Weakly Supervised Training for Hologram Verification in Identity Documents

Glen Pouliquen, Guillaume Chiron, Joseph Chazalon, Thierry Géraud, Ahmad Montaser Awal

Weakly Supervised Training for Hologram Verification in Identity Documents

Glen Pouliquen[1,2], Guillaume Chiron[1], Joseph Chazalon[2], Thierry Géraud[2], and Ahmad Montaser Awal[1]

[1] IDnow AI & ML Center of Excellence, Cesson-Sévigné, France
name.surname@idnow.io
[2] EPITA Research Lab. (LRE), Le Kremlin-Bicêtre, France
name.surname@epita.fr

Abstract. We propose a method to remotely verify the authenticity of Optically Variable Devices (OVDs), often referred to as "holograms", in identity documents. Our method processes video clips captured with smartphones under common lighting conditions, and is evaluated on two public datasets: MIDV-HOLO and MIDV-2020. Thanks to a weakly-supervised training, we optimize a feature extraction and decision pipeline which achieves a new leading performance on MIDV-HOLO, while maintaining a high recall on documents from MIDV-2020 used as attack samples. It is also the first method, to date, to effectively address the photo replacement attack task, and can be trained on either genuine samples, attack samples, or both for increased performance. By enabling to verify OVD shapes and dynamics with very little supervision, this work opens the way towards the use of massive amounts of unlabeled data to build robust remote identity document verification systems on commodity smartphones. Code is available at `https://github.com/EPITAResearchLab/pouliquen.24.icdar`.

Keywords: Know Your Consumer (KYC) · Identity Documents · Hologram Verification · Weakly Supervised Learning · Contrastive Loss

1 Introduction

Often called KYC (Know Your Customer), remotely verifying the authenticity of identity documents is a critical point for building online trust. This is an increasingly regulated process which relies on identity documents, among other proofs, to establish the link between an online identity and a real state-backed one. This linking usually requires checking that the document is original and was not altered. The photography of the bearer is of paramount importance here to ensure that the user of a remote system is the intended one. Optically variable devices (OVDs), commonly referred to as "holograms" and illustrated in Figure 2, are powerful tools to secure physical documents in line with the recommendation of the EU council [14], among others. Built using elaborated

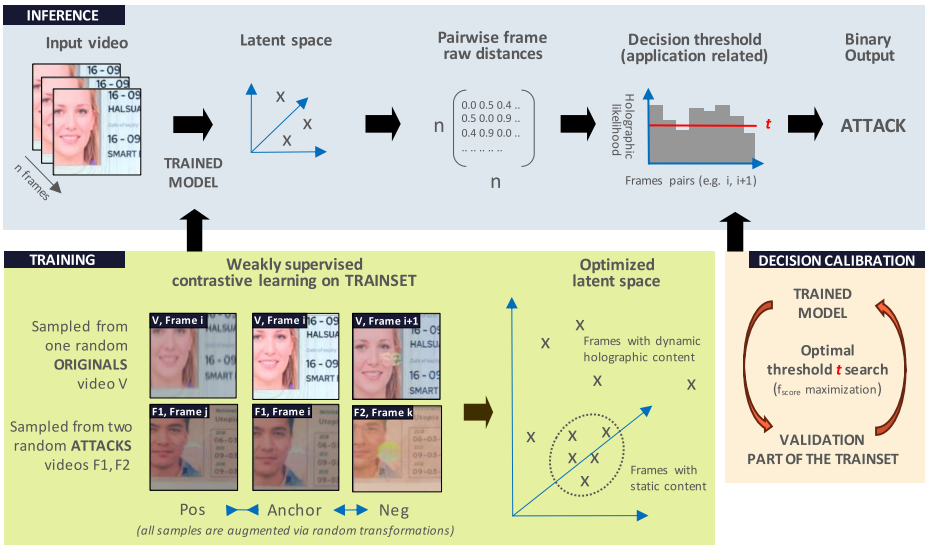

Fig. 1. Proposed approach overview, involving 1) the weakly supervised TRAINING with a specific data selection strategy over the trainset; 2) the INFERENCE pipeline extracting optimized features used afterward to compute the final "Original/Attack" decision based on a thresholding of pairwise distances. The DECISION illustrates how the threshold is calibrated over the validation part of the train set.

and undisclosed optical techniques, these devices exhibit rich visual behaviors when viewing and/or illuminating conditions (angle, light color, etc.) change. They are embedded in a wide spectrum of sensitive elements, i.e. not only identity documents but also banknotes or tamper-proof labels (e.g. for medical drugs), and contribute to ensure: 1. **Integrity:** they cannot be removed without altering their properties, making tampering very challenging; 2. **Authenticity:** they are very difficult to forge, making the creation of fraudulent documents equally hard.

Despite the widespread use of holograms, automating their remote verification in the context of an automatic enrollment, whether it is to open an account in an online bank or to contract a loan, poses many challenges. Indeed, such visual objects were primarily designed for manual inspection, sometimes using special tools like magnifiers or dedicated light sources. As a result, automated remote validation is limited in many aspects: acquisition is often performed using commodity smartphones under uncontrolled ambient light to capture macroscopic and visible patterns, while following simple interactive scenarios. Nevertheless, verifying holographic devices from a video is possible to some extent, and many recent works and datasets contributed to this effort.

While, in the general forgery detection literature, several approaches try to detect falsification clues [16], others follow the opposite (yet complementary) direction of checking whether clues of authenticity and integrity are present [7]. This work contributes to the latter: we propose a method to control the presence

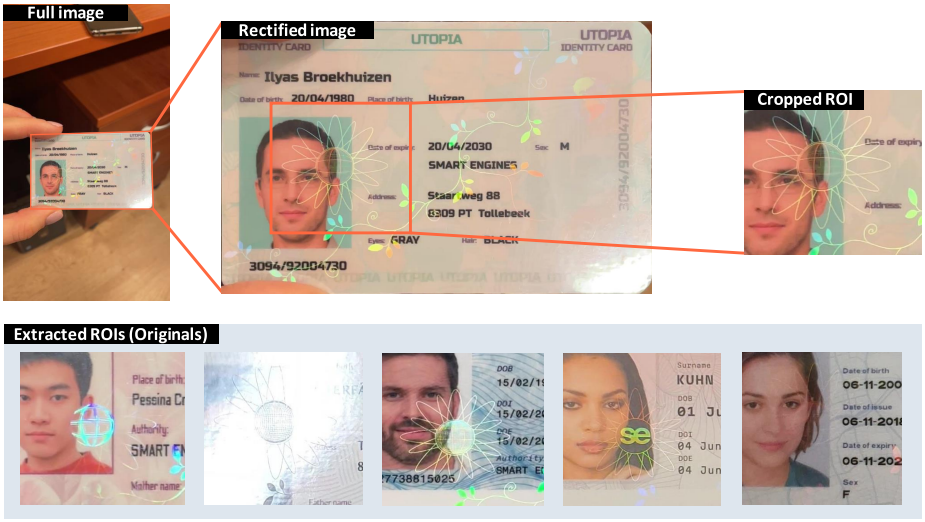

Fig. 2. In MIDV Holo dataset, documents are captured in different places involving various backgrounds and lightning conditions (left). Document quads are annotated on all images allowing rectifications (center). Additionally, we propose to define a region of interest containing part of the face and the holograms in charge of securing it (right). Extracted Regions of Interest (ROIs) from sampled labeled as "Originals" (below) contain more or less visible holographic content. *Identities* (names and faces) are synthetic.

of some holographic content at specific positions of a document (e.g. photography area), and address the problem of photography replacement, which was introduced in MIDV-Holo [18] but not yet addressed (to our knowledge). After a detailed review of related approaches and datasets (Section 2), we introduce our key contribution: a new method to detect and validate holographic content, whose feature extraction is trained in a weakly-supervised fashion (i.e., not requiring a precise labeling of each video frame with the particular visual appearance of a hologram), and which outperforms the original approach on public datasets (Section 3). For practical reasons, we also propose an updated experimental protocol which specifies, among others, training, validation and test sets for the MIDV-Holo dataset, as well as a public, open-source reimplementation of the approach proposed in the original MIDV-Holo publication [18], with systematic optimization of the calibration of the decision function (Section 4). Our approach (illustrated in Figure 1) is carefully evaluated on several public datasets, over several runs, and an ablation study is conducted to challenge the benefits of every aspect of our method (Section 5). The code to reproduce our results is publicly available at `https://github.com/EPITAResearchLab/pouliquen.24.icdar`.

2 Related works

Optically Variable Devices (OVDs) are often built using polarized inks or diffraction grating — a network of microscopic reflective structures engraved in the thickness of some transparent layer. We refer the reader to the MIDV-Holo publication [18] for a concise introduction on the optical design of these objects. These OVDs exhibit continuous, sometimes rapid, transitions among a virtually infinite set of visual states when changing the relative positions of the light source(s), the camera and the document. We can consider such space of visual appearances as a sort of manifold which we navigate by changing visualization conditions. Such model is valuable to identify the 3 fundamental visual features an automated method can check:

1. **Appearance Conformity:** *Can a particular visual appearance (shape and color) be generated by a genuine OVD?* — This can be viewed as assessing how far a particular sample (usually an image) is from the real manifold of a given hologram, and use a distance threshold as verification criterion. Implementing this control enables to detect attacks with no hologram or with a different hologram shape but, if used alone, would be tricked by simple static copies of the expected hologram.
2. **Appearance Coverage:** *How well do a particular set of visual appearances (usually captured as a video) matches the set of possible ones?* — This can be viewed as measuring how well the samples obtained cover the real manifold of a given hologram. Implementing this control enables to detect attacks with static holograms but, if used without control of state conformity, would be tricked by any random holographic layer. Approaches checking only color distributions are vulnerable to this attack.
3. **Transitions Validity:** *Are the transitions between observed visual appearances consistent with expected ones?* — This can be viewed as checking whether samples obtained describe valid paths on the real manifold of a given hologram. Implementing this control enables to detect attacks with imperfect hologram imitations or rapid swapping of static holograms. The low frame rates utilized in current real remote authentication applications present a significant challenge that has not been adequately explored in the research literature.

Early approaches like the one of Hartl et al. [2] identified a discrete set of visual appearances to check for during a manual inspection of identity documents. Expected visual states were acquired and validated using a robotic arm with controlled light. While lacking automated verification, this approach proposed a practical protocol to assist a human operator during its work to validate 1. visual state conformity thanks to visual comparison, and 2. visual behavior completeness by checking every expected visual state is seen.

The work of Chapel et al. [17] proposes a way to automate the verification of visual states conformity thanks to a learned classifier based on local binary patterns (LBP) features. However, training this system requires to label each video frame with target class (visual state), which is both too expensive for

real application in our experience, and also challenging because of the frequent "mixing" of visual shapes in real OVDs. Furthermore, robustness of the static feature extractor may be a concern when background are not constant like in the case of face photos: a learned feature extractor seems necessary here.

To overcome the need for labeled frames when training a visual state classifier, some approaches relaxed the control on visual state conformity to focus instead on the validation of visual behavior completeness. The work of Kada et al. [15] opened an interesting direction by restating the problem as a semantic segmentation problem where images of documents captured as a video are first carefully registered, then a pixel-level classification is performed to predict whether a particular region belongs to a hologram. Such prediction is mainly based on statistics on the distribution of pixel color values. While the final segmentation map may be used as some visual appearance clue, this method does not check whether inter-pixel behavior is consistent, nor visual appearance conformity for a particular frame, and lacks a global decision stage.

A major step was made thanks to the work of Koliaskina et al. [18]. The authors not only reuse the same idea of semantic segmentation (also based on a static, handcrafted feature extraction) to produce a map of pixels which exhibit some "holographic behavior", but also provide a global decision stage based on a variance threshold and a first public dataset, MIDV-Holo, containing document with holographic contents along with several presentation attacks, as illustrated in Figure 2. While validated on full-size documents, the approach was not tested on the particular case of the face photo region, and was not evaluated against the photo replacement attack. Furthermore, as detailed in Section 4, this milestone publication required us to reimplement the proposed approach and specify training, validation and test splits to conduct a fair comparison with our proposed approach.

Another family of approaches tackled the problem of learning a useful embedding space, thanks to which it may be possible to both check visual state conformity and visual behavior completeness. A first example is the work of Soukup et al. [4] targeting hologram verification on banknotes. Their approach extract representations from video frames using a Convolutional Neural Network (CNN) trained with a supervised classification task. Target classes represent different visual appearances, and were captured using a LED ring that illuminated the hologram in various directions to automate the annotation process. Because of the changing nature of the background for some OVDs in identity documents (such as in the area of the face picture), such approach may need an important amount of training data to be applied in our context, which makes it impractical since it requires physical access to real documents.

Finally, a last related work is the one of Ay et al. [12] which proposed to train a Generative Adversarial Network (GAN) to capture the visual properties of some hologram. While the generative properties of such approach are very attractive, successfully training such architecture to model thin holograms on non-constant backgrounds like face pictures is a great challenge, as the network may more easily capture and generate facial feature which cover a larger pro-

portion of the area of interest. Furthermore, as final decision is performed using the discriminator network, a proper calibration would require attack samples. Another limitation of using the discriminator is that, despite the rich representation learned, this approach is limited to controlling the visual state conformity of isolated frames, and cannot check visual behavior completeness as differences between visual states cannot be measured.

Looking at these existing works, we can sketch out desirable features an automated verification method should provide, which our proposed approach tends to incorporate:

1. It should be based on a learned representation which can be tied to a particular OVD, in order to be able to capture both visual appearance and behavior information, as opposed to handcrafted, static, pixel-based feature extraction techniques.
2. Such representation should be learned in a weakly supervised way to avoid requiring manual labelling of existing frames, or physical access to a large amount of original documents and presentation attacks.
3. The learning objective should be able to guide the training even in the presence of non-constant backgrounds and thin holographic objects, like in the case of the face picture area.

3 Contrastive Learning of Hologram Representation

This section introduces our key contribution, summarized in Figure 1: a new method to detect and validate holographic content, whose representation (feature extraction) is trained in a weakly-supervised fashion; i.e. it only requires a single label ("original" or "attack", as per MIDV-Holo [18] terminology) for each video clip. For this purpose, we use a particular kind of contrastive loss which enables the training to be driven by intrinsic data properties. This relies on certain assumptions about the video clips, such as their ability to capture varied perspectives of the document. It also involves various transformations to enhance the data. The resulting representation can be shown to effectively focus on hologram regions, and can be used to assess both appearance consistency and coverage in a final decision stage considering as many video frames as necessary.

3.1 Learning Objective

To avoid the need for assigning labels to every video frame of the training set, we employ a contrastive learning objective which enables us to drive model training using intrinsic data properties (described in the next subsection). More specifically, we use a triplet loss [1] defined on a minibatch of N elements as

$$\mathcal{L}(a,p,n) = \frac{1}{N}\sum_{i}^{N} l_i(a_i,p_i,n_i), \quad l_i(a_i,p_i,n_i) = \max\left(d(a_i,p_i) - d(a_i,n_i) + m, 0\right) \tag{1}$$

where a_i is the projected representation of an *anchor* sample whose distance from the representation of a *positive* (similar) sample p_i is minimized, while the distance to the representation of a *negative* (dissimilar) sample n_i is maximized. Each of these representations are computed from augmented inputs to improve training. An extra margin term $m = 1$ is used to enforce a minimal distance to negative samples. We use $d(x_i, y_i) = ||xi - y_i||_2$ as distance function, and train using an AdamW optimizer [8] with default PyTorch parameters.

3.2 Triplet Sample Selection Strategy

The selection of the samples which constitute the triplets is the cornerstone of our approach. It is guided by weak labels provided at the video clip level, i.e. "original" or "attack", which exhibit different properties in the MIDV-Holo dataset. In the case of video clips labeled as "originals", we assume the visual appearance of the hologram throughout a major part of the recording. Conversely, for video clips labeled as "attacks", we assume that there will be no change in its visual appearance. This requires to remove cases of "photo replacement" attacks from our training set, as they exhibit the behavior of the real hologram except at the position of the replaced face picture. These assumptions led to the following selection process, illustrated in Figure 3:

- **Original**: Given that the document is always moving in the videos of the dataset (at 5 frames per second), we assume that the hologram is changing. Thus, frame t and frame $t + 1$ are expected to contain two different visual states of the hologram. The anchor and the positive samples are generated from the same frame t with different augmentations. Frame $t + 1$, with augmentation, is used as the negative sample.
- **Attack**: In the case of an attack, we know that all the frames from a same video contain the exact same visual state of the hologram. Therefore, we take uniformly selected frames from this video as anchor and positive sample. For the negative sample, we select a uniformly selected frame from another video with the same identity.

In both cases, the anchor, positive, and negative frames all represent the same identity (face picture). Consequently, the network aims to minimize the distance between the embeddings of two frames depicting the same visual state of a hologram while maximizing the difference between a frame showing the same face but with different hologram states. The assumption that the viewpoint continuously changes is generally valid in the MIDV-Holo dataset and can be easily enforced in a real industrial scenario. This is because a document detection stage is typically required during capture to localize, classify, and rectify documents, providing indications about the camera pose relative to the document.

3.3 Augmentations

To diversify anchor, positive, and negative samples (initially resized to 256 px), we apply transformations with specified probabilities: – Rotation, horizontal or

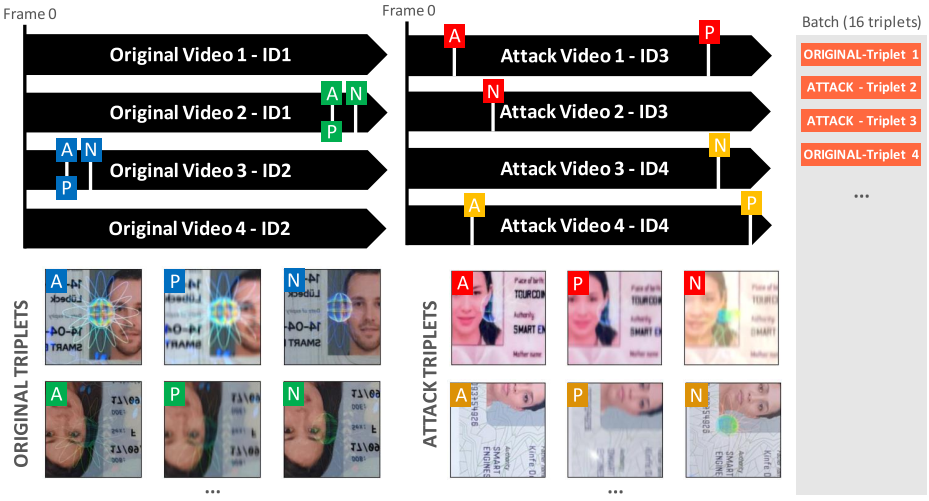

Fig. 3. Frame sampling strategy for building triplets [A]nchor/[P]ositive/[N]egative from *original* and *attack* videos. Each *original* triplet is sampled from a unique *original* video (A and P at t, N at $t+1$). Each *attack* triplet is sampled from 2 different videos of a same identity with A and P, both belonging to a common video, and N belonging to a different one. All samples are transformed with uniformly selected augmentations.

vertical mirroring ($p = 0.5$) applied equally to anchor, positive and negative samples. – Crop by a random 80% ROI, resized to 224 px ($p = 1$). – Gaussian blur ($p = 0.4$, kernel: $3 <$ kernel < 11, $2 < \sigma < 10$). – Color jittering ($p = 0.4$, $0.7 <$ brightness < 1.3, $0.9 <$ contrast < 1.1, $0.95 <$ saturation < 1.05). Images are then normalized to ImageNet's mean and standard deviation.

3.4 Qualitative Validation: Feature Attribution Maps

We employed the Integrated Gradients method by Sundararajan et al. [5], implemented in Captum [10], to identify the focal elements of our approach. The results, illustrated in Figure 4, showcase the efficacy of our weakly supervised training method. Specifically, the model ($mobilevit_{xxs}$) trained using this approach assigns significant importance to the hologram area. This stands in stark contrast to the same network architecture trained exclusively on ImageNet, which lacks a similar level of focus on the hologram. This observation underscores the value of our training strategy in guiding the model's attention towards pertinent features.

3.5 Final Decision Stage

Finally, thanks to the representation produced by the feature extraction network, learned in a weakly-supervised manner as previously presented, we can extract and compare vector representations for each frame of a video clip to

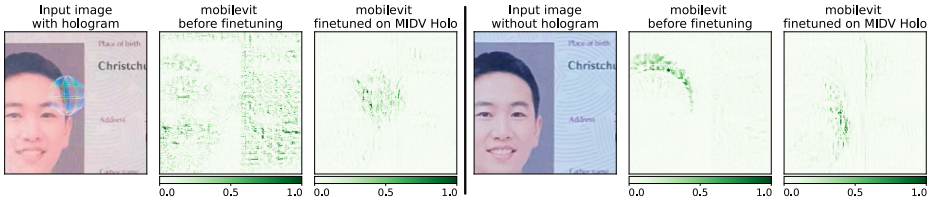

Fig. 4. Integrated Gradients [5] visualizes a training sample, emphasizing our method's effectiveness in directing the network's attention towards the hologram. In contrast, the ImageNet-trained model lacks this focused attribution to the hologram, highlighting the significance of our training approach.

control. Subsequently, we compute pairwise *cosine* similarity between the representations of the frames. We consider two scenarios: analyzing the full video (more resource-intensive but theoretically more reliable), or adopting the incremental cumulative mode introduced in the original MIDV-Holo publication. This last mode deems a video clip as original as soon as it meets the acceptable criterion. By computing the mean of these differences, it becomes possible to obtain an indicator of the expected visual behaviors' coverage. We use a single threshold calibrated on the validation set to make the final decision on accepting the video clip as original or triggering an alert for a potential attack. It is important to note that this approach does not directly inspect each individual visual appearance of the hologram. However, as the representation is trained to project non-hologram content to the same representation, such content tends to be constant while frames containing hologram content, on the other hand, typically exhibit variability. This results in a simultaneous control of *Appearance Conformity* and *Appearance Coverage*.

4 Extensions to MIDV-Holo

This section introduces extensions to the original MIDV-Holo dataset and evaluation protocol that we needed to benchmark our contribution. To compare the performance of our proposed approach with the MIDV-Holo baseline, we re-implemented and open-sourced the latter for the "no tracking" mode, i.e., without the frame alignment preliminary stage. Then, we propose an improved protocol enabling cross-validation to cope with the variance we observe in experimental results. This requires to revise the metrics used and to ways to separate training, validation, and test sets, over several runs.

4.1 Reproduction of the MIDV-Holo Baseline Approach

The authors of the original MIDV-Holo publication [18] introduced a baseline approach for semantic segmentation of video frames, identifying pixels within a holographic area and computing their ratio as a proxy to verify the hologram's shape, resulting in a binary decision: – *negative*: the video clip is deemed to

Table 1. Reproduction of Table 1 from the original MIDV-Holo [18] publication, comparing ROC AUC values in "no tracking" mode between our re-implementation and the original, validating its accuracy.

S_{thresh}	30			40			50		
h_{thresh}	0.01	0.02	0.03	0.01	0.02	0.03	0.01	0.02	0.03
Original MIDV-Holo [18]	0.795	0.825	0.832	0.828	0.841	0.832	**0.847**	0.838	0.807
Our re-implementation	0.838	0.846	0.844	0.855	0.844	0.831	**0.857**	0.826	0.790

contain the expected hologram, considered "original"; – *positive*: no hologram is found, raising an alert for a potential "attack". No public implementation of this baseline approach existed, so we created a public, open-source re-implementation following the authors' guidance. We evaluated our approach using the same conditions and metrics outlined in Table 1 before integrating it into our experiments.

The baseline approach, originally calibrated and tested on the entire MIDV-Holo dataset without clear separation between calibration and test sets, relied on three parameters: S_{thresh}, h_{thresh} and T. Our implementation differs from the original in two notable aspects: 1. resizing images to 1123×709 pixels and 2. imposing a minimum buffer of five frames before returning a result, compared to the original's theoretical requirement of two frames. We recalculated Table 1 of the original paper for the "no tracking" mode and observed nearly identical performance in terms of ROC AUC as reported in Table 1. Specifically, we identified the same optimal threshold configuration ($S_{\text{thresh}} = 50$ and $h_{\text{thresh}} = 0.01$) across values of the T parameter.

4.2 Enabling Cross-Validation on MIDV-Holo

While being an important contribution with a first public dataset with "holograms" in identity documents, MIDV-Holo still contain little data: 700 video clips which can be broken down as follows: – 2 types of documents, equally shared: identity card-like and passport-like, – 10 model variants for each type, also equally shared, – 5 "identities", i.e., fake holder for each model variant, and – 3 originals and 4 presentation attacks (actual video clips) for each "identity". These presentation attacks can either contain static content, in the case of the "copy without holo" (no hologram at all), "pseudo holo copy" (static imitation using an image editor), and "photo holo copy" (photocopy of the document) attacks; or dynamic content as in the case of the "photo replacement" attack where an original document is physically altered to change the face picture. In this latter case, no hologram is visible over the face picture, but it is still present on the rest of the document. All original document variants exhibit the *same hologram*, with a small translation between identity card and passport documents.

Because we need to be able to compare methods trained and calibrated on this dataset and reduce variance in the experiments, we propose to specify training, validation and test sets for 5 different splits. The process for generating such splits is illustrated in Figure 5, and aims at challenging the generalization to new identities rather than the generalization to new documents. Therefore, all

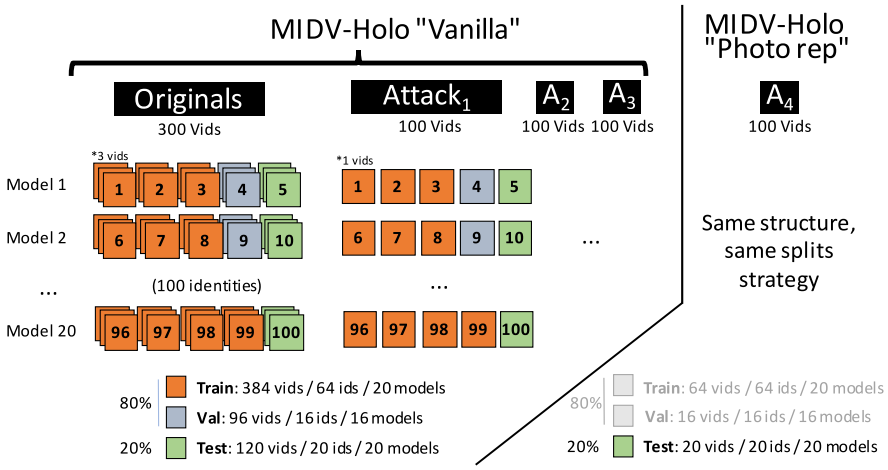

Fig. 5. Proposed split over the MIDV-Holo dataset (64% train, 16% validation and 20% test). MIDV-Holo "Vanilla" refers to the part tackled in the original paper. "Photo replacement" attacks are exclusively used for testing in our experiments.

identities in the train, validation, and test sets are distinct, while the document models (i.e. identity cards, passports) are common across the different sets. We proceed as follows: we stratify the dataset by document model (20 cases), then for each document model, we select 1 out of 5 identity for testing, the 4 remaining ones being used for training, except for 1 document every 5 items where an identity is kept for validation instead of training. All video clips for the selected identities go into the same target subset for a given split. This results in the following data partition: 64% training, 16% validation and 20% test. No identity can be present in two subsets for a given split. The *"photo replacement"* attack case is handled specially as we never use the corresponding samples for training or validation, and only use them for testing.

A last modification to the original protocol is that we use the F_{score} (harmonic mean of Precision and Recall) as the final metric, while MIDV-Holo authors preferred to report Recall values for a false positive rate close to 10%. Reporting ROC curves computed on the test set would be possible, but would only give a hint about the expected performance in production while hiding calibration uncertainty. In order to avoid an extra level of complexity during the training of a learned feature extractor (to favor Recall over Precision), we encourage the use of a simpler metric which provides a total ordering.

5 Experiments and Results

This section proposes an experimental evaluation of our proposed approach (described in Section 3) compared to the MIDV-Holo baseline [18], along with an ablation study. Contrary to the MIDV-Holo original experiment applied to the

whole rectified documents (see reproduced results in Table 2), our experiments exclusively focus on the critical region of the document containing the face picture of the bearer. We believe it is important to be able to leverage prior knowledge about the documents controlled, and challenged this idea by cropping registered document image to a particular region of interest for each document model, as illustrated in Figure 2.

5.1 Experimental Protocol

Our experiments utilized three publicly available datasets. Initially, both our proposed method and the MIDV-Holo baseline were trained and calibrated using the MIDV-Holo "Vanilla" training and validation sets, as defined in the preceding section. Subsequently, they were assessed on three distinct test sets from the MIDV series to gauge their generalization capabilities:

- **MIDV-Holo "Vanilla"** (120 test videos) originally introduced in [18], selecting only test set elements as defined in Section 4.2.
- **MIDV-Holo "Photo Replacement"** (20 test videos) is a specific subset of MIDV-Holo, and represents a distinct task from the "Vanilla" set, as it was not addressed in the original paper. While the complete set comprises 100 videos, our experiments solely involve the 20 test videos at each run, as our approach does not entail training on this dataset.
- **MIDV 2020 "Clips"** (1000 videos, test only): To assess the method's generalization to various document types, we utilized images from the "Clips" category of the MIDV 2020 [13] dataset. Following document rectification, similar Regions of Interest (ROIs) were extracted as for MIDV-Holo. As clips were sampled at 10 frames per second (fps), we dropped one frame out of two to match the frame rate of MIDV-Holo (5 fps).

It's important to note that the MIDV-Holo dataset features a single form of holographic layer, consistent across all 20 document models, with minor translations between identity-card-like and passport-like models. Consequently, our system effectively trains to detect and validate this specific holographic device.

We compared several variants of the approach. For the feature extraction stage, we tested the following models, all implemented using the timm library [9]: $resnet_{18}$ [3], $mobilevit_{xxs}$ [11] and $mobilenet_{small0.5}$ [6]. By default, all experiments utilized models initialized with weights pretrained on ImageNet. Regarding the global binary decision stage, we considered the following strategies as mentioned in Section 3.5 :

- **Whole video**: Decision made using all video frames. It is a greedy strategy which prevents any eventual bias related to frame selection or video duration.
- **Cumulative**: This strategy, utilized by MIDV-Holo [18], involves iteratively updating a cumulative metric over the sequence. If the metric surpasses a predefined threshold, the video is deemed original, potentially leading to an early stop. Otherwise, if the threshold is not met by the sequence's end, the video is classified as attack.

Table 2. Comparison of results between the MIDV-Holo baseline (reimplemented by us) and our proposed method. Metrics (F_{score} or *Recall* for attacks-only datasets) are presented across three distinct test datasets, for 2 decision strategies: Whole video and Cumulative. Both methods utilize exclusively our proposed MIDV-Holo "Vanilla" train/validation sets for training and calibration. * denotes the original MIDV-Holo configuration (applied on full rectified documents), albeit using train-validation and test splits. "MIDV-Holo ROI" and our method both focus on the same ROI.

	Test dataset →	**MIDV-HOLO "Vanilla" (120 mixed vids)**	**MIDV-HOLO "Photo repl." (20 attack vids)**	**MIDV 2020 "Clips" (1k attack vids)**
Decision	**Metric →** / **Method ↓**	F_{score} (%)	*Recall* (%)	*Recall* (%)
Whole video	MIDV-Holo ROI	80 ± 3	63 ± 10	92 ± 2
	OUR - $mobilevit_{xxs}$	90 ± 2	87 ± 14	93 ± 6
Cumulative	MIDV-Holo FULL DOC *	77 ± 1	27 ± 12	84 ± 5
	MIDV-Holo ROI	82 ± 4	66 ± 10	93 ± 0
	OUR - $mobilevit_{xxs}$	86 ± 5	84 ± 11	94 ± 4
Dummy	Perfectly random	50	50	50
	Always positive (attack)	67	100	100
	Always negative (original)	0	0	0

For our weakly supervised approach, network features are trained on the train set, the best epoch is selected based on the validation set, and the decision threshold is calibrated on the validation set. Calibration involves selecting the best F_{score} for both the *whole video* and *cumulative* decision strategies. For the MIDV-Holo baseline reproduction, parameters are calibrated on the union of training and validation sets. Each operation is repeated 5 times with different train/validation/test splits, utilizing various random seeds for generation to mitigate potential biases. Results presented in Tables 2 and 3 represent averages and standard deviations across these 5 runs.

5.2 Results and Ablation Study

Table 2 presents the outcomes achieved with the various configurations. For brevity, only the result for the best feature extraction architecture is reported here. The final rows of the table serve as a baseline, indicating the performance metrics for completely random and constant decision processes. Notably, the constant prediction of attacks reaches an F_{score} of 67% on MIDV-Holo "Vanilla", setting a lower bound for acceptable results.

These first results show the superiority of our proposed approach on MIDV-Holo test sets for both decision strategies. While achieving similar performance to the baseline on the MIDV 2020 test set, the consistently high scores may suggest a bias towards predicting attacks, contrasting with our method's robustness shown in the mixed dataset.

Finally, we conducted an ablation study to challenge the benefits of various aspects of our method. Table 3 summarizes the results, with the second row

being the reference for our non-ablated approach (Augmentations, Contrastive, Full train set). The ablated components are described here below.

Data Augmentation: *What is the impact of data augmentations?* The triplet loss, along with our sampling strategy, essentially relies on augmentations. The difference in performance between the first two rows of Table 3 confirms that this key component helps generalize across different datasets. Let's note that for "Originals" triplets, disabling augmentations nullifies the term $d(a_i, p_i)$ in Equation (1) as a_i and p_i are equal.

Training Strategy: *Is a contrastive loss competitive against direct decision optimization?* We trained a simple classifier under the same conditions (pretrained on ImageNet, same augmentations) to distinguish between original and attack frames. Then, the evaluation was done at the video level based on the average prediction for each frame, and the final outcome was calculated using a threshold calibrated on the validation set, similar to other methods. Results were surprisingly good on the MIDV-Holo "Vanilla" test set but showed a significant drop on other datasets. This underscores that while MIDV-Holo is useful as it's the first academic one of its kind, it must be handled with care. It also emphasizes the necessity of using multiple datasets to demonstrate the relevance of each method. Furthermore, it must be noted that a binary classifier cannot individually control the coverage of expected visual appearances of a hologram.

Training Set: *Are attack samples required to train our approach?* The proposed method operates on the assumption that there are equal numbers of frauds and origins. However, in a real-world scenario, it is challenging to access attack samples. Thus, it makes sense to study the impact of training only on original samples. For this specific experiment, during training, the best model was selected based on a minimum validation loss criterion (over Originals only). However, the final decision threshold calibration was computed on the extended validation set (Originals and Attacks). Training only with MIDV-Holo "Vanilla" Originals results in slightly lower performance on the test set, but still remains better than the MIDV-Holo baseline.

Model Architecture and Tuning: *How important is the model architecture and are pretrained weights sufficient?* As our approach is not tied to a particular architecture, we trained and tested several lightweight ones that can match industrial processing speed requirements. We also checked whether finetuning actually improved performance, as pretrained weights can already exhibit sensitivity to saturated colors present in holograms. Results show similar performance for the architectures tested, with $mobilevit_{xxs}$ and $mobilenet_{small0.5}$ being slightly superior when trained on mixed samples and originals only, respectively. Furthermore, the poor performance in the last row of Table 3 proves that generic features do not provide an adequate representation for our problem.

Table 3. Ablation study showing the contribution of 3 essential components of the proposed method: 1) data augmentation, 2) contrastive learning strategy, 3) training data. All configurations are tested on 3 different model architectures. * best configuration with all the features enabled (reported in Table 2).

ABLATED ELEMENTS OF THE PIPELINE				**Test dataset →**	**MIDV-Holo "Vanilla" (120 mixed vids)**	**MIDV-Holo "Photo replace" (20 fake vids)**	**MIDV 2020 "Clips" (1k fake vids)**
Data aug.	**Train strategy**	**Training Dataset**	**Decision**	**Metric →** / **Archi ↓**	F_{score} (%)	***Recall*** (%)	***Recall*** (%)
On	Contrast (triplet loss)	MIDV-Holo "Vanilla" full train set (Originals & Attacks)	Whole video	$mobilenetv3_{s50}$	88 ± 3	93 ± 8	92 ± 5
				$mobilevit_{xxs}$ *	90 ± 2	87 ± 14	93 ± 6
				$resnet18$	88 ± 2	91 ± 7	93 ± 5
Off				$mobilenetv3_{s50}$	83 ± 6	75 ± 17	86 ± 7
				$mobilevit_{xxs}$	87 ± 12	65 ± 20	87 ± 7
				$resnet18$	88 ± 6	81 ± 13	83 ± 5
On	Classifier (softmax)			$mobilenetv3_{s50}$	89 ± 3	77 ± 12	44 ± 7
				$mobilevit_{xxs}$	94 ± 3	85 ± 11	59 ± 4
				$resnet18$	92 ± 1	76 ± 10	76 ± 14
	Contrast (triplet loss)	Originals only		$mobilenetv3_{s50}$	82 ± 7	89 ± 11	94 ± 4
				$mobilevit_{xxs}$	84 ± 4	87 ± 18	89 ± 9
				$resnet18$	83 ± 2	84 ± 13	87 ± 8
None (pretrained weights)				$mobilenetv3_{s50}$	73 ± 6	81 ± 15	61 ± 19
				$mobilevit_{xxs}$	67 ± 1	92 ± 10	82 ± 7
				$resnet18$	77 ± 7	76 ± 19	59 ± 16

6 Conclusion

We have presented a novel approach for verifying both Appearance Conformity and Appearance Coverage of Optically Variable Devices (OVDs, or "holograms") in identity documents using video clips recorded from commodity smartphones. This approach leverages a feature extraction network trained with a contrastive loss, which can be specialized to a given hologram while requiring only video-level annotations, rather than individual frame labels. Furthermore, we have demonstrated that this approach can achieve attractive results using original video samples alone, which are abundantly obtained in industrial pipelines. Thanks to the separate calibration of its decision stage, our approach can be easily tuned to specific security requirements.

The evaluation of this approach necessitated the introduction of several extensions to the original MIDV-Holo dataset and the reimplementation of the proposed baseline. Our experiments have revealed the superiority of our approach over the previous baseline and its robust generalization capabilities across both the MIDV-Holo "Photo Replacement" and MIDV 2020 "Clips".

Lastly, our ablation study has uncovered a significant insight: while the MIDV-Holo "Vanilla" dataset yields intriguingly good results when tested with a simple binary classifier trained at the frame level, its generalization to other datasets is poor, as expected. This raises the question: *"What does the binary classifier actually learn?"* for future investigation.

References

[1] Vassileios Balntas et al. "Learning local feature descriptors with triplets and shallow convolutional neural networks". In: *Procedings of the British Machine Vision Conference 2016*. York, UK: British Machine Vision Association, 2016, pp. 119.1–119.11. ISBN: 978-1-901725-59-9. DOI: 10.5244/C.30.119.

[2] Andreas Daniel Hartl et al. "Efficient Verification of Holograms Using Mobile Augmented Reality". In: *IEEE Transactions on Visualization and Computer Graphics* 22.7 (July 2016), pp. 1843–1851. ISSN: 1941-0506. DOI: 10.1109/TVCG.2015.2498612.

[3] Kaiming He et al. "Deep Residual Learning for Image Recognition". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 770–778. DOI: 10.1109/CVPR.2016.90.

[4] Daniel Soukup and Reinhold Huber-Mörk. "Mobile hologram verification with deep learning". In: *IPSJ Transactions on Computer Vision and Applications* 9 (Dec. 1, 2017). DOI: 10.1186/s41074-017-0022-7.

[5] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. "Axiomatic attribution for deep networks". In: *Proceedings of the 34th International Conference on Machine Learning - Volume 70*. ICML'17. Sydney, NSW, Australia: JMLR.org, 2017, pp. 3319–3328. arXiv: 1703.01365.

[6] Andrew Howard et al. "Searching for MobileNetV3". In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. 2019, pp. 1314–1324. DOI: 10.1109/ICCV.2019.00140.

[7] Bofan Lin et al. "Face Liveness Detection by rPPG Features and Contextual Patch-Based CNN". In: *Proceedings of the 2019 3rd International Conference on Biometric Engineering and Applications*. ICBEA 2019. Stockholm, Sweden: Association for Computing Machinery, 2019, pp. 61–68. ISBN: 9781450363051. DOI: 10.1145/3345336.3345345.

[8] Ilya Loshchilov and Frank Hutter. *Decoupled Weight Decay Regularization*. Jan. 4, 2019. arXiv: 1711.05101.

[9] Ross Wightman. *PyTorch Image Models*. https://github.com/rwightman/pytorch-image-models. 2019. DOI: 10.5281/zenodo.4414861.

[10] Narine Kokhlikyan et al. *Captum: A unified and generic model interpretability library for PyTorch*. 2020. arXiv: 2009.07896 [cs.LG].

[11] Sachin Mehta and Mohammad Rastegari. "MobileViT: Light-weight, General-purpose, and Mobile-friendly Vision Transformer". In: *International Conference on Learning Representations*. 2021. arXiv: 2110.02178.

[12] Betul Ay. "Open-Set Learning-Based Hologram Verification System Using Generative Adversarial Networks". In: *IEEE Access* 10 (2022). Conference Name: IEEE Access, pp. 25114–25124. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2022.3155870.

[13] Konstantin Bulatov et al. "MIDV-2020: A Comprehensive Benchmark Dataset for Identity Document Analysis". In: *Computer Optics* 46.2 (Apr. 2022). ISSN: 01342452, 24126179. DOI: 10.18287/2412-6179-CO-1006. arXiv: 2107.00396.

[14] Council of the European Union. *PRADO - Public Register of Authentic travel and identity Documents Online, v. 12344/22*. 2022. URL: https://www.consilium.europa.eu/prado/en/prado-glossary.html (visited on 01/25/2024).

[15] Oumayma Kada et al. "Hologram Detection for Identity Document Authentication". In: *Pattern Recognition and Artificial Intelligence*. Ed. by Mounîm El Yacoubi et al. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2022, pp. 346–357. ISBN: 978-3-031-09037-0. DOI: 10.1007/978-3-031-09037-0_29.

[16] Yuval Nirkin et al. "DeepFake Detection Based on Discrepancies Between Faces and Their Context". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44.10 (2022), pp. 6111–6121. DOI: 10.1109/TPAMI.2021.3093446.

[17] Marie-Neige Chapel, Musab Al-Ghadi, and Jean-Christophe Burie. "Authentication of Holograms with Mixed Patterns by Direct LBP Comparison". In: *IEEE 25th International Workshop on Multimedia Signal Processing (MMSP) 2023*. ISSN: 2473-3628. Sept. 2023, pp. 1–6. DOI: 10.1109/MMSP59012.2023.10337669.

[18] L. I. Koliaskina et al. "MIDV-Holo: A Dataset for ID Document Hologram Detection in a Video Stream". In: *Document Analysis and Recognition - ICDAR 2023*. Ed. by Gernot A. Fink et al. Cham: Springer Nature Switzerland, 2023, pp. 486–503. ISBN: 978-3-031-41682-8. DOI: 10.1007/978-3-031-41682-8_30.

Acknowledgements

The SOTERIA project, partially supporting this work, was funded by the European Union's Horizon 2020 research and innovation program under grant agreement No 101018342.