

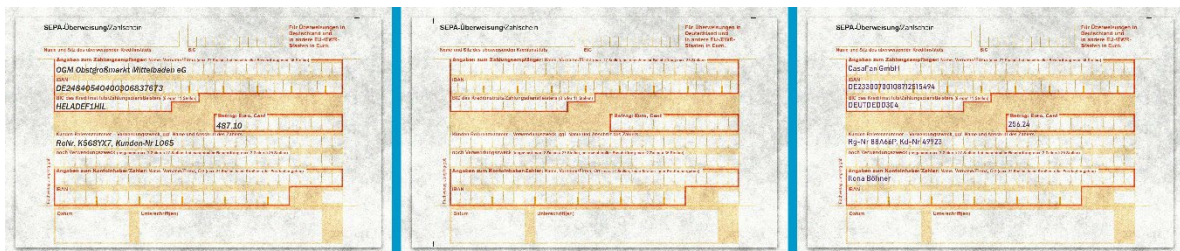
Proposal for a Bachelor- / Master-Thesis

Photorealistic Document Anonymization

Automatic extraction and interpretation of information from documents is crucial in many fields, ranging from financial to medical. It streamlines tasks, enhances decision-making, and also improves accessibility for individuals with visual impairments. Visual Document Understanding (VDU) offers a solution by automating this process, making it more efficient and enabling quick retrieval of relevant data.

To build a robust and reliable VDU a significant amount of data is needed. However, the General Data Protection Regulation (GDPR) poses challenges for gathering large amounts of data for VDU due to its stringent regulations on data privacy and protection. Therefore, document anonymization, while retaining the usefulness of the data, is a key factor when building datasets for VDU development.

The objective of this thesis is to develop a method to realistically erase certain text fields from a photographed or scanned document. In addition, those fields shall be replaced with artificial yet realistic data. The following figure offers an example of the anonymization pipeline.



Optionally, the inpainting step could also be skipped so the information replacement is carried out as an end-to-end process.

Implementation and evaluation are to be done using Python. A well-structured and well-commented code has to be handed in at the end of the thesis.

The thesis is offered in collaboration with Gini GmbH, a market-leading company that simplifies payment for millions of users from banks, insurance, and e-commerce companies in Germany.

Contact: Jürgen Seiler, 06.034, juergen.seiler@fau.de
Professor: Prof. Dr.-Ing. André Kaup
Prerequisites: Experience with python programming, computer vision, and machine learning.
Available: Immediately