



Aalborg Universitet

AALBORG UNIVERSITY  
DENMARK

## Seasons in Drift: A Long-Term Thermal Imaging Dataset for Studying Concept Drift

Nikolov, Ivan Adriyanov; Philipsen, Mark Philip; Liu, Jinsong; Dueholm, Jacob Velling; Johansen, Anders Skaarup; Nasrollahi, Kamal; Moeslund, Thomas B.

*Published in:*

Thirty-fifth Conference on Neural Information Processing Systems 2021

*Publication date:*

2021

*Document Version*

Accepted author manuscript, peer reviewed version

[Link to publication from Aalborg University](#)

*Citation for published version (APA):*

Nikolov, I. A., Philipsen, M. P., Liu, J., Dueholm, J. V., Johansen, A. S., Nasrollahi, K., & Moeslund, T. B. (2021). Seasons in Drift: A Long-Term Thermal Imaging Dataset for Studying Concept Drift. In *Thirty-fifth Conference on Neural Information Processing Systems 2021* Neural Information Processing Systems Foundation.  
<https://openreview.net/forum?id=LjjqegBNTPi>

### General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal -

### Take down policy

If you believe that this document breaches copyright please contact us at [vbn@aub.aau.dk](mailto:vbn@aub.aau.dk) providing details, and we will remove access to the work immediately and investigate your claim.

---

# Seasons in Drift: A Long-Term Thermal Imaging Dataset for Studying Concept Drift

---

Ivan Nikolov<sup>1</sup>, Mark P. Philipsen<sup>1</sup>, Jinsong Liu<sup>1</sup>, Jacob V. Dueholm<sup>1</sup>, Anders S. Johansen<sup>1</sup>, Kamal Nasrollahi<sup>1,2</sup>, and Thomas B. Moeslund<sup>1</sup>

<sup>1</sup> Visual Analysis and Perception Lab, Aalborg University, Aalborg, Denmark

<sup>2</sup> Research, Milestone Systems, Brøndby, Denmark

## Abstract

1 The time dimension of datasets and long-term performance of machine learning  
2 models have received little attention. With extended deployments in the wild,  
3 models are bound to encounter novel scenarios and concept drift that cannot be  
4 accounted for during development and training. In order for long-term patterns and  
5 cycles to appear in datasets, the datasets must cover long periods of time. Since  
6 this is rarely the case, it is difficult to explore how computer vision algorithms cope  
7 with changes in data distribution occurring across long-term cycles such as seasons.  
8 Video surveillance is an application area clearly affected by concept drift. For this  
9 reason we publish the Long-term Thermal Drift (LTD) dataset. LTD consists of  
10 thermal surveillance imaging from a single location across 8 months. Along with  
11 thermal images we provide relevant metadata such as weather, the day/night cycle  
12 and scene activity. In this paper we use the metadata for in-depth analysis of the  
13 causal and correlational relationships between environmental variables and the  
14 performance of selected computer vision algorithms used for anomaly and object  
15 detection. Long-term performance is shown to be most correlated with temperature,  
16 humidity, the day/night cycle and scene activity level. This suggests that the  
17 coverage of these variables should be prioritised when building datasets for similar  
18 applications. As a baseline, we propose to mitigate the impact of concept drift by  
19 first detecting points in time where drift occurs. At this point we collect additional  
20 data that is used to retraining the models. This improves later performance by an  
21 average of 25% across all tested algorithms.

## 22 1 Introduction

23 Once computer vision algorithms step outside the lab and are deployed in real-life outdoor applica-  
24 tions, their performance tends to drop significantly due to conditions changing over time, i.e. concept  
25 drift [90, 24, 85]. Concept drift can materialize as gradual, recurring or sudden changes in the visual  
26 representation of the scene. Existing datasets, in general, favour coverage of multiple locations  
27 [32, 75] for short periods of time [46, 45, 83]. Such datasets are ill suited for exploring long-term  
28 effects such as concept drift and algorithms developed on their basis are unlikely to show robustness  
29 to long-term phenomena. Research studying concept drift [28, 55], uses synthetic datasets or datasets  
30 augmented in order to introduce drift. This does not necessarily completely represent real-world  
31 concept drift.

32 Our work presents a novel real-world dataset covering the 8 months from January to August. This  
33 time span means that the dataset encompasses a wide range of weather conditions, human activity,  
34 seasonal transitions, and recurring cycles such as weekdays, weekends, mornings and evenings.

35 Along with the thermal images, timestamped metadata has been gathered. The metadata includes  
36 weather data such as temperature, humidity, precipitation, etc. as well as metrics for scene activity  
37 level. We use the dataset to study concept drift by exploring contributing factors and demonstrating  
38 their effects on algorithmic performance. By publishing the dataset, we seek to aid the community  
39 in evaluating existing algorithms against a long-term benchmark and in the development of algorithms  
40 that show greater robustness to long-term phenomena.

41 To explore the dataset, two common tasks are chosen, namely anomaly and people detection. These  
42 tasks tend to suffer strong performance degradation when exposed to long-term concept drift [77].  
43 Object detection in general or detecting people in particular is a fundamental task involved in  
44 many use cases such as autonomous driving [86, 10, 8], tracking [6, 67, 73, 19] and re-identification  
45 [40, 41, 26]. Common for many of the use cases is the application of object detection in unconstrained  
46 environments and across long spans of time. Anomaly detection, where the goal is to detect unusual  
47 behavioral patterns, is another task that is exposed to concept drift. These algorithms must be able to  
48 distinguish irrelevant changes due to e.g. concept drift from emergencies such as burglaries or assaults  
49 [75], car accidents [39], loitering and suspicious behaviour [89], indoor [27] and outdoor [15, 36, 43]  
50 falls.

51 We select representative algorithms for each task and evaluate their performance across time and in  
52 relation to environmental factors. As expected, all models exhibit performance degradation, as the  
53 test data diverges from the training set. Temperature and humidity proves to influence the models  
54 the most, followed by the change between day and night and the activity level of the scene. On the  
55 other hand, variation in precipitation and wind do not influence the performance of the models. In  
56 general, methods that learn from solving tasks that consider the entirety of the image are likely to be  
57 less impacted by drift, compared to methods that consider small regions or individual pixels [76].  
58 An example could be object detectors vs. autoencoders, where something like brightness is likely to  
59 impact the autoencoder’s reconstruction significantly, but won’t effect the class or position of objects.  
60 By including both autoencoders and object detectors we ensure that both ends of this spectrum are  
61 covered in our analysis.

62 Finally, a baseline algorithm is presented to reduce the consequences of concept drift. This algorithm  
63 provides additional training data from points in time where concept drift is detected. This baseline  
64 is intended to encourage researchers to develop other methods of reducing the impact of concept  
65 drift. We believe that our findings on this novel dataset generalize to other environments and use  
66 cases, as well as other modalities and therefore will be an example to follow for future definition and  
67 collection of datasets. This in turn will help the community getting closer to deploying long-term  
68 computer vision algorithms for real-life outdoor applications. The main contributions of this paper  
69 can be summarized as follows:

- 70 • The Long-term Thermal Drift (LTD) dataset - the longest-spanning systematically collected  
71 thermal dataset comprised of 8 months of video data, containing both timestamp and weather  
72 condition metadata;
- 73 • In-depth analysis of the correlational and causal relationships between the performance of  
74 models and environmental factors;
- 75 • A baseline algorithm for reducing the effects of concept drift.

## 76 **2 Related Work**

### 77 **2.1 Concept Drift Detection**

78 As many systems need to be deployed and work stably for long periods of time and with input data  
79 which can change both gradually and suddenly, the presence of drift and ways to deal with it is a  
80 topic that has been widely studied. In computer vision it is normally studied by either focusing on  
81 specific real-world use cases or synthetically augmenting existing datasets. Real-world cases can be  
82 taken from egocentric video [53] or industrial inspection [52]. These cases present both examples  
83 of the problem and detection methods, but have limited use outside of the specific environments.  
84 Augmented versions of popular datasets such as MNIST and CIFAR can also be used. The works by  
85 [55] and [61] focus on methods for detecting data shifts using differences between the training and  
86 testing data, utilizing dimensionality reduction and statistical tests like Maximum Mean Discrepancy  
87 and Kolmogorov-Smirnov test. The benefit of using synthetically augmented data for testing is that

Table 1: Existing urban computer vision stationary and changing location datasets. The *Location* can be either changing denoting moving camera like the ones on self-driving cars or stationary like on surveillance cameras. The *Type* of the datasets can be either RGB, thermal or LiDAR, the *Duration* is the size of the dataset in hours, the *Period* is the capturing time span and the *Metadata* is any additional information

| Name                     | Year        | Location          | Type           | Duration [hours] | Period          | Metadata                             |
|--------------------------|-------------|-------------------|----------------|------------------|-----------------|--------------------------------------|
| KAIST [32]               | 2015        | Changing          | RGB/Thermal    | 43.41            | -               | -                                    |
| CVC-14 [20]              | 2016        | Changing          | RGB/Thermal    | 11.8             | -               | -                                    |
| Oxford RobotCar [48]     | 2017        | Changing          | RGB/LiDAR      | -                | 1 year          | GPS, IMU, Day/Night, Weather         |
| Aachen Day-Night [70]    | 2018        | Changing          | RGB            | -                | -               | GPS, Day/Night, Weather              |
| Gated2Depth [23]         | 2019        | Changing          | RGB/LiDAR      | -                | -               | GPS, IMU, Day/Night, Weather         |
| Dark Zurich [68]         | 2019        | Changing          | RGB            | -                | -               | GPS, Day/Night                       |
| ACDC [69]                | 2020        | Changing          | RGB            | -                | several days    | GPS, Weather                         |
| Ford AV [1]              | 2020        | Changing          | RGB/LiDAR      | -                | 1 year          | GPS, IMU Day/Night, Weather, Time    |
| Bdd100k [87]             | 2020        | Changing          | RGB            | -                | -               | Weather, Time                        |
| UCSD [49]                | 2010        | Stationary        | RGB            | 3.1              | -               | -                                    |
| Caltech Pedestrian [13]  | 2011        | Stationary        | RGB            | 10               | -               | -                                    |
| VIRAT [54]               | 2011        | Stationary        | RGB            | 29               | -               | -                                    |
| Avenue [46]              | 2013        | Stationary        | RGB            | 0.5              | -               | -                                    |
| ShanghaiTech Campus [45] | 2018        | Stationary        | RGB            | 3.6              | -               | -                                    |
| Surveillance Videos [75] | 2018        | Stationary        | RGB            | 128              | -               | -                                    |
| Street Scene [62]        | 2020        | Stationary        | RGB            | 4                | 2 summers       | -                                    |
| ADOC [60]                | 2020        | Stationary        | RGB            | 24               | 1 day           | -                                    |
| AU-AIR [5]               | 2020        | Stationary        | RGB            | 2                | -               | Time, Positions                      |
| MEVA [12]                | 2021        | Stationary        | RGB/Thermal    | 144              | 3 weeks         | GPS, Time                            |
| <b>LTD (Our)</b>         | <b>2021</b> | <b>Stationary</b> | <b>Thermal</b> | <b>298</b>       | <b>8 months</b> | <b>GPS, Day/Night, Weather, Time</b> |

88 different types of shifts can easily be simulated - from gradual drift to adversarial attacks [28]. But  
 89 these simulated shifts do not always correspond to real-world ones. Some more robust methods also  
 90 exist [77], aimed at using real-world drift in wider variety of use cases. The need for more research  
 91 into concept drift, paired with a long-term real-world dataset is evident, as the effects from it can  
 92 limit long term deployment of vision systems [72, 2].

## 93 2.2 Datasets

94 We can separate previous work roughly in two types of use cases - datasets that contain a scenes from  
 95 a stationary location, like the ones captured from CCTV and surveillance cameras and datasets with  
 96 constantly changing locations, like the ones specifically directed towards autonomous cars, robots  
 97 and human egocentric footage. The two types of datasets are used for different tasks, like vehicle and  
 98 pedestrian detection and environmental segmentation for changing datasets [32, 87, 1] and pedestrian  
 99 tracking and anomaly detection for stationary ones [45, 62, 12]. The changing datasets also benefit  
 100 from more diverse data coming from different sensors, compared to more image based stationary  
 101 datasets. Our proposed LTD dataset is directed towards advancing the state-of-the-art in stationary  
 102 location outdoor urban datasets by providing a longer duration, larger variation and rich metadata. A  
 103 comparison in Table 1 shows how the dataset stacks against previous work.

104 Datasets used for autonomous driving with changing locations [87, 70, 23, 1], which contain multiple  
 105 modalities like LiDARs, RGB, depth cameras, as well as GPS and IMU data. They also contain  
 106 data with longer duration from multiple days [69] to a whole year [48]. These datasets also focus  
 107 on presenting adverse weather conditions, which can be used for domain adaptation and making  
 108 autonomous driving and robotics application more robust [68, 1, 69]. Thermal datasets are less  
 109 prevalent but still widely used [17, 20]. These moving location car datasets normally do not contain  
 110 explicit information of their duration, as they are captured from many cars and the data is sampled.

111 On the other hand stationary location datasets do not contain any information about the period over  
 112 which they were collected. This combined with the relative short duration of many of the widely  
 113 used datasets ([49, 45, 13, 44]) makes it impossible for them to be used for studying long-term  
 114 effects on deployed machine learning solutions. The duration of some of these datasets is taken from  
 115 the research presented in [60]. Some larger datasets are gathered from internet videos [75], which  
 116 lack the needed continuity for testing gradual concept drift in the data. More recent datasets have  
 117 been produced with the goal to capture larger variations in the environments [12, 60], but with a  
 118 limited scope. The lack of metadata is another problem, limiting the study of factors causing concept

119 drift, as only some of the investigated datasets provide insufficient metadata [5, 12, 66]. Most of  
 120 the investigated datasets focus on RGB data, with only some containing both RGB and thermal  
 121 data [32, 12]. However, thermal imaging is better at preserving preserving people’s anonymity as it  
 122 does not capture facial and body detail. This removes the need for post-processing like blurring or  
 123 pixelating faces to protect personal data [88, 47, 37], which is a crucial requirement for complying  
 124 with the European general data protection regulations (GDPR).The thermal imaging market has seen  
 125 significant growth [14] and is forecast to expand even more in the following years [65, 34], which  
 126 makes it necessary for long-term public thermal datasets to be easily accessible

### 127 3 The Long-term Thermal Drift (LTD) Dataset

128 To address the gaps seen in the stationary surveillance state-of-the-art and to leverage the need for  
 129 more thermal data, a new dataset is proposed. It consists of thermal videos with resolution  $288 \times 384$   
 130 captured through the period of **8 months** using a Hikvision DS-2TD2235D-25/50 thermal camera  
 131 [30]. The camera is a long wavelength infrared (LWIR) unit, capturing wavelengths between 8 and 14  
 132  $\mu m$ . Raw data is captured through the day and saved in a mp4 format as 8-bit uncalibrated grayscale  
 133 videos. A pre-processing algorithm is then run through the data. It first cuts the raw files into days  
 134 starting from 00 : 00 and separates them into folders. Each folder is timestamped with the year,  
 135 month and day timestamp. The videos for each day are then cut into **2-minute** clips selected every  
 136 30 minutes through the day, for a total of **298 hours**. These videos are additionally timestamped with  
 137 hour and minute timestamp. The starting point of the data is May 2020 until September 2020, together  
 138 with a second part from January 2021, up until May 2021. This gives the data a large weather variation  
 139 through the winter, spring and summer seasons. The images were taken on the harbor front in Aalborg,  
 140 Denmark. The approximate longitude and latitude coordinates are given as (9.9217, 57.0488).  
 141 We provide the dataset - <https://www.kaggle.com/ivannikolov/longterm-thermal-drift-dataset>,  
 142 together with the code to extract the necessary data and to reproduce the experimental pipeline  
 143 <https://github.com/IvanNik17/Seasonal-Changes-in-Thermal-Surveillance-Imaging>.

144 Some examples of seasonal and day and night variation of the captured data, together with weather  
 145 and human activity variation can be seen in Figure 1. These large variations, together with a total size  
 146 almost twice as large as other datasets in Section 2.2, allows for studying the effects of concept drift  
 147 on trained models.

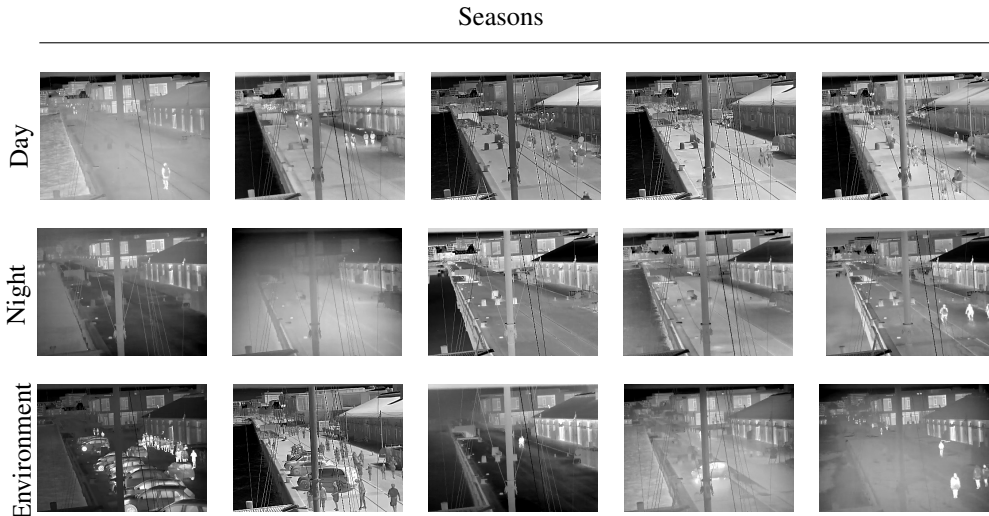


Figure 1: Examples of extreme changes in the image data contained in the proposed dataset. From left to right the day and night rows show example changes from data of February, March, April, June and August. The third row shows changes based on weather conditions and human activity.

148 Figure 1 depicts issues stemming from the natural thermal data concept drift, such as grayscale  
 149 inversion in the background and people in different seasons, view limitation and reflections caused by  
 150 weather like fog, rain, snow, view cluttering from multiple people and vehicles.

Table 2: Average metadata for each month. From left - temperature, humidity, precipitation, dew point, wind direction, wind speed, sun radiation and minutes of sunshine in a 10-minute interval.

|      | Temp.<br>[°C] | Hum.<br>[%] | Precip.<br>[kg/m <sup>2</sup> ] | Dew P.<br>[°C] | Wind Dir.<br>[degrees] | Wind Sp.<br>[m/s] | Sun Rad.<br>[W/m <sup>2</sup> ] | Sun<br>[min] |
|------|---------------|-------------|---------------------------------|----------------|------------------------|-------------------|---------------------------------|--------------|
| Jan. | -0.48         | 90.10       | 0.01                            | -1.96          | 161.91                 | 2.58              | 23.97                           | 0.90         |
| Feb. | -0.54         | 85.15       | 0.01                            | -2.83          | 131.00                 | 2.95              | 51.12                           | 1.42         |
| Mar. | 3.75          | 83.61       | 0.01                            | 0.93           | 218.80                 | 3.58              | 99.35                           | 1.85         |
| Apr. | 4.47          | 97.25       | 0.13                            | 4.10           | 126.50                 | 2.97              | 67.31                           | 2.23         |
| May  | 10.74         | 75.46       | 0.01                            | 6.07           | 217.32                 | 3.04              | 256.76                          | 3.66         |
| June | 16.36         | 71.46       | 0.01                            | 10.57          | 151.27                 | 2.37              | 256.46                          | 3.63         |
| July | 12.91         | 75.32       | 0.01                            | 8.46           | 268.15                 | 3.97              | 270.17                          | 3.62         |
| Aug. | 16.93         | 79.17       | 0.02                            | 12.69          | 163.18                 | 2.08              | 197.86                          | 3.15         |

### 151 3.1 Metadata Analysis

152 Besides video data we also provide metadata in the form of weather data, gathered using the open  
 153 source Danish Meteorological Institute (DMI) weather API [33] in 10-minute intervals. The selected  
 154 properties are - temperature, measured in [°C], relative humidity percentage measured 2m over  
 155 terrain, accumulated precipitation in [kg/m<sup>2</sup>], dew point temperature in [°C] measured 2m over  
 156 terrain, wind direction in degrees orientation, wind speed in [m/s], both measured 10m over terrain,  
 157 mean sun radiation in [W/m<sup>2</sup>] and minutes of sunshine in the measured interval. These properties  
 158 are selected, as it is speculated that they would be useful to explain changes in the captured image  
 159 data. An overview of the average weather metadata measurements of the dataset can be seen in Table  
 160 2. Temperature and relative humidity have been shown to affect thermal cameras, when detecting  
 161 surface defects in concrete structures [80], measuring skin temperature changes on athletes [35],  
 162 getting accurate readings for volcanology [3] and inspecting food [21]. Precipitation and dew point  
 163 temperature can indicate the presence of rain, fog or high moisture and condensation. These can  
 164 increase attenuation of infrared light and change the produced camera response [4, 11]. The build-up  
 165 of moisture can create puddles in the images, which would change the scene reflectivity and reflected  
 166 temperature [7]. The sun radiation and amount of sunshine can affect the captured images by rapidly  
 167 changing the intensity of the infrared light. Finally wind speed and direction can cause movement of  
 168 background parts of the scene like water ripples, ropes, etc., as well as movement of the camera itself.

## 169 4 Long-term Performance Experiment

170 We study the effects of concept drift on six machine learning models - two autoencoders, two object  
 171 detectors and anomaly detectors. For these experiments only weather parameters not found to have  
 172 significant correlation to other parameters are considered, namely - temperature, humidity, wind speed,  
 173 wind direction and precipitation. More information on the correlation between weather parameters is  
 174 given in the Appendix.

### 175 4.1 Data Selection Protocol

176 In order to keep the experiments and labelling effort manageable, samples across the full data set  
 177 are selected based on the following protocol. This is done to minimize the number of frames and  
 178 maximize the variation covered by the selection. For the sampling temperature metadata is used, as  
 179 it is proven to directly correlate with changes to thermal images [80, 35, 21]. The protocol can be  
 180 summarized as follows:

- 181 1. Every **2-minute** clip in the dataset is sampled with a frequency of one frame per second,  
 182 resulting in **120 frames per clip**;
- 183 2. Based on the temperature metadata, we select a cold month for the training set and another  
 184 cold month, a median temperature one, and a warm month for the test set;
- 185 3. The training set exists in three variants: coldest day 13th of February, the corresponding  
 186 week 13-20 of February, and the entirety of February;
- 187 4. The test sets consist of data from January (similar cold month), April (month with median  
 188 temperature), and August (warmest month).

189 From each of the thus created subsets, a greedy furthest point sampling is used for selecting frames.  
190 The frames for each day are sampled by calculating the farthest distances in the 2D feature space  
191 of the frame numbering and the temperature. A visual example of the sampling can be seen in  
192 the Appendix. The amounts of selected samples vary for the training data depending on the used  
193 algorithm. This is further discussed in the next sections.

## 194 4.2 Tested Models

195 Six deep learning models are tested. All six are originally designed to work with RGB data, so their  
196 input channel is reduced from 3 to 1, corresponding to a change to the grayscale thermal data. No  
197 additional changes were made, as the focus of the paper is not algorithm performance but change in  
198 performance over time.

199 Two of tested models are autoencoders, as representatives for dimensional reduction, noise removal,  
200 concept drift detection and anomaly detection methods. Autoencoders are well suited for researching  
201 concept drift in long-term datasets, as their reconstruction performance is inherently tightly connected  
202 to the training data. The first autoencoder follows a simple fully convolutional architecture with  
203 symmetric 5-layer encoder and decoder. The implementation is based on the autoencoder used in a  
204 previous work [43]. It is theorized that its simplicity will make it sensitive to concept drift in the input  
205 data. The second autoencoder is the latest version of the Vector Quantised Variational Autoencoder  
206 (VQVAE2) [63]. This autoencoder uses collections of multi-scale hierarchical discrete tensors, called  
207 codebooks, to map its latent space. This gives it more robustness compared to regular autoencoders.  
208 The VQVAE2 implementation used here is closely based on [50]. Both autoencoders are trained for  
209 200 epochs.

210 Two versions of the anomaly detector method MNAD [57] are also tested. They extend traditional  
211 autoencoders, by introducing memory-guided normality detection. We look at the typical reconstruc-  
212 tion based comparison (MNAD\_recon), as well as the prediction approach (MNAD\_pred), using  
213 the preceding four consecutive frames to predict the future frame. The backbone consists of the  
214 U-Net structure, without skip-connections for the MNAD\_recon variant. In between the encoder and  
215 decoder of U-Net is a memory module, storing prototypical events, concatenated with the original  
216 encoder output. The memory is primarily learned during training, but also updates during testing.  
217 Both versions are trained for 100 epochs.

218 Lastly two supervised object detectors are also tested - the YOLOv5 and Faster R-CNN[64]. The  
219 chosen hyperparameters for YOLOv5 remain the same as the work in [82], except that the initial  
220 learning rate is set to 0.00075 and trained for 200 epochs. The Faster R-CNN is trained for 200  
221 epochs as well with SGD, with initial learning rate set as 0.005, the weight decay as 0.005 and the  
222 momentum kept at 0.9. Both object detectors have previously been successfully applied to outdoor  
223 thermal imaging [38, 31, 9, 18].

224 The autoencoders are trained on a NVIDIA GTX1070 Super, the anomaly detectors on a NVIDIA  
225 RTX3080 and the object detectors on a NVIDIA RTX2080Ti.

## 226 4.3 Drift Algorithmic Performance Analysis

227 This experiment aims to see how the performance of the selected algorithms changes depending on  
228 the variation of the training data.

229 The training sets for the autoencoders and the anomaly detectors contain 5000 frames per subset, sam-  
230 pled using the method discussed in subsection 4.1, where 20% are used for validation. Performance  
231 is reported as the average MSE across every image in each of the three test sets. The performance of  
232 the two autoencoders and anomaly detectors is listed in Table 3. We can see that the MSE for the  
233 CAE, VQVAE2 and MNAD\_recon increases the farther away the test data goes from the training data.  
234 It can also be seen that the larger temporal pool provided for sampling for the weekly and monthly  
235 training data helps with keeping the MSE lower through the different months. The MNAD\_pred is  
236 the only model keeping a consistent performance through the months without any noticeable drift.  
237 This is most likely due to the U-Net skip connections being able to reconstruct the background scene  
238 with a very low reconstruction error.

239 For the object detectors, because of the necessary data-labeling a smaller number of images are  
240 used for training and testing - both having 100 frames per subset. In addition to these a validation

241 set comprising of 51 images evenly sampled from a previous annotated dataset [43] collected in  
 242 February 2020 is used. All of the subsets are annotated with bounding boxes around people seen  
 243 in each frame using the LabelImg open source program [81]. The annotations are also part of the  
 244 LTD dataset. Since the performance of object detector is based on detected bounding boxes, mAP is  
 245 used to evaluate it. The performance of the object detectors is given in Table 4. The accuracy of both  
 246 object detectors, drastically drops in the month of April. To prevent overfitting the smaller amount of  
 247 training data, we observe the validation and test loss.

248 As a conclusion from the performance analysis the higher variation provided by sampling from  
 249 the week and month data, has been translated to better and more stable models in all the tested  
 250 models. We can still see the effects of the seasonal drift, so additional analysis will be provided in the  
 251 following sections.

Table 3: Results are reported as the average of the MSE across every frame in the test set. Higher results show worse performance.

| Methods     | Train    |        | Test   |        |
|-------------|----------|--------|--------|--------|
|             | Feb.     | Jan.   | Apr.   | Aug.   |
| CAE         | Day 5k   | 0.0096 | 0.0202 | 0.0242 |
|             | Week 5k  | 0.0061 | 0.0167 | 0.0212 |
|             | Month 5k | 0.0042 | 0.0109 | 0.0147 |
| VQVAE2      | Day 5k   | 0.0051 | 0.0072 | 0.0068 |
|             | Week 5k  | 0.0039 | 0.0066 | 0.0061 |
|             | Month 5k | 0.0021 | 0.0039 | 0.0035 |
| MNAD Recon. | Day 5k   | 0.0028 | 0.0057 | 0.0069 |
|             | Week 5k  | 0.0065 | 0.0066 | 0.0062 |
|             | Month 5k | 0.0015 | 0.0041 | 0.0048 |
| MNAD Pred.  | Day 5k   | 0.0008 | 0.0007 | 0.0009 |
|             | Week 5k  | 0.0007 | 0.0006 | 0.0007 |
|             | Month 5k | 0.0007 | 0.0006 | 0.0007 |

Table 4: Results are reported as the mAP<sub>50</sub> across every frame in the test set. Lower results show worse performance.

| Method       | Train     |        | Test   |        |
|--------------|-----------|--------|--------|--------|
|              | Feb.      | Jan.   | Apr.   | Aug.   |
| YOLOv5       | Day 100   | 0.8010 | 0.5390 | 0.5240 |
|              | Week 100  | 0.7940 | 0.4540 | 0.4860 |
|              | Month 100 | 0.7930 | 0.4860 | 0.4830 |
| Faster R-CNN | Day 100   | 0.6760 | 0.3230 | 0.3370 |
|              | Week 100  | 0.6740 | 0.2790 | 0.3060 |
|              | Month 100 | 0.6400 | 0.2560 | 0.3180 |

## 252 5 Drift Analysis

253 In this section we look at the possible relations between the observed model performance drift and  
 254 the changes in the captured metadata. Looking through the data examples given in Figure 1, two  
 255 main visual change types are identified - seasonal and day/night. These types can be caused by either  
 256 changes in the weather conditions, the human activity or a combination between the two. The relation  
 257 between the model performance metrics and metadata features representing these changes is analysed.  
 258 As discussed in section 3.1, we choose temperature, humidity, precipitation, wind direction and wind  
 259 speed as weather data features. For analysing the day/night changes the timestamp data is used to  
 260 calculate hours of the day, as well as to calculate the sunrise and sunset times [74, 51]. To quantify  
 261 the activity in the scene the difference between each testing frame and the previous frame from the  
 262 main dataset is calculated. The mean value from this difference is selected. To focus only on scene  
 263 activity everything in the background that moves like the waterfront and the visible ropes and masts  
 264 is masked out. More information on this can be found in the Appendix.

265 We choose to use the results only from the models trained on the monthly February data, for easier  
 266 visualization. The correlation between each of these features and the measured performance metric  
 267 for each of the methods is first calculated. For the autoencoders and anomaly detectors this is the  
 268 MSE, while for the object detectors we calculate the F1-score from all images containing people, as it  
 269 gives a good overview of the precision and recall of the models. Both the basic Pearson’s correlation,  
 270 as well as the more sensitive to non-linear relations Distance correlation [78, 16] are calculated. The  
 271 statistical significance p-values are also calculated with threshold at 0.05. The calculated correlation  
 272  $r$  values are given in Table 5, where those with p-values below the threshold are shown in red.

273 From Table 5 it can be seen that temperature and humidity have both the largest correlation values to  
 274 most of the metrics, as well as the most consistently statistically significant results, followed by the  
 275 scene activity and day/night features. We focus on these four features in the following analysis.

276 To get a better understanding of not only the correlational, but also causal relations between the  
 277 models’ performance metrics and the chosen features, we look at the Granger causality test [22].



Table 5: Correlation between the model’s measured performance values MSE and F1-score and the weather, time and scene activity features. Two correlation measures are used - Pearson’s (P.C.) and Distance (D.C.) correlation. Measures which do not meet the statistical significance threshold of their p-values are shown in red and marked **X**. The Day/Night features is specified as D./N.

|                         | Measure | Temp.          | Hum.  | Wind Dir.      | Wind Sp.       | Precip.         | Activ.         | D./N.          | Hour           |
|-------------------------|---------|----------------|-------|----------------|----------------|-----------------|----------------|----------------|----------------|
| CAE - MSE               | P. C.   | 0.679          | 0.636 | 0.018 <b>X</b> | 0.157          | 0.109 <b>X</b>  | 0.270          | 0.545          | 0.166          |
|                         | D. C.   | 0.682          | 0.588 | 0.158          | 0.170          | 0.126 <b>X</b>  | 0.291          | 0.538          | 0.287          |
| VQVAE2 - MSE            | P. C.   | 0.381          | 0.690 | 0.001 <b>X</b> | 0.194          | 0.172           | 0.217          | 0.403          | 0.124          |
|                         | D. C.   | 0.347          | 0.639 | 0.174          | 0.201          | 0.224           | 0.217          | 0.382          | 0.213          |
| MNAD Recon. - MSE       | P. C.   | 0.607          | 0.672 | 0.016 <b>X</b> | 0.173          | 0.126           | 0.220          | 0.509          | 0.156          |
|                         | D. C.   | 0.617          | 0.629 | 0.188          | 0.177          | 0.155           | 0.252          | 0.501          | 0.273          |
| MNAD Pred. - MSE        | P. C.   | 0.107 <b>X</b> | 0.277 | 0.064 <b>X</b> | 0.152          | 0.072 <b>X</b>  | 0.677          | 0.369          | 0.137          |
|                         | D. C.   | 0.231          | 0.348 | 0.154          | 0.172          | 0.086 <b>X</b>  | 0.665          | 0.462          | 0.312          |
| YOLOv5 - F1-score       | P. C.   | 0.261          | 0.258 | 0.102 <b>X</b> | 0.011 <b>X</b> | 0.096 <b>X</b>  | 0.124 <b>X</b> | 0.047 <b>X</b> | 0.009 <b>X</b> |
|                         | D. C.   | 0.293          | 0.283 | 0.146 <b>X</b> | 0.094 <b>X</b> | 0.135 <b>X</b>  | 0.255          | 0.113 <b>X</b> | 0.174 <b>X</b> |
| Faster R-CNN - F1-score | P. C.   | 0.354          | 0.456 | 0.115 <b>X</b> | 0.135 <b>X</b> | 0.0124 <b>X</b> | 0.199          | 0.147          | 0.001 <b>X</b> |
|                         | D. C.   | 0.334          | 0.460 | 0.228          | 0.149 <b>X</b> | 0.065 <b>X</b>  | 0.231          | 0.163          | 0.118 <b>X</b> |

278 The test only guarantees a predictive causality between variables, but would be able to point out  
 279 any possible connections. The Granger causality tests the null hypothesis that the past values of one  
 280 variable do not cause another. The p-value threshold is set to 0.05, below that the null hypothesis  
 281 can be rejected, with the conclusion that there is a predictive causality between the variables. As the  
 282 normal Granger causality test as presented in [71] is used on data with linear relations, we also use  
 283 the more robust non-linear Neural Granger test [79]. Two best performing versions are used, based  
 284 on long-short term memory networks (LSTM) and multi-level perceptron (MLP). Both models were  
 285 trained using proximal gradient descent [56], with  $\lambda = 0.002$ , ridge regression coefficient 0.01 and  
 286 learning rate of 0.005. The results from the Granger causality tests are given in Table 6, where cells  
 287 shown with green indicate a statistically significant presence of Granger causality and the ones with  
 288 red - no presence.

Table 6: Results from calculating linear and non-linear (LSTM and MLP) Granger causality tests. The cells marked with **✓** show positive predictive causality, while cells marked with **X** show no significant causality.

|                         | Temp. |      |     | Hum.  |      |     | Activ. |      |     | D./N. |      |     |
|-------------------------|-------|------|-----|-------|------|-----|--------|------|-----|-------|------|-----|
|                         | Basic | LSTM | MLP | Basic | LSTM | MLP | Basic  | LSTM | MLP | Basic | LSTM | MLP |
| CAE - MSE               | ✓     | ✓    | ✓   | ✓     | ✓    | X   | X      | X    | X   | ✓     | ✓    | ✓   |
| VQVAE2 - MSE            | ✓     | ✓    | ✓   | ✓     | ✓    | X   | X      | X    | X   | ✓     | ✓    | ✓   |
| MNAD Recon. - MSE       | ✓     | ✓    | ✓   | ✓     | ✓    | X   | X      | X    | X   | ✓     | ✓    | ✓   |
| MNAD Pred. - MSE        | ✓     | ✓    | X   | X     | X    | X   | X      | ✓    | X   | ✓     | ✓    | ✓   |
| YOLOv5 - F1-score       | ✓     | X    | X   | ✓     | X    | ✓   | X      | X    | X   | X     | X    | X   |
| Faster R-CNN - F1-score | X     | X    | X   | X     | ✓    | X   | X      | X    | X   | ✓     | ✓    | ✓   |

289 The results show that the human activity has no predictive causality towards the performance of the  
 290 models, which combined with the results from the correlation analysis, can point towards a second-  
 291 hand relation. Our hypothesis is that the change in weather conditions and the day/night cycle are  
 292 related to the change in human activity. From the other features, temperature has stronger predictive  
 293 causality towards the autoencoders and anomaly detectors, while humidity and the day/night cycle  
 294 have a more balanced predictive causality.

295 Figure 2 shows the relationship between the features and the model metrics. As a processing step  
 296 before plotting the temperature and humidity they are first smoothed using a mean filter with a kernel  
 297 size of 20 and then the MSE is normalized between 0 and 1. This is done as they are not compared,  
 298 but the trend of their change is visualized. We plot the average values for the training month of  
 299 February, as a vertical red line, to indicate a "threshold".

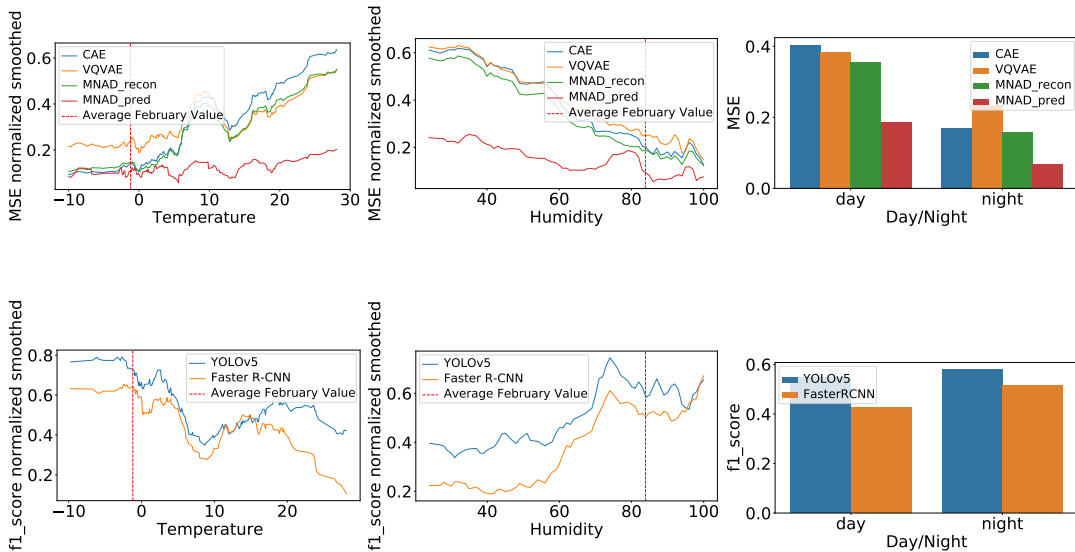


Figure 2: Visual representation of the changes of MSE and F1-score for the tested models compared to the temperature, humidity and day/night cycle.

## 300 6 Drift Prediction Baseline

301 As a baseline for exploring and mitigating the effects of concept drift a reference algorithm for  
 302 predicting drift is presented. We use three strongest features - temperature, humidity and day/night  
 303 cycle, together with MSE from our convolutional autoencoder (CAE) trained on the February monthly  
 304 data. The CAE is chosen, as it is the most sensitive to changes in the dataset and is strongly correlated  
 305 to the performance of all other tested models, except Faster R-CNN. The CAE MSE results from  
 306 the training data are used together with the chosen features to train two widely used novelty/outlier  
 307 detection models - isolation forests [42] and one-class SVM [59], available as part of scikit-learn  
 308 [58]. The isolation forest has 100 base estimators, the one-class SVM has a radial basis function  
 309 (RBF) kernel and  $\gamma$  of 0.03. We then test the results from each day from the full LTD dataset to detect  
 310 points where many outliers emerge in both predictors. The first large concentration of outliers in 7  
 311 consecutive days is selected, which in our case is 5th of March.

312 To test if taking in consideration data from the found drift point can help with the performance of the  
 313 models against concept drift, training data from one week starting after the 5th of March is sampled.  
 314 The new data is used together with the previous training data from February to retrain the tested  
 315 models. The results, together with the month results from Table 3 and 4 for comparison, are given in  
 316 Table 7 and Table 8. By adding the March data, all tested models achieve better results. We can see  
 317 that the outlier detection models trained on the CAE MSE, together with the temperature, humidity  
 318 and day/night cycle can be used together as an indicator for the amount of drift present in the input  
 319 data.

## 320 7 Conclusion and Future Work

321 In this paper we introduced the Long-term Thermal Drift (LTD) dataset spanning 8 months for  
 322 detecting concept drift in deep learning models. The dataset and the accompanying metadata can be  
 323 used to document performance degradation as data drifts from the training set. These effects were  
 324 studied on anomaly and object detection models, as well as autoencoders. It was demonstrated that  
 325 more diverse training data lowers the effects of concept drift. The performance of the models showed  
 326 a strong correlational and causal relationship to the change in temperature and humidity. A less  
 327 pronounced relationship was observed to the day/night cycle and scene activity. Lastly, we showed

Table 7: The MSE results from the full month in Table 3, compared to the ones using the new training datasets containing a combination of February and the week in March where drift is detected. Higher results show worse performance.

| Methods     | Train             | Test          |               |               |
|-------------|-------------------|---------------|---------------|---------------|
|             |                   | Jan.          | Apr.          | Aug.          |
| VQVAE2      | Feb. 5k           | 0.0021        | 0.0039        | 0.0035        |
|             | Feb. 5k + Mar. 5k | <b>0.0020</b> | <b>0.0033</b> | <b>0.0030</b> |
| MNAD Recon. | Feb. 5k           | 0.0015        | 0.0041        | 0.0048        |
|             | Feb. 5k + Mar. 5k | <b>0.0006</b> | <b>0.0015</b> | <b>0.0025</b> |
| MNAD Pred.  | Feb. 5k           | 0.0007        | 0.0006        | 0.0007        |
|             | Feb. 5k + Mar. 5k | <b>0.0007</b> | <b>0.0005</b> | <b>0.0006</b> |

Table 8: The mAP<sub>50</sub> Results from the full month in Table 4, compared to the ones using the new training datasets containing a combination of February and the week in March where drift is detected. Lower results show worse performance.

| Method       | Train               | Test          |               |               |
|--------------|---------------------|---------------|---------------|---------------|
|              |                     | Jan.          | Apr.          | Aug.          |
| YOLOv5       | Feb. 100            | 0.7930        | 0.4860        | 0.4830        |
|              | Feb. 100 + Mar. 100 | <b>0.8690</b> | <b>0.6640</b> | <b>0.6110</b> |
| Faster R-CNN | Feb. 100            | 0.6400        | 0.2560        | 0.3180        |
|              | Feb. 100 + Mar. 100 | <b>0.6990</b> | <b>0.3910</b> | <b>0.3380</b> |

328 how the concept drift can be further mitigated by detecting when it starts to manifest and providing  
 329 additional data to the training process.

330 The proposed LTD dataset contains a combination of diverse environmental images and granular  
 331 metadata. The equally spaced long-term data can be used to test the change in performance of deep  
 332 learning models at different data scenarios - only day or night data, changes between activity in the  
 333 weekday and weekends, summer and winter scenarios. The influence of weather conditions like rain,  
 334 snow or fog can also be explored. The possibility of training more robust models and predicting  
 335 when steps need to be taken, before their performance degrades, is only possible with such long-term  
 336 sequential datasets.

337 Possible negative social impacts of such long-term datasets concentrating on a single location is that  
 338 they can be used to track the habits, interactions and movements of people. We offset this by providing  
 339 a thermal dataset, which provides greater protection of people’s anonymity than conventional RGB  
 340 and does not require post-processing for blurring facial features.

341 The long-term nature of the dataset can also be used, as demonstrated in this paper, to utilize time-  
 342 series analysis procedures on the outputs from different layers of deep learning models. From simple  
 343 time-series analysis and forecasting models like Vector Autoregressive (VAR) Models [29] to more  
 344 complex and data agnostic models like STRIPE [25] or Adversarial Sparse Transformers [84].

345 We believe that the proposed dataset and the accompanied analysis would help researchers understand  
 346 the causes for performance drift in models and hence enable easier deployment of long-term solutions  
 347 in outdoor environments.

## 348 References

- 349 [1] Siddharth Agarwal, Ankit Vora, Gaurav Pandey, Wayne Williams, Helen Kourous, and James  
 350 McBride. Ford multi-av seasonal dataset. *The International Journal of Robotics Research*, 39  
 351 (12):1367–1376, 2020.
- 352 [2] Paulo RL Almeida, Luiz S Oliveira, Alceu S Britto Jr, and Robert Sabourin. Adapting dynamic  
 353 classifier selection for concept drift. *Expert Systems with Applications*, 104:67–85, 2018.
- 354 [3] M Ball and Harry Pinkerton. Factors affecting the accuracy of thermal imaging cameras in  
 355 volcanology. *Journal of Geophysical Research: Solid Earth*, 111(B11), 2006.
- 356 [4] Erwan Bernard, Nicolas Rivière, Mathieu Renaudat, Michel Péalat, and Emmanuel Zenou.  
 357 Active and thermal imaging performance under bad weather conditions. 2014.
- 358 [5] Ilker Bozcan and Erdal Kayacan. Au-air: A multi-modal unmanned aerial vehicle dataset  
 359 for low altitude traffic surveillance. In *2020 IEEE International Conference on Robotics and  
 360 Automation (ICRA)*, pages 8504–8510. IEEE, 2020.
- 361 [6] Antonio Brunetti, Domenico Buongiorno, Gianpaolo Francesco Trotta, and Vitoantonio Bevilac-  
 362 qua. Computer vision and deep learning techniques for pedestrian detection and tracking: A  
 363 survey. *Neurocomputing*, 300:17–33, 2018.

- 364 [7] DM Bulanon, TF Burks, and V Alchanatis. Study on temporal variation in citrus canopy using  
365 thermal imaging for citrus fruit detection. *Biosystems Engineering*, 101(2):161–171, 2008.
- 366 [8] Li Chen, Nan Ma, Patrick Wang, Jiahong Li, Pengfei Wang, Guilin Pang, and Xiaojun Shi.  
367 Survey of pedestrian action recognition techniques for autonomous driving. *Tsinghua Science  
368 and Technology*, 25(4):458–470, 2020.
- 369 [9] Yung-Yao Chen, Sin-Ye Jhong, Guan-Yi Li, and Ping-Han Chen. Thermal-based pedestrian  
370 detection using faster r-cnn and region decomposition branch. In *2019 International Symposium  
371 on Intelligent Signal Processing and Communication Systems (ISPACS)*, pages 1–2. IEEE, 2019.
- 372 [10] Zhilu Chen and Xinming Huang. Pedestrian detection for autonomous vehicle using multi-  
373 spectral cameras. *IEEE Transactions on Intelligent Vehicles*, 4(2):211–219, 2019.
- 374 [11] JOSEFINE CORNÉ and ULRIKA HELANDER SJÖBLOM. *Investigation of IR transmittance  
375 in different weather conditions and simulation of passive IR imaging for flight scenarios*. PhD  
376 thesis, MS thesis, KTH Royal Institute of Technology, Stockholm, Sweden, 2019.
- 377 [12] Kellie Corona, Katie Osterdahl, Roderic Collins, and Anthony Hoogs. Meva: A large-scale  
378 multiview, multimodal video dataset for activity detection. In *Proceedings of the IEEE/CVF  
379 Winter Conference on Applications of Computer Vision*, pages 1060–1068, 2021.
- 380 [13] Piotr Dollar, Christian Wojek, Bernt Schiele, and Pietro Perona. Pedestrian detection: An  
381 evaluation of the state of the art. *IEEE transactions on pattern analysis and machine intelligence*,  
382 34(4):743–761, 2011.
- 383 [14] Yole Développement. Thermal imagers and detectors 2020 - covid-  
384 19 outbreak impact – preliminary report. [http://www.yole.fr/  
385 Thermal\\_Imagers\\_And\\_Detectors\\_Covid19\\_Outbreak\\_Impact.aspx](http://www.yole.fr/Thermal_Imagers_And_Detectors_Covid19_Outbreak_Impact.aspx), 2020. Accessed:  
386 2021-08-11.
- 387 [15] Faten A Elshwemy, Reda Elbasiony, and Mohamed Talaat Saidahmed. A new approach  
388 for thermal vision based fall detection using residual autoencoder. *International Journal of  
389 Intelligent Engineering and Systems*, 13(2):250–258, 2020.
- 390 [16] Wladston Filho. Distance correlation. <https://gist.github.com/wladston>, 2020. Ac-  
391 cessed: 2021-07-22.
- 392 [17] FLIR. Flir thermal dataset for algorithm training. [https://www.flir.com/oem/adas/adas-  
393 dataset-form/](https://www.flir.com/oem/adas/adas-dataset-form/), 2019. Accessed: 2021-09-26.
- 394 [18] Debasmita Ghose, Shasvat M Desai, Sneha Bhattacharya, Deep Chakraborty, Madalina Fiterau,  
395 and Tauhidur Rahman. Pedestrian detection in thermal images using saliency maps. In *Pro-  
396 ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*,  
397 pages 0–0, 2019.
- 398 [19] Özgür Göçer, Kenan Göçer, Barış Özcan, Mujesira Bakovic, and M Furkan Kıraç. Pedestrian  
399 tracking in outdoor spaces of a suburban university campus for the investigation of occupancy  
400 patterns. *Sustainable cities and society*, 45:131–142, 2019.
- 401 [20] Alejandro González, Zhijie Fang, Yainuvis Socarras, Joan Serrat, David Vázquez, Jiaolong Xu,  
402 and Antonio M López. Pedestrian detection at day/night time with visible and fir cameras: A  
403 comparison. *Sensors*, 16(6):820, 2016.
- 404 [21] AA Gowen, BK Tiwari, PJ Cullen, K McDonnell, and CP O’Donnell. Applications of thermal  
405 imaging in food quality and safety assessment. *Trends in food science & technology*, 21(4):  
406 190–200, 2010.
- 407 [22] Clive WJ Granger. Investigating causal relations by econometric models and cross-spectral  
408 methods. *Econometrica: journal of the Econometric Society*, pages 424–438, 1969.
- 409 [23] Tobias Gruber, Frank Julca-Aguilar, Mario Bijelic, and Felix Heide. Gated2depth: Real-time  
410 dense lidar from gated images. In *The IEEE International Conference on Computer Vision  
411 (ICCV)*, 2019.

- 412 [24] Mingyang Guan, Changyun Wen, Mao Shan, Cheng-Leong Ng, and Ying Zou. Real-time  
413 event-triggered object tracking in the presence of model drift and occlusion. *IEEE Transactions*  
414 *on Industrial Electronics*, 66(3):2054–2065, 2018.
- 415 [25] Vincent Le Guen and Nicolas Thome. Probabilistic time series forecasting with structured shape  
416 and temporal diversity. *arXiv preprint arXiv:2010.07349*, 2020.
- 417 [26] Hua Han, MengChu Zhou, Xiwu Shang, Wei Cao, and Abdullah Abusorrah. Kiss+ for rapid and  
418 accurate pedestrian re-identification. *IEEE Transactions on Intelligent Transportation Systems*,  
419 22(1):394–403, 2020.
- 420 [27] Fouzi Harrou, Nabil Zerrouki, Ying Sun, and Amrane Houacine. An integrated vision-based  
421 approach for efficient human fall detection in a home environment. *IEEE Access*, 7:114966–  
422 114974, 2019.
- 423 [28] Manzoor Ahmed Hashmani, Syed Muslim Jameel, Hitham Al-Hussain, Mobashar Rehman,  
424 and Arif Budiman. Accuracy performance degradation in image classification models due to  
425 concept drift. *Int. J. Adv. Comput. Sci. Appl*, 10, 2019.
- 426 [29] Jonas MB Haslbeck, Laura F Bringmann, and Lourens J Waldorp. A tutorial on estimating  
427 time-varying vector autoregressive models. *Multivariate Behavioral Research*, 56(1):120–149,  
428 2021.
- 429 [30] Hikvision. Ds-2td2235d-25/50. [https://us.hikvision.com/en/products/more-](https://us.hikvision.com/en/products/more-products/discontinued-products/thermal-camera/thermal-network-bullet-camera-ds)  
430 [products/discontinued-products/thermal-camera/thermal-network-bullet-](https://us.hikvision.com/en/products/more-products/discontinued-products/thermal-camera/thermal-network-bullet-camera-ds)  
431 [camera-ds](https://us.hikvision.com/en/products/more-products/discontinued-products/thermal-camera/thermal-network-bullet-camera-ds), 2015. Accessed: 2021-05-27.
- 432 [31] Noor Ul Huda, Bolette D Hansen, Rikke Gade, and Thomas B Moeslund. The effect of a diverse  
433 dataset for transfer learning in thermal person detection. *Sensors*, 20(7):1982, 2020.
- 434 [32] Soonmin Hwang, Jaesik Park, Namil Kim, Yukyung Choi, and In So Kweon. Multispectral  
435 pedestrian detection: Benchmark dataset and baseline. In *Proceedings of the IEEE conference*  
436 *on computer vision and pattern recognition*, pages 1037–1045, 2015.
- 437 [33] Danish Meteorological Institute. Dmi api. [https://confluence.govcloud.dk/display/](https://confluence.govcloud.dk/display/FDAPI)  
438 [FDAPI](https://confluence.govcloud.dk/display/FDAPI), 2019. Accessed: 2021-05-27.
- 439 [34] Mordor Intelligence. Ir camera market - growth, trends, covid-19 impact, and fore-  
440 casts (2021 - 2026). [https://www.mordorintelligence.com/industry-reports/ir-](https://www.mordorintelligence.com/industry-reports/ir-camera-market)  
441 [camera-market](https://www.mordorintelligence.com/industry-reports/ir-camera-market), 2021. Accessed: 2021-08-11.
- 442 [35] CA James, AJ Richardson, PW Watt, and NS Maxwell. Reliability and validity of skin  
443 temperature measurement by telemetry thermistors and a thermal camera during exercise in the  
444 heat. *Journal of thermal biology*, 45:141–149, 2014.
- 445 [36] Iason Katsamenis, Eftychios Protopapadakis, Athanasios Voulodimos, Dimitris Dres, and  
446 Dimitris Drakoulis. Man overboard event detection from rgb and thermal imagery: Possibilities  
447 and limitations. In *Proceedings of the 13th ACM International Conference on PErvasive*  
448 *Technologies Related to Assistive Environments*, pages 1–6, 2020.
- 449 [37] My Kieu, Andrew D Bagdanov, and Marco Bertini. Bottom-up and layerwise domain adaptation  
450 for pedestrian detection in thermal images. *ACM Transactions on Multimedia Computing,*  
451 *Communications, and Applications (TOMM)*, 17(1):1–19, 2021.
- 452 [38] Mate Krišto, Marina Ivacic-Kos, and Miran Pobar. Thermal object detection in difficult weather  
453 conditions using yolo. *IEEE Access*, 8:125459–125476, 2020.
- 454 [39] Santhosh Kelathodi Kumaran, Debi Prosad Dogra, and Partha Pratim Roy. Anomaly detection  
455 in road traffic using visual surveillance: A survey. *arXiv preprint arXiv:1901.08292*, 2019.
- 456 [40] Aske Rasch Lejbølle, Kamal Nasrollahi, Benjamin Krogh, and Thomas B Moeslund. Person  
457 re-identification using spatial and layer-wise attention. *IEEE Transactions on Information*  
458 *Forensics and Security*, 15:1216–1231, 2019.

- 459 [41] Hui Li, Meng Yang, Zihui Lai, Weishi Zheng, and Zitong Yu. Pedestrian re-identification based  
460 on tree branch network with local and global learning. In *2019 IEEE International Conference*  
461 *on Multimedia and Expo (ICME)*, pages 694–699. IEEE, 2019.
- 462 [42] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation-based anomaly detection. *ACM*  
463 *Transactions on Knowledge Discovery from Data (TKDD)*, 6(1):1–39, 2012.
- 464 [43] Jinsong Liu, Mark Philip Philipsen, and Thomas B Moeslund. Supervised versus self-supervised  
465 assistant for surveillance of harbor fronts. In *16th International Conference on Computer Vision*  
466 *Theory and Applications (VISAPP)*, 2021.
- 467 [44] Qiao Liu, Zhenyu He, Xin Li, and Yuan Zheng. Ptb-tir: A thermal infrared pedestrian tracking  
468 benchmark. *IEEE Transactions on Multimedia*, 22(3):666–675, 2019.
- 469 [45] W. Liu, D. Lian W. Luo, and S. Gao. Future frame prediction for anomaly detection – a new  
470 baseline. In *2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
- 471 [46] Cewu Lu, Jianping Shi, and Jiaya Jia. Abnormal event detection at 150 fps in matlab. In  
472 *Proceedings of the IEEE international conference on computer vision*, pages 2720–2727, 2013.
- 473 [47] Chao Ma, Ngo Thanh Trung, Hideaki Uchiyama, Hajime Nagahara, Atsushi Shimada, and  
474 Rin-ichiro Taniguchi. Adapting local features for face detection in thermal image. *Sensors*, 17  
475 (12):2741, 2017.
- 476 [48] Will Maddern, Geoffrey Pascoe, Chris Linegar, and Paul Newman. 1 year, 1000 km: The oxford  
477 robotcar dataset. *The International Journal of Robotics Research*, 36(1):3–15, 2017.
- 478 [49] Vijay Mahadevan, Wei-Xin LI, Viral Bhalodia, and Nuno Vasconcelos. Anomaly detection  
479 in crowded scenes. In *Proceedings of IEEE Conference on Computer Vision and Pattern*  
480 *Recognition*, pages 1975–1981, 2010.
- 481 [50] Alex McKinney. Vqvae2 implementation, 2021. URL [https://github.com/vvvm23/vqvae-](https://github.com/vvvm23/vqvae-2)  
482 [2](https://github.com/vvvm23/vqvae-2). last accessed: June 7, 2021.
- 483 [51] Jean Meeus. *Astronomical algorithms*. Richmond, 1991.
- 484 [52] Carlos Mera, Mauricio Orozco-Alzate, and John Branch. Incremental learning of concept  
485 drift in multiple instance learning for industrial visual inspection. *Computers in Industry*, 109:  
486 153–164, 2019.
- 487 [53] Pravin Nagar, Mansi Khemka, and Chetan Arora. Concept drift detection for multivariate data  
488 streams and temporal segmentation of daylong egocentric videos. In *Proceedings of the 28th*  
489 *ACM International Conference on Multimedia*, pages 1065–1074, 2020.
- 490 [54] Sangmin Oh, Anthony Hoogs, Amitha Perera, Naresh Cuntoor, Chia-Chih Chen, Jong Taek Lee,  
491 Saurajit Mukherjee, JK Aggarwal, Hyungtae Lee, Larry Davis, et al. A large-scale benchmark  
492 dataset for event recognition in surveillance video. In *CVPR 2011*, pages 3153–3160. IEEE,  
493 2011.
- 494 [55] Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin,  
495 Joshua V Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model’s uncer-  
496 tainty? evaluating predictive uncertainty under dataset shift. *arXiv preprint arXiv:1906.02530*,  
497 2019.
- 498 [56] Neal Parikh and Stephen Boyd. Proximal algorithms. *Foundations and Trends in optimization*,  
499 1(3):127–239, 2014.
- 500 [57] Hyunjong Park, Jongyoun Noh, and Bumsub Ham. Learning memory-guided normality for  
501 anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and*  
502 *Pattern Recognition*, pages 14372–14381, 2020.
- 503 [58] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel,  
504 P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher,  
505 M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine*  
506 *Learning Research*, 12:2825–2830, 2011.

- 507 [59] John Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized  
508 likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.
- 509 [60] Mantini Pranav, Li Zhenggang, et al. A day on campus-an anomaly detection dataset for events  
510 in a single camera. In *Proceedings of the Asian Conference on Computer Vision*, 2020.
- 511 [61] Stephan Rabanser, Stephan Günnemann, and Zachary C Lipton. Failing loudly: An empirical  
512 study of methods for detecting dataset shift. *arXiv preprint arXiv:1810.11953*, 2018.
- 513 [62] Bharathkumar Ramachandra and Michael Jones. Street scene: A new dataset and evaluation  
514 protocol for video anomaly detection. In *Proceedings of the IEEE/CVF Winter Conference on  
515 Applications of Computer Vision*, pages 2569–2578, 2020.
- 516 [63] Ali Razavi, Aaron van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images  
517 with vq-vae-2. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and  
518 R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Cur-  
519 ran Associates, Inc., 2019. URL [https://proceedings.neurips.cc/paper/2019/file/  
520 5f8e2fa1718d1bbcadf1cd9c7a54fb8c-Paper.pdf](https://proceedings.neurips.cc/paper/2019/file/5f8e2fa1718d1bbcadf1cd9c7a54fb8c-Paper.pdf).
- 521 [64] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time  
522 object detection with region proposal networks. *arXiv preprint arXiv:1506.01497*, 2015.
- 523 [65] Allied Market Research. Global thermal imaging camera market by 2030. [https:  
524 //www.globenewswire.com/news-release/2021/08/09/2277188/0/en/Global-  
525 Thermal-Imaging-Camera-Market-is-Expected-to-Reach-7-49-Billion-by-  
526 2030-Says-AMR.html](https://www.globenewswire.com/news-release/2021/08/09/2277188/0/en/Global-Thermal-Imaging-Camera-Market-is-Expected-to-Reach-7-49-Billion-by-2030-Says-AMR.html), 2021. Accessed: 2021-08-11.
- 527 [66] Royston Rodrigues, Neha Bhargava, Rajbabu Velmurugan, and Subhasis Chaudhuri. Multi-  
528 timescale trajectory prediction for abnormal human activity detection. In *Proceedings of the  
529 IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2626–2634, 2020.
- 530 [67] Yahia Fahem Said and Mohammad Barr. Pedestrian detection for advanced driver assistance  
531 systems using deep learning algorithms. *IJCSNS*, 19(10), 2019.
- 532 [68] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Guided curriculum model adaptation and  
533 uncertainty-aware evaluation for semantic nighttime image segmentation. In *Proceedings of the  
534 IEEE/CVF International Conference on Computer Vision*, pages 7374–7383, 2019.
- 535 [69] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Acdc: The adverse conditions dataset with  
536 correspondences for semantic driving scene understanding. *arXiv preprint arXiv:2104.13395*,  
537 2021.
- 538 [70] Torsten Sattler, Will Maddern, Carl Toft, Akihiko Torii, Lars Hammarstrand, Erik Stenborg,  
539 Daniel Safari, Masatoshi Okutomi, Marc Pollefeys, Josef Sivic, et al. Benchmarking 6dof  
540 outdoor visual localization in changing conditions. In *Proceedings of the IEEE Conference on  
541 Computer Vision and Pattern Recognition*, pages 8601–8610, 2018.
- 542 [71] Skipper Seabold and Josef Perktold. statsmodels: Econometric and statistical modeling with  
543 python. In *9th Python in Science Conference*, 2010.
- 544 [72] Tegjyot Singh Sethi and Mehmed Kantardzic. Handling adversarial concept drift in streaming  
545 data. *Expert systems with applications*, 97:18–40, 2018.
- 546 [73] Guojiang Shen, Linfeng Zhu, Jihan Lou, Si Shen, Zhi Liu, and Longfeng Tang. Infrared  
547 multi-pedestrian tracking in vertical view via siamese convolution network. *IEEE Access*, 7:  
548 42718–42725, 2019.
- 549 [74] Krzysztof Stopa, Andrey Kobyshev, Matthias, and Hadrien Bertrand. Suntime. [https://  
550 github.com/SatAgro/suntime](https://github.com/SatAgro/suntime), 2019. Accessed: 2021-07-22.
- 551 [75] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance  
552 videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*,  
553 pages 6479–6488, 2018.

- 554 [76] Yu Sun, Eric Tzeng, Trevor Darrell, and Alexei A Efros. Unsupervised domain adaptation  
555 through self-supervision. *arXiv preprint arXiv:1909.11825*, 2019.
- 556 [77] Abhijit Suprem, Joy Arulraj, Calton Pu, and Joao Ferreira. Odin: automated drift detection and  
557 recovery in video analytics. *arXiv preprint arXiv:2009.05440*, 2020.
- 558 [78] Gábor J Székely, Maria L Rizzo, Nail K Bakirov, et al. Measuring and testing dependence by  
559 correlation of distances. *The annals of statistics*, 35(6):2769–2794, 2007.
- 560 [79] Alex Tank, Ian Covert, Nicholas Foti, Ali Shojaie, and Emily Fox. Neural granger causality.  
561 *arXiv preprint arXiv:1802.05842*, 2018.
- 562 [80] Quang Huy Tran, Dongyeob Han, Choonghyun Kang, Achintya Haldar, and Jungwon Huh.  
563 Effects of ambient temperature and relative humidity on subsurface defect detection in concrete  
564 structures by active thermal imaging. *Sensors*, 17(8):1718, 2017.
- 565 [81] Tzutalin. Labelimg. <https://github.com/tzutalin/labelImg>, 2015. Accessed: 2021-06-  
566 06.
- 567 [82] Ultralytics. Yolov5, 2020. URL <https://github.com/ultralytics/yolov5>. last accessed:  
568 April 15, 2021.
- 569 [83] Meng Wang, Wei Li, and Xiaogang Wang. Transferring a generic pedestrian detector towards  
570 specific scenes. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages  
571 3274–3281. IEEE, 2012.
- 572 [84] Sifan Wu, Xi Xiao, Qianggang Ding, Peilin Zhao, WEI Ying, and Junzhou Huang. Adversarial  
573 sparse transformer for time series forecasting. 2020.
- 574 [85] Limin Yang, Wenbo Guo, Qingying Hao, Arridhana Ciptadi, Ali Ahmadzadeh, Xinyu Xing, and  
575 Gang Wang. {CADE}: Detecting and explaining concept drift samples for security applications.  
576 In *30th {USENIX} Security Symposium ({USENIX} Security 21)*, 2021.
- 577 [86] Zhiheng Yang, Jun Li, and Huiyun Li. Real-time pedestrian and vehicle detection for au-  
578 tonomous driving. In *2018 IEEE Intelligent Vehicles Symposium (IV)*, pages 179–184. IEEE,  
579 2018.
- 580 [87] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht  
581 Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous mul-  
582 titask learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern  
583 recognition*, pages 2636–2645, 2020.
- 584 [88] Yupeng Zhang, Yuheng Lu, Hajime Nagahara, and Rin-ichiro Taniguchi. Anonymous camera  
585 for privacy protection. In *2014 22nd International Conference on Pattern Recognition*, pages  
586 4170–4175. IEEE, 2014.
- 587 [89] Joey Tianyi Zhou, Jiawei Du, Hongyuan Zhu, Xi Peng, Yong Liu, and Rick Siow Mong Goh.  
588 Anomalynet: An anomaly detection network for video surveillance. *IEEE Transactions on  
589 Information Forensics and Security*, 14(10):2537–2550, 2019.
- 590 [90] Indrè Žliobaitė, Mykola Pechenizkiy, and Joao Gama. An overview of concept drift applications.  
591 *Big data analysis: new algorithms for a new society*, pages 91–114, 2016.