ChaLearn LAP Seasons in Drift Challenge

Dataset, Design and Results

Johansen, Anders Skaarup; Jacques Junior, Julio C.S.; Nasrollahi, Kamal; Escalera, Sergio; Moeslund, Thomas B.

Published in:
Computer Vision - ECCV 2022 Workshops, Proceedings

DOI (link to publication from Publisher):
10.1007/978-3-031-25072-9_50

Publication date:
2023

Document Version
Accepted author manuscript, peer reviewed version

ChaLearn LAP Seasons in Drift Challenge: Dataset, Design and Results

Anders Skaarup Johansen[1], Julio C. S. Jacques Junior[2], Kamal Nasrollahi[1,3], Sergio Escalera[2,4], and Thomas B. Moeslund[1]

[1] Aalborg University, Denmark,{asjo,kn,tmb}@create.aau.dk
[2] Computer Vision Center, Spain, jjacques@cvc.uab.cat
[3] Milestone Systems, Denmark, kna@milestone.dk
[4] University of Barcelona, Spain, sergio@maia.ub.es

Abstract. In thermal video security monitoring the reliability of deployed systems rely on having varied training data that can effectively generalize and have consistent performance in the deployed context. However, for security monitoring of an outdoor environment the amount of variation introduced to the imaging system would require extensive annotated data to fully cover for training and evaluation. To this end we designed and ran a challenge to stimulate research towards alleviating the impact of concept drift on object detection performance. We used an extension of the Long-Term Thermal Imaging Dataset, composed of thermal data acquired from 14th May 2020 to 30th of April 2021, with a total of 1689 2-minute clips with bounding-box annotations for 4 different categories. The data covers a wide range of different weather conditions and object densities with the goal of measuring the thermal drift over time, from the coldest day/week/month of the dataset. The challenge attracted 184 registered participants, which was considered a success from the perspective of the organizers. While participants managed to achieve higher mAP when compared to a baseline, concept drift remains a strongly impactful factor. This work describes the challenge design, the adopted dataset and obtained results, as well as discuss top-winning solutions and future directions on the topic.

1 Introduction

In the context of thermal video security monitoring the sensor type that is responsible of quantifying the observed infrared-radiation as a thermograph can be split into two groups: sensors that produce relative thermographs and sensors that produce absolute thermographs. Absolute thermographs can correlate the observed radiation directly with temperature, whereas relative thermographs produce observations relative to the "coldest" and "warmest" radiation. In security monitoring contexts the absolute temperature readings produced by an absolute thermograph are not necessary and can potentially suppress thermal details when observing thermally uniform environment. Furthermore the price of absolute thermal cameras are much higher than their relative counterpart.

When performing image recognition tasks the visual appearance of objects and their surroundings is very important, and in an outdoor context that is subjected to changes in temperature, weather, sun-radiation, among others, the visual appearance of objects and their surroundings change quite drastically. This is further expanded by societal factors like the recent pandemic which could introduce mandatory masks. This is known as "Concept Drift" where objects remain the same however the concept definition which is observed through representation changes. While in theory it could be possible to collect a large enough dataset encompassing the weather conditions, the actors, usually people, within the context also dress and act differently. Furthermore the cost of producing such a dataset would be quite extensive as potentially years worth of data would have to be annotated. Typically deployment of object detectors would have a pretrained baseline, and the model would have to be retrained when the observed context drifts too far away from the training context. The reliability in such a system is questionable as deployed algorithms tend not to have a way to quantify the performance during deployment and extra data would have to be routinely annotated to verify that the system is still performing as expected. To address this issue and foster more research into long-term reliability of deployed learning based object detectors a benchmark for classifying the impact of concept drift could greatly benefit the field.

The ECCV 2022 ChaLearn LAP Seasons in Drift Challenge aims to propose a setting for evaluating the impact of concept drift at a month to month basis and evaluating the impact of concept drift in a weighted manner. The problem of concept drift is exacerbated with limited training data, particularly when the distribution of the visual appearance in the data is similar. To explore the consistency of performance across varied levels of concept drift particularly of object detection algorithms, an extended set of frames were annotated spanning several months. The challenge attracted a total of 184 participants on its different tracks. With a total of 691 submissions at the different challenge stages and tracks, from over 180 participants, the challenge managed to successfully establish a benchmark for thermal concept drift. Top-wining solutions outperformed the baseline by a large margin following distinct strategies, detailed in Sec. 4.

The rest of the paper is organized as follows. In Sec. 2 we present the related work. The Challenge design, which includes a short description of the adopted dataset, evaluation protocol and baseline are detailed in Sec. 3. Challenge results and top-winning solutions are discussed in Sec. 4. Finally, conclusion and suggestions for future research directions are drawn in Sec. 5.

2 Related Work

Popular thermal detection and segmentation datasets, such as KAIST [13] and FLIR-ADAS [24], provide thermal and visible images. The focus of a large part of academic research have been focused on leveraging a multi-modal input [16,29,30,10] or using the aligned visible/thermal pairs as a way to do un-

supervised domain adaptation between the visible and thermal [25,28,7,10]. Approaches that leverage the multi-modal input directly typically use siamese style networks to perform modality specific feature extraction, subsequently leveraging a fusion scheme to combine the information in a learned manner [16,29,25], alternatively simple concatenation or addition is performed after initial feature extraction [10,30]. In contrast, a network can be optimized to be domain agnostic. HeatNet [25] and DANNet [28] leverage an adversarial approach to guide the network to extract domain agnostic features.

It has been proven that in security monitoring contexts fusion of visible and thermal images outperforms any modality alone [17,14], however in a real-world scenario camera setups tend to be single sensor setups. While thermal cameras are robust to changes in weather and lighting conditions, they still struggle with the change of visual appearance of objects due to the change of scene temperature [17,15,6,8,9]. Early work [9] leveraged edges to highlight objects, making detection possible robust to the variation when the relative contrast between objects and their surroundings were consistent. Recent studies leverage research in the visible imaging domain, and directly apply it to the thermal domain [17,6]. Until recently thermal specific detection methods have been a rarity and recently it was proven that contextual information is important to increase robustness to day/night variation [15,23] for thermal only object detection. By employing a conditioning of the latent representation guided by an auxiliary day/night classification head, the accuracy of day and night accuracy can be significantly increased [15]. Similar increase in performance can also be gained with a combination of a shallow feature-extractor and residual FPN-style connections [8]. Most notably the residual connections are leverage during training to enforce learning of discriminative features throughout the network, and serve no purpose during inference, and as such can be removed.

3 Challenge Design

The ECCV 2022 Seasons in Drift Challenge[5] aimed to spotlight the problem of concept drift in a security monitoring context and highlight the challenges and limitations of existing methods, as well as to provide a direction of research for the future. The challenge used an extension of the LTD Dataset [21] which consists of thermal footage that spans multiple seasons, detailed in Sec. 3.1. The challenge was split into 3 different tracks associated with thermal object detection. Each track having the same evaluation criteria/data but varying the amount of train data as well as the time span of the data, as detailed next.

- ***Track 1 - Detection at day level:*** Train on a predefined and single day data and evaluate concept drift across time[6]. The day is the 13th of February 2020 as it is the coldest day in the recorded data, due to the relative thermal

[5] Challenge - `https://chalearnlap.cvc.uab.cat/challenge/51/description/`
[6] Track 1 (on Codalab) - `https://codalab.lisn.upsaclay.fr/competitions/4272`

appearance of objects being the least varied in colder environments this is our starting point.
- ***Track 2 - Detection at week level:*** Train on a predefined and single week data and evaluate concept drift across time[7]. The week selected is the week of the 13th – 20th of February 2020 - (i.e. expanding from our starting point)
- ***Track 3 - Detection at month level:*** Train on a predefined and single month data and evaluate concept drift across time[8]. The selected month is the entire month of February.

The training data is chosen by selecting the coldest day, and surrounding data as cold environments introduce the least amount of concept drift. Each track aims at evaluating how robust a given detection method is to concept drift, by training on limited data from a specific time period (day, week, month in February) and evaluating performance across time, by validating and testing performance on months of unseen data (Jan., Mar., Apr., May., Jun., Jul., Aug. and Sep.). The February data is only present in the training set and the remaining months are equally split between validation and test.

Each track is composed of two phases, i.e., development and test phase. At the development phase, public train data was released and participants needed to submit their predictions with respect to a validation set. At the test (final) phase, participants needed to submit their results with respect to the test data, which was released just a few days before the end of the challenge. Participants were ranked, at the end of the challenge, using the test data. It is important to note that this competition involved the submission of results (and not code). Therefore, participants were required to share their codes and trained models after the end of the challenge so that the organizers could reproduce the results submitted at the test phase, in a "code verification stage". At the end of the challenge, top ranked methods that pass the code verification stage were considered as valid submissions.

3.1 The dataset

The dataset used in the challenge is an extension of the Long-Term Thermal Imaging [21] dataset, and spans 188 days in the period of 14th May 2020 to 30th of April 2021, with a total of 1689 2-minute clips sampled at 1fps with associated bounding box annotations for 4 classes (Human, Bicycle, Motorcycle, Vehicle). The collection of this dataset has included data from all hours of the day in a wide array of weather conditions overlooking the harborfront of Aalborg, Denmark. In this dataset depicts the drastic changes of appearance of the objects of interest as well as the scene over time in a static security monitoring context to develop robust algorithms for real-world deployment. Figure 1 illustrates the camera setup and two annotated frames of the dataset, obtained at different time intervals.

[7] Track 2 (on Codalab) - `https://codalab.lisn.upsaclay.fr/competitions/4273`
[8] Track 3 (on Codalab) - `https://codalab.lisn.upsaclay.fr/competitions/4276`

(a) (b) (c)

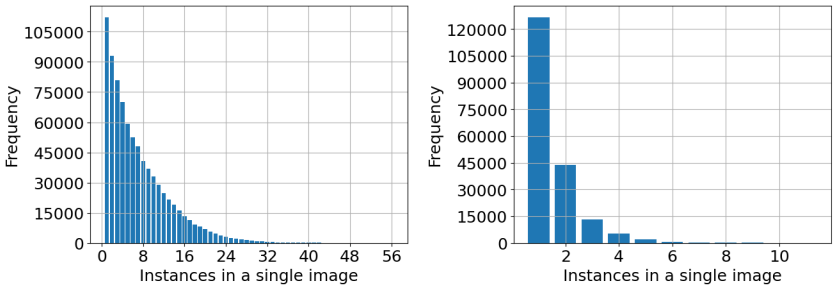
Fig. 1: Illustration of the camera setup (a) and two annotated frames of the dataset, captured at different time intervals (b-c).

For a detailed explination of the datasets weather contents, an overview can be found in the original dataset paper [21]. As for the extended annotations provided with this challenge, we can observe that the distribution of classes is heavily skewed towards the classes that are most commonly observed in the context. As can be seen in Table 1 the total number of occourances of each class is heavily scewed towards the *Person* class. Furthermore, as can be seen in Figure 2, each class follows roughly the same trend in terms of the density of which they occur. While the most common for all classes is a single count of the given object present in a given image is 1, the range of occurrences are greater for the *Person* category.

The camera used for recording the dataset was elevated above the observed area, and objects often appear very distant with regards to the camera, in combination with the resolution of the camera most objects appear very small in the image (see Figure 1). Table 1 summarizes the amount of objects from each class pertaining to each size category. The size is classified using the same scheme as used in the COCO dataset[19], where objects with areas $area < 32^2$, $32^2 < area < 96^2$ and $area > 96^2$ are considered small, medium and large respectively. The density of object sizes are also illustrated in Figure 3, where it can be more clearly seen that the vast majority of objects fall within the small category for classes. This holds true for classes *Person*, *Bicycle* and *Motorcycle*, where as the *Vehicle* class more evenly covers all size categories. This is a result of larger vehicles only being allowed to drive in the area closest to the camera.

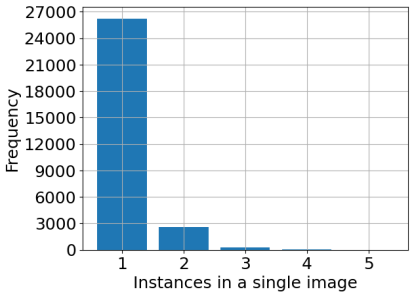
Table 1: Object frequency observed for each COCO-style size category.

	Class			
Size	Person	Bicycle	Motorcycle	Vehicle
Small	5.663.804	288.081	27.153	113.552
Medium	454	7	0	37.007
Large	176.881	5.192	5.240	550.696
Total	5.841.139	293.280	32.393	701.255

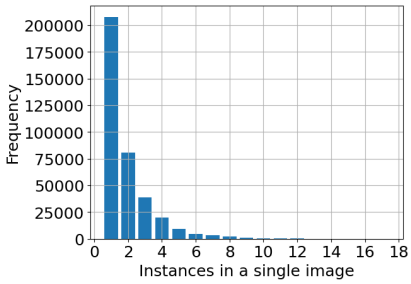


(a) Person

(b) Bicycle



(c) Vehicle



(d) Motorcycle

Fig. 2: Histogram of object density, across the dataset, density of objects (x-axis) and occurrences (y-axis).

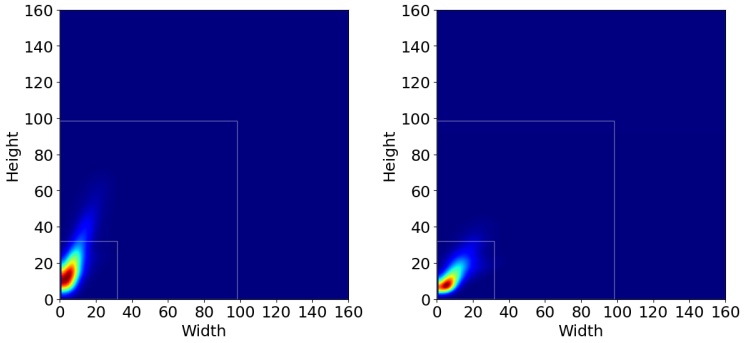
3.2 Evaluation protocol

The challenge followed the COCO evaluation[9] scheme for mAP. The primary metric is, mAP across 10 different IoU thresholds (ranging from 0.5 to 0.95 at 0.05 increments). This is calculated for each month in the validation/test set and the model is then ranked based on a weighted average of each month (more distant months having a larger weight as more concept drift is present), referred to as mAP_w in the analysis of the results (Table 2). The evaluation is performed leveraging the official COCO evaluation tools[10].

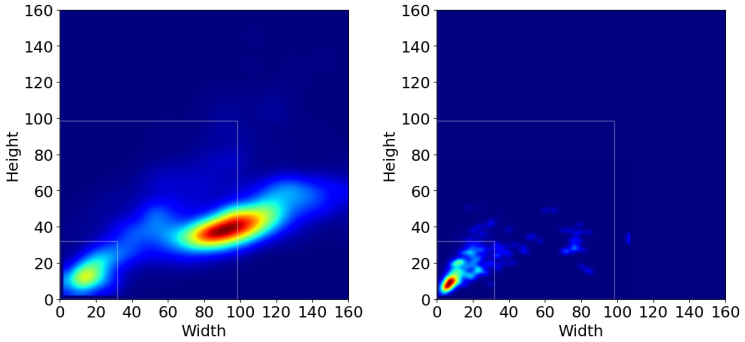
3.3 The baseline

The baseline is a YOLOv5 with the default configuration from the Ultralytics[11] repository, including augmentations. It was trained with a batch size of 64 for 300 epochs, with an input image size of 384×288 and the best performing model is chosen. Naturally, the labels were converted to the normalized yolo format

[9] `https://cocodataset.org/#detection-eval`
[10] `https://github.com/cocodataset/cocoapi`
[11] `https://github.com/ultralytics/yolov5`



(a) Person (b) Bicycle



(c) Vehicle (d) Motorcycle

Fig. 3: Illustration of object size (height×width, in pixels) across the dataset. The white outlines seperate the areas that would be labeled as small, medium and large following COCO standards.

([cls] [c_x] [c_y] [w] [ht]) for both training and evaluation. For submission on the Codalab platform they were converted back to the ([cls] [tl_x] [tl_y] [br_x] [br_y]) coordinates. The models were all trained on the same machine with 2x Nvidia RTX 3090 GPUs, all training is also conducted as multi GPU training using the pytorch distributed learning module.

4 Challenge Results and Winning Methods

The challenge ran from 25 April 2022 to 24 June 2022 through Codalab[12] [22], a powerful open source framework for running competitions that involve result or code submission. It attracted a total of 184 registered participants, 82, 52 and 50 on track 1, 2 and 3, respectively. During development phase we received 267 submissions from 17 active teams in track 1, 117 submissions from 6 teams in

[12] Codalab - `https://codalab.lisn.upsaclay.fr`

track 2, and 96 submissions from 4 teams in track 3. At the test (final) phase, we received 84 submissions from 23 active teams in track 1, 55 submissions from 22 teams in track 2, and 72 submissions from 24 teams in track 3. The reduction in the number of submissions from the development to the test phase is explained by the fact that the maximum number of submissions per participant on the final phase was limited to 3, to minimize the change of participants to improve their results by trial and error.

Table 2: Codalab leaderboards* at the test (final) phase.

Participant	mAP_w	mAP	Jan	Mar	Apr	May	Jun	Jul	Aug	Sep
Track 1 (day level)										
Team GroundTruth*	**.2798**	**.2832**	.3048	**.3021**	**.3073**	**.2674**	**.2748**	**.2306**	**.2829**	**.2955**
Team heboyong*	.2400	.2434	**.3063**	.2952	.2905	.2295	.2318	.1901	.2615	.1419
Team BDD	.2386	.2417	.2611	.2775	.2744	.2383	.2371	.1961	.2365	.2122
Team Charles	.2382	.2404	.2676	.2848	.2794	.2388	.2416	.2035	.2446	.1630
Team Relax	.2279	.2311	.2510	.2642	.2556	.2138	.2336	.1856	.2214	.2235
Baseline*	.0870	.0911	.1552	.1432	.1150	.0669	.0563	.0641	.0835	.0442
Track 2 (week level)										
Team GroundTruth*	**.3236**	**.3305**	.3708	.3502	**.3323**	.2774	**.2924**	**.2506**	**.3162**	.4542
Team heboyong*	.3226	.3301	.3691	.3548	.3279	**.2827**	.2856	.2435	.3112	.4662
Team Hby	.3218	.3296	.3722	.3556	.3256	.2806	.2818	.2432	.3067	**.4714**
Team PZH	.3087	.3156	**.3999**	**.3588**	.3212	.2596	.2744	.2502	.3013	.3592
Team BDD	.3007	.3072	.3557	.3367	.3141	.2562	.2735	.2338	.2936	.3942
Baseline*	.1585	.1669	.2960	.2554	.2014	.1228	.0982	.1043	.1454	.1118
Track 3 (month level)										
Team GroundTruth*	**.3376**	**.3464**	**.4142**	**.3729**	**.3414**	**.3032**	**.2933**	**.2567**	.3112	**.4779**
Team heboyong*	.3241	.3316	.3671	.3538	.3289	.2838	.2864	.2458	**.3132**	.4735
Team BDD	.3121	.3186	.3681	.3445	.3248	.2680	.2843	.2450	.3062	.4076
Team PZH	.3087	.3156	.3999	.3588	.3212	.2596	.2744	.2502	.3013	.3592
Team BingDwenDwen	.2986	.3054	.3565	.3477	.3241	.2702	.2707	.2337	.2808	.3598
Baseline*	.1964	.2033	.3068	.2849	.2044	.1559	.1535	.1441	.1944	.1827

Top solutions are highlighted in bold, and solutions that passed the "code verification stage" are marked with a *.

4.1 The Leaderboard

The leaderboards at the test phase for the different tracks are shown in Table 2. Note that we only show here the top-5 solutions (per track), in addition to the baseline results. Top solutions that passed the "code verification stage" are highlighted in bold. The full leaderbord of each track can be found in the respective Codalab competition webpage.

As expected, Table 2 shows that overall better results are obtained with more train data. That is, a model trained at the month level is overall more accurate than the same model trained at the week level, which is overall more accurate than the one trained at the day level. Therefore, the differences in performance improvement when training the model at the month level (compared to week level) are smaller than those obtained when training the model at the week level (compared to day level), particularly when a large shift in time is observed (e.g., from Jun. to Sep.), suggesting that the increase of train data from week to month level may have a small impact when large shifts are observed. This was

also observed by the *Team heboyong* (described in Sec. 4.3), which reported to have only used week level data to train their model (i.e., on Tracks 2 and 3), based on the observation that using more data was not improving the final result. This raises an interesting point in that even for winning approaches the variation of the training data is much more important than the amount of training data, a further analysis of what causes the loss of mAP across will be discussed in 4.4.

Table 3 shows some general information about the top winning approaches. As it can be seen from Table 3, common strategies employed by top-winning solutions are the use of pre-trained models combined with data augmentation. Next, we briefly introduce the top-winning solutions that passed the code verification stage based on the information provided by the authors. For a detailed information, we refer the reader to the associated fact sheets, available for download in the challenge webpage5[0]. Two participants (i.e., *Team GroundTruth* and *Team heboyong*) ranked best on all tracks. Each participant applied the same method on all tracks, but trained at day, week or month level, detailed as follows.

Table 3: General information about the top winning approaches.

	Top-1 *Team GroundTruth*	**Top-2** *Team heboyong*
Pre-trained model	✓	✓
External data	✗	✗
Data augmentation	✓	✓
Use of the provided validation set as part of the training set at the final phase	✗	✗
Handcrafted features	✗	✗
Spatio-temporal feature extraction	✗	✗
Object tracking	✗	✗
Leverage timestamp information	✗	✗
Use of empty frames present in the dataset	✗	✗
Construct any type of prior to condition for visual variety	✗	✗

4.2 Top-1: *Team GroundTruth*

The *Team GroundTruth* proposed to take benefit of temporal and contextual information to improve object detection performance. Based on Scaled-YOLOv4 [26], they first perform sparse sampling at the input. The best sampling setting is defined based on experiments given different sampling methods (i.e., average sampling, random sampling, and active sampling). Mosaic [1] data augmentation is then used to improve the detector's recognition ability and robustness to small objects. To obtain a more accurate and robust model at inference stage, they adopt Model Soups [27] for model integration, given the results obtained by Scaled-YOLOv4p6 and Scaled-YOLOv4p7 detectors trained using different hyperparameters, also combined with horizontal flip data augmentation to further improve the detection performance. Given a video sequence of region proposals and their corresponding class scores, Seq-NMS [12] associates bounding boxes in adjacent frames using a simple overlap criterion. It then selects boxes to maximize a sequence score. Those boxes are used to suppress overlapping

boxes in their respective frames and are subsequently re-scored to boost weaker detections. Thus, Seq-NMS [12] is applied as post-processing to improve the performance further. An overview of the proposed pipeline is illustrated in Figure 4.

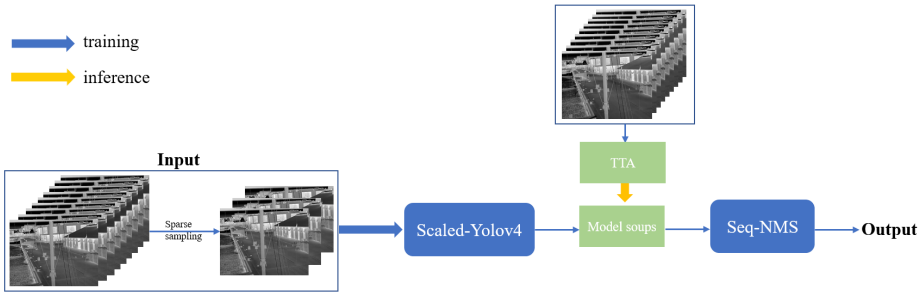


Fig. 4: Top-1 winning solution pipeline: *Team GroundTruth*.

4.3 Top-2: *Team heboyong*

The *Team heboyong* employed Cascade RCNN [4], a two-stage object detection algorithm, as the main architecture for object detection, with Swin Transformer [20] as backbone. According to the authors, Swin Transformer gives better results when compared with other CNN-based backbones. CBNetv2 [18] is used to enhance the Swin Transformer to further improve accuracy. MMdetection [5] is adopted as the main framework. During training, only 30% of the train data is randomly sampled, to reduce overfitting, combined with different data augmentation methods, such as Large Scale Jitter, Random Crop, MixUp [31], Albu Augmentation [3] and CopyPaste [11]. At inference stage, they use Soft-NMS [2] and flip augmentation to further enhance the results. An overview of the proposed pipeline is illustrated in Figure 5. They also reported to have not addressed well the long-tail problem caused by the extreme sparsity of the bicycle and motorcycle categories, which resulted in low mAP for these two categories.

4.4 What challenge the models the most?

In this section we analyze the performance of the baseline, Team GroundTruths and Team heboyongs models on the test set. Particularly, we inspect the performance of each model with regards to temperature, humidity object area and object density. Temperature and humidity are chosen as they were discovered that these two factors have the highest correlation with visual concept drift [21]. Additionally, because of the uneven distribution of object densities across dataset, the impact of the object density is also investigated.

Impact of temperature. can be observed in Figure 7, as the temperature increases the performance of the model degrades. This is expected as the available

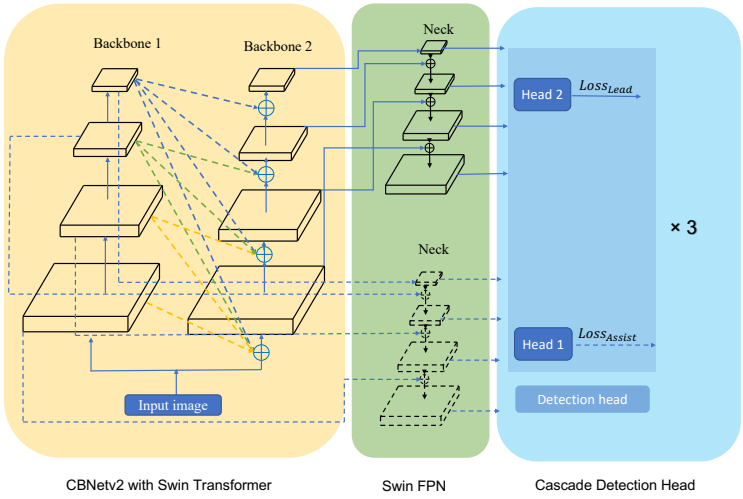


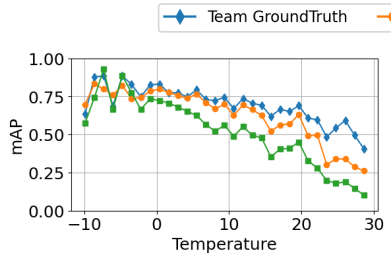
Fig. 5: Top-2 winning solution pipeline: *Team heboyong*.

training data has been picked from the coldest month and as such warmer scenes are not properly represented in the training data, and as mentioned in 3 this is deliberately done as temperature is one of the most impactfull factors of concept drift in thermal images [21]. The performance of the baseline model shows severe degradation when compared to the winner and *Team heboyong*, while the performance consistently degrades for all models. Interestingly, *Team heboyong* method is distinctly more sensitive to concept drift with the smaller training set, while the winning solutions seems to perform consistently regardless of the amount of data trained on.

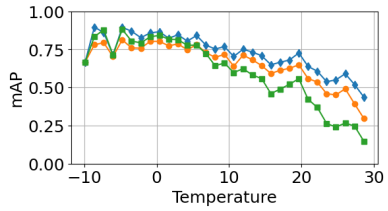
Impact of humidity. According to the initial paper [21], humidity is one of the most impactfull factors of concept drift, as it tends to correlate positively with the different types of weather. This leads to a quite interesting observation, which can be made across all tracks with regards to the impact of humidity. As can be observed in Figure 8, the mAP of detectors increases with the humidity across all tracks. This could be because higher humidity tends to correlate with the level of rain-clouds, which would explain partially cloudy being more difficult for the detectors as the visual appearance in the image is less uniform.

Impact of object size. As would be expected the models converge towards fitting bounding-boxes to the most dominant object size of the training data (see Table 1). As shown in Figure 10, the models obtain very good performance on the most common of object sizes and struggle with objects as they increase in size and rarity. In this case the participants see strong improvement over baseline, and also manage to become more robust towards rarer cases. As can

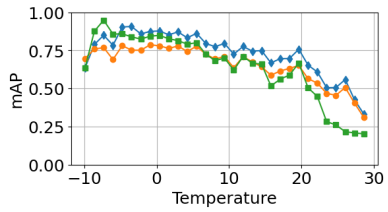
also be observed in the figure this problem is increasingly alleviated with the increase of training data.



(a) Track 1: Day

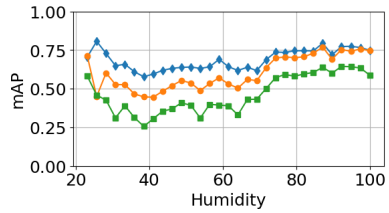


(b) Track 2: Week

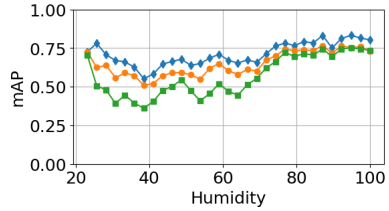


(c) Track 3: Month

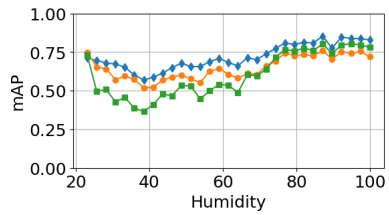
Fig. 7: Overview of performance with samples separated with regards to the temperature recorded for the given frame.



(a) Track 1: Day



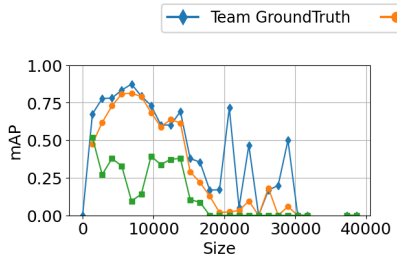
(b) Track 2: Week



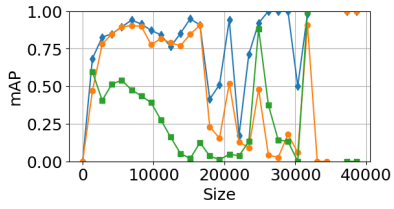
(c) Track 3: Month

Fig. 8: Overview of performance with samples separated with regards to the humidity recorded for the given frame.

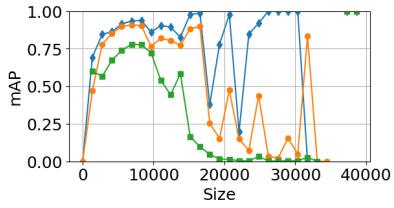
Impact of object density. As shown in Figure 2, the density of objects for the majority of the images is towards the lower end, as such one would expect the detectors' mAP to degrade when a scene becomes more crowded and the individual objects become more difficult to detect due to occlusions. However what is observed is the mAP of highlighted methods are consistent as density increases, while the performance across densities also correlate to the amount of training data.



(a) Track 1: Day

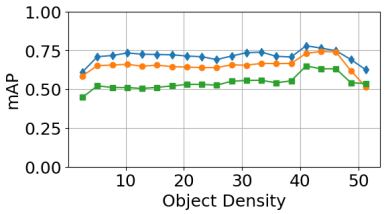


(b) Track 2: Week

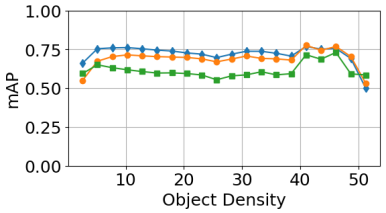


(c) Track 3: Month

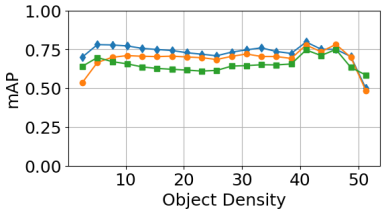
Fig. 10: Overview of performance with samples separated with regards the size of objects bounding-box



(a) Track 1: Day



(b) Track 2: Week



(c) Track 3: Month

Fig. 11: Overview of performance with samples separated with regards to the object density of the frame

5 Conclusions

The Seasons in Drift challenge attracted over 180 participants whom made 480 submissions during validation and 211 submissions for test set and a potential place on the finale leaderboard. While the concept of measuring the impact of thermal drift on detection performance in a security monitoring context is a very understudied field, a lot of people participated. Many of the participants managed to beat the proposed baseline by quite a large margin, especially with limited training data, and achieved more robust solutions when compared to the degradation of the baseline in terms of performance with respect to drift. Allthough great improvements can be observed, the problem of concept drift still negatively affects the performance of participating methods. Interestingly while the winner and *Team heboyong* methods use different architectures, the impact of

concept drift seems to transcend the choice of SotA object detectors. This lends merit investigating methods that could condition layers of the network given the input image, and introduce a venue for the model to learn an adaptable approach as opposed to learning a generalized model specific to the thermal conditions of the training context. As can be observed in Figures 10 and 11 the size of the observed objects seem to be a more challenging factor than the density of which they occour in. Detection of small objects is a known and well documented problem, and despite the nature of thermal cameras, still persist as an issue in the thermal domain. Further research could be done to learn more scale invariant object detectors or rely entirely on other methods than an RPN or Anchors to produce object proposals.

Acknowledgements

This work has been partially supported by Milestone Research Program at AAU, the Spanish project PID2019-105093GB-I00 and by ICREA under the ICREA Academia programme.

References

1. Bochkovskiy, A., Wang, C., Liao, H.M.: Yolov4: Optimal speed and accuracy of object detection. CoRR **abs/2004.10934** (2020)
2. Bodla, N., Singh, B., Chellappa, R., Davis, L.S.: Soft-NMS - improving object detection with one line of code. In: ICCV (2017)
3. Buslaev, A.V., Iglovikov, V.I., Khvedchenya, E., Parinov, A., Druzhinin, M., Kalinin, A.A.: Albumentations: Fast and flexible image augmentations. Information **11**(2) (2020)
4. Cai, Z., Vasconcelos, N.: Cascade R-CNN: Delving into high quality object detection. In: CVPR (2018)
5. Chen, K., Wang, J., Pang, J., Cao, Y., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Xu, J., Zhang, Z., Cheng, D., Zhu, C., Cheng, T., Zhao, Q., Li, B., Lu, X., Zhu, R., Wu, Y., Dai, J., Wang, J., Shi, J., Ouyang, W., Loy, C.C., Lin, D.: MMDetection: Open MMLab detection toolbox and benchmark. CoRR **abs/1906.07155** (2019)
6. Chen, Y.Y., Jhong, S.Y., Li, G.Y., Chen, P.H.: Thermal-based pedestrian detection using faster r-cnn and region decomposition branch. In: ISPACS (2019)
7. Dai, D., Van Gool, L.: Dark model adaptation: Semantic image segmentation from daytime to nighttime. In: ITSC (2018)
8. Dai, X., Yuan, X., Wei, X.: Tirnet: Object detection in thermal infrared images for autonomous driving. Applied Intelligence **51**(3) (2021)
9. Davis, J.W., Keck, M.A.: A two-stage template approach to person detection in thermal imagery. In: WACV-W (2005)
10. Devaguptapu, C., Akolekar, N., M Sharma, M., N Balasubramanian, V.: Borrow from anywhere: Pseudo multi-modal object detection in thermal imagery. In: CVPR-W (2019)
11. Ghiasi, G., Cui, Y., Srinivas, A., Qian, R., Lin, T.Y., Cubuk, E.D., Le, Q.V., Zoph, B.: Simple Copy-Paste is a strong data augmentation method for instance segmentation. In: CVPR (2021)

12. Han, W., Khorrami, P., Paine, T.L., Ramachandran, P., Babaeizadeh, M., Shi, H., Li, J., Yan, S., Huang, T.S.: Seq-NMS for video object detection. CoRR **abs/1602.08465** (2016)
13. Hwang, S., Park, J., Kim, N., Choi, Y., So Kweon, I.: Multispectral pedestrian detection: Benchmark dataset and baseline. In: CVPR (2015)
14. Jia, X., Zhu, C., Li, M., Tang, W., Zhou, W.: Llvip: A visible-infrared paired dataset for low-light vision. In: ICCV (2021)
15. Kieu, M., Bagdanov, A.D., Bertini, M., Bimbo, A.d.: Task-conditioned domain adaptation for pedestrian detection in thermal imagery. In: ECCV (2020)
16. Kim, J., Kim, H., Kim, T., Kim, N., Choi, Y.: Mlpd: Multi-label pedestrian detector in multispectral domain. Robotics and Automation Letters **6**(4) (2021)
17. Krišto, M., Ivasic-Kos, M., Pobar, M.: Thermal object detection in difficult weather conditions using yolo. IEEE access **8** (2020)
18. Liang, T., Chu, X., Liu, Y., Wang, Y., Tang, Z., Chu, W., Chen, J., Ling, H.: Cbnetv2: A composite backbone network architecture for object detection. CoRR **abs/2107.00420** (2021)
19. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV (2014)
20. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: ICCV (2021)
21. Nikolov, I., Philipsen, M., Liu, J., Dueholm, J., Johansen, A., Nasrollahi, K., Moeslund, T.: Seasons in drift: A long-term thermal imaging dataset for studying concept drift. In: NeurIPS (2021)
22. Pavao, A., Guyon, I., Letournel, A.C., Baró, X., Escalante, H., Escalera, S., Thomas, T., Xu, Z.: CodaLab Competitions: An open source platform to organize scientific challenges. Ph.D. thesis, Université Paris-Saclay, FRA. (2022)
23. Siris, A., Jiao, J., Tam, G.K., Xie, X., Lau, R.W.: Scene context-aware salient object detection. In: ICCV (2021)
24. Telodyne: FLIR AADAS Dataset, `https://www.flir.com/oem/adas/adas-dataset-form/`
25. Vertens, J., Zürn, J., Burgard, W.: Heatnet: Bridging the day-night domain gap in semantic segmentation with thermal images. In: IROS (2020)
26. Wang, C., Bochkovskiy, A., Liao, H.: Scaled-YOLOv4: Scaling cross stage partial network. In: CVPR (2021)
27. Wortsman, M., Ilharco, G., Gadre, S.Y., Roelofs, R., Gontijo-Lopes, R., Morcos, A.S., Namkoong, H., Farhadi, A., Carmon, Y., Kornblith, S., Schmidt, L.: Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. CoRR **abs/2203.05482** (2022)
28. Wu, X., Wu, Z., Guo, H., Ju, L., Wang, S.: Dannet: A one-stage domain adaptation network for unsupervised nighttime semantic segmentation. In: CVPR (2021)
29. Zhang, H., Fromont, E., Lefevre, S., Avignon, B.: Multispectral fusion for object detection with cyclic fuse-and-refine blocks. In: ICIP (2020)
30. Zhang, H., Fromont, E., Lefèvre, S., Avignon, B.: Guided attentive feature fusion for multispectral pedestrian detection. In: WACV (2021)
31. Zhang, H., Cissé, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. CoRR **abs/1710.09412** (2017)