Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

OIDIQ: Open identity document image quality

Gregor Grote

gregor.grote@h-da.de

Hochschule Darmstadt

Juan E. Tapia

Hochschule Darmstadt

Christian Rathgeb

Hochschule Darmstadt

Research Article

Keywords: Computer vision, Image quality, Identity documents, Document image quality assessment, Remote identity verification

Posted Date: November 26th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11729550/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

OIDIQ: Open identity document image quality

Gregor Grote [1*], Juan E. Tapia [1] and Christian Rathgeb [1]

[1*]da/sec, Hochschule Darmstadt, Schöfferstraße 3, Darmstadt, 64295, Hessen, Germany.

*Corresponding author(s). E-mail(s): gregor.grote@h-da.de;
Contributing authors: juan.tapia-farias@h-da.de; christian.rathgeb@h-da.de;

Abstract

Remote identity verification relies on camera images of identity documents (IDs), but reflections, shadows, blur, compression, occlusions, and poor framing can reduce the performance of subsequent processing. Existing document image quality assessment largely targets generic text pages and often predicts a single optical character recognition (OCR) or subjective-oriented score, while face-oriented frameworks such as open face image quality (OFIQ) are not directly transferable to ID images. We present open identity document image quality (OIDIQ), an open framework for assessing the front of ID images. OIDIQ detects and normalizes the document, masks occlusions, segments the portrait, and reports 14 independent quality measures on a 0–100 scale divided into three categories. The measures cover common capture problems, such as glare, shadows, blur, compression and poor exposure, as well as properties of the document itself, including its alignment, visibility and framing. OIDIQ also reports an estimate of the performance of its own preprocessing steps. We developed and evaluated the components required for these assessments, including document localization, occlusion masking, glare and shadow detection, and ID-specific image quality measures. Since existing ID datasets do not provide comprehensive quality annotations, we trained and evaluated the individual components using a combination of real captures, manually annotated examples and synthetically generated degradations. The implementation and model weights will be made publicly available upon acceptance.

Keywords: Computer vision, Image quality, Identity documents, Document image quality assessment, Remote identity verification

1 Introduction

Banks, insurances, and public services increasingly verify the identity of their customers remotely. The customer captures their identity document (ID) with a smartphone camera, and the service reads the text fields, compares the portrait with a selfie and checks whether the ID is pristine, i.e. bona fide [1]. All of these steps depend on the quality of the captured image. A glare can hide the portrait or a text field, blur makes small print unreadable, and strong compression removes fine security patterns. Presentation attack detection (PAD) for IDs is also sensitive to the capture conditions [2]. Assessing the ID image quality right after the capture allows a service to reject an unusable image and to ask the user for a new presentation.

For face images, image quality assessment is standardized in ISO/IEC 29794-5, and the open framework OFIQ [3] implements it. OFIQ reports a set of separate quality measures, such as sharpness, exposure, compression artifacts and occlusion, each mapped to a scalar score in the range 0

to 100. Such separate scores tell the operator and the user what is potentially wrong with an image. For IDs, no comparable open framework exists to the best of our knowledge. Document image quality assessment mostly predicts the OCR accuracy of generic text pages with a single score [4], and quality checks in commercial ID verification systems are not public.

In our previous work [5], we applied the measures of OFIQ to ID images. This showed that several of them are not suited for IDs, as their models were trained on faces and their score conversions were tuned to face images.

In this work, we present OIDIQ, an open framework for ID image quality assessment. Our main contributions are:

- OIDIQ, a modular extendable framework for the preprocessing and quality measure calculation of ID, similar to OFIQ, which will be available on GitHub[1] upon acceptance (Section 3).
- The proposal and rigorous evaluation of dedicated preprocessing and quality measure algorithms for IDs, most notable,
  - a corner detector that generalizes to unseen document types, rejects other documents, and predicts the accuracy of each corner, compared with a classical and a learned document localizer (Section 4.1);
  - an occlusion segmentation model for IDs, evaluated on MIDV-500 and on a new set of passport images with large occlusions against SAM 2.1 and MobileSAM (Section 4.3);
  - a glare and shadow detector evaluated on real captures against SAM 3 and threshold baselines (Section 5.2.1).
- An adaptation of existing quality measure concepts including the OFIQ compression and sharpness models for IDs, with a new estimate of the compression strength, and a new sharpness model, which we show can predict OCR legibility (Sections 5.2.2 and 5.2.3).

The remainder of this paper is structured as follows. Section 2 reviews related work, and Section 3 gives an overview of OIDIQ. Section 4 describes the preprocessing, Section 5 the quality measures, and Online Resource 1 the scalar conversion of each measure. Section 7 concludes the paper, and Section 6 discusses its limitations.

[1] available upon acceptance.

2 Related Work

Document image quality assessment. Document image quality assessment (DIQA) usually defines quality either as subjective quality scores [7, 8] or as objective measures such as OCR accuracy [4, 9]. Learned methods regress this accuracy from image patches [10], and sharpness measures for natural images were shown to transfer poorly to text [11]. These methods target generic text pages and give a single score, which does not tell which defect causes a low quality. For IDs, Shemiakina et al. [12] assess whether the projective distortion of a captured document still allows text recognition, as a binary decision, tested on synthetic distortions of MIDV-2019 [13] document templates. To the best of our knowledge, no published work assesses the image quality of IDs with a set of separate measures designed and tuned specifically for ID images.

General image quality assessment. No-reference image quality assessment predicts the perceived quality of natural images, with handcrafted statistics [14] or deep networks [15–17]. These methods are trained on human opinion scores of natural images, give a single score, and do not consider which region of the image shows the ID.

Biometric sample quality. In biometrics, the quality of a sample is defined by its utility, that is, how well it predicts the recognition performance [18]. NFIQ 2 [19] implements this for fingerprints and OFIQ [3] for faces, see [20, 21]. OFIQ combines a preprocessing (face detection, landmarks, alignment and segmentation) with separate measures, such as sharpness, compression artifacts, exposure, illumination uniformity and occlusion, and converts each native score to a scalar score from 0 to 100. OIDIQ follows the same structure for IDs. As IDs are used for several tasks, such as OCR, face comparison and PAD, there is no single utility, so we validate each measure against the defect it measures or an example task.

Identity documents and our previous work. Datasets of ID images such as MIDV-500

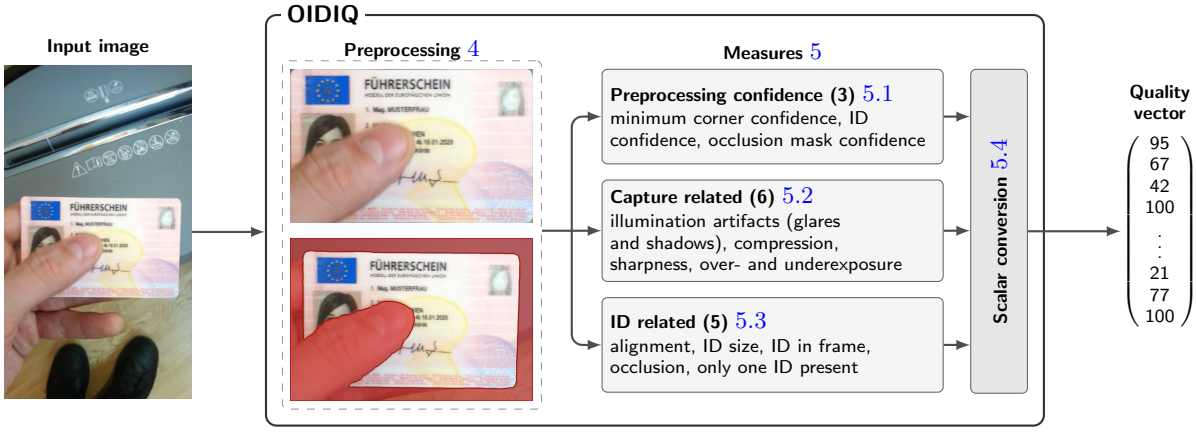


Fig. 1 Overview of OIDIQ. The preprocessing normalizes the ID and masks occlusions (red). The quality measures of the three categories are converted to scores from 0 to 100, which form the quality vector. Image from MIDV-500 [6]

[6] and MIDV-2019 [13] were collected for document localization and recognition, DLC-2021 [22] for document liveness detection, and synthetic datasets such as SynID [23], FantasyID [24] and IDNet [25] for PAD and fraud detection [26, 27]. None of them contains quality labels. Yáñez and Tapia [28] and Schulz et al. [29] predict a single quality score for ID cards, from subjective ratings and from face and OCR confidences. In our previous work [5], we built a preliminary quality framework for ID cards solely based on applicable quality measures of OFIQ. Related work on the individual steps, such as document localization, shadow and glare detection, is discussed in the corresponding sections.

3 OIDIQ overview

OIDIQ takes a single camera image that should show one ID and returns a set of quality scores. It works in two stages:

First, the preprocessing (Section 4) finds the ID and the regions of it that the measures should assess. It detects the corners of the ID, crops the ID and transforms it to a top-down view of fixed size, masks occlusions such as fingers, and segments the portrait.

In the second stage, the measures (Section 5) assess different aspects of quality independently. We divide them into three different categories:

- *Preprocessing confidence*: the minimum corner confidence, the ID confidence and the occlusion mask confidence.
- *Capture-related quality defects*: illumination artifacts (glares and shadows), compression artifacts, sharpness, and over- and underexposure.
- *ID-related quality defects*: alignment, ID size, ID in frame, occlusion, and whether only one ID is present.

Similar to OFIQ [3], each measure first computes a native score, such as a share of pixels, a model output or an angle, and then converts it to a scalar score from 0 (worst) to 100 (optimal) (Section 5.4). The scores are kept separate, so that a low score shows which aspect is the problem, for example to guide the user with actionable feedback during the capture process.

We evaluated each step and measure on its own, with the image data that contains the corresponding ground truth (Table 1). As no dataset of ID images comprise quality labels, we trained on synthetic artifacts where real annotations are tedious to obtain (compression, blur, glares and shadows), and annotated real captures where possible (corners, occlusions, exposure). To test the generalization to unseen documents, we used test documents that were not part of the training data or held out document types.

4 Preprocessing

In this section we first describe the preprocessing steps. Before the quality measures are computed, each image passes through the following preprocessing steps: the corners of the ID are detected

Table 1 Main datasets used in this work

Dataset	Type	Used for
DLC-2021 [22]	Real captures	Corner and occlusion training, glare test
SynID [23]	Synthetic	Corner and illumination training
FantasyID [24]	Synthetic, recaptured	Corner and illumination training, corner and compression test
IDNet [25]	Synthetic	Corner and illumination training
PAD competition [27]	Real, recaptured	Occlusion training
MIDV-500 [6]	Real captures	Corner, occlusion, compression, sharpness and exposure test
MIDV-2019 [13]	Real captures	Corner test

Table 2 Composition of the corner detection data (number of images; IDs = annotated IDs in the training images). Derived sets are generated from DLC-2021 frames; CORD and the business cards are oversampled

Source	Train	Val	IDs
With IDs			
DLC-2021 [22]	6,881	994	6,881
COCO objects overlaid	4,000	250	4,000
Second ID pasted	3,000	200	6,000
Same ID pasted repeatedly	2,500	200	7,563
Cropped and padded	5,000	300	5,000
FantasyID pasted [24]	4,500	300	5,224
IDNet pasted [25]	1,000	100	1,150
SynID pasted [23]	2,000	150	2,304
Without IDs (negatives)			
ID removed by inpainting	4,500	300	–
COCO [31]	1,500	200	–
CORD receipts [33]	1,500	99	–
Business cards [34]	2,000	100	–
Total	38,381	3,193	38,122

(Section 4.1), the ID is cropped and normalized (Section 4.2), and occlusions are masked (Section 4.3).

4.1 Corner detection

The first step of the preprocessing is the detection of the corners of the ID. For this, we use a YOLO26 pose model of size m [30]. The input is the image padded by 10% on all sides, so that corners that lie a little bit outside the original image can still be predicted. The output is a quadrilateral per detected ID with its four corners, clockwise from the corner closest to the top left of the image, and a confidence per corner as well as a confidence for the entire quadrilateral.

Dataset

The training data consists of 38,381 images and the validation data of 3,193 images (Table 2, Figure 2, top row). Real captures come from DLC-2021 [22], which includes video frames of ten mock document types.

As ten designs are too few for a detector that has to work on unseen IDs, we pasted synthetic IDs of further designs onto background images: the ten training designs of FantasyID [24], IDNet images [25] and passports of SynID [23], including their printed and screen recaptures.

We then augmented the images by partially occluding the IDs with objects from COCO [31], pasting a second ID into the image or the same ID multiple times, or applying random crops.

A detector trained only on IDs is expected to falsely detect any rectangular document. About a quarter of the training images therefore contain no ID: DLC-2021 frames with the ID removed by inpainting [32], which keep the backgrounds of the positive images, as well as COCO images, receipts from CORD [33], and business cards [34].

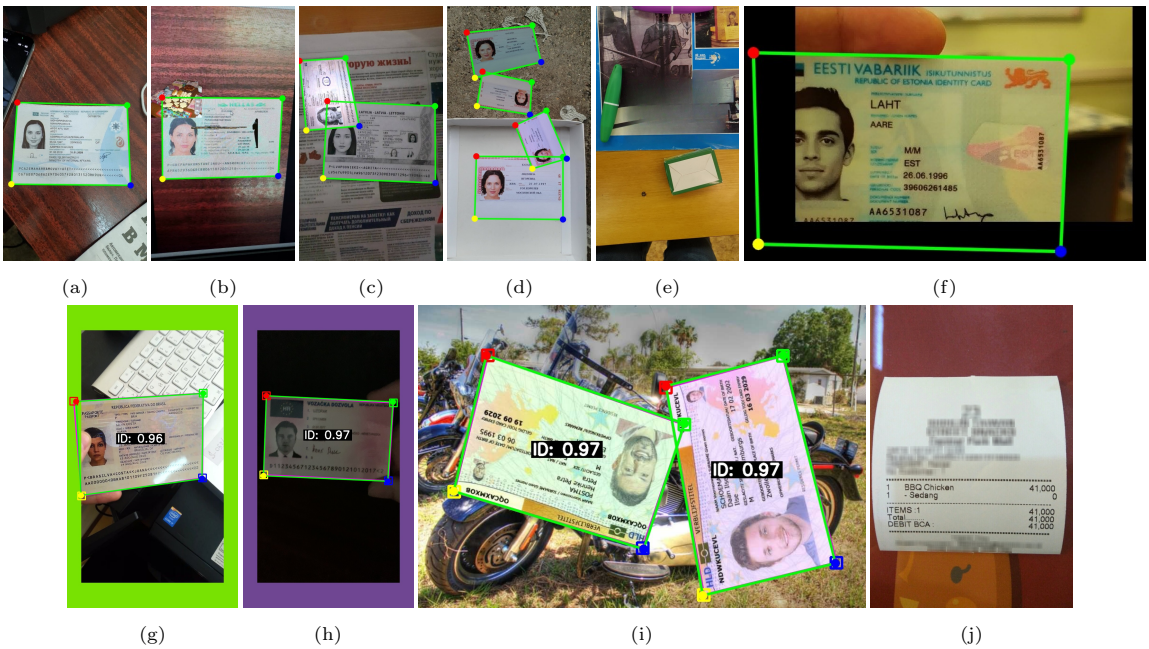
Model and training

We trained the model from scratch at 640×640 pixels for 250 epochs.

We changed the standard training of a YOLO keypoint detector, because we want to be able to use a confidence per corner as a quality measure (Section 5.1.1). We therefore trained the confidence of a corner to predict its accuracy instead of its visibility, with the target $\exp\left(-d^2 / \left(2A\tau^2\right)\right)$, where d is the error of the predicted corner, A the area of the bounding box and $\tau = 0.03$. This is the keypoint similarity of the object keypoint similarity (OKS) [35]. Similar to CCNet [36], which learns this similarity per keypoint in a separate branch added after training, we train the confidence output of the detector with it directly.

Evaluation

The test set consists of 14,400 frames of 48 document types from MIDV-500 [6] and 5,760 frames of MIDV-2019 [13]. We left out the types without a portrait. Eleven of the types match the ten types of DLC-2021, so we report them separately as *seen* designs and the other 37 as *unseen* designs. We additionally used 900 images with the three



(a) (b) (c) (d) (e) (f)

(g) (h) (i) (j)

Fig. 2 Top row: training data with the annotated corners: (a) DLC-2021 frame, (b) COCO objects overlaid, (c) second ID pasted, (d) same ID pasted repeatedly, (e) ID removed by inpainting (negative) and (f) crop that cuts off the ID, padded. Bottom row: corner predictions with the detection confidence on unseen designs of (g) MIDV-500, (h) MIDV-2019 and (i) FantasyID, and (j) on a negative image

FantasyID [24] test designs, pasted onto held-out COCO images, and 584 held-out negative images from COCO, CORD [33] and the business cards.

Table 3 reports the results and Figure 2 (bottom row) shows examples. On unseen MIDV-500 designs, the model reaches a precision of 99.9 %, a recall of 90.9 % and a mean corner error of 0.78 % of the ID diagonal, so the model transfers to unseen designs.

The main limitation is corners far outside the frame. On MIDV-500, the recall is 98.4 % with all corners inside the frame, 94.7 % with corners within the inference padding and 43.9 % with corners beyond it, with mean corner errors of 0.45 %, 0.96 % and 4.45 %. The minimum corner confidence reflects this, so such images can be filtered out (Section 5.1.1): its mean is 0.94 with all four corners inside the frame and 0.67 with at most two.

We compared the model with three baselines (Table 3). For the Hough-based method of Tropin et al. [37], we used their published per-frame outputs on MIDV-500, as neither the code nor outputs for other data are available. DocAligner [38] is a public document localizer, which we ran on the same images as our model. We also ran the corner model of our previous work [5] on these images.

Our previous model has the highest recall and IoU (95.5 % and 0.924 vs. 90.9 % and 0.888 on unseen designs), but more false detections (1.6 % vs. 0.1 %) and a higher corner error (1.12 % vs. 0.78 %). DocAligner has the lowest corner error (0.38 % on unseen designs). It also finds almost all IDs with the corners inside the frame or within the padding. The Hough-based method returns a result for almost every frame, which gives it the most false detections.

On FantasyID, our model has a recall of 99.7 % compared to 52.2 % for DocAligner and 36.8 % for our previous model, which both return at most one document per image and often no result.

DocAligner localizes single documents most accurately, with a mean corner error of about 4.5 pixels on the normalized image compared to about 9 pixels for our model on unseen MIDV-500 designs. However, a quality framework also has to handle images without an ID and with several IDs, and it needs a confidence per corner, which DocAligner does not provide. Our model is the only compared method that covers all of this,

Table 3 Corner detection results on the test sets. A prediction matches an annotated ID at an IoU of at least 0.5, IoU: mean over the IDs. Error: mean corner error of the matched IDs in % of the ID diagonal. FD: images with a false detection in %. Hough: Tropin et al. [37], only available on MIDV-500. DocAligner: [38]. Previous: our previous model [5]. Best result per test set in bold

Method	P	R	IoU	Error	FD
MIDV-500 unseen (11,100 images, 37 types)					
Ours	**99.9**	90.9	0.888	0.78	**0.1**
Previous	98.4	**95.5**	**0.924**	1.12	1.6
Hough	92.9	92.9	0.902	1.08	7.1
DocAligner	98.9	91.7	0.906	**0.38**	1.0
MIDV-500 seen (3,300 images, 11 types)					
Ours	**99.1**	94.8	0.928	0.76	**0.9**
Previous	96.4	**95.3**	**0.929**	0.84	3.6
Hough	92.2	92.2	0.898	0.96	7.8
DocAligner	98.4	92.2	0.910	**0.41**	1.5
MIDV-2019 unseen (4,440 images, 37 types)					
Ours	99.5	98.4	0.963	0.68	0.5
Previous	98.7	98.0	0.934	1.67	1.3
DocAligner	**100.0**	**100.0**	**0.981**	**0.56**	**0.0**
MIDV-2019 seen (1,320 images, 11 types)					
Ours	98.7	99.8	0.978	0.59	1.4
Previous	99.0	98.2	0.933	1.93	1.0
DocAligner	**100.0**	**100.0**	**0.984**	**0.45**	**0.0**
FantasyID unseen (900 images, 3 types)[a]					
Ours	**99.4**	**99.7**	**0.976**	**0.74**	**0.7**
Previous	78.2	36.8	0.335	3.29	11.8
DocAligner	98.4	52.2	0.504	1.37	1.0
Negatives (584 images)[a]					
Ours	–	–	–	–	**0.5**
Previous	–	–	–	–	38.2
DocAligner	–	–	–	–	35.4

[a]Previous and DocAligner are made for documents in general and return one document per image, so this comparison favors our model.

which is why we use it, at the cost of a higher corner error. But this is not really an issue in this setting, as when the error is high, the confidence of the model is low (Spearman correlation -0.53 and -0.71 per test set), the image has not been properly captured, and the user will be prompted to recapture the ID.

4.2 Cropping and normalization

Afterwards, we use these detected corners to crop the ID from the image (from now on referred to as "cropped image") and apply a perspective transformation to obtain a top-down view of the ID (from now on referred to as "normalized image"). The normalized image has a fixed resolution of 1000×631 pixels, which corresponds to the aspect ratio of the ID-1 format (85.60×53.98 mm) [39].

4.3 Occlusion masking

We detect occlusions on the ID, to be used for masking occluded regions in subsequent processing as well as for reporting the amount of occlusion as a quality measure (Section 5.3.4). To this end, we segment the visible part of the ID. We trained our own model, as we did not find any open weight models for this task. Available models, like the U-Net trained on MIDV-500 [40], learn the filled corner polygon as the target and therefore mark only 24 % of the occluded pixels on our passports, compared to 88 % for our model.

Dataset

The training data consists of 2,016 images: 594 images from the dataset of the second competition on PAD on ID cards [27], which includes printed and screen-recaptured IDs, 580 images from DLC-2021 [22], and 842 Mexican and Chilean IDs from a private dataset. For annotation, we used SAM 2.1 [41] prompted with a box around the annotated corners to provide a pre-annotation, which we then corrected manually.

Model and training

We use a Mobile U-Net [42]. As an input, we crop the image around the detected corners, convert it to grayscale and resize it to 512×320 pixels. In a second channel, we provide the filled quadrilateral of the detected corners. The output is the probability of each pixel to belong to the visible ID, which we threshold at 0.5.

During training, an occluder is pasted onto the ID (an object cut out from COCO [31] or a synthetic shape) and removed from the target mask, followed by geometric and photometric augmentation (Figure 3a–d). We trained the model on a random 90 % of the training data (1,814 images) for 825 epochs. On the remaining 202 images, the mean IoU of the visible region is 0.985. As these contain frames of the same documents as the training images, we report the results on separate test data.

Evaluation

The test data consists of 302 images: 206 from MIDV-500 [6] and 96 that we captured ourselves. We used two datasets for testing, as the occlusions in MIDV-500 are mostly fingers and only on ID

cards. Our dataset adds a wider variety of occluders on passports: we photographed physical mock passports from the set of SynID [23] with a phone, 24 images for each of the four available countries (Colombia, Poland, Portugal and Spain). We used the annotated corners as input, so that errors of the corner detection do not affect the results.

We compare our model with SAM 2.1 (base variant, on the same input crop in color, prompted with the box around the corners instead of the filled polygon, keeping the mask with the highest IoU to the ideal document shape) with the much smaller MobileSAM [43] (same input and prompt), and with an pretrained U-Net trained on MIDV-500 [40] (on the full image). We compute the IoU of the occlusion target in the box around the corners, extended by 6 % of its longer side, and pad the image where the box exceeds it. In addition, we report the AUC for detecting images in which more than 1 % of the ID-1 shape is occluded (Section 5.3.4). Table 4 reports the results and Figure 3 e–j shows examples.

On MIDV-500, our model reaches an IoU of 0.957 and detects occluded frames with an AUC of 0.93. SAM performs better on this test set (IoU of 0.970, AUC of 0.98), while MobileSAM performs worse (IoU of 0.941, AUC of 0.89). The U-Net reaches an IoU of 0.948, but it hardly detects occluded frames (AUC of 0.64), although MIDV-500 is its training data, as it marks most occluders as part of the ID. Our model often marks glares as occlusions (Figure 3e, f): 39 % of its falsely occluded pixels on the card lie in predicted glares.

The passports have much larger occlusions. Here, our model is on par with SAM, with an IoU of 0.953 compared to 0.956 and an AUC of 0.995 compared to 0.971, with overlapping intervals. MobileSAM again performs worse (IoU of 0.925, AUC of 0.95), and the U-Net reaches only an IoU of 0.833 and an AUC of 0.83.

We use SAM only as a baseline due to its computational cost: it has about 81 million parameters and takes about 1.4 s per image on a CPU, compared to 13 million parameters and 62 ms for our model.

4.4 Face detection

As in our previous work [5], the portraits on the normalized image are segmented with a pretrained YOLO11 instance segmentation model of size *m*.

Table 4 Occlusion masking on the two test sets with annotated corners: mean IoU per image with the occlusion target in the extended box around the corners and AUC for images with more than 1 % of the ID-1 shape occluded, with 95 % bootstrap confidence intervals over the images in brackets

Method	IoU	AUC
MIDV-500 (206 frames)		
Ours	0.957 [0.951, 0.963]	0.928 [0.890, 0.962]
SAM 2.1	**0.970** [0.968, 0.972]	**0.984** [0.968, 0.995]
MobileSAM	0.941 [0.933, 0.948]	0.886 [0.838, 0.929]
U-Net [40][a]	0.948 [0.940, 0.956]	0.643 [0.561, 0.718]
Passports (96 images)		
Ours	0.953 [0.944, 0.961]	**0.995** [0.985, 1.000]
SAM 2.1	**0.956** [0.946, 0.964]	0.971 [0.936, 0.997]
MobileSAM	0.925 [0.916, 0.934]	0.954 [0.908, 0.986]
U-Net [40]	0.833 [0.809, 0.856]	0.826 [0.741, 0.898]

[a] Trained on MIDV-500, so these images are part of its training data.

This is used to prevent biases in the subsequent quality measures that rely on luminance, ensuring that the face region does not skew the results. We kept this step in this version.

5 Quality measures

Each measure computes a native measure, such as a share of pixels, a model output or an angle. We describe the measures by category, and the conversion of the native measures to scalar scores in Section 5.4.

5.1 Preprocessing confidence

The preprocessing confidence measures report an estimate on how reliable the preprocessing is. An error of the preprocessing propagates to the other measures, as they assess the normalized image and the region without occlusions. A low score therefore indicates that the presented ID is somehow not properly detectable.

5.1.1 Minimum corner confidence

The minimum corner confidence measures how reliable the detected corners are, as a misplaced corner distorts the cropped and normalized image. It is the lowest of the four corner confidences of the ID (Section 4.1).

We evaluated the measure on the 19,936 IDs of the five test sets that our detector found (Table 3), as an unmatched detection has no annotated corners to compare with. The measure follows the



Fig. 3 Occlusion masking on training examples with a COCO object (a, b) or a synthetic shape (c, d), on MIDV-500 (e–g) and on the passports (h–j): predicted ID (green) and target (yellow outline). The model marks glares as occlusions (e, f)

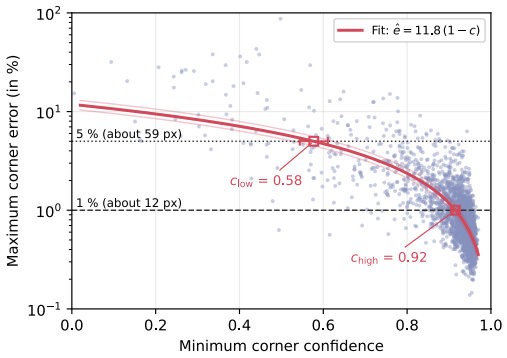


Fig. 4 Maximum corner error of an ID in % of the ID diagonal against its minimum corner confidence on the five test sets with IDs, with the fit per test set (thin) and on all IDs with equal weights for the test sets (bold), and the two anchor points of the score (Online Resource 1) with their standard deviation over the fits. Of the 19,936 IDs, 600 per test set are shown

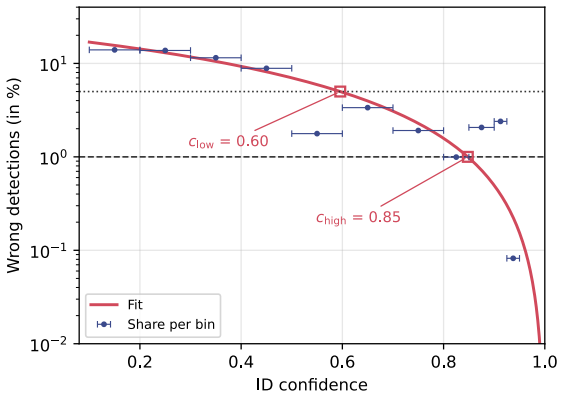


Fig. 5 Share of wrong detections against the ID confidence, with the fit and the two anchor points of the score (Online Resource 1). Dots: share per confidence bin, horizontal lines: width of the bin. Bins without wrong detections lie below the axis

maximum corner error of the ID (in % of the ID diagonal), with a Spearman correlation between −0.53 and −0.71 per test set (Figure 4). The error stays small at high confidence and increases quickly below a certain confidence.

5.1.2 ID confidence

The ID confidence measures how certain the detector is that an ID is present. It is the confidence of the most confident detection in the image, and 0 if nothing was detected.

We evaluated the measure on all test sets, including the negative images, as it should separate IDs from other content. Of the 20,087 detections with a confidence of at least 0.1, 85 are wrong, that is, they have an IoU below 0.5 with every annotated ID. The share of wrong detections decreases with higher confidence (Figure 5). However, the detector finds IDs in the test sets very well and rarely fires on other content, so nearly all detections have a high confidence, and the utility of this measure is not as large as that of the other measures.

5.1.3 Occlusion mask confidence

The occlusion mask confidence measures how reliable the predicted occlusion mask is, as a wrong mask excludes valid regions from the other measures or leaves occluded regions in them. We use

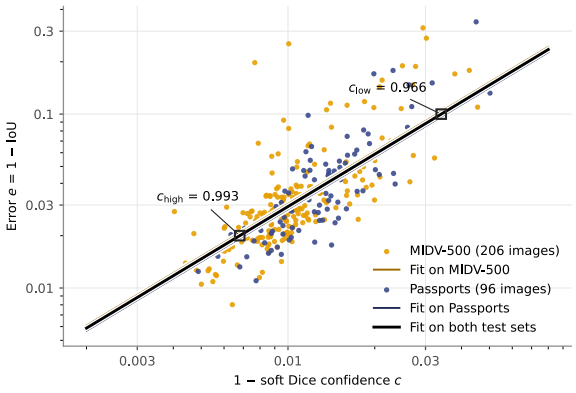


Fig. 6 Error of the occlusion mask against the soft Dice confidence on both test sets, with the fit per test set (thin) and on all images with equal weights for the test sets (bold), and the two anchor points of the score (Online Resource 1). Both axes are logarithmic

the Soft Dice Confidence (SDC) [44] as a measure of segmentation confidence:

$$c = \frac{2\sum_i q_i y_i}{\sum_i (q_i + y_i)}, \tag{1}$$

where q_i is the predicted probability that pixel i is occluded and y_i is its thresholded prediction.

We evaluated the measure on the 302 images of the two test sets (Table 4). The measure follows the IoU, with a Spearman correlation of 0.73 on MIDV-500 and 0.79 on the passports (bootstrap 95 % confidence intervals 0.65 to 0.81 and 0.67 to 0.86). The error of the mask decreases with the confidence (Figure 6).

The measure only sees the output of the model, so it cannot detect an occlusion that the model misses with high confidence. Two MIDV-500 frames with more than 10 % of the ID occluded still reach scores of 88 and 97 (Online Resource 1, Section S3) with an IoU of 0.75 and 0.80.

5.2 Capture-related measures

The capture-related measures assess defects that come from the capture conditions: illumination artifacts, compression artifacts, blur, and over- and underexposure. They assess the normalized image, except for the compression artifacts, which we uses on the cropped image.

5.2.1 Illumination artifact detection

Glares and shadows can partially or completely hide the printed content. In our previous work [5], illumination was only assessed as the standard deviation of the median luminance of 3×4 blocks. To prevent dark or bright faces as well as dark text from being misclassified as illumination artifacts, these regions were masked out before calculating the standard deviation. This leaves out many regions that may still contain illumination artifacts, and these regions are more relevant for downstream tasks. For example, a glare on the face would cause face detection to fail, but a glare in a region without any faces or text will not harm most downstream tasks. Existing glare detectors for documents work on coarse blocks [45], and related datasets and methods mainly target glare only [46, 47] or the removal of weak shadows instead of the detection of only strong ones [48, 49]. Since real glares and shadows are hard to annotate consistently at pixel level due to their soft borders, we train a detector for both artifacts on images with synthetic glares and shadows and validate it on real captures. For this, we regress a per-pixel severity map of each artifact with a Mobile U-Net [42], the same architecture as for occlusion masking (Section 4.3). As an input we use the normalized image. The output has one channel for glares and one for shadows.

Dataset

We use clean, normalized images from three datasets: SynID [23], FantasyID [24] and IDNet [25]. For FantasyID, we only use the scanned documents, as the recaptured ones may already contain illumination artifacts. For each dataset, we generate about 10,000 images with glares or shadows at a resolution of 256×160. For negative examples, we added the same number of clean images without any artifacts to the training set.

Each artifact is modeled as a randomly placed, rotated and skewed Gaussian intensity field. Glares are rendered by adding white or colored light, and shadows by a multiplicative reduction of the light. Examples are shown in Figure 7.

As the target, we use a per-pixel map of the perceived severity of each artifact.

We trained our model for 150 epochs with the mean squared error loss.



Fig. 7 Examples of synthetic glares (first column) and shadows (third column) with the corresponding severity labels (second and fourth column). Images from IDNet [25]

Evaluation

We tested on one dataset per artifact, annotated manually with binary masks. For glares, we used 290 images from DLC-2021 [22], of which 81 contain glares. Its shadows are mostly weak and leave the print readable.We therefore captured 128 images ourselves, using the physical passports of the occlusion test set (Section 4.3), this time with shadows instead of occlusions.

To obtain artifact regions from our model, we threshold the predicted map and clean it with a morphological operation. We match these regions with area constraints similar to DetEval [50]: a predicted region is a true positive if at least 50 % of it is annotated, and an annotated region is found if at least 50 % of it is predicted, which tolerates the fuzzy borders of glares and shadows. In addition, we report the pixel-level IoU over all pixels of a test set and the AUC for detecting images with an artifact.

We select the morphological operation, its kernel size and the threshold by the region-level F1 score on a random sample of the synthetic data (500 images per artifact type and 500 clean images per dataset). We apply the selected parameters (Table 5) unchanged to the real images.

During testing, we compare the results with simple pixel scores as baselines, using the same post-processing and parameter selection: for glares $\min(R, G, B)$ and the CIELAB lightness L, for shadows their dark counterparts, L and $\max(R, G, B)$. We also compare with SAM 3 [51] on the normalized image, prompted with the text "glare" or "shadow", with the same parameter selection. Finally, we compare with the illumination uniformity measure of our previous work [5], which only gives a score per image, so we report only its AUC. Table 6 reports the results and Figure 8 shows examples.

Table 5 Post-processing parameters selected on the synthetic data. Glare thresholds apply to the score of the method, shadow thresholds to the remaining brightness (one minus the predicted map for our model and SAM 3, otherwise L and $\max(R, G, B)$). Time: median per image on a CPU for the map of the method, without post-processing

Set	Method	Morphology	Threshold	Time
Glare	Ours	opening 7×7	≥ 0.29	30 ms[a]
	$\min(R, G, B)$	opening 13×13	≥ 0.89	< 1 ms
	L	opening 13×13	≥ 0.94	< 1 ms
	SAM 3	opening 7×7	≥ 0.17	12 s
Shadow	Ours	opening 15×15	≥ 0.11	30 ms[a]
	$\max(R, G, B)$	opening 15×15	≤ 0.38	< 1 ms
	L	opening 15×15	≤ 0.38	< 1 ms
	SAM 3	closing 15×15	≥ 0.13	12 s

[a] Our model outputs the glare and the shadow mask at the same time.

On DLC-2021, our model detects 61% of glares with a precision of 0.65, and an AUC of 0.94. It mainly misses small glares, detecting only 8 of 31 covering less than 0.6 % of the card, compared to 48 of 61 larger ones. The simple baselines achieve lower F1 scores (0.44 and 0.43) and IoU (0.34). SAM 3 performs comparably, with an AUC of 0.96. Our previous uniformity measure performs poorly (AUC of 0.56).

On the passports, our model detects shadows with a perfect AUC of 1.00. Missed regions are mostly thin. The threshold baselines achieve lower

F1 scores (0.59 and 0.64), mainly due to false detections on hair, skin and dark print. In contrast, our model produces no false detections on the portrait in the shadow-free example. SAM 3 also achieves an AUC of 1.00, but has lower F1 and IoU scores (0.73 and 0.59). The previous uniformity measure reaches an AUC of 0.87.

We use SAM 3 only as a baseline due to its computational cost: it has approximately 850 million parameters and takes about 12 s per image and per type on a CPU, compared to 13 million parameters and 30 ms for our model, which also produces both masks at the same time.

As the native measure of each artifact, we use the maximum of the predicted map over the ID without the occluded regions, which also gives the AUC in Table 6.

5.2.2 Compression artifacts prediction

ID images in remote verification are usually compressed before they are transmitted, and strong compression can remove fine details. OFIQ [3] estimates compression artifacts with a CNN trained on face images. As compression artifacts are local and largely independent of the image content, we tested whether this model works on ID images without retraining. We use the cropped image as input instead of the normalized image, as the perspective transformation may alter the artifacts.

Dataset generation

We used two datasets for test, as they cover synthetic IDs and real captures: FantasyID [24] and frames of physical documents from MIDV-500 [6]. We compressed 3,000 images per dataset with each of five settings, so that each dataset has 15,000 compressed images in total: JPEG 2000, JPEG with 4:2:0, 4:2:2 and 4:4:4 chroma subsampling, and WebP. For each image and setting, we drew a quality parameter q uniformly from 1 to 100. For JPEG 2000, we mapped q linearly to a target PSNR between 30 and 50 dB.

Evaluation

As the original images are already compressed, we evaluated the change of the prediction between the original and the compressed image. Figure 9 shows this change against the compression strength, with a logistic fit per dataset. The curves stay flat for weak compression and drop for stronger compression. The variance is only high for q below about 10, which is unrealistic in practice. For WebP, the change follows the quality much less, a limitation of the model, which was not trained on WebP. We therefore use the OFIQ model without retraining, as it captures JPEG and JPEG 2000, the most common formats in practice.

Compression strength estimation

OFIQ [3] converts the raw model output x to a scalar score with a sigmoid function:

$$S^{\mathrm{OFIQ}}_{\mathrm{compression}} = -0.0278 + \frac{103}{1 + \exp\left(\frac{0.3308 - x}{0.092}\right)}, \quad (2)$$

where $S^{\mathrm{OFIQ}}_{\mathrm{compression}}$ is rounded to the nearest integer and limited to the range from 0 to 100. This conversion does not work for ID images, as every raw output above about 0.64 maps to 100, while the median raw output of the original images is 0.93 for both datasets. As a result, 99 % of the compressed images still map to 100, even at strong compression.

We therefore estimate the strength of the compression as the quality parameter $\hat{q}$ that explains the raw output. We first fitted an exponential curve to the change of the raw output between the original image with output x_o and the compressed image with output x:

$$x - x_o \approx \Delta(q) = -B\left(e^{-q/C} - e^{-100/C}\right). \quad (3)$$

We optimized the two parameters B and C by least squares on all images of all five compression settings. Solving for q gives the estimated quality of the compressed image relative to the original:

$$\hat{q}_{\mathrm{ref}} = -C \ln\left(\frac{x_o - x}{B} + e^{-100/C}\right), \quad (4)$$

where $\hat{q}_{\mathrm{ref}} = 100$ for $x \geq x_o$, and $\hat{q}_{\mathrm{ref}}$ is otherwise limited to the range from 0 to 100.

To test the estimate on unseen documents, we fitted the curve on all images of one dataset and estimated q for the images of the other. For JPEG and JPEG 2000, the mean absolute error between $\hat{q}_{\mathrm{ref}}$ and q is between 9.6 and 12.5. For WebP, it is 26.7 and 20.2, as the model does not detect WebP artifacts well.

Table 6 Detection results on the real test images of our model and of the baselines with the parameters of Table 5, with 95 % bootstrap confidence intervals over the images in brackets. Precision (P), recall (R) and F1 are computed on regions with the DetEval area constraints (see text), IoU on pixels, and the AUC separates images with and without the artifact by the maximum of the predicted map (for the uniformity measure of our previous work by its score)

Method	P	R	F1	IoU^a	AUC
Glare (DLC-2021, 290 images)					
Ours	0.65 [0.53, 0.78]	0.61 [0.51, 0.72]	0.63 [0.54, 0.72]	***0.48*** *[0.37, 0.59]*	0.94 [0.91, 0.97]
$\min(R, G, B)$	0.65 [0.47, 0.87]	0.34 [0.25, 0.43]	0.44 [0.34, 0.54]	*0.34 [0.24, 0.44]*	0.95 [0.91, 0.99]
L	0.67 [0.53, 0.82]	0.32 [0.22, 0.41]	0.43 [0.32, 0.52]	*0.34 [0.23, 0.45]*	**0.97** [0.93, 0.99]
SAM 3	0.70 [0.60, 0.82]	0.60 [0.50, 0.69]	**0.65** [0.57, 0.72]	***0.48*** *[0.41, 0.55]*	0.96 [0.92, 0.99]
Uniformity [5]	–	–	–	–	0.56 [0.49, 0.63]
Shadow (passports, 128 images)					
Ours	0.92 [0.86, 0.97]	0.83 [0.75, 0.91]	**0.87** [0.82, 0.92]	***0.77*** *[0.73, 0.81]*	**1.00** [1.00, 1.00]
L	0.46 [0.38, 0.53]	0.83 [0.74, 0.90]	0.59 [0.52, 0.65]	*0.72 [0.67, 0.76]*	0.86 [0.80, 0.92]
$\max(R, G, B)$	0.54 [0.44, 0.63]	0.79 [0.70, 0.87]	0.64 [0.56, 0.71]	*0.72 [0.66, 0.77]*	0.91 [0.86, 0.96]
SAM 3	0.67 [0.59, 0.75]	0.79 [0.70, 0.88]	0.73 [0.67, 0.78]	*0.59 [0.54, 0.65]*	**1.00** [1.00, 1.00]
Uniformity [5]	–	–	–	–	0.87 [0.80, 0.92]

[a] As pixel-accurate annotations of glares and shadows are very hard to obtain, the IoU may not fully reflect the detection performance within small intervals.

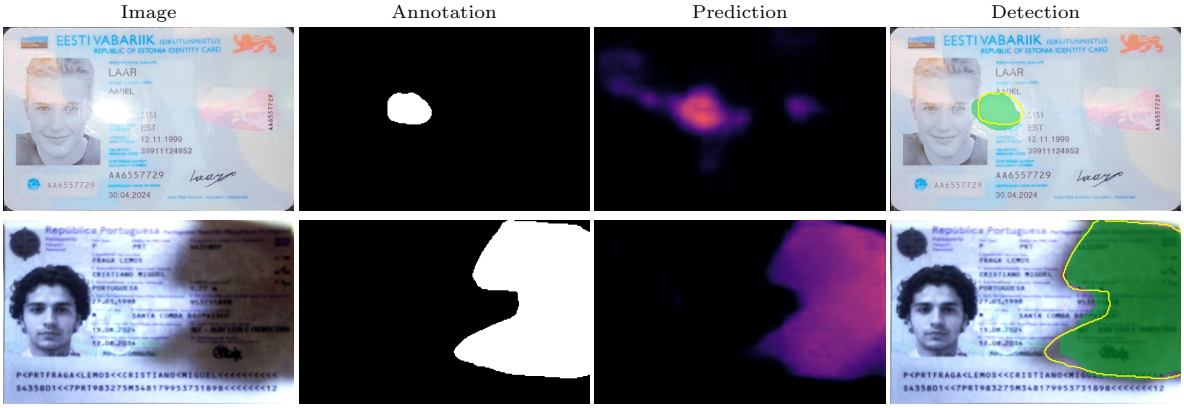


Fig. 8 Detection examples for glares on DLC-2021 (top row) and shadows on the passport dataset (bottom row): normalized image, annotation, predicted map and detection after post-processing with the parameters of Table 5 (green), with the annotation outlined (yellow)

As the original image is not available at inference, we replaced its output with the mean raw output $\bar{x}_o$ of the original images used for the fit:

$$\hat{q} = -C \ln\left(\frac{\bar{x}_o - x}{B} + e^{-100/C}\right), \qquad (5)$$

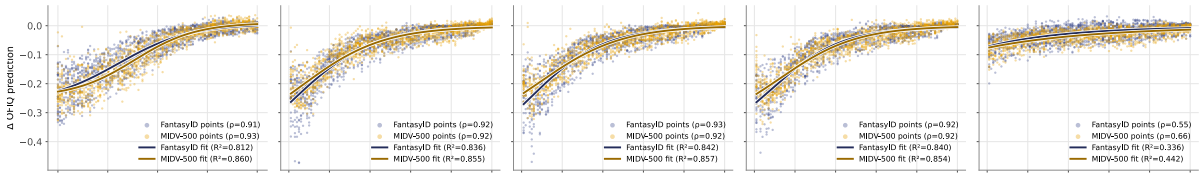
where $\hat{q} = 100$ for $x \geq \bar{x}_o$, and $\hat{q}$ is otherwise limited to the range from 0 to 100. We use $\hat{q}$ as the native measure and tested it in the same way (Figure 10). For JPEG and JPEG 2000, $\hat{q}$ follows q with a Spearman correlation between 0.89 and 0.92 and a mean absolute error between 10.5 and 13.9, so the estimate transfers between synthetic IDs and real captures. This error is only slightly larger than with the original image, although $\hat{q}$ also reflects the unknown compression of the original. For WebP, the correlation is only 0.67 on FantasyID and 0.61 on MIDV-500, as the OFIQ model detects WebP artifacts poorly, and strongly compressed images are estimated too high. The estimate is therefore suited for JPEG and JPEG 2000, while WebP would require retraining the model.

We then fitted the final curve on all images of both datasets, which gives $B = 0.246$, $C = 38.79$ and $\bar{x}_o = 0.932$:

$$\hat{q} = -38.79 \ln\left(\frac{0.932 - x}{0.246} + 0.076\right). \qquad (6)$$

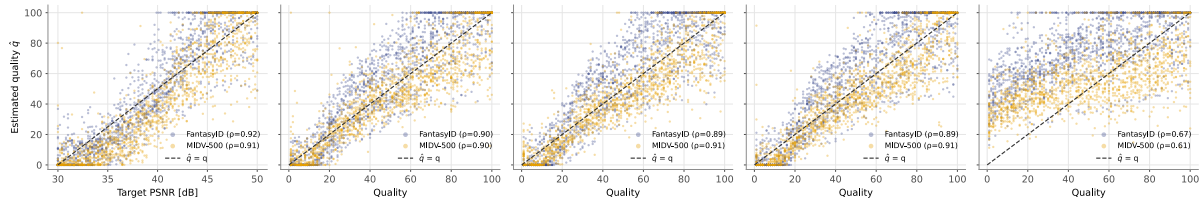
5.2.3 Sharpness estimation

Blur caused by defocus, camera shake or a low capture resolution can make small text unreadable.



(a) JPEG 2000 (b) JPEG 4:2:0 (c) JPEG 4:2:2 (d) JPEG 4:4:4 (e) WebP

Fig. 9 Change of the OFIQ compression artifact prediction against the quality parameter q (for JPEG 2000: the target PSNR in dB) for each compression setting. The legends give the Spearman correlation ρ with q per dataset, and the top row shows a logistic fit $y = a + b/(1 + e^{(c-x)/d})$ and its R^2 per dataset. For readability, each plot shows a random subset of 1,500 images per dataset, while ρ and the fits are computed on all images



(a) JPEG 2000 (b) JPEG 4:2:0 (c) JPEG 4:2:2 (d) JPEG 4:4:4 (e) WebP

Fig. 10 Estimated quality $\hat{q}$ without the original image (curve fitted on the other dataset) against the quality parameter q (for JPEG 2000: the target PSNR in dB) for each compression setting, with the Spearman correlation ρ per dataset. The dashed line marks $\hat{q} = q$. For readability, each plot shows a random subset of 1,500 images per dataset, while ρ is computed on all images

We measure this loss as the SSIM [52] between the image and its sharp original. OFIQ [3] estimates sharpness with a random forest on handcrafted features from Laplacian, Sobel and mean difference filters, trained on face images. Similar to the compression artifacts (Section 5.2.2), we first tested whether this model predicts the SSIM on ID images without retraining.

Dataset

We used the frames of physical documents from MIDV-500 [6]. We used only real captures, as synthetic IDs hold much more fine detail, and a light blur reduces the SSIM significantly. We blurred each frame with one or two stages, each of which applies a Gaussian blur (defocus), a down- and upscaling with a random interpolation method (resolution loss) or a motion blur (camera shake). As the ground truth, we computed the SSIM between the normalized original and the normalized blurred image (1000×631 pixels). This gives 14,160 pairs of original and blurred images, as we left out the 240 frames in which the ID lies almost completely outside the image. Figure 11 shows examples of the blur.

As for the compression artifacts, we evaluated the change of the score between the original and the blurred image on all 14,160 pairs, as the originals are not perfectly sharp. The change of the OFIQ score only moderately follows the SSIM, with a Spearman correlation of 0.54 (Figure 13). We therefore trained a new model.

Model and training

We tried to first fit a new random forest to the same features as the OFIQ model. As the absolute sharpness of the original images is unknown, we first trained a pairwise random forest that predicts which of two images is sharper (Figure 12). The SSIM of the MIDV-500 pairs varies little, so we scaled it linearly to s, with the 5th percentile of the training pairs at 0 and their 95th percentile at 1, and clipped it to $[0, 1]$. A pair of an original and its blurred image has the target $1 - s/2$, the reverse order $s/2$, and an image paired with itself 0.5. The sharpness of an image is then its mean pairwise prediction against a reference set of 3,000 random original and blurred images. As this needs 3,000 predictions per image, we distilled the pairwise model into a second random forest that predicts this score from the features of a single image, as done for intra-class comparison-based face image quality assessment [53]. The distilled model takes 24 ms per image on a CPU. To test



(a) Original (b) SSIM 0.83 (c) SSIM 0.90 (d) SSIM 0.73

Fig. 11 Detail of a normalized MIDV-500 frame and its blurred versions, with the SSIM to the original: (a) original, (b) Gaussian blur with $\sigma = 3$ pixels, (c) down- and upscaling by a factor of 0.25 with bicubic interpolation, and (d) motion blur with a kernel size of 25×25 pixels

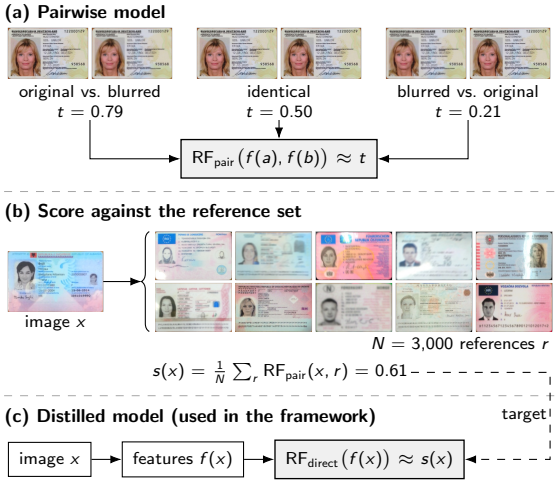


Fig. 12 Training of the new sharpness model on MIDV-500: (a) a pairwise random forest predicts which of two images is sharper, with targets from the scaled SSIM s of original and blurred images and 0.5 for identical images; (b) the sharpness of an image is its mean pairwise prediction against a reference set; (c) a second random forest is distilled from these scores and predicts the sharpness from a single image

on unseen documents, we used a five-fold cross-validation over the document types, so that each pair is scored by models that were not trained on its document type. The final model included in the framework is trained on all pairs.

Evaluation

We compared our model with OFIQ and with Tenengrad [54], a classic focus measure that is the mean squared Sobel gradient magnitude of the grayscale normalized image. The change of the score of our model follows the SSIM with a Spearman correlation of 0.81, which is higher than OFIQ (0.54) and Tenengrad (0.65) (Figure 13). Our model scores the original above its blurred version in 98 % of the pairs, compared to 95 % for Tenengrad and 86 % for OFIQ.

Example task: text legibility

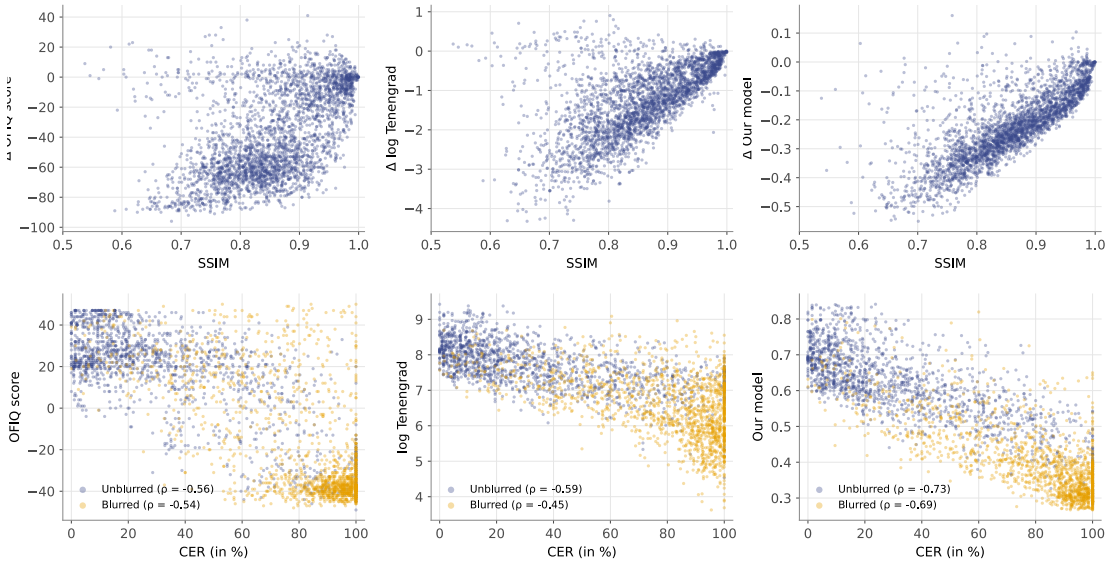
As an example of how the score can be used and converted into a scalar score, we compared it with the text legibility on a random sample of 2,000 pairs. We drew the sample from the 45 document types with annotated Latin text. We read the normalized original and blurred images with the Latin model of EasyOCR [55]. As the legibility score, we used the character error rate (CER), which is the edit distance computed by edlib [56] divided by the length of the annotated text fields of the document. We compared each field with $l \geq 4$ letters or digits to the OCR detection with the lowest CER over all consecutive substrings. The unblurred frames vary in focus, motion blur and resolution, so they show how the score works on real captures. On these frames, our model follows the CER with a Spearman correlation of -0.73, compared to -0.56 for OFIQ and -0.59 for Tenengrad (Figure 13, bottom row). On the original and blurred images together, the correlation is -0.86, compared to -0.74 and -0.71.

5.2.4 Over- and underexposure

Over- and underexposure reduce the contrast of the print, and strong overexposure clips bright regions, so that their content is lost. In our previous work [5], we used the exposure measures of OFIQ [3], which count the share of pixels with a luminance in $[0, 25]$ for underexposure and in $[247, 255]$ for overexposure. We did not test these ranges on ID images, so we evaluated them on labeled real captures and compared them with the mean gray value of the ID.

Evaluation

We used 500 normalized frames of physical documents from MIDV-500 [6]. We labeled each frame as underexposed, clean or overexposed, which



(a) OFIQ (b) Tenengrad (c) Ours

Fig. 13 Change of the score from the original to the blurred image against the SSIM on the 14,160 MIDV-500 pairs, with the Spearman correlation ρ: (a) OFIQ ($\rho = 0.54$), (b) Tenengrad (log ratio, $\rho = 0.65$), and (c) our model ($\rho = 0.81$), with scores from the validation folds. Bottom row: score of the unblurred and blurred images of the OCR example against the OCR character error rate (CER). For readability, each plot shows 3,000 data points

gives 28 underexposed, 378 clean and 94 overexposed frames.

All measures use the normalized image without the portrait (Section 4.4), the occluded regions (Section 4.3) and the detected glares and shadows (Section 5.2.1). As the gray value, we use the mean of the R, G and B values instead of the luminance, as the luminance weights blue by only 0.07, so that a pure blue print has a luminance of 18 and would count as underexposed. The underexposed frames are dim rather than black. Only 0.1 % of their pixels lie in the OFIQ range $[0, 25]$, compared to less than 0.1 % for the other frames, so this range hardly separates them (AUC 0.58, bootstrap 95 % confidence interval 0.46 to 0.69). For overexposure, the OFIQ range $[247, 255]$ holds 25.1 % of the pixels of overexposed frames and 0.8 % of the other frames (AUC 0.83, 0.77 to 0.88). The largest gaps lie at $t = 168$ for underexposure and $t = 211$ for overexposure, so the labels differ in the overall brightness and not only in clipped pixels. We therefore use the mean gray value m of the ID as the native measure, which separates both better, with an AUC of 0.92 (0.89 to 0.95) for underexposure and 0.88 (0.82 to 0.93) for overexposure. Excluding the glares and shadows did not change the AUC of the mean gray value, but we excluded them regardless, so that a glare or shadow is assessed only by the illumination artifact measure.

5.3 ID-related measures

The ID-related measures assess how the ID is placed in the image: its alignment with the camera, its size, whether it lies completely inside the image, whether it is occluded, and whether only one ID is present. We compute them from the detected corners (Section 4.1) and the occlusion mask (Section 4.3).

5.3.1 Alignment

The alignment measures how strongly perspective distorts the identity document. We use the local scaling of Shemiakina et al. [12]. From the detected corners (Section 4.1), we compute the homography that maps the normalized image to the original image. A small circle at a point p of the normalized image maps to an ellipse in the original image, with semi-axes proportional to the singular values $\sigma_{\min}(p) \leq \sigma_{\max}(p)$ of the Jacobian of the homography at p. A tilt by the angle θ shortens the circle by $\cos\theta$ in one direction, so we estimate the local tilt as $\arccos(\sigma_{\min}(p)/\sigma_{\max}(p))$. In their original work, Shemiakina et al. [12] used

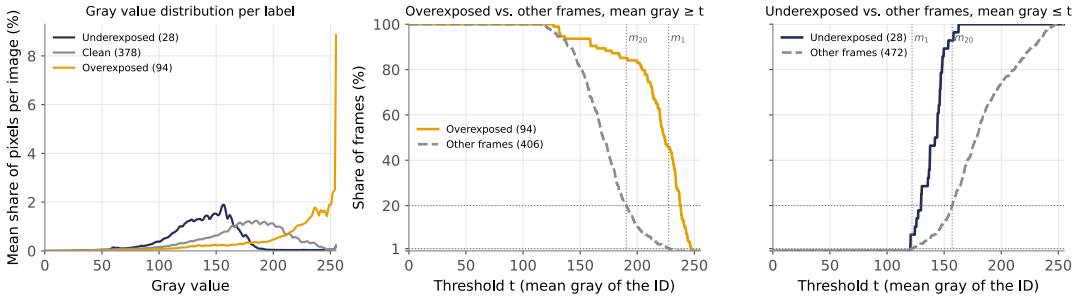


Fig. 14 Gray values of the 500 labeled MIDV-500 frames: (left) mean share of pixels per frame at each gray value for each label, (middle) share of overexposed and of all other frames with a mean gray value of the ID $\geq t$, and (right) share of underexposed and of all other frames with a mean gray value $\leq t$. The dotted lines mark the thresholds m_1 and m_{20} of the scalar conversion (Online Resource 1, Section S7), at which 1 % and 20 % of the other frames are flagged (FPR)

the minimum of all $\sigma_{\min}(p)/\sigma_{\max}(p)$ for the maximum distortion of any p. As we want to measure how well the ID is aligned with the camera, we instead take the mean of the angle θ over the normalized image and denote this native measure as α, which does not depend on the size or the in-plane rotation of the ID. However, it depends on the aspect ratio of the document. We therefore calculate α for the ID-1 and the ID-3 aspect ratio [39] and keep the smaller value. We tested α on 3,000 quadrilaterals with a simulated camera at random tilts up to 75° and with Gaussian noise of 3 pixels on the corners. α has a Spearman correlation of 0.99 with the true tilt and a mean absolute error of 2.7° (without noise: 2.4°).

5.3.2 ID size

The ID size measures how much of the image the ID fills, as a small ID has a low resolution and a very large ID is likely cut off by the image border. As the shape and orientation of an ID limit how much of an image it can fill, we compare its size to the largest size possible in the original image. For this, we move the quadrilateral spanned by the detected corners (Section 4.1) to the center of the image and scale it around its center until the first corner touches the image border. With the scaling factor s, the native measure is $r = 1/s^2$, the area of the quadrilateral divided by the area of the scaled one. This works for any aspect ratio and rotation, and r exceeds 100 % if the ID does not fit into the image.

5.3.3 ID in frame

The ID in frame measure reports how much of the ID lies outside the image. The native measure x is the share of the area of the same quadrilateral that lies outside the original image.

5.3.4 Occlusion

The occlusion measure reports how much of the ID is covered, as occluded regions of the ID cannot be read or assessed. Similar to the face occlusion measure of OFIQ [3], we use the occluded share of the ID as the native measure, with the ID-1 shape instead of the landmarked face region:

$$o = \frac{|F \setminus M|}{|F|}, \tag{7}$$

where F is the ID-1 shape, warped to the corners of the ID, and M is the visible region predicted by the occlusion masking (Section 4.3).

5.3.5 Only one ID present

A further ID in the image makes it unclear which ID is to be assessed. Similar to the single face present measure of OFIQ [3], we use the areas of the detected ID quadrilaterals instead of the face boxes. The native measure is the ratio a_2/a_1, where a_1 is the area of the largest and a_2 the area of the second largest detected ID, with $a_2 = 0$ if only one ID is detected. We count every detection with a confidence of at least 0.1, the threshold of our detector (Section 4.1).

5.4 Scalar conversion

Similar to OFIQ [3], we convert each native measure x to a scalar score S from 0 (worst) to 100 (optimal). We use a linear function between two anchor points for all measures, as it is easy to interpret and to adapt:

$$S = 100 \cdot \frac{x_0 - x}{x_0 - x_{100}}, \tag{8}$$

where x_{100} is the native value at which the score reaches 100 and x_0 the value at which it reaches 0, and S is clipped to $[0, 100]$. The ID size uses two such pieces on both sides of an ideal size. We set the anchor points in one of four ways:

- *Expected error*: for the preprocessing confidence, we fitted the error of the preprocessing step against the confidence and set the anchors at at tolerable error and at a not tolerable error.
- *False positive rate (FPR)*: for illumination artifacts and exposure, we set the anchors at the native values that yield a specific FPR on out test set.
- *Example task*: for sharpness, we set them at two character error rates of an OCR model.
- *Direct limits*: for compression and the ID-related measures, the native measure has a direct meaning, such as a quality parameter, an angle or a share of the ID, so we set the anchors directly.

Online Resource 1 gives the conversion of each measure. As the acceptable level of a defect depends on the application and the exact values are dataset depended, these anchor points are examples that an operator can adapt.

6 Limitations

OIDIQ assesses a single image of the front side of an ID with a portrait. Documents without a portrait, such as the back side of many ID cards, need a new corner detector trained on them. We also did not evaluate the measures on presentation attacks, where some defects, such as moiré or glares on a screen, may be the evidence of an attack rather than a quality problem.

The parameters of the score conversions are examples, derived from the error of the native measure, the false positive rate on our data, or one example task, and an operator has to adapt them to the tolerable error of the application. Apart from text legibility for sharpness, we did not relate the scores to the performance of downstream tasks, as we wanted to keep the framework general. OIDIQ reports the scores separately and does not combine them into a unified quality score, as OFIQ does for faces.

Future work should address these points: relate the measures to downstream tasks, such as face comparison and PAD, and study how to combine the separate scores into a unified quality score.

7 Conclusion

We presented OIDIQ, an open framework that assesses the quality of ID images with separate measures, each converted to a score from 0 to 100, similar to OFIQ for face images. The preprocessing detects the ID, normalizes it and masks occlusions, and reports its own reliability through confidence measures. Our corner detector generalizes to unseen document types and rarely fires on other documents, while learned document localizers detect many of them. Our occlusion model comes close to SAM 2.1 at a fraction of its runtime, and our glare and shadow detector, trained only on synthetic artifacts, works on real captures and is on par with SAM 3 for glares and better for shadows. Of the OFIQ models, the compression model transfers to IDs for JPEG and JPEG 2000 with a new estimate of the compression strength, while the sharpness model does not follow the blur of IDs well, so we trained a new model that also predicts the OCR error better. For exposure, the mean gray value of the ID separates under- and overexposed captures much better than the OFIQ pixel ranges.

Supplementary information. Online Resource 1: Scalar conversion of each quality measure, with the fits and data on which the anchor points are based.

Statements and Declarations

Funding. This research work has been funded by the European Union (EU) under G.A. no. CarMen (101168325), and the German Federal Ministry of Education and Research and the Hessian Ministry of Higher Education, Research, Science and the Arts within their joint support of the

National Research Center for Applied Cybersecurity ATHENE.

Competing interests. The authors have no competing interests to declare that are relevant to the content of this article.

Data availability. The code of OIDIQ and all models will be made publicly available on GitHub upon acceptance. The public datasets used in this work (Table 1) are available from their respective authors. All newly created datasets and annotations used in this work will be available upon acceptance.

References

[1] Temoshok, D., et al.: Digital identity guidelines: Identity proofing and enrollment. Technical Report NIST SP 800-63A-4 (2nd Public Draft), National Institute of Standards and Technology (August 2024). https://doi.org/10.6028/NIST.SP.800-63A-4.2pd . https://doi.org/10.6028/NIST.SP.800-63A-4.2pd

[2] Ruiz, E.M., et al.: Identity Card Presentation Attack Detection: A Systematic Review (2025). https://arxiv.org/abs/2511.06056

[3] Open source face image quality (ofiq) – projektabschlussbericht. Project final report, Bundesamt für Sicherheit in der Informationstechnik (BSI) (july 2024). Reference implementation for ISO/IEC 29794-5. https://www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/OFIQ/Projektabschlussbericht_OFIQ_1_0.pdf

[4] Alaei, A., et al.: Document image quality assessment: A survey. ACM Comput. Surv. **56**(2) (2023) https://doi.org/10.1145/3606692

[5] Grote, G., Tapia, J.E., Rathgeb, C.: Image quality assessment of identity cards using measures from open face image quality. In: 14th Int. Workshop on Biometrics and Forensics (IWBF), pp. 1–6 (2026). https://doi.org/10.1109/IWBF68042.2026.11558141

[6] Arlazarov, V.V., *et al.*: MIDV-500: a dataset for identity document analysis and recognition on mobile devices in video stream. Computer Optics **43**(5), 818–824 (2019) https://doi.org/10.18287/2412-6179-2019-43-5-818-824

[7] Alaei, A., *et al.*: "quality" vs. "readability" in document images: Statistical analysis of human perception. In: 13th IAPR Int. Workshop on Document Analysis Systems (DAS), pp. 363–368 (2018). https://doi.org/10.1109/DAS.2018.42

[8] Shahkolaei, A., Beghdadi, A., Cheriet, M.: Blind quality assessment metric and degradation classification for degraded document images. Signal Processing: Image Communication **76**, 11–21 (2019) https://doi.org/10.1016/j.image.2019.04.009

[9] Ye, P., Doermann, D.: Learning features for predicting ocr accuracy. In: Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012), pp. 3204–3207 (2012)

[10] Kang, L., *et al.*: A deep learning approach to document image quality assessment. In: IEEE Int. Conf. on Image Processing (ICIP), pp. 2570–2574 (2014). https://doi.org/10.1109/ICIP.2014.7025520

[11] Kumar, J., Chen, F., Doermann, D.: Sharpness estimation for document and scene images. In: Proc. of the 21st Int. Conf. on Pattern Recognition (ICPR2012), pp. 3292–3295 (2012)

[12] Shemiakina, J., et al.: A method of image quality assessment for text recognition on camera-captured and projectively distorted documents. Mathematics **9**(17) (2021) https://doi.org/10.3390/math9172155

[13] Bulatov, K., Matalov, D., Arlazarov, V.V.: MIDV-2019: challenges of the modern mobile-based document OCR. In: Osten, W., Nikolaev, D.P. (eds.) Twelfth Int. Conf. on Machine Vision (ICMV), vol. 11433, p. 114332. SPIE, Amsterdam, Netherlands (2020). https://doi.org/10.1117/12.2558438 . Int. Society for Optics and Photonics. https://doi.org/10.1117/12.2558438

[14] Mittal, A., Moorthy, A.K., Bovik, A.C.: No-reference image quality assessment in the spatial domain. IEEE Transactions on Image Processing **21**(12), 4695–4708 (2012) https://doi.org/10.1109/TIP.2012.2214050

[15] Liu, X., Van De Weijer, J., Bagdanov, A.D.: RankIQA: Learning from rankings for no-reference image quality assessment. In: IEEE

Int. Conference on Computer Vision (ICCV), pp. 1040–1049 (2017). https://doi.org/10.1109/ICCV.2017.118

[16] Ke, J., *et al.*: Musiq: Multi-scale image quality transformer. In: 2021 IEEE/CVF Int. Conf. on Computer Vision (ICCV), pp. 5128–5137 (2021). https://doi.org/10.1109/ICCV48922.2021.00510

[17] Sun, W., *et al.*: Blind quality assessment for in-the-wild images via hierarchical feature fusion and iterative mixed database training. IEEE Journal of Selected Topics in Signal Processing **17**(6), 1178–1192 (2023) https://doi.org/10.1109/JSTSP.2023.3270621

[18] Grother, P., Tabassi, E.: Performance of biometric quality measures. IEEE Transactions on Pattern Analysis and Machine Intelligence **29**(4), 531–543 (2007) https://doi.org/10.1109/TPAMI.2007.1019

[19] Tabassi, E., et al.: NIST fingerprint image quality 2 (NFIQ 2). NIST Interagency/Internal Report 8382, National Institute of Standards and Technology, Gaithersburg, MD (2021). https://doi.org/10.6028/NIST.IR.8382

[20] Schlett, T., et al.: Face image quality assessment: A literature survey. ACM Comput. Surv. **54**(10s) (2022) https://doi.org/10.1145/3507901

[21] Merkle, J., et al.: State of the Art of Quality Assessment of Facial Images (2022). https://arxiv.org/abs/2211.08030

[22] Polevoy, D.V., et al.: Document liveness challenge dataset (dlc-2021). Journal of Imaging **8**(7) (2022) https://doi.org/10.3390/jimaging8070181

[23] Tapia, J.E., *et al.*: Syn-IDPass: Passport synthetic dataset for presentation attack detection. In: IEEE Int. Joint Conf. on Biometrics (IJCB), pp. 1–9 (2025). https://doi.org/10.1109/IJCB65343.2025.11411263

[24] Korshunov, P., *et al.*: Fantasyid: A dataset for detecting digital manipulations in id-documents. In: IEEE Int. Joint Conf. on Biometrics (IJCB), pp. 1–9 (2025). https://doi.org/10.1109/IJCB65343.2025.11411292

[25] Xie, L., *et al.*: IDNet: A novel identity document dataset via few-shot and quality-driven synthetic data generation. In: IEEE Int. Conf. on Big Data (BigData), pp. 2244–2253 (2024). https://doi.org/10.1109/BigData62323.2024.10825017

[26] Benalcazar, D., *et al.*: Synthetic id card image generation for improving presentation attack detection. IEEE Transactions on Information Forensics and Security **18**, 1814–1824 (2023) https://doi.org/10.1109/TIFS.2023.3255585

[27] Tapia, J.E., *et al.*: Second competition on presentation attack detection on id card. In: IEEE Int. Joint Conf. on Biometrics (IJCB), pp. 1–10 (2025). https://doi.org/10.1109/IJCB65343.2025.11411073

[28] Yáñez, C., Tapia, J.: Image quality assessment on identity documents. In: Int. Conf. of the Biometrics Special Interest Group (BIOSIG), pp. 1–5 (2021). https://doi.org/10.1109/BIOSIG52210.2021.9548294

[29] Schulz, D., *et al.*: Identity documents image quality assessment. In: 30th European Signal Processing Conf. (EUSIPCO), pp. 1017–1021 (2022). https://doi.org/10.23919/EUSIPCO55093.2022.9909773

[30] Jocher, G., et al.: Ultralytics YOLO26: Unified Real-Time End-to-End Vision Models (2026). https://arxiv.org/abs/2606.03748

[31] Lin, T.-Y., *et al.*: Microsoft coco: Common objects in context. In: Fleet, D., *et al.*(eds.) Computer Vision – ECCV 2014, pp. 740–755. Springer, Cham (2014)

[32] Suvorov, R., *et al.*: Resolution-robust large mask inpainting with fourier convolutions. In: IEEE/CVF Winter Conf. on Applications of Computer Vision (WACV), pp. 3172–3182 (2022). https://doi.org/10.1109/WACV51458.2022.00323

[33] Park, S., et al.: CORD: A consolidated receipt dataset for post-ocr parsing (2019)

[34] sc: business cards Dataset. Roboflow. visited on 2026-10-29 (2026). https://universe.roboflow.com/sc-8qaqk/business-cards-n3hbf

[35] Ronchi, M.R., Perona, P.: Benchmarking and error diagnosis in multi-instance pose estimation. In: IEEE Int. Conf. on Computer Vision (ICCV), pp. 369–378 (2017). https://doi.org/10.1109/ICCV.2017.48

[36] Gu, K., *et al.*: On the calibration of human pose estimation. In: Salakhutdinov, R., *et al.*(eds.) Proc. of the 41st Int. Conf. on Machine Learning. Proceedings of Machine Learning Research, vol. 235, pp. 16530–16547 (2024). https://proceedings.mlr.press/v235/gu24a.html

[37] Tropin, D.V., et al.: Advanced hough-based method for on-device document localization. Computer Optics **45**(5) (2021) https://doi.org/10.18287/2412-6179-co-895

[38] Yuan, Z.: DocAligner. GitHub. GitHub repository (2023). https://github.com/DocsaidLab/DocAligner

[39] International Organization for Standardization: Identification cards – physical characteristics. Standard ISO/IEC 7810:2019, International Organization for Standardization, Geneva, Switzerland (2019)

[40] Iglovikov, V.: ternaus/midv-500-models v0.0.2. Zenodo (2020). https://doi.org/10.5281/zenodo.4263532

[41] Ravi, N., *et al.*: SAM 2: Segment anything in images and videos. In: Int. Conf. on Learning Representations (ICLR) (2025)

[42] Jing, J., *et al.*: Mobile-unet: An efficient convolutional neural network for fabric defect detection. Textile Research Journal **92**(1-2), 30–42 (2022) https://doi.org/10.1177/0040517520928604 https://doi.org/10.1177/0040517520928604

[43] Zhang, C., et al.: Faster Segment Anything: Towards Lightweight SAM for Mobile Applications (2023)

[44] Borges, B.L.C., Pacheco, B.M., Silva, D.: Soft dice confidence: A near-optimal confidence estimator for selective prediction in semantic segmentation. Machine Learning **115**(7), 176 (2026) https://doi.org/10.1007/s10994-026-07096-w

[45] Rodin, D., Orlov, N.: Fast glare detection in document images. In: Int. Conf. on Document Analysis and Recognition Workshops (ICDARW), vol. 7, pp. 6–9 (2019). https://doi.org/10.1109/ICDARW.2019.60123

[46] Hou, S., *et al.*: Text-aware single image specular highlight removal. In: Pattern Recognition and Computer Vision: 4th Chinese Conference, PRCV 2021, Beijing, China, October 29 – November 1, 2021, Proceedings, Part IV, pp. 115–127. Springer, Berlin, Heidelberg (2021). https://doi.org/10.1007/978-3-030-88013-2_10

[47] Li, F., *et al.*: A comprehensive survey of specularity detection: state-of-the-art techniques and breakthroughs. Artificial Intelligence Review **58**(7), 218 (2025) https://doi.org/10.1007/s10462-025-11233-7

[48] Bell, S., Bala, K., Snavely, N.: Intrinsic images in the wild. ACM Trans. Graph. **33**(4) (2014) https://doi.org/10.1145/2601097.2601206

[49] Li, J.: A deep learning method for document shadow removal with sobel prior under mask supervision, vol. 39, pp. 29584–29586 (2025). https://doi.org/10.1609/aaai.v39i28.35333

[50] Wolf, C., Jolion, J.-M.: Object count/area graphs for the evaluation of object detection and segmentation algorithms. Int. J. Doc. Anal. Recognit. **8**(4), 280–296 (2006)

[51] Carion, N., *et al.*: SAM 3: Segment anything with concepts. In: Int. Conf. on Learning Representations (ICLR) (2026)

[52] Wang, Z., *et al.*: Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing **13**(4), 600–612 (2004) https://doi.org/10.1109/TIP.2003.819861

[53] Ou, F.-Z., *et al.*: Sdd-fiqa: Unsupervised face image quality assessment with similarity distribution distance. In: IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 7666–7675 (2021). https://doi.org/10.1109/CVPR46437.2021.00758

[54] Krotkov, E.: Focusing. Int. Journal of Computer Vision **1**(3), 223–237 (1988) https://doi.org/10.1007/BF00127822

[55] Jaided AI: EasyOCR: Ready-to-use OCR with 80+ supported languages. https://github.com/JaidedAI/EasyOCR (2020)

[56] Šošić, M., Šikić, M.: Edlib: a C/C++ library for fast, exact sequence alignment using edit distance. Bioinformatics **33**(9), 1394–1395 (2017) https://doi.org/10.1093/bioinformatics/btw753

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- onlineresource1.pdf