MS-EVS: Multispectral event-based vision for deep learning based face detection

Saad Himmi* Vincent Parret† Ajad Chhatkuli* Luc Van Gool*

Abstract

Event-based sensing is a relatively new imaging modality that enables low latency, low power, high temporal resolution and high dynamic range acquisition. These properties make it a highly desirable sensor for edge applications and in high dynamic range environments. As of today, most event-based sensors are monochromatic (grayscale), capturing light from a wide spectral range over the visible, in a single channel. In this paper, we introduce multispectral events and study their advantages. In particular, we consider multiple bands in the visible and near-infrared range, and explore their potential compared to monochromatic events and conventional multispectral imaging for the face detection task. We further release the first large scale bimodal face detection datasets, with RGB videos and their simulated color events, N-MobiFace and N-YoutubeFaces, and a smaller dataset with multispectral videos and events, N-SpectralFace. We find that early fusion of multispectral events significantly improves the face detection performance, compared to the early fusion of conventional multispectral images. This result shows that multispectral events carry relatively more useful information about the scene than conventional multispectral images do, with respect to their grayscale equivalent. To the best of our knowledge, our proposed method is the first exploratory research on multispectral events, specifically including near infrared data.

1. Introduction

Event cameras, also known as neuromorphic cameras or Event-based Vision Sensors (EVS), are novel vision sensors that operate differently from conventional cameras, e.g. Active Pixel Sensors (APS). While conventional cameras capture a series of frames at a fixed rate, event cameras capture asynchronous and quasi-continuous streams of events [15,33,46]. Each event is generated independently, at the pixel level, when the pixel intensity changes by a certain threshold value, resulting in a high temporal resolution (microsecond) and potentially a very low power consumption, as much less data is transmitted. Some other properties of the event cameras are the absence of motion blur and their high dynamic range: it can operate well in both very bright and very dark environments simultaneously. This makes event cameras particularly suitable for high-speed and low-latency applications such as robotics, autonomous vehicles, and virtual/mixed reality.

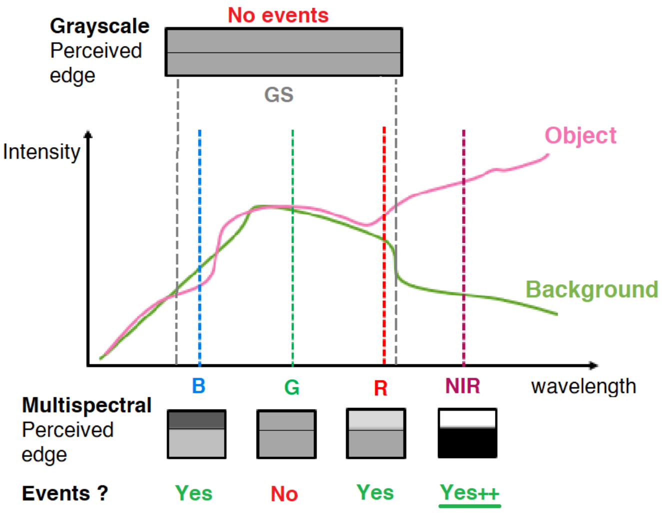


Figure 1. Illustration of the additional information retrieved when considering multispectral event-based data. Multispectral events enable the perception of color contrast and provide more selective information about the scene. Additionally, extra-visible bands could be more adequate in some scenarios, yielding the biggest contrast. Best viewed in color.

Despite their advantages, event cameras have been mostly limited to monochromatic vision, which limits the range of information they can capture. Development is at least partly hampered by the fact that little amount of research exists on the advantages of using different or multiple channels of event acquisition for computer vision systems. Indeed, existing event-based cameras are primarily based on grayscale intensity changes. Considering a scenario like in figure 1, different objects have a different pixel intensity, depending on the scene illumination, the cam-

[*] Computer Vision Laboratory, ETH Zürich, Switzerland
[†] Stuttgart Laboratory 1, Sony Semiconductor Solutions Europe, Sony Europe B.V.

era sensitivity and more importantly the object's reflectance spectrum (eq. 1). We represent the object and background's intensity across wavelengths in the figure, for a single pixel. In this scene, a regular monochromatic grayscale event camera (on top) would not perceive any edge (intensity change) because the overall intensity in the selected range is the same. As the operation of event cameras depends on brightness changes, it leads on purpose to less data being captured, especially with low sensitivity parameters. If the contrast is too low, like the Grayscale row in fig. 1, no event data will be generated except for noise, whereas conventional cameras can capture slight differences, especially for high bit depth cameras. To alleviate this issue, one could consider a multispectral event camera that outputs events only for narrow bands in the visible (e.g. Blue, Green, Red). From the different channels, we could now perceive intensity differences, in particular for the Blue and Red channel and events will be generated. Moreover, grayscale event cameras would miss out on information beyond the visible spectrum: if the object and background have very different reflectance in the near infrared, it would be a relevant band to use for our application. Our goal in this work is to explore how informative multispectral event-based data could be for face detection and understand if it shows any advantage over regular monochromatic events.

In this paper, we propose an explorative study of multispectral events. We formalize their concept and setup experiments to compare their performance against regular grayscale events and conventional images for face detection. Face detection is an important task in computer vision and has a wide range of applications, such as surveillance, biometrics, and human-robot interaction, among many. Also, the human skin spectrum is well studied [2, 11, 36] and it has interesting properties beyond the visible spectrum that generalize to any skin type. Finally, event cameras have a lot of potential in handheld devices, mobile robots or in surveillance and monitoring applications, often interacting with humans. For these reasons, we chose face detection as our use-case task but we believe that our observations and conclusions may carry over to other recognition tasks. Our main contributions are:

1. We propose the 3 **first ever bimodal datasets with multispectral events**, from simulation, labeled for face detection. Two of them are converted from open-source large-scale grayscale and color (RGB) datasets [34, 55]; the third one is smaller but features data over 10 spectral bands, from blue to near-infrared. By bimodal, we mean temporally and spatially aligned conventional images (APS) and their simulated event-based data (EVS).

2. We propose the first **fair comparison of APS and EVS-based face-detection**, comparing the robustness of events over conventional images when both use the same neural architecture and have been trained on the same data. This differs from usual approaches where APS and EVS models are either trained on different imbalanced data or using hardly comparable algorithms.

3. We introduce the novel idea of **multispectral events**, capturing events over multiple spectral bands, and **we explore what benefits they could have over monochromatic grayscale events**. We show that event-based face detection is significantly improved by the use of multispectral events, beyond the effects we observe with conventional multispectral images. Additionally, even though we perfom all our experiments with simulated events, we validate our observations on a small set of real multispectral event-based sequences.

In this paper, we first take a look back on existing work for event-based face detection, the use of color or multispectral bands for face detection and the early attempts of multispectral event-based sensing, in section 2. Then, in section 3, we formalize the multispectral events concept and describe a set of experiments to evaluate what advantages multispectral events could bring, either in separate or combined channels. To perform our experiments, and because multispectral events are a novel idea, we had to build different datasets that we also release in this paper. Finally, we report the results of our experiments in section 4, and discuss these results in section 5.

2. Related work

Event-based face detection. While conventional face detection is a mature problem in Computer Vision, event-based face detection is in very early stages of research. It has a lot of potential for always-on applications thanks to the low latency and low power consumption properties of event cameras [5, 15] but it is a new problem to solve due to the change in data paradigm (frames to events). We distinguish two approaches: detection over reconstructed images, or detection over events. *Barua et al. (2016)* [4], the earliest work on event-based face detection, tried both approaches. While detection over reconstructed images is convenient and enables the use of well-known face detectors (e.g. [53]), it negates most advantages of event cameras - although the reconstructed video has a higher dynamic range and framerate. Equivalently to image-based face detection, over the time, research shifted from the use of hand-crafted features (as in [4, 29]) to deep learning approaches. First, researchers [6, 8, 12, 24, 54] tried to directly apply well-studied object detection architectures (SSD [35], YOLO [42] or Faster R-CNN [43]) to 2D event representations (e.g. [17, 28, 48]), mostly encoding temporal information too. More recent works [18, 26, 30, 39, 44] used the fact that events intrinsically only account for changes in the scene and added memory to their architecture. As our goal is to explore the effects of multispectral events for face detection, we chose not to run our experiments on the latest event-based face detection model but on a network that can be used for both modalities so that the only differences lie in the input data. We decided to use RetinaFace [13], a non-recurrent CNN architecture, known for face detection, to perform our experiments. Although we need to keep in mind that the network we used was designed for APS images and that the overall results with EVS might be improved with recurrency, we argue that our approach is valid to quantify the benefit of multiple spectral bands for each modality and should generalize to event-based recurrent neural networks.

Color/Multispectral face detection. Most of the literature involving multispectral data and faces are focused on two areas, either face recognition using multispectral biometrics (e.g. [10, 32, 49]) or face detection in thermal infrared (e.g. [7]). Due to event cameras' limited sensitivity in the far infrared [50], thermal infrared face detection papers are not relevant to multispectral event-based sensing. A few works [10, 32, 56] tried to tackle the task of near-infrared (NIR) face detection, mainly to solve illumination challenges such as light-variations or low light for systems in dark environments (c.f. [38, 49]). To the best of our knowledge, [58] is the only work that look at a deep learning method for face detection using multispectral bands. Also, hyperspectral imaging, the use of many narrow spectral bands, is relatively unexplored for face detection compared to land/vegetation classification or segmentation [3, 9, 31], or even chemometrics [25].

In our work, we only explore early fusion of multispectral data (images/events) by stacking the different channels in a single input tensor. Another approach would have been to perform late fusion, i.e. compute features on each hyperspectral channel and fuse them in a deeper layer. The late fusion approach is quite common in the multispectral deep learning research [19, 27] and effective ways to fuse the features have been found. *Jiang et al. (2019)* [27] showed that late fusion of RGB and NIR features significantly improves the performance over RGB alone for classification. We perform a similar experiment using early fusion and explore if it also generalizes to event-based data. Finally, *Singh et al. (2020)* [47] explored the role of color in CNNs by training on color data (as a human does) and evaluate on congruent (realistic), grayscale and incongruent images (false colors). This evaluation method is common in psychology tests to evaluate the importance of color for object recognition. They showed that the testing accuracy on color images is way better than for grayscale images and incongruent images [47]. However, we believe a neural network should be validated and tested on a similar data distribution to the training data, else, we cannot conclude that color is the real factor improving object recognition. To prevent this, we always fine-tune our network on the multispectral data of interest (same channels during training and testing), therefore, if a model performs better or worse, it is only explainable by the different input (spectral bands).

Multispectral event-based sensing. In their work, *Hansen and Gegenfurtner (2009,2017)* [21, 22] have proven that luminance and chromatic edges are statistically independent and therefore color provides extra information for edge detection in conventional images [21]. They additionally showed that color clearly contributes to object-contour perception by comparing the performance of luminance edges alone to luminance and color edges combined [22], using conventional images. As we know that event cameras mainly capture edges in the scene, it suggests that color could particularly benefit event-based imaging. Unfortunately, almost all the existing commercial event cameras are monochromatic (grayscale) [5, 15] except for the Color-DAVIS346 from iniVation [37, 45, 50]. *Scheerlinck et al. (2019)* presented the only color event-based dataset available, the Color Event Dataset (CED) [45], that consists of 50 minutes of footage with both color frames and color events for image reconstruction. Only a few sequences in the dataset could have been used for our face detection task but these are definitely not enough to train a face detection network from scratch. By opposition to *Tomy et al. (2020)* [52], we do not try to fuse conventional images and event data together as we believe this approach is suboptimal for object detection. When doing so, only the motion and illumination robustness properties of event cameras are kept while the very low latency and low energy consumption advantages are dropped when fusing. Very similar to our approach, *Marcireau et al. (2018)* tried to generate real RGB events using a set of beamsplitters and color filters for event-based color segmentation. In our work, we explore the face detection task instead, and use simulated events to generate significantly more RGB events, systematically labeled, and to efficiently generate multispectral event-based data, beyond the visible spectrum. To the best of our knowledge, our work is the first to introduce the concept of hyperspectral, multispectral or infrared events and explore its implications compared to multispectral conventional images, here for face detection.

3. Method

3.1. Multispectral events

The light intensity received by a pixel, $I(x, y)$, follows equation 1, where $L(\lambda)$ is the scene illumination, $R(\lambda)$ is the object reflectance and $S(\lambda)$ is the camera sensitivity for that pixel, at a particular wavelength λ. An event (x, y, t, p) is generated, at time t, whenever $\log(I(x, y))$ increases or decreases by a constant (defining the polarity p).

$$I(x,y) = \int_{0}^{\infty} R(\lambda)S(\lambda)L(\lambda)d\lambda \tag{1}$$

Most commercial event cameras are monochromatic, this means their pixel sensitivity $S(\lambda)$ spans across all visible wavelengths and often also the beginning of the near-infrared range [50]. Our idea is to play with the spectral sensitivity of the sensor to create *multispectral events*. For example, if we place an ideal long-pass filter in front of the event camera with a cut-off frequency λ_c, the new intensity equation becomes:

$$I(x,y) = \int_{\lambda_c}^{\infty} R(\lambda)S(\lambda)L(\lambda)d\lambda \tag{2}$$

, because $S(\lambda)$ would be null or negligible for $\lambda < \lambda_c$.

This concept can be generalized to other kind of filters and the sensitivity of the sensor can be shaped as it is already done on regular RGB or multispectral cameras.

In conventional imaging, multispectral (MS) images can look very different from grayscale (GS). This difference in intensity contrast certainly affects the count and distribution of events in a scene.

Hardware feasibility is further discussed in the Supplementary Material, but the already existing application of polarization filters on EVS pixels [20] or RGB Bayer pattern on EVS pixels [37, 41] would suggest that a multispectral EVS camera would be feasible, also considering that the lower sensitivity in narrower bands would be compensated by the relatively higher sensititivy of the EVS pixel.

3.2. Multispectral event-based datasets

One issue when working with event-based data is the lack of available datasets, particularly in our case, as we want to explore multispectral events while no multispectral event-based camera exists. Moreover, as we use a learning-based approach for face detection, we need a large-scale dataset, with enough samples, diversity and challenging poses, to train a robust model from scratch. Finally, if we want to fairly compare our event-based face detector to an image-based one, they should be trained on a similar domain. A scalable way of building a large event-based face detection dataset is to use **simulated events** [16]. One could use open-source event simulation software, e.g. ESIM [40] or v2e [23], but for our datasets we used a proprietary software based on the exact same principle. These simulators all rely on linearly interpolating the intensity signals to generate events, based on a generative model (e.g. [33]). In our case, we produce an event when the brightness increases or decreases by 30% (threshold).

Unfortunately, we could not find a suitable video-based face detection dataset with multispectral data (including NIR). Here is our approach:

- To train from scratch, we select open-source datasets with color videos (RGB) and convert them to events.
- For fine-tuning, we capture our own dataset with two multispectral cameras and simulate MS events.
- To validate our work with simulated events, we also capture a few sequences with real events (GS and IR).

3.2.1 Training: large-scale color (RGB) data

To build **N-YoutubeFaces** [55] and **N-MobiFace** [34], we simulate events for each color channel in the videos of the original datasets. Note that the initial frame rate of each video is at least 24 fps, which is in line with the simulator's requirements. YoutubeFaces [55] consists of 3.4k very short color videos of 1.6k different people. It perfectly suits our need of a diverse dataset for face detection in the wild. At the same time, the data is not too challenging as the sequences mainly come from movies or TV studio broadcasts, where the image standards are high. On the other hand, MobiFace [34] consists of 80 unedited mobile live color video recordings (95k frames) by smartphone users. It is challenging by design but all these challenges (fast motion, camera rotations, motion blur, scale and illumination variations etc...) are situations where event cameras excel. However, because N-MobiFace is made of simulated events from conventional images, if the source image has motion blur, the simulated events will also suffer from it. We did not try to filter out these sequences as we observed it improved the generalization capabilities of our face detector. After conversion to events, we automatically relabeled both datasets to complete for the missing ground truths, using existing face detection networks [14, 57] and Non-Maximum Suppression over all predictions. More information on the data format, the labeling and the data cleaning in the Supplementary Material.

All in all, N-Youtubefaces and N-Mobiface are the first large-scale bimodal datasets, with synchronized conventional RGB videos and their RGB simulated event-based data, labeled for face detection. We hope that these two datasets, with easy and challenging data, will help the event-based sensing community to push the capabilities of event-based cameras for face detection, tracking or recognition. It is also a nice platform to explore efficient ways to combine information from conventional cameras and event-based cameras, or to explore handheld device applications with N-MobiFace.

3.2.2 Fine-tuning: multispectral data

The goal of this paper is to explore multispectral events for deep-learning based face detection. The blue, green and red channels of N-MobiFace and N-YoutubeFaces are a good

Channel	Blue	460nm	500nm	Green	570nm	610nm	Red	700nm	Gray	NIR
APS	**0.672**	**0.670**	**0.666**	**0.667**	**0.673**	**0.670**	**0.663**	**0.672**	**0.668**	0.582
EVS	**0.589**	**0.590**	**0.593**	0.579	0.577	0.576	**0.589**	0.559	**0.592**	0.521

Table 1. Mean Average Precision (mAP) of all the single-channel face detection models, after finetuning on N-SpectralFace. Higher is better. Best model per row is in **bold** (with 1.0% tolerance). No single channel clearly outperforms others, only Near-Infrared is worse.

start but not enough to draw conclusions on multispectral events and it does not help with exploring infrared bands. Therefore, we need specifically multispectral data, enough to fine-tune our baseline model and carry our multispectral experimeents. The dataset **N-SpectralFace** consists of 69 sequences captured around the office in different places, with different faces and lighting conditions. Each sequence is recorded simultaneously with two cameras part of a setup including a beamsplitter: a SILIOS CMS-C [51] multispectral camera and a Basler dart [1] grayscale camera with a long-pass IR filter in front. In total, 10 multispectral bands are captured in parallel: 8 narrow bands in the visible, 1 grayscale band over all the visible spectrum (from 430nm to 700nm) and 1 near-infrared band (from 850nm up to around 1000nm). The data from both cameras are synced temporally and spatially through calibration so that the bands can be combined. N-SpectralFace is converted and labeled following the same approach as for N-MobiFace and N-YoutubeFaces.

3.2.3 Validation: Real multispectral events

All the event data we use for training or fine-tuning are simulated events. *How could we make sure that our trained models and the conclusions we derive are also valid for real events ?* To answer this, we captured a few sequences with real multispectral events, in the dataset **Real-SpectralFace**. One DAVIS [5] camera is mounted with an Infrared (IR) long-pass filter, the other one with an IR cut-off filter. Therefore, each sequence captures 2 event-based channels simultaneously: grayscale (visible spectrum) and near-infrared, which will be useful to validate some of our observations on multispectral events. Moreover, it is important to consider that DAVIS cameras provide simultaneous frame and event output, facilitating a direct comparison between real EVS observations and APS.

3.3. Experimental Setup

To examine the advantages of multispectral events for face detection, we use the RetinaFace architecture [13] and adapt the training for our experiments. To make annotation easier and because it does not significantly improve the mAP in the original paper, we do not use face landmarks. For simplicity and because other representations did not show better performance, we use the Binary image 2D representation of events. For each pixel, if there was an event within the last time window (50ms), we assign the value 1 to it regardless of the polarity, else we assign it 0. We have access to RGB and multispectral datasets (frames and events) but the RGB data is by far the most abundant. Thus, we first train a RetinaFace baseline model from scratch using the RGB bands from N-MobiFace and N-YoutubeFaces, then we fine-tune it on the multispectral data of interest from N-SpectralFace. We perform two main experiments: Single channel and Multi channel.

In the Single channel experiment, we compare the performance of individual channels one to another by training a model for each band and modality (APS and EVS). In the Multi channel experiment, we are interested in the effect of combining different single channels together. In particular, we select the channel combinations to evaluate two aspects: the effect of increasing the number of channels and the effect of near-infrared on the face detection performance. As we have bimodal datasets (APS and EVS), all of the 42 final models are trained on the same number of samples, in the same order, for fairness. Finally, our work can only benefit from an evaluation on real events to show that our results can generalize to a real multispectral event-based imaging sensor, if it is developed. Therefore, we also test our grayscale (GS), infrared (IR) and GS+IR models on real multispectral events from Real-SpectralFace.

4. Results

In the next subsections, we compare the face detection performance of different APS and EVS models, using mean Average Precision (mAP). Note that, by mAP, we mean Average Precision (AP) averaged over 10 IoU thresholds (0.5:0.05:0.95) as we perform single-class object detection. In the further analysis, AP at single IoU thresholds are reported (e.g. AP@.5 or AP@.75).

4.1. Single channel experiment

First, we investigate if any of the single channels is best suited for face detection, out of the 10 multispectral channels of N-SpectralFace. In table 1, we report the metrics for all 20 single channel models. Overall, APS models are performing better than EVS models. No single visible channel seems better than the others for face detection when used alone: the mAP only varies by 1 to 2% for both APS and EVS data. However, for both data modalities, the Near Infrared (NIR) channel consistently performs worse, respectively by 9% and 7% for APS and EVS.

Channel	N-SpectralFace		Real-SpectralFace	
	APS	EVS	APS	EVS
Grayscale (GS)	0.668	0.592	0.413	0.578
Infrared (IR)	0.582	0.521	**0.548**	0.535
GS+IR	**0.693**	0.626	0.473	**0.607**
Blue+Red (BR)	0.670	0.615	-	-
BR+IR	0.679	0.639	-	-
Color (BGR)	0.665	0.628	-	-
BGR+IR	**0.684**	**0.647**	-	-
BR+GS	0.672	0.626	-	-
BR+GS+IR	**0.688**	0.645	-	-
Hyperspectral (8ch)	0.664	0.645	-	-
8ch+IR	0.681	**0.656**	-	-
8ch+GS (9ch)	0.677	0.645	-	-
9ch+IR	0.671	**0.657**	-	-

Table 2. Mean Average Precision (mAP) of all the multi-channel face detection models, after finetuning on N-SpectralFace. Higher is better. Best model per column is in **bold** (with 1.0% tolerance). Multi channel combinations that include infrared are highlighted in light gray. Results on Real-SpectralFace are also reported.

4.2. Multi channel experiment

Although no single channel showed a superior performance for face detection, these channels might possibly complete each other if combined. To investigate what happens when increasing the number of channels and when fusing NIR and visible data, 12 different channel combinations are examined in table 2. Six of these combinations are visible bands only, from grayscale alone (GS) to all 9 visible bands (9ch). The other six combinations are identical visible bands but with infrared added (+IR suffix). In the first two columns of table 2, we report the validation mAP for all multi channel models.

Increasing the number of channels. In APS visible-only combinations (rows with white background), the number of channels does not seem to impact the face detection performance. Whether it is grayscale alone (GS) or all 9 visible channels together (9ch), the mAP stays within a 1% range (66 to 67%). The best APS model is only two channels (GS+IR) with only 3% better mAP than the other visible-only models. On the other hand, for EVS, there is a clear dependency on the number of channels. The mAP consistently improves with the number of channels for both the visible-only models and the models with NIR (rows with gray background), but we notice a diminishing increase: the highest increase is between 1 and 2 channels ($> 3.4\%$) and there is almost no increase anymore between 8 and 9 channels ($< 0.1\%$). In summary, EVS benefits more of combining different visible bands than APS.

Introducing infrared. *What happens when we combine visible and infrared bands?* For both APS and EVS models, adding infrared to any visible-only combination improves the face detection performance. In table 2, we intentionally alternate the visible-only models and their +IR pair for an easier read of the results. Note that for EVS models, the effect of adding infrared also seems to have diminishing returns. We observe a 3.4% increase for the EVS GS+IR model (2 channels), a 2.4% increase for the EVS BR+IR model (3 channels) and only 1.2% for the EVS 9ch+IR model (10 channels). In summary, by increasing the number of channels and including IR, APS face detection improves by only 2.5% while it increases by 6.5% for EVS. In figure 2, we show some examples of face detection on multispectral events from a single scene. Multispectral events show a greater level of details in the scene compared to the relatively sparse single-channel EVS inputs (in particular EVS IR), which in this example allow to detect more accurately the face (see EVS GS+IR or BGR) or more faces (see EVS BGR+IR). We believe this helps the EVS models filtering out false positives, better estimating object contour and make the system more robust overall. More samples in the Supplementary Material.

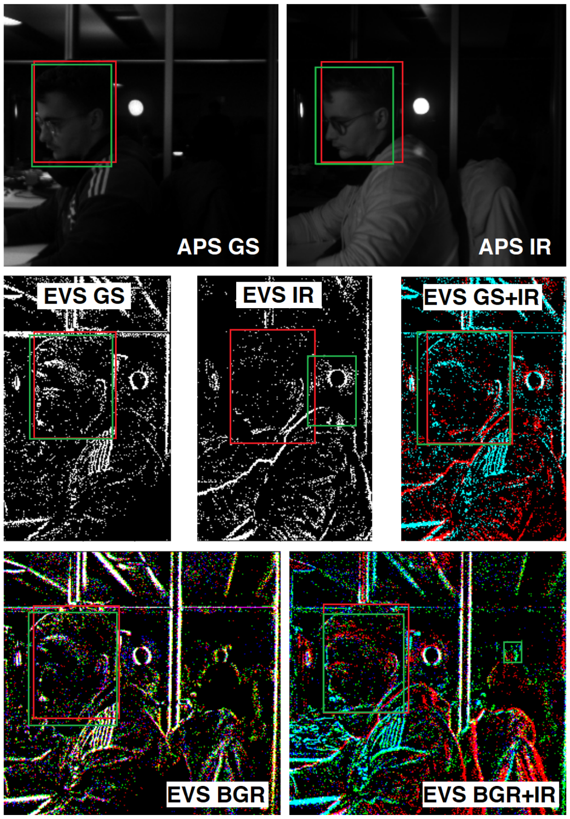
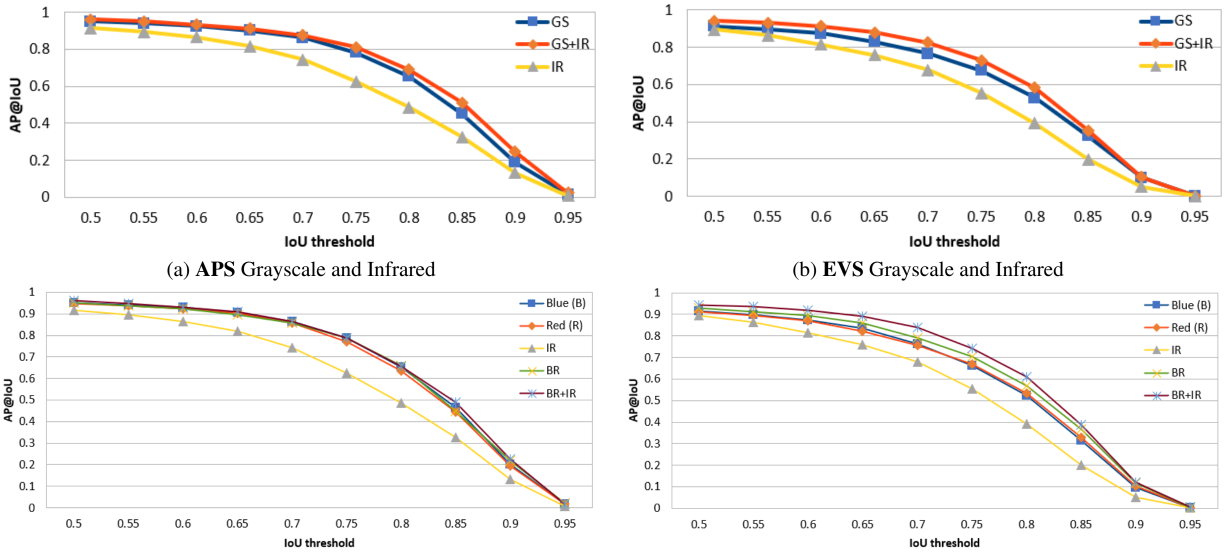


Figure 2. Multispectral APS and EVS samples from N-SpectralFace. For multispectral events, the infrared channel is always represented in red. Ground truth bounding boxes are in red, face predictions are in green. Notice the increased details in multispectral EVS samples. Best viewed in color.



(a) **APS** Grayscale and Infrared

(b) **EVS** Grayscale and Infrared

(c) **APS** Blue, Red and Infrared

(d) **EVS** Blue, Red and Infrared

Figure 3. Average Precision (AP) over IoU thresholds. Left is APS, Right is EVS. Each row is a different set of multispectal models. Multispectral EVS models show a clear AP improvement across the different thresholds, compared to APS models.

Comparison across IoU thresholds. The mean Average Precision (mAP) metric is convenient to compare multiple models' general ability to detect faces but it loses some of the performance granularity by averaging over all Intersection-over-Union (IoU) thresholds. In fig. 3, we evaluate the performance of each multispectral model over different IoU thresholds: higher IoU thresholds inform us about the ability of a model to correctly regress the corners of a face while lower thresholds only evaluate the ability to detect (or not) a face at all. The two APS plots (fig. 3a and 3c) show that increasing the number of channels or adding infrared does not have much effect on the face detection performance at lower IoU thresholds and it has a moderate effect for stricter IoU thresholds (> 0.75). On the other hand, the two EVS plots (fig. 3b and 3d) show that the curves are clearly differentiated. In particular, from smaller IoU thresholds (0.5) up to 0.85, combining visible channels together or introducing infrared substantially improve face detection for event-based models. From these plots, it seems that for APS, combining visible bands is similar to averaging the individual performance of single channels. Adding IR to visible channels (either GS or BR in the example), shows a small improvement in AP for stricter thresholds. In EVS, the behaviour is different, combining visible bands is already quite beneficial, showing a clear edge over the performance of single channels (BR vs B or R). Moreover, adding IR also clearly improves the performance of the initial model. Therefore, when going from the B or R model to BR then BR+IR, in EVS, we get two significant improvements, resulting in the 5% mAP improvement for EVS, while APS improves by less than 1%.

Improvement from combining channels. Finally, we quantify how much the models benefit from combining multispectral bands by reporting the metric difference of each multispectral input and their best individual channel. In table 3, these new metrics are reported for all APS and EVS multi channel models. The distribution of the table colors shows that APS does not benefit as much from early fusion of multispectral bands as EVS. Some APS models even suffer from the channel combination while almost all EVS models show above 3% mAP improvement over their

Channel	mAP (%)		AP@.5 (%)		AP@.75 (%)	
	APS	EVS	APS	EVS	APS	EVS
GS+IR	2.5	3.5	1.0	2.9	2.9	5.4
BR	-0.1	2.6	-0.0	1.2	-0.1	3.5
BR+IR	0.8	5.0	0.8	2.7	0.1	7.4
BGR	-0.7	3.9	0.1	1.9	-2.1	5.3
BGR+IR	1.2	5.7	0.6	3.0	1.3	9.2
BR+GS	0.1	3.4	0.7	1.8	-0.5	4.4
BR+GS+IR	1.6	5.3	1.1	2.8	0.6	8.2
8ch	-1.0	5.2	-0.7	2.8	-1.4	7.3
8ch+IR	0.7	6.3	-0.7	3.3	-0.0	9.4
9ch	0.4	5.3	-0.9	2.9	-0.6	7.3
9ch+IR	-0.3	6.4	-0.7	3.4	-1.4	9.3

Table 3. Absolute difference between multi channel models and their best single channel model. The cells are colored based on four different ranges. Negative numbers, i.e. when the channel combination performs worse than its best single channel, are in red. Improvements up to 2% are in orange, between 2% and 5% are in light green, and above 5% are in dark green. Best viewed in color. EVS multispectral models show a clear face detection improvement over their best single channel, compared to APS.

best single channel. For event-based data, models with infrared show a bigger gap with their best indiviudal channel: the ”+IR” models show an mAP improvement 1.5 to 2 times better than their ”visible only” pair, with diminishing gap when increasing the number of channels. Overall, the best improvement over single channels is achieved by the EVS 8ch+IR, 9ch+IR and BGR+IR models for the AP@.75 metric with more than 9% improvement. With table 3, we confirm that EVS not only benefits from a multispectral input, but the improvement is also substantially higher than for APS. Combining both visible and infrared EVS bands always further improves the face detection performance.

4.3. Real multispectral events

To confirm that the multi-channel experiments' results are not only valid for simulated events, the GS, IR and GS+IR models are validated on real multispectral events from Real-SpectralFace. The results are reported in the last column in table 2. One can observe that the EVS mAP is of similar magnitude for both N-SpectralFace and Real-SpectralFace data, in the 50 to 60% range, so the models are still able to detect faces on real events even though trained on simulation. Moreover, EVS models' mAP trend for Real-SpectralFace is the same as for N-SpectralFace. The EVS infrared model is worse than the EVS grayscale one but the combination of both (GS+IR) has the best face detection performance. For the APS models, the face detection performance in Real-SpectralFace is generally lower and the mAP trend is different (the best model is now IR). This can be explained by Real-SpectralFace APS data being captured with a DAVIS event camera [5], which has a lower resolution, lower dynamic range and more noise.

5. Discussion

Since no multispectral event cameras currently exist, the experiments have been carried mainly with simulated events, allowing us to compare APS and EVS fairly. The behaviour of the multispectral EVS models is always compared to multispectral APS models to draw conclusions. It is quite intuitive that color or multispectral data would improve the face detection performance. However, the main contribution of this paper is showing that the EVS improvement from combining channels is substantially better than for conventional imaging sensors. In other words, if we note $I(x)$ as the useful information for face detection contained in an input x, our experiments show that:

$$I(MS_{EVS}) - I(GS_{EVS}) > I(MS_{APS}) - I(GS_{APS}) > 0 \tag{3}$$

i.e., not only we show that event-based sensing, as conventional imaging, can improve by using multispectral data (more information is captured, the difference for both modalities is > 0), but we show that the gain in information from using multispectral bands is greater for EVS.

One could argue that a fair comparison of APS and EVS models would actually be to choose a state-of-the-art (SOTA) neural network for both image-based and event-based data, instead of applying the RetinaFace [13] architecture to both. The main addition of SOTA models for event-based object/face detection is the more clever use of temporal information through recurrent neural networks or visual transformers, e.g. in [18, 39]. We argue that the temporal information is independant from the color information and therefore our results could generalize to new SOTA models. Moreover, showing that the event-based face detection performance improves on a ”sub-optimal” neural network (designed for APS) demonstrates that multispectral events have an inherent substantial advantage over monochromatic events. We believe our result is stronger than if we had found it through the design of a specific neural architecture to efficiently fuse multispectral features. This is in line with our decision to only explore early fusion of the multispectral data. Note that our conclusions are consistent with the works from *Hansen and Gegenfurtner (2009,2017)* [21, 22] on object-contour perception in color images. Event-based sensing mainly captures edges in the scene and it could justify why EVS particularly benefits from the addition of color or infrared bands. For this reason, we are curious to know if our observations generalize to other tasks than face detection, and we encourage the research community to reproduce our results and extend them to other computer vision tasks, when possible.

6. Conclusion

In conclusion, this work demonstrates that event cameras substantially benefit from multispectral sensing. Combining events captured in different spectral bands improves the face detection performance by a significantly larger margin than it does for conventional multispectral images. These results point toward a better efficiency of event-based data to store color information and use it effectively. The inherent sparsity of event-based data and the fact that events capture intensity differences in the scene could be factors that explain this phenomenon. To obtain these results, we built and shared two large-scale image-based and event-based face detection datasets with color (RGB) data and a third dataset with multispectral APS and EVS data for fine-tuning. We took advantage of event simulation to perform experiments on a hypothetical sensor and presented explorative comparisons that could motivate the development of a prototype. To the best of our knowledge, our research is the first to tackle the evaluation of multispectral event-based sensing, especially in the infrared.

References

[1] Basler AG. Basler dart – area scan cameras, 2023. Available at https://www.baslerweb.com/en/products/cameras/area-scan-cameras/dart/. Accessed on 28.06.2023. 5

[2] Elli Angelopoulou. The reflectance spectrum of human skin. Technical report, University of Pennsylvania, December 1999. 2

[3] Nicolas Audebert, Bertrand Le Saux, and Sebastien Lefevre. Deep learning for classification of hyperspectral data: A comparative review. *IEEE Geoscience and Remote Sensing Magazine*, 7(2):159–173, 2019. 3

[4] Souptik Barua, Yoshitaka Miyatani, and Ashok Veeraraghavan. Direct face detection and video reconstruction from event cameras. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–9, 2016. 2

[5] Christian Brändli, Raphael Berner, Minhao Yang, Shih-Chii Liu, and Tobi Delbruck. A 240 × 180 130 db 3 µs latency global shutter spatiotemporal vision sensor. *Solid-State Circuits, IEEE Journal of*, 49:2333–2341, 10 2014. 2, 3, 5, 8

[6] Marco Cannici, Marco Ciccone, Andrea Romanoni, and Matteo Matteucci. Event-based convolutional networks for object detection in neuromorphic cameras. *CoRR*, abs/1805.07931, 2018. 2

[7] Zhicheng Cao, Heng Zhao, Shufen Cao, and Liaojun Pang. Face detection in the darkness using infrared imaging: a deep-learning-based study. page 49, 08 2021. 3

[8] Nicholas F. Y. Chen. Pseudo-labels for supervised learning on dynamic vision sensor data, applied to object detection under ego-motion, 2018. 2

[9] Yushi Chen, Zhouhan Lin, Xing Zhao, Gang Wang, and Yanfeng Gu. Deep learning-based classification of hyperspectral data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 7(6):2094–2107, 2014. 3

[10] Yi-Ting Chou and Peter Bajcsy. Toward face detection, pose estimation and human recognition from hyperspectral imagery. Technical Report NCSA-ALG04-0005, University of Illinois at Urbana-Champaign, October 2004. 3

[11] Catherine Cooksey, Benjamin Tsai, and David Allen. Spectral reflectance variability of skin and attributing factors. (9461), 2015-05-21 2015. 2

[12] Joubert Damien, Konik Hubert, and Chausse Frederic. Convolutional neural network for detection and classification with event-based data. *Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, 2019. 2

[13] Jiankang Deng, Jia Guo, Yuxiang Zhou, Jinke Yu, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-stage dense face localisation in the wild. *CoRR*, abs/1905.00641, 2019. 3, 5, 8

[14] Glenn Jocher et. al. ultralytics/yolov5: v6.0 - YOLOv5n 'Nano' models, Roboflow integration, TensorFlow export, OpenCV DNN support, Oct. 2021. 4

[15] Guillermo Gallego, Tobi Delbrück, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew J. Davison, Jörg Conradt, Kostas Daniilidis, and Davide Scaramuzza. Event-based vision: A survey. *CoRR*, abs/1904.08405, 2019. 1, 2, 3

[16] Daniel Gehrig, Mathias Gehrig, Javier Hidalgo-Carrió, and Davide Scaramuzza. Video to events: Bringing modern computer vision closer to event cameras. *CoRR*, abs/1912.03095, 2019. 4

[17] Daniel Gehrig, Antonio Loquercio, Konstantinos G. Derpanis, and Davide Scaramuzza. End-to-end learning of representations for asynchronous event-based data. *CoRR*, abs/1904.08245, 2019. 2

[18] Mathias Gehrig and Davide Scaramuzza. Recurrent vision transformers for object detection with event cameras, 2023. 2, 8

[19] Shreyank N. Gowda and Chun Yuan. Colornet: Investigating the importance of color spaces for image classification. *CoRR*, abs/1902.00267, 2019. 3

[20] Germain Haessig, Damien Joubert, Justin Haque, Moritz B. Milde, Tobi Delbruck, and Viktor Gruev. Pdavis: Bio-inspired polarization event camera. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3963–3972, 2023. 4

[21] THORSTEN HANSEN and KARL R. GEGENFURTNER. Independence of color and luminance edges in natural scenes. *Visual Neuroscience*, 26(1):35–49, Jan. 2009. 3, 8

[22] Thorsten Hansen and Karl R. Gegenfurtner. Color contributes to object-contour perception in natural scenes. *Journal of Vision*, 17(3):14, Mar. 2017. 3, 8

[23] Yuhuang Hu, Shih-Chii Liu, and Tobi Delbrück. V2E: from video frames to realistic DVS event camera streams. *CoRR*, abs/2006.07722, 2020. 4

[24] Massimiliano Iacono, Stefan Weber, Arren Glover, and Chiara Bartolozzi. Towards event-driven object detection with off-the-shelf deep learning. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1–9, 2018. 2

[25] Garima Jaiswal, Arun Sharma, and Sumit Yadav. Critical insights into modern hyperspectral image applications through deep learning. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 11, 11 2021. 3

[26] Weiheng Liu Dongqing Zou Qiang Wang Paul-K.J. Park Jia Li, Feng Shi and Hyunsurk Eric Ryu. Adaptive temporal pooling for object detection using dynamic vision sensor. In Gabriel Brostow Tae-Kyun Kim, Stefanos Zafeiriou and Krystian Mikolajczyk, editors, *Proceedings of the British Machine Vision Conference (BMVC)*, pages 40.1–40.12. BMVA Press, September 2017. 2

[27] Jionghui Jiang, Xi'an Feng, Fen Liu, Yingying Xu, and Hui Huang. Multi-spectral rgb-nir image classification using double-channel cnn. *IEEE Access*, 7:20607–20613, 2019. 3

[28] Xavier Lagorce, Garrick Orchard, Francesco Galluppi, Bertram E. Shi, and Ryad B. Benosman. Hots: A hierarchy of event-based time-surfaces for pattern recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(7):1346–1359, 2017. 2

[29] Gregor Lenz, Sio-Hoi Ieng, and Ryad Benosman. Event-based face detection and tracking using the dynamics of eye blinks. *Frontiers in Neuroscience*, 14, 2020. 2

[30] Jianing Li, Jia Li, Lin Zhu, Xijie Xiang, Tiejun Huang, and Yonghong Tian. Asynchronous spatio-temporal memory network for continuous event-based object detection. *IEEE Transactions on Image Processing*, 31:2975–2987, 2022. 2

[31] Shutao Li, Weiwei Song, Leyuan Fang, Yushi Chen, Pedram Ghamisi, and Jón Atli Benediktsson. Deep learning for hyperspectral image classification: An overview. *IEEE Transactions on Geoscience and Remote Sensing*, 57(9):6690–6709, 2019. 3

[32] Stan Z. Li, Rufeng Chu, Shengcai Liao, and Lun Zhang. Illumination invariant face recognition using near-infrared images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(4):627–639, 2007. 3

[33] Patrick Lichtsteiner, Christoph Posch, and Tobi Delbruck. A 128 $\times$ 128120db15 μ s latency asynchronous temporal contrast vision sensor. *IEEE journal of solid-state circuits*, 43(2):566–576, 2008. 1, 4

[34] Yiming Lin, Shiyang Cheng, Jie Shen, and Maja Pantic. Mobiface: A novel dataset for mobile face tracking in the wild, 2019. 2, 4

[35] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott E. Reed, Cheng-Yang Fu, and Alexander C. Berg. SSD: single shot multibox detector. *CoRR*, abs/1512.02325, 2015. 2

[36] Michael J. Mendenhall, Abel S. Nunez, and Richard K. Martin. Human skin detection in the visible and near infrared. *Appl. Opt.*, 54(35):10559–10570, Dec 2015. 2

[37] Diederik Paul Moeys, Chenghan Li, Julien N.P. Martel, Simeon Bamford, Luca Longinotti, Vasyl Motsnyi, David San Segundo Bello, and Tobi Delbruck. Color temporal contrast sensitivity in dynamic vision sensors. In *2017 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 1–4, 2017. 3, 4

[38] E. M. Nowara, T. K. Marks, H. Mansour, and A. Veeraraghavan. Near-infrared imaging photoplethysmography during driving. *IEEE Transactions on Intelligent Transportation Systems*, pages 1–12, 2020. 3

[39] Etienne Perot, Pierre de Tournemire, Davide Nitti, Jonathan Masci, and Amos Sironi. Learning to detect objects with a 1 megapixel event camera. *CoRR*, abs/2009.13436, 2020. 2, 8

[40] Henri Rebecq, Daniel Gehrig, and Davide Scaramuzza. Esim: an open event camera simulator. 10 2018. 4

[41] Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. High speed and high dynamic range video with an event camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(6):1964–1980, 2021. 4

[42] Joseph Redmon, Santosh Kumar Divvala, Ross B. Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. *CoRR*, abs/1506.02640, 2015. 2

[43] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks, 2016. 2

[44] Cian Ryan, Brian O'Sullivan, Amr Elrasad, Joe Lemley, Paul Kielty, Christoph Posch, and Etienne Perot. Real-time face & eye tracking and blink detection using event cameras. *CoRR*, abs/2010.08278, 2020. 2

[45] Cedric Scheerlinck, Henri Rebecq, Timo Stoffregen, Nick Barnes, Robert E. Mahony, and Davide Scaramuzza. CED: color event camera dataset. *CoRR*, abs/1904.10772, 2019. 3

[46] Rafael Serrano-Gotarredona, Teresa Serrano-Gotarredona, Antonio Acosta-Jimenez, and Bernab Linares-Barranco. A neuromorphic cortical-layer microchip for spike-based event processing vision systems. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 53(12):2548–2566, 2006. 1

[47] Aditya Singh, Alessandro Bay, and Andrea Mirabile. Assessing the importance of colours for cnns in object recognition. *CoRR*, abs/2012.06917, 2020. 3

[48] Amos Sironi, Manuele Brambilla, Nicolas Bourdis, Xavier Lagorce, and Ryad Benosman. HATS: histograms of averaged time surfaces for robust event-based object classification. *CoRR*, abs/1803.07913, 2018. 2

[49] Diego Socolinsky. *Multispectral Face Recognition*, pages 293–313. 10 2007. 3

[50] Gemma Taverni, Diederik Paul Moeys, Chenghan Li, Celso Cavaco, Vasyl Motsnyi, David San Segundo Bello, and Tobi Delbruck. Front and back illuminated dynamic and active pixel vision sensors comparison. *IEEE Transactions on Circuits and Systems II: Express Briefs*, 65(5):677–681, 2018. 3, 4

[51] Silios Technologies. Cms series: Multispectral cameras. Available at `https://www.silios.com/cms-series`. Accessed on 28.06.2023. 5

[52] Abhishek Tomy, Anshul Paigwar, Khushdeep S. Mann, Alessandro Renzaglia, and Christian Laugier. Fusing event-based and rgb camera for robust object detection in adverse conditions. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 933–939, 2022. 3

[53] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 1, pages I–I, 2001. 2

[54] Jixiang Wan, Ming Xia, Zunkai Huang, Li Tian, Xiaoying Zheng, Victor Chang, Yongxin Zhu, and Hui Wang. Event-based pedestrian detection using dynamic vision sensors. *Electronics*, 10(8), 2021. 2

[55] Lior Wolf, Tal Hassner, and Itay Maoz. Face recognition in unconstrained videos with matched background similarity. In *CVPR 2011*, pages 529–534, 2011. 2, 4

[56] Chao Yan and Yuanqing Wang. A novel multi-user face detection under infrared illumination by real adaboost. In *2009 International Conference on Computational Intelligence and Software Engineering*, pages 1–6, 2009. 3

[57] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multi-task cascaded convolutional networks. *CoRR*, abs/1604.02878, 2016. 4

[58] Zhiwei Zhang, Dong Yi, Zhen Lei, and Stan Z. Li. Regularized transfer boosting for face detection across spectrum. *IEEE Signal Processing Letters*, 19(3):131–134, 2012. 3