes Bild ausschließlich als schwarz ausgegeben wurde, sind die Ergebnisse als sehr schlecht zu bewerten, bzw. der Versuch, Ergebnisse zu erzielen, ist gescheitert. Auch der Ansatz, den Anteil der zu segmentierenden Objekte im Bild zu vergrößern durch das Zerschneiden der Bilder, hat nichts gebracht.

Wir vermuten, dass der in GitHub angegebene Checkpoint nicht geeignet ist.

3.4 Fazit

Auto-DeepLab verfolgt einen interessanten Ansatz, Bilder zu segmentieren. In der Theorie wird Auto-DeepLab für die sehr geringe Trainings- / Suchzeit gelobt. Auffällig ist jedoch, dass der Link in dem Paper zu Auto-DeepLab [11] zu einer Umsetzung von DeepLab [15] führt. So haben wir damit gearbeitet, konnten allerdings keine Ergebnisse erzielen und haben bei der Suche nach Lösungen festgestellt, dass viele Andere ebenfalls Probleme hatten, mit dem Framework zu arbeiten.

4 | nnU-Net

4.1 Funktionsweise / Theorie

Das nnU-Net ist ein Framework, welches sich mit der Segmentierung von medizinischen 3D-Aufnahmen mit Hilfe von automatisiertem maschinellem Lernen beschäftigt. Es wurde im Rahmen des Medical Segmentation Decathlon Wettbewerb entwickelt und gewann diesen sowie im Anschluss auch viele weitere Segmentierungs-Wettbewerbe. Die folgenden Ausführungen beruhen auf dem Paper [17] und auf dem Paper [18], beide von Fabian Isensee et al.

Das nnU-Net verwendet eine klassische und nicht neue U-Net Architektur (not new U-Net). Es konzentriert sich kaum auf Architekturdesign und -suche, sondern vorrangig auf die Suche von guten Hyperparametern. Dies sei entscheidend für den Erfolg von nnU-Net (siehe Abbildung 4.5). Es wird eine gleiche oder sehr ähnliche U-Net Struktur immer durch eine individuell auf die individuellen Daten angepasste Trainingspipeline zur Optimierung geschickt (siehe Abbildung 4.1). Das Training der Parameter des Netzes ist also individuell zugeschnitten, während die Architektur sich bei unterschiedlichen Daten nicht oder kaum unterscheidet. Es wird sich also vorrangig auf das Training des Netzes und die Suche individueller Hyperparameter für die Trainingspipeline konzentriert und nicht auf die Suche nach der Architektur.

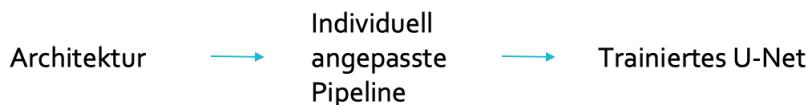


Abbildung 4.1

Das nnU-Net verwendet 3 Standardarchitekturen, welche 2D U-Net, 3D full resolution U-Net und 3D U-Net Cascade sind. Vor dem Training kann man einstellen, wie viele und welche Architekturen man trainieren möchte. Per default trainiert nnU-Net alle und wählt am Ende die beste oder die beste Kombination aus maximal zwei Architekturen aus. 2D U-Net eignet sich besonders für 2D-Daten und läuft gut auf anisotropen Daten. Es arbeitet auf den Bildern in Originalauflösung. 3D full resolution U-Net eignet sich für kleine 3D-Daten und arbeitet auch auf den Bildern in Originalauflösung. Bei größeren Bildern werden jedoch die Patches sehr klein, was zum immer größer werdenden Verlust

von Kontextdaten führt. Daher gibt es das 3D U-Net Cascade für große 3D-Daten. Es besteht aus 2 hintereinander gereihten U-Nets. Das erste U-Net arbeitet auf den Bildernals ganzes, also ohne Aufteilung in Patches, in geringerer Auflösung. Diese grobe Vorsegmentierung wird zusammen mit dem Bild in Originalgröße an das zweite U-Net weitergegeben. Dieses arbeitet dann wieder auf der vollen Bildauflösung und mit Patches und erstellt eine endgültige und verfeinerte Segmentierung. Durch diesen Übergabeschritt zwischen den beiden U-Nets bleiben die Kontextdaten erhalten (siehe Abbildung 4.2).

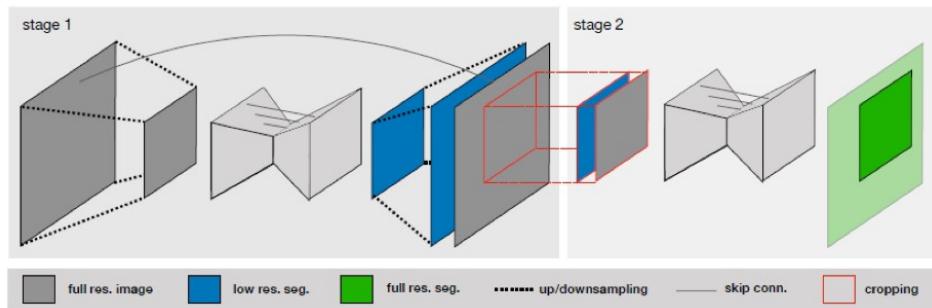


Abbildung 4.2: nnU-Net Cascade [17]

Um die Konzentration auf die Anpassung der Hyperparameter der Trainingspipeline an den individuellen Datensatz zu erreichen, wird zunächst ein Datafingerprint aus den Eigenschaften der Trainingsdaten erstellt (siehe Abbildung 4.3). Die hierbei genutzten Eigenschaften sind unter anderem die Imgasize, das Pixelvolumen oder die Farbkanäle, die spacing Anisotropie sowie die Anzahl der Klassen und deren Häufigkeitsverteilung.

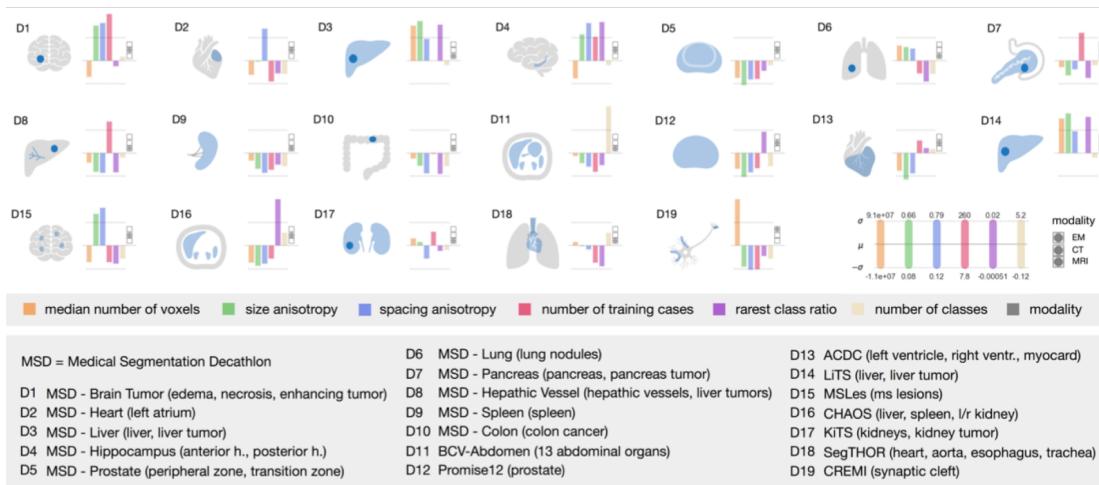


Abbildung 4.3: Datafingerprint [18]

Aus dem Datafingerprint werden, mit Hilfe von heuristischen Regeln, die Inferred Parametern

berechnet. Die Inferred Parameter umfassen die Patch Size, die Batch size, wichtige Parameter zur dynamischen Anpassung der Netzwerktopologie, wie zum Beispiel die Anzahl der Max. Poolings und Downsamplings, sowie Parameter zur Bild Vorverarbeitung.

Die Bestimmung der Patch size erfolgt zunächst initial über den Median der Bildgröße nach dem Resampling. Anschließend wird mit dieser Patch Size die Architektur konfiguriert und geschaut ob ausreichend GPU-Memory zur Verfügung steht. Steht nicht ausreichend GPU-Memory zur Verfügung, so wird die Patch Size reduziert und die Architektur darauf aufbauend neu konfiguriert. Dies wird so oft wiederholt, bis ausreichend GPU-Memory verfügbar ist. Anschließend wird die Batch Size angepasst und das Netzwerk abschließend konfiguriert (Siehe Abbildung 4.4). Dabei muss beachtet werden, dass die Patch Size immer durch 2^i teilbar sein muss (mit $i = \text{Anzahl an DownSampling Operationen}$) da sich die Patch Size pro DownSampling Operation halbiert. Ist das nicht gegeben, so wird die Patch-Size entsprechend vergrößert oder verkleinert bis sie in allen Dimensionen durch 2^i teilbar ist.

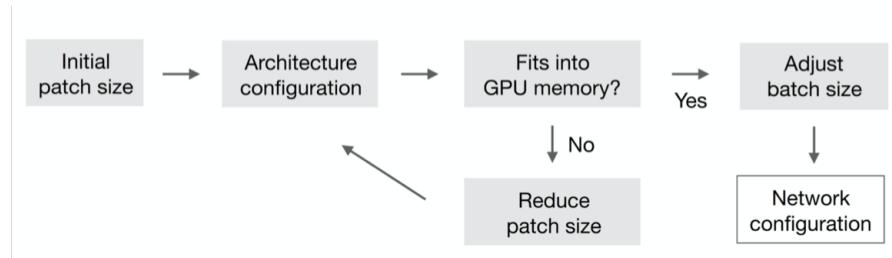


Abbildung 4.4: Patch Size Ermittlung [18]

Im Anschluss an die Inferred Parameter wird der Pipelinefingerprint erstellt, welcher sich aus den Inferred Parametern, den Blueprint Parametern und den empirischen Parametern zusammensetzt. Während die bereits beschriebenen Inferred Parameter für die entscheidende Anpassung an einen neuen Datensatz sorgen, sind die Blueprint Parameter unabhängig von dem Datensatz. Sie enthalten die drei möglichen Architekturen, sowie Hyperparameter mit festen default Werten, wie Verlustfunktion, Training Schedule, Data Augmentation, Normalisierung, stochastic Gradient oder Aktivierungsfunktion (siehe Abbildung 4.5). Die Verlustfunktion wird als die Summe von Dice-Verlustfunktion und Cross-Entropy-Verlustfunktion gewählt. Dies wird gemacht, da medizinische Bilddaten oft Probleme mit einer großen Disbalance im Vorkommen der einzelnen Klassen haben und darum im Training seltener vorkommende Klassen unterrepräsentiert sind und gleichzeitig durch die Lösung dieses Problems die Verteilung der Klassen verzerrt wird. An der Zusammensetzung dieser beiden Verlustfunktionen könnte man also auch arbeiten, wenn man das Framework auf andere Arten von Datensätzen anpassen wollte. Das Training läuft über 1000 Epochen mit jeweils 250 Trainingsiterationen.

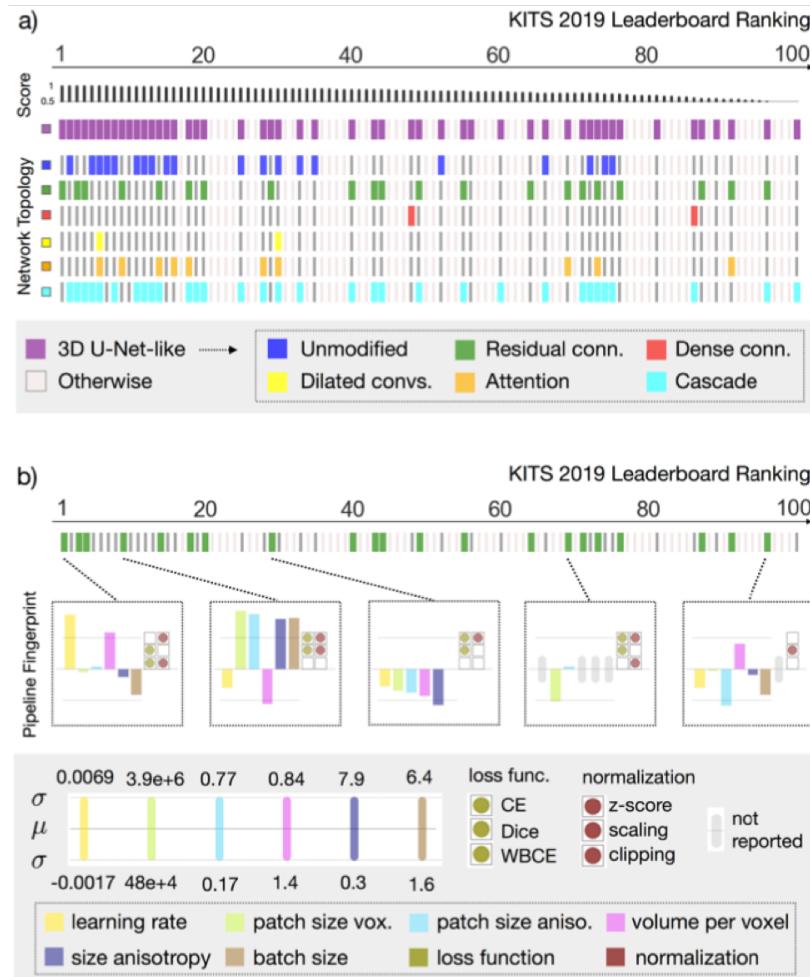


Abbildung 4.5: Leaderboard KITS 2019 und Pipelinefingerprint [18]. a) Platzierungen der KITS 2019 Challenge [19] dargestellt nach Netzwerk-Architektur. Die Verteilung verstrt die Annahme, dass die Architekturauswahl wenig entscheidend ist, die Auswahl der Hyperparameter dagegen sehr wichtig. b) Pipelinefingerprint einiger residual connections Netze aus der Challenge.

Da die empirischen Parameter nicht direkt aus dem Datensatz erschlossen werden knnen, werden sich nach dem Training empirisch bestimmt. Sie werden zur Nachbearbeitung und bei der Auswahl der besten Netzstruktur genutzt.

4.2 Unsere Arbeit / Praxis

Datensatz	Split (Train:Test)	verwendete nnU-Net-Variante	Trainingszeit (h)	Train- Accuracy (Dice)	Test- Accuracy (Dice)
Larven [14]	265:0 \approx 1:0	2D	8:30	0.99970	-
Larven [14]	173:92 \approx 2:1	2D	6:45	0.99982	0.94459
Pascal VOC12 [3]	2516:340 \approx 7:1	2D	26:00	0.90266	0.34953
Retina-2D [20] (manuelle Data-Augmentation)	56:34 \approx 2:1	2D	23:20	0.99977	0.93606
Retina-2D [20] (minimal)	13:32 \approx 1:2	2D	21:15	0.99999	0.83013
CT [21] (2000 Epochen)	19:0 \approx 1:0	3D_fullres (2000 Epochen)	\approx 86:00	0.67197	-
CT [21]	19:0 \approx 1:0	2D	29:00	0.00109	-
CT [21]	19:0 \approx 1:0	3D_cascade	\approx 39:00 + 48:00 =87:00	0.20865	-
Retina-3D [22]	14:7 \approx 2:1	3D_fullres	45:15	0.91863	0.83759
Retina-3D [22]	14:7 \approx 2:1	2D	19:00	0.98574	0.78931
Retina-3D [22] (Ensemble)	14:7 \approx 2:1	2D & 3D_fullres	-	0.97775	0.82363

Abbildung 4.6: Datensätze, auf denen wir trainiert haben (jeweils 1000 Epochen auf GPUv100)

Grundsätzlich haben wir für jeden folgenden Datensatz die Ursprungsdateien in Nifti-Dateien umgewandelt und meistens zufällig in einen Train- und Testsplit aufgeteilt, außer bei dem 3D-CT Datensatz, da wird dort auch mit allen Samples im Trainsplit keine sonderlich guten Ergebnisse erzielen konnten, und dem ersten Versuch auf dem Larvendatensatz.

Der generelle Arbeitsablauf bestand bei uns aus der Befehlsabfolge:

```
nnUNet_plan_and_preprocess -t <TASK-ID> --verify_dataset_integrity
```

bzw. für 2D-Datensätze, da dort kein 3D-Modell anwendbar ist:

```
nnUNet_plan_and_preprocess -t <TASK-ID> -pl3d None
```

Dieser Befehl bereitet das Training vor und prüft, ob der gegebene Datensatz als Nifti-Dateien korrekt ist, indem Wertebereiche und das Vorhandensein aller Dateien geprüft wird. Da dieser Befehl das Training vorbereitet, muss er mit den gleichen verfügbaren Ressourcen wie auch später das Training aufgerufen werden, in unserem Fall auf dem GPUv100 Knoten von Palma II.

Danach kann das Training für die einzelnen Netzvarianten (2d, 3d_fullres, 3d_cascade gestartet werden:

```
# 2d
nnUNet_train 2d nnUNetTrainerV2 <TASK-ID> all --npz
# 3d_fullres
nnUNet_train 3d_fullres nnUNetTrainerV2 <TASK-ID> all --npz
# Cascade
nnUNet_train 3d_lowres nnUNetTrainerV2 <TASK-ID> all --npz
nnUNet_train 3d_cascade_fullres nnUNetTrainerV2CascadeFullRes
    <TASK-ID> all --npz
```

Dabei verwenden wir für 3D-Datensätze immer 3d_fullres und die 2d-Variante, für 2D-Datensätze immer nur die 2d Variante. Falls das Framework es für einen 3D-Datensatz notwendig hält bzw. überhaupt zulässt, wird auch 3d_cascade verwendet. Der Parameter –npz sorgt dafür, dass die Softmax-Ausgaben zusätzlich gespeichert werden, was zwar sehr viel Festplattenspeicher benötigt, uns aber später ein eventuelles Ensembling der Predictions ermöglicht.

Der Parameter „all“ gibt an, welcher der 5 Folds, die das Framework automatisch erstellt, zur Validierung benutzt wird, während die anderen 4 dem Training dienen. Der Autor des Frameworks vermutet, dass wenn man statt „all“ alle Werte 0 bis 4 verwendet und hinterher aus den 5 verschiedenen Folds ein Ensemble bildet die finale Performance besser ist im Vergleich zu „all“, jedoch hat auch er keine empirischen Beweise dafür [23]. Wir haben uns für „all“ entschieden, da die Handhabung dann etwas einfacher wird und wir vermuten, dass die zum Training benötigte Zeit deutlich geringer ist, da lediglich ein Modell aus allen Trainingsdaten trainiert wird, und nicht 5 verschiedene basierend auf einer verschiedener Aufteilung der Folds. Außerdem sind unsere Ergebnisse trotz der nicht optimalen Wahl der Parameter ziemlich gut (s. Tabelle 4.6).

Nach dem Beenden des Trainings lassen wir uns von dem Framework die Predictions zu dem Train- und Testsplit erzeugen:

```
nnUNet_predict -i <Pfad zu Original-Niftis> -o <Prediction-Pfad>
-m <2d, 3d_fullres oder 3d_cascade_fullres> -t <TASK-ID> -f all -z
```

Der Parameter -z sorgt auch hier wieder für das Speichern der Softmax-Werte, um später eventuell ein Ensembling aus verschiedenen Netzvarianten zu bilden.

Abschließend erstellen wir mit dem Framework eine Auswertung der Performance auf dem Datensatz, indem wir für den Train- und Testsplit die Ground-Truth Segmentierung mit den Predictions vergleichen:

```
nnUNet_evaluate_folder -ref <Ground-Truth-Pfad>
```

```
-pred <Prediction-Pfad> -l <Klassennummern>
```

<Klassennummern> ist hierbei eine Liste aller Klassennummern, die in der Auswertung berücksichtigt werden. Da uns die Performance auf dem Hintergrund (0) nicht interessiert, starten wir bei 1 und gehen z.B. bei Pascal VOC12 [3] alle Klassen durch `-l 1 2 3 4 ... 20`. Die erzeugt in dem Ordner, in dem die Predictions liegen eine JSON-Datei mit ausführlichen Informationen über die Güte der Predictions je Sample und je vorhandener Klasse, aus der wir einen Scatterplot erstellen.

4.2.1 Datensätze aus dem Paper

Nach unseren schlechten vorherigen Erfahrungen mit (Auto-) Deeplab [15] und insbesondere NAS-Unet [9] haben wir zuerst versucht die bemerkenswert guten Ergebnisse [17, Kapitel 4, Table 2] auf den Datensätzen der Medical Segmentation Decathlon Challenge [24] mit dem Framework zu reproduzieren. Wir haben die 3D-Datensätze Spleen, Lung und Heart [24] ausprobiert und konnten auf allen ähnliche Ergebnisse wie im Paper [17] erzielen.

4.2.2 Larven-Datensatz

Larven-Datensatz ohne Testsplit

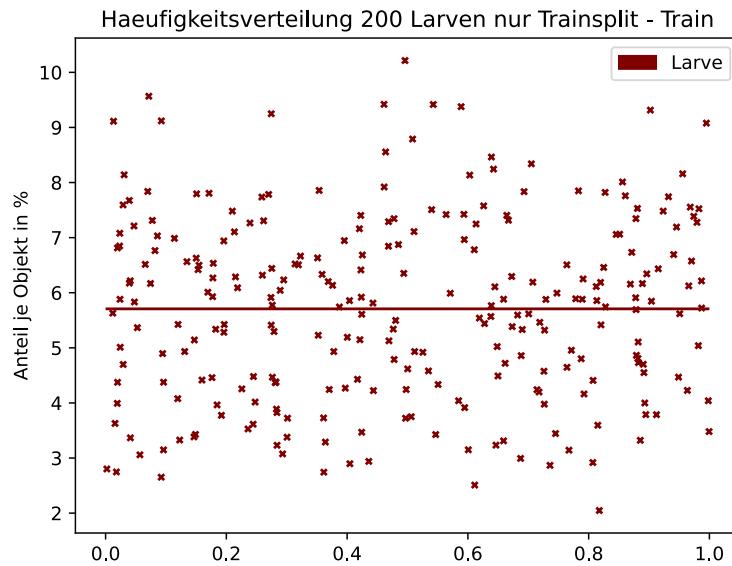
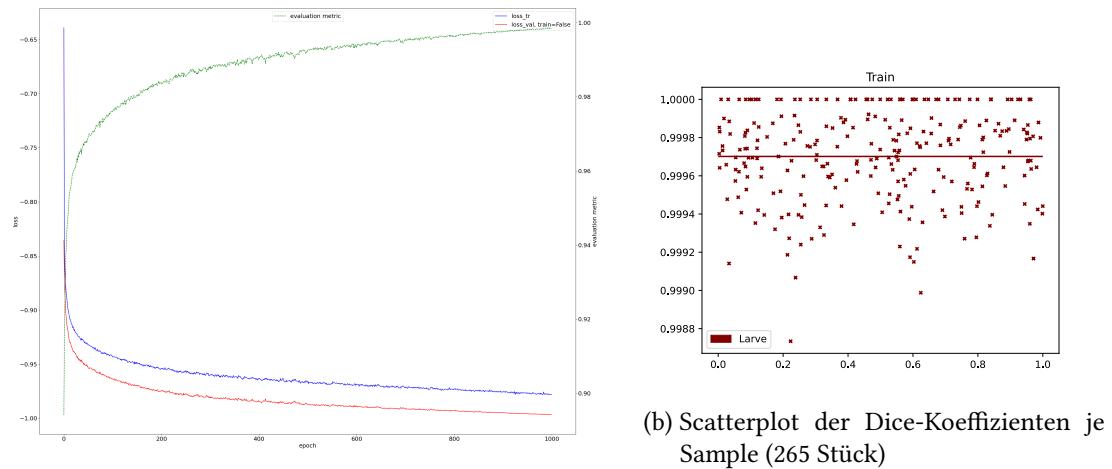


Abbildung 4.7: Anteil von Objekt (Larve) je Sample im Trainsplit (alle 265 Samples) mit Durchschnitt
 $\approx 5,8\%$

Nachdem wir die Ergebnisse im Paper erfolgreich reproduzieren konnten haben wir versucht einen eigenen Datensatz in das Framework zu geben. Dabei ist zu erwähnen, dass unser Larven-Datensatz [14] ein 2D Datensatz ist. Jedoch ist nnU-Net nicht dafür entworfen worden auf 2D-Datensätze angewendet zu werden, besonders wenn die Datensätze aus der „non-biomedical domain“ [25] stammen.

Dies ist jedoch nicht als Einschränkung des Funktionsumfangs zu verstehen, sondern lediglich als Vorwarnung, dass die Ergebnisse eventuell nicht gut ausfallen werden.

Wir haben das zum Framework gehörige Python-Script [26] nur leicht modifizieren müssen und konnten den 2D Datensatz in Nifti-Dateien (.nii.gz) konvertieren und so in das Framework geben. Da es sich hierbei um unseren ersten Test des Frameworks mit eigenem Datensatz handelte haben wir zuerst keinen Test-Split vorgesehen und alle 265 Bilder als Trainingsdaten benutzt.



(a) Verlauf des Dice-Koeffizienten beim Training über 1000 Epochen

Abbildung 4.8: Dice-Koeffizienten auf dem Trainsplit zum Larvendatensatz ohne Testsplit

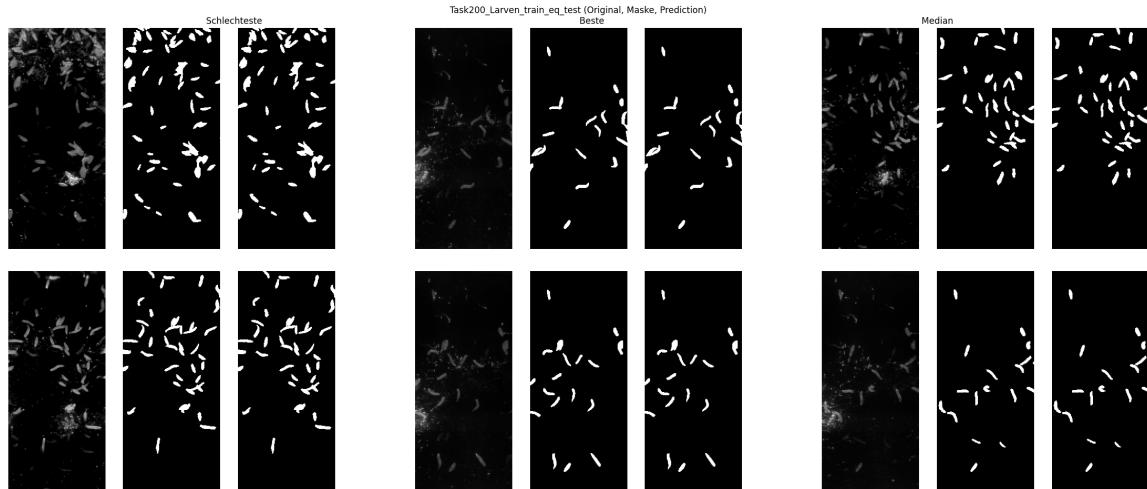


Abbildung 4.9: Visualisierung des Trainsplits auf dem Larvendatensatz ohne Testsplit (links: schlechteste Ergebnisse, mitte: beste Ergebnisse, rechts: Ergebnisse im Median; jeweils Original, Ground-Truth und Prediction)

Wir konnten bei unserem ersten Versuch einen eigenen Datensatz in das Framework zu geben einen Dice-Wert von durchschnittlich > 0.999 erzielen und auch die am schlechtesten segmentierten Samples liegen weit über 0.99, wie in Abbildung 4.8b zu erkennen ist. Auch bei der Visualisierung der besten, schlechtesten und mittleren Ergebnisse (Abbildung 4.9) kann man zwischen Ground-Truth und der Prediction mit bloßem Auge keine Unterschiede erkennen.

Larven-Datensatz mit $\frac{2}{3}$ Train- und $\frac{1}{3}$ Testsplit

Anschließend haben wir die Larvenbilder zufällig in $\frac{2}{3}$ Trainingsdaten und $\frac{1}{3}$ Testdaten aufgeteilt und erneut trainieren lassen. Wir haben uns vergewissert, dass Train- und Testsplit möglichst allgemein und zueinander ähnlich sind, und nicht zufällig in einem Split nur die Samples mit hohem Objektanteil und im anderen mit wenig Objektanteil vorhanden sind. Sowohl im Train- als auch im Testsplit ist der Objektanteil in den Samples ähnlich (s. Abbildung 4.10).

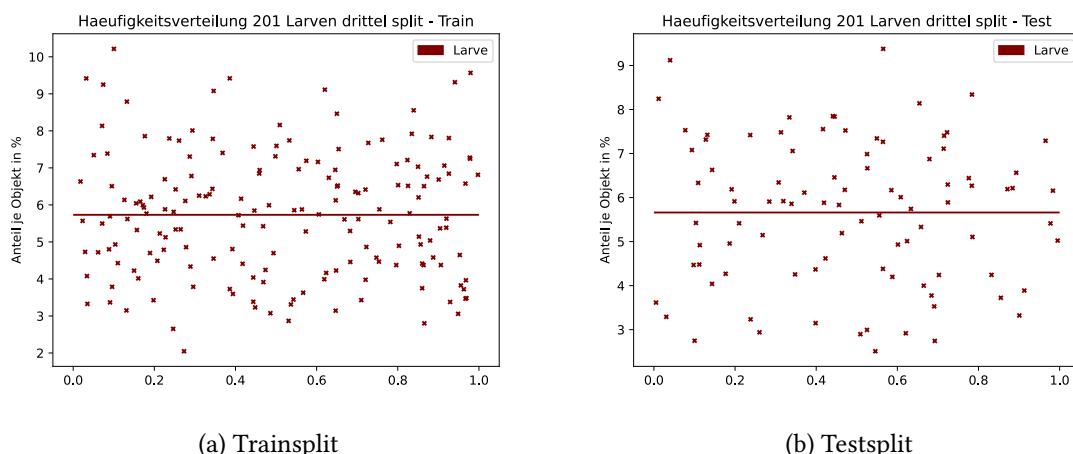
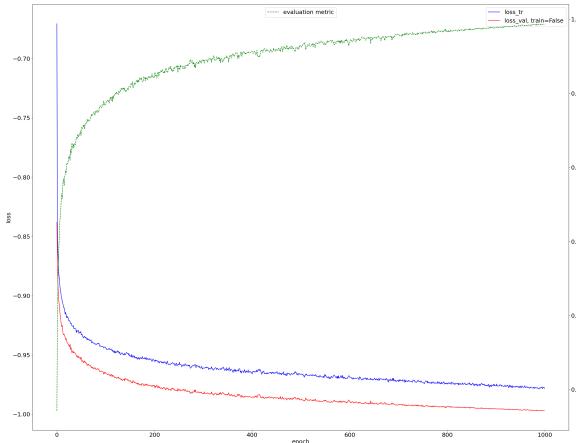
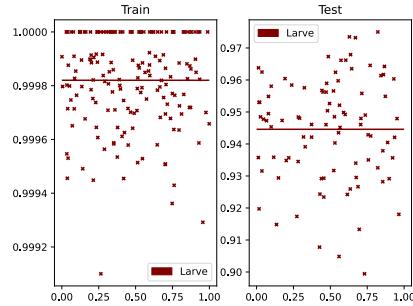


Abbildung 4.10: Anteil von Objekt (Larve) je Sample mit Durchschnitt je Split $\approx 5,8\%$

Hierbei konnten wir nach dem Training mit Hilfe des Testsplits die Performance des erlernten Modells evaluieren und prüfen, ob das Framework tatsächlich gelernt hat die Strukturen und Merkmale einer Larve zu erkennen und von ähnlich aussehender Verschmutzung zu unterscheiden. Auf dem Testsplit ist die Performance etwas schlechter als auf dem Trainsplit, aber mit einem Dice-Koeffizienten von über 0.94 im Durchschnitt trotzdem noch erstaunlich gut und selbst das am schlechtesten segmentierte Sample im Testsplit hat mit 0.9 auch noch eine akzeptable Genauigkeit (s. Abbildung 4.11b).



(a) Verlauf des Dice-Koeffizienten beim Training über 1000 Epochen



(b) Scatterplot der Dice-Koeffizienten je Sample für Train- und Testsplitt

Abbildung 4.11: Dice-Koeffizienten zum Larvendatensatz mit einem Drittel als Testsplitt

Bei der Visualisierung der Predictions fällt erneut auf, dass auf dem Trainsplit mit bloßem Auge keine Unterschiede zu Ground-Truth vorhanden sind (s. Abbildung 4.12), bei dem Testsplitt kommt es jedoch bei den schlechtesten Beispielen zu Fehlern, die auch deutlich erkennbar sind. Es werden teilweise komplett Larven nicht oder nur teilweise erkannt und zudem wird auch besonders in den stark verschmutzen Bildern die Verschmutzung als Larve erkannt. Hierbei stellt sich die Frage, ob die Ground-Truth Segmentierung korrekt ist, da besonders die angeblichen Verschmutzungen, die das Modell als Larve erkannt hat, im Originalbild tatsächlich eher wie eine Larve aussehen als eine Verschmutzung. Intuitiv würde man die betroffenen Stellen im Originalbild vermutlich auch, wie das Modell, als Larve interpretieren. Die Larven die nicht erkannt wurden sind sehr schwach bis gar nicht mit dem Auge erkennbar bzw. ähneln stark einer Verschmutzung (s. Abbildung 4.13).

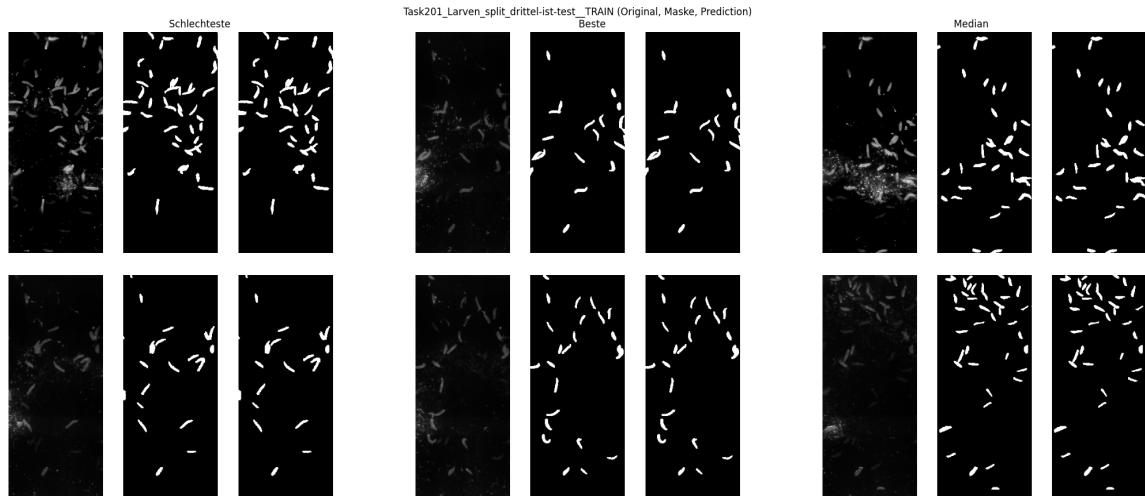


Abbildung 4.12: Visualisierung des Trainsplits auf dem Larvendatensatz mit $\frac{1}{3}$ Testsplit (links: schlechteste Ergebnisse, Mitte: beste Ergebnisse, rechts: Ergebnisse im Median; jeweils Original, Ground-Truth und Prediction)

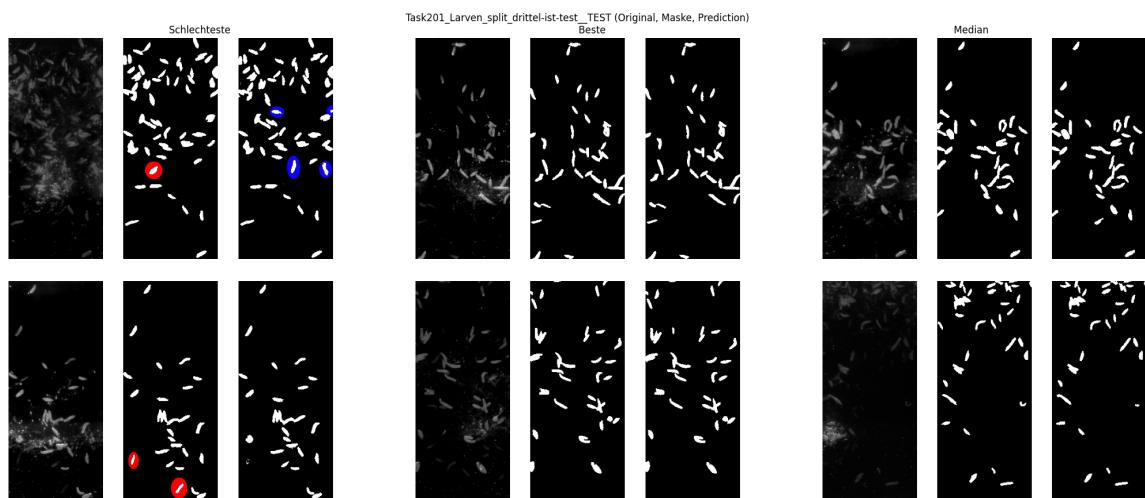


Abbildung 4.13: Visualisierung des Testsplits auf dem Larvendatensatz mit $\frac{1}{3}$ Testsplit (links: schlechteste Ergebnisse, Mitte: beste Ergebnisse, rechts: Ergebnisse im Median; jeweils Original, Ground-Truth und Prediction). Rot eingefärbt sind nicht erkannte Larven, blau als Larven erkannte Verschmutzungen

4.2.3 Retina 2D-Datensatz

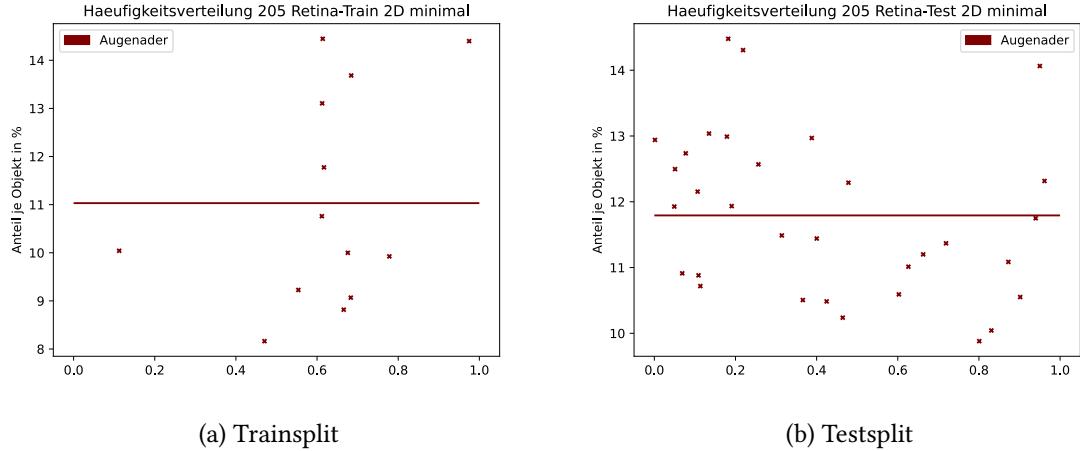


Abbildung 4.14: Anteil von Objekt (Ader) je Sample mit Durchschnitt je Split $\approx 11\text{-}12\%$

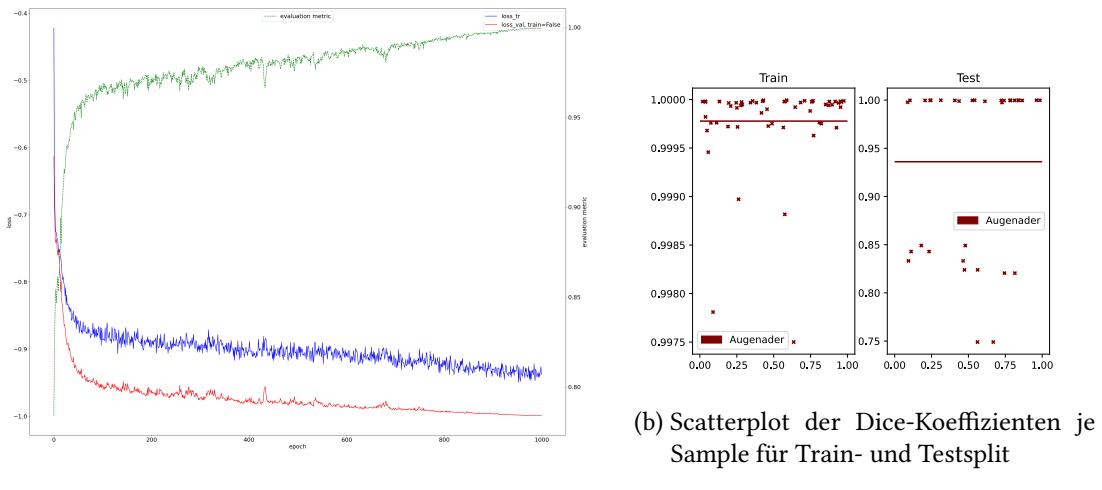
Da wir auf dem Larven-Datensatz [14] (Graustufen mit einer einzigen Objekt-Klasse) so gute Ergebnisse erzielen konnten, obwohl das Framework für solche Aufgaben eigentlich nicht gemacht ist, wollten wir einen Schritt weiter gehen und statt graustufen Bildern farbige Bilder verwenden. Dazu verwenden wir den Retina-2D Datensatz [20]. Dieser besteht aus jeweils 15 Aufnahmen von gesunden Retinae, Retinae von Augen mit Glaukom und diabetischer Retinopathie, also insgesamt 45 hochauflösenden RGB-Bildern. Unser Ziel der Segmentierung ist es, unabhängig von der Erkrankung, die Adern in der Retina zu markieren.

Diese Bilder konnten auch wie bei den Larven mit dem zur Verfügung gestellten Python-Script [26] in Nifti-Dateien konvertiert werden. Jedoch ergab sich beim Ausführen des Trainings das Problem, dass die automatisch ermittelte Batch-Size des Frameworks angeblich zu niedrig ist. Nach etwas Ausprobieren und Nachschauen im Code sind wir auf eine „estimated GPU-RAM consumption“ [27] gestoßen, die die Batch-Size vorgibt. Durch die hohe Auflösung der Bilder ($\approx 3500 \times 2300$) und den 3 Farbkanälen wird dieser geschätzte GPU-Ram Verbrauch zu groß und als Folge dessen die Batch-Size mit 1 zu klein, da in einem Batch per Definition des Frameworks immer mindestens 2 Samples enthalten sein müssen.

Durch Ausgeben des geschätzten GPU-Ram Verbrauchs haben wir herausgefunden, dass dieser ungefähr linear mit der Anzahl an Pixeln wächst und konnten so ausrechnen, dass eine Verkleinerung der Bilder auf mindestens 42% ausreicht, damit 2 Samples in ein Batch gelangen können. Diese Verkleinerung der Auflösung muss nur für die Trainingsdaten vorgenommen werden. Auf den Testdaten, von denen lediglich eine Prediction erstellt werden muss, kann die Auflösung höher sein.

Außerdem haben wir, bevor wir mit dem geschätzten GPU-Ram Verbrauch gespielt haben, manuelle

Data-Augmentation betrieben indem wir die Bilder rotiert und gespiegelt haben, in der Hoffnung dadurch mehr Samples in einem Batch zu erhalten. Dies hat sich im Nachhinein jedoch als nicht nötig herausgestellt und hat dem Framework das Training im 1. Durchlauf eventuell unnötig erschwert. Später im 2. Durchlauf mit minimaler Trainingssample-Anzahl haben wir diesen Fehler nicht gemacht.



(a) Verlauf des Dice-Koeffizienten beim Training über 1000 Epochen

Abbildung 4.15: Dice-Koeffizienten zum Retina-2D Datensatz [20] mit einem Drittel als Testsplitt

Da wir auch hier relativ gute Ergebnisse erzielen konnten, wollten wir das Framework an seine Grenzen bringen und so wenig Trainingsamples wie möglich zur Verfügung stellen. Durch Ausprobieren und schrittweises Annähern haben wir herausgefunden, dass nnU-Net, jedenfalls bei diesem Datensatz, mindestens 13 Trainingsbeispiele benötigt, da bei weniger Trainingsbeispielen später beim Training ein *IndexOutOfBoundsException* Fehler auftritt. Leider ist beim Aufteilen in Train- und Testsplit der Objektanteil in den Samples nicht ganz ausbalanciert, da im Trainsplit nur 11% der Pixel Adern sind und im Testsplit knapp 12% (s. Abbildung 4.14).

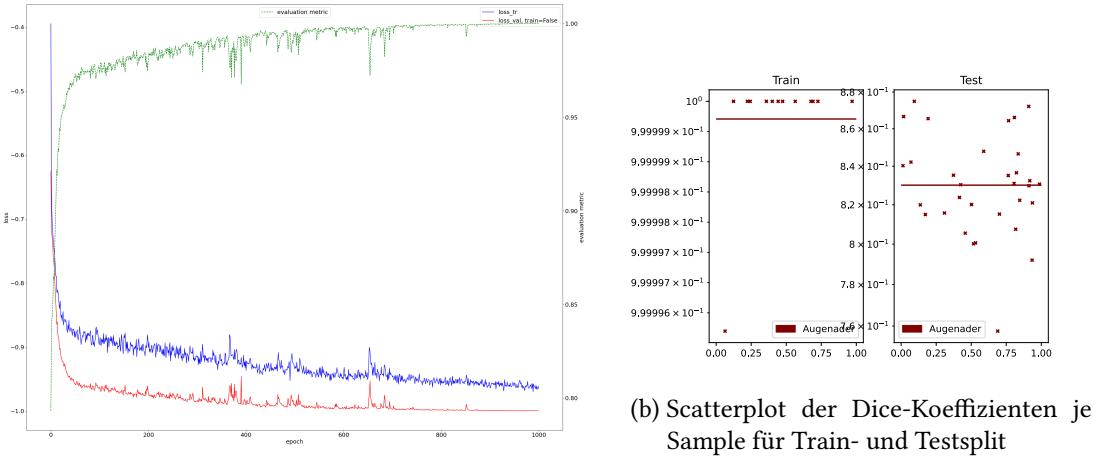


Abbildung 4.16: Dice-Koeffizienten zum Retina-2D Datensatz mit einem minimalen Trainsplit von 13 Samples

Es fällt auf, dass bei dem minimalen Trainingssplit Overfitting stattfindet, da alle Samples bis auf ein einziges einen Dice-Koeffizienten von genau 1 besitzen (s. Abbildung 4.16b). Auch der Progress-Graph (Abbildung 4.16a) steigt am Anfang schneller als bei $\frac{2}{3}$ Trainingssplit.

Auf dem Testsplit fällt auf, dass bei Durchlauf 1 (Abbildung 4.15b) eine Gruppierung stattfindet. Viele Samples liegen sehr nah bei 1 und eine zweite, etwa gleich große Gruppe liegt um 0,85 herum. Dies kommt sehr wahrscheinlich von unserer, fälschlicherweise durchgeföhrten, manuellen Data-Augmentation, bei der wir das gleiche Bild mehrmals in rotierter und gespiegelter Form in das Framework gegeben haben. Beim 2. Durchlauf mit minimalem Trainsplit sind die Samples im Testsplit relativ gleichmäßig um den Durchschnitt von 0,83 verteilt (s. Abbildung 4.16b).

Beim Betrachten der Visualisierung der besten, schlechtesten und mittleren Predictions je Split (Abbildungen 4.17, 4.18) fällt auf, dass die Adern in der Predictions generell breiter sind als in Ground-Truth. Das kommt von der leider notwendigen Verkleinerung der Auflösung der Bilder, da dann bei feinen Adern anstatt nur schwarzer oder weißer Pixel in der Ground-Truth Segmentierung auch graue Pixel entstehen, da z.B. Adern mit einer Breite von einem Pixel nicht weiter in der Auflösung reduziert werden können. Wir haben uns dafür entschieden, die dann grauen Pixel auch als weiße Pixel, also als Ader-Segmentierung, zu zählen, da ansonsten feine Adern Lücken bekommen und die Segmentierung insgesamt schlechter ausfällt. Wenn das Framework wie geplant zu einem späteren Zeitpunkt Patches in der 2D-Variante erlaubt, wäre eine Verkleinerung der

Auflösung nicht mehr nötig und das Problem der zu breiten Adern nicht mehr vorhanden [28]. Ansonsten findet man keine groben Fehler, die meisten Adern werden sehr genau erkannt.

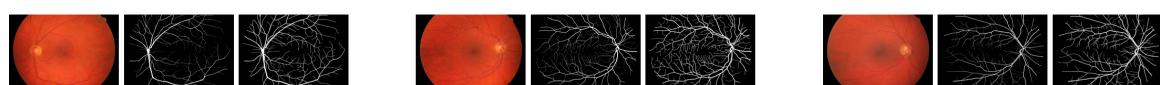
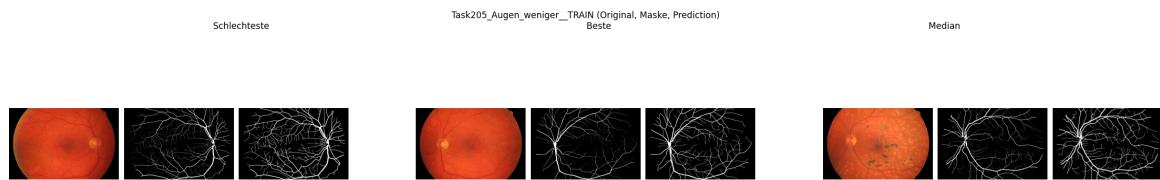


Abbildung 4.17: Visualisierung des Trainsplits auf dem Retina-Datensatz mit minimalem Trainsplit (links: schlechteste Ergebnisse, mitte: beste Ergebnisse, rechts: Ergebnisse im Median; jeweils Original, Ground-Truth und Prediction)

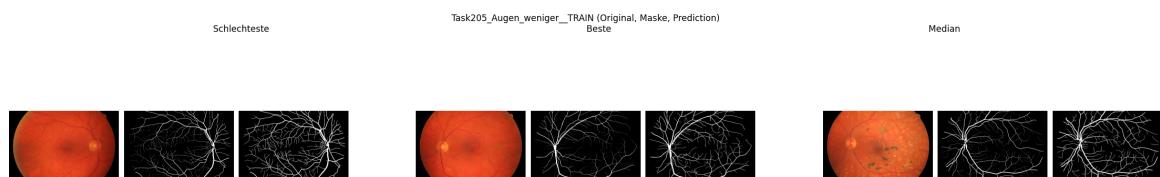


Abbildung 4.18: Visualisierung des Testsplits auf dem Retina-Datensatz mit minimalem Trainsplit (links: schlechteste Ergebnisse, mitte: beste Ergebnisse, rechts: Ergebnisse im Median; jeweils Original, Ground-Truth und Prediction)

Da wir bei diesem Datensatz relativ gute Ergebnisse produzieren konnten, haben wir dem Framework komplett fremde Bilder [29] in verschiedenen Auflösungen und Zoom-Stufen zum Segmentieren gegeben. Zu diesen Bildern gab es leider keine Ground-Truth Segmentierung, jedoch kann man auch so grob abschätzen wie robust das Modell ist, und dass es tatsächlich die Merkmale einer Ader in der Retina erlernt hat und sogar mit verschiedenen Auflösungen, Ausschnitten und *Färbungen* der Retina umgehen kann (s. Abbildung 4.19 und 4.20).

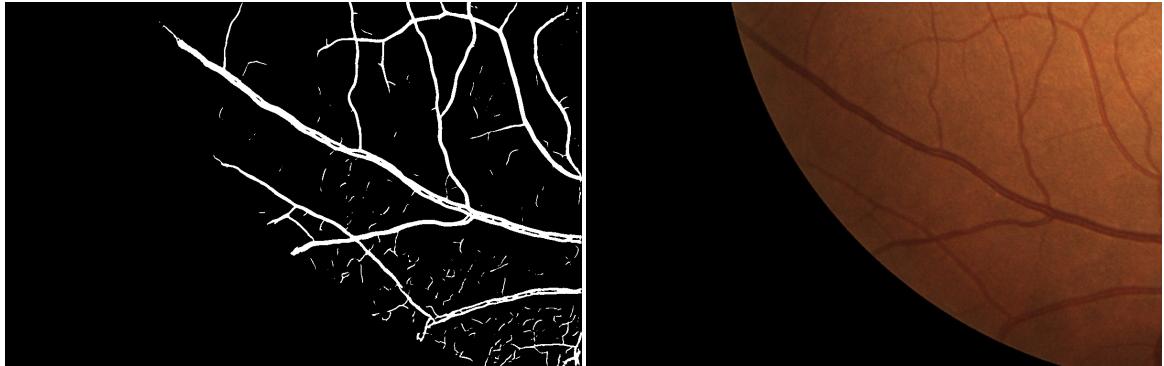


Abbildung 4.19: Bild 13a_right aus [29] in starkem Zoom - Prediction und Original

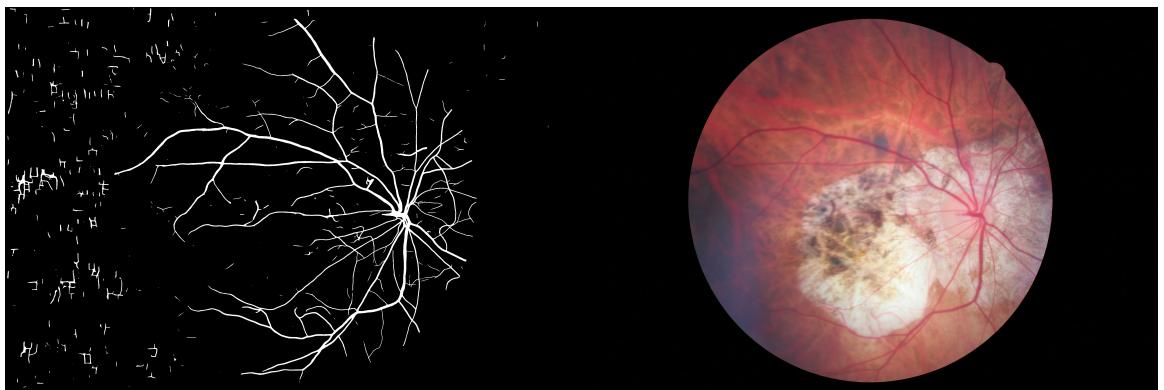


Abbildung 4.20: Bild 1170_right aus [29] mit Wucherung im Hintergrund - Prediction und Original

4.2.4 CT-Datensatz

Da wir bisher nur eigene 2D-Datensätze in das Framework gegeben haben und es nur wenig öffentlich zugängliche 3D-Datensätze zur Segmentierung gibt, die nicht Teil der MSD-Challenge [24] sind, wurde uns ein Datensatz mit 19 Ganzkörper CT-Aufnahmen zur Verfügung gestellt [21], in denen Kalzium-Ablagerungen am Rand der Gefäße segmentiert werden sollen. Auffällig ist hierbei, dass trotz der hohen Auflösung des Datensatzes (512×512 mit $\approx 400\text{-}600$ Slices je Sample) nur *sehr* wenige Pixel in der Segmentierung markiert sind. Bestenfalls sind ≈ 13000 Pixel im kompletten

3D Volumen bestehend aus 570 Slices mit je einer Auflösung von 512x512, was einem Anteil von *maximal* 0,008% entspricht. Außerdem gibt es auch Samples mit deutlich weniger oder gar keinen markierten Pixeln in der Ground-Truth Segmentierung (s. Abbildung 4.21).

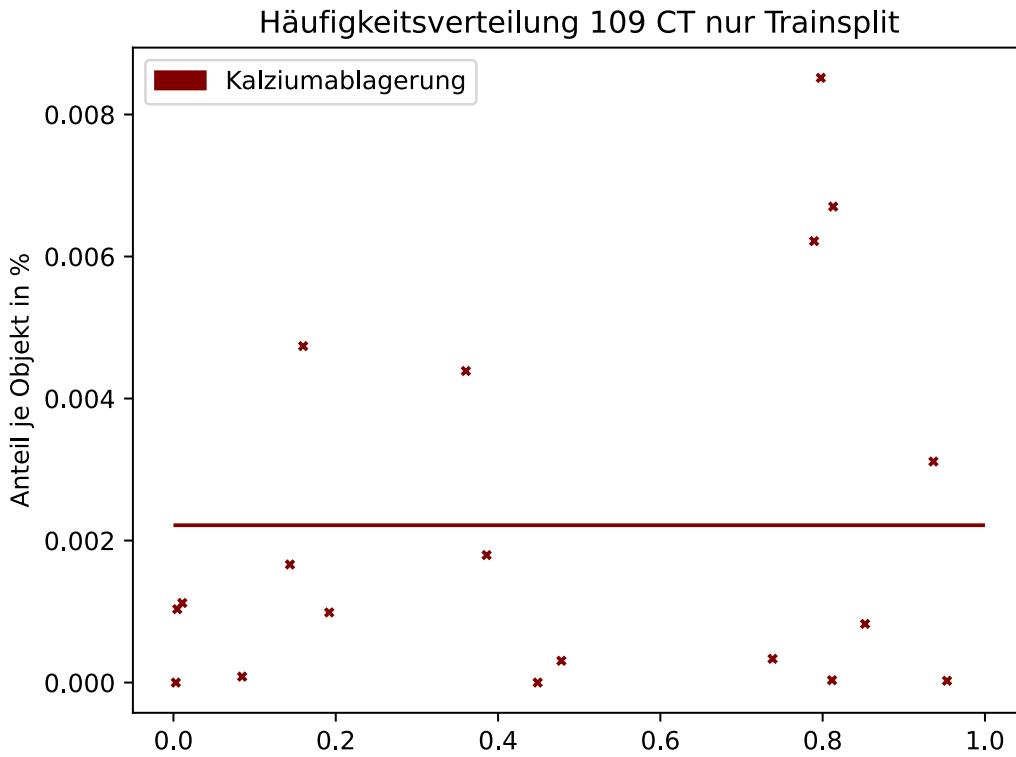


Abbildung 4.21: Anteil von Objekt (Kalziumablagerung) je Sample mit Durchschnitt $\approx 0.002\%$

Dies erschwert das Training und könnte eine Begründung für die nicht sonderlich guten Ergebnisse sein.

Um den Datensatz in nnU-Net zu geben, mussten wir erst die zur Verfügung gestellten .nrrd (Ground-Truth) und .dcom (3D-CT-Scan) Dateien mit eigenen kleinen Pythonscripten [30] in Nifti-Dateien umwandeln. Dabei sind wir auf sehr viele Probleme gestoßen und hatten erst kurz vor Ende des Projektseminars einigermaßen gute Ergebnisse. Zuerst hatten wir das Problem, dass beim Einlesen der .dcom und .nrrd Dateien die Drehung nicht intuitiv und nicht bei beiden gleich ist. Somit passte die Drehung der Ground-Truth Dateien nicht mehr zu den Original CT-Dateien, was aufgrund der geringen Häufigkeit an Pixeln in den Ground-Truth Daten und den mit dem Auge schwer ausfindig zu machenden zugehörigen Stellen im Original-Bild nicht direkt ersichtlich war. Nachdem wir die Dateien passend gedreht haben und sie in allen Dimensionen übereinstimmten, konnten wir beim Training mit allen Netzvarianten keine sonderlich guten Ergebnisse erzielen. Wir

haben lange ausprobiert und gesucht, woran das liegen könnte, bis wir dann schließlich festgestellt haben, dass wir beim Konvertieren in Nifti-Dateien versehentlich den Wertebereich jedes Pixels von 16 Bit auf 8 Bit einschränken. Außerdem ist uns beim Trainieren erst nach zweimaligem Neustarten ohne Erfolg, mit sehr schlechten Ergebnissen bei genauerem Lesen des Papers aufgefallen, dass wir das Preprocessen in nnU-Net bisher immer auf dem Login-Node des Clustercomputers Palma II durchgeführt haben, der keine Grafikkarte besitzt, um Wartezeiten in der Warteschlange zu vermeiden. Das Preprocessen ist aber abhängig von den zu dem Zeitpunkt zur Verfügung gestellten Ressourcen, insbesondere des verfügbaren GPU Speichers. Nachdem wir diesen Denkfehler behoben haben und den CT-Datensatz sowohl auf einer GPUv100 Karte preprocessed und trainiert haben, gelang es uns einigermaßen akzeptable Ergebnisse zu erzielen.

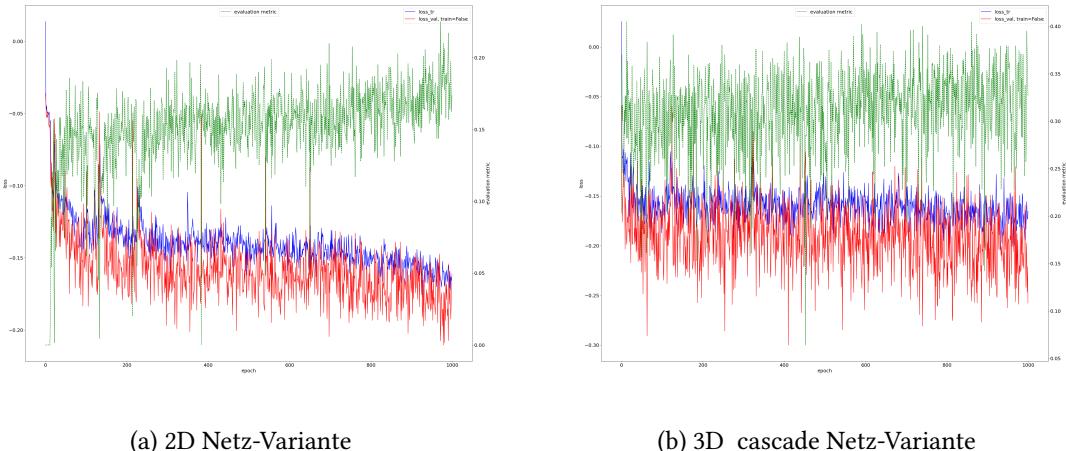


Abbildung 4.22: Verlauf des Dice-Koeffizienten beim Training über 1000 Epochen

Bei dem Training der 2D und 3D_cascade Netz-Variante fällt auf, dass kaum eine Steigung vorhanden ist und die Ergebnisse mit einem Dice-Koeffizienten von ≈ 0.2 bzw. 0.35 gleichbleibend schlecht sind (s. Abbildung 4.22). Bei der 3D_fullres Netz-Variante haben wir eine deutliche Steigung festgestellt, waren mit dem Ergebnis nach den standardmäßigen 1000 Epochen aber noch nicht zufrieden. Da die Steigung des Progress-Graphen (Abbildung 4.23) auch bei Epoche 1000 noch relativ stark ist, haben wir das Training um weitere 1000 Epochen fortgesetzt und so insgesamt 2000 Epochen, also doppelt so lange wie die Standardeinstellung, trainieren lassen und konnten einen Dice-Koeffizienten von ca. 0.6 erreichen (s. Abbildung 4.23).

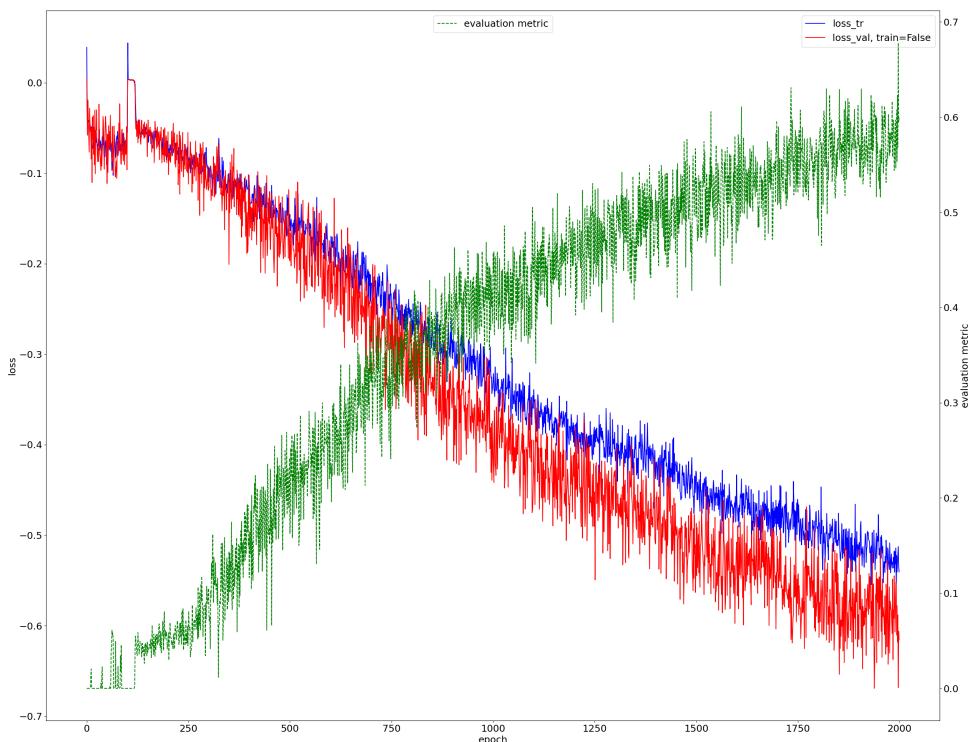


Abbildung 4.23: Verlauf des Dice-Koeffizienten beim Training über 2000 Epochen mit 3D_fullres

Bei den Ergebnissen fällt auf, dass die Güte der Predictions bei 3D_cascade stark streut und nur in einem Fall sehr gut ist (s. Abbildung 4.24b). Die 2D Netz-Variante ist nicht zu gebrauchen (s. Abbildung 4.24a). Dies liegt daran, dass dem 2D-Netz die Information über die Tiefe fehlt und nur die Slices einzeln für sich betrachtet werden.

Lediglich mit 3D_fullres nach 2000 Epochen konnten wir einigermaßen akzeptable Ergebnisse erzielen (s. Abbildung 4.25). Hierbei wurden Samples, die schon in Ground-Truth keinen einzigen markierten Pixel besitzen, ausgelassen, da dort kein Dice-Koeffizient berechnet werden kann.

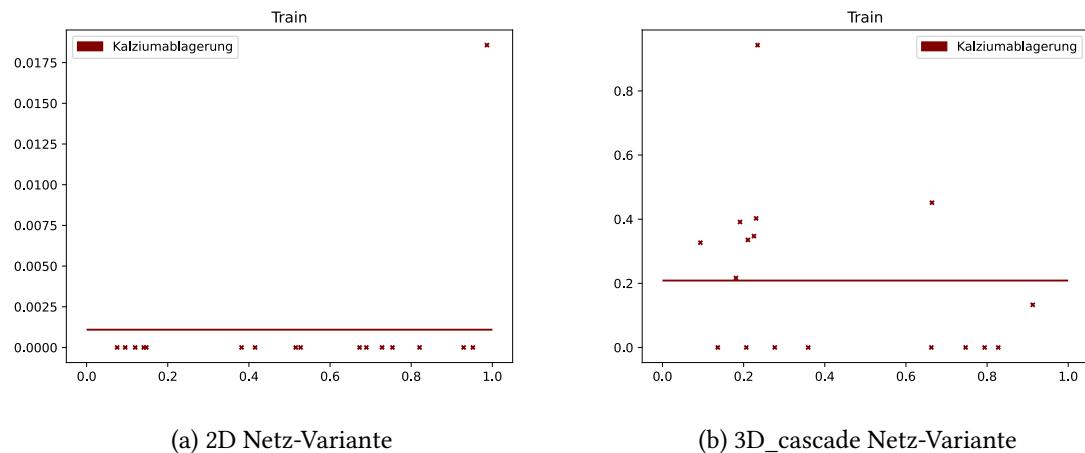


Abbildung 4.24: Scatterplot der Dice-Koeffizienten je Sample in der 2D bzw 3D_cascade Netz-Variante

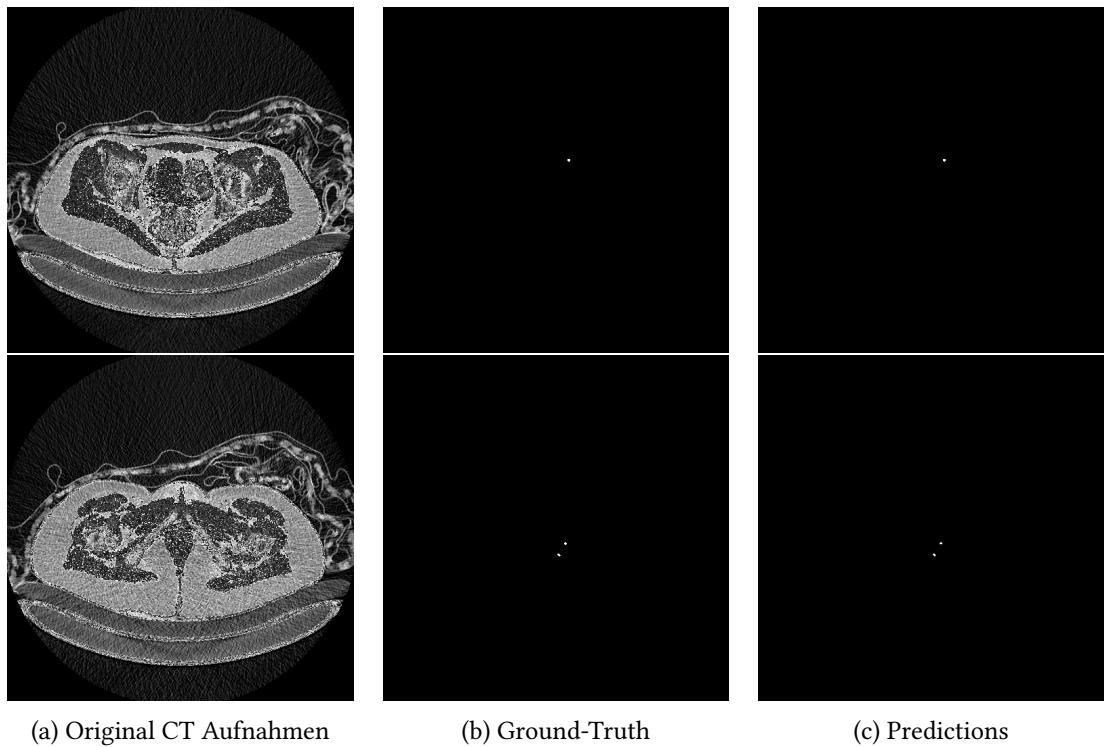
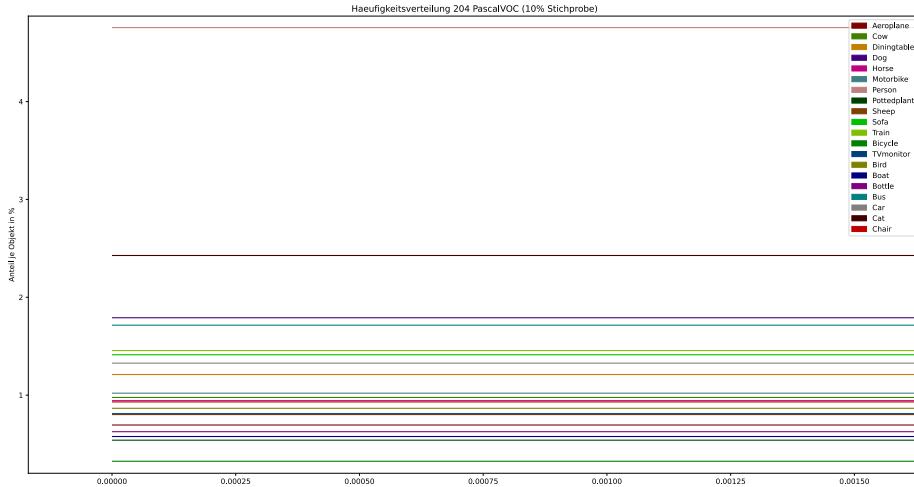


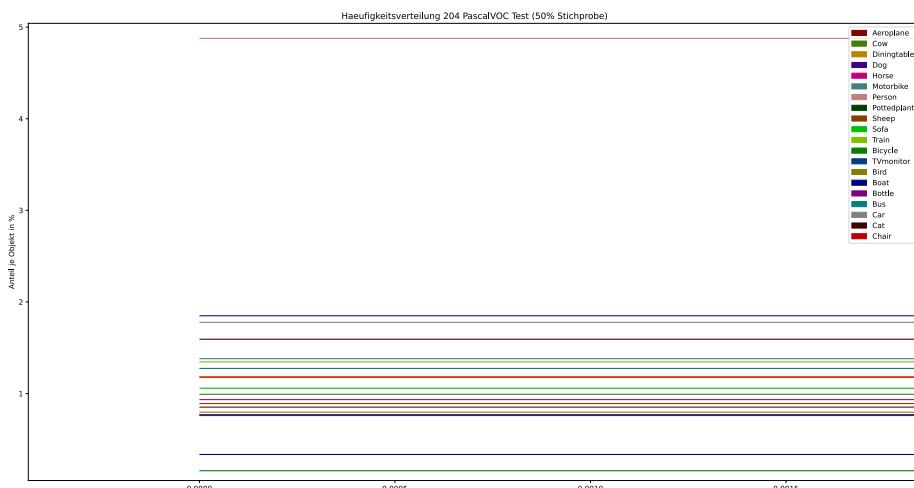
Abbildung 4.26: Visualisierung eines Samples (003)

Bei der Visualisierung des zweitbesten Samples (Nummer 003, Dice=0,906) sieht man, wie wenig Pixel (insgesamt über alle Slices 1628 Stück) markiert sind und wie schwer zu erkennen die zugehörigen Stellen im Original CT-Bild sind (s. Abbildung 4.26). In dem besten Sample (Nummer 018, Dice=1) wären es noch deutlich weniger Pixel gewesen (34 Stück).

4.2.5 Pascal VOC2012



(a) Trainsplit vergrößert



(b) Testsplit vergrößert

Abbildung 4.27: Durchschnittswerte für jede Klasse und ihre zugehörige Auftrittshäufigkeit im gesamten Datensatz

Nachdem wir bei den bisherigen 2D-Datensätzen nur Graustufen-Bilder mit einer Klasse (Larven [14]) und farbige Bilder mit einer Klasse (Retina 2D [20]) verwendet haben wollten wir auch noch einen 2D-Datensatz in Farbe mit mehreren Objekt-Klassen ausprobieren. Die Entscheidung fiel

relativ schnell auf Pascal VOC2012 [3], da dieser sehr viele Klassen (20 Klassen + Background) und viele segmentierte Bilder (2856 Stück) in verschiedenen Formaten und Auflösungen zur Verfügung stellt und von vielen anderen Frameworks zum Vergleichen verwendet wird. Außerdem ist dieser Datensatz mit Fotografien nochmal wesentlich weiter von der „biomedical-domain“ [25], für die das Framework eigentlich erstellt wurde, entfernt und dient somit als Test, wie robust das Framework mit verschiedenen Daten umgeht.

Um die Pascal-VOC 2012 Bilder in nnU-Net zu geben, mussten erst die Farbkodierungen in der Ground-Truth-Segmentierung in Indizes umgewandelt werden [30], da nnU-Net bei 0=Background beginnend aufsteigende Integer für die Klassen erwartet, aber in dem Pascal-Datensatz [3] die Segmentierung zur besseren Erkennbarkeit farblich gekennzeichnet ist.

Außerdem mussten 57 Bilder, die nicht farbig sondern nur in Graustufen in Pascal-VOC 2012 [3] vorhanden sind entfernt werden, da das Framework nicht mit einer variablen Anzahl an (Farb-)Kanälen umgehen kann.

Bei dem Datensatz ist auffällig, dass die Klassenhäufigkeiten sich untereinander unterscheiden und manche Klassen wesentlich häufiger auftreten als andere. Im Train- und im Testsplit ist diese Unausgeglichenheit jedoch ähnlich (s. Abbildung 4.27).

Um die Gesamt-Performance auf diesem Datensatz auszuwerten mussten wir erstmalig über mehrere Klassen einen Durchschnitt bilden. Naiv könnte man die Dice-Koeffizienten jeder Klasse addieren und am Ende durch die Anzahl der Klassen teilen. Dies hätte jedoch die Folge, dass gute bzw. schlechte Klassen, die sehr selten vorkommen, das Endergebnis unverhältnismäßig nach oben bzw. unten ziehen. Wir haben uns für eine Gewichtung der Form

$$\text{average} = \frac{\sum_i d_i * n_i}{\sum_i n_i}$$

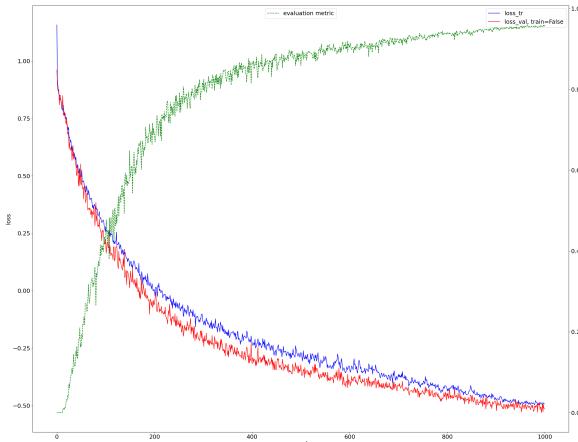
mit d_i : Dice-Koeffizient eines Samples für Klasse i

n_i : Anzahl zu Klasse i gehöriger Pixel im Ground-Truth des Samples.

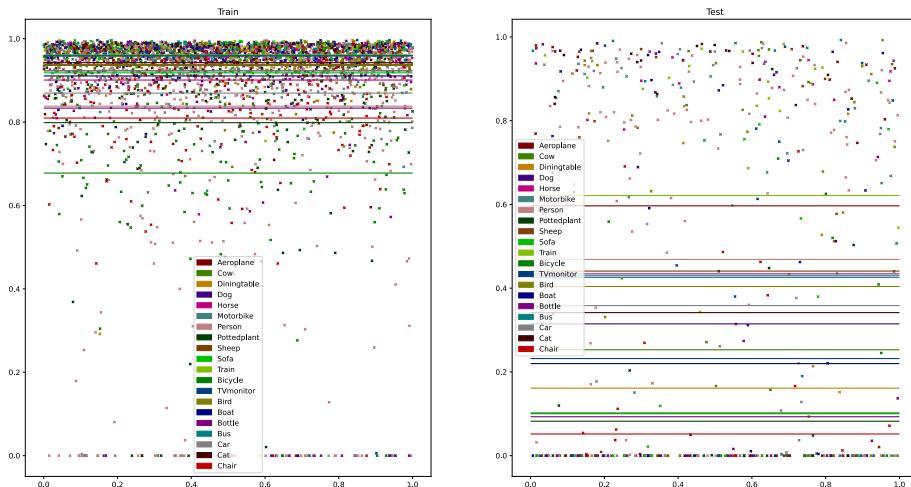
Diese Formel kann auch auf die Durchschnittswerte, die automatisch vom Framework je Klasse generiert werden, angewandt werden um einen Durchschnittswert über alle Klassen hinweg zu bilden.

Das Training verlief gut und es konnten, wie von den anderen Datensätzen gewohnt, relativ schnell gute Ergebnisse auf dem Trainsplit bzw. im Progress-Graph verzeichnet werden. Im Testsplit sind die Ergebnisse jedoch weit von denen aus dem Trainsplit entfernt. Dies zeigt unserer Meinung nach die Grenzen dieses Frameworks auf, da es bisher sehr gut mit regelmäßigen Mustern und Strukturen wie z.B. bei den Larven oder den Retina Adern umgehen konnte, jedoch jetzt bei Fotografien aus verschiedenen Blickwinkeln mit Objekten von Klassen in verschiedenen Größen, die sich zudem noch sehr stark ähneln (z.B. Schaf und Hund) schlecht abschneidet. Außerdem

ist auffällig, dass bestimmte Klassen wesentlich schlechter auf dem Testsplit sind als andere (s. Abbildung 4.28a und 4.28b).



(a) Verlauf des Dice-Koeffizienten beim Training über 1000 Epochen



(b) Scatterplot der Dice-Koeffizienten je Sample für Train- und Testsplitt

Abbildung 4.28: Dice-Koeffizienten zum PascalVOC 2012 [3] Datensatz mit $\frac{7}{8}$ Trainsplit und $\frac{1}{8}$ Testsplitt

Bei der Visualisierung fällt auf, dass auf dem Trainsplit die meisten Objekte grob erkannt werden, nur bei feineren Strukturen (z.B. dem kleinen Fahrrad) wird ungenau segmentiert. Ansonsten sind keine groben Fehler vorhanden und es werden keine Klassen verwechselt (s. Abbildung 4.29).

Auf dem Testsplitt hingegen sieht man, dass ähnlich aussehende Klassen wie z.B. Person, Schaf, Kuh und Hund häufig verwechselt werden, aber im Median eine ziemlich solide Performance gegeben ist (s. Abbildung 4.30).

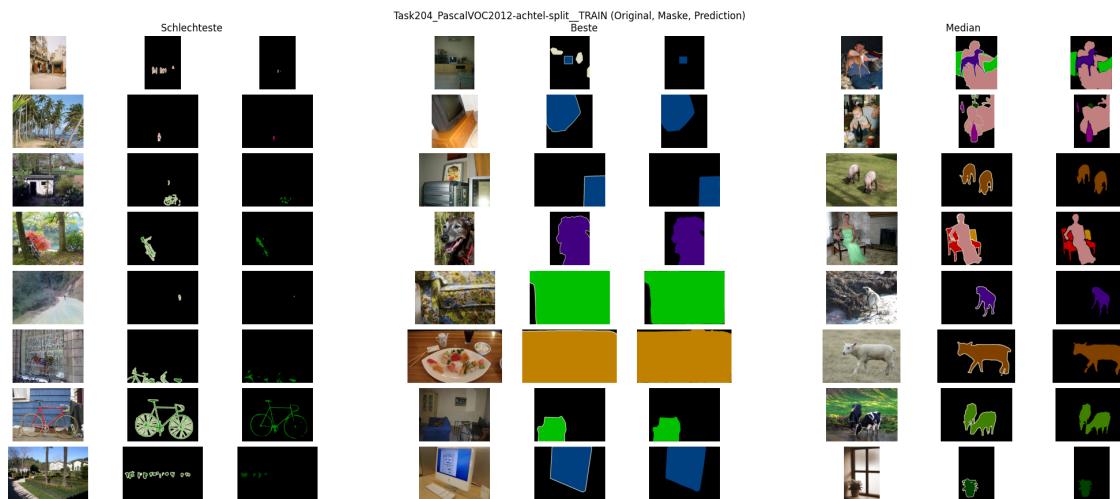


Abbildung 4.29: Visualisierung des Trainsplits auf dem PascalVOC 12 [3] Datensatz (links: schlechteste Ergebnisse, mitte: beste Ergebnisse, rechts: Ergebnisse im Median; jeweils Original, Ground-Truth und Prediction)

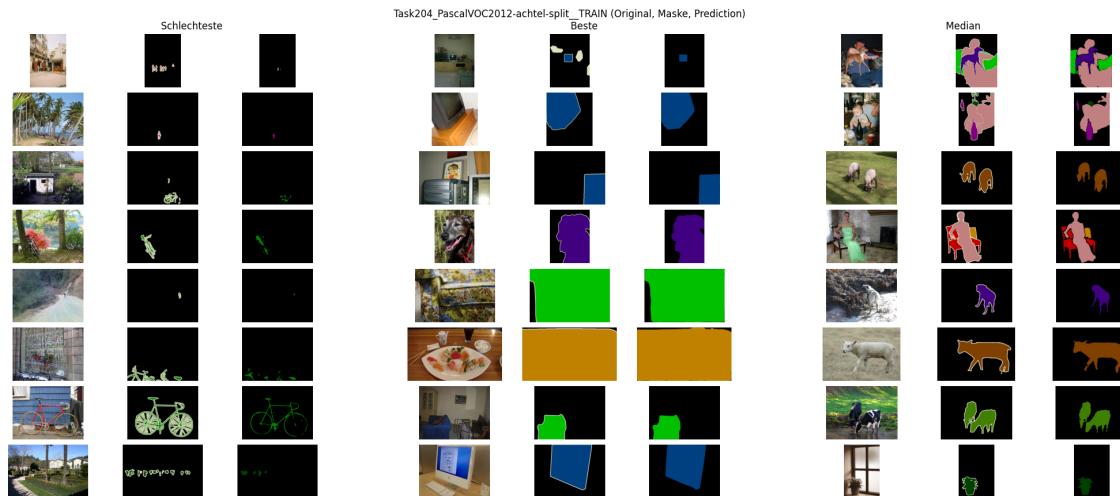


Abbildung 4.30: Visualisierung des Testsplitts auf dem PascalVOC 12 [3] Datensatz (links: schlechteste Ergebnisse, mitte: beste Ergebnisse, rechts: Ergebnisse im Median; jeweils Original, Ground-Truth und Prediction)

4.2.6 Retina 3D-Datensatz

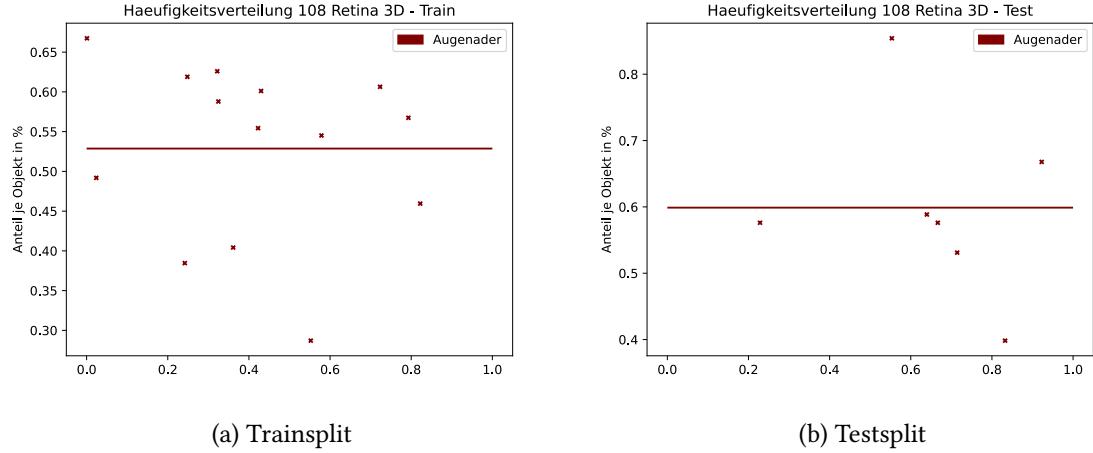


Abbildung 4.31: Anteil von Objekt (Ader) je Sample mit Durchschnitt je Split $\approx 0.55\%$

Abschließend haben wir, da wir auf dem anderen 3D-Datensatz mit CT-Aufnahmen keine besonders guten Ergebnisse erzielen konnten, einen weiteren 3D-Datensatz mit Retinae zur Verfügung gestellt bekommen [22]. Er besteht aus 21 Samples von 3D-Scans der Retina mit Segmentierungen. Diese Samples liegen sowohl in ihrer ursprünglichen, gekrümmten Form vor als auch in einer geplätteten Form, die die Krümmung der Netzhaut heraus rechnet. Wir haben uns auf die gekrümmte Version beschränkt.

Wir haben wieder die zur Verfügung gestellten .mat Dateien in Niftis konvertiert [30] und das Training gestartet. Der Objektanteil in dem Datensatz ist sowohl im Train- als auch im Testsplit im Vergleich zu dem 2D-Retina Datensatz [20] um ein Vielfaches geringer, jedoch deutlich höher als bei dem CT-Datensatz (s. Abbildung 4.31)

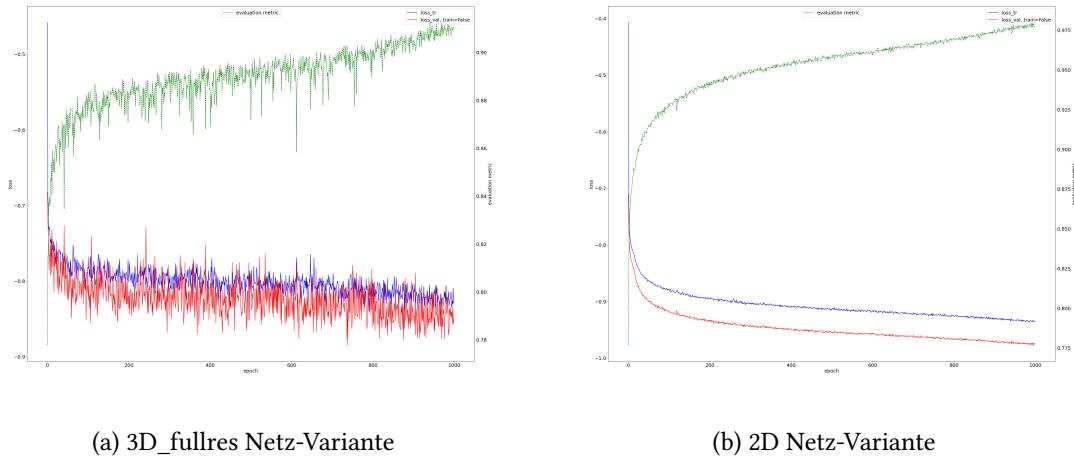


Abbildung 4.32: Progress-Graphen für Retina 3D in den Netzvarianten 3D_fullres und 2D.

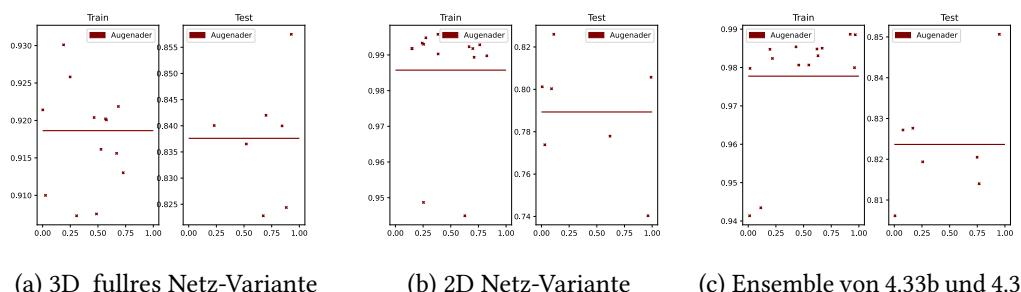


Abbildung 4.33: Scatterplot mit Dice-Koeffizienten je Sample für Retina 3D in den Netzvarianten 3D fullres, 2D und ihrem Ensemble

Es fällt auf, dass insgesamt die 2D Netz-Variante auf dem Trainsplit am besten abschneidet, während die 3D_fullres Netz-Variante auf dem Testsplit am besten abschneidet. Das Ensemble aus den beiden Varianten ist im Trainsplit schlechter als als 2D, aber besser als 3D_fullres und auf dem Testsplit besser als 2D aber schlechter als 3D_fullres (s. Abbildung 4.33).

Anhand von Abbildung 4.34 sieht man beispielhaft, dass das 3D_fullres Netz im Testsplit feine Adern ein bisschen besser erkennt als das 2D Netz. Insgesamt sieht man bei beiden Netz-Varianten und dem Ensemble keine groben Fehler.

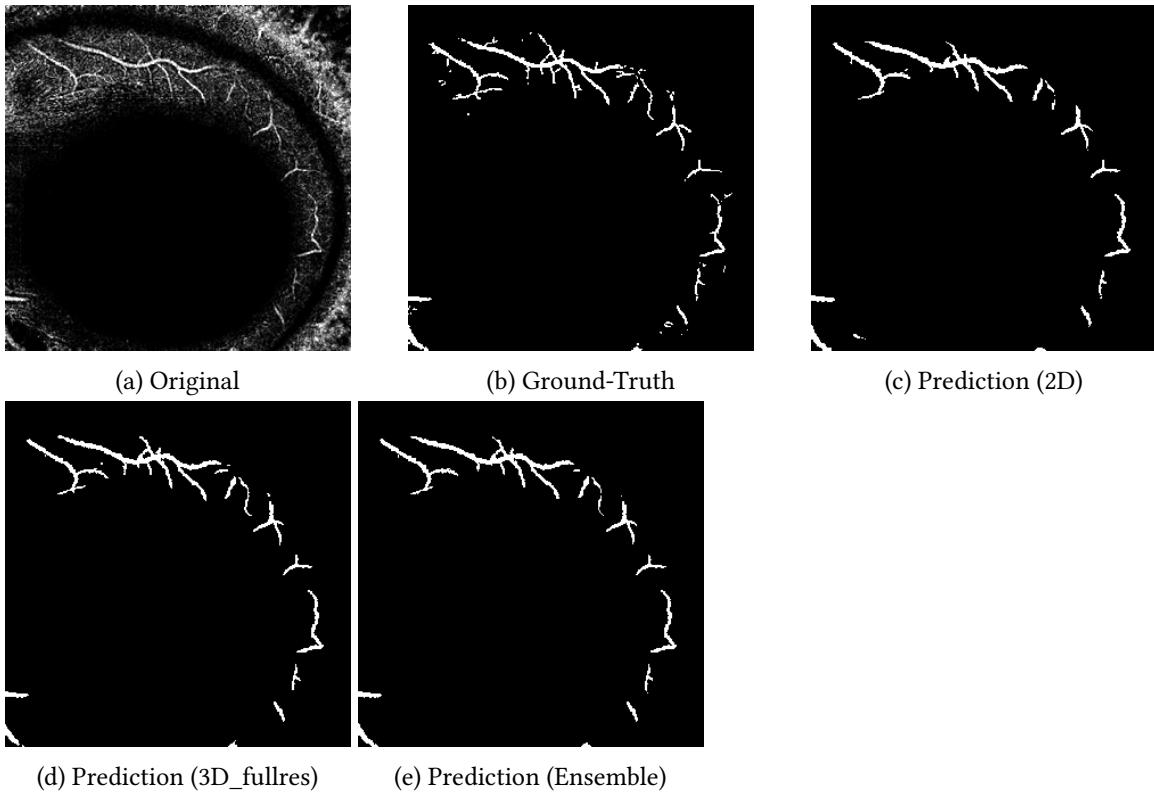


Abbildung 4.34: Visualisierung eines zufällig ausgewählten Samples aus dem Testsplit (Nummer 008) und dessen Prediction

4.3 Fazit

Für die Installation des Frameworks von nnU-Net ist eine vollständige und gut funktionierende Installationsanleitung vorhanden, welche uns den Einstieg und die Installation des Frameworks sehr erleichtert hat. Die Entwickler von nnU-Net liefern außerdem eine gute Einführung in die Arbeit mit dem Framework, in welcher sie erklären welches Dateiformat vom Framework erwartet wird und in welchen Ordnerstrukturen welche Dateien eingefügt werden müssen. Sie liefern außerdem guten Support bezüglich der Wahl wichtiger Parameter. Auch positiv aufgefallen sind uns die schnellen Antworten der Entwickler auf Fragen, welche in den Github-Issues gestellt wurden. Auch wird das Framework auf Github ständig weiterentwickelt und es gibt schnelle Bug-Fixes.

Der Code von nnU-Net ist übersichtlich gestaltet und gut kommentiert, wodurch man rasch einen guten Überblick darüber bekommt, was wo passiert. Daher lässt er sich auch gut nachvollziehen und so auch gut damit arbeiten, da eindeutig klar ist, welcher Code für was zuständig ist. Außerdem ist der Code darauf ausgelegt, allgemein und für verschiedene Benutzer zu funktionieren, sodass zum Beispiel keine Dateipfade angepasst werden müssen. Das Framework gibt gute und sinnvolle

Fehlermeldungen zurück. Der Code arbeitete für uns absolut zuverlässig und tat genau das was er sollte.

Die Einarbeitung in das Framework war daher sehr leicht umsetzbar und die Arbeit gestaltet sich sehr nutzerfreundlich und effizient.

In den zwei Papern von nnU-Net erhält man sowohl einen guten Überblick über die Grundidee und die Ziele des Papers, als auch gut verständliche Informationen über die Details der Idee und der Implementierung. Auf diese Art sind die verarbeiteten Ideen und Überlegungen gut verständlich und eindeutig.

Bezüglich der Performance hält das Framework, was es auf den Datensätzen verspricht und konnte bei uns die gleichen guten Ergebnisse erzielen wie im Paper angegeben. Das Framework arbeitet sehr gut und zuverlässig auf medizinischen Bildern, von denen bei uns alle sehr ähnliche Muster innerhalb des jeweiligen Datensatzes hatten. Die Ergebnisse auf diesen Datensätzen waren gut bis sehr gut und blieben auch dann gut, wenn das Verhältnis aus Hintergrund und Maske nicht gut war, so zum Beispiel auf dem Retina-3D Datensatz [22]. Bei einem sehr schlechten Verhältnis zwischen Maske und Hintergrund, wie zum Beispiel in unserem CT-Datensatz [21] zu den Verkalkungen in Blutgefäßen, werden die Ergebnisse dann aber auch bei diesem Framework merklich schlechter, bleiben aber aus unserer Sicht in Anbetracht der erschwerten Umstände akzeptabel, wenn auch deutlich Verbesserungswürdig. Die Ergebnisse des Frameworks auf mehreren Klassen sind jedoch nicht so gut. Auffällig ist hier, dass der Datensatz PascalVOC12 [3], welchen wir als Test für mehrere Klassen verwendet haben, deutlich unregelmäßiger in sich selbst war, als die medizinischen Datensätze. Unseren Ergebnissen zu Folge kommt das Framework entweder mit dieser Unregelmäßigkeit, oder mit den mehreren Klassen nicht so gut klar. Wie gut das Framework auf medizinischen, regelmäßigen Datensätzen mit mehreren Klassen klarkommt, könnte man in der Zukunft noch versuchen herauszufinden.

Insgesamt können wir feststellen, dass das Framework im Bereich der Segmentierung medizinischer Datensätze gute Arbeit macht. Aufgrund der positiven und immer aktuellen Aktivität der Entwickler auf Github, der großen Nutzerfreundlichkeit des Frameworks und der guten Ergebnisse erscheint eine Nutzung und Arbeit mit dem Framework für uns als sehr sinnvoll. Wegen der guten Verständlichkeit des Codes, der Grundidee und der Nutzung der verwendeten Konzepte eignet sich das Netz durchaus auch zur Weiterarbeit und Weiterentwicklung. Aufgrund des guten Verständnisses und der guten Übersicht über verwendete Ideen und Methoden fällt es leicht sich über mögliche Anpassungen Gedanken zu machen. So zum Beispiel über eine mögliche andere Kombination von Verlustfunktionen, wenn man nicht mit medizinischen Datensätzen arbeiten möchte und die Empfindlichkeit auf Hintergrundklassen anpassen möchte.

5 | Gesamtfazit

Abschließend können wir nach unserer Arbeit mit den verschiedenen Frameworks sagen, dass sehr große Unterschiede in der Nutzerfreundlichkeit festzustellen sind. Während die Benutzung von NAS-Unet extremst zeitaufwändig und unglaublich schwierig zu erarbeiten war, gestaltete sie sich bei DeepLab leichter und war bei nnU-Net extrem gut machbar. Schade ist dabei, dass uns Auto-DeepLab leider nicht zur Verfügung stand. Wir kommen hier zu dem Schluss, dass wir beim nächsten Mal eher auf den Code und die Dokumentation achten und einem Framework wie NAS-Unet nicht so viel Zeit widmen würden.

Auch im Performancebereich zeigt nnU-Net mit Abstand die besten und stabilsten Ergebnisse. Daraus, und aus den sehr guten Erfolgen von nnU-Net in den diversen Wettbewerben, folgern wir daher, dass zumindest im Bezug auf medizinische Datensätze, der Ansatz von nnU-Net, besonderen Wert auf die Hyperparameter und weniger Wert auf die Architektur zu legen, ein sehr erfolgreicher und weiterzuverfolgender Ansatz ist. Insbesondere im Vergleich zu dem Ansatz von NAS-Unet, der vor Allem Wert auf den Architektursuchraum legt, scheint er eindeutig die bessere Wahl zu sein.

Abschließend können wir sagen, dass es uns sinnvoll erscheint mit dem Ansatz und den Ideen von nnU-Net weiterzuarbeiten und sie eventuell sogar weiter zu entwickeln oder sie auf andere Anwendungsbereiche anzupassen. Letzteres zum Beispiel durch Arbeit an der Kombination der Lossfunktionen.

Literatur

- [1] *High Performance Computing PALMA II.* Adresse: <https://www.uni-muenster.de/IT-Technik/Server/HPC.html>.
- [2] Y. Weng, T. Zhou, Y. Li und X. Qiu, „NAS-Unet: Neural Architecture Search for Medical Image Segmentation,“ 2019. Adresse: https://www.researchgate.net/publication/332216927_NAS-Unet_Neural_Architecture_Search_for_Medical_Image_Segmentation.
- [3] M. Everingham, S. Eslami, L. V. Gool, C. Williams, J. Winn und A. Zisserman, *The Pascal visual object classes challenge: A retrospective*, Jan. 2015. Adresse: <http://host.robots.ox.ac.uk/pascal/VOC/voc2012/index.html>.
- [4] G. L. et al., *Evaluation of prostate segmentation algorithms for MRI: the PROMISE12 challenge*, Feb. 2014. Adresse: <http://www.sciencedirect.com/science/article/pii/S1361841513001734>.
- [5] *CHAOS-Combines (CT-MR) Healthy Abdominal Organ Segmentation*. 2019. Adresse: https://chaos.gand-challenge.org/Combined_Healthy_Abdominal_Organ_Segmentation/.
- [6] *Ultrasound Nerve Segmentation Kaggle*, 2016. Adresse: <https://www.kaggle.com/c/ultrasound-nerve-segmentation>.
- [7] *NAS-Unet, Fraunhofer, Kommentar*. Adresse: <https://gitlab.itwm.fraunhofer.de/sharad/nasunet/-/commit/872273db3153fdb8354ddf60deb73702c0fc126d>.
- [8] *NAS-U-Net Github Issue 11*. Adresse: <https://github.com/tianbaochou/NasUnet/issues/11>.
- [9] *NAS-Unet Github Repository*. Adresse: <https://github.com/tianbaochou/NasUnet>.
- [10] *NAS-U-Net Github Issue 31*. Adresse: <https://github.com/tianbaochou/NasUnet/issues/31>.
- [11] C. Liu, L.-C. Chen, F. Schroff, H. Adam, W. Hua, A. Yuille und L. Fei-Fei, „Auto-DeepLab: Hierarchical Neural Architecture Search for Semantic Image Segmentation,“ 2019. Adresse: <https://arxiv.org/pdf/1901.02985.pdf>.
- [12] H. Noh, S. Hong und B. Han, „Learning Deconvolution Network for Semantic Segmentation,“ 2015. Adresse: <https://arxiv.org/pdf/1505.04366.pdf>.

- [13] A. Newell, K. Yang und J. Deng, „Stacked Hourglass Networks for Human Pose Estimation,“ 2016. Adresse: <https://arxiv.org/pdf/1603.06937.pdf>.
- [14] D. Berth, B. Risse, T. Michels, N. Otto, X. Jiang und C. Klämbt, *An FIM-Based Long-Term In-Vial Monitoring System for Drosophila Larvae*, 2016. Adresse: <https://ieeexplore.ieee.org/document/7742314>.
- [15] *Deeplab Github Repository*. Adresse: <https://github.com/tensorflow/models/tree/master/research/deeplab>.
- [16] N. Rosenberg, *Auto-DeepLab Umsetzung Noam Rosenberg - Github Repository*. Adresse: <https://github.com/NoamRosenberg/autodeeplab>.
- [17] F. Isensee, J. Petersen, A. Klein, D. Zimmerer, P. F. Jaeger, S. Kohl, J. Wasserthal, G. Köhler, T. Norajitra, S. Wirkert und K. H. Maier-Hein, „nnU-Net: Self-adapting Framework for U-Net-Based Medical Image Segmentation,“ 2018. Adresse: <https://arxiv.org/pdf/1809.10486.pdf>.
- [18] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen und K. H. M. Hain, „Automated Design of Deep Learning Methods for Biomedical Image Segmentation,“ 2020. Adresse: <https://arxiv.org/pdf/1904.08128.pdf>.
- [19] *Kidney Tumor Segmentation Challenge 2019*. Adresse: <https://kits19.grand-challenge.org/>.
- [20] A. Budai und J. Odstrcilik, *Robust Vessel Segmentation: 2D-Retina Aufnahmen von gesunden und kranken Augen*, 2013. Adresse: <https://www5.cs.fau.de/research/data/fundus-images/>.
- [21] *CT Datensatz von Kalziumablagerungen in den Gefäßen*.
- [22] *Retina 3D Datensatz*.
- [23] *nnU-Net Github Repository - Issue zur Verwendung des Fold-Parameters*. Adresse: <https://github.com/MIC-DKFZ/nnUNet/issues/29>.
- [24] *Medical Segmentation Decathlon*. Adresse: <http://medicaldecathlon.com/>.
- [25] *nnU-Net Github Repository - Verwendung von 2D-Datensätzen*. Adresse: https://github.com/MIC-DKFZ/nnUNet/blob/master/documentation/dataset_conversion.md#how-to-use-2d-data-with-nnu-net.
- [26] *nnU-Net Github Repository - Skript zum Konvertieren von 2D-Datensätzen in Niftis*. Adresse: https://github.com/MIC-DKFZ/nnUNet/blob/master/nunet/dataset_conversion/Task120_Massachusetts_RoadSegm.py.
- [27] *nnU-Net Github Repository*. Adresse: <https://github.com/MIC-DKFZ/nnUNet>.

- [28] *nnU-Net Github Repository - Fehlermeldung zu geplanten Patches in 2D-Variante (Z. 69).* Adresse: https://github.com/MIC-DKFZ/nnUNet/blob/d396fb702dc43d73f674d2fdfb11d4782381558/nunet/experiment_planning/experiment_planner_baseline_2DUNet.py#L69.
- [29] *Diabetic Retinopathy Detection*, 2015. Adresse: <https://www.kaggle.com/c/diabetic-retinopathy-detection/data>.
- [30] M. Blunk, P. Nümann und M. Wolff, *AutoML für Segmentierung - Github Repository*. Adresse: https://github.com/pian001/AutoML_Projektseminar_Segmentierung.