Tabelle 4.1.: Überblick über die Arten der 367 ausgewählten Wäschestücke.

Art des Wäschestücks	Anzahl
Struktur 2	1
Muster: kleine Blumen	2
Muster: Quadrate	2
Stoff 1	2
Struktur 5	2
unifarben	4
Muster: große Blumen	6
Frottee	54
unregelmäßig liniert	109
Struktur 1	185

ein Bild von den drei genutzten Wäschearten mit den zugehörigen GTs. Zudem gibt Abbildung 4.2 einen Überblick über die Größe der Flecken und Löcher in den beiden Datensätzen. Abbildungen A.1 und A.2 zeigen des Weiteren, wie viele Flecken und Löcher in den Bildern jeweils vorkommen. Grundsätzlich fällt dabei auf, dass die Bilder des Trainingsdatensatzes deutlich mehr Flecken pro Bild enthalten als die des Validierungsdatensatzes. Zudem gibt es auf den Bildern des Trainingsdatensatzes deutlich mehr kleine Flecken (47,6 % der Flecken sind kleiner als 500 Pixel) als auf denen des Validierungsdatensatzes (14,9 % der Flecken sind kleiner als 500 Pixel), was auf die Stoffeigenschaften zurückgeführt werden kann.

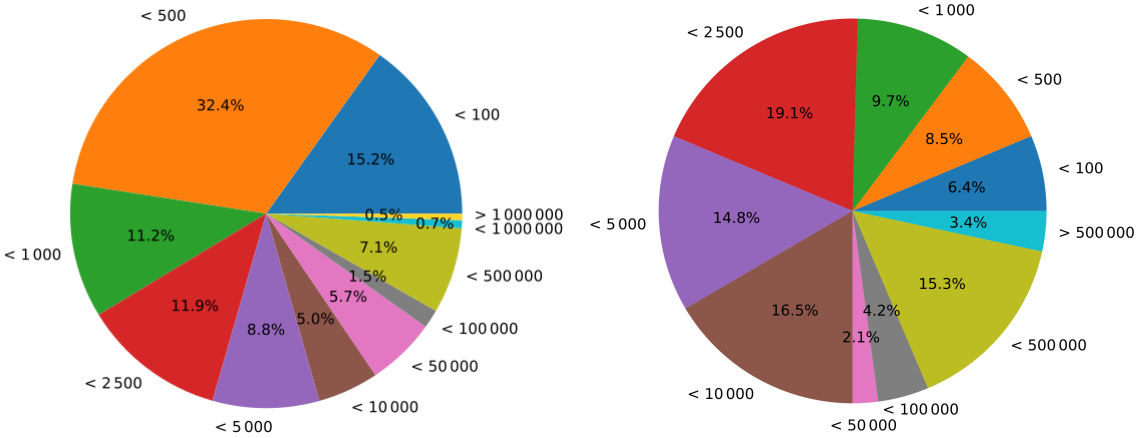
Aufgrund der geringen Größe der Datensätze wurde im Laufe des Trainings zum Teil eine Trainingsdatenerweiterung (engl. *training data augmentation*) durchgeführt. Diese umfasste ein zufälliges Rotieren um 90°, 180° oder 270° sowie ein zufälliges Spiegeln entlang der x- oder y-Achse. Eine weitere Trainingsdatenerweiterung wurde durch ein zufälliges Gaußsches Rauschen mit einem Mittelwert von 0 und einer Standardabweichung von 5 erreicht. Die Umsetzung in Caffe erfolgte mittels eines eigenen Python-Layers.

4.2. Metriken zur Evaluation salienzbasierter Verfahren

Zur Evaluation von Modellen zur Detektion salienter Objekte haben sich im Laufe der Zeit im Wesentlichen vier Metriken etabliert [5]. Diese werden genutzt, um die Übereinstimmung zwischen Modellvorhersage und Annotation zu bestimmen.

Nachfolgend werden diese Metriken beschrieben. S steht dabei für die vorhergesagte Salienzkarte, die auf $[0, 255]$ normiert wurde. Die binäre Annotation (GT) wird durch G dargestellt. $|\cdot|$ repräsentiert die Anzahl der Werte ungleich 0 in einem Binärbild.

Die ersten beiden Metriken basieren auf den überlappenden Regionen zwischen Annotation und Vorhersage. Dazu wird die Salienzkarte S zunächst in ein Binärbild M umgewandelt. Somit kann sowohl die Genauigkeit (engl. *precision*) als auch die Trefferquote (engl. *recall*)



(a) Trainingsdatensatz (b) Validierungsdatensatz

Abbildung 4.2.: Datensatzstatistiken über die Größe der Regionen, aufgeteilt nach Datensätzen. Die Label geben die obere Grenze für die Regionengröße in Pixel an. Die untere Grenze ist durch die obere Grenze der nächst kleineren Region vorgegeben, d.h. im Trainingsdatensatz haben beispielsweise 32,4 % der Flecken und Löcher eine Größe zwischen 100 und 499 Pixeln.

wie folgt berechnet werden:

$$Genauigkeit = \frac{|M \cap G|}{|M|} \tag{4.1}$$

$$Trefferquote = \frac{|M \cap G|}{|G|} \tag{4.2}$$

Zur Binarisierung der Salienzkarte S gibt es drei verschiedene Methoden. Die erste Methode nutzt einen adaptiven Schwellwert, der doppelt so hoch ist wie der Durchschnittswert der Salienzkarte S:

$$T_a = \frac{2}{W \cdot H} \sum_{x=1}^{W} \sum_{y=1}^{H} S(x, y) \tag{4.3}$$

Bei der zweiten Methode wird für jeden möglichen Schwellwert zwischen 0 und 255 aus der Salienzkarte ein Binärbild erzeugt. Für jeden Schwellwert wird sowohl die Genauigkeit als auch die Trefferquote berechnet, die dann in einem PR-Plot (engl. *Precision-Recall-Plot*) gegeneinander aufgetragen werden. Bei der dritten Methode wird der SaliencyCut-Algorithmus [24] genutzt. Initial wird dabei ein Schwellwert gewählt, der zu einer hohen Trefferquote, aber im Gegenzug zu einer schlechten Genauigkeit führt. Dieses Binärbild wird nach und nach mithilfe der GrabCut-Segmentierungsmethode [25] verbessert. Auf dem finalen Binärbild werden dann für die Genauigkeit und Trefferquote die endgültigen Werte berechnet.

Da im Regelfall weder die Genauigkeit noch die Trefferquote die Qualität des Modells ausreichend beschreibt, wird zusätzlich der F_β-Wert berechnet, der die Genauigkeit und Treffer-

quote in einer Formel gewichtet:

$$F_\beta = \frac{(1+\beta^2) \cdot Genauigkeit \cdot Trefferquote}{\beta^2 \cdot Genauigkeit + Trefferquote} \tag{4.4}$$

Bei der Evaluation zur Vorhersage salienter Objekte wird häufig ein β^2-Wert von 0,3 verwendet, um die Genauigkeit stärker zu gewichten als die Trefferquote. Bei der ersten und dritten Binarisierungs-Methode wird für jedes Bild ein F_β-Wert berechnet. Bei der zweiten Methode wird für jedes Bild bei jedem Schwellwert ein F_β-Wert berechnet, von denen der maximale Wert für das Bild ausgewählt wird. Der in allen Methoden finale F_β-Wert ist letztlich der Durchschnitt aller F_β-Werte.

Als weitere Metriken kommen die Richtig-Positiv-Rate (TPR von engl. *true positive rate*) und Falsch-Positiv-Rate (FPR von engl. *false positive rate*) in Betracht:

$$TPR = \frac{|M \cap G|}{|G|} \tag{4.5}$$

$$FPR = \frac{|M \cap \bar{G}|}{|\bar{G}|} \tag{4.6}$$

$\bar{G}$ stellt dabei das zu G komplementäre Binärbild dar. In einer sogenannten ROC-Kurve (engl. *receiver operating characteristic*) werden TPR und FPR bei verschiedenen Schwellwerten zur Binarisierung gegeneinander aufgetragen. Der AUC-Score (engl. *area under ROC curve*) ist dabei ein weiteres Maß, das die Fläche unter der ROC-Kurve angibt.

In den bisherigen Metriken werden Pixel, die richtigerweise als nichtsalient markiert werden, nicht beachtet. Diese werden jedoch in dem MAE-Score (engl. *mean absolute error*) berücksichtigt, der den durchschnittlichen absoluten Fehler zwischen der kontinuierlichen Salienzkarte, normiert auf den Bereich $[0, 1]$ ($\hat{S}$), und dem binären GT, ebenfalls normiert auf den Bereich $[0, 1]$ ($\hat{G}$), berechnet:

$$MAE = \frac{1}{W \cdot H} \sum_{x=1}^{W} \sum_{y=1}^{H} |\hat{S}(x, y) - \hat{G}(x, y)| \tag{4.7}$$

In dieser Arbeit wurden die kontinuierlichen Salienzkarten mit Schwellwerten zwischen 0 und 255 binarisiert, wobei nicht immer für alle möglichen Schwellwerte Binärbilder berechnet wurden, um die Evaluation zu beschleunigen. Meist wurde nur jeder vierte Schwellwert genutzt.

Da die Salienz-Detektion als binäres Segmentierungsproblem aufgefasst werden kann, bieten sich prinzipiell auch Metriken zur Evaluation von Segmentierungsalgorithmen wie die *intersection over union* (IoU) an [26]:

$$IoU = \frac{|M \cap G|}{|M \cup G|} \tag{4.8}$$

Diese Metrik enthält jedoch gegenüber der Betrachtung von Genauigkeit (Formel 4.1) und Trefferquote (Formel 4.2) keine zusätzliche Information, weshalb sie nicht verwendet wurde.

4.3. Spezifische Evaluationsmetrik zur Fleck- und Locherkennung

Neben der pixelweisen Evaluation wurde des Weiteren auch eine eigens entwickelte regionenbasierte Metrik verwendet, für die anwendungsrelevantere Kriterien erfüllt werden müssen. So reicht es beispielsweise aus, wenn ein Großteil des Flecks mit gewissen Toleranzen vorhergesagt wird, um ihn speziell behandeln zu können. Bei dieser Metrik wird zunächst die regionenbasierte Detektionsrate (RDR) pro Bild berechnet:

$$RDR = \frac{\text{detektierte Regionen}}{\text{detektierte Regionen} + \text{nicht detektierte Regionen}} \tag{4.9}$$

Eine im GT als salient markierte Region gilt als detektiert, wenn

1. mindestens 50 % der Fläche im vorhergesagten Binärbild als salient klassifiziert wird **und**
2. der Anteil des tatsächlichen Flecks mindestens 25 % des vorhergesagten Flecks ausmacht **oder**
3. der maximale Wert der minimalen Abstände zwischen Konturpixeln eines vorhergesagten Flecks zu Konturpixeln eines tatsächlichen Flecks unterhalb von 100 Pixeln[1] liegt.

Der Grund für die Auswahl des ersten Kriteriums ist, dass ein Großteil des Flecks vorhergesagt werden soll. Gleichzeitig sollen die vorhergesagten Flecken jedoch auch nicht zu viele makellose Bereiche beinhalten, was durch das zweite und dritte Kriterium ausgeschlossen wird. Grundsätzlich ist ein zu groß vorhergesagter Fleck in der Anwendung weniger tragisch als ein nicht detektierter Fleck, weshalb das als zweite Kriterium genutzte relative Maß mit 25 % vergleichsweise schwach gewählt wurde. Da Vorhersagen von insbesondere schmalen, länglichen Makeln wie Haaren oder Rissen dieses zweite Kriterium häufig nicht erfüllen (vgl. Abbildung 4.3c), obwohl der Makel gut vorhergesagt wurde, wurde mit dem dritten Kriterium ein weiteres, absolutes Maß eingeführt.

Die RDR wird für jedes Bild für alle gewählten Schwellwerte zwischen 0 und 255 berechnet. Final werden die Detektionsraten über alle Bilder gemittelt, so dass pro gewähltem Schwellwert eine Gesamt-RDR berechnet wird. In einem Plot wird die Gesamt-Detektionsrate gegen die Schwellwerte aufgetragen. Ein zusätzliches Balkendiagramm wurde genutzt, um die maximal erreichte RDR abhängig von der Regionengröße darzustellen.

Regionen im GT, deren Fläche kleiner als 100 Pixel groß ist, werden von dieser Metrik ignoriert, da es sich um Artefakte in der Annotation oder zunächst irrelevante Flecken (z.B. sehr kleine Haare) handelt.

Abbildung 4.3 veranschaulicht diese Metrik durch Beispiele.

[1] 100 Pixel entsprechen auf den in der Arbeit genutzten Bildern etwa 1 cm.

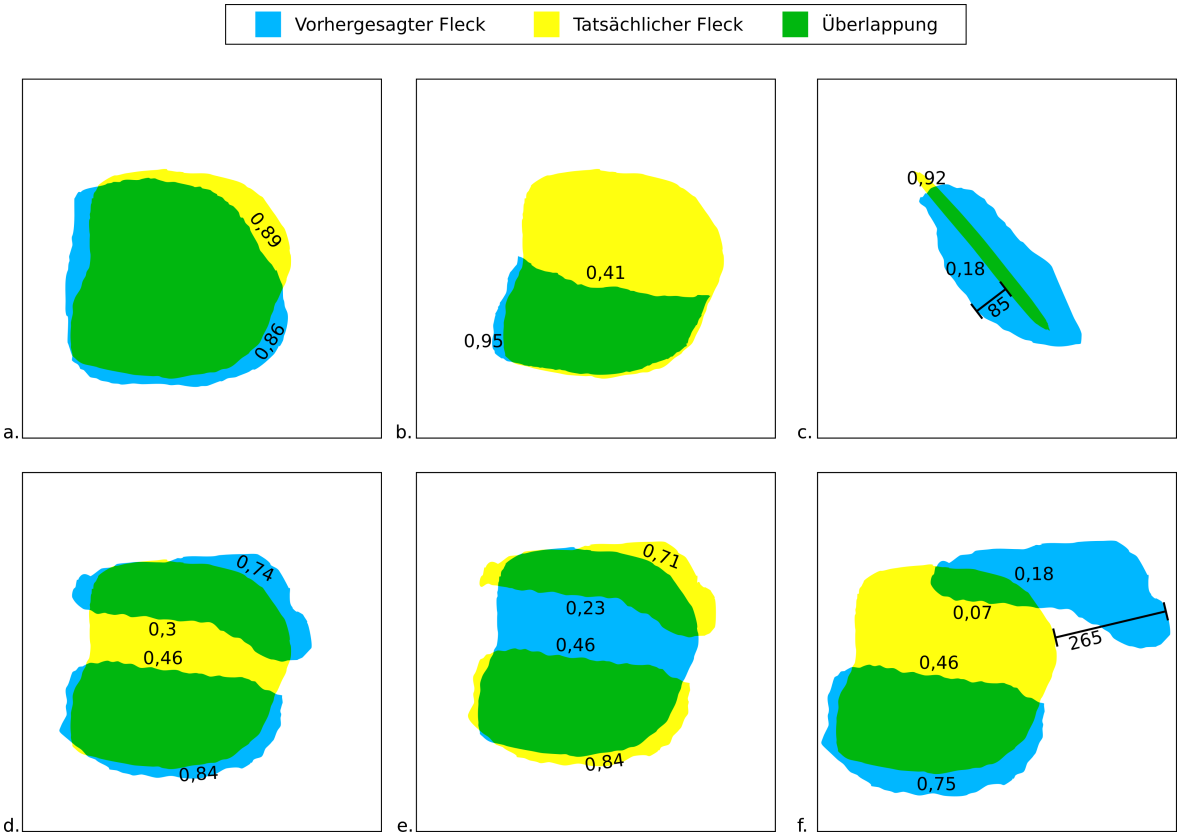


Abbildung 4.3.: Beispiele für die regionenbasierte Evaluationsmetrik. Die Dezimalzahlen geben den relativen Anteil der zugehörigen grünen Fläche (Überlappung) zur jeweiligen Region an. Die ganzzahligen Werte geben den maximalen Wert der minimalen Abstände zwischen Konturpixeln eines vorhergesagten Flecks zu Konturpixeln eines tatsächlichen Flecks an. **a.** Der tatsächliche Fleck wird als detektiert gewertet, da 89 % des Flecks vorhergesagt wurden und der Fleck gleichzeitig 86 % der Vorhersage ausmacht und somit über dem Schwellwert (25 %) liegt. **b.** Der Fleck wird nicht als detektiert gelten, da nur 41 % des Flecks vorhergesagt wurden. **c.** 92 % des Flecks werden richtig vorhergesagt. Da der annotierte Fleck nur 18 % der Vorhersage ausmacht, ist das zweite Kriterium jedoch nicht erfüllt. Der maximale Abstand von den Konturpixeln liegt mit 85 Pixeln allerdings unterhalb des Schwellwerts, weshalb der tatsächliche Fleck als detektiert gewertet wird. **d.** Der tatsächliche Fleck wird als detektiert gewertet, da die Vorhersage 30 % + 46 % = 76 % des Flecks ausmacht. Gleichzeitig liegt keine der detektierten Regionen deutlich außerhalb des Flecks, d.h. beide relativen Anteile (74 % und 84 %) liegen über dem Schwellwert. **e.** Beide tatsächlichen Flecken gelten als detektiert, da 71 % bzw. 84 % der Flecken vorhergesagt wurden. Gleichzeitig ist 69 % der Vohersage tatsächlich Fleck. Wäre der untere tatsächliche Fleck nicht vorhanden, würde der obere tatsächliche Fleck nicht als detektiert gewertet werden, da er nur 23 % des vorhergesagten Flecks ausmacht. **f.** Auch wenn 53 % des Flecks durch 2 Flecken vorhersagt werden, wird der tatsächliche Fleck als nicht detektiert gewertet, da nur 18 % eines vorhergesagten Flecks tatsächlich Fleck ist und der maximale Abstand der Konturpixel 265 Pixel beträgt.