Received: 28 December 2021 | Revised: 30 May 2022 | Accepted: 2 June 2022

IET Image Processing

DOI: 10.1049/ipr2.12562

The Institution of Engineering and Technology WILEY

ORIGINAL RESEARCH

Attention transfer from human to neural networks for road object detection in winter

Jonathan Boisclair | **Sousso Kelouwani** | **Follivi Kloutse Ayevide** | **Ali Amamou** | **Muhammad Zeshan Alam** | **Kodjo Agbossou**

Hydrogen Research Institute, Université du Québec á Trois-Rivières, Trois-Rivières, 3351 des Forges, Trois-Rivières, QC G9A 5H7, Canada

Correspondence
Jonathan Boisclair, Hydrogen Research Institute, Université du Québec á Trois-Rivières, Trois-Rivières, 3351 des Forges, Trois-Rivières, QC G9A 5H7, Canada.
Email: jonathan.boisclair@uqtr.ca

Funding information
Canada Research Chair Program; Natural Sciences and Engineering Research Council of Canada

Abstract

As an essential feature of autonomous road vehicles, obstacle detection must be executed on a real-time onboard platform with high accuracy. Cameras are still the most commonly used sensors in autonomous driving. Most detections using cameras are based on convolutional neural networks. In this regard, a recent teacher–student approach, called transfer learning, has been used to improve the neural network training process. This approach has only been used with a neural network acting as a teacher to the best of our knowledge. This paper proposes a novel way of improving training data based on attention transfer by getting the attention map from a human. The proposed method allows the dataset size reduction by 50%, which leads to up to a 60% decline in the training time. The experimental results indicate that the proposed method can enhance the F1-score of the network by up to 10% in winter conditions.

1 | INTRODUCTION

In order to achieve a level 4 or 5 of the international standard J3016 [1], cars must be able to detect road elements in all possible conditions, especially harsh winter conditions. More than 200 million people are living in countries[1] where snow is present for more than 60 days per year. The most commonly used sensor for road elements detection is the monocular camera [3, 4]. However, aerosols such as snow and rain scatter light through a wide range of angles and disrupt the vision making it difficult for object detections [5]. Such scattering may degrade the detection of outdoor vision systems. [6] Moreover, slim objects such as pedestrians become hard to detect as they become heavily altered and occluded with such aerosols [7]. Current car wipers leave raindrops and snow on the windshield, obstructing the drivers' visions and cameras inside the vehicle. As a result, the visibility of road objects such as traffic signs, vehicle, and pedestrians are gravely reduced [8]. R. Sato et al. [8] demonstrate an example of image alteration done by rain in [8, Fig. 2]. This statement can also be applied to snow, which deposits on the windshield. Ershadi et al. [9] have demonstrated that traditional approaches' accuracy decreases from 95% [9, Table 5] to 75% [9, Table 7] in snowy conditions. It supports the dramatic impact of winter on object detection. Ziadia et al. [10] also observed that winter conditions reduce pedestrian and distant vehicles' detection rate. Chebrolu et al. [11] have mentioned that most models fail to predict pedestrians in a dark environment. Winter increases dark environments' durations. As camera-based environment detection mainly uses neural networks (NNs) for detecting the environment around autonomous vehicles, it is highly affected by such conditions. As NNs are based on camera systems, adverse conditions heavily affect their detections. However, a high detection rate is required for path planning [12, 13] and maneuvers [14] of these vehicles. Improving the recognition in adverse conditions is possible by training the NN with specialized deformation, like snow or rain [15]. However, each possible occlusion of the image must be learned to improve the network. Learning all these variations could lower the F_1-score (17) [16, 17] and also results in fewer detections in stable conditions or a slower inference time for more extensive networks. For an occlusion like snow, an infinite number of image deformations exist depending on the location and shape of the snowflakes. Since each possible adjustment must be learned, the

[1] Russia (144.4), Canada (37.6), Sweden (10.3), Finland (5.5), Norway (5.3) and every country above the 43rd meridian [2].

infinite number places the learner in an impossible position. While training a network in normal conditions, where the visibility is not obstructed by snow or ice, the large number of datasets [18–32] available for standard cameras allow the training for broad recognition. Per contra, the preparation of such datasets is itself a big challenge because it requires the collection of the images and their time-consuming labelling [33]. The easiest way to upgrade detection accuracy is to increase the training set size. However, increasing the size of the dataset results in an augmentation of training time, which is not suitable for all applications. For instance, long-time training is not viable for autonomous cars in which the situation may constantly change by confronting new cars of different shapes and pedestrians with new clothes. When working with a very specific recognition set, the datasets that can be used become very scarce. Thus, increasing the amount of data in the training set becomes expensive. The last way of improving the training data is to use artificial data. Artificial data can be a solution to create data from simulations. However, utilizing artificial data may cause a drop in quality as the transformation may not suit the current problem. Hence, the size of the datasets must be kept small enough to be trained in a reasonable time.

In order to reduce the size of a dataset, researchers around the globe have worked on transfer learning [34–36]. The most popular transfer learning technique is the use of a sizeable generic dataset as a pre-training, then a small application-specific dataset for the top-up training [35]. The pre-training of the network using the data outside the recognition set can increase the classification performance [34] by inducing specific patterns shared with the detection set. These patterns, such as corners, are common to many real-life objects. This technique has the drawback of high computational time caused by the vast training size. Moreover, shapes with rare occurrences will have lower detection accuracy. Xi et al. [35] state that selecting an appropriate subset of the dataset for pre-training could produce a high-performance trained network in the same way as a large amount would do.

Transfer learning has also evolved into a second branch. The previously reserved task-to-task transfer using pre-training has been extended for a network-to-network transfer. Network-to-network transfer can be done via automatic image annotation [37–40] or weight transfer such as pre-training and sharing a backbone. Srinivas et al. [41] adapted this concept for pedestrian detection, allowing the research community to access the teacher–student networks for autonomous driving. In short, by using a high discriminative network as a teacher to a high capacity one, it is possible to obtain a high-capacity network. Seeing that the datasets can be small again, the problem shifts to recognition accuracy. This accuracy can be enhanced by changing the training method, dataset quality, and dataset size. Regarding the improvement of recognition accuracy in harsh conditions, the choice of the source network is vital for the transfer. A particular approach, named human-attention transfer from the transfer learning family, can be used with fewer drawbacks.

The human-attention transfer is still used to teach a more specific and smaller network [42]. There have been many papers that consider attention in NNs [43–46]. Howbeit, these papers approach attention as a detection problem. Attention during detection cannot solve severe condition limitations as it is not ground truth. During detection, attention acts as a heuristic. It is preceded by Grad-CAM++ [47] which states that with a slight decrease in detection accuracy, it is possible to hide parts that are not highly activated. In Grad-CAM++, the process is applied to the validation part of the network where no training is involved. Unlike Grad-Cam++, which uses a neural network to generate attention maps, human-attention transfer acts on the training step instead of understanding the network like Grad-CAM++. As proposed by Xiao et al. [48], attention can be described in two levels, object-level attention and part-level attention. Object-level attention is a high-level consideration also called saliency in the literature. These methods are based on detecting which part of the image is essential. Part-level attention methods are based on the decomposition of the object into smaller ones. As such, it is considered highly supervised. Each part will be highly activated on the map.

Human-attention transfer works at a previously undefined third level. That level is located between the object level and the part level. It contains feature-level saliency with only object-level annotation. Human-attention is located on that layer and brings relevant parts in the decision process to higher weight without identifying named bounding boxes for those[2].

In the light of the discussed papers, the main contribution of this paper is to put forward a new annotation process for transferring the knowledge from humans to a NN. This new transfer format is easily obtained from humans and dramatically helps the computer learn the appropriate pattern. That novel annotation format improves each picture's value in the dataset, allowing for smaller datasets. This effective transferring increases the accuracy or decreases the training time even under challenging conditions, like a winter storm. This increase in accuracy facilitates autonomous car driving under challenging winters.

The rest of the paper is organized as follows. Section 2 elaborates on the proposed approach. Section 3 demonstrates the results, and finally, the conclusion is provided in Section 4.

2 | PROPOSED ATTENTION PROCESS

A tool named Digital Representation of Attention Labeler (DRAL) is proposed in this work for handling attention. A training method that best utilizes the new pre-processed data set will be presented afterward.

2.1 | An overview of the novel approach

Typically, object detection is performed by drawing bounding boxes around objects of different classes to guide the network

[2] Video example available at https://youtu.be/rAnMiux725Y

according to the object shape. This approach is complemented by an attention pre-step, similar to semantic labelling. Semantic labelling tags every pixel in the image with the correct category instead of bounding boxes that guide the network according to the object shape. Instead of the class, attention adds a weight of importance to each pixel. It guides the network toward learning which pixels in the object are the most essential. Although semantic attention is achievable, the proposed method uses attention with bounding boxes. The proposed approach is a teacher and student system based on transfer learning. Typical teacher and student processes use one or more NNs as teachers. In this case, the teacher will be an ensemble of humans which will be named *masters* further in this article. Multi-source training is needed since there are multiple *masters*. By collecting the attention maps from several humans, it should be possible to extract a common attention map by averaging [44]. Once the maps are created, the following steps are equivalent to the traditional method, as seen in Figure 1. The proposed method adds attention to the labelling step, which comes first. This approach is based on humans as attention must be extracted from a human line of sight. The learned pattern must be the same since the network mimics the brain. Using humans prevents adding training data dynamically in the transfer, as the process is not automated. In the case of having a large enough sample of images, dynamically adding images can be neglected.

The usage of attention is highly suitable for winter, in which many occlusions are not part of the detected object. Attention could hide these pixels by assigning them a weight close to zero. Here, attention is coming from humans. They tend to stop for a short period on important regions of an object while skipping the less essential ones. During winter, when there are many occlusions, humans tend to look for the remaining parts of the object for longer, creating a subset of essential elements. Attention is extracted from an image instead of a video to create the attention-images from humans. Image-based attention finds the essential parts of the images.

Hiding some unnecessary details, such as a license plate, advertisements on a bus, or snow covering windows during wintertime, reduces the number of counterexamples to show the network to prevent learning unnecessary information. As opposed to video-based methods, image-based methods do not use Long Short Term Memory (LSTM) networks and have multiple important points. Video-based methods generally extract the most important point of the image, as seen in BDD-Attention [31].

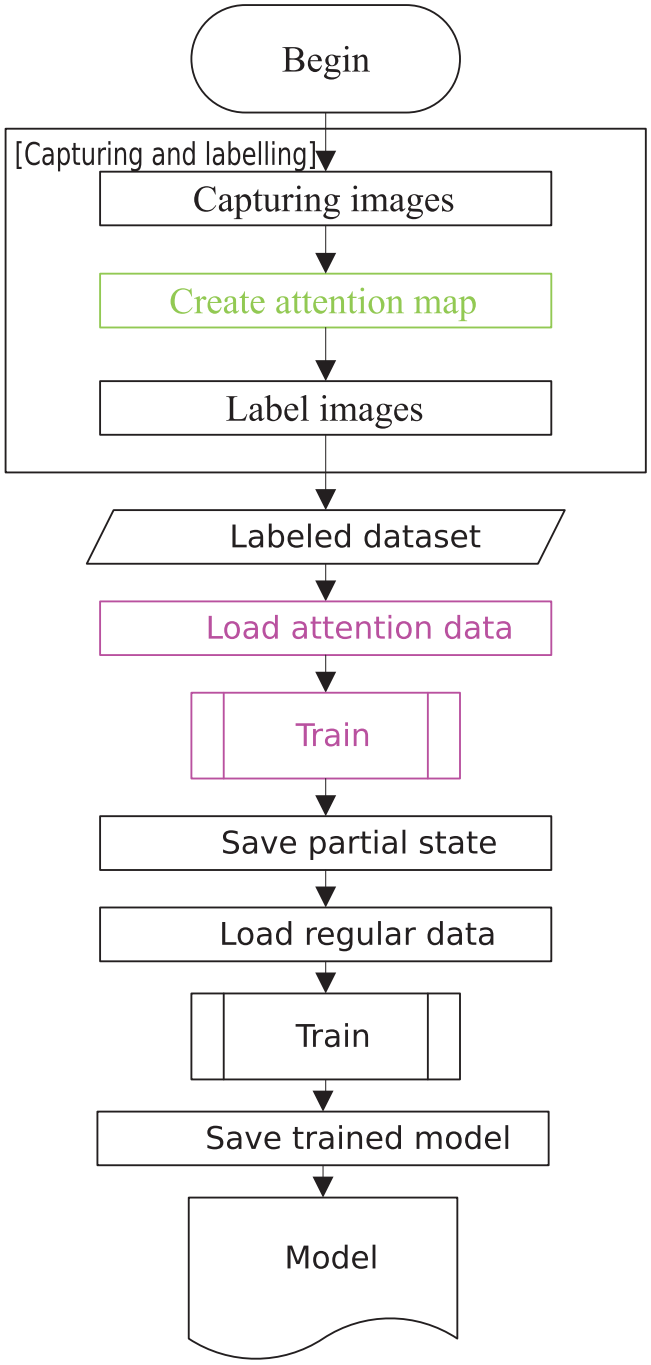


FIGURE 1 Flowchart of the training process

2.2 | Novel way of creating the attention maps

2.2.1 | Definition of an attention value

By tracking a human line of sight, it is possible to learn what parts of the image are the most important. However, analyzing the human gaze through a camera is challenging since it depends on the disposition of the devices [49]. By using a camera, the gaze can only be approximated with a high margin of errors [49]. In this regard, a computer interface that simulates the gaze is used to create the attention map. In human vision, neurons are the most sensitive units, located in a small region in the centre of the view. This region appears to be circular and is called *fovea* [50]. There is no standard formula to describe the density of the fovea. Hence, this paper proposes an empirical equation, based on a sigmoid, for this purpose as follows:

The inverse of a sigmoid function on the absolute value of the distance follows the density of the cones in the human

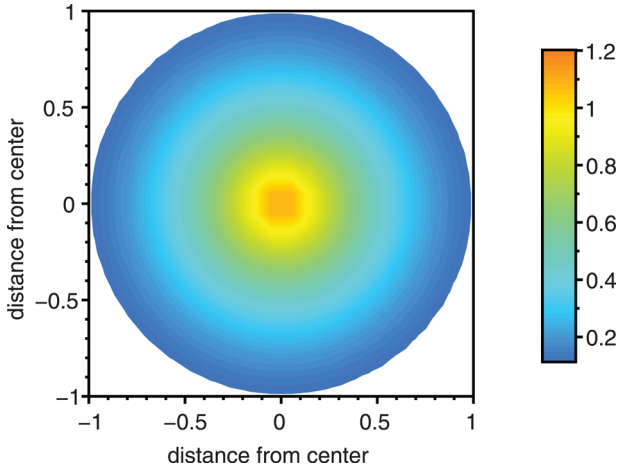


FIGURE 2 Visual representation of the attention factor from equation 1

eye.

$$\alpha_{\Delta s} = \begin{cases} 2 \times \left(1 - \dfrac{1.5^{(8 \times \Delta s - 1)}}{1.5^{(8 \times \Delta s - 1)} + 1}\right) & \text{if } \Delta s < 1 \\ 0 & \text{if } \Delta s \geq 1 \end{cases}, \quad (1)$$

where $\alpha_{\Delta s}$ represents the distance factor, Δs represents the normalized distance from the centre of the circle, $1.5^{(8 \times \Delta s - 1)}$ is a sigmoid argument with experimental values, and one represents the outer edge of that circle as follows:

$$\Delta s = \frac{\sqrt{(\Delta x)^2 + (\Delta y)^2}}{r}, \quad (2)$$

where Δx represents the normalized horizontal difference on a scale of minus one to one, Δy denotes the normalized vertical difference of location, and r is the radius of vision (a constant defined by the size of the vision hole). It should be noted that r is represented as $100px$, but can be any values small enough to display only a part of the image. The resulting factor can be presented in the form of a map, as shown in Figure 2. The choice of $100px$ was made for a 1920×1080 screen for the smallest object detectable to be slightly smaller than the vision hole. This size allows for proper labelling of the attention while still being fast enough for the *master*.

2.2.2 | Attention map

This method improves the training set to transfer knowledge from human to machine. First, all inputs are displayed to the *masters* under a specially developed software (DRAL). The entire image is covered with a black mask, and the human user scans the image with the computer cursor. A small circle (vision hole) is shown to the user, and it follows the cursor. In order to add a context view, a thumbnail of the picture is also shown on the right side. Its size is not large enough to view details. The weight of the pixels overlapping the vision hole in the attention map is

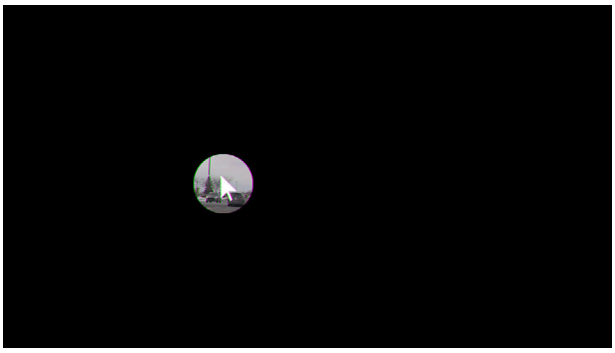


FIGURE 3 Example of the vision hole

based on the time cursor spent on this location. A short video has been prepared to demonstrate the attention map generation process effectively[3]. Contrary to camera-based methods, using the cursor eliminates the imprecision on the target [49] since only a tiny portion of the image is displayed (Figure 3). The whole image is scanned by the user using a pattern of his choice. As humans tend to look at essential objects for longer, the vision hole will stay at that location for a more significant duration.

The software records the image's part in that circle and shows its duration. It then creates an attention map of where the person was looking according to the distance factor ($\alpha_{\Delta s}$). If the same picture is presented to multiple *masters*, an average of the attention map is calculated for even further improvement [44]. An example of the output of the process is available in Figure 4. In this figure, the effects of snow are demonstrated. Partially melted snow is present in the middle of the image, making frontal vision blurry and distorted. Distant objects are hidden under a white cloak similar to fog. Cars are partially covered in snow, and all road markings are hidden. Only the objects are considered in this figure since the current approach was meant for this. Road boundaries are not in the attention map extracted considering object detection. They use a different kind of network and would require a separate training process and separate dataset for training. It is viable to use attention on road boundaries; however, attention is focused on object detection. Each point (x, y) in the attention image has a time (T) in seconds defined by a weighted sum obtained from the following equation:

$$T_{x,y} = \begin{cases} \sum_k \Delta t_k \times \alpha_{\Delta s}(x, y, C_{x,k}, C_{y,k}) & if < A_{max} \\ A_{max} & else \end{cases}, \quad (3)$$

where $T_{x,y}$ is the time of attention in seconds for a point located at x, y, k is a movement of the cursor, Δt_k is the amount of time the cursor stays there for the movement k multiplied by $\alpha_{\Delta s}$, $C_{x,k}$ and $C_{y,k}$ are the centres of the vision hole on both horizontal and vertical axes, and A_{max} is a constant defined as the maximum of attention at a single location. For the learning

[3] https://youtu.be/rAnMiux725Y

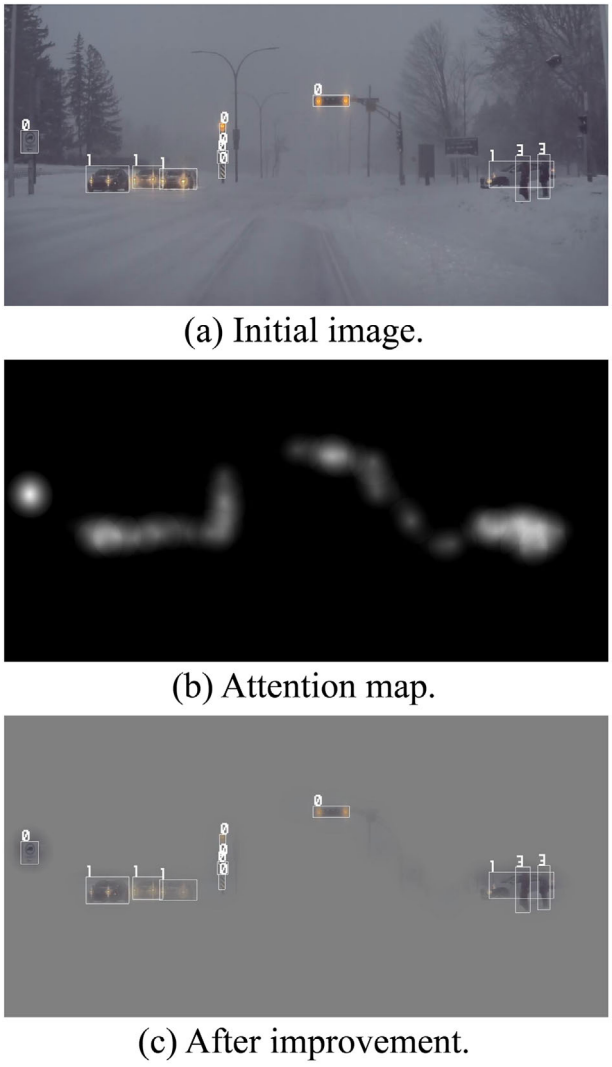


FIGURE 4 Application of attention transfer from human; images from the proposed dataset

purpose, it was supposed that $A_{max} = 5$ sec. Placing the maximum prevents the annotated parts from disappearing on the normalization filter in case the *master* gazes at a fixed point for a long time.

Once the attention map is created, it is normalized between zero and one to serve as the alpha band of the input:

$$N_{x,y} = \frac{T_{x,y}}{\max T}, \tag{4}$$

where $N_{x,y}$ is the normalized point at location x, y, and T represents the ensemble of all $T_{x,y}$. Normalization acts as a way of letting each of the 16 *masters* be equally useful while averaging. The input image is then overlaid on a gray inactivated background of 0.5. Images are fed to the training set with the original annotations of bounding boxes. In this way, the NN can only learn the contents that should be learned instead of learning all the content in the bounding box, including the background and occlusions. During winter, parts like license plates and icicles will have an attention weight close to zero, transforming the output as inactivated (0.5).

Each channel c in the image, such as *red*, *blue*, *green*, is fed into the following weighted sum:

$$F_{x,y,c} = N_{x,y} \times O_{x,y,c} + 0.5 \times (1 - N_{x,y}), \tag{5}$$

where O represents the original image, N is the normalized attention map, F is the final improved image using attention, and x and y represent the location of a single point. This equation allows the image to be used in any NN without editing the network itself. This weighted sum requires a single weight between zero and one.

2.2.3 | Bounding boxes

The second part of the transfer is refining the bounding boxes. By using the already labelled data, such as the Berkeley Deep Drive (BDD), it is possible to improve the boxes with attention transfer. It allows the conversion of poor datasets into high-quality ones. X-Means [51] is used to cluster the attention map into object proposals. X-Means has been utilized as it is simple and, unlike the NN-based clustering methods, does not require any training. On the other hand, X-Means is slower than other options, which is not a problem for this work. While using X-Means as the region proposal method, it is possible to separate the attention image into multiple clusters. Those distance-based and attention-based clusters must be filtered to become usable semantically. The first filter will separate any clusters containing multiple non-touching regions. The second filter will merge smaller clusters embedded inside the larger ones. X-Means needs vectors as input. The attention map is converted into three-dimensional vectors where X, Y represent its position in the image, and Z represents the normalized attention ranging from 0 to 1. X-Means expects different parameters to be equal in range. A weighted function is required to recreate that assumption by normalization. The distance function used for these points is the weighted Minkowski:

$$\Delta_{MINKOWSKI} = \sqrt[p]{\omega_x \times |X_2 - X_1|^p + \omega_y \times |Y_2 - Y_1|^p + \omega_z \times |Z_2 - Z_1|^p} \tag{6}$$

where

$$0 \le \omega_z \le (I_w^2 + I_h^2) \tag{7}$$

$$|X_1 - X_2| \le I_w \tag{8}$$

$$|Y_1 - Y_2| \le I_h \tag{9}$$

$$|Z_1 - Z_2| \le 1 \tag{10}$$

$$\omega_x = 1 \tag{11}$$

$$\omega_y \in \{1, I_w/I_h\} \tag{12}$$

where I_w is the image width and I_h is the image height.

Ideally, each vector component must have the same or similar scale for X-Means to work without bias. However, a bias is required in the current case since attention and locations are fundamentally different. A ω_z greater than or equal to the diagonal length of the image would make the X-Means only consider the attention. The bias against Z would be consequentially larger than X or Y. ω_z is ideal at $\omega_z = I_w$ in a way that the representation becomes close to a cube, eliminating the bias. A perfect cube would require that the size follows either $I_w = I_h$ or $\omega_y = I_w/I_h$. The algorithm is run for five combinations of parameters as described in Table 3. These five pairs of parameters are chosen based on the similarity of their output to the ground truth and the number of boxes they introduced. In short, a low ω_z has a bias against the location. A high ω_z, on the other hand, has a bias against the intensity of the attention. It is needed to divide those patches since our attention cannot jump from one object to another. Moreover, it may let a meager activation trail that a threshold cannot always filter without removing important parts.

Adjustments to labels are made by consolidating the original bounding boxes from the dataset with the proposed regions from the X-Means operation. Since the region proposal can separate objects more than needed. A merging algorithm is required to compare the objects with the manually marked ones. That algorithm can be defined in four steps:

1) Identifying the proposed boxes that have an Intersection over union (IoU from (13)) of more than a specified threshold (TU) and accepting them as new boxes.
2) Identifying the proposed boxes in which the Intersection over smallest (IoS from (14)) is above a second threshold (T_S) and expanding the initial bounding box to completely include the smallest one.
3) Identifying all boxes that are inside the original one and shrinking the original one to fit them. Shrink if it is less than T_{Shrink}.
4) Concerning the boxes that do not match with the proposal, keep the original bounding box

$$IoU\ (A, B) = \frac{A \cap B}{A \cup B} \tag{13}$$

$$IoS\ (A, B) = \frac{A \cap B}{\min\{A, B\}} \tag{14}$$

The fourth step allows boxes that do not have a proposal for improvement to be considered sufficient. Those objects are essential to detect, but the region proposal did not produce any suggestions for improvement. No improvement will be made to those boxes. However, improvements made by attention will remain.

Similarly, this method can also be used for unlabelled datasets. It was previously stated that the method could improve the bounding boxes because X-Means act as a region proposal. The region proposal can be used even with some loosely tagged labels by defining the attention map before labelling. The user labels images faster since the proposal is present, and it does not need to be precise. Using these box proposals and an attention protocol makes it possible to label images efficiently. The protocol may be as simple as one sentence: "Look for cars in the image." The attention map created by that instruction can then be used for the automatic tagging of cars. Moreover, the existing NNs function as secondary teachers and verify whether a car with high confidence is in reality or not. By using attention, it is possible to employ multiple teachers. The first step would form an attention map for a specified class by each teacher. Then that map would be merged with all *masters* of that class by utilizing an average. The latter junctions would create a final attention map for an image based on all classes. The attention for the second junction would be $F_{x,y} = \max(A_{i_{x,y}})$, where A_i is a category of objects, i is the index of the category, and $F_{x,y}$ is the resulting value for that location. Proposals may be either before or after the second merge. Proposing boxes before the second merge would help the algorithm produce accurate bounding boxes for intricate groups of objects with different categories.

In short, the creation of an attention map using DRAL was demonstrated. Subsequently, a method for improving the labels was proposed. Finally, a particular way for labelling images so that the work overhead would be minimal was presented. With the adequately prepared data, the network can be efficiently trained.

2.3 | Training

Within the deep learning framework, many convolutional NNs exist [52–57].

To test attention, YOLOv3 [58], SSD [59] and Faster-RCNN [60] are used in conjunction with almost all of PyTorch's optimizers. This approach should work with any network that learns patterns with many neurons, including any recent convolutional NN. YOLOv3 and SSD have been chosen because they are small networks with good results that can operate at a frequency faster than a real-time camera [4]. Faster RCNN has been chosen for its completeness and great performance. PyTorch is used for its simplicity, completeness, popularity, and open-source availability.

As for a training protocol, four types of input data exist. By combining those input data, 15 different types of training data are obtained. However, only six of them are deemed valuable, which are defined in Table 1, and are used for the training purposes of the CNN presented in this article. As an input, there are the regular from BDD (BDDR), attention-based from BDD (BDDP), regular from UWD (UWDR), and attention-based from the new dataset (UWDP). The set of all attention-based images is DSP, and all regular is DSR. DSA consists of every

TABLE 1 Number of images per dataset used. The Source column represents where the data is coming from, Berkeley Deep Drive(BDD), UQTR(Data captured by our team), or Russian traffic sign dataset(RTSD). The R/P column represents regular or proposed data. The size represents the number of pictures in each data section. The conditions column represents the weather conditions of the specific section of the dataset.

ID	Source	R/P	Size	Max Precision	Max Recall	Max F1	Conditions
DSA	all	R+P	14381	19	65	29	Clear/Summer Winter (Light+Heavy)
BDD	BDD	R+P	10999	16	72	26	Clear/Summer Winter (Light)
BDDP	BDD	P	1000	18	72	29	Clear/Summer Winter (Light)
DSP	all	P	2691	–	–	–	Clear/Summer Winter (Light+Heavy)
UWDP	UQTR +RTSD	P	1691	13	64	22	Clear/Winter (Heavy) Clear/ Winter (Heavy)
UWDR	UQTR+RTSD	R	1691	17	66	27	Clear/Winter (Heavy) Clear/ Winter (Heavy)
BDDR	BDD	R	9999	18	72	24	Clear/Summer Winter (Light)
DSR	all	R	11690	–	–	–	Clear/Summer Winter (Light+Heavy)

image from attention-based and regular. All images and attention maps are available on our github[4]. The novel part of the dataset named UWD will be presented in Section 2.4.

The network is trained up to a threshold β of epochs on the improved dataset and then $(1-\beta)$ on the regular dataset (DSR). This first step allows learning the appropriate filters.[5] It is then trained on the regular labelled data for the remaining epochs to reduce the number of false positives. This method is the same as a pre-training [62] except that in the current situation, the pre-training dataset is the same as the training one with the alteration. The $(1-\beta)$ regular epochs let the network adapt to real-world images without attention to guide it. A β too low would cancel all benefits that our attention-based images did; a β too high would create overfitting. For the test purposes, β has been set to 80% following multiple tests. It was the percentage resulting in the best training. For the epochs preceding the β cut, only $\lambda_{pre} = 80\%$ of the selected dataset is used for training, and in the later epochs, $\lambda_{late} = 80\%$ of BDDR is used. λ is separated into two parameters since both datasets could share different proportions of data in training. The validation metrics are always calculated against $\nu = 20\%$ of DSR. BDDR and UWDR are the only datasets sharing images between training and validation. For those two, BDDR is altered to omit the used images, which are scarce. As attention-based data can drive particular patterns into training, overfitting is in its nature. The latter part of the training, where BDDR is pushed back into the training, reduces overfitting resulting from those patterns by presenting all the information. Too much regular data and second overfitting will be created based on the full images and lose every benefit of those forced patterns. Attention works with any convolutional NN.

For the proof of concept, the basics of the networks were not changed except for some minor required adjustments. One of the applied changes was to remove batch normalization to allow the network to run multiple sub-batch before running an optimizer. A batch size of 32 was used, with a sub-batch of the highest number of images that fit in the VRAM[6] up to 32.

As for the calculation of the loss, the original loss calculation was kept for each network. In YOLOv3, the loss function is a combination of a mean squared error (MSE), binary cross entropy (BCE), and cross entropy (CE) loss. The loss for the boxes' locations was calculated using MSE, confidence was computed using BCE, and classes using CE. Layers up to 75 are part of the Darknet-53 network, which is the backbone of YOLOv3. For Faster-RCNN and SSD, the Multibox loss was used. Multibox loss is a loss function that combines a smooth L1 loss for the regression of the boxes and a BCE for classification.

The metrics used for evaluating the network are the standard precision (15), recall (16) and F_1-score (17) from a confusion matrix. Even if more metrics exist for detection networks, such as *mean average precision* (mAP), they are more complex and generate similar results compared to the simple metrics. Precision, recall, and F_1-score (17) are defined by the following equations:

$$Precision = \frac{Relevant \cap Retrieved}{Retrieved} \tag{15}$$

$$Recall = \frac{Relevant \cap Retrieved}{Relevant} \tag{16}$$

[4] Available at https://github.com/irh-ca-team-car/attention-data
[5] Appropriate to be classified as expert input [61].
[6] Video ram

$$F_1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (17)$$

where retrieved is the set of detected objects and relevant is the ground-truth objects.

Training is done for every required epoch and each compatible optimizer. Each step of training is defined in Figure 1 where parts in black are untouched, parts in green are added, and a two-step training inspires parts in pink.

It is feasible to upgrade data by using attention and merging an attention map with the training data. It is also possible to improve the data training at a considerable small cost. That cost could be even further reduced by labelling the images in an attention format [33].

Creating the attention map for a single image takes an average of 31 seconds from the display to the generation on disk. The average for the bounding boxes calculation is around three minutes of computing time on a 3.9 GHz Xeon W-2133 with a single core. Using another region proposal, this 3-min computation can almost be reduced to none. X-Means takes an extended processing time but does not require training. This paper has used it to demonstrate the proof of concept regarding the proposed method. Other region proposals, such as NNs, can detect objects almost instantly but require training. The third step of labelling takes an average of 14 s per image, which is faster than labelling alone. This is because the attention step obscures parts of the images, resulting in less work for the individual tagging. Moreover, since the box proposals are generated utilizing attention, the process of person labelling does not need to be precise. Bounding boxes will be adapted to the proposals as well. Labelling time is evaluated using regular images, and labelling alone takes an average of 27 s, which is almost twice as slow as the proposed method. This approach adds an operational overhead of 18 human seconds per image while tagging and 180 computer seconds for the region proposal. These times are calculated based on the tagging of the first 500 images from BDD and the first 252 new images for a proposed dataset.

2.4 | New winter dataset: UWD

Following the proposed method, a small dataset named UWD has been set up for training. An electric Kia Soul 2017, as visible in Figure 5, is used for creating the new dataset. For capturing the dataset, 5160 HD frames at 60 Hz were obtained from a GoPro as training requires a higher resolution than operation. One image is taken every 20 frames for the dataset from these frames, resulting in 258 images. On top of these images, more than 500 frames were taken from the RTSD dataset [63], and more than 500 from the dashcam of the author's Tesla during snowstorms.

In the first step, attention maps are developed using 16 *masters*. Then, the computer determines the proposals, and finally, the images are labelled by 16 other people for bounding boxes. 16 *masters* are sufficient to remove human subjectivity. The attention map made by the 16 *masters* has less than 5% difference from the ones produced by the neural network experts. This

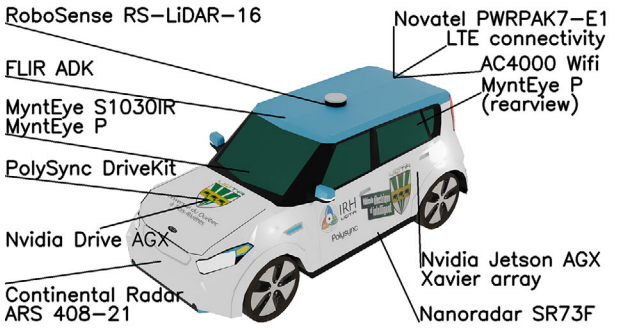


FIGURE 5 The car used for the UWD dataset

TABLE 2 Classes index

Class Index	Class descriptions
0	traffic-related objects
1	engine
2	bi-wheeled
3	humanoid

TABLE 3 Minkowski distance parameters used for box proposal. Two hundred and fifty-five represent the maximum value of a grey in an eight-bit image. ω_y has a weight calculated to convert the image to a square so a cube can be achieved

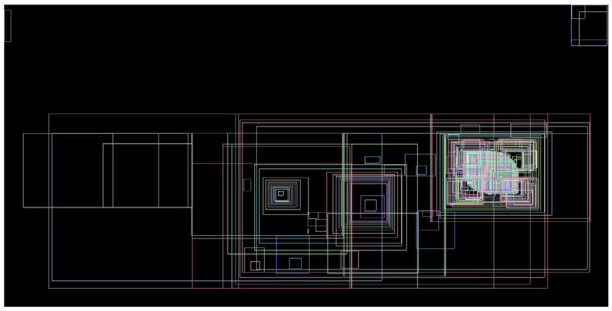
p	wx	wy	wz
1	1	1	3×255
1	1	I_w/I_h	$1 \times I_w$
4	1	1	3×255
4	1	1	4×255
4	1	I_w/I_h	$1 \times I_w$

dataset is driving-oriented and contains four classes to learn. The four classes are traffic-related objects (signs, traffic lights, constructions, barriers), engine (car, bus, truck), bi-wheeled (bicycle and motorcycle), and finally humanoid (person, mannequins). The four categories are ordered by the ascending level of threat to the target, meaning that traffic-related objects are harmful to the car and dangerous to humanoids. These classes are determined by the type of action for the ego vehicle instead of visual clues. Miscategorizing a car as a truck does not penalize the detection process. Those classes are presented in Table 2.

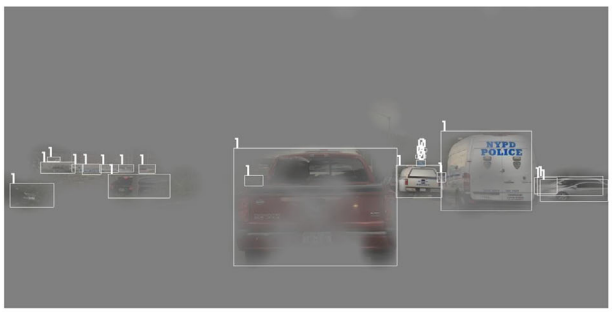
3 | RESULTS

The results of the attention transfer from an individual to a machine will be presented first, followed by examples from UWD in winter.

(a) Initial image.



(b) Attention map.



(c) After improvement.

FIGURE 6 Labels improvement for $T_U = 0.9$, $T_S = 0.9$, $T_3 = 0.25$

3.1 | Bounding boxes

DRAL (Digital Representation of Attention Labeler) recalculates bounding boxes utilizing the attention map. This recalculation allows for more accurate bounding boxes exploiting only still discernible elements. However, since the details of the objects are omitted, it is obvious that the bounding boxes will not be useful for a human anymore. Omitted details could be the front vent of the three cars, as shown in Figure 4.

These new bounding boxes are visible in Figure 6. Most of the bounding boxes in the image have been modified by less than 10% of their area and location. In Figure 6b, all the proposals from the region proposal are visible. Since X-Means output an enormous amount of proposals and to improve figure clarity, color code has been used to differentiate different boxes from each other. The region proposal produced many proposals in the top right corner, mostly unused as they did not match with actual bounding boxes. The police minivan in Figure 6 also has a noticeable change in its box. The air vent and tires are now mostly excluded. Most smaller vehicles have not been modified as no proposal has been found for them. The number over the boxes are the classes as defined in Table 2.

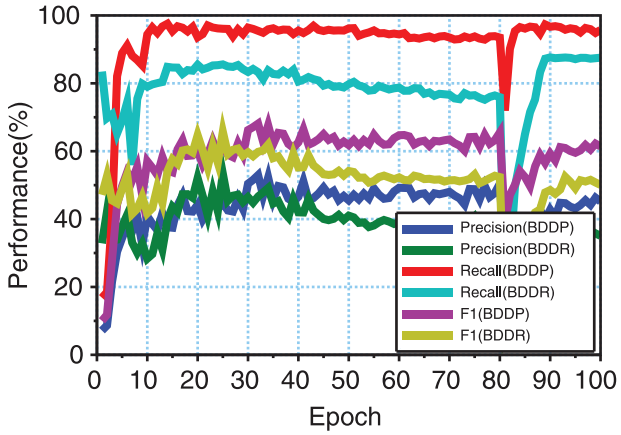


FIGURE 7 Precision and recall evolution based on epochs for both regular (BDDR) and attention-based (BDDP) data, calculated on training set. $\beta = 80\%$

3.2 | Training

The network learning uses the annotated images with an improved version based on the proposed tool (DRAL) for the training enhancement. In the validation set, the trained network on the upgraded data converges faster than the original data. It also has lower precision and recall on the training set. The size of datasets and subsets are defined in Table 1.

The network that is trained on original data maximizes at a precision of 59% and recall of 87% (Figure 7). The network trained on the attention-based data maximizes at a precision of 74% and a recall of 98%. The comparison between BDDR and BDDP presented in Figure 7 is valid as both datasets are composed of the same images with or without the attention process applied.

An example of training results using the Adamax optimizer is presented in Figures 8 and 9. Figure 8b shows the recall of the same set. Finally, Figure 9a represents the F_1-score (17) on the validation set. All those figures are the average of the training of YOLOv3, SSD, and Faster-RCNN. The blue (BDDR) and pink (UWDR) curves mark the original data, and the black (BDDP) and red (UWDP) curves represent the data created by the improved attention-based method. The turquoise curve (DSP) illustrates both Berkeley Deep Drive in attention format (BDDP) and UWDP. Finally, the green curve (DSA) represents all data from Berkeley and all UWD. It is important to note that BDD contains some rare winter images, but UWD contains winter images exclusively. Moreover, DSA and BDDR contain images taken during winter, while UWDP and UWDR are exclusively winter-based. BDD has images taken by the University of Berkeley in New York city. New York winters are not as harsh as Canadian winters, resulting in non-optimal detection during Canadian winter. Harsh winter images are required to test the autonomous cars that will operate in that condition. For

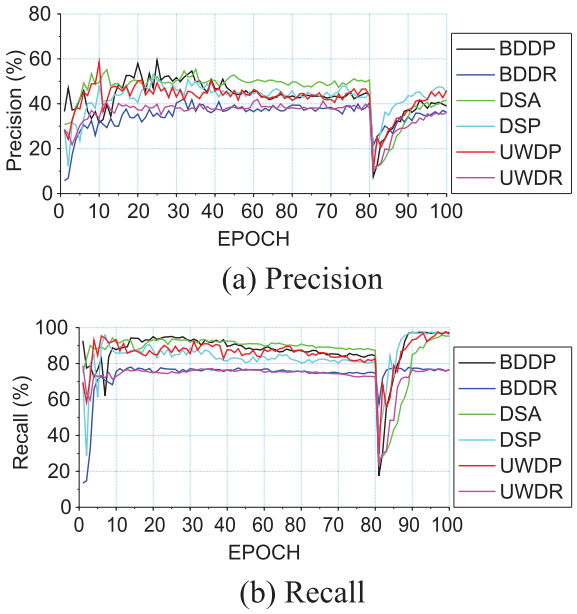


FIGURE 8 Average validation metrics across networks. DSA is split equally between matching regular images and attention-based images. Half the amount of images is used as they are doubled by the attention process in this dataset

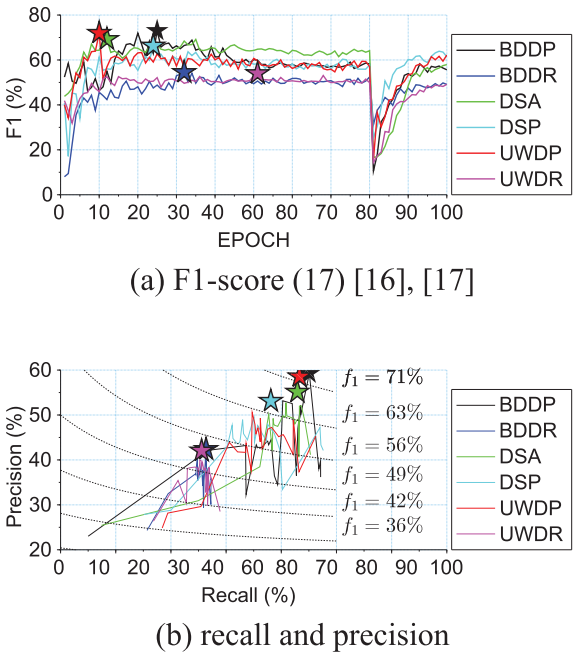


FIGURE 9 Average validation summary across networks. DSA is split equally between matching regular images and attention-based images. Half the amount of images is used as they are doubled by the attention process in this dataset

this training to be fair compared to datasets, 1000 images from each were used. Since the DSA dataset contains both regular and attention-based images, they are separated equally between each data type. This means that only half the amount of data is necessary when using attention-based mixed with regular

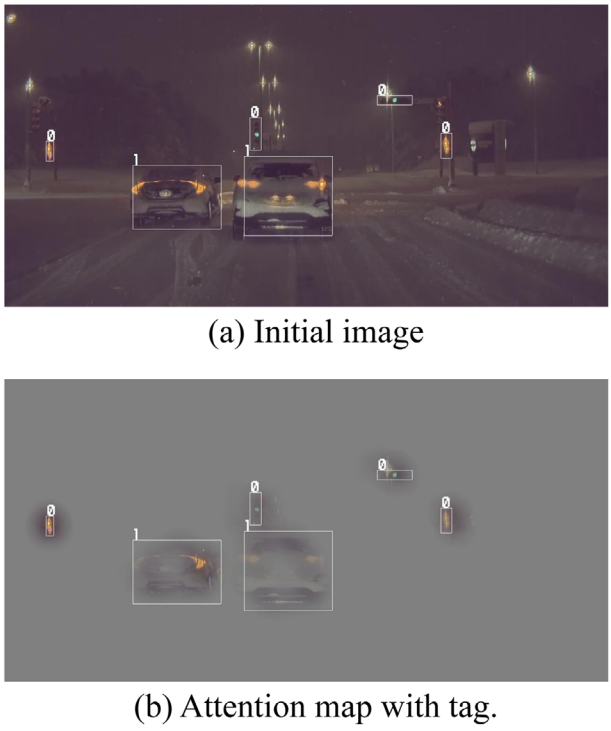


FIGURE 10 Newly tagged image from our winter dataset

images, as each image is present two times, with and without attention. Datasets that contains attention or are exclusively attention such as DSP, DSA, BDDP and UWDP results in higher F_1-score (17). The usage of winter attention such as UWDP results in an improvement in the training time from 25 epochs to 9 epochs to obtain the maximum F_1-score (17). DSA, which both attention and regular data in both summer and winter contexts rank as one of the top F_1-score (17) in a concise amount of epochs. It is also noted that the sudden change in all graphs for epoch 80 is caused by the change of dataset at $\beta = 80\%$. To validate that the network will perform better during winter conditions, all training was validated against the **R** dataset. Figure 7 shows that the proposed method has been able to improve the accuracy of recognition as opposed to regular training. Summer data using attention topped up 60% F_1-score (17), while regular data only topped 55% for the same amount of data.

An example of a newly tagged winter image is included in Figure 10. This image has been captured and tagged directly in the attention format instead of improving the existing labels. As seen in Figure 10, the boxes created by labelling directly in the attention format are similar to the boxes from Figure 4. Such similarity between attention applied before and after labelling the boxes indicates that both existing and new datasets could be used with the proposed method. However, labelling images with the attention method first increases the labelling speed. This image clearly shows that the contents of both the window and the license plates have been hidden to improve the image quality since these parts contain patterns that must not be learned. The number over the bounding box are the classes associated with

the object as defined in Table 2. To validate this method, the network has been trained on the four classes, driving-inspired dataset introduced in Section 2.4.

4 | CONCLUSION

This paper puts forward an approach for improving object detection in autonomous vehicles based on the attention mechanism. The proposed tool, called DRAL, has shown convenient utilization and compelling results for extracting attention from a person's gaze through several tests. Attention has been merged with regular images and bounding boxes recalculated to develop an artificial high-quality training set. This approach has been motivated by the closeness of neural networks to the human brain, and humans normally drive with almost full recognition of the environment. The proposed technique has been used to create a 1691-image driving oriented dataset published on Github[7].

As a medium for transfer learning, attention has demonstrated compelling accuracy in object detection for autonomous cars, even during harsh conditions like winter. Even with an increase of 50% of labelling time per picture, the proposed method can cut a dataset size by half. This reduced-sized dataset results in 25% less total time passed in labelling. On top of the reduction in labelling time, the proposed method also reduces training time by reaching the maximum F_1-score (17) in about half the number of epochs. This approach has successfully trained multiple neural networks with higher detection accuracy than its regular counterparts. A total of 1691 images have been proposed in a novel dataset regarding the attention for autonomous land vehicles.

ACKNOWLEDGEMENTS

This research was funded by the Natural Sciences and Engineering Research Council of Canada and the Canada Research Chair Program.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in attention-data at https://github.com/irh-ca-team-car/attention-data.

ORCID

Jonathan Boisclair https://orcid.org/0000-0001-6688-2761

REFERENCES

1. SAE: Levels of driving automation. https://www.sae.org/news/2019/01/saeupdates-j3016-automated-driving-graphic (2019)
2. worldometer.: Countries in the world by population. https://www.worldometers.info/world-population/population-by-country/ (2020)
3. Zhe, T., Huang, L., Wu, Q., Zhang, J., Pei, C., Li, L.: Inter-vehicle distance estimation method based on monocular vision using 3d detection. IEEE Trans. Veh. Technol. 69(5), 4907–4919 (2020). https://www.scopus.com/inward/record.uri?eid=2-s2.0-85085168451&doi=10.1109%2fTVT.2020.2977623&partnerID=40&md5=1ad2271037929f51e330a4cff9677bb6
4. Zhao, Z., Zheng, P., Xu, S., Wu, X.: Object detection with deep learning: A review. IEEE Trans. Neural Networks Learn. Syst. 30(11), 3212–3232 (2019)
5. Malik, M., Majumder, S.: An integrated computer vision based approach for driving assistance to enhance visibility in all weather conditions. In: Proceedings of the 1st International and 16th National Conference on Machines and Mechanisms (iNaCoMM2013), IIT Roorkee, India, 18-20 December 2013
6. Kang, L.W., Chou, K.L., Fu, R.H.: Deep learning-based weather image recognition. In: Proceedings - 2018 International Symposium on Computer, Consumer and Control, IS3C 2018, pp. 384–387. IEEE, Piscataway, NJ (2019)
7. Zhang, S., Yang, J., Schiele, B.: Occluded pedestrian detection through guided attention in cnns. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6995–7003. IEEE, Piscataway, NJ (2018)
8. Sato, R., Domany, K., Deguchi, D., Mekada, Y., Ide, I., Murase, H., et al.: Visibility estimation of traffic signals under rainy weather conditions for smart driving support. In: 2012 15th International IEEE Conference on Intelligent Transportation Systems, pp. 1321–1326. IEEE, Piscataway, NJ (2012)
9. Ershadi, N.Y., Menendez, J.M., Jimenez, D.: Robust vehicle detection in different weather conditions: Using mipm. PloS One 13(3), e0191355 (2018)
10. Ziadia, M., Kelouwani, S., Amamou, A., Dube, Y., Agbossou, K.: Using an intelligent vision system for obstacle detection in winter condition. In: VEHITS 2019 - Proceedings of the 5th International Conference on Vehicle Technology and Intelligent Transport Systems, pp. 562–568, Heraklion, Crete, Greece (2019)
11. Chebrolu, K.N.R., Kumar, P.N.: Deep learning based pedestrian detection at all light conditions. In: Proceedings of the 2019 IEEE International Conference on Communication and Signal Processing, ICCSP 2019, pp. 838–842. IEEE, Piscataway, NJ (2019)
12. Ji, J., Khajepour, A., Melek, W.W., Huang, Y.: Path planning and tracking for vehicle collision avoidance based on model predictive control with multiconstraints. IEEE Trans. Veh. Technol. 66(2), 952–964 (2017)
13. Nakamura, A., Liu, Y.C., Kim, B.: Short-term multi-vehicle trajectory planning for collision avoidance. IEEE Trans. Veh. Technol. 69(9), 9253–9264 (2020)
14. Yuan, C., Liu, H., Shen, J., Chen, L., Jiang, H.: Design and analysis of an autobraking system controller for autonomous vehicles under the influence of perturbation. IEEE Trans. Veh. Technol. 67(3), 1923–1931 (2018)
15. Samuel Dodge, L.K.: Understanding how image quality affects deep neural networks. In: Conference on the QUality of Multimedia Experience(QoMEX). IEEE, Piscataway, NJ
16. Michaelis, C., Mitzkus, B., Geirhos, R., Rusak, E., Bringmann, O., Ecker, A.S., et al.: Benchmarking robustness in object detection: Autonomous driving when winter is coming. arXiv:190707484v1 [csCV] (2019)
17. Musicant, D.R., Kumar, V., Ozgur, A.: Optimizing f-measure with support vector machines. In: Proceedings of the International FLAIRS Conference, pp. 356–360. Haller AAAI Press (2003)
18. Singha, A., Bhowmik, M.K.: Tu-vdn: Tripura university video dataset at night time in degraded atmospheric outdoor conditions for moving object detection. In: 2019 IEEE International Conference on Image Processing (ICIP), pp. 2936–2940. IEEE, Piscataway, NJ (2019)
19. Pham, Q.H., Sevestre, P., Pahwa, R.S., Zhan, H., Pang, C.H., Chen, Y., et al.: A*3d dataset: Towards autonomous driving in challenging environments. arXiv:190907541v1 [csCV] (2019)
20. Pavlov, A.L., Karpyshev, P.A., Ovchinnikov, G.V., Oseledets, I.V., Tsetserukou, D.: IceVisionSet: lossless video dataset collected on russian winter roads with traffic sign annotations. In: 2019 International

[7] Available at https://github.com/irh-ca-team-car/attention-data

Conference on Robotics and Automation (ICRA), pp. 9597–9602. IEEE, Piscataway, NJ (2019)
21. Shakhuro, V.I., Konushin, A.: Russian traffic sign images dataset. Computer Optics 40, 294–300 (2016)
22. Maddern, W., Pascoe, G., Linegar, C., Newman, P.: 1 year, 1000km: The oxford robotcar dataset. The International Journal of Robotics Research (IJRR) (2016)
23. Li, X., Flohr, F., Yang, Y., Xiong, H., Braun, M., Pan, S., et al.: A new benchmark for vision-based cyclist detection. In: 2016 IEEE Intelligent Vehicles Symposium (IV), pp. 1028–1033. IEEE, Piscataway, NJ (2016)
24. Hwang, S., Park, J., Kim, N., Choi, Y., So Kweon, I.: Multispectral pedestrian detection: Benchmark dataset and baseline. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1037–1045. IEEE, Piscataway, NJ (2015)
25. Xu, Y., Wang, H., Liu, X., He, H.R., Gu, Q., Sun, W.: Learning to see the hidden part of the vehicle in the autopilot scene. Electronics 8(3), 331 (2019)
26. Li, Y., Tong, G., Gao, H., Wang, Y., Zhang, L., Chen, H.: Pano-RSOD: A dataset and benchmark for panoramic road scene object detection. Electronics 8(3), 329 (2019)
27. Dominguez Sanchez, A., Cazorla, M., Orts Escolano, S.: A new dataset and performance evaluation of a region-based CNN for urban object detection. Electronics 7(11), 301 (2018). https://www.mdpi.com/2079-9292/7/11/301
28. Jensen, M.B., Philipsen, M.P., Møgelmose, A., Moeslund, T.B., Trivedi, M.M.: Vision for looking at traffic lights: Issues, survey, and perspectives. IEEE Trans. Intell. Transp. Syst. 17(7), 1800–1815 (2016)
29. Hodan, T., Haluza, P., S, O., Matas, J., Lourakis, M., Zabulis, X.: T-less: An RGB-D dataset for 6D pose estimation of texture-less objects. In: 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 880–888. IEEE, Piscataway, NJ (2017)
30. Chowdhuri, S., Pankaj, T., Zipser, K.: Multinet: Multi-modal multi-task learning for autonomous driving. In: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1496–1504. IEEE, Piscataway, NJ (2019)
31. Fisher, Y., Wenqi, X., Yingying, C., Fangchen, L., Mike, L., Vashisht, M., et al.: Bdd100k: A diverse driving video database with scalable annotation tooling. arXiv:180504687 [csCV], 2018. https://arxiv.org/abs/1805.04687
32. Matthew, P., Danson, G., Jason, R., Michael, S., Carlos, W., Krzysztof, C., et al.: Canadian adverse driving conditions dataset. arXiv:200110117 [csCV], 2020. https://arxiv.org/abs/2001.10117
33. Faizov, B.V., Shakhuro, V.I., Sanzharov, V.V., Konushin, A.S.: Classification of rare traffic signs. Computer Optics 44(2), 236–245 (2020)
34. Pan, S.J., Yang, Q.: A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering 22(10), 1345–1359 (2010)
35. Xi, Y., David, A., Sanja, F.: Neural data server: A large-scale search engine for transfer learning data. arXiv:200102799 [csCV], 2020. https://arxiv.org/abs/2001.02799
36. Wang, K., Gao, X., Zhao, Y., Li, X., Dou, D., Xu, C.Z.: Pay attention to features, transfer learn faster CNNS. In: International Conference on Learning Representations.
37. Zhang, D., Islam, M.M., Lu, G.: A review on automatic image annotation techniques. Pattern Recognit 45(1), 346–362 (2012). https://www.sciencedirect.com/science/article/pii/S0031320311002391
38. Jeon, J., Lavrenko, V., Manmatha, R.: Automatic image annotation and retrieval using cross-media relevance models. In: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval. SIGIR '03, p. 119–126. Association for Computing Machinery, New York, NY (2003). https://doi.org/10.1145/860435.860459
39. Chen, Y., Zeng, X., Chen, X., Guo, W.: A survey on automatic image annotation. Applied Intelligence 50(10), 3412–3428 (2020)
40. Ma, Y., Liu, Y., Xie, Q., Li, L.: Cnn-feature based automatic image annotation method. Multimedia Tools and Applications 78(3), 3767–3780 (2019)
41. Chebrolu, K.N.R., Kumar, P.: Deep learning based pedestrian detection at all light conditions. In: 2019 International Conference on Communication and Signal Processing (ICCSP), pp. 0838–0842. IEEE, Piscataway, NJ (2019)
42. Ding, Z., Shao, M., Fu, Y.: Incomplete multisource transfer learning. IEEE Trans. Neural Networks Learn. Syst. 29(2), 310–323 (2018)
43. Sergey, Z., Nikos, K.: Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. arXiv:161203928 [csCV], 2016. https://arxiv.org/abs/1612.03928
44. Monroy, R., Lutz, S., Chalasani, T., Smolic, A.: Salnet360: Saliency maps for omni-directional images with CNN. Signal Process. Image Commun. 69, 26–34 (2018)
45. He, X., Peng, Y., Zhao, J.: Fine-grained discriminative localization via saliency-guided faster R-CNN. In: Proceedings of the 25th ACM International Conference on Multimedia (MM '17), pp. 627–635. Association for Computing Machinery, New York, NY, USA (2017). https://doi.org/10.1145/3123266.3123319
46. Haohang, X., Hongkai, X., Guojun, Q.: Flat: Few-shot learning via autoencoding transformation regularizers. arXiv:191212674 [csCV], 2019. https://arxiv.org/abs/1912.12674
47. Chattopadhay, A., Sarkar, A., Howlader, P., Balasubramanian, V.N.: Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 839–847. IEEE, Piscataway, NJ (2018)
48. Xiao, T., Xu, Y., Yang, K., Zhang, J., Peng, Y., Zhang, Z.: The application of two-level attention models in deep convolutional neural network for fine-grained image classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 842–850. IEEE, Piscataway, NJ (2015)
49. Pingmei, X., Krista, A.E., Yinda, Z., Adam, F., Sanjeev, R.K., Jianxiong, X.: Turkergaze: Crowdsourcing saliency with webcam based eye tracking. arXiv:150406755 [csCV] 2015. https://arxiv.org/abs/1504.06755
50. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Trans. Pattern Anal. Mach. Intell. (11), 1254–1259 (1998)
51. Pelleg, D., Moore, A.: X-means: Extending k-means with efficient estimation of the number of clusters. In: Proceedings of the Seventeenth International Conference on Machine Learning, p. 727–734. Morgan Kaufmann, San Francisco, CA (2000)
52. Chao, Y.W., Vijayanarasimhan, S., Seybold, B., Ross, D.A., Deng, J., Sukthankar, R.: Rethinking the faster R-CNN architecture for temporal action localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1130–1139. IEEE, Piscataway, NJ (2018)
53. Zhang, L., Lin, L., Liang, X., He, K.: Is faster R-CNN doing well for pedestrian detection? European Conference on Computer Vision. Lecture Notes in Computer Science, vol 9906, pp. 443–457. Springer, Cham (2016)
54. Kandylakis, Z., Vasili, K., Karantzalos, K.: Fusing multimodal video data for detecting moving objects/targets in challenging indoor and outdoor scenes. Remote Sensing 11(4), 446 (2019)
55. Zhu, X., Hu, H., Lin, S., Dai, J.: Deformable convnets v2: More deformable, better results. arXiv:181111168 [csCV] (2018)
56. Zhou, Y., Lyu, Y., Huang, X.: Roadnet: An 80-mw hardware accelerator for road detection. IEEE Embedded Sys. Lett. 11(1), 21–24 (2019)
57. Xiang, W., Mao, H., Athitsos, V.: Thundernet: A turbo unified network for real-time semantic segmentation. In: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1789–1796. IEEE, Piscataway, NJ (2019)
58. Redmon, J., Farhadi, A.: Yolov3: An incremental improvement. arXiv preprint arXiv:180402767 (2018)
59. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., et al.: Ssd: Single shot multibox detector. In: European Conference on Computer Vision 2016, vol. 9905, pp. 21–37. Springer, Cham (2016)
60. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell 28(6), 1137–1149 (2017). https://doi.org/10.1109/TPAMI.2016.2577031

61. Williams Jr, D.R., Gutzwiller, L.R., Hazen, M.U., Anderson, B.S., McIntyre, A., Abeles, T.: Classifying data with deep learning neural records incrementally refined through expert input' [Generic]. Google Patents, (2016)
62. He, K., Girshick, R., Dollar, P.: Rethinking imagenet pre-training. In: Proceedings of the IEEE international Conference on Computer Vision, pp. 4918–4927. IEEE, Piscataway, NJ (2019)
63. Shakhuro, V., Konushin, A.: Russian traffic sign images dataset. Computer Optics 40(2), 294–300 (2016)

How to cite this article: Boisclair, J., Kelouwani, S., Ayevide, F.K., Amamou, A., Alam, M.Z., Agbossou, K.: Attention transfer from human to neural networks for road object detection in winter. IET Image Process. 16, 3544–3556 (2022). https://doi.org/10.1049/ipr2.12562