DriveFace: A Cross-Spectral Through-Glass Face Dataset for On-the-Move Vehicular Border Control

Anjith George[1] Luis Luevano[1] Alain Komaty[1] Zeina Al Amine[1] Vidit Vidit[1] Sébastien Marcel[1,2]
[1]Idiap Research Institute, Rue Marconi 19, 1920 Martigny
[2]University of Lausanne (UNIL), Lausanne, Switzerland
{anjith.george, luis.luevano, alain.komaty,
zeina.alamine, vidit.vidit, sebastien.marcel}@idiap.ch

Abstract

The continuous growth in cross-border mobility places increasing pressure on existing border control infrastructures, motivating on-the-move *biometric authentication, in which travellers are identified directly inside their vehicles at checkpoints. Face recognition is well-suited to this setting, as it can be acquired passively and at a distance. Its development, however, is hindered by the lack of representative datasets: existing benchmarks are collected in controlled environments and do not capture the challenges inherent to vehicular acquisition, including motion blur, variable illumination, occlusions, and cross-spectral enrollment. To address this gap, we introduce a dataset for on-the-move face recognition in border-control scenarios, comprising NIR vehicle-crossing videos paired with smartphone-based pre-enrollment data. Baseline evaluations with state-of-the-art models show clear performance limitations under these realistic conditions, highlighting the need for dedicated methods to advance the field.*

1. Introduction

Face recognition [20] is a highly suitable biometric modality for automated border control (ABC) applications, owing to its non-contact acquisition capability and seamless integration with established workflows [25]. Nevertheless, achieving robust performance at land border crossings remains a significant challenge. Under operational conditions, subjects are frequently captured with non-frontal head pose, motion blur, variable ambient illumination, and partial facial occlusion. These difficulties are further exacerbated when image acquisition is performed through automotive windows, which introduces additional optical degradations that materially degrade recognition accuracy.

A particularly underexplored yet operationally significant scenario is *in-vehicle border screening*, in which a traveler is observed while seated inside a vehicle through a windshield or side window. In addition to the typical challenges inherent to unconstrained face recognition, automotive windows introduce complex optical phenomena. Modern windshields frequently incorporate infrared-reflective coatings engineered to attenuate solar heat gain; such coatings can reflect in excess of 50% of near-infrared (NIR) radiation, substantially reducing the signal available to NIR-based imaging systems. Variability in window tint across vehicle makes and models further increases variability [24], while specular reflections arising from ambient lighting introduce additional noise and artifacts.

VIS-NIR face recognition further suffers from the cross-

(a) iPhone reference (b) Sensor-vehicle outdoor setup (90°)

(c) Outdoor (45°) (d) Indoor car (e) Simulated tinted window

Figure 1. Overview of the acquisition setup in DriveFace , including a reference image, the probe VIDAR PAX [2] sensor-vehicle setup, and samples from the three main probe capture settings.

spectral domain gap [3]: enrollment references are typically acquired in the visible (VIS) spectrum (*e.g.*, passport or mobile photographs), whereas checkpoint probes are captured in NIR, yielding substantial differences in facial texture and reflectance that challenge models trained on VIS data alone. Although benchmarks such as CASIA NIR-VIS 2.0 [27] have advanced heterogeneous face recognition, they are collected in controlled settings and do not reproduce the combined effects of through-glass imaging, subject motion, and unconstrained outdoor environments. Datasets acquired in operational border settings remain largely proprietary [7].

To address these limitations, we introduce DriveFace , a NIR dataset that captures the end-to-end biometric workflow at vehicular border crossings and enables systematic study of cross-spectral matching, through-glass degradation, and adverse illumination within a single operationally representative benchmark collected over two months. A dedicated presentation attack subset further supports anti-spoofing research under realistic deployment conditions. Figure 1 illustrates the capture settings.

The main contributions of this work are listed below:

- DriveFace , a publicly available dataset for face recognition at vehicular border crossings, pairing VIS pre-enrollment images with in-vehicle NIR probes acquired through automotive windows, and a presentation attack subset (DriveFace-PAD).
- Rich per-acquisition metadata: tint level, illumination, head pose, and vehicle speed, enabling fine-grained analysis of individual degradation sources.
- Standardized face recognition and PAD protocols with baseline results establishing reference performance on DriveFace . The dataset is available at: [1] [2].

2. Related Work

While there are a large number of face recognition datasets publicly available [4, 18, 8, 9, 49], most are collected in controlled environments and do not reflect the challenges of vehicular border control. Below, we review datasets relevant to this setting.

In-Vehicle Face Datasets In-vehicle face datasets are typically introduced for driver monitoring rather than identity recognition. Representative examples include Drive&Act [31], DMD [33], and DriverMVT [34], which capture in-cabin subjects under varying pose, occlusion, illumination, and seating position for tasks such as distraction, drowsiness, gaze, and head-pose estimation. Datasets more oriented toward in-vehicle authentication include AVICAR [26], VFPAD [22], and iCarB [23]. While valuable for automotive facial analysis, none target heterogeneous cross-session identity matching from *outside* the vehicle through tinted or clear windows with profile-pose variation, leaving a clear gap for external-view, cross-modal recognition.

NIR Face Datasets Near-infrared heterogeneous datasets support face recognition beyond the visible spectrum, particularly under low light and for RGB-to-NIR matching. Representative datasets include CASIA NIR-VIS 2.0 [27], PolyU-NIRFD [44], Oulu-CASIA NIR&VIS [48], and BUAA-VisNir [19], alongside broader multimodal datasets such as the Tufts Face Database [35] and LAMP-HQ [43]. Although these have advanced heterogeneous face recognition, they are collected indoors under controlled conditions and do not capture through-glass acquisition, in-vehicle deployment, motion, or outdoor settings.

Face Acquisition Through Glass Through-glass face acquisition remains relatively underexplored. The most related works are the through-windshield driver recognition study of Cornett *et al*. [7], the Face Image Reflection Removal (FIRR) dataset [39], and PP4AV [38], which cover subjects viewed through windshields, real outdoor driving, and reflection-degraded imagery, respectively. However, [39] and [38] target different tasks and do not model heterogeneous mobile-RGB to external-NIR matching, while [7] captures multiple occupants through tinted windows. Public benchmarks for through-glass face recognition therefore remain limited.

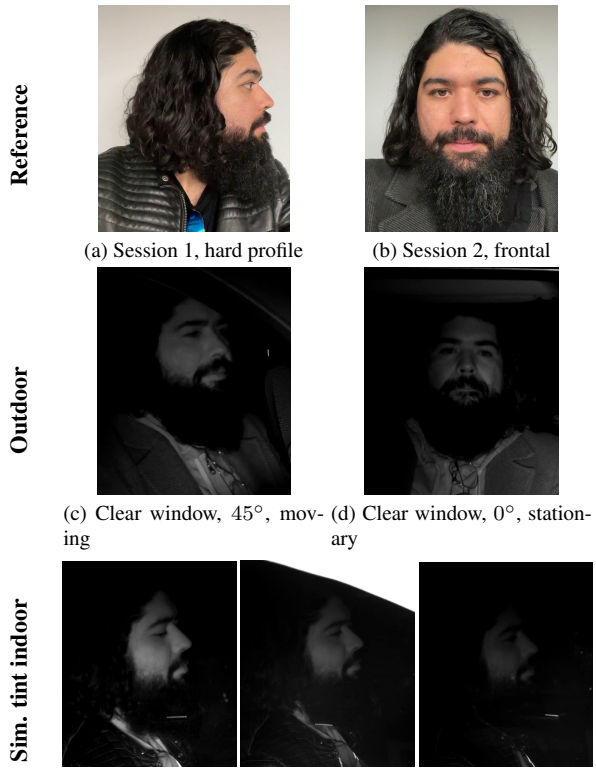
Presentation Attack Detection Datasets PAD datasets assess robustness to print, replay, and mask attacks. Established benchmarks such as Replay-Attack [6], CASIA-FASD [47], MSU-MFSD [40], OULU-NPU [5], and SiW [30] rely on RGB capture in mobile or desktop settings, while more recent datasets such as CASIA-SURF [45], CASIA-SURF CeFA [28], CelebA-Spoof [46], WMCA [16], and HiFiMask [29] extend toward multimodal sensing, cross-ethnicity evaluation, and 3D masks. However, none of them jointly address PAD in NIR, through-glass acquisition, and external-view capture under tint, reflections, and motion. The in-vehicle datasets VFPAD [22] and iCarB [23] are closer: VFPAD targets in-vehicle NIR PAD and iCarB supports in-car PAD-oriented evaluation, but both focus on in-cabin capture rather than external probes acquired through automotive glass.

3. The DriveFace Dataset

The proposed dataset was designed to support biometric research in operational border-control settings. It consists of three components: *(i) pre-enrollment face captures*, *(ii) in-vehicle face captures*, and *(iii) presentation attacks*. Together, these components reflect a realistic work-

[1] https://www.idiap.ch/paper/driveface

[2] https://www.idiap.ch/en/scientific-research/data/driveface



(a) Session 1, hard profile (b) Session 2, frontal

(c) Clear window, 45°, moving (d) Clear window, 0°, stationary

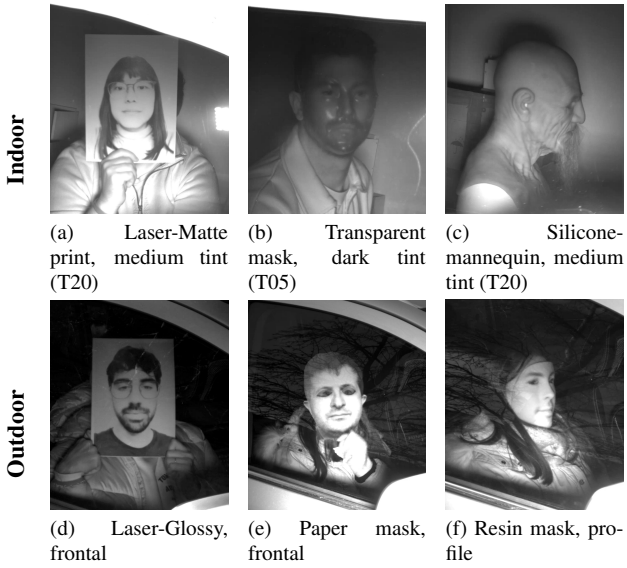
(e) Clear window (f) Med. tint (T35) (g) Dark tint (T05)

Figure 2. Samples from a single subject in DriveFace . The figure shows visible-spectrum reference captures from two sessions, infrared probe samples acquired through simulated indoor window conditions with different tint levels and real-world outdoor infrared acquired through clear vehicle windows at different viewing angles and moving conditions.

Table 1. Summary of the main components of DriveFace .

Subset	Mod.	# Var.	Face pose
Reference	RGB	4	frontal to profile
Outdoor probes	NIR	24	frontal, 3/4, profile
Sim. tint indoor	NIR	36	frontal, profile
Indoor car	NIR	6	frontal, profile

flow in which a traveler first provides reference biometric data prior to arrival and is subsequently observed under challenging checkpoint conditions. The dataset was collected from **70 consenting subjects** over **two recording sessions**, with corresponding reference samples and probes. Figure 2 shows examples of the pre-enrollment face captures for reference and probe images. For most participants, the two sessions were collected approximately two months apart in order to capture realistic short-term variation in appearance, including changes in facial hair, hairstyle, grooming, makeup, and other mild changes in facial appearance.



(a) Laser-Matte print, medium tint (T20) (b) Transparent mask, dark tint (T05) (c) Silicone-mannequin, medium tint (T20)

(d) Laser-Glossy, frontal (e) Paper mask, frontal (f) Resin mask, profile

Figure 3. Example presentation attack samples from DriveFace-PAD . The figure shows variation in attack types under indoor and outdoor conditions, with examples covering different presentation poses and levels of glass tint.

Table 2. Summary of the presentation attack subset in DriveFace-PAD .

Attack	Identity	Instruments / materials
Print	Bona fide IDs	Printer: Inkjet / Laser Matte / Normal / Glossy
Replay	Bona fide IDs	Laptop display
Mask	External IDs	Paper, resin, silicone, plastic, plastic + makeup, thin colored plastic, rubber, silicone, mannequin

3.1. Demographics

The dataset was collected from *70 consenting volunteers*. The same participant pool was used across the data collection campaign, which allows direct association between pre-enrollment reference samples and in-vehicle operational captures. This design supports both verification and identification experiments in which clean enrollment data are matched against challenging probes. This dataset effort contains a demographically diverse group of participants. In terms of gender, the dataset is relatively balanced, with 45.7% female and 54.3% male subjects. The age distribution spans a broad range from 18 to 85 years old. The dataset also includes variation in apparent skin tone, which we summarize into three broad groups: light, medium, and dark, containing 51, 11, and 6 subjects, respectively, based on the available annotations. Figure 4 illustrates the age and skin-tone distributions of the subjects in the dataset.

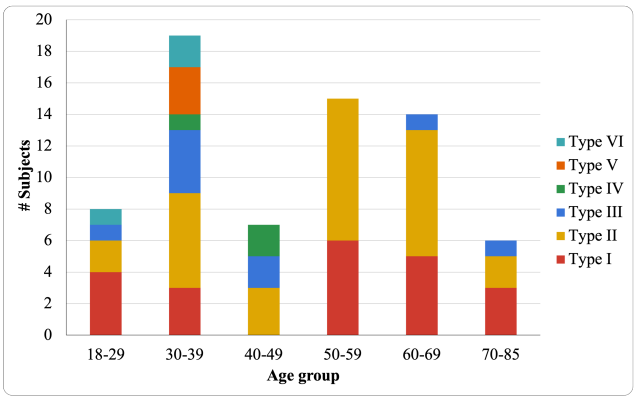


Figure 4. Age group distribution labeled by Fitzpatrick skin color tones.

3.2. Pre-enrollment Face Captures

The pre-enrollment subset was designed to simulate a trusted-traveler or pre-registration procedure performed before arrival at a checkpoint. For each subject and each session, close-range facial videos were acquired using the front-facing cameras of two consumer smartphones: an *iPhone 12* and a *Samsung Galaxy S9*. This subset provides high-quality reference data intended to serve as identity ground truth for subsequent comparisons with more challenging operational captures. During acquisition, subjects were instructed to perform controlled head movements in order to introduce natural pose variation rather than a single static frontal view. This subset is therefore suitable as an enrollment/reference set for verification and identification experiments under cross-device and cross-condition settings.

3.3. In-vehicle Infrared Face Captures

The in-vehicle subset targets face acquisition and recognition for subjects seated inside vehicles and observed through automotive glass. This scenario represents a particularly challenging condition for border biometrics, due to the combined effects of reflections, window tint, low transmission, viewpoint changes, motion blur, and reduced illumination. The in-vehicle data were acquired primarily in the *near-infrared (NIR)* spectrum under two complementary protocols using a VIDAR PAX infrared sensor [2] with an external illuminator.

Real-vehicle outdoor captures: In the first protocol, subjects were recorded inside a real vehicle from an acquisition distance of approximately *1–2 m* from the glass. Recordings were captured from multiple viewpoints, including approximately $0°$, $45°$, and $90°$, and under both *stationary* and *moving-vehicle* conditions. This protocol was designed to reproduce realistic checkpoint approach conditions, including motion-related degradation and perspective distortion.

Controlled indoor simulated-car captures: In the second protocol, a controlled indoor setup was used to isolate the effects of glass transmission and scene illumination. Recordings were acquired under *five window tint levels* corresponding to *5, 15, 20, 30, and 45 Visible Light Transmission (VLT) levels*, and under *three illumination levels* of *5%*, *20%*, and *30%*. Instead of using real cars, we mounted automotive glass panels with different VLT levels on wooden frames simulating these variations. This protocol provides a structured benchmark for studying the limits of face recognition systems under through-glass conditions.

Real-vehicle indoor captures: For the third protocol, a smaller set of real-vehicle indoor captures was also collected to study face acquisition in a more controlled environment while preserving the in-car setting. In this protocol, subjects were recorded inside a real vehicle at approximately $90°$ relative to the vehicle direction through the *clear window*. Two facial pose conditions were considered, namely *CAM* and *FWD*, and recordings were acquired at *5%*, *20%*, and *30%* illuminator power. This part of the dataset contains *418 NIR video clips*.

Overall, the in-vehicle subset provides a challenging and structured benchmark for evaluating robustness to through-glass degradation, viewpoint variation, and adverse lighting in border-control scenarios.

3.4. Presentation Attacks

The dataset also includes a presentation attack subset designed to evaluate the robustness of face anti-spoofing systems in the proposed vehicular border-control setting. Attacks are presented in outdoor and indoor conditions from inside the vehicle and captured from the outside using the infrared sensor positioned at approximately $45°$ relative to the vehicle direction.

The PAD subset includes three attack categories: *print*, *replay*, and *mask*. Print and replay attacks are generated from subjects for whom bona fide samples are available in other parts of the dataset, while mask attacks are collected from external subjects. All attack types are recorded under the same factors: *four infrared illumination levels* (*5%, 10%, 20%, 30%*) and *two attack poses*, namely *CAM* and *FWD*. The subset also includes variation within each attack category. For *print attacks*, the dataset contains different combinations of *printer type* and *paper material*, including *inkjet* and *laser* printers, and *matte*, *normal*, and *glossy* paper. For replay attacks, it contains two *laptop display* types. For *mask attacks*, the dataset includes *paper*, *resin*, *silicone*, *transparent plastic*, *transparent plastic with makeup*, *thin colored plastic*, *rubber*, *silicone over mannequin*, and *mannequin*. Figure 3 shows example attacks, while Table 2 summarizes the main factors covered by the subset.

3.5. Annotations and Metadata

Automated detection and tracking were used to isolate consenting participants and remove bystanders, followed

by a final *human-in-the-loop* check to ensure that no non-consenting identities remained visible. The dataset provides annotations and metadata at the *subject*, *session*, and *video* levels, with most acquisition settings encoded in the filename. *Identity labels* include the subject ID and, for in-vehicle captures, the identities and seat positions of all occupants. *Session labels* are defined for both reference and probe data: reference filenames encode subject ID, date, and session number, while the parent folder indicates the mobile device; in the probe subset, session 1 denotes outdoor captures and session 2 indoor captures. Additional *subject-level* metadata include age, gender, and skin tone. *Viewpoint labels* specify camera direction relative to the vehicle ($0°$, $45°$, $90°$) and face pose (*CAM*, *FWD*). *Tint and illumination metadata* include illuminator setting, scenario, speed, weather, and window tint level. For the PAD subset, *attack labels* also encode the attack category, instrument, and material.

4. Experiments

This section describes the benchmarking performed on the DriveFace dataset, including both FR and PAD experiments.

4.1. Face Recognition Experiments

Protocols In the operational setting, RGB videos are used for enrollment. These enrollment videos capture subjects performing controlled head movements (left–right and up–down). In the probe scenario, subjects are recorded using an NIR camera while seated inside vehicles under different operational conditions. For enrollment, a frontal face image is selected for each subject using a pose estimator. For the probe set, we use face crops produced by a face detector and manually annotated with identity labels corresponding to the operational scenarios, selecting up to 10 samples per subject from each video.

For training and evaluation, the dataset is partitioned into disjoint train and test sets based on identity. Specifically, 60% of identities are used for training and the remaining 40% for testing. Each subject has two capture sessions, where samples from different sessions are used for enrollment and probe sets to ensure cross-session evaluation.

The dataset includes three operational conditions: (1) outdoor in-vehicle scenarios, (2) indoor in-vehicle scenarios, and (3) indoor simulation scenarios with varying automotive glass tint levels. These correspond to the protocols *outdoor*, *indoor_car*, and *simulation*. Additional metadata, such as vehicle speed, tint level, weather, and illumination conditions are provided to support more detailed analysis. The dataset statistics are shown in Table 3.

Metrics To evaluate face recognition performance, we report several standard metrics commonly used in the literature, including Area Under the Curve (AUC), Equal Error Rate (EER), Rank-1 identification rate, and Verification Rates (VR a.k.a 1-FNMR) at specific False Acceptance Rate (FARs a.k.a, FMRs) of 1%.

Table 3. Face recognition protocol statistics for the DriveFace dataset.

	Train			Test		
Protocol	**IDs**	**Videos**	**Frames**	**IDs**	**Videos**	**Frames**
outdoor	41	1178	11699	28	909	9057
indoor_car	42	245	2050	28	166	1410
simulation	42	1450	14100	28	928	8980

Baselines To benchmark performance on the dataset, we consider several publicly available open-source face recognition models. Although there exists a modality gap between enrollment (RGB) and probe (NIR), this gap is relatively small compared to other heterogeneous settings [15]. We evaluate a set of baseline models that are originally trained for standard face recognition tasks, and additionally include models that are explicitly trained on the training split of DriveFace . AdaFace [21] introduced an adaptive margin-based loss for face recognition that uses the feature norm as a proxy for image quality. In our experiments, we used the publicly available AdaFace model with a ResNet-100 backbone trained on the WebFace12M dataset. LVFace [42] introduced a Vision Transformer (ViT) based FR model designed to better exploit large vision models for face recognition. In our experiments, we used the publicly available LVFace model based on a ViT-L backbone trained on the Glint360K dataset. EdgeFace [10] is a lightweight face recognition model built on a hybrid CNN–Transformer backbone. It is trained on RGB face images from the WebFace dataset [49]. In our experiments, we used the publicly available EdgeFace-Base variant. xEdgeFace [14] is a contrastive self-distillation framework designed to adapt a pretrained face recognition backbone to heterogeneous face recognition. The method retains the original EdgeFace [10] architecture and introduces cross-modal alignment along with teacher–student supervision to enable effective adaptation while preserving performance on the source (RGB) task. As a result, the model gains robust heterogeneous recognition capability without catastrophic forgetting, while remaining lightweight.

Experiment details For AdaFace, EdgeFace, and LVFace, we use publicly available pretrained models and perform baseline evaluations by comparing each probe sample against all gallery templates using cosine distance. All images are first preprocessed using face detection and alignment, and the resulting 112×112 crops are fed into the pretrained networks for feature extraction and evaluation. For xEdgeFace, we adopt the EdgeFace backbone and further tune it using the training split of DriveFace following

Table 4. Model performance comparison across different protocols

Protocol	Model	AUC	EER	VR@FAR=1%	Rank-1
Outdoor	AdaFace	99.45	2.69	**96.01**	97.42
	LVFace	98.50	5.26	90.17	92.26
	EdgeFace	98.38	5.37	91.35	93.29
	xEdgeFace	98.94	4.18	93.12	94.22
Simulation	AdaFace	96.23	8.26	88.27	90.46
	LVFace	95.19	11.42	79.68	83.95
	EdgeFace	94.09	11.73	82.93	84.72
	xEdgeFace	96.03	8.09	**88.63**	89.68
Indoor Car	AdaFace	98.59	3.11	**96.24**	97.11
	LVFace	98.52	5.29	90.23	93.27
	EdgeFace	98.75	4.48	93.92	93.34
	xEdgeFace	97.98	3.47	95.95	96.16

the method proposed in [14], specifically optimizing it for the cross-spectral setting. The performance reported corresponds to evaluation on the test set.

Experimental results Tab. 4 summarizes the face recognition performance across the three evaluation protocols.

Outdoor protocol. The outdoor protocol achieves the strongest overall performance across all models, indicating that despite the RGB-NIR modality gap, robust FR models perform reasonably well. AdaFace achieves the best performance with an AUC of 99.45% and the lowest EER of 2.69%, along with the highest Rank-1 accuracy (97.42%). This suggests that its adaptive margin mechanism generalizes well even under mild modality shifts. xEdgeFace improves over its backbone (EdgeFace), reducing the EER from 5.37% to 4.18% and increasing verification rates, demonstrating the effectiveness of cross-spectral adaptation.

Simulation protocol. The simulation protocol is the most challenging scenario, with all models exhibiting a noticeable drop in performance. This can be attributed to the presence of controlled variations such as different glass tint levels and potentially more severe spectral distortions. AdaFace and xEdgeFace achieve comparable performance, with AUC values around 96% and EERs near 8%, significantly higher than in other protocols. Importantly, xEdgeFace slightly outperforms AdaFace in VR@FAR=1%, highlighting the advantage of explicit cross-spectral adaptation under challenging conditions. EdgeFace and LVFace show the largest degradation, indicating that models trained purely on RGB data struggle to generalize in this scenario.

Indoor car protocol. In the indoor in-vehicle setting, performance remains high but shows a slight degradation compared to outdoor conditions, likely due to more constrained lighting and reflections from vehicle interiors. AdaFace achieves the best overall results in this protocol, with the lowest EER (3.11%) and highest Rank-1 accuracy (97.11%), indicating strong robustness to controlled in-car conditions. Notably, xEdgeFace again improves over EdgeFace, confirming that domain adaptation consistently benefits cross-spectral matching. However, the gains are less pronounced than in the simulation scenario, suggesting that the domain gap is smaller in this setting.

Across all protocols, the results highlight two key aspects: (i) the impact of cross-spectral and environmental variations on recognition performance, and (ii) the benefit of adapting models to the target domain. Pretrained RGB models (AdaFace, EdgeFace, and LVFace) demonstrate strong baseline performance, indicating that the RGB–NIR gap in this dataset is small, although it still introduces noticeable degradation under more challenging conditions. Among these, AdaFace consistently achieves the most robust performance, likely due to its quality-aware training strategy. At the same time, the consistent improvements of xEdgeFace over its backbone EdgeFace across all protocols emphasize the importance of domain adaptation for heterogeneous face recognition, with the gains being particularly evident in the more challenging simulation scenario. It should be noted that, EdgeFace and xEdgeFace are significantly more lightweight compared to AdaFace and LVFace, yet they remain competitive, highlighting an important trade-off between efficiency and performance. Overall, while strong pretrained models generalize reasonably well, explicit cross-spectral adaptation is crucial for achieving robust performance under varying operational conditions, especially when illumination changes and sensor differences introduce larger domain shifts.

Figure 5 shows representative failure cases in the simulation protocol. From the images, it is evident that most failures occur under conditions such as extreme tint, low contrast, extreme profile views, and the presence of occlusions.

Effect of Tint Level We performed a set of experiments on the controlled subset, reporting AUC separately for each VLT level and for clear glass (Tab. 5). Across the evaluated models, performance generally improves as light transmission increases, with clear glass giving the best results. This directly quantifies the effect of tint.

Table 5. AUC values for different FR models across tint levels.

VLT	AdaFace	LVFace	EdgeFace	xEdgeFace
T05	98.36	98.72	98.42	98.20
T15	98.48	98.09	98.12	98.60
T20	99.31	99.04	98.23	98.72
T35	99.62	99.14	98.75	99.35
clear	100.00	99.86	99.87	100.00

4.2. Presentation Attack Detection

As noted in the study in [13], replay attacks do not constitute a significant challenge in the near-infrared (NIR) domain, as they are largely invisible in NIR. Therefore, this work focuses on more challenging and operationally rele-

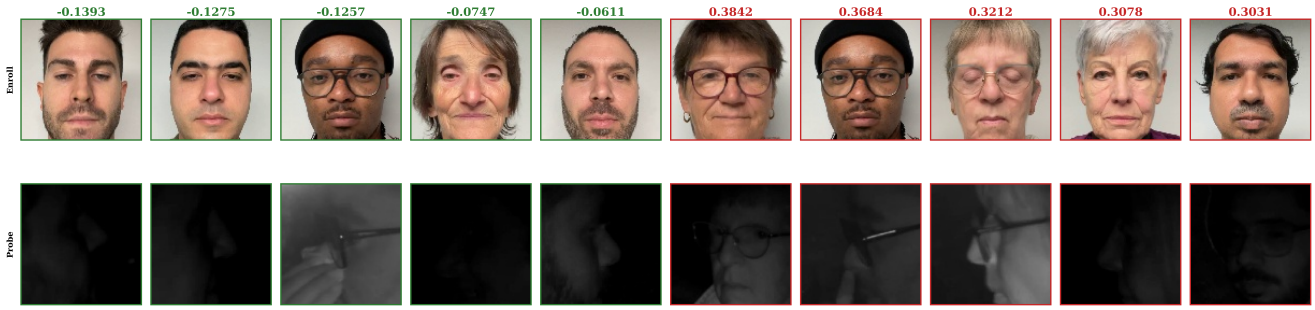


Figure 5. Worst-case verification pairs. **Left (green):** genuine pairs with the lowest match scores, indicating failure to recognize true matches. **Right (red):** impostor pairs with the highest match scores, indicating false acceptances. Top row shows enrollment images; bottom row shows corresponding probe images.

vant attack types, namely print and mask-based presentation attacks [12], which exhibit greater attack potential.

Protocols We define three evaluation protocols to assess model robustness under different attack conditions: *grandtest*, where all attack types appear across training, development, and evaluation splits, representing a known-attack setting with disjoint subjects and attack instruments; *unseen_print*, where print attacks are excluded from training and development and appear only at evaluation, simulating an unseen print-attack scenario; and *unseen_mask*, which follows the same setup for mask attacks to measure generalization to unseen mask-based attacks. A detailed summary of the dataset statistics for each protocol is provided in Table 6.

Table 6. Summary of PAD protocol splits.

Protocol	Split	IDs	Frames	Bonafide	Attack	Print	Mask
grandtest	train	42	13 375	9 856	3 519	1 495	2 024
	dev	14	4 845	3 488	1 357	561	796
	eval	14	5 501	4 098	1 403	664	739
unseen_print	train	36	11 843	9 263	2 580	0	2 580
	dev	12	3 944	2 965	979	0	979
	eval	22	7 934	5 214	2 720	2 720	0
unseen_mask	train	44	13 786	11 546	2 240	2 240	0
	dev	14	3 781	3 301	480	480	0
	eval	12	6 154	2 595	3 559	0	3 559

Metrics To evaluate the performance of presentation attack detection (PAD) systems, we adopt the standardized metrics defined in ISO/IEC 30107-3 [1]. Specifically, we report the Attack Presentation Classification Error Rate (APCER) and the Bona Fide Presentation Classification Error Rate (BPCER), where BPCER corresponds to the rate at which bona fide samples are incorrectly classified as attacks. Additionally, we compute the Average Classification Error Rate (ACER), defined as the mean of APCER and BPCER. The decision threshold is determined on the development set of each protocol using the Equal Error Rate (EER) criterion.

Baselines We evaluate several baseline models across the protocols. DeepPixBiS [11] employs a DenseNet-based architecture with pixel-wise binary supervision to enhance PAD performance. In our implementation, we utilize only the pixel-wise binary loss during training, and final predictions are obtained by averaging the resulting pixel-wise score maps. For CLIP [36], we utilize the vision encoder of the CLIP model, specifically the ViT-B/32 variant. Two training configurations are considered: (i) *CLIP (fc only)*, where only the final fully connected layer is trained while keeping the backbone frozen, and (ii) *CLIP (full)*, where all model parameters are fine-tuned. The DinoV2 [32] baseline utilizes the DinoV2 backbone, specifically the ViT-B/14 variant. Only the final fully connected classification layer is fine-tuned, while the backbone remains frozen. ConvNeXtV2-Tiny and EfficientNet-B0 are lightweight convolutional architectures based on ConvNeXt V2 [41] and EfficientNet [37], respectively. Both models are pretrained on ImageNet and subsequently adapted for the binary PAD classification task.

Experimental Settings All images were first preprocessed with face detection using the SCRFD [17] model and alignment, after which they were cropped to a fixed spatial resolution of 224×224 pixels. All models were trained for 100 epochs using a learning rate of 1×10^{-4} and a weight decay of 1×10^{-6}, with a batch size of 64. The training process was implemented in `PyTorch` and used NVIDIA RTX 3090 GPUs. To improve generalization, data augmentation was applied during training, including random horizontal flipping, RandAugment, random rotations up to 30°, and color jitter with brightness, contrast, and saturation variations of 0.15. To mitigate class imbalance, the training data was sampled in a balanced manner.

Experimental Results on PAD Table 7 summarizes the PAD performance across all three protocols. In the grandtest protocol all models achieve low error rates, with EfficientNet-B0 performing best (ACER = 0.50%), followed by ConvNeXtV2-Tiny (0.70%). This indicates that

Table 7. Summary of PAD results.

Protocol	Model	EER		
		APCER	BPCER	ACER
grandtest	DeepPixBiS [11]	0.00	3.30	1.60
	CLIP (fc only)	4.40	2.60	3.50
	DinoV2 (fc only)	3.50	1.00	2.20
	EfficientNet-B0	0.10	0.90	**0.50**
	CLIP (full)	1.60	6.50	4.00
	ConvNeXtV2-Tiny	0.30	1.00	<u>0.70</u>
unseen_mask	DeepPixBiS [11]	91.60	0.00	45.80
	CLIP (fc only)	78.80	0.00	39.40
	DinoV2 (fc only)	52.30	0.10	**26.20**
	EfficientNet-B0	54.50	0.00	<u>27.20</u>
	CLIP (full)	83.90	1.60	42.70
	ConvNeXtV2-Tiny	73.10	0.00	36.60
unseen_print	DeepPixBiS [11]	36.70	0.00	**18.40**
	CLIP (fc only)	67.80	0.10	34.00
	DinoV2 (fc only)	53.00	0.20	26.60
	EfficientNet-B0	70.70	0.00	35.40
	CLIP (full)	64.70	3.40	34.00
	ConvNeXtV2-Tiny	42.80	0.10	<u>21.50</u>

the evaluated models perform well in the known-attack setting. CLIP-based models show comparatively higher errors, particularly in BPCER.

In contrast, performance drops sharply in the unseen attack scenarios. For unseen_mask, all models exhibit very high APCER, with DinoV2 (fc only) performing best (ACER = 26.20%), suggesting better generalization from self-supervised (SSL) features. However, overall results remain poor, highlighting the difficulty of mask attack detection under domain shift. For unseen_print, the degradation is less severe. DeepPixBiS achieves the best result (ACER = 18.40%), followed by ConvNeXtV2-Tiny (21.50%).

These results motivate further research on domain adaptation and self-supervised representations for generalization to unseen attacks.

5. Discussion

The experiments on DriveFace demonstrate that both FR and PAD remain challenging in realistic vehicular border-control conditions, even when strong pretrained models are used. On the FR side, the results show that the RGB-NIR modality gap is manageable but still non-negligible, especially under more difficult conditions such as the simulation protocol, where glass tint and controlled illumination changes introduce a larger domain shift. AdaFace consistently provides the strongest overall baseline, while xEdgeFace shows clear gains over EdgeFace across all protocols, confirming the value of explicit cross-spectral adaptation. Importantly, EdgeFace and xEdgeFace are significantly smaller and more lightweight than AdaFace and LVFace, yet they remain competitive, making them particularly attractive for deployment in practical resource-constrained systems. On the PAD evaluations, the grandtest protocol results indicate that known attacks can be detected reliably, but performance degrades substantially in unseen attack settings, especially for mask attacks, revealing a clear generalization gap. These findings suggest that DriveFace is a challenging and realistic benchmark that exposes limitations not only in cross-spectral face recognition, but also in robust spoof detection under operational conditions, highlighting the need for models that jointly address efficiency, domain shift, and generalization to unseen scenarios.

6. Conclusion

In this work, we introduced the new DriveFace and DriveFace-PAD datasets for cross-spectral through-glass face analysis in on-the-move vehicular border-control scenarios, together with benchmark protocols for both face recognition and presentation attack detection. The datasets capture realistic operational challenges, including RGB-to-NIR matching, through-glass acquisition, unconstrained viewpoints, varying illumination, and presentation attacks. Experimental results show that while modern pretrained face recognition models already provide strong baselines, their performance still drops under more challenging environmental and spectral conditions, and explicit cross-spectral adaptation further improves robustness. Similarly, PAD results reveal that detecting known attacks is relatively easy, whereas generalization to unseen attacks remains difficult. Overall, DriveFace and DriveFace-PAD provide a valuable benchmark for advancing research on robust and efficient biometric systems for real-world border-control applications. To support reproducibility and further research, the dataset, evaluation protocols, and code are publicly available.

7. Ethics Statement

All participants in this data collection volunteered and provided informed consent, agreeing to the collection and use of their data for the specified research purposes. The project under which the data was collected was approved by the institution's Data Research Ethics Committee (DREC).

8. Acknowledgments

The project leading to this work has received funding from Frontex under the Frontex Research Grants Programme. Call for Proposals 2024/CFP/INNOVATE/01 Grant Agreement No. 2025/280. This work reflects only the authors' view. Neither the European Union nor Frontex are responsible for any use that may be made of the information it contains. This research was also partly funded by the European Union project CarMen (Grant Agreement No. 101168325).

References

[1] ISO/IEC 30107-3:2023 Information technology — Biometric presentation attack detection — Part 3: Testing and reporting, 2023.

[2] Adaptive Recognition Inc. *VIDAR User Manual*. Adaptive Recognition Inc., Nov. 2025. Accessed: 2026-04-17.

[3] D. Anghelone, C. Chen, A. Ross, and A. Dantcheva. Beyond the visible: A survey on cross-spectral face recognition. *Neurocomputing*, 611:128626, 2025.

[4] J. R. Beveridge, P. J. Phillips, D. S. Bolme, B. A. Draper, G. H. Givens, Y. M. Lui, M. N. Teli, H. Zhang, W. T. Scruggs, K. W. Bowyer, P. J. Flynn, and S. Cheng. The challenge of face recognition from digital point-and-shoot cameras. In *2013 IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, pages 1–8, Sep. 2013.

[5] Z. Boulkenafet, J. Komulainen, L. Li, X. Feng, and A. Hadid. OULU-NPU: A mobile face presentation attack database with real-world variations. In *12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017)*, pages 612–618, 2017.

[6] I. Chingovska, A. Anjos, and S. Marcel. On the effectiveness of local binary patterns in face anti-spoofing. In *2012 BIOSIG - Proceedings of the International Conference of Biometrics Special Interest Group (BIOSIG)*, pages 1–7, 2012.

[7] D. Cornett, A. Yen, G. Noyola, D. Montez, C. Johnson, S. Baird, H. Santos-Villalobos, and D. Bolme. Through the windshield driver recognition. Technical report, Oak Ridge National Laboratory (ORNL), Oak Ridge, TN (United States), 2018.

[8] J. Deng, J. Guo, N. Xue, and S. Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[9] J. Deng, J. Guo, D. Zhang, Y. Deng, X. Lu, and S. Shi. Lightweight face recognition challenge. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, Oct 2019.

[10] A. George, C. Ecabert, H. O. Shahreza, K. Kotwal, and S. Marcel. Edgeface: Efficient face recognition model for edge devices. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 6(2):158–168, 2024.

[11] A. George and S. Marcel. Deep pixel-wise binary supervision for face presentation attack detection. In *2019 international conference on biometrics (ICB)*, pages 1–8. IEEE, 2019.

[12] A. George and S. Marcel. Robust face presentation attack detection with multi-channel neural networks. In *Handbook of Biometric Anti-Spoofing: Presentation Attack Detection and Vulnerability Assessment*, pages 261–286. Springer, 2023.

[13] A. George and S. Marcel. The invisible threat: Evaluating the vulnerability of cross-spectral face recognition to presentation attacks. *arXiv preprint arXiv:2505.00380*, 2025.

[14] A. George and S. Marcel. xedgeface: Efficient cross-spectral face recognition for edge devices. *arXiv preprint arXiv:2504.19646*, 2025.

[15] A. George, A. Mohammadi, and S. Marcel. Prepended domain transformer: Heterogeneous face recognition without bells and whistles. *IEEE transactions on information forensics and security*, 18:133–146, 2022.

[16] A. George, Z. Mostaani, D. Geissenbuhler, O. Nikisins, A. Anjos, and S. Marcel. Biometric face presentation attack detection with multi-channel convolutional neural network. *IEEE Transactions on Information Forensics and Security*, 15:42–55, 2020.

[17] J. Guo, J. Deng, A. Lattas, and S. Zafeiriou. Sample and computation redistribution for efficient face detection. *arXiv preprint arXiv:2105.04714*, 2021.

[18] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. *ArXiv*, abs/1607.08221, 2016.

[19] D. Huang, J. Sun, and Y. Wang. The buaa-visnir face database instructions. *School Comput. Sci. Eng., Beihang Univ., Beijing, China, Tech. Rep. IRIP-TR-12-FR-001*, 3(3):8, 2012.

[20] M. Kim, A. Jain, and X. Liu. 50 years of automated face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

[21] M. Kim, A. K. Jain, and X. Liu. Adaface: Quality adaptive margin for face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18750–18759, 2022.

[22] K. Kotwal, S. Bhattacharjee, P. Abbet, Z. Mostaani, H. Wei, X. Wenkang, Z. Yaxi, and S. Marcel. Domain-specific adaptation of cnn for detecting face presentation attacks in nir. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 4(1):135–147, 2022.

[23] V. Krivokuca, J. Maceiras, A. Komaty, P. Abbet, and S. Marcel. in-car biometrics (icarb) datasets for driver recognition: Face, fingerprint, and voice. *arXiv*, 2024.

[24] P. Kuchár, R. Pirník, J. Ďurišová, M. Skuba, T. Mizera, and J. Kafková. Effect of window tinting on passenger detection and enforcement in road transport. *Transportation Research Procedia*, 74:938–945, 2023.

[25] R. D. Labati, A. Genovese, E. Muñoz, V. Piuri, F. Scotti, and G. Sforza. Biometric recognition in automated border control: a survey. *ACM Computing Surveys (CSUR)*, 49(2):1–39, 2016.

[26] B. Lee, M. Hasegawa-Johnson, C. Goudeseune, S. Kamdar, S. Borys, M. Liu, and T. Huang. AVICAR: audio-visual speech corpus in a car environment. In *Interspeech 2004*, pages 2489–2492, 2004.

[27] S. Li, D. Yi, Z. Lei, and S. Liao. The casia nir-vis 2.0 face database. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 348–353, 2013.

[28] A. Liu, Z. Tan, J. Wan, S. Escalera, G. Guo, and S. Z. Li. Casia-surf cefa: A benchmark for multi-modal cross-ethnicity face anti-spoofing. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1179–1187, January 2021.

[29] A. Liu, C. Zhao, Z. Yu, J. Wan, A. Su, X. Liu, Z. Tan, S. Escalera, J. Xing, Y. Liang, et al. Contrastive context-aware

learning for 3d high-fidelity mask face presentation attack detection. *IEEE Transactions on Information Forensics and Security*, 2022.

[30] Y. Liu, A. Jourabloo, and X. Liu. Learning deep models for face anti-spoofing: Binary or auxiliary supervision. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 389–398, 2018.

[31] M. Martin, A. Roitberg, M. Haurilet, M. Horne, S. Reiß, M. Voit, and R. Stiefelhagen. Drive&act: A multi-modal dataset for fine-grained driver behavior recognition in autonomous vehicles. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2019.

[32] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[33] J. D. Ortega, N. Kose, P. Cañas, M.-A. Chao, A. Unnervik, M. Nieto, O. Otaegui, and L. Salgado. Dmd: A large-scale multi-modal driver monitoring dataset for attention and alertness analysis. In *Computer Vision – ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part IV*, page 387–405, Berlin, Heidelberg, 2020. Springer-Verlag.

[34] W. Othman, A. Kashevnik, A. Ali, and N. Shilov. Drivermvt: In-cabin dataset for driver monitoring including video and vehicle telemetry information. *Data*, 7(5), 2022.

[35] K. Panetta, Q. Wan, S. Agaian, S. Rajeev, S. Kamath, R. Rajendran, S. P. Rao, A. Kaszowska, H. A. Taylor, A. Samani, and X. Yuan. A comprehensive database for benchmarking imaging systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(3):509–520, 2020.

[36] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[37] M. Tan and Q. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

[38] L. Trinh, P. Pham, H. Trinh, N. Bach, D. Nguyen, G. Nguyen, and H. Nguyen. Pp4av: A benchmarking dataset for privacy-preserving autonomous driving. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1206–1215, 2023.

[39] R. Wan, B. Shi, H. Li, L.-Y. Duan, and A. C. Kot. Face Image Reflection Removal. *International Journal of Computer Vision*, 129(2):385–399, Feb. 2021.

[40] D. Wen, H. Han, and A. K. Jain. Face Spoof Detection With Image Distortion Analysis. *IEEE Transactions on Information Forensics and Security*, 10(4):746–761, Apr. 2015.

[41] S. Woo, S. Debnath, R. Hu, X. Chen, Z. Liu, I. S. Kweon, and S. Xie. Convnext v2: Co-designing and scaling convnets with masked autoencoders. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16133–16142, 2023.

[42] J. You, S. Li, Y. Sun, J. Wei, M. Guo, C. Feng, and J. Ran. Lvface: Progressive cluster optimization for large vision models in face recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11840–11849, 2025.

[43] A. Yu, H. Wu, H. Huang, Z. Lei, and R. He. Lamp-hq: A large-scale multi-pose high-quality database and benchmark for nir-vis face recognition. *International Journal of Computer Vision*, 129:1467 – 1483, 2019.

[44] B. Zhang, L. Zhang, D. Zhang, and L. Shen. Directional binary code with application to polyu near-infrared face database. *Pattern Recognition Letters*, 31(14):2337–2344, 2010.

[45] S. Zhang, A. Liu, J. Wan, Y. Liang, G. Guo, S. Escalera, H. J. Escalante, and S. Z. Li. Casia-surf: A large-scale multi-modal benchmark for face anti-spoofing. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2(2):182–193, 2020.

[46] Y. Zhang, Z. Yin, Y. Li, G. Yin, J. Yan, J. Shao, and Z. Liu. Celeba-spoof: Large-scale face anti-spoofing dataset with rich annotations. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII*, page 70–85, Berlin, Heidelberg, 2020. Springer-Verlag.

[47] Z. Zhang, J. Yan, S. Liu, Z. Lei, D. Yi, and S. Z. Li. A face antispoofing database with diverse attacks. In *2012 5th IAPR International Conference on Biometrics (ICB)*, pages 26–31, 2012.

[48] G. Zhao, X. Huang, M. Taini, S. Z. Li, and M. PietikäInen. Facial expression recognition from near-infrared videos. *Image and vision computing*, 29(9):607–619, 2011.

[49] Z. Zhu, G. Huang, J. Deng, Y. Ye, J. Huang, X. Chen, J. Zhu, T. Yang, J. Lu, D. Du, et al. Webface260m: A benchmark unveiling the power of million-scale deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10492–10502, 2021.