# EDGE20: A Cross Spectral Evaluation Dataset for Multiple Surveillance Problems

Ha Le, Christos Smailis, Lei Shi, and Ioannis Kakadiaris
Computational Biomedicine Lab
Dept. of Computer Science, University of Houston
{hale4, csmailis, lshi22, ikakadia}@central.uh.edu

## Abstract

*Surveillance-related datasets that have been released in recent years focus only on one specific problem at a time (*e.g.*, pedestrian detection, face detection, or face recognition), while most of them were collected using visible spectrum (VIS) cameras. Even though some cross-spectral datasets were presented in the past, they were acquired in a constrained setup, which limited the performance of methods for the aforementioned problems under a cross-spectral setting. This work introduces a new dataset, named EDGE20, that can be used in addressing the problems of pedestrian detection, face detection, and face recognition in images captured using trail cameras under the VIS and NIR spectra. Data acquisition was performed in an outdoor environment, during both day and night, under unconstrained acquisition conditions. The collection of images is accompanied by a rich set of annotations, consisting of person and facial bounding boxes, unique subject identifiers, and labels that characterize facial images as frontal, profile, or back faces. Moreover, the performance of several state-of-the-art methods was evaluated for each of the scenarios covered by our dataset. The baseline results we obtained highlight the difficulty of current methods in the tasks of cross-spectral pedestrian detection, face detection, and face recognition due to unconstrained conditions, including low resolution, pose variation, illumination variation, occlusions, and motion blur.*

## 1. Introduction

The problems of pedestrian detection, face detection, and face recognition have attracted significant attention from the research community in the past. Previous datasets for pedestrian and face detection have explored scenarios of cross-spectral detection in images captured from the visible (VIS), as well as the far-infrared (FIR) [11, 29] and thermal spectra [14], while to the best of our knowledge none have explored the use of images captured in the near-infrared (NIR) spectrum. On the other hand, datasets for cross-spectral face recognition, have explored the near-infrared (NIR) spectrum [17, 15, 13, 18, 20]. However, these datasets are usually captured in well-lit indoor settings, under controlled acquisition conditions and are thus insufficient for a real surveillance scenario.

To address the aforementioned gaps in the literature, we introduce EDGE20, a dataset composed of images collected from trail cameras. Data acquisition was performed under unconstrained conditions, in an outdoor environment, using VIS and NIR cameras, during both day and night time. Subjects were photographed from frontal, profile, and back-face views using four different cameras while wearing head-related accessories that occlude their facial features. The dataset is accompanied by two annotation sets from two different annotators, consisting of person and facial bounding boxes along with subject IDs.

To establish baseline performance for the different problems contained in EDGE20, we evaluated several state-of-the-art methods for pedestrian detection [22], face detection [24], and face recognition [6]. Our experimental results demonstrated the difficulty of the state-of-the-art methods for person and face detection, as well as face recognition to perform well with images captured in an unconstrained outdoor environment and especially when captured with NIR cameras.

The contributions of this work are: (i) A dataset, namely EDGE20, that enables the evaluation of methods for pedestrian and face detection, as well as face recognition in VIS and NIR images. To the best of our knowledge, this is the first dataset that attempts to cover all of the aforementioned problems under the described settings (Section 3). (ii) Baseline results for person detection, face detection, and face recognition by evaluating state-of-the-art models from the literature (Section 5).

The remainder of this paper is organized as follows: Section 2 reviews the related work about datasets for pedestrian detection, face detection, and face recognition. Section 3 discusses the data acquisition and annotation procedures.

Table 1: A summary of pedestrian detection datasets published in the last decade.

| Datasets | Year | # Images | # Pedestrians | VIS | NIR | FIR | Thermal |
|---|---|---|---|---|---|---|---|
| Caltech [7] | 2009 | 249K | 347K | ✓ | | | |
| TUD-Brussels [28] | 2009 | 0.5K | 1.3K | ✓ | | | |
| GM-ATCI [25] | 2014 | 137K | 200K | ✓ | | | |
| KAIST [14] | 2015 | 95K | 86K | ✓ | | | ✓ |
| CVC [11] | 2016 | 10K | 18K | ✓ | | ✓ | |
| CityPersons [31] | 2017 | 5K | 35K | ✓ | | | |
| SCUT [29] | 2018 | 211K | 352K | | | ✓ | |
| **EDGE20** | 2019 | 3.5K | 3.7K | ✓ | ✓ | | |

Table 2: A summary of head and face detection datasets, published within the last decade.

| Datasets | Year | # Images | # Faces/Heads | Backface | VIS | NIR |
|---|---|---|---|---|---|---|
| FDDB [26] | 2010 | 2.8K | 5.2K | | ✓ | |
| AFW [32] | 2012 | 0.2K | 0.5K | | ✓ | |
| MALF [2] | 2015 | 5.3K | 11.9K | | ✓ | |
| HollywoodHeads [27] | 2015 | 224.7K | 369.8K | ✓ | ✓ | |
| WIDER Face [30] | 2016 | 33.2K | 393.7K | | ✓ | |
| SCUT-HEAD [21] | 2018 | 4.4K | 111.3K | ✓ | ✓ | |
| Crowdhuman [23] | 2018 | 19.4K | 470K | ✓ | ✓ | |
| **EDGE20** | 2019 | 3.5K | 3.7K | ✓ | ✓ | ✓ |

Section 4 presents statistics of the EDGE20 dataset and the evaluation protocols used for the different problems that the dataset covers. In Section 5, baseline experimental results from state-of-the-art methods for pedestrian detection [22], face detection [24], and face recognition [6] are reported. Finally, conclusions are drawn in Section 6.

## 2. Related Work

In this section, we provide a brief overview of well-established datasets for different types of surveillance problems that were published in the last decade.

**Datasets for Pedestrian Detection**: A summary of datasets for pedestrian detection released in the past decade is provided in Table. 1. The Caltech pedestrian dataset [7] is one of the most popular and challenging datasets for pedestrian detection. Its contents come from approximately ten hours of video footage recorded by a car traversing on the streets in the greater Los Angeles metropolitan area. Caltech contains around 347K of pedestrians annotated from 249K video frames. In the same year, the TUD-Brussels dataset is released with 508 images of $1,326$ annotated pedestrians. Despite being small in size, the TUD-Brussels is considered challenging because pedestrians appear from multiple viewpoints at small scales. Shai *et al*. [25] presented GM-ATCI datasets collected by a fisheye camera mounted on a vehicle. The dataset contains 250 clips with a total duration of 76 minutes and over 200K annotated pedestrians. CityPersons dataset [31] is a subset of CityScapes dataset [4], which is recorded in street scenes from 50 cities and has high-quality pixel-wise annotations of $5,000$ images.

The aforementioned datasets contain only color images captured under the visible light spectrum. Datasets with images captured under multiple spectra include KAIST [14] and CVC [11]. Data for KAIST were collected using two cameras: a regular color camera and a thermal camera. The two cameras are mounted on a car and captured footage simultaneously from both day and night traffic scenes. KAIST contains around 95K images with 86K annotated pedestrians. Similarly, data for the CVC dataset were collected using two cameras mounted on a car, but a far-infrared (FIR) camera was used instead of a thermal camera. The CVC dataset contains approximately 10K images with 18K annotated pedestrians. Recently, Xu *et al*. [29] introduced a large scale FIR pedestrian detection dataset, named SCUT. The dataset was captured by a monocular FIR camera mounted on a car. SCUT contains 11-hour long image sequences, with 211K frames annotated for a total of 352K pedestrians. In contrast to these datasets, EDGE20 is a multiple spectra dataset collected by a regular color camera and a near-infrared (NIR) camera.

**Datasets for Face and Head Detection**: Table 2 provides a summary of datasets for head and face detection. Older datasets for face detection include FDDB [26], AFW [32], and MALF [2]. However, with the progress of face detection algorithms, these datasets were overcome by a more challenging dataset, known as the WIDER-Face dataset [30], which includes faces of multiple scales in complex environments. WIDER-Face became the most popular dataset for evaluating face detection algorithms. While there has been impressive progress towards face detection in the past, the more general problem of detecting human heads in images remains challenging. Several datasets for head detection have been introduced to assess the performance of related models. HollywoodHeads [27] is the largest head detection dataset with 224.7K images and 369.8K annotated heads. After HollywoodHeads, two large-scale head datasets were published, including SCUT-HEAD [21] and Crowdhuman [23]. Unlike the previously mentioned datasets, EDGE20 contains head images from both VIS and NIR spectra.

**Datasets for NIR-VIS Face Recognition**: Several datasets for cross-spectral face recognition have been published in the past. A comparative list of all the datasets for VIS-NIR face recognition and their characteristics are summarized in Table 3. CASIA NIR-VIS 2.0 [18] contains images from 725 different subjects captured in an indoor environment. Some of the subjects in the CASIA NIR-VIS 2.0 dataset wore glasses that acted as a form of facial occlusion. Other datasets with similar acquisition conditions include Oulu-CASIA NIR-VIS [15], BUAA-VisNir [13], and HFB [17], but they contain fewer subjects (80, 150, and 100, respectively). The Oulu-CASIA NIR-VIS dataset was designed with the additional goal of introducing expression recognition as a challenge. The dataset contains 80 subjects captured under six different expressions. ND-NIVL dataset is one of the most recent ones, which contains 24,605 im-

Table 3: A summary of NIR-VIS face recognition datasets.

| Datasets | Year | # Subjects | # Faces | Indoor | Outdoor | Resolution Variations | Pose Variations | Expression Variations | Facial Occlusions | Motion Blur |
|---|---|---|---|---|---|---|---|---|---|---|
| HFB [17] | 2009 | 100 | 992 | ✓ | | | | | | |
| Oulu-CASIA NIR-VIS [15] | 2009 | 80 | 64,912 | ✓ | | | | ✓ | | |
| BUAA-VisNir [13] | 2012 | 150 | 3,900 | ✓ | | | | | | |
| LDHF-DB [20] | 2012 | 100 | 1,600 | ✓ | ✓ | ✓ | | | | |
| CASIA NIR-VIS 2.0 [18] | 2013 | 725 | 17,580 | ✓ | | | | | ✓ | |
| ND-NIVL [1] | 2015 | 574 | 24,605 | ✓ | | | | | | |
| **EDGE20** | 2019 | 197 | 3,724 | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

ages captured from 574 subjects. One of the most challenging datasets for NIR-VIS Face recognition is LDHF-DB [20]. Despite containing a relatively small number of subjects (100), this dataset includes several challenges, as it comprises of images captured both in indoor and outdoor environments, under multiple distances (60m, 100m, and 150m). In comparison to previous datasets, EDGE20 attempts to offer a wider variety of challenges for the NIR-VIS face recognition problem, using images captured under unconstrained conditions.

## 3. The EDGE20 Dataset

EDGE20 is a dataset that covers problems related to cross-spectral surveillance scenarios in open spaces, using images captured from trail cameras, during day and night. Specifically, the dataset consists of images captured through the VIS and NIR spectra. The dataset provides rich sets of annotations to support the evaluation of pedestrian detection, face detection, and face recognition with different settings.

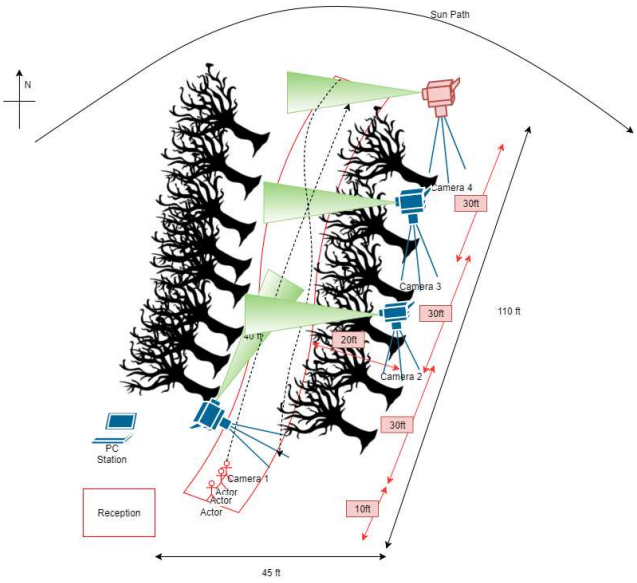



Figure 1: Diagram of the data acquisition scene. The scene was set-up to mimic a bi-directional trail path. The green cones represent field of view of the cameras. All cameras were mounted at eight to 10 ft high. Subjects were asked to walk freely inside the trail from a beginning point to the end of the trail and then return back.

### 3.1. Scene Configuration

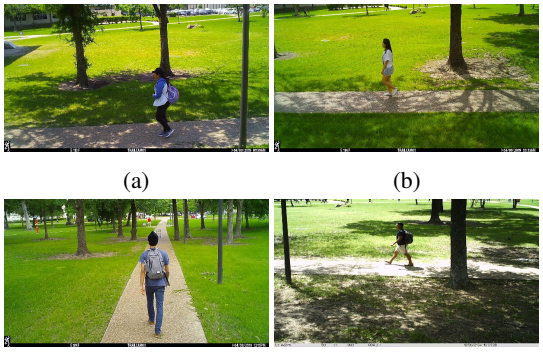
The scene used for the acquisition resembled a two-way trail path that allowed bi-directional flow for the subjects. The width of the scene was 45 ft, while its diagonal length was 110 ft. Subjects were asked to walk back and forth between the beginning and the end of the path. After reviewing and agreeing to sign an Institutional Review Board (IRB) approved participation consent form, each subject was asked to walk for two rounds. In the first round, the subjects were assigned random accessories to wear or carry, which included hats, beanies, backpacks, and poster tubes. In the second round, the subjects walked without accessories. The acquisition of images for the EDGE20 dataset was performed using four NIR-VIS trail cameras (three Browning Strike Force trail cameras and one Ltl Acorn 6310MG camera). The four cameras were positioned close to the path, as noted in Figure 1, adhering to the following requirements:

- Cameras were placed within a 40 ft distance from the trail.
- Three of the cameras had a line of view perpendicular to the direction of the path, while one of them was placed to have a line of view that was horizontal to the path.
- Cameras were positioned to avoid facing the sun directly.
- Cameras were placed higher than a person's eye level.

In Figure 1, the green cone next to each camera represents the field of view of that camera. The cameras were positioned at different heights using tripods to mimic the variation that could exist in a real trail camera setup.

### 3.2. Data Acquisition

Images for the EDGE20 dataset were collected from 197 subjects during two sessions (day and night). From these



(a) (b) (c) (d)

Figure 2: Sample images captured during the day session of the EDGE20 dataset: (a) image from Camera 1, (b) image from Camera 2, (c) image from Camera 3, and (d) image from Camera 4.

(a) (b) (c) (d)

Figure 3: Sample NIR images captured during the night session of the EDGE20 dataset: (a) image from Camera 1, (b) image from Camera 2, (c) image from Camera 3, and (d) image from Camera 4.

subjects, 197 participated in the day session while 68 participated in the night session. In total, 3,494 images were collected. Sample images captured during the day and night sessions are depicted in Figures 2 and 3, respectively. Some of the images contain only one individual subject, while others contain multiple subjects. Figure 4 depicts images of multiple subjects in the EDGE20 dataset. Since data acquisition occurred in an open area during crowded times, individuals who were captured in images without having signed a consent form were manually blurred in the acquired images to preserve their privacy. Since both day an night sessions took place during two different days, many of the subjects wear the same clothes.

### 3.3. Annotation

To annotate the collected images, we developed a web-based annotation tool to aid users in producing image annotations for a variety of computer vision tasks. The tool allows users to incorporate the output of separate automated person detection and pose estimation methods (e.g., AlphaPose by Fang *et al*. [9]) to speed up the annotation process. Users can create new pedestrian and face bounding boxes along with pose annotations or refine the proposals produced by automated methods while focusing on tasks such as providing labels related to persons, objects, and scene attributes of an image.

The annotations provided by EDGE20 consist of the following types of information: (i) subject IDs, (ii) person bounding boxes, (iv) face bounding boxes, (v) face labels: frontal, profile, back. Two human annotators were involved in the task of image annotation. Statistics from the data of two annotators, for both day and night sessions, are summarized in Table 4.

(a) (b) (c) (d)

Figure 4: Sample images of pedestrian groups in the EDGE20 dataset: (a) visible image acquired from the frontal view, (b) visible image acquired from the profile view, (c) NIR image acquired from the frontal view, and (d) NIR image acquired from the profile view.

## 4. Evaluation

The annotations provided by the EDGE20 dataset, are used to define image sets, for different evaluation configurations, for the problems of (i) Pedestrian Detection (PD), (ii) Face Detection (FD) and (iii) Face Recognition (FR).

### 4.1. Image Sets

The full list of image sets available in the EDGE20 dataset is summarized in Table 4. The two main sets of images from which the rest are derived, are *"VIS Images"* (**V**), and *"NIR Images"* (**N**). The image sets **V** and **N** contain full scene images captured from the VIS and NIR spectra, respectively. The *"VIS Images"* set is split into three subsets, namely *"VIS Frontal Face ROIs"* (**VF**), *"VIS Profile Face ROIs"* (**VP**), and *"VIS Back Face ROIs"* (**VB**). Sim-

Table 4: Annotation summary for the EDGE20 dataset. The different types of images contained in the dataset are denoted by the "*Set*" column.

| Set | Description | # Images | # Subjects | Annotator 1 | | Annotator 2 | |
|---|---|---|---|---|---|---|---|
| | | | | # Pedestrians | # Faces | # Pedestrians | # Faces |
| V | VIS Full Scene | 2,697 | 197 | 2,931 | 2,473 | 2,952 | 2,500 |
| VF | VIS Frontal Face ROIs | - | 68 | - | 448 | - | 448 |
| VP | VIS Profile Face ROIs | - | 197 | - | 2,025 | - | 2,052 |
| VB | VIS Back Face ROIs | - | 197 | - | 427 | - | 430 |
| N | NIR Full Scene | 797 | 68 | 806 | 746 | 802 | 741 |
| NF | NIR Frontal Face ROIs | - | 68 | - | 143 | - | 143 |
| NP | NIR Profile Face ROIs | - | 68 | - | 603 | - | 598 |
| NB | NIR Back Face ROIs | - | 68 | - | 56 | - | 53 |
| VG | VIS Gallery Face ROIs | - | 197 | - | 197 | - | 197 |
| VFP | VIS Probe Face ROIs | - | 197 | - | 252 | - | 252 |
| NG | NIR Gallery Face ROIs | - | 68 | - | 68 | - | 68 |
| NFP | NIR Probe Face ROIs | - | 68 | - | 96 | - | 95 |

(a) (b) (c)

Figure 5: Annotated examples demonstrating pedestrian and face bounding boxes: (a) a frontal shot from an NIR image, (b) a profile shot from a VIS image, and (c) a shot from a VIS image (in this case the annotated face bounding box is assigned the "back" label).

ilarly, the *"NIR Images"* is split into three subsets: *"NIR Frontal Face ROIs"* (**NF**), *"NIR Profile Face ROIs"* (**NP**), and *"NIR Back Face ROIs"* (**NB**). These subsets contain facial ROIs, and as the naming suggests, they were labeled according to the view under which the face was captured (Frontal, Profile, and Back).

For face recognition experiments, a subset of frontal, non-occluded, high-resolution face images was manually selected from the **VF** to form the gallery. This subset is then named *"VIS Gallery Face ROIs"* (**VG**) and contains only one image per subject. The VIS face images that are not contained in the **VG** set are used to form the set of VIS probe images, named *"VIS Face Probe ROIs"* (**VFP**). Similarly, a subset of frontal non-occluded and high-resolution face images is manually selected from the set **NF** to represent the *"NIR Face Gallery ROIs"* (**NG**). Images that were not included in **NG** from the set **NF** are used in the *"NIR Face Probe ROIs"* set, abbreviated as (**NFP**).

Table 5: Evaluation protocols of EDGE20: two protocols for pedestrian detection, three protocols for face detection, and eight protocols for face recognition. For each face recognition protocol, the first image set is the gallery, and the second image set is the probe.

| Protocol | Problem | Image Sets |
|---|---|---|
| PD1 | VIS PD | V |
| PD2 | NIR PD | N |
| FD1 | VIS Frontal FD | VF |
| FR1 | VIS FR | VG - VFP |
| FR2 | VIS FR - Profile Probes | VG - VP |
| FR3 | VIS-NIR FR | VG - NFP |
| FR4 | VIS-NIR FR - Profile Probes | VG - NP |
| FR5 | NIR-VIS FR | NG - VFP |
| FR6 | NIR-VIS FR Profile Probes | NG - VP |
| FR7 | NIR-NIR FR | NG - NFP |
| FR8 | NIR-NIR FR Profile Probes | NG - NP |

## 4.2. Protocols

Table 5 demonstrates 11 evaluation protocols for surveillance problems: (i) two for pedestrian detection, (ii) one for face detection, and (iii) eight for face recognition. The evaluation procedure for pedestrian detection and face detection problems involves the comparison between the detected and the groundtruth bounding boxes. Similarly, the eight evaluation protocols of face recognition assess the performance of algorithms using data captured from the combinations of the frontal and profile views, for VIS-VIS, NIR-NIR, VIS-

NIR, and NIR-VIS matching tasks. Each of these protocols is equivalent to a 1:N identification problem using a gallery and a probe. To ensure a close-set scenario in the FR5 and FR6 protocols, only subjects that appeared in the gallery are kept in the corresponding probe. All evaluation protocols are performed two times with the groundtruth provided by two annotators to model the bias of each annotator separately.

### 4.3. Performance Metrics

#### 4.3.1 Pedestrian Detection and Face Detection

The general object detection problem can be separated into two subtasks: classification and localization. A simple accuracy-based metric that ignores misclassifications is biased and insufficient for the evaluation of the classification task. For that purpose, the average precision score, which considers both precision and recall, is used to evaluate object classification models. In order to evaluate the localization component of a detection model, Intersection over Union (IoU) is used to summarize how well the ground truth object overlaps the predicted boundary. By taking the mean value of average precision over all classes at different IoU thresholds, mean Average Precision (mAP) [8] has become the most common way to evaluate object detection performance.

The log-average miss rate ($\text{MR}^{-2}$) [31] is leveraged for the evaluation of pedestrian detection. It is computed by averaging the miss rate at nine false positives per image (FPPI) rates that are evenly distributed in the log-space from $10^{-2}$ to $10^{0}$.

The detection rate (DR) is used to evaluate the performance of face detection on different subsets. It is defined as the percentage of the number of detected faces given all annotated faces in a subset where the IoU threshold is set to $0.5$ [33]. In the face detection protocol FD1, besides mAP, the DR of frontal faces (DR-FF) and the DR of profile faces (DR-PF) are also computed to assess the performance of face detector at different views.

#### 4.3.2 Face Recognition

The Cumulative Match Characteristic (CMC) curve [16] is used to evaluate face recognition protocols. CMC is the standard metric for closed-set identification protocols. It represents the percentage of probe matches that have at least one true-positive match within the top $k$ sorted ranks within the gallery. The value of $k$ can vary from 1 to the number of images in the gallery. However, $k$ is commonly bounded to a fixed number. For EDGE20, $k$ is bounded to 1, which means that only the most similar pair between an image in the probe and images in the gallery is taken into account.

Table 6: The evaluation results of Faster R-CNN on EDGE20 for the task of pedestrian detection with two protocols **PD1** and **PD2**. In a cell, the first number is **mAP**, and the second number is $\textbf{MR}^{-2}$.

| Protocol | PD1 | PD2 |
|---|---|---|
| **Annotator 1** | 72.93 / 36.48 | 74.01 / 35.08 |
| **Annotator 2** | 79.22 / 29.33 | 90.07 / 27.83 |

Table 7: The evaluation results of SANet on EDGE20 for the task of face detection with the protocol **FD1**. We report the detection rate of frontal faces (**DR-FF**), the detection rate of profile faces (**DR-PF**), and **mAP** score.

| Metrics | DR-FF | DR-PF | mAP |
|---|---|---|---|
| **Annotator 1** | 91.66 | 76.22 | 69.58 |
| **Annotator 2** | 92.05 | 74.22 | 70.49 |

## 5. Results

### 5.1. Pedestrian Detection

To establish a baseline for pedestrian detection, we employed Faster R-CNN [22] as it has been adopted as a baseline detector for many pedestrian detection datasets in the past. Specifically, for our Faster R-CNN baseline, we used the implementation offered by the MMDetection toolbox [3]. For this baseline, Faster R-CNN used a ResNet-50 model [12] and was pre-trained using the COCO dataset [19]. The evaluation results for protocols PD1 and PD2 can be found in Table 6. We can observe that the performance obtained for VIS images in PD1 is higher than the performance for NIR images in PD2 by $1.4\%$ and $1.50\%$ for Annotators 1 and 2, respectively. The qualitative results for the task of pedestrian detection that are depicted in Figure 6 also confirm the point that detecting pedestrians in NIR images is more complicated than detecting pedestrians in VIS images. As can be observed from Figure 6 NIR images contain challenges related to low resolution and motion blur.

### 5.2. Face Detection

As a baseline for face detection, we employed SANet [24]. First, we trained and evaluated SANet using the WIDER Face dataset. On the WIDER Face dataset, SANet achieved mAP values of $88.3\%$ and $88.2\%$ on the hard subsets of the validation and testing sets, respectively. To better understand the challenges of EDGE20, detection rates are computed for the subsets of frontal faces and profile faces. A lower detection rate indicates a more challenging subset. Table 7 shows the evaluation results of SANet on EDGE20 for different annotators. Qualitative results of SANet on EDGE20 are depicted in Figure 7. The detection rate results from SANet on frontal faces outperform that on the profile faces by $15.44\%$ for Annotator 1 and $17.83\%$ for An-

Figure 6: Qualitative results of Faster R-CNN on EDGE20 dataset for the pedestrian detection problem. In each image, red boxes denote groundtruth, green boxes denote results of Faster R-CNN. Images in the dashed-green box represent success detection results, and the images in the dashed-red box represents failure cases.

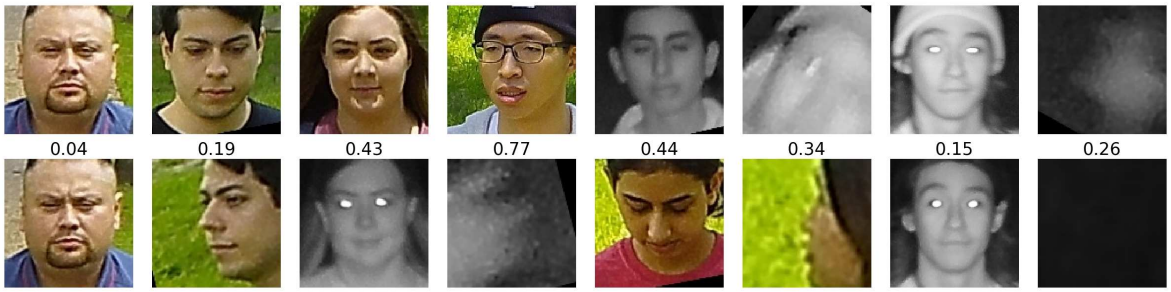
Figure 7: Qualitative results of SANet on EDGE20 dataset for the face detection problem. In each image, red boxes denote groundtruth, and green boxes denote results of SANet. Images in the dashed-green box represent success detection results, and the images in the dashed-red box represent failure cases.

notator 2. The result values from Table 7 showed that it is more difficult for SANet to detect profile faces than frontal faces, which also has been demonstrated in Figure 7. Finally, the mAP score is being used for evaluating SANet on both frontal and profile faces, achieving scores of $69.58\%$ and $70.49\%$ for Annotator 1 and Annotator 2, respectively.

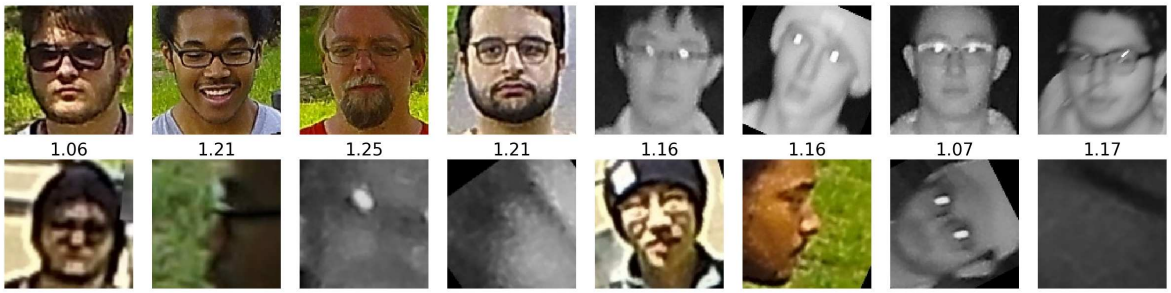
### 5.3. Face Recognition

To establish a baseline for face recognition, we employed ArcFace [6]. Following the recent trend of incorporating margins in well-established loss functions for maximizing face class separability in Deep Convolutional Neural Networks (DCNNs), the ArcFace method adopted an Additive Angular Margin Loss (ArcFace) to obtain highly discriminative features for face recognition. We trained ArcFace on the Deep Glint dataset [5]. We assessed the performance of our trained ArcFace model for the task of NIR-VIS face recognition using the CASIA NIR-VIS 2.0 dataset and achieved the state-of-the-art performance, 99.97% rank-1 identification rate. To establish our results on EDGE20, we used the groundtruth face bounding boxes and applied PRNet [10] for landmark detection. Cosine similarity is used to measure the similarity of two feature vectors generated by ArcFace. Given two feature vectors $u$ and $v$, their cosine similarity is computed as $1 - \cos(u, v)$. The similarity ranges from 0, which indicates perfect similarity, to 2, that corresponds that two faces are completely different.

The evaluation results of ArcFace on EDGE20 are shown in Table. 8. The results showed that even in the same domain, it is harder to perform face recognition for profile face images. In the VIS domain, performance dropped for at least 55% (the difference between FR1 and FR2) when switching from a frontal face probe to a profile face probe. In the NIR domain, the performance gap is at least 25% (the difference between FR5 and FR6). The lowest rank-1 identification rates are achieved for FR4 and FR8, which means that there is plenty of room for the improvement of identification of profile faces in the NIR spectrum. The low perfor-




(a)




(b)

Figure 8: Qualitative results of ArcFace on EDGE20 dataset for the face recognition problem. The number in between each pair is the cosine similarity: (a) true positive pairs, (b) false negative pairs

Table 8: Rank-1 identification rate of ArcFace on EDGE20.

| **Protocols** | **FR1** | **FR2** | **FR3** | **FR4** | **FR5** | **FR6** | **FR7** | **FR8** |
|---|---|---|---|---|---|---|---|---|
| **Annotator 1** | 87.94 | 33.33 | 43.88 | 1.20 | 36.54 | 11.31 | 47.96 | 1.89 |
| **Annotator 2** | 87.64 | 31.19 | 45.83 | 1.56 | 41.41 | 11.24 | 46.88 | 2.25 |

mance on profile face images is caused by self-occlusion, occlusions by accessories, the low resolution of face images, and the blur caused by the motion of subjects. We also observe that although the performance was adequate for the VIS-VIS settings, the task of performing face recognition in NIR-VIS and NIR-NIR settings is more complicated. Qualitative results from the evaluation of ArcFace with the EDGE20 dataset are depicted in Figure 8. Figure 8a contains examples of true positive pairs with relatively small cosine similarity, while Figure 8b has examples of false negative pairs with high cosine similarity. The eight pairs of images from left to right depict results from the eight face recognition protocols from FR1 to FR8, respectively. From Figure 8b, we can observe that the failure cases are due to low resolution (the first pair), profile pose (the second pair), motion blur (the third pair), and misalignment (the fourth pair). Similar challenges are observed in the last four pairs.

## 6. Conclusion

In this paper, we presented EDGE20, a dataset with VIS and NIR images. EDGE20 is accompanied by detailed groundtruth labels, including pedestrian bounding boxes, face bounding boxes, and face labels. Moreover, EDGE20 specifies 11 evaluation protocols for variations of the problems of (i) pedestrian detection, (ii) face detection, and (iii) face recognition. These protocols allow the evaluation of the different models that can act as parts of a surveillance system under both the VIS and NIR modalities. The experiments we performed with baseline methods demonstrated the variety of challenges related to unconstrained and cross-domain detection and recognition problems that are part of the EDGE20 dataset. EDGE20 will become publicly available for research purposes.

**Acknowledgement** This material is based upon work supported by the U.S. Department of Homeland Security under Grant Award Number 2017-ST-BTI-0001-0201. This grant is awarded to the Borders, Trade, and Immigration (BTI) Institute: A DHS Center of Excellence led by the University of Houston, and includes support for the project EDGE awarded to the University of Houston. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Department of Homeland Security.

## References


[1] J. Bernhard, J. Barr, K. W. Bowyer, and P. Flynn. Near-IR to visible light face matching: Effectiveness of pre-processing options for commercial matchers. In *Proc. IEEE International Conference on Biometrics Theory, Applications and Systems*, pages 1–8, Arlington, VA, 9 2015.

[2] Bin Yang, Junjie Yan, Zhen Lei, and S. Z. Li. Fine-grained evaluation on face detection in the wild. In *Proc. IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, pages 1–7, Ljubljana, Slovenia, 5 2015.

[3] K. Chen, J. Wang, J. Pang, Y. Cao, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Xu, Z. Zhang, D. Cheng, C. Zhu, T. Cheng, Q. Zhao, B. Li, X. Lu, R. Zhu, Y. Wu, J. Dai, J. Wang, J. Shi, W. Ouyang, C. C. Loy, and D. Lin. MMDetection: Open MMLab detection toolbox and benchmark. *arXiv preprint*, 6 2019.

[4] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The Cityscapes dataset for semantic urban scene understanding. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 3213–3223, Las Vegas, NV, 6 2016.

[5] DeepGlint. Trillionpairs (http://trillionpairs.deepglint.com), 2018.

[6] J. Deng, J. Guo, and S. Zafeiriou. ArcFace: Additive angular margin loss for deep face recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–12, Long Beach, CA, 6 2019.

[7] P. Dollar, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: A benchmark. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 304–311, Miami, FL, 6 2009.

[8] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The Pascal visual object classes (VOC) challenge. *International Journal of Computer Vision*, 88(2):303–338, 6 2010.

[9] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu. RMPE: Regional multi-person pose estimation. In *Proc. IEEE International Conference on Computer Vision*, pages 2353–2362, Venice, Italy, 10 2017.

[10] Y. Feng, F. Wu, X. Shao, Y. Wang, and X. Zhou. Joint 3D face reconstruction and dense alignment with position map regression network. In *Proc. European Conference in Computer Vision*, Munich, Germany, 9 2018.

[11] A. González, Z. Fang, Y. Socarras, J. Serrat, D. Vázquez, J. Xu, and A. López. Pedestrian detection at day/night time with visible and FIR cameras: A comparison. *Sensors*, 16(6):820, 6 2016.

[12] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, Las Vegas, NV, 6 2016.

[13] D. Huang, J. Sun, and Y. Wang. The BUAA-VisNir face database instructions. Technical report, Beihang University, Beijing, China, 2012.

[14] S. Hwang, J. Park, N. Kim, Y. Choi, and I. S. Kweon. Multispectral pedestrian detection: Benchmark dataset and baseline. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 1037–1045, Boston, MA, 6 2015.

[15] Jie Chen, D. Yi, Jimei Yang, Guoying Zhao, S. Z. Li, and M. Pietikainen. Learning mappings for face synthesis from near infrared to visual light images. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 156–163, Miami, FL, 6 2009.

[16] N. D. Kalka, J. A. Duncan, and J. Dawson. IARPA Janus benchmark multi-domain face. In *Proc. IEEE International Conference on Biometrics: Theory, Applications and Systems*, pages 1–9, Tampa, FL, 9 2019.

[17] S. Z. Li, Z. Lei, and Meng Ao. The HFB face database for heterogeneous face biometrics research. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–8, Miami, FL, 6 2009.

[18] S. Z. Li, D. Yi, Z. Lei, and S. Liao. The CASIA NIR-VIS 2.0 face database. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 348–353, Portland, OR, 6 2013.

[19] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: Common objects in context. In *Proc. European Conference in Computer Vision*, pages 740–755, Zurich, Switzerland, 9 2014.

[20] H. Maeng, S. Liao, D. Kang, S.-W. Lee, and A. K. Jain. Nighttime face recognition at long distance: Cross-distance and cross-spectral matching. In *Proc. Asian Conference on Computer Vision*, pages 708–721, Daejeon, Korea, 11 2012.

[21] D. Peng, Z. Sun, Z. Chen, Z. Cai, L. Xie, and L. Jin. Detecting heads using feature refine net and cascaded multi-scale architecture. In *Proc. International Conference on Pattern Recognition*, pages 2528–2533, Beijing, China, 8 2018.

[22] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Proc. Advances in Neural Information Processing Systems 28*, pages 91–99, Montreal, Canada, 12 2015.

[23] S. Shao, Z. Zhao, B. Li, T. Xiao, G. Yu, X. Zhang, and J. Sun. CrowdHuman: A benchmark for detecting human in a crowd. *arXiv preprint*, 4 2018.

[24] L. Shi, X. Xu, and I. A. Kakadiaris. SANet: Smoothed attention network for single stage face detector. In *Proc. International Conference on Biometrics*, Crete, Greece, 6 2019.

[25] S. Silberstein, D. Levi, V. Kogan, and R. Gazit. Vision-based pedestrian detection for rear-view cameras. In *Proc. IEEE Intelligent Vehicles Symposium Proceedings*, pages 853–860, Dearborn, MI, 6 2014.

[26] Vidit Jain and E. Learned-Miller. FDDB: A benchmark for face detection in unconstrained settings. Technical report, University of Massachusetts, Amherst, Amherst, MA, 2010.

[27] T.-H. Vu, A. Osokin, and I. Laptev. Context-aware CNNs for person head detection. In *Proc. IEEE International Conference on Computer Vision*, pages 2893–2901, Santiago, Chile, 12 2015.

[28] C. Wojek, S. Walk, and B. Schiele. Multi-cue onboard pedestrian detection. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 794–801, Miami, FL, 6 2009.

[29] Z. Xu, J. Zhuang, Q. Liu, J. Zhou, and S. Peng. Benchmarking a large-scale FIR dataset for on-road pedestrian detection. *Infrared Physics & Technology*, 96:199–208, 1 2019.

[30] S. Yang, P. Luo, C. C. Loy, and X. Tang. WIDER FACE: A face detection benchmark. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 5525–5533, Las Vegas, NV, 6 2016.

[31] S. Zhang, R. Benenson, and B. Schiele. CityPersons: A diverse dataset for pedestrian detection. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 4457–4465, Honolulu, HI, 7 2017.

[32] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 2879–2886, Providence, RI, 6 2012.

[33] C. L. Zitnick and P. Dollár. Edge boxes: Locating object proposals from edges. In *Proc. European Conference in Computer Vision*, pages 391–405, Zurich, Switzerland, 9 2014.