Wenn aus Worten **Bilder werden**

TEXT Prof. Dr. Martin Steinebach

Ein paar Begriffe eintippen, kurz warten und schon entsteht ein Bild, das nie eine Kamera aufgenommen hat: Text-to-Image-Verfahren stellen die Mediensicherheit vor neue Herausforderungen. Ein Bildforensiker erklärt, was die neue Technologie kann und was – noch – nicht.

Täuschend echt und schwer zu stoppen: Deepfakes stehen nicht erst seit dem Ukraine-Krieg im Mittelpunkt vieler Diskussionen rund um gezielte Desinformation. Wer die Technologie beherrscht, kann vorhandene Medien so verändern, dass sie neue Botschaften senden – oder einzelnen Personen nie getätigte Aussagen in den Mund legen.

Mit Text-to-Image-Verfahren wie Stable Diffusion, Midjourney oder Dall-E ist eine neue Art von Werkzeugen entstanden, deren Möglichkeiten perspektivisch sogar noch weiter reichen. Hier erzeugt eine Maschine Bilder, deren Inhalt von einem Text vorgegeben wird. Ermöglicht wird dies durch umfangreiche Trainingsdaten, anhand der die Maschine lernen konnte, welche Bilder mit welchen Worten beschrieben werden. Das birgt das Risiko, dass an einem gewissen Punkt jede und jeder nach individuellen Vorstellungen Inhalte erzeugen und diese missbräuchlich nutzen kann – etwa als Beleg für eine Falschmeldung.

Wer die derzeit verfügbaren Lösungen testet, merkt schnell, dass sie aus Datensätzen gelernt haben, die breitflächig aus dem Internet gewonnen wurden – und daher nicht immer die Realität widerspiegeln. Das führt dazu, dass die Inhalte geschlechtlich nicht ausgewogen sind: Den Begriff „Flugbegleiter“ stellen Text-to-Image-Verfahren fast ausschließlich durch weibliche Personen dar. Keine Überraschung: Sucht man im Internet nach Bildern zu demselben Begriff, ist der Frauenanteil ebenfalls hoch. Aber auch tiefer gehende Phänomene sind zu erkennen. Lässt man eine weibliche Person ohne nähere Angaben zur Kleidung zeichnen, besteht die Wahrscheinlichkeit, dass das Bild eine nackte Person zeigt. Selbst bei anderslautenden Vorgaben versäumt es das Verfahren teilweise, Kleidung zu zeichnen.

Beobachten lässt sich dieses Phänomen allerdings nur dann, wenn der Zugang zum Verfahren unzensiert ist. Der Großteil der offen verfügbaren Methoden unterbindet bereits Texteingaben, die zu Nacktheit führen oder Personen des aktuellen öffentlichen Lebens betreffen. Auch die erzeugten Bilder werden noch einmal überprüft, weil ihr Inhalt selbst bei unverfänglichen Angaben gegen selbst gesteckte Richtlinien verstoßen

Neuschwanstein in Flammen: Eine Bildfälschung ist schnell erstellt. Ein frei verfügbares Foto des Bauwerks dient als Vorlage, um eine Explosion vorzutäuschen, die es niemals gab.

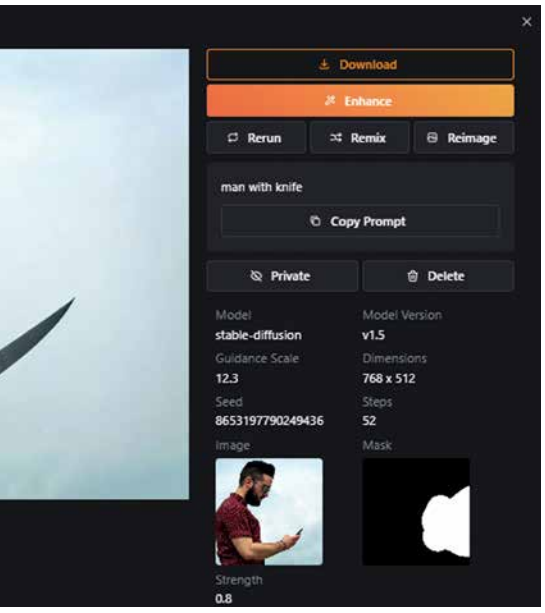
kann. Da die Methoden teilweise als Open Source zur Verfügung stehen, können Menschen mit der nötigen Expertise solche Schutzmaßnahmen jedoch recht einfach umgehen. Auch stehen entsprechend angepasste Varianten bereits als Onlinedienste zur Verfügung. Zudem ist es möglich, spezialisierte Bild-Generatoren nachzutrainieren. Obwohl die Technologie erst wenige Monate zur Verfügung steht, haben sich schon Gruppen gebildet, die sich auf einzelne Bereiche wie das Erzeugen japanischer Manga-Zeichnungen spezialisiert haben.

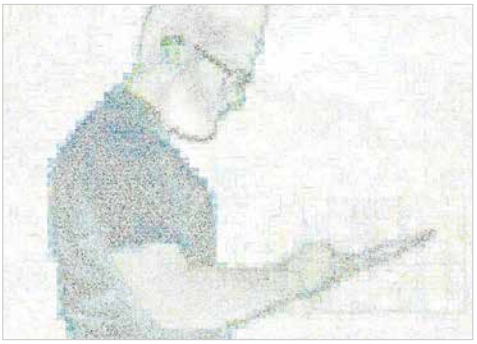
Vermutlich wird es noch dauern, bis diese Verfahren auch für Desinformationen verwendet werden. Der Einsatz der oben erwähnten Deepfakes war schon vor zahlreichen früheren politischen Ereignissen befürchtet worden und hat doch länger auf sich warten lassen. Viele unterschätzen den Aufwand, um wirklich glaubwürdige hochwertige Inhalte zu erstellen. Dies gilt auch für Text-to-Image-Verfahren: Bilder mit gewünschten Inhalten sind schnell erstellt, wenn diese einzelne Orte, Gegenstände oder Personen betreffen. Interaktionen glaubwürdig zu erzeugen ist deutlich schwieriger. Bilder von Politikerinnen oder Politikern zu erstellen ist einfach, wenn diese zum Zeitpunkt der Erfassung der Trainingsdaten bereits bekannt waren. Vorzugeben, dass sich zwei Prominente die Hände schütteln, gelingt den Verfahren meist noch nicht so gut. Ein Foto zu erzeugen, das zwei bekannte Persönlichkeiten in einer Tiefgarage bei der Übergabe eines Geldkoffers zeigt, ist derzeit mit Texteingaben unrealistisch.

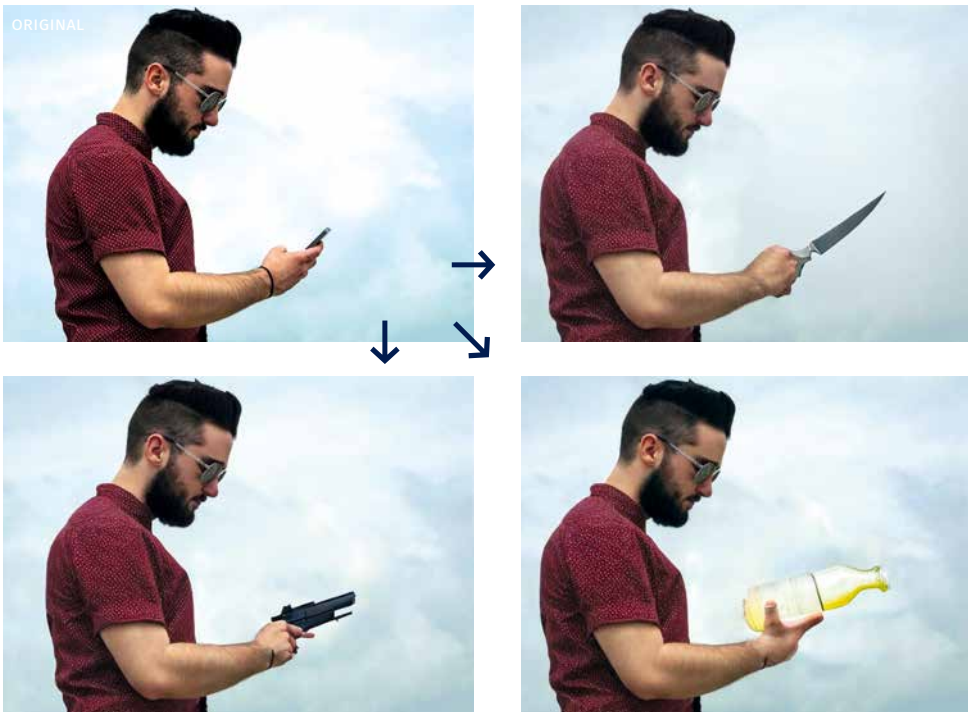
Das sogenannte Inpainting erleichtert diese Aufgabe aber bereits deutlich: Dabei nimmt das Verfahren an einem bestehenden Foto durch Texteingabe und eine Maske die gewünschten Veränderungen vor. So lässt sich etwa der Aufbau des Bildes konkret vorgeben und nur die Köpfe der gezeigten Personen werden neu gezeichnet. Das Ausgangsfoto lässt sich ganz einfach mit Statistinnen oder Statisten stellen und dann per Kommando verändern.

Zu viele Finger als Erkennungsmerkmal

Wie bei Bildmanipulationen mit Werkzeugen wie Photoshop und Deepfakes stellt sich auch bei Text-to-Image-Verfahren die Frage, wie sich diese erkennen lassen. Die Erfahrung zeigt, dass zum Anfang einer technischen Entwicklung oft das menschliche Auge ausreicht, um Fälschungen zu entlarven. Da die Verfahren aber schnell reifen und anfängliche Fehler behoben werden, ist dieser Ansatz wahrscheinlich nur eine kurz- bis mittelfristige Lösung. Bei Deepfakes gab es anfangs immer wieder Fehler in der Erkennung und beim Ersetzen eines Gesichts, inzwischen sind die Ergebnisse deutlich besser und daher schwerer zu erkennen. Text-to-Image-Verfahren neigen derzeit noch dazu, Fehler beim Erstellen der Bilder zu machen und beispielsweise zu viele Finger an einer Hand zu zeichnen oder Objekte ineinander übergehen zu lassen. Anhand von solchen Fehlern ist es leicht, das Bild als synthetisch zu erkennen. Außerdem weisen die Bilder manchmal noch verschwommene Bildbereiche auf, die bei einer aufmerksamen Betrachtung entdeckt werden können. Dies gilt vor allem bei fotorealistischen Darstel-







LINKS Ein Bild, drei Varianten: oben das Original, darunter eine einfache händische Retusche, um die Tätowierung am rechten Arm zu entfernen. Ganz unten wurde der Kopf der Frau für Inpainting mittels Stable Diffusion/mage.space markiert. Die Aufforderung lautete schlicht: Merkel.

RECHTS Das Fälschen ist mit einem Werkzeug wie mage.space oder einer Cloud-Lösung auf Basis von Stable Diffusion einfach. Das Originalbild wird hochgeladen, der zu verändernde Bereich markiert und im Text eingegeben, was gezeigt werden soll: Je nach Objekt wird auch die Handhaltung anders ausgespielt.

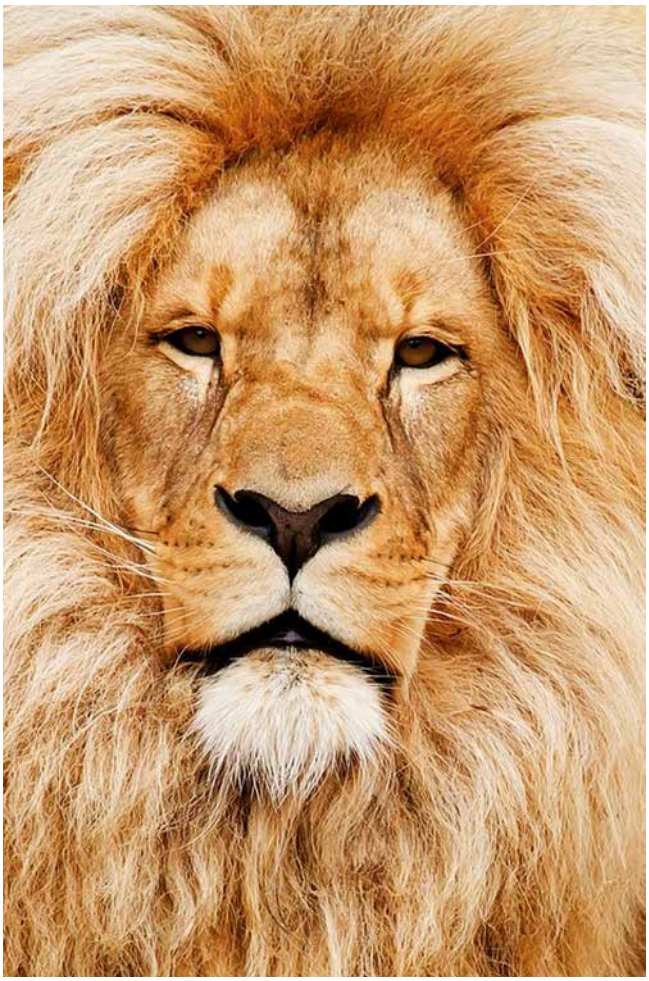
RECHTS UNTEN Bei einer sogenannten Error-Level-Analyse zeigt sich, wo das Bild verändert wurde: Das Rauschen des Himmels rund um das Messer unterscheidet sich vom Rauschen des restlichen Himmels.

lungen, die für gezielte Desinformation besonders relevant sind.

Aus technischer Sicht gibt es mehrere Ansätze, Bilder zu erkennen, die aus Text-to-Image-Verfahren stammen. Zumindest ein Teil der Verfahren[1] bringt beim Erstellen ein digitales Wasserzeichen ein, das Rückschlüsse auf die Herkunft des Bildes zulässt. Allerdings muss man davon ausgehen, dass diese Wasserzeichen bei einer missbräuchlichen Nutzung entweder im Source Code deaktiviert oder erfolgreich entfernt werden können. Wie schnell das gehen kann, haben Wettbewerbe wie BOWS (Break our watermarking system) gezeigt.[2] Gerade wenn Profis Zugang zum System haben, können sie die Marker recht einfach entfernen. Dies wäre beispielsweise bei Stable Diffusion der Fall, da der Code des Wasserzeichenverfahrens ebenfalls verfügbar ist.

Synthetische Bilder erschweren die Spurensuche

Geht man davon aus, dass keine Markierung zur Überprüfung ausgelesen werden kann, kommen die Methoden der Multimedia-Forensik zum Einsatz. Denkbar ist hier die Manipulationserkennung, die verräterische Spuren der Verfahren aufdeckt. Solche Ansätze sind bei Deepfakes erfolgreich, da diese üblicherweise einen kleinen Bildbereich eines Videos verändern, zum Beispiel durch Ersetzen eines Gesichts. Dieser veränderte Bildausschnitt weist andere Signaleigenschaf-



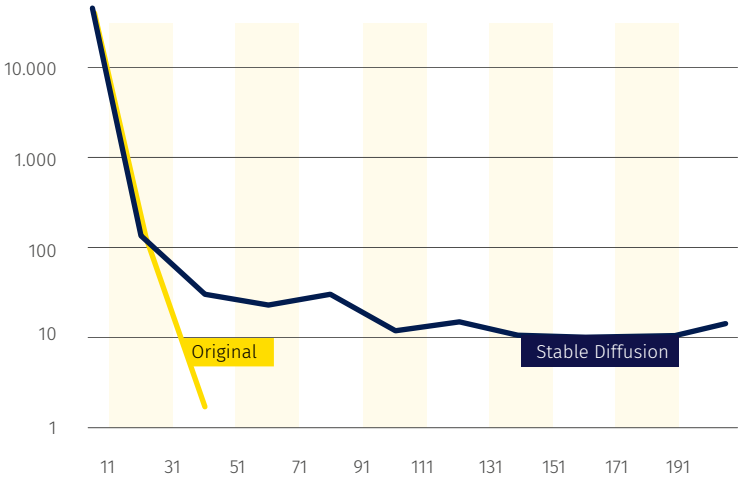
ten auf als der Rest des Videoframes, da hier eine andere Historie an Kompression und Skalierung vorliegt. Vereinfacht gesagt: Der Bereich des Deepfakes „rauscht“ anders als der Rest. Bei Text-to-Image-Verfahren, bei denen ganze Bilder neu erzeugt werden, ist dies nicht der Fall. Hier gibt es keinen Bereich, der zu einem Original gehört und mit einem veränderten Bereich verglichen werden kann. Im Fall von Inpainting kommt es darauf an, wie das Bild erzeugt wird: Wenn nur der zu ersetzende Bereich neu erstellt und in den Rest eines vorhandenen Bildes kopiert wird, lassen sich die Unterschiede forensisch erkennen. Wird durch Inpainting nur ein Bereich inhaltlich verändert, aber das ganze Bild (also der neue und der unveränderte Teil) nach dem Vorbild des vorhandenen Originals neu gezeichnet, dann weisen alle Bildteile die gleichen Eigenschaften auf und können durch bildforensische Methoden nicht voneinander unterschieden werden.

Um synthetische Bilder zu erkennen, könnten auch Methoden aus der Steganalyse gut geeignet sein. Ursprünglich wurden diese entwickelt, um das Vorhandensein steganographischer – also versteckter – Nachrichten in Bildern zu erkennen. Solche Methoden stützen sich häufig auf den Nachweis statistischer Auffälligkeiten, die durch das Einbetten der Nachricht entstehen. Sie sind daher gut geeignet, Bilder mit einem untypischen statistischen Verhalten zu erkennen. Das wäre beispielsweise der Fall, wenn synthetische Bilder grundsätzlich andere Rauscheigenschaften haben als Fotos: eventuell, weil Fotos immer verschiedene Rauschquellen wie die Kamera, die ISO-Empfindlichkeit und die Verarbeitung vereinen – und diese bei der Synthese nicht auftreten.

Zwischen Fingerabdrücken und verwischten Spuren

Und auch die Kameraballistik könnte helfen. Wer den Begriff Ballistik nur aus Fernsehsendungen wie CSI und Tatort kennt, liegt nicht ganz daneben: Auch Kameras haben eine Art Fingerabdruck, etwa durch defekte Pixelsensoren. Anhand dieser individuellen Profile lässt sich herausfinden, mit welcher Kamera ein Foto aufgenommen wurde. Sind entsprechende Fingerabdrücke gar nicht nachzuweisen, kann dies für ein künstliches Bild sprechen. Die Erfahrung zeigt allerdings, dass alle oben genannten Ansätze immer nur einen gewissen Grad an Erkennung erlauben. Zudem gehen sie häufig davon aus, dass bei der Bilderstellung nicht gezielt Spuren verschleiert wurden. Dies kann zum Beispiel durch künstliches Rauschen, Weichzeichnen oder das Kopieren eines Kamera-Fingerabdrucks geschehen.

Möchte man Bildern vertrauen, ist es perspektivisch wahrscheinlich schwieriger, die Echtheit eines Bildes nachzuweisen, als eine Fälschung zu entlarven. Eine Kennung für synthetische Bilder zu erzwingen ist wenig sinnvoll – zu einfach lässt sie sich umgehen. Doch auch eine forensische Untersuchung kann Verwirrung stiften: Selbst Profis passiert es, dass sie gefälschte Bilder nicht als solche erkennen oder Originale für Fälschungen halten. Eine zuverlässigere Lösung wäre es, eine Infrastruktur aufzubauen, die Originale anhand



MITTE Erkennen Sie, welcher der beiden Löwen synthetisch erstellt wurde?*

RECHTS Ein Vergleich der Lauflängen auf der niedrigsten Bitebene der beiden Löwen zeigt, dass der künstlich erzeugte Löwe deutlich längere identische Bitfolgen, also Aneinanderreihungen mehrerer Bits hintereinander, aufweist.

einer Kette von Nachweisen immer referenziert. Auf diese könnte im Falle einer Überprüfung zurückgegriffen werden, beispielsweise auf der Website einer Nachrichtenagentur. Eine Signatur des Originals durch eine vertrauenswürdige Instanz wie beispielsweise eine Nachrichtenagentur, eine bekannte Fotografin oder eine Behörde würde zusätzliche Sicherheit bringen. Doch dies zieht einen hohen organisatorischen Aufwand nach sich. Zumindest wäre hier aber klar, dass es sich um vertrauenswürdige Bilder handelt. Bei allen übrigen Bildern muss wie bisher die Betrachterin oder der Betrachter Vorsicht walten lassen und Inhalte kritisch hinterfragen: Nur weil Neuschwanstein auf meinem Smartphone in Flammen steht, heißt das noch lange nicht, dass es auch wirklich brennt – und auch Angela Merkel wird sich so schnell vermutlich nicht tätowieren lassen. Fest steht: Die Forschung rund um Text-to-Image-Verfahren wird voranschreiten – nicht nur zur Erstellung, sondern auch zur Erkennung synthetischer Bilder. In Zukunft ist auch mit neuen technischen Ansätzen zu ihrer Aufdeckung zu rechnen. I

Über den Autor

Prof. Dr. Martin Steinebach forscht seit 1999 zur Sicherheit digitaler Medien. Er ist Abteilungsleiter Multimedia-Sicherheit und IT-Forensik am Fraunhofer SIT und Forschungsbereichsleiter für Sicherheit für und durch maschinelles Lernen am Nationalen Forschungszentrum für angewandte Cybersicherheit Athene. Zudem erhielt er eine Honorarprofessur Multimedia-Sicherheit und IT-Forensik an der TU Darmstadt und war langjähriger Sprecher der GI-Fachgruppe Steganographie und digitale Wasserzeichen.

[1] *https://medium.com/@steinsfu/stable-diffusion-the-invisible-watermark-in-generated-images-2d68e2ab1241*

[2] Westfeld, Andreas (2006), *Lessons from the BOWS contest. In Proceedings of the 8th Workshop on Multimedia and Security*, S. 208–213

* Der rechte Löwe wurde mittels Stable Diffusion erstellt.