Who Cares about the Weather? Inferring Weather Conditions for Weather-Aware Object Detection in Thermal Images

Anders Skaarup Johansen * , Kamal Nasrollahi , Sergio Escalera , Thomas B. Moeslund

Posted Date: 7 September 2023

doi: 10.20944/preprints202309.0499.v1

Keywords: thermal; object detection; conditioning; weather-aware

Article

Who Cares about the Weather?

Inferring Weather Conditions for Weather-Aware Object Detection in Thermal Images

Anders Skaarup Johansen [1,†*], Kamal Nasrollahi [1,2], Sergio Escalera [1,3] and Thomas B. Moeslund [1]

[1] Aalborg University; asjo, kn@create.aau.dk
[2] Milestone Systems;
[3] Universitat de Barcelona and Computer-Vision Center; sescalera@ub.edu
* Correspondence: Asjo@create.aau.dk
† Current address: Rendsburggade 14, DK-9000, Aalborg
‡ These authors contributed equally to this work.

Featured Application: This work focuses on achieving a weather-agnostic approach for real-world deployment of object-recognition algorithms. Particularly object recognition in thermal data exposed to long-term thermal drift.

Abstract: Deployments of real-world object-detection systems often experience a degradation in performance over time due to concept drift. Systems that leverage thermal cameras are especially susceptible because the respective thermal signatures of objects and their surroundings are highly sensitive to environmental changes. In this study, a conditioning method is investigated. The method aims to guide the training loop of thermal object detection systems by leveraging an auxiliary branch to predict the weather, while directly or indirectly conditioning the baseline detection system. Leveraging such an approach to train detection networks does not necessarily improve the performance of native architectures, however, it can be observed that conditioned networks manage to extract a signal from thermal images that guides the network to detect objects that baseline models miss. As the extracted signal appears to be quite noisy and very challenging to regress accurately, further work is needed to identify an ideal optimization vector.

Keywords: thermal; object detection; concept drift; conditioning; weather recognition

1. Introduction

Deploying thermal image-recognition deep-learning models for long-term analysis of a scene becomes increasingly difficult over time due to concept drift. Not only does the visual signature of the scene and objects within it change with seasons, but it also changes significantly between day and night. Thermal concept drift is an increasingly researched topic [1,2], and is crucial for real-world deployments of thermal vision systems. Typically, the aim is focused on identifying distinct concept-drift factors or assuming the presence of distinct distributions. Establishing effective methods to combat concept drift is a vital component when deploying computer vision systems in real-world environments. Traditional evaluation methods do not provide accurate descriptions of the impact concept drift has on performance during long-term deployments[1]. Changes in contextual parameters, such as weather conditions, can be somewhat related to the degradation of performance observed with long-term concept drift [1]. For example, the relationship between the degradation of object detection models and changes in temperature and humidity is statistically significant [1]. As weather conditions are somewhat correlated with changes in the visual appearance of captured thermal footage, they could be leveraged to guide the model toward learning two different representations: one that varies with the weather (weather-aware), and one that doesn't (weather-agnostic).

Multi-task learning has become an increasingly popular method for training generalized image-recognition models [2–7], but it mostly focuses on using auxiliary branches that are somewhat task-adjacent, where an intuitive connection to the primary task can be drawn. Each task contributes to converting the latent representation into a more generalized representation, which often increases performance for all tasks[4,6]. Given that the signals induced by the auxiliary tasks help to achieve a more robust representation, similar approaches could be leveraged to extract a contextually aware signal through auxiliary conditioning.

1.1. Estimating weather

Directly leveraging weather information would require a vision system to directly infer weather conditions from the captured data [8,9]. By treating it as a classification problem, deep learning methods have shown great promise at classifying categories of weather [8–10]. This shows a weather signal can be somewhat extracted from single images and categorized into distinct classes. Most weather classification approaches focus on binary classification of distinct weather conditions (i.e. cloudy, sunny, raining, etc.) and lack the granularity observed during long-term deployment. To address this, datasets like RFS [11] and MWD [12] propose treating it as a multi-label classification problem, to capture the ambiguity between different weather phenomena and transitive weather conditions[11,12]. When estimating weather conditions from a single image, all regions are not created equal [12,13]: thus some methods isolate predetermined regions, such as the sky[9,14], or leverage region-proposal networks [12,13], to extract region specific features.

1.2. Adapting to weather

Adverse weather conditions, particularly those that are not present in the training dataset, present a real challenge for deployed computer vision systems that are exposed to the weather. Typically, approaches to concept drift rely on detecting drift, then adapting accordingly [15,16]. With deep-learning based approaches, this typically means a general system will be trained to establish a baseline, then subsequently be exposed to unseen data. Depending on the task, an evaluation metric will be used as a method to detect drift [17]. When adapting to weather-related drift, some have tried to remove the distracting elements directly [18–22], train weather-agnostic models by simulating various weather conditions and including that in the training loop[23–26] or train several models in an ensemble and leveraging a weighted approach to determine the final prediction[27–30]. Moreover, in situations where an unsatisfactory amount of variation can be captured in the training data, continual-learning [16,31] or domain-adaptation [15,16] approaches are often leveraged in an attempt to obtain consistent performance as the visual appearance of the context changes.

1.3. Leveraging metadata for recognition

In recent years, including auxiliary optimization tasks has been shown to greatly improve the performance of the downstream task, whether used as a pre-text task (as often seen with vision transformers [32–35]), or jointly optimized with the downstream task [36,37]. Using auxiliary tasks to guide a primary task by introducing aspects that cannot be properly captured in the downstream task's optimization objective, has shown great promise in improving the performance and generalization of a downstream task [2]. Depending on the model architecture and desired purpose of this weather-conditioned representation, it can be leveraged as a constraining parameter to enforce the inclusion of the auxiliary representation directly [2], thereby forcing the network to adjust to being aware of the contextual information induced. Alternatively, the auxiliary representation could be seen as purely supplemental information, which potentially consists of redundant elements, and, as such, should only be leveraged to indirectly guide the network.

 doi:10.20944/preprints202309.0499.v1

1.4. Qualitative vs. Quantitative thermal cameras

Thermal cameras work by capturing the amount of infrared radiation emitted by objects within a scene. Though they all aim to capture the same type of information (namely heat), there are two types of thermal cameras. First is qualitative thermography (sometimes referred to as "relative thermal imaging"), where the goal is to show the relative differences of infrared radiation throughout the camera's field of view. These are often used for inspection and security purposes, as they often provide distinct contrast between colder and hotter elements in the field of view, regardless of absolute temperature. Second is quantitative thermography (sometimes refered to as "absolute thermal imaging"), where each sampling point in the field of view is mapped to an absolute temperature measurement. This enables accurate capture of absolute thermal differences between elements in the scene, and consistent visual response for any thermal signature. In recent years the advances in thermal imaging technology have made the use of thermal cameras increasingly popular, either in isolation or in conjunction with traditional CCTV-cameras. Quantitative thermal cameras could be seen as the ideal solution, as they provide essentially the same functionality as qualitative thermal cameras, but with the added benefit of accurate thermal readings. The technology required to construct an absolute thermograph is significantly more complicated than that of a relative thermography, and, as such, are much more costly to produce and purchase. For the purpose of many tasks the absolute temperature readings are redundant for the purpose of the thermal camera, and as such do not justify the cost, thus making qualitative thermal cameras much more commmon in deployed vision systems.

In this paper, a methodology is detailed, which employs predicting continuous weather-related meta-variables for auxiliary guidance, and provides an overview of the Long-term Thermal Drift (LTD) which contains both object-centric annotations as well as fine-grained weather information for each sample. Further methods describe how fine-grained weather prediction can be leveraged to condition the network during training to guide the network to become weather-aware. Particularly this will be divided into direct- and indirect-conditioning methods. Finally, this is followed by a discussion of extensive experiments, which evaluate the impact of the aforementioned methodology (conditioned on temperature, humidity, time-of-day), and the impact on performance metrics with respect to the respective weather conditions. While the analysis does not show a direct improvement in accuracy metrics, it does show that auxiliary conditioning in this way allows the networks to extract and somewhat model the underlying weather signal.

2. Methodology

While more fluid prediction schemes have become available for weather estimation, the methods for predicting weather conditions found in the literature are still predominantly done using a binary scheme. Using a binary scheme as a conditioning method assumes that there is a fixed amount of distribution to the model. In an uncontrolled environment, this is a potentially false assumption as unknown variables could introduce noise to the signal that would make it difficult to distinguish ground truth close to the bin edges [38,39]. This is potentially further exacerbated when processing thermal video from cameras with a relative internal thermograph. As detailed in Section 1.4, the prevalence of relative thermal cameras makes it a promising modality to investigate, particularly for a real-world context.

To our knowledge, the only existing work that performs auxiliary conditioning for task-specific improvements is presented in [2]. The authors leverage a direct-conditioning approach (detailed in Section 2.3) on the KAIST Multispectral Pedestrian Detection (KAIST) dataset, and manage to achieve a decrease in Miss-Rate (MR). However, the KAIST dataset contains thermal images from an absolute thermal camera, resulting in a fairly similar thermal signature to that of pedestrians (as seen in Figure 1).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 September 2023 doi:10.20944/preprints202309.0499.v1

(a) Day **(b)** Night **(c)** Day **(d)** Night

Figure 1. Examples of thermal images in the KAIST dataset, where a similar thermal signature of people can be observed at different times of the day, due to the use of quantitative thermography as well as the limited periods of captured data.

In this section an overview of the object-centric annotations of the LTD-Dataset and the associated meta-data are detailed. Furthermore, the method proposed in [2] can be adapted for prediction of a continuous auxiliary variable is described. Finally, the architecture of a direct-conditioning approach (similar to [2]) as well as an indirect-conditioning approach, using a State of the Art (SotA) transformer-based model is detailed.

2.1. Dataset

In the original LTD dataset benchmark, [1] and the subsequent challenge[40] the performance impact concept drift has on object detectors is correlated with the absolute change in mean Average Precision (mAP) across different concept drift related meta variables (most notably: temperature, humidity, time of day). Subsequently, the dataset has been extended with additional object-centric annotations. The dataset was uniformly sampled with a .5 frames per second sample rate, resulting in over 900.000 images with over 6.000.000 annotated objects.

As can be seen in Figure 2, the thermal signature of people varies significantly more in the LTD dataset, as does the contrast between objects and background.

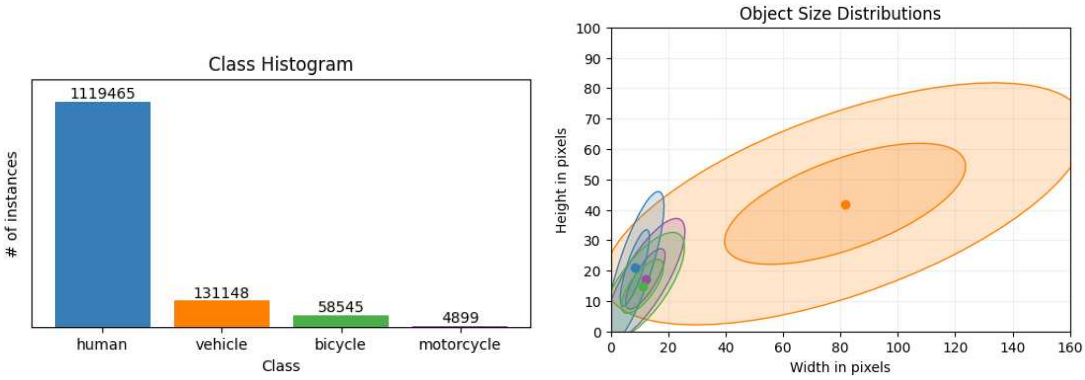
(a) Day **(b)** Night **(c)** Day **(d)** Night

Figure 2. Examples of images containing people from the LTD Dataset. Where, drastically different thermal signatures for objects can be observed, due to the use of qualitative thermography, as well as the dataset spanning 9 months.

The objects are represented as four classes, namely; namely; person, bicycle, motorcycle and vehicle. The LTD dataset is captured in real-world, unconstrained context, and is thus susceptible to associated biases, such as: skewed object distribution (As seen in Figure 3a), frames without objects of interest, highly varied object densities, and uneven distribution of weather conditions (As seen in Figures 4a to 4c). Furthermore, sizes of objects are affected by the camera being suspended roughly 6 meters above the ground and aimed downwards, which results in most objects being small (As seen in Figures 3b and A1). However, this might be expected for deployment in a real-world security context.

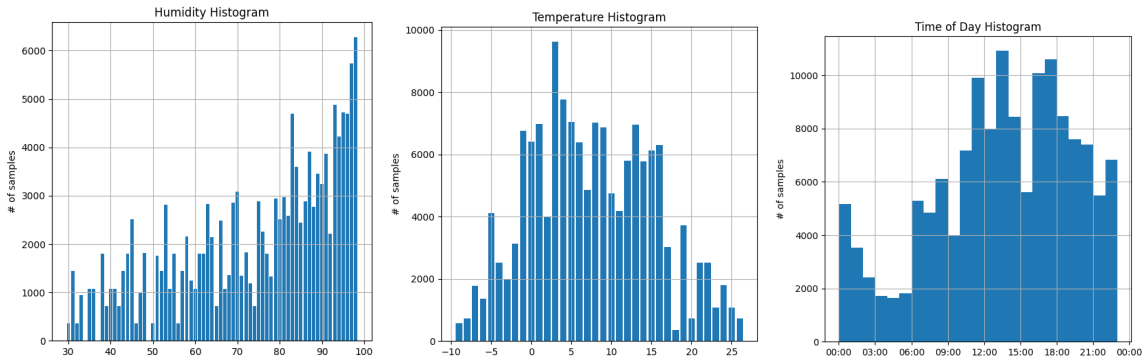
Furthermore, as shown in Figure 3b, while each class has its own unique distribution, the distributions predominantly contain very small objects, with an exception of the vehicle class. This adds an additional degree of difficulty as most object detectors tend to struggle with smaller objects [32,41,42]. Additionally, a heatmap with absolute counts of object sizes can be found in Figure A1.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 September 2023 doi:10.20944/preprints202309.0499.v1



(a) **(b)**

Figure 3. On 3a the total amount of instances from each of the given classes can be observed, while on 3b the mean object size can be seen as a dot with additional rings drawn at 1, 2 standard deviations respectively



(a) **(b)** **(c)**

Figure 4. Histograms showing the distribution of meta-variables across the entire dataset

2.2. From discrete to continuous meta-prediction

In the KAIST dataset [43], the data falls into two distinct categories, daytime and nighttime. However, in a real-world deployment the system would observe a gradual change between daytime and nighttime, which is not accurately represented by such a binary grouping. However, in the LTD dataset [1] each clip has an extensive, highly granular set of metadata. This allows us to evaluate the impact of auxiliary task-conditioning in a real-world with more diverse samples.

In [2] the authors propose guiding the conditioning branch with a binary classification head (classifying day or night). However, to perform fine-grained, continuous weather prediction, the auxiliary optimization task and loss must be adjusted accordingly. The problem with binary classification is that it treats all false positives equally, regardless of the magnitude of the error. For continuous classifications however, a severity of the misclassification can be assessed by determining the absolute difference between the prediction and the ground-truth. Naively, an $L1$-loss can be used to punish/reward the network, based on the difference in absolute distance. However, due to the data being captured by a relative thermal camera, identical visual appearance cannot be guaranteed between calibrations. During capture of the data for the LTD dataset the camera would routinely undergo automatic calibration, resulting in an inconsistent profile over time. This induces a noise signal which could result in the optimization converging towards a global mean rather than an acceptable guess. We combat this by employing an exponential L1 tuned to allow a pre-determined degree of deviation, before approaching the values of the primary task loss or losses.

$$L1_e(x,y) \equiv L \equiv \{l_1, ..., l_N\}, l_n \equiv |x_n - y_n|^{\frac{k}{|x_n - y_n|}} \tag{1}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 September 2023 doi:10.20944/preprints202309.0499.v1

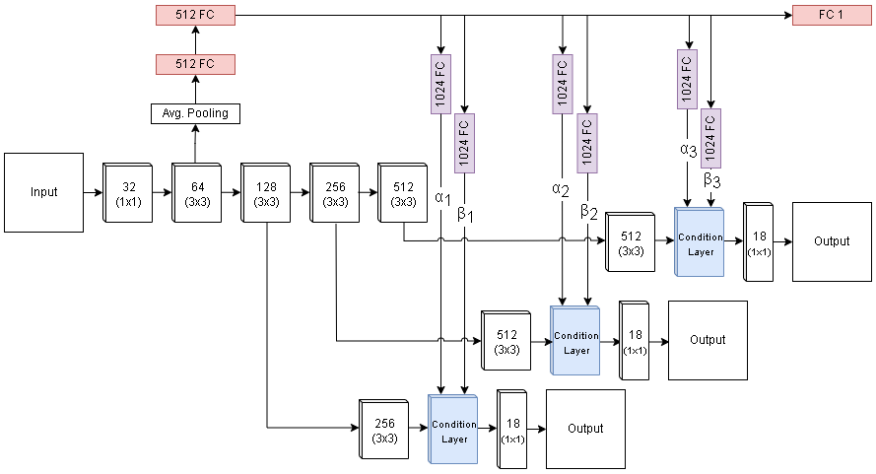
When establishing baseline performance for each model, the minimum, maximum and standard deviation of the primary loss was noted down for the final epoch. A corresponding k that would approach the expected loss values of the primary task was selected. Where, the resulting weighting of the auxiliary in the optimization process would be approximately equal to that of the primary task at the border of the desired deviation k, while exponentially increasing when deviating further from the allowed k

Due to the thermal images of the LTD dataset being recorded with a relative thermal camera, the visual appearance of a scene might be slightly different, even under similar meta-conditions. Thus, predicting exact values from visual data would be an ill-posed problem, as any given state inherits some degree of variance from the calibration of the relative thermograph.

2.3. Directly conditioning

In [2] the authors propose a method of directly conditioning the latent representation of each predictive branch through a conditioning layer. The conditioning element is part of an auxiliary classification network, which is aimed at predicting whether the given sample belongs to the daytime distribution or the nighttime distribution. The latent representation used for this auxiliary prediction is derived from an intermediate representation of the entire image. Thus, the representation must be able to extract a notion of day and night, which can make the network 'aware' and adapt accordingly.

In the proposed method, the overall mAP, does not improve significantly at higher Intersection over Union (IoU)s, however the weather-conditioned network shows reduction in object MR. By directly conditioning the intermediate representation, the network is forced to directly incorporate the weather information in its semantically rich representation. We employ the original implementation on the YOLOv5 model. As shown in Figure 5a the standard YOLO architecture is extended with an auxiliary branch, extracted from one of the early stages of the feature extractor. Subsequently, a series of fully connected layers condense the representation to feed it to a prediction head. the prediction head produces a single value that is regressed following the exponential L1 loss described in Section 2.2. Individual fully-connected layers feed the representation to the conditioning layer in the different stages of the network, prior to the given stage's prediction head. The conditioning layer (shown in Figure 5b), takes in the a set of feature maps and an element-wise multiplication and summation with separated auxiliary representations α_n and β_n respectively.



(a) YOLOv5 with conditioning branch

(b) Conditioning layer

Figure 5. YOLO-styled conditioning network (Figure 5) and internals of conditioning layer (Figure 5b). Red, purple and blue denote auxiliary branch-, mapping-, and conditioning layers respectively.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 September 2023 doi:10.20944/preprints202309.0499.v1

2.4. Indirectly imposed conditioning

Vision-Transformers have proven to effectively leverage global reasoning to solve various vision tasks. By calculating an all-to-all affinity mapping, i.e. self-attention, between input elements, i.e. tokens, transformers can effectively relate elements, even when they belong to separate modalities. For classification this is often employed with an additional learnable element, which is then mapped to a prediction head. The repeated self-attention allows the classification token to extract information from the entire input without directly imposing changes to other inter-token relationships. DETR [32,41,44] is a common state-of-the-art transformer-based object detector. Subsequent variants have shown to greatly improve convergence and stability of optimization [41,44], by extending the deformable-DETR[44] with a learnable classification token, and using the encoding of the classification token to perform prediction of the auxiliary task, namely weather condition prediction. While the optimization could potentially drive the transformer to learn embeddings that are optimized towards affinity with the classification token, the network should be able to disregard weather-related embeddings in cases where weather does not provide any significant optimization benefit. Unlike the directly-imposed approach, the network could learn to dynamically disregard regions of the image that do not provide contextual information.

Inspired by the use of a [CLS] token in the original BERT [45] paper, we include an additional token with every input sample, which is propagated through every encoder layer of the transformer. This way global information from a given sample can be continuously aggregated in a single representation. Prior to reaching the decoder layers, the [CLS] token is separated and passed to an auxiliary branch (as seen in Figure 6). The auxiliary branch consists of a series of fully connected layers (sizes; 512, 512, 1), acting as a mapping from latent-representation to a single value that can be regressed.

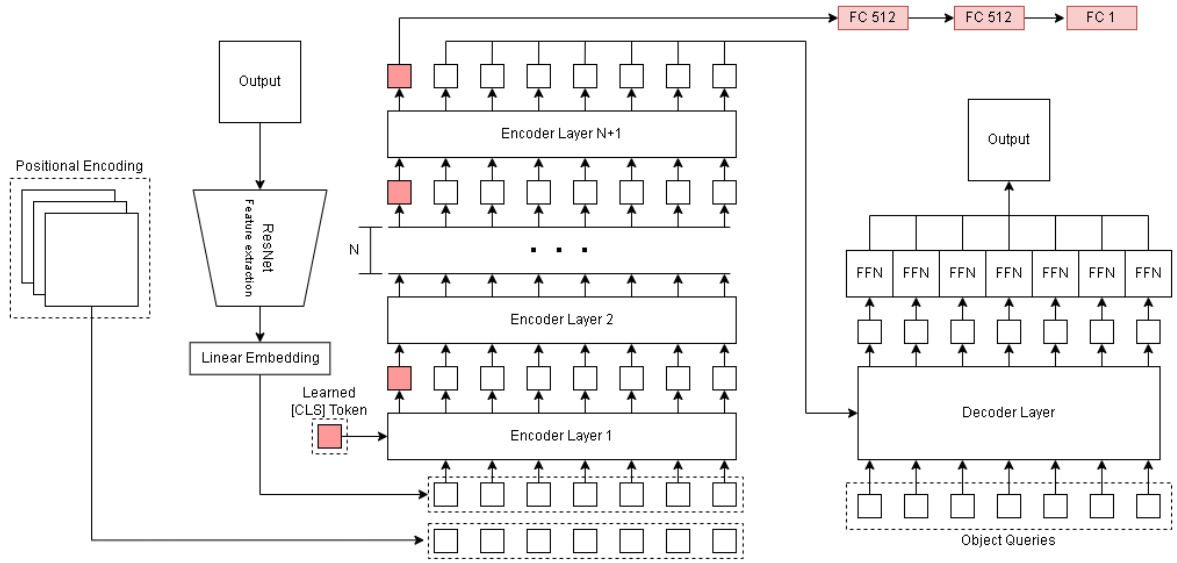


Figure 6. DETR-style transformer network with indirect conditioning. Red modules and tokens denote components used for auxiliary prediction

3. Results

3.1. Experimental setting

To establish a baseline the models were trained as described in their respective papers and implementations. Since none of the models contained a thermal variant, they were all trained from scratch, which required increased training time in order to expect convergence, as "standard" configurations implemented loading an image-net pretrained feature extraction network. For this reason we set the maximum allowed epochs to 250 for all models. The batch size for all models was

 doi:10.20944/preprints202309.0499.v1

also set to 8 per GPU, resulting in 3.8 total iterations. All models were trained in the same pytorch environment (torch 1.7, torchvision 0.8.1), on two Nvidia RTX 3090 cards. Class-wise losses were weighted to reflect the frequency of each class in their respective subsets. The complete dataset was split into 3 roughly equal parts. Two were employed for training and validation, and the third test set remained hidden to allow for future challenges similar to [40]. Further information data availability is described in the 'Data Availability section at the end of the paper.

3.2. Evaluating weather conditioning

Due to the auxiliary branch being trained in a supervised manner, it has to be exposed to the variety observed in the training set. As such all clips in the dataset were evenly distributed across equally-sized training, test, and validation sets. Because this potentially allows a naive approach to generalize easily, due to the inclusion of the full variation present in the dataset, the proposed method is compared to an equally trained naive approach, without the auxiliary meta-prediction branch.

To evaluate the potential impact of each of the three meta-conditions (namely, temperature, humidity and time of day), each model was trained naively (i.e. according to the training loop described in the respective paper [2,44]) as well as with the auxiliary conditioning branch (direct- and indirect- conditioning for YOLO- and DETR-variants, respectively). To allow for fair optimization each model is trained for the same amount of epochs as their respective baseline.

Likewise we observe the performance when compared to temperature and object size, to investigate if any categories potentially suffer, in order to reach a more general improvement of the system. While these correlations might not be intuitively tied to weather, the latent representation learned could inadvertently favor certain aspects of the object distributions.

Because conditions are quantified using different metrics, the ground truth ranges vary significantly. To normalize their representation, the values are remapped so that the observed values fall roughly within the range $[-2, 2]$. This range is chosen to avoid the network having to also learn a mapping between arbitrary ranges, while keeping in line with the normalization done internally in the networks, which is done to avoid unstable variances in the activations [32,33,42].

3.3. Accuracy

Table 1 details the overall mAP and MR for all of their models across validation set. This is used as a metric of overall object detection performance similarly to what is commonly done for other object detection datasets, and to retain a fair comparison with the original LTD dataset evaluation [1]. Additionally Table 2 details the Mean Average Error (MAE) of auxiliary prediction branch, as well as the Standard Deviation (Std.) of the the prediction error. This is listed to provide insight into the performance of the auxiliary branch.

As can be seen in Table 1 the baseline models which are naively trained without any auxiliary guidance tend to perform better on primary task metrics (mAP), however weather-conditioned variants (particularly temperature variants) display reduced miss-rates, indicating that while their accuracy is generally lower, they recognize more objects than the baseline-counterpart.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 September 2023 doi:10.20944/preprints202309.0499.v1

Table 1. In this table the mean Average Precision (mAP), and Miss-Rate (MR) of direct- (YOLOv5) and indirect-conditioning (DETR) variants are detailed. Highlighted with **bold** is the best performing across all models and highlighted with underline is the best performing model for a given architecture. mAP_{VOC} denotes mAP where IoU is atleast 0.5, mAP_{COCO} denotes mAP at varying IoUs (i.e. {0.50, 0.55, 0.60, ..., 0.95}). mAP_L, mAP_M and mAP_S denote mAP of objects with $\{area < 32^2, area > 32^2 < 96^2$ and $area > 96^2\}$ respectively.

Model	mAP_{voc}	mAP_{coco}	mAP_L	mAP_M	mAP_S	MR
YOLOv5 (Baseline)	**0.604**	**0.465**	**0.825**	**0.640**	0.491	0.342
YOLOv5 (Pretrain)	0.600	0.454	0.831	0.621	0.489	0.324
YOLOv5 (Temp.)	0.584	0.410	0.796	0.590	0.468	0.322
YOLOv5 (Hum.)	0.493	0.293	0.675	0.560	0.268	0.357
YOLOv5 (ToD)	0.549	0.439	0.805	0.566	0.431	0.356
DN-DETR Baseline	0.378	0.348	0.123	0.344	0.563	0.421
DN-DETR (Temp.)	0.225	0.148	0.100	0.190	**0.682**	0.389
DN-DETR (Hum.)	0.191	0.132	0.100	0.160	0.671	0.415
DN-DETR (ToD)	0.219	0.142	0.00	0.169	0.661	0.410
Def. DETR Baseline	0.332	0.202	0.005	0.051	0.637	0.383
Def. DETR (Temp.)	0.297	0.184	0.001	0.045	0.620	**0.351**
Def. DETR (Hum.)	0.213	0.114	0.000	0.020	0.517	0.416
Def. DETR (ToD)	0.289	0.178	0.001	0.040	0.619	0.395

Table 2. Accuracy of the predicted auxiliary prediction value, *Dir.* and *Indir.* denotes the direct- and indirect-conditioning models respectively, while the model row denotes the variant used.

	Model	MAE	Std.
Dir.	Temperature	7.1	3.7
	Humidity	18.9	9.4
	Time of Day	7.3	7.1
Indir.	Temperature	5.1	2.9
	Humidity	15.3	8.9
	Time of Day	8.3	7.9

3.4. Accuracy compared to weather

To evaluate the impact of the conditioning branch on the performance with respect to the different weather conditions used for auxiliary prediction and optimization, Figures 9 to 11 detail the the relation between mAP and the three meta-variables chosen, (namely Temperature, Humidity and time of day). Visual examples to accompany the accuracy overview of Tables 1 and 2, can be seen in Figures 7 and 8 (Ground truth labels and the image without bounding boxes can be found Figure A2), while visualizations of accuracy with respect to the different weather variables can be seen in Figures 9 to 11.

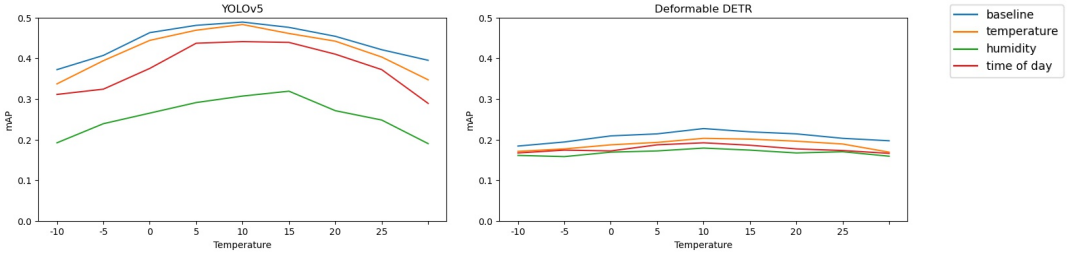
(a) Baseline **(b)** Temperature **(c)** Humidity **(d)** Time of Day

Figure 7. Example of Direct-Conditioning performance for each conditioned model. Bounding boxes marked in green, red, yellow are considered True Positives, False Positives, and False Negatives respectively.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 September 2023 doi:10.20944/preprints202309.0499.v1

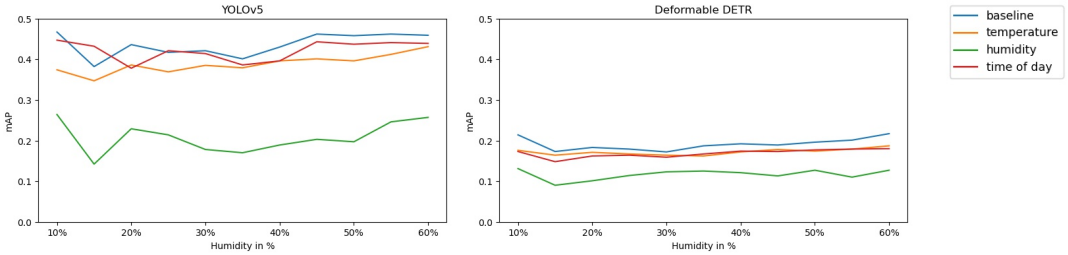
(a) Baseline **(b)** Temperature **(c)** Humidity **(d)** Time of Day

Figure 8. Example of Indirect-Conditioning performance for each conditioned model. Bounding boxes marked in green, red, yellow are considered True Positives, False Positives, and False Negatives respectively.



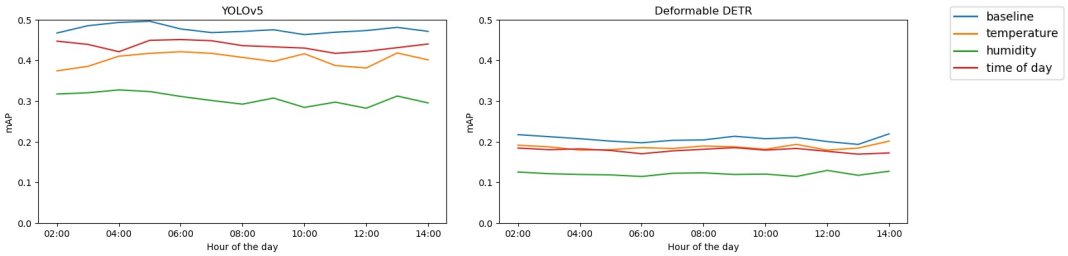
(a) **(b)**

Figure 9. Accuracy of models with regards to temperature



(a) **(b)**

Figure 10. Accuracy of models with regards to humidity



(a) **(b)**

Figure 11. Accuracy of models with regards to the time of day

In Figure 9 it can be observed that training models with an temperature-focused auxiliary branch, does not change the performance of said model in any significant way (other than generally lowered mAP). It can be seen that all models follow a curve that is similar to the distribution of samples seen in Figure 4b, it can be expected that this is happening as the models optimization is simpler when regressing to the mean of the dataset. In addition it can be observed that the indirect-conditioning method is generally more agnostic to variation in the meta variable. Similarly to the temperature focused auxiliary branch, humidity- and time-of-day-conditioning does not seem to improve overall performance of the models. However interestingly the models seem to be generally agnostic to the

 doi:10.20944/preprints202309.0499.v1

distribution of samples (shown in Figures 4a and 4c). This indicates that the model has trouble extracting meaningful information with regards to the auxiliary optimization task. This is also present in Table 2 where it can be seen that the networks have difficulty with accurately predicting their respective weather condition (specifically humidity and time-of-day), whereas temperature prediction is rather accurate, and falls close to the acceptable deviation of the $L1_e$ loss.

4. Discussion

Previous work has shown that traditional Convolutional Neural Networks (CNNs) able to predict weather categories, and in some contexts can help guide the network to be aware of the distribution a given sample belongs to and adjust accordingly. While this has not been shown to increase the accuracy in terms of mAP, it has been shown to decrease false negative predictions. Thermal images with significant concept drift could introduce artifacts that would look appropriate for a given object in one distribution but would be undesired for another. Intuitively, during training, the model would either adjust to over-predict (i.e. increased false positives), or under-predict (i.e. increased false negatives) when concept drift occurs. Essentially, the model is tasked with learning an unknown set of distributions, and optimized toward learning to recognize patterns common to the mean of the cumulative distribution. Therefore, one could hypothesize that guiding the network towards being aware of a variable correlated with the observed concept drift would allow the network to potentially establish connections between the conditioning representation and the semantic representation used for object prediction.

While it can be observed in Table 1 that the mAP scores do not improve over baseline when conditioned with the auxiliary branch, the change in MR indicates that the auxiliary branch is enforcing a signal it relates to the auxiliary task. Particularly, the temperature-conditioned variant manages to detect objects which the baseline fails to detect. However, the weather-conditioned method also produces an increase in false-positives. Because the visual appearance changes are gradual, resulting in lower accuracy, potentially the weather-conditioned learns a more varied representation of given objects which allows it to detect more objects at the cost of false activations in other places. Additionally, it can be observed (in Figure 9 is that the transformer-based model performs significantly more uniformly across temperatures. However, it is not certain that this is entirely an aspect of the auxiliary predictive branch, or the nature of transformers input-dependent attention. Another surprising detail can be found in Table 1, which shows the DETR-variants, in general, seem to work really well on small objects, which is counter to what is observed in the original and subsequent papers [32,41,44].We might reasonably conclude that this is partially due to the decoder module, which has a fixed amount of learnable query tokens, which should naturally converge towards spatial and latent features that are the most prominent, i.e. the person class. Initially, an experiment was conducted with regards to the amount of queries to produce, and while increasing them drastically(300->600) would improve performance by roughly 0.3%, the performance would increase significantly as well. The increased amount of query tokens could have been kept as a baseline. However, for the sake of keeping baseline models (i.e. YOLOv5 and Deformable DETR) somewhat comparable with other work, the hyperparameters described in their respective repository and paper was kept.

In previous work, namely [2], evaluated performance on a dataset that had two distinct thermal distributions (day and night clips). Our approach assumed that a more continual representation could be learned, however perhaps this cannot be learned fully without an additional proxy that enforces a strict set of conditions forcing formation of distinct distributions.

The appearance shift induced when the thermal camera calibrates to adjust the internal thermograph, perhaps, induces some noise into the signal, making it difficult to learn a robust approximation of the signal. It could be the visual noise induced partially obfuscates clear delineations between visual groups of visually similar samples, resulting in a regression to the mean being the simplest convergence, or perhaps the optimal solution for the downstream task. In such a situation naively training without the auxiliary branch would be the optimal solution if the goal is simply to

optimize accuracy. While the auxiliary task does seem to induce noise, both methods (direct- and indirect-conditioning) seem to also somewhat guide the network towards containing a more continual representation, as seen by the reduction in MR. Perhaps trying to construct distributions as a series of k overlapping distributions, and leveraging a model-soup style approach [46] could provide a more distinct learning of each sub-distribution, while still achieving a generalized model of all distributions.

An alternative solution to the regression approach could be a smooth classification approach, where the prediction is considered a smooth positive if it predicts a value within a pre-determined bin-size for each ground-truth number.

5. Conclusion

Thermal concept drift poses a challenging hurdle to overcome when deploying object recognition systems. Drawing from contextual clues that impact the visual appearance of the scene could be beneficial. Using auxiliary metrics to condition a network directly or indirectly does not seem to improve the overall performance of the system with regards to mAP. However, it does result in a consistent decrease in MR. While not resulting in a direct improvement, this shows that a signal can be extracted from the conditioning meta-variable, which can guide the representations learned. The difficulty in accurately modelling the objects across thermal signatures seems, similarly to a naively trained baseline, to prefer representations that favor the most frequent representations, and as such could be simply seen as a regression to the mean. However, due to the networks consistently being able to extract a signal related to the auxiliary task, it could imply that deliberately splitting the data into a set of K distributions based on a combination of meta-variables or visual appearance could provide more stable guidance.

Author Contributions: Conceptualization, A.S.J., K.N. and S.E.; methodology, A.S.J., K.N. and S.E; software, A.S.J.; validation, A.S.J.; formal analysis, A.S.J.; investigation, A.S.J.; resources, A.S.J. and K.N.; data curation, A.S.J. and K.N.; writing—original draft preparation, A.S.J.; writing—review and editing, A.S.J., K.N. and S.E.; visualization, A.S.J.; supervision, K.N. and S.E.; project administration, K.N. and A.S.J.; funding acquisition, K.N. All authors have read and agreed to the published version of the manuscript.

Funding: This research is partially funded by Milestone Systems A/S as a part of the Milestone Research Programme at Aalborg University (MRPA)

Institutional Review Board Statement: Not Applicable

Informed Consent Statement: Not Applicable

Data Availability Statement: The data used for this work is partially available on Kaggle[1] and the remaining object annotations will become available soon

Acknowledgments: Milestone systems and collaboration with Sergio

Conflicts of Interest: The authors declare that there are no conflict of interest regarding the publication of this paper

Acronyms

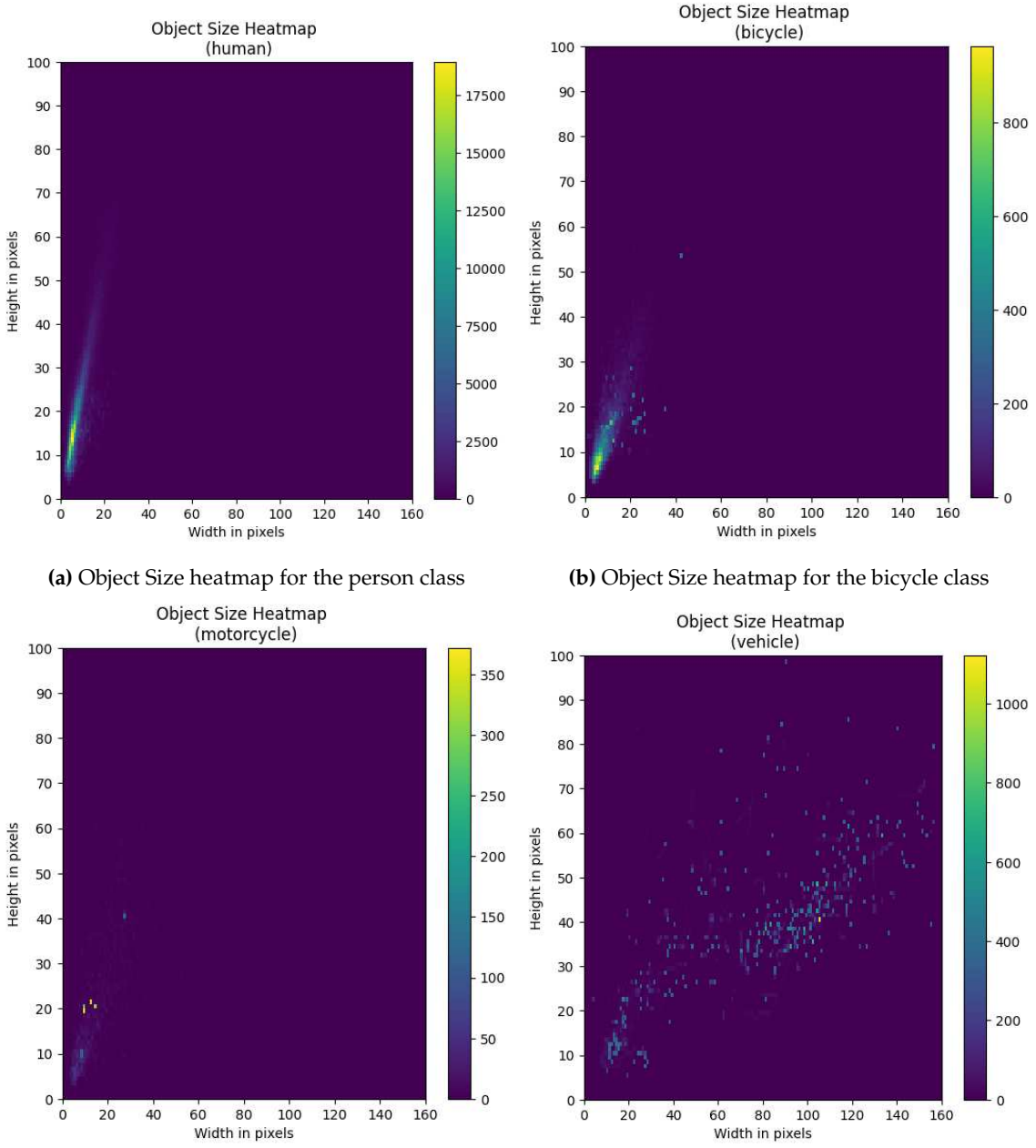
CNN	Convolutional Neural Network 11
IoU	Intersection over Union 6, 9
KAIST	KAIST Multispectral Pedestrian Detection 3–5
LTD	Long-term Thermal Drift 3–6, 8
MAE	Mean Average Error 8
mAP	mean Average Precision 4, 6, 8, 9, 11, 12
MR	Miss-Rate 3, 6, 8, 9, 11, 12
SotA	State of the Art 4
Std.	Standard Deviation 9

[1] https://www.kaggle.com/datasets/ivannikolov/longterm-thermal-drift-dataset

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 September 2023 doi:10.20944/preprints202309.0499.v1

Appendix A Additional Dataset Figures

Appendix A.1 Class-wise object distributions



(a) Object Size heatmap for the person class

(b) Object Size heatmap for the bicycle class

(c) Object Size heatmap for the motorcycle class

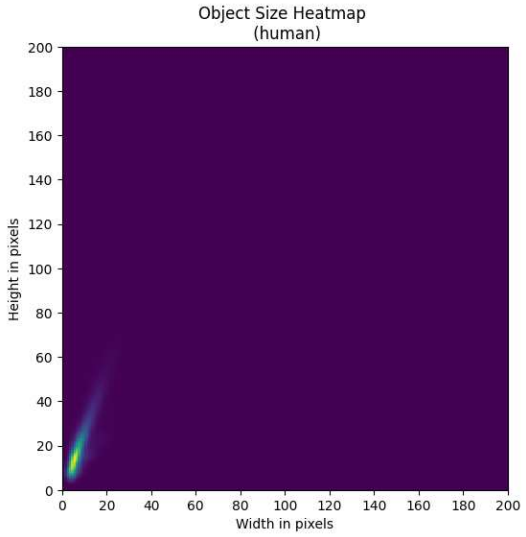
(d) Object Size heatmap for the vehicle class

Figure A1. Figures A1a to A1d show a detailed heatmap of the object size distributions for each class individually

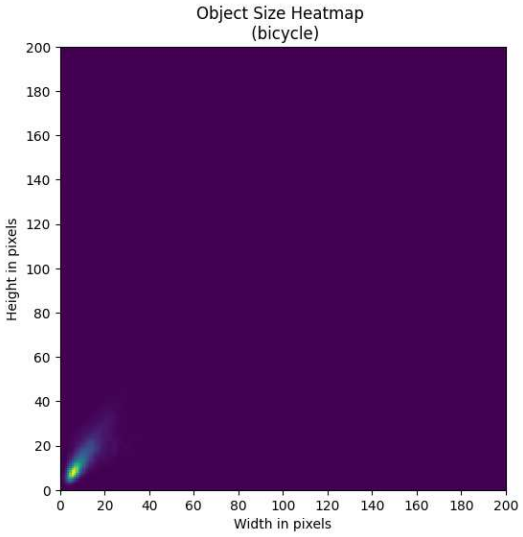
Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 September 2023 doi:10.20944/preprints202309.0499.v1

(a) **(b)**

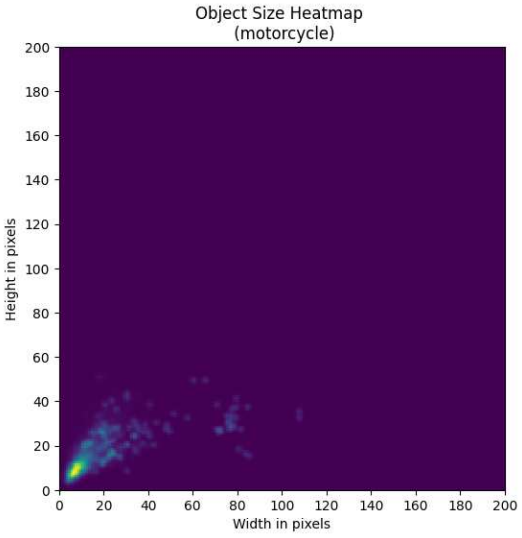
Figure A2. The example image used in Figures 7 and 8 without without bounding boxes(Figure A2a) and with bounding boxes(Figure A2b). In Figure 7 green, yellow and red refer to person, bicycle and vehicle classes respectively



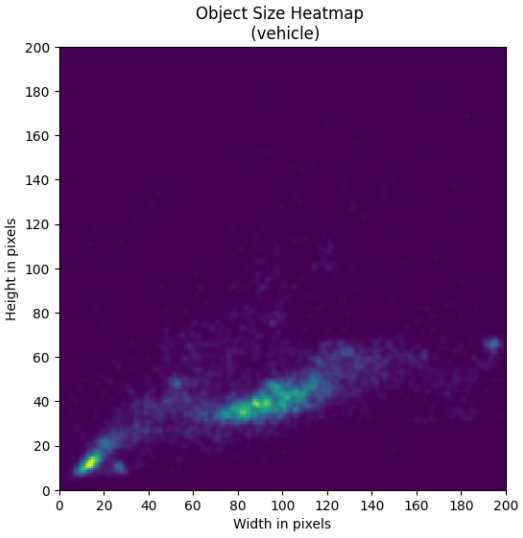
(a) Object Size heatmap for the person class



(b) Object Size heatmap for the bicycle class



(c) Object Size heatmap for the motorcycle class



(d) Object Size heatmap for the vehicle class

Figure A3. Histograms of training datasplit

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 September 2023 doi:10.20944/preprints202309.0499.v1

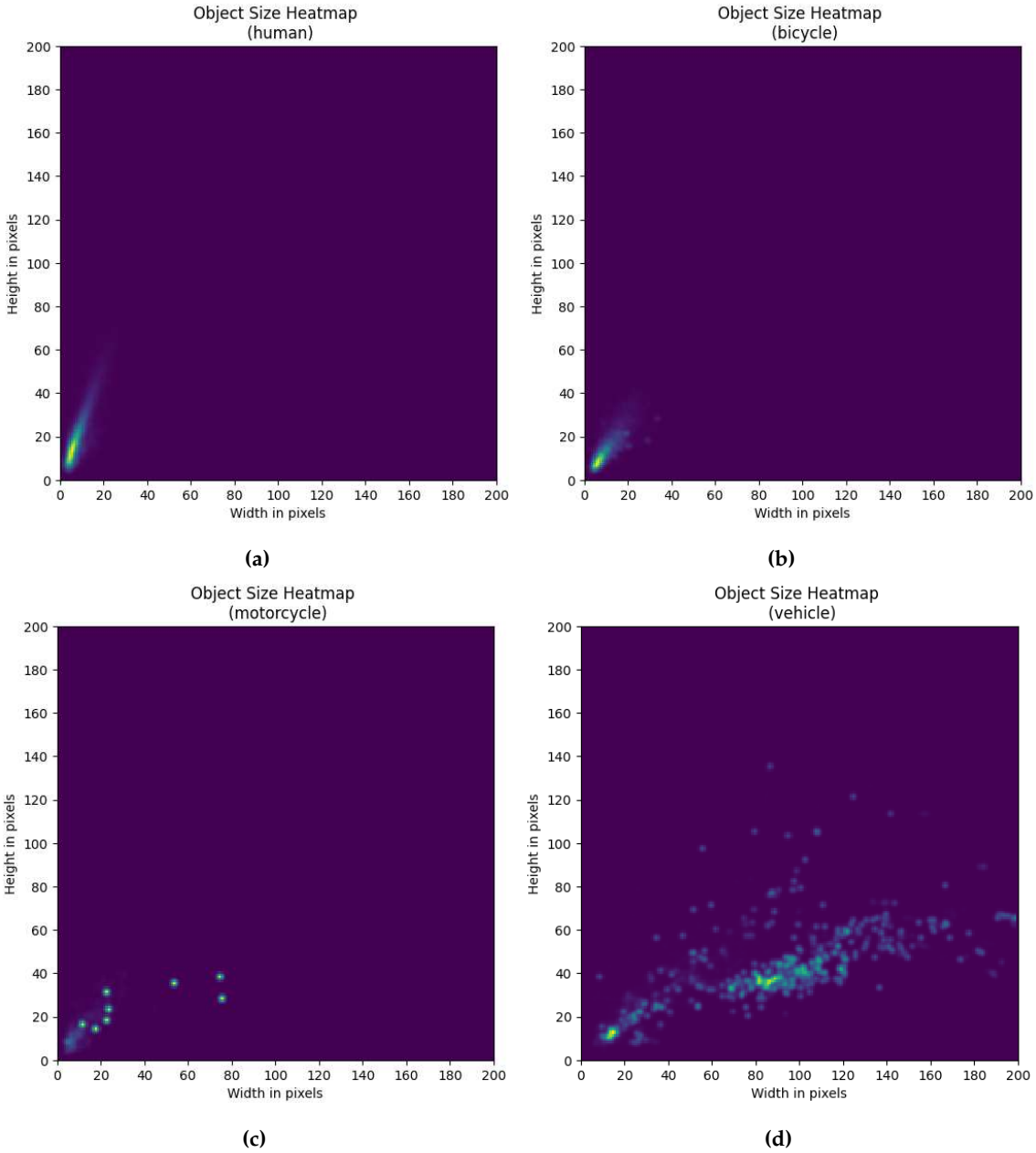


Figure A4. Histograms of Valid datasplit

References

1. Nikolov, I.A.; Philipsen, M.P.; Liu, J.; Dueholm, J.V.; Johansen, A.S.; Nasrollahi, K.; Moeslund, T.B. Seasons in drift: A long-term thermal imaging dataset for studying concept drift. In Proceedings of the Thirty-fifth Conference on Neural Information Processing Systems. Neural Information Processing Systems Foundation, 2021.
2. Kieu, M.; Bagdanov, A.D.; Bertini, M.; Del Bimbo, A. Task-conditioned domain adaptation for pedestrian detection in thermal imagery. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16. Springer, 2020, pp. 546–562.
3. Hu, R.; Singh, A. Unit: Multimodal multitask learning with a unified transformer. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 1439–1449.
4. Heuer, F.; Mantowsky, S.; Bukhari, S.; Schneider, G. Multitask-centernet (mcn): Efficient and diverse multitask learning using an anchor free approach. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 997–1005.

5. Bhattacharjee, D.; Zhang, T.; Süsstrunk, S.; Salzmann, M. Mult: an end-to-end multitask learning transformer. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 12031–12041.
6. Perreault, H.; Bilodeau, G.A.; Saunier, N.; Héritier, M. Spotnet: Self-attention multi-task network for object detection. In Proceedings of the 2020 17th Conference on Computer and Robot Vision (CRV). IEEE, 2020, pp. 230–237.
7. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the Proceedings of the IEEE international conference on computer vision, 2017, pp. 2961–2969.
8. Dahmane, K.; Duthon, P.; Bernardin, F.; Colomb, M.; Chausse, F.; Blanc, C. Weathereye-proposal of an algorithm able to classify weather conditions from traffic camera images. *Atmosphere* **2021**, *12*, 717.
9. Bhandari, H.; Palit, S.; Chowdhury, S.; Dey, P. Can a camera tell the weather? In Proceedings of the 2021 36th International Conference on Image and Vision Computing New Zealand (IVCNZ). IEEE, 2021, pp. 1–6.
10. Chu, W.T.; Zheng, X.Y.; Ding, D.S. Camera as weather sensor: Estimating weather information from single images. *Journal of Visual Communication and Image Representation* **2017**, *46*, 233–249.
11. Guerra, J.C.V.; Khanam, Z.; Ehsan, S.; Stolkin, R.; McDonald-Maier, K. Weather Classification: A new multi-class dataset, data augmentation approach and comprehensive evaluations of Convolutional Neural Networks. In Proceedings of the 2018 NASA/ESA Conference on Adaptive Hardware and Systems (AHS). IEEE, 2018, pp. 305–310.
12. Lin, D.; Lu, C.; Huang, H.; Jia, J. RSCM: Region selection and concurrency model for multi-class weather recognition. *IEEE Transactions on Image Processing* **2017**, *26*, 4154–4167.
13. Glasner, D.; Fua, P.; Zickler, T.; Zelnik-Manor, L. Hot or not: Exploring correlations between appearance and temperature. In Proceedings of the Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 3997–4005.
14. Ye, R.; Yan, B.; Mi, J. BIVS: Block Image and Voting Strategy for Weather Image Classification. In Proceedings of the 2020 IEEE 3rd International Conference on Computer and Communication Engineering Technology (CCET). IEEE, 2020, pp. 105–110.
15. Gama, J.; Žliobaitė, I.; Bifet, A.; Pechenizkiy, M.; Bouchachia, A. A survey on concept drift adaptation. *ACM computing surveys (CSUR)* **2014**, *46*, 1–37.
16. Lu, J.; Liu, A.; Dong, F.; Gu, F.; Gama, J.; Zhang, G. Learning under concept drift: A review. *IEEE transactions on knowledge and data engineering* **2018**, *31*, 2346–2363.
17. Xiang, Q.; Zi, L.; Cong, X.; Wang, Y. Concept Drift Adaptation Methods under the Deep Learning Framework: A Literature Review. *Applied Sciences* **2023**, *13*, 6515.
18. Bahnsen, C.H.; Moeslund, T.B. Rain removal in traffic surveillance: Does it matter? *IEEE Transactions on Intelligent Transportation Systems* **2018**, *20*, 2802–2819.
19. Wei, W.; Meng, D.; Zhao, Q.; Xu, Z.; Wu, Y. Semi-Supervised Transfer Learning for Image Rain Removal. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.
20. Wang, H.; Yue, Z.; Xie, Q.; Zhao, Q.; Zheng, Y.; Meng, D. From rain generation to rain removal. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 14791–14801.
21. Li, S.; Ren, W.; Zhang, J.; Yu, J.; Guo, X. Single image rain removal via a deep decomposition–composition network. *Computer Vision and Image Understanding* **2019**, *186*, 48–57.
22. Chen, J.; Tan, C.H.; Hou, J.; Chau, L.P.; Li, H. Robust video content alignment and compensation for rain removal in a cnn framework. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 6286–6295.
23. Li, K.; Li, Y.; You, S.; Barnes, N. Photo-Realistic Simulation of Road Scene for Data-Driven Methods in Bad Weather. In Proceedings of the Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops, 2017.
24. Rao, Q.; Frtunikj, J. Deep learning for self-driving cars: Chances and challenges. In Proceedings of the Proceedings of the 1st international workshop on software engineering for AI in autonomous systems, 2018, pp. 35–38.
25. Tremblay, M.; Halder, S.S.; De Charette, R.; Lalonde, J.F. Rain rendering for evaluating and improving robustness to bad weather. *International Journal of Computer Vision* **2021**, *129*, 341–360.

26. Halder, S.S.; Lalonde, J.F.; Charette, R.d. Physics-based rendering for improving robustness to rain. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 10203–10212.
27. Gao, J.; Wang, J.; Dai, S.; Li, L.J.; Nevatia, R. Note-rcnn: Noise tolerant ensemble rcnn for semi-supervised object detection. In Proceedings of the Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 9508–9517.
28. Solovyev, R.; Wang, W.; Gabruseva, T. Weighted boxes fusion: Ensembling boxes from different object detection models. *Image and Vision Computing* **2021**, *107*, 104117.
29. Körez, A.; Barışçı, N.; Çetin, A.; Ergün, U. Weighted ensemble object detection with optimized coefficients for remote sensing images. *ISPRS International Journal of Geo-Information* **2020**, *9*, 370.
30. Walambe, R.; Marathe, A.; Kotecha, K.; Ghinea, G.; et al. Lightweight object detection ensemble framework for autonomous vehicles in challenging weather conditions. *Computational Intelligence and Neuroscience* **2021**, *2021*.
31. Dai, R.; Lefort, M.; Armetta, F.; Guillermin, M.; Duffner, S. Self-supervised continual learning for object recognition in image sequences. In Proceedings of the Neural Information Processing: 28th International Conference, ICONIP 2021, Sanur, Bali, Indonesia, December 8–12, 2021, Proceedings, Part V 28. Springer, 2021, pp. 239–247.
32. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16. Springer, 2020, pp. 213–229.
33. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* **2020**.
34. 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 Proceedings of the Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 10012–10022.
35. Han, K.; Xiao, A.; Wu, E.; Guo, J.; Xu, C.; Wang, Y. Transformer in transformer. *Advances in Neural Information Processing Systems* **2021**, *34*, 15908–15919.
36. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* **2020**.
37. Tian, Y.; Bai, K. End-to-End Multitask Learning With Vision Transformer. *IEEE Transactions on Neural Networks and Learning Systems* **2023**.
38. Singh, S.; Khim, J.T. Optimal Binary Classification Beyond Accuracy. *Advances in Neural Information Processing Systems* **2022**, *35*, 18226–18240.
39. Ghosh, S.; Delle Fave, F.; Yedidia, J. Assumed density filtering methods for learning bayesian neural networks. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2016, Vol. 30.
40. Johansen, A.S.; Junior, J.C.J.; Nasrollahi, K.; Escalera, S.; Moeslund, T.B. Chalearn lap seasons in drift challenge: Dataset, design and results. In Proceedings of the Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part V. Springer, 2023, pp. 755–769.
41. Li, F.; Zhang, H.; Liu, S.; Guo, J.; Ni, L.M.; Zhang, L. Dn-detr: Accelerate detr training by introducing query denoising. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 13619–13627.
42. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767* **2018**.
43. Hwang, S.; Park, J.; Kim, N.; Choi, Y.; So Kweon, I. Multispectral pedestrian detection: Benchmark dataset and baseline. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 1037–1045.
44. Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; Dai, J. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159* **2020**.

45. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* **2018**.
46. Wortsman, M.; Ilharco, G.; Gadre, S.Y.; Roelofs, R.; Gontijo-Lopes, R.; Morcos, A.S.; Namkoong, H.; Farhadi, A.; Carmon, Y.; Kornblith, S.; et al. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In Proceedings of the International Conference on Machine Learning. PMLR, 2022, pp. 23965–23998.