Computational Photography

Unterstützung der digitalen Fotografie durch künstliche Intelligenz

Oliver Hummel | oh021@hdm-stuttgart.de
Hochschule der Medien Stuttgart | Wintersemester 2018/19

Einführung

Der Begriff Computational Photography bezeichnet an sich nur die Unterstützung der Fotografie durch Computer aller Art. Streng genommen zählen also auch einfache Digitalkameras dazu, da hier das Bild auch durch digitale Prozesse erzeugt wird. Die einzige klare Abtrennung besteht zu vollständig analoger Fotografie, bei der das Bild auf chemischem Weg entsteht. Da laut dieser klassischen Definition kein Unterschied zu digitaler Fotografie besteht, hat sich die Definition etwas gewandelt und legt nun stärker Wert auf die Rolle des Computers bei der Bildkomposition sowie der nachträglichen Bearbeitung.

Dies bietet für die Nutzer der Technologie auch den größten Vorteil gegenüber der klassischen digitalen Fotografie. Durch Algorithmen, die unter anderem aus dem Bereich von Machine Learning stammen, können Bilder in vielerlei Hinsicht automatisiert verbessert werden. Für die Endverbraucher sind Kameras mit solchen Fähigkeiten attraktiv, da langwierige Bildbearbeitung entfällt und sofort ein nahezu optimales Bild entsteht. Da viele Endverbraucher im Umgang mit Fotografie und Bildbearbeitung auch nicht sehr vertraut sind, erhalten sie mit diesen Kameras ein subjektiv besseres Ergebnis als z.B. mit einer hochwertigen Spiegelreflexkamera. Die Auflösung sowie das Linsensystem einer Spiegelreflexkamera ist einem Smartphone natürlich weit überlegen, doch in der subjektiven Beurteilung eines Fotos spielt dies bei einem Großteil der Nutzer kaum eine Rolle. Kriterien wie die Sättigung und Dynamik des Bildes sind weitaus wichtiger. Gründe wie dieser sind einer der Hauptursachen warum Computational Photography mittlerweile bei Smartphone Herstellern so hoch auf der Agenda steht und auch aktiv beworben wird.

Beispiele

HDR

High Dynamic Range (HDR) Fotografie wird oft als Beispiel für Computational Fotografie genannt. Dabei wird versucht, eine Szene mit großen Helligkeitsunterschieden in einem einzigen Bild einzufangen. Mit nur einer Aufnahme einer Kamera ist das oft unmöglich, da die Blendeneinstellung sowie die Belichtungszeit auf einen Helligkeitsbereich festgelegt werden müssen. Bereiche in der Szene die deutlich heller oder dunkler sind, erscheinen überbelichtet oder zu abgeschattet in der Aufnahme. Die Lösung liegt in der Kombination mehrerer Aufnahmen. Für jeden Helligkeitsbereich erscheint das Bild dadurch scharf und detailreich.

Abbildung 1: Links überstrahlt der Hintergrund die Aufnahme. Die HDR Aufnahme rechts zeigt alle Bildbereiche.[1]

Im Bereich der mobilen Fotografie hat Google hier seit der Einführung ihrer Pixel Smartphones Schlagzeilen gemacht. Der HDR+ Modus deren Kameraanwendung verbessert Aufnahmen bei schwachem Licht und erzeugt bei Tageslicht Aufnahmen mit kraftvollen Farben. Auch hier werden mehrere Aufnahmen kombiniert. Allerdings werden keine unterschiedlichen Belichtungszeiten verwendet, sondern alle Aufnahmen mit einer kurzen Belichtungszeit gemacht. Dadurch werden helle Bereiche nicht überblendet. Dunkle Bereiche werden nachträglich digital aufgehellt. Dies würde im Normalfall zu Bildrauschen führen. Da dies aber für alle Aufnahmen einer Serie durchgeführt wird, verschwindet mögliches Bildrauschen sobald der Durchschnitt über alle Aufnahmen gebildet wird.

Night Sight

Bei Night Sight[2] handelt es sich um eine Erweiterung des HDR+ Algorithmus, welche nochmals deutlich hellere Bilder bei Dunkelheit ermöglicht. Dazu werden die Aufnahmen deutlich stärker künstlich aufgehellt als bei HDR+. In der Folge können selbst in sehr dunklen Bildbereichen noch einzelne Details ausgemacht werden. Ein Problem stellt allerdings der automatische Weißabgleich bei Dunkelheit dar. Dieser versucht, eine realistische Repräsentation der Farben einer Aufnahme sicherzustellen. Bei schwacher Beleuchtung gestaltet sich dies allerdings als schwierig.

Daher wird bei Night Sight ein auf Machine Learning basierender Algorithmus für den Weißabgleich eingesetzt. Er bekam als Trainingsdaten Bilder, die einen korrekt durchgeführten Weißabgleich zeigen, sowie Bilder, bei denen der Weißabgleich zu einer unnatürlich aussehenden Aufnahme geführt hat. Dadurch konnte der Algorithmus lernen, welche Farbwahl eine Aufnahme haben muss um für einen menschlichen Betrachter natürlich zu wirken.

Portrait Mode

Hierbei handelt es sich um Aufnahmen mit geringer Schärfentiefe und großer Blendenöffnung. Dadurch können Personen oder Objekte in Szene gesetzt werden, in dem der Hintergrund unscharf gehalten wird. Dies erfordert normalerweise ein größeres Objektiv als in einem Smartphone verbaut werden kann. Daher führen Smartphones eine Unterscheidung des Bildes in Vorder- und Hintergrund durch und wenden einen Weichzeichner auf den Hintergrund an, um künstlich für Unschärfe zu sorgen.

Diese Einteilung kann entweder auf Basis der Parallaxe erfolgen, wenn das Smartphone zwei Kameras besitzt, oder durch gelernte Modelle. Dazu wurde ein Machine Learning Algorithmus trainiert, welcher aus einem zweidimensionalen Bild eine Tiefenkarte erzeugt. Die Trainingsdaten kamen bei Googles

[1] https://ai.googleblog.com/2014/10/hdr-low-light-and-high-dynamic-range.html [Zugriff am 10.02.2019]
[2] https://ai.googleblog.com/2018/11/night-sight-seeing-in-dark-on-pixel.html [Zugriff am 10.02.2019]

Implementierung von einer Kombination aus fünf Smartphones. Aus den perspektivischen Unterschieden dieser fünf Blickwinkel kann mithilfe eines Algorithmus eine Tiefenkarte berechnet werden. Diese geht in Kombination mit einem der Ausgangsbilder in ein neuronales Netz ein. Das gelernte Modell reicht dann aus, um auf Bildern die Tiefeninformation zu approximieren.

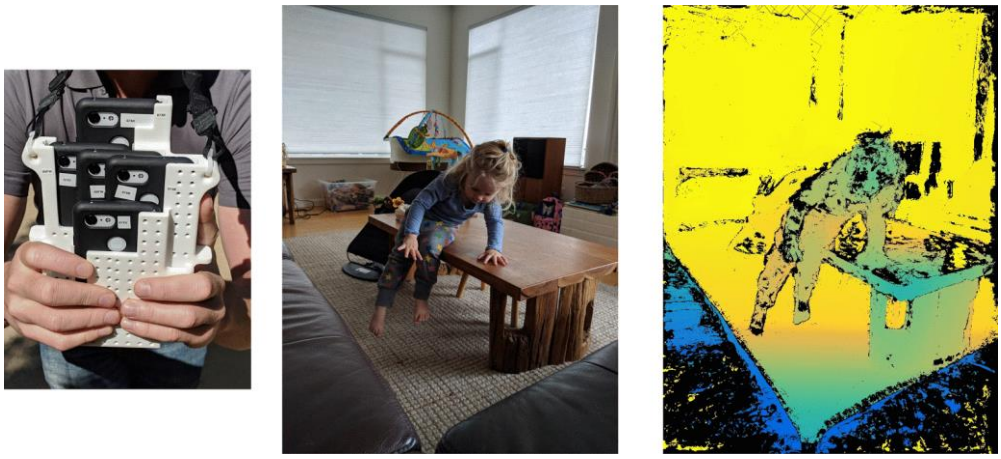

Abbildung 2: Kamerakonstruktion, Einzelaufnahme und berechnete Tiefenkarte (von links)[3]

Convolutional Neural Networks

Die Analyse von Bildern durch neuronale Netze erfolgt meist durch Convolutional Neural Networks (CNNs). Sie können auch große Bilder effizient analysieren und Objekte und deren Position darin erkennen. In tiefen neuronalen Netzen gibt es mehrere Schichten mit Neuronen, die untereinander verbunden sind. Die Neuronen einer Schicht bestimmen für alle Neuronen der folgenden Schicht deren Ausgabewert. Aus anfänglichen Schichten mit vielen Neuronen wird deren Anzahl mit fortschreitender Tiefe immer geringer. Am Ausgang des Netzes sind dann nur noch einzelne Neuronen vorhanden, welche direkt die Wahrscheinlichkeit oder Position unterschiedlicher Features wiederspiegeln.

CNNs können gut mit Bildern umgehen, da zusätzlich zu regulären Schichten noch Convolution und Pooling Schichten eingesetzt werden. Im Convolution Layer wird ein Bereich von Pixeln (üblicherweise 3x3 oder 5x5) mit einem Filter verrechnet, der die gleiche Größe besitzt. Das so entstandene Skalarprodukt fasst nun diesen Bereich als einen Wert zusammen. Nach der Berechnung wird der Filter im Normalfall um einen Pixel verschoben, so dass er teilweise einen neuen Wertebereich umfasst. Damit das Bild dabei nicht schrumpft wird bei einem 3x3 Filter ein Padding mit einer Breite von einem Pixel um das Bild herum ergänzt. Bestimmte Filter können so z.B. Kanten erkennen.

Die zweite Schicht die speziell bei CNNs zum Einsatz kommt, sind Pooling Layer. Diese verkleinern aktiv die Größe des Bildes und helfen dabei, auch größere Bilder zu analysieren ohne dass die Berechnung der Ausgabeneuronen zu rechenintensiv wird. Pooling Layer kombinieren Bildbereiche zu einem Wert. Dabei gibt es keine Überlappungen wodurch die Größe verringert wird. Eine gebräuchliche Art des Poolings ist das Max Pooling, bei dem nur der größte Wert eines Bereichs übernommen wird.

[3] https://ai.googleblog.com/2018/11/learning-to-predict-depth-on-pixel-3.html [Zugriff am 10.02.2019]

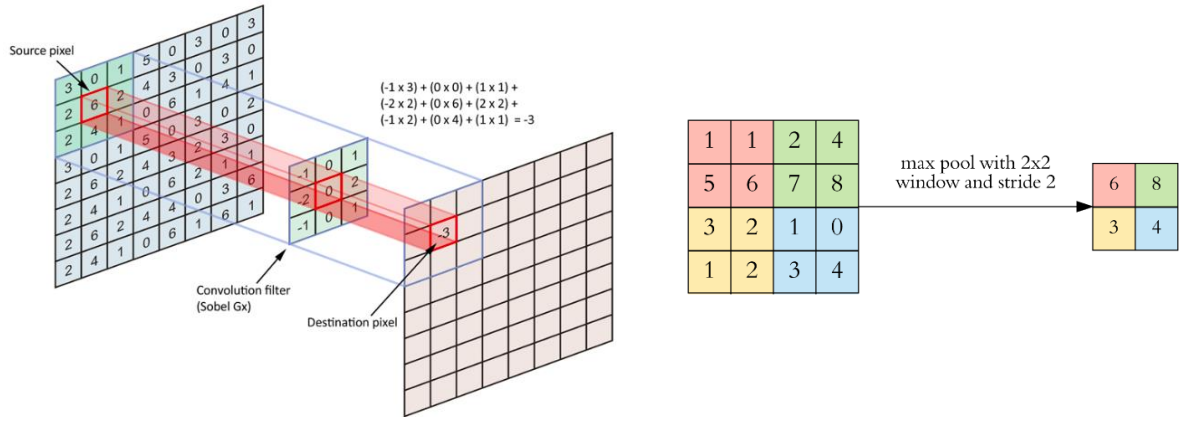

Abbildung 3: Convolution Layer[4] (links) und Pooling Layer[5] (rechts)

In CNNs wird im Normalfall immer ein Convolution und ein Pooling Layer hintereinandergeschaltet. Diese Kombination wird mehrmals wiederholt. Dadurch nimmt die Anzahl der Pixel eines Bildes stark ab, bevor sie in vollverbundene Schichten des neuronalen Netzes einfließen, um schlussendlich eine Ausgabe zu generieren.

Einsatz von Künstlicher Intelligenz

Künstliche Intelligenz in Form von CNNs kann aber nicht nur für einfache Bildbearbeitung genutzt werden, sondern ermöglicht auch die Ergänzung von Details in Bildern, die davor noch gar nicht existierten. Dazu muss das Bild oftmals erst analysiert werden um Features wie Gesichter, Personen oder Schriftzüge zu erkennen. Sobald die KI diese Informationen erkannt hat, kann sie diese Features beliebig ergänzen, transponieren oder komplett austauschen.

Gesichtserkennung

Die Erkennung von Gesichtern in Bildern oder Videos markiert oftmals den Einstieg für komplexe Bildtransformationen. Aber auch zum automatischen Fokussieren kann sie verwendet werden. Wenn ausreichend Trainingsdaten vorhanden sind, ist dies für KI mittlerweile eine einfache Aufgabe. Die Trainingsdaten bestehen aus Bildern mit Gesichtern (meistens als Graustufenbild) sowie den dazugehörigen Koordinaten relevanter Merkmale wie Augen, Nase und Mund. Ein CNN wie es oben beschrieben wurde, bekommt als Eingabe ein Bild und gibt die Koordinaten der Gesichtsmerkmale aus.

Bildvergrößerung

Um bei großen Ausdrucken wie Plakaten eine ausreichende Bildqualität sicherzustellen, muss das Originalbild eine sehr hohe Auflösung aufweisen. Die Kameras aktueller Smartphones liefern oft eine Auflösung im Bereich von 12 bis 16 Megapixeln. Deren Aufnahmen sind zwar scharf genug für die Betrachtung auf den dafür üblichen Medien, aber bei Ausdrucken von der Größe mehrerer Quadratmeter wird etwas an Schärfe eingebüßt. Manchmal möchte man auch einen Ausschnitt eines Bildes vergrößern, was ebenfalls eine ausreichende Ursprungsgröße erfordert.

Herkömmlicherweise werden Bilder durch Bikubische Interpolation vergrößert. Dabei wird versucht aus benachbarten Pixeln die Farbwerte für neue Pixel zu berechnen. Je nach Algorithmus können damit schon gute Ergebnisse erzielt werden, aber ab einem gewissen Vergrößerungsfaktor sehen die

[4] https://medium.freecodecamp.org/an-intuitive-guide-to-convolutional-neural-networks-260c2de0a050 [Zugriff am 10.02.2019]
[5] https://towardsdatascience.com/applied-deep-learning-part-4-convolutional-neural-networks-584bc134c1e2 [Zugriff am 10.02.2019]

erzeugten Bilder nicht mehr natürlich aus. Dieses Verfahren vergrößert zwar die Anzahl der Pixel, kann aber keine neuen Informationen in das Bild einfließen lassen.

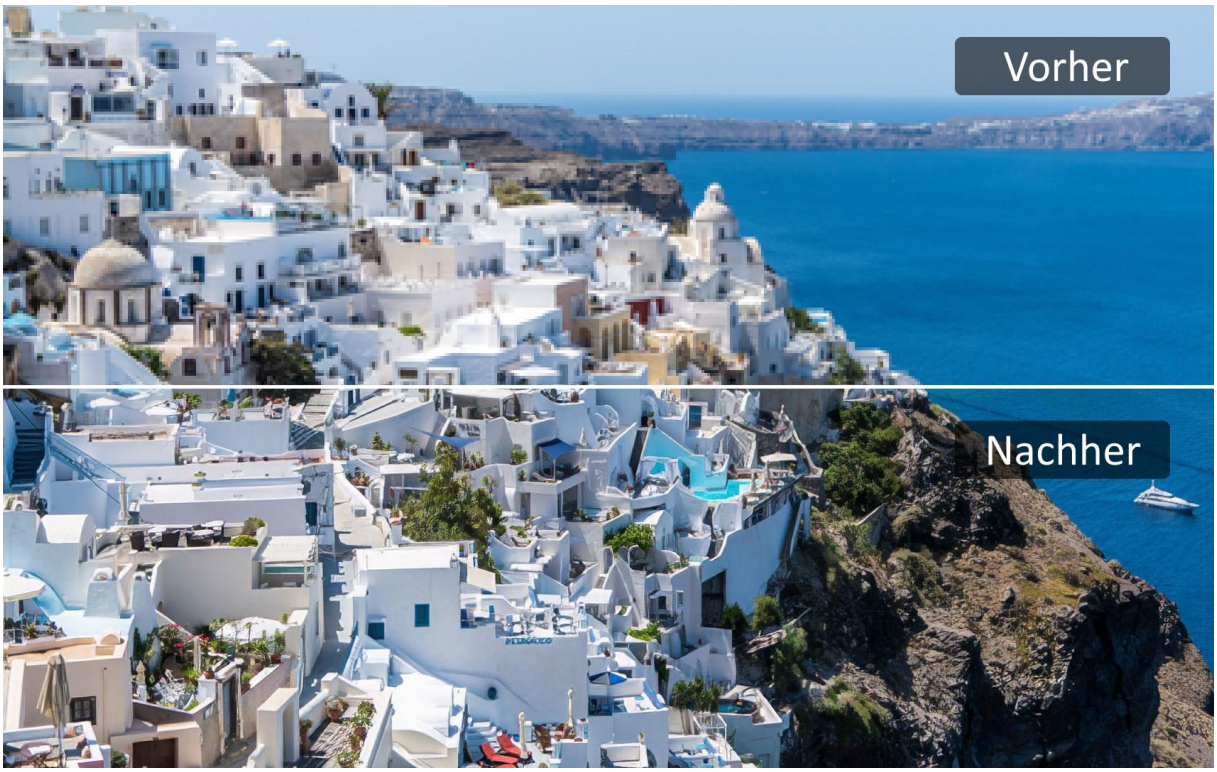

Abbildung 4: Bildvergrößerung durch KI Verfahren[6]

Hier sind KI Verfahren im Vorteil. Sie haben gelernt, was in einem Bild zu sehen ist und wie diese Objekte auszusehen haben. Wird ein Bild, welches Text enthält, klassisch vergrößert sind die Kanten der Buchstaben irgendwann verwaschen. Die KI allerdings weiß um welche Buchstaben es sich handelt und wie diese auszusehen haben. Für das vergrößerte Bild erzeugt die KI die Buchstaben in höherer Auflösung und setzt diese einfach in das Bild ein. Auch Bilder von Objekten wie einem Baum kann die KI so vergrößern. Sie hat gelernt, dass ein Baum Blätter hat und kann diese im fertigen Bild ergänzen. Die KI erzeugt also Details in Bildern, die gar nicht existieren. Aber für einen Betrachter der dies nicht weiß, bleibt der Eindruck einer stimmigen Gesamtaufnahme erhalten.

Bildergänzung und Objektentfernung

Bei der Bildvergrößerung hatte die KI bereits ein Konzept von Objekten in Bildern. Wenn man dieses Verfahren nun auf die Spitze treibt und ganze Bildbereiche entfernt, ist die KI gezwungen die Bereiche mit Informationen zu ergänzen, die sie gar nicht wissen kann. Allein auf der Basis der umgebenden Objekte muss die Lücke also sinnvoll gefüllt werden. Im Umkehrschluss können dadurch auch Objekte aus Bildern entfernt werden. Hat man ein Bild einer Landschaft mit einer einzelnen Person, kann diese großzügig ausgeschnitten werden. Die KI versucht nun dieses Loch zu füllen. Da sie nur sieht, dass es sich um ein Bild einer Landschaft handelt, versucht sie den Rest der Landschaft zu ergänzen. Die Person wurde entfernt, da die KI aufgrund des Gesamtbilds darauf schließt, dass sich an dieser Stelle keine Person befindet.

[6] https://letsenhance.io/ [Zugriff am 10.02.2019]

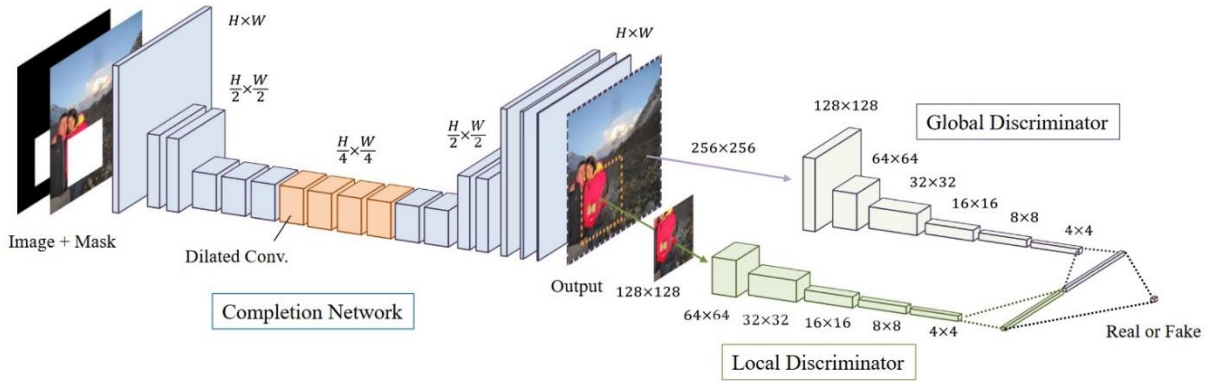

Abbildung 5: Drei neuronale Netze werden für die realistische Bildergänzung benötigt[7]

Forscher der Waseda Universität in Japan haben dieses Konzept noch verbessert, indem sie noch zwei weitere neuronale Netze trainiert haben. Diese schauen sich das Ergebnis der Bildergänzung an und entscheiden aufgrund gelernter Modelle, ob es sich um ein echtes oder generiertes Bild handeln könnte. Dabei betrachtet ein Netz (Local Discriminator) nur den generierten Bereich und überprüft diesen auf Fehler. Das andere Netz (Global Discriminator) entscheidet ob der generierte Bereich in das Gesamtbild passt. Für die finale Entscheidung werden die Ergebnisse dieser beiden Netze kombiniert. Bekommt das Bildergänzungsnetzwerk (Completion Network) das Signal, dass ein Bild als unecht eingestuft wurde, wagt es einen neuen Versuch mit leicht geänderten Gewichten innerhalb des Netzes. Mit der Zeit entstehen somit immer bessere Ergebnisse.

Fazit

Computational Photography setzt nicht unbedingt Machine Learning ein, kann dadurch aber in einigen Anwendungsfällen sehr viel bessere Ergebnisse erzielen als klassische Algorithmen. Im Zeitalter der mobilen Fotografie wird von Smartphones eine sehr hohe Bildqualität erwartet, die schon fast mit professioneller Kameraausrüstung vergleichbar ist. In Zukunft wird KI vermutlich noch sehr viel stärker eingesetzt werden. Es ist durchaus möglich, dass der Sensor einer Smartphone Kamera nur noch den groben Input liefert aber die komplette Ausarbeitung des Bildes durch neuronale Netze erfolgt, die das Ergebnis so tunen, dass es am ehesten der Vorstellung eines perfekten Bildes für den Benutzer entspricht.

Allerdings bietet KI auch ungeahnte Möglichkeiten zur Manipulation von Bildern und Videos. Schon heute werden diese oft nur mangelhaft geprüft und direkt als Nachrichten weitergegeben. Da KI schon bald perfekte Fälschungen liefern könnte, die auch andere künstliche Intelligenzen nicht mehr als solche entlarven können, müssen wir uns Gedanken darüber machen inwiefern wir Bildern und Videos noch trauen können. In einer Zeit in der Fake News eine immer größere Rolle spielen, kann gefälschtes Bildmaterial schnell zu Propaganda oder zur Denunzierung politischer Gegner eingesetzt werden. Dies muss uns immer bewusst sein.

[7] S. Iizuka, E. Simo-Serra, H. Ishikawa. Globally and Locally Consistent Image Completion. 2017. http://hi.cs.waseda.ac.jp/~iizuka/projects/completion/data/completion_sig2017.pdf [Zugriff am 10.02.2019]