

OWL AUTONOMOUS IMAGING • 2023

CONVOLUTIONAL NEURAL NETWORKS

for THERMAL IMAGING



Vol. 3
01/2023

THE POWER OF CONVOLUTIONAL NEURAL NETWORKS AT NIGHT

Increasing regulatory concentration on improving the protection of Vulnerable Road Users (VRUs) against vehicle collisions at night has led to new evaluations of an assortment of imaging modalities that might quickly, effectively, and economically identify VRUs and measure their positions relative to moving vehicles.

This white paper explains how the Owl Thermal Ranger™ system can locate and classify VRUs in the dark from their own thermal signatures using just one infrared camera and carefully trained convolutional neural networks.

IMAGING AT NIGHT

On a clear, moonless night out in the countryside, the illuminance at the surface of a road, indicated in Figure 1, is about 1/1,000 lux - very dark - 500 to 10,000 times less than the recommended road illumination needed for drivers to see VRUs soon enough to avoid collisions. Headlights can improve visibility on straightaways but on curves and on rises and falls of the pavement darkness encroaches until no reaction time remains. To improve detection where well-distributed artificial lighting is not installed some other strategy is needed.




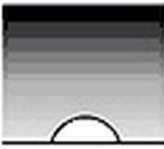
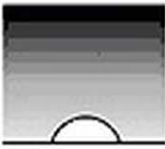
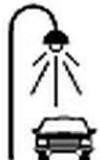
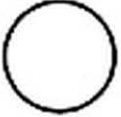



| | | | | | |
|----------------|--|--|---|---|--|
| Outdoor | Sunny Day (Direct Sunlight)  10000-1000000 Lux | Sunny Day (Shade)  2000-4000 Lux | Overcast Day  100-1000 Lux | Dawn  5-25 Lux | Dusk  5-25 Lux |
| | Street Lighting  0.5-10 Lux | Full Moon  0.1 Lux | Quarter Moon  0.01 Lux | No Moon  0.001 Lux | Overcast Night  0.001 Lux |

Figure 1 Darkness at night

CREDIT: EAGLEVIEW SECURITY, NEW ZEALAND

In another of our white papers, “Infrared for ADAS and Robotic Mobility Applications”, we discussed the utility of thermal infrared in the identification of VRUs. However, detection alone is not enough to support decision-making – each VRU must be classified by type and tagged with a distance from the vehicle. To accomplish both these functions, Owl AI has developed for its Thermal Ranger system a complex of convolutional neural networks (CNNs) that can extract from a single thermal image all the information required for automatic emergency braking decisions.

DETECTION, RECOGNITION, AND IDENTIFICATION (DRI)

The images supplied to the CNN must contain sufficient detail to allow extraction of the desired information. Starting with human viewers decades ago, the U. S. Army Night Vision and Electronic Sensors Directorate established three useful levels of image functionality, Detection, Recognition, and Identification (DRI), to use as a basis for specifying and characterizing thermal cameras and optics. These categories are defined in the Night Vision Thermal Imaging Systems Performance Model [1], also referred to as the Johnson criteria, to provide a basis for computing the distance at which a proposed system can produce an image of a specific target usable for an intended task. Figure 2 shows an example.

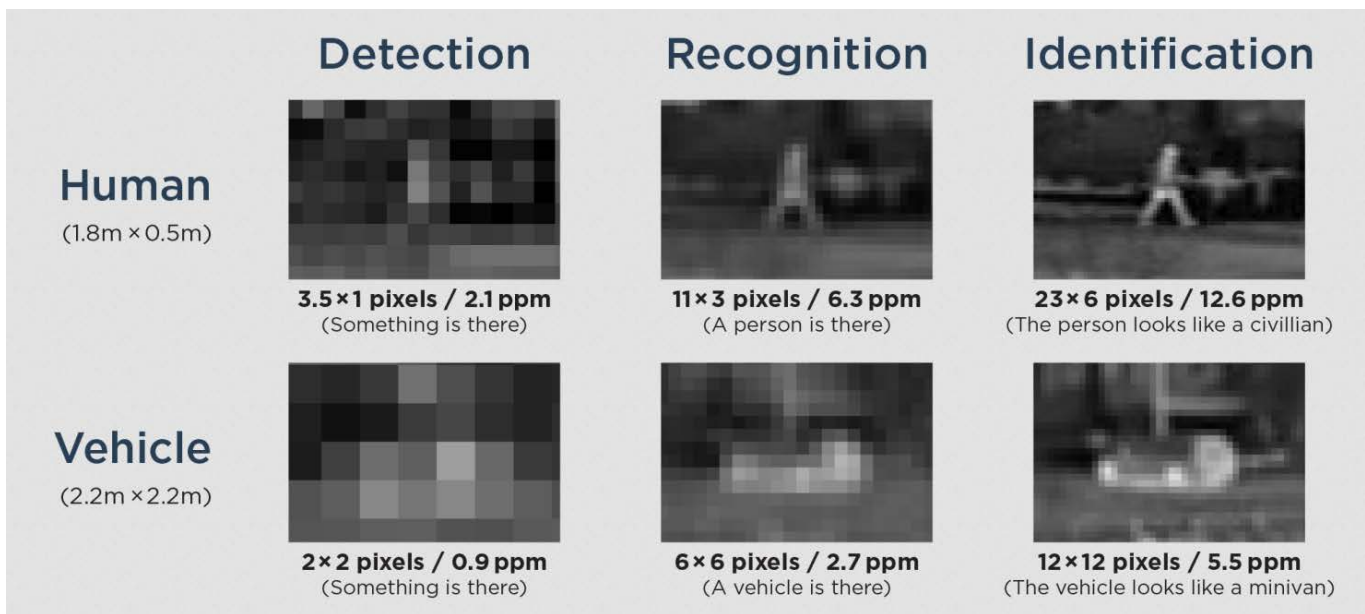


Figure 2 Typical DRI requirements

CREDIT: INFINITI ELECTRO-OPTICS

The DRI criteria can be applied to any image classification system including thermal, visible, or LiDAR. Understanding the Identification requirements is helpful when making resolution, range, and field of view (FoV) trade-offs when specifying an imaging system. Typically, Detection is insufficient for ADAS, which requires Recognition or Identification of VRUs for reliable emergency breaking decisions.

CNN FUNDAMENTALS

Convolutional neural networks are computer simulations of pliable groups of neurons that operate by configuring themselves to produce a positive response when they detect a strong correlation between objects in a new image and objects in a series of images that they have processed in training. CNN training is similar to the method used to train humans for recognition tasks. In training radiologists to read x-rays, for instance, the saying is, “show interns 50,000 images and tell them what they mean and then they will be able to quickly decide what any new images show”. The trick is to pick the right set of images for training to minimize both missed pathology and overinterpretation.

CNN training curves profile how the recognition accuracy of a typical CNN change with exposure to an image set. In training, the CNN is presented with images which are identified according to what the CNN should recognize. Initially, the CNN knows nothing, so its accuracy is very low and its failure rate (loss) is very high. With additional training, the CNN improves in accuracy and the loss drops. Typically, the training data set is presented to the CNN many times in sessions known as epochs to reinforce the selectivity of the CNN.

However, it is possible that the repeated training data presentations can cause the CNN to ultimately recognize only those images that very closely resemble the training images. During training, the loss drops more and more as the CNN becomes more accurate in recognizing the training images. The reversal in loss only appears when images that were not used for training are presented to validate the function of the CNN. This phenomenon, called overfitting, can be seen in Figure 2 [2]. Attempts to make the CNN perform perfectly during training almost always lead to additional loss with new images. Care must be taken to select for use a stage in training when the CNN operation is near its optimum.

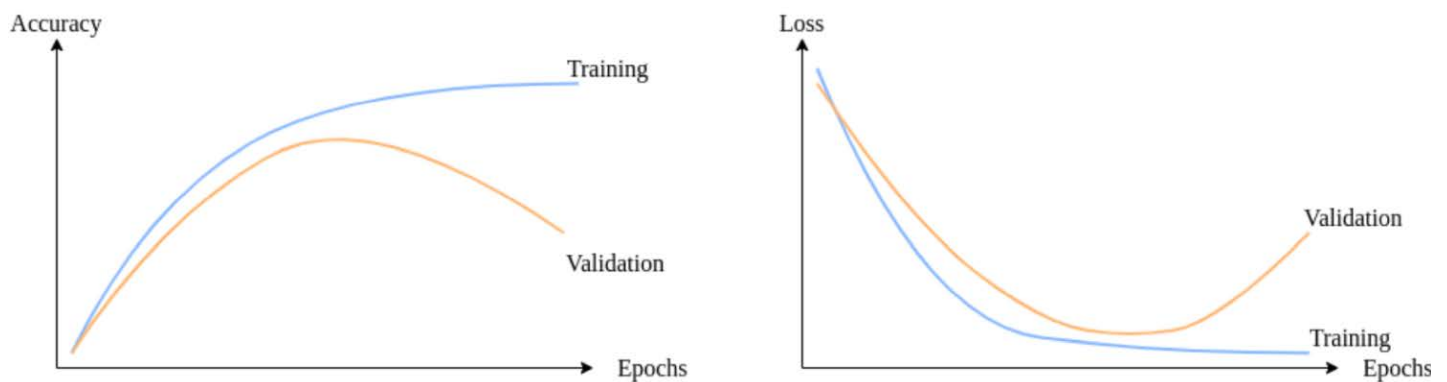


Figure 3 CNN training curves

CREDIT: BAELDUNG.COM

CNN SUPERVISION

Neural networks can be told what reality is. On their own, they can determine discrimination features, but they do not know what these features mean. Train a CNN on pedestrian images and when it is shown a new pedestrian image, the “pedestrian” feature set combination will light up but without supervision, the CNN will not be able to tell an operator or another computer that it has detected a pedestrian. Supervision, informing a CNN what it has found, must be applied during training for the CNN to be equipped to report findings in real terms.

In the Thermal Ranger system, one of these findings is the distance assigned to each pixel. When the CNN provides a results map for all the pixels it must be told during training what these results mean in terms of real distances, typically in meters. Supervision is not strictly necessary because a separate algorithm could be applied to accept the output of the CNN and convert it to distance, but this method requires careful analysis of the CNN output and is, therefore, a source of potential uncertainty.



Figure 4 Base Localization Depth Map

CREDIT: S. MAHDI, ET. AL.

TRAINING THE THERMAL RANGER SYSTEM

The Thermal Ranger system needs two types of scene information to formulate its report on the scene contents, the identification of each VRU with its location in the image and the distance from the camera to each identified object. Both of these determinations are performed by CNNs.

Recognition and locating objects are a widely implemented CNN operation. For this operation, the CNN accepts images and outputs a map of features, roughly the location and orientation of edges indicating transitions in gray level or color. This edge map is passed to a second stage of the CNN to recognize, in the arrangement of edges, specified types of objects. The details of this process are explained in the videos referenced at the end of this white paper.

Using a CNN to generate a depth map from single monochrome thermal images, however, is not common. Thermal images have very different spatial frequency content from visible monochrome or RGB images, requiring special proprietary training and calibration techniques to be developed. The goal of building a depth map in the system is to label each of the significant objects in the image identified by the recognition CNN with a depth that matches the distance from the camera as measured by conventional means - the “ground truth”.

Using a CNN to generate a depth map from single monochrome thermal images, however, is not common. Thermal images have very different spatial frequency content from visible monochrome or RGB images, requiring special proprietary training and calibration techniques to be developed.

In practice, what can be extracted from a single image is a map of disparity, a measure proportional to the inverse of the distance. Typically, disparity maps are dependent on the local area around each pixel, generating data that corresponds to the known contours of target objects in the training data set. For example, Figure 3 shows a group of men sitting on a girder high in the sky with a row of buildings below [3]. Note that while the depth contours of the groups of men and buildings, as encoded in color, are each reasonably determined, the CNN reports that the groups of men and buildings are at roughly the same distance from the camera and cover the same depth range.

buildings, as encoded in color, are each reasonably determined, the CNN reports that the groups of men and buildings are at roughly the same distance from the camera and cover the same depth range.

Preventing this behavior requires careful training on the target objects so that objects of different types in the same image are placed correctly relative to each other. The training regimen used by Owl AI meets the requirement partly by using a carefully designed training set and partly by incorporating disparity indicators in the images intrinsic to the camera itself.

WHAT THE THERMAL RANGER SYSTEM NEEDS TO SEE

What is desirable is a system capable of recognizing and locating VRUs in clear and degraded visual environments, day and night. Recognizing and locating are two quite different functions having a certain hierarchy. The order of events goes something like this:

1. Accept raw data from a camera and perform any required geometric normalization so that each type of object appears the same in all image locations.
2. Apply a CNN to the task of recognizing all objects of interest and identifying them according to type, this process is commonly referred to as classification.
3. Using the classification data, assign the locations for the sides of a two-dimensional bounding box and a class to each object.
4. Simultaneously with step 2, apply another CNN to produce a range map for the entire image, assigning to every pixel in the image a value representing the distance from the camera to the object imaged by that pixel.
5. Convert the values determined by the range CNN into distances in real units, typically meters.
6. Combine the information from the two CNNs to add a third dimension (depth) to the bounding boxes and assign a distance to the nearest face of the bounding box.
7. Use color to code the depth map for the entire image to produce an informative picture for display.
8. Assemble the type, location, size and range of all objects of interest in the image for reporting to equipment intended to take appropriate action.

Preventing this behavior requires careful training on the target objects so that objects of different types in the same image are placed correctly relative to each other. The training regimen used by Owl AI meets the requirement partly by using a carefully designed training set and partly by incorporating disparity indicators in the images intrinsic to the camera itself.

Figure 4 shows a block diagram of a sample implementation of this sequence.

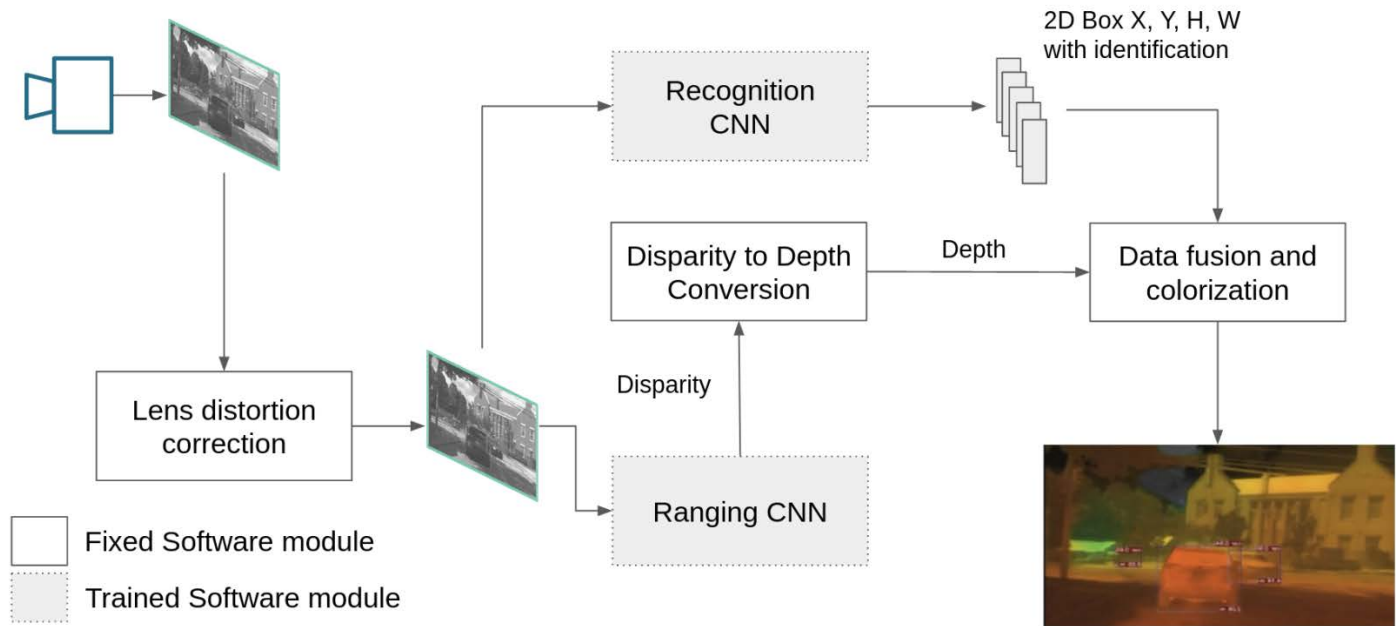


Figure 5 Sample process pipeline

This entire set of processes combining CNN and conventional computation is called an inference pipeline (IP). Partitioning the functions between the CNN and conventional sections is directed toward maximizing accuracy and reliability of the results. Since the Thermal Ranger system is intended for deployment on real vehicles in real situations, the entire IP is capable of field updating whenever better training data and conversion algorithms become available.

WHY MEASURING DISTANCE IS BETTER THAN GUESSING IT

Figure 3 above shows the danger in depending on a CNN to determine distance by examining objects in the scene. To simplify, the distance assigned to each of the men is dependent on their relative size – smaller means farther away. The algorithm used smooths out the variations from the random size variations of the subjects and determines an average slope to the distance then superimposes distance details from each subject. The same process produces similar results with the buildings. While this produces smooth estimates of depth for the two groups, it provides no information on the relative depth between them. This occurs because the image is already reduced to a flat plane – a photograph – so there is no real disparity data in the image to detect.

Most monocular systems intended for automotive use apply the same technique – analysis of the detected objects to determine their distance. Generally, this involves developing a table of sizes for the various types of target objects and comparing the size of the bounding boxes around detected objects with values in the tables – so-called “pixel counting”.

The Thermal Ranger™ system avoids size comparison errors by making direct distance measurements, generating its disparity map using a CNN trained to match the characteristics of objects at various distances and the modifications to these images caused by intrinsic properties of the camera itself.

This might be reasonably reliable if all the VRUs of each type were the same size but a man bending down to tie his shoelace is about the same size as a standing 10-year-old. If the system recognizes this target as a man, his distance will be significantly overestimated. Unfortunately, almost all the errors these systems make are dangerous distance overestimations.

The Thermal Ranger™ system avoids size comparison errors by making direct distance measurements, generating its disparity map using a CNN trained to match the characteristics of objects at various distances and the modifications to these images caused by intrinsic properties of the camera itself. Further, Owl AI is well along in the development of a thermal sensor designed specifically to maximize CNN depth estimation performance in automotive settings.

HOW THE RIGHT THERMAL CAMERA CONTRIBUTES TO GOOD RESULTS

To maximize the object fusion success, the location and range data should be as closely matched as possible in resolution, frame rate and scene perspective with minimum interruptions in scene acquisition for recalibration. Thus, a single shutterless thermal camera coupled with image processing capable of extracting both location and range data would meet all these challenging requirements.

MONOCULAR THERMAL RANGING

Practical and cost-effective thermal ranging calls for developing a camera and processor that are optimized for the application: detection and ranging of VRUs from a moving automobile. The requirements are straightforward:

- Detection of humans (and animals) by their own emitted body heat using thermal imaging in the 8-14 μm thermal band with a single low-cost sensor
- Sufficient resolution to enable detection and ranging curb to curb
- A camera that does not need to be shuttered periodically for recalibration

Continued on page 9 _____

Continued from page 8

- **CNNs trained to extract location and range information from a monocular thermal image**
- **Output data formatted for use in automotive AEB systems**
- **Optionally, output video for driver use including markers showing detected VRUs with their ranges.**

These requirements are intended to reduce any variability in the images sent to the CNN not part of the target objects themselves to reduce the need for normalization processing before the images are presented to the CNN. While images with various distortions could be included in the training set the overall effect of increasing variety during training is to produce a larger minimum loss figure in operation, an outcome clearly at odds with the purpose of the system.

Currently, no commercial sensor incorporates all the features needed to supply the operating modes and the data quality demanded by the requirements above, so design of a custom sensor was undertaken by Owl. The final design consists of a VOx microbolometer array incorporating features that provide continuous thermal reference signals, allowing uninterrupted shutterless operation, an imperative for safe application within moving vehicles. Further, the 1280 (H) x 800 (V) resolution of this array is sufficient to support a wide horizontal field of view with a single camera. To support simple implementation, the array itself includes many of the traditional image signal processing (ISP) functions, providing clean, continuous images to the CNN without complex external processors.

THE OWL THERMAL RANGER CNN IN ACTION

Three steps in the identification and ranging process are illustrated in Figure 5. The original monocular thermal image provided to the IP is shown at the lower right. In the center is a bird's eye view of the point clouds representing target objects extracted from the image with a blank field for all areas not trained for identification by the CNN. The identified target objects (pedestrians and automobiles) are overlaid with color-coded range estimates and surrounded by 3D bounding boxes. At this stage, all coordinates (x,y,z) have been converted from the pixel domain to a 3D domain in meters. At the upper left is the driver view representation of the output image showing the recognized objects with bounding boxes and range labels set in their natural surroundings.

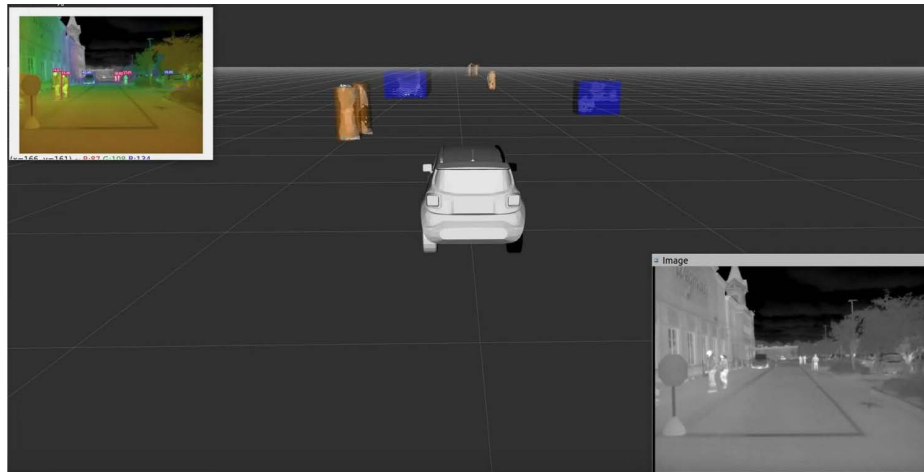


Figure 6 Stages in identification and labeling

Figure 7 is an expanded view of the driver view output image from the upper left of Figure 6. In this view, the pedestrian thermal images can be seen with their associated 2D bounding boxes and range labels. Notice, on the left, the distinction in range between two pedestrians walking together. In this particular example, the software labels pedestrian ranges up to 50 meters with specific values while pedestrians outside that range are labeled as pedestrians (ped) but do not have range labels. Note that the automobile headlights in the distance are not found to be pedestrians and that other automobiles are detected and marked but not ranged.

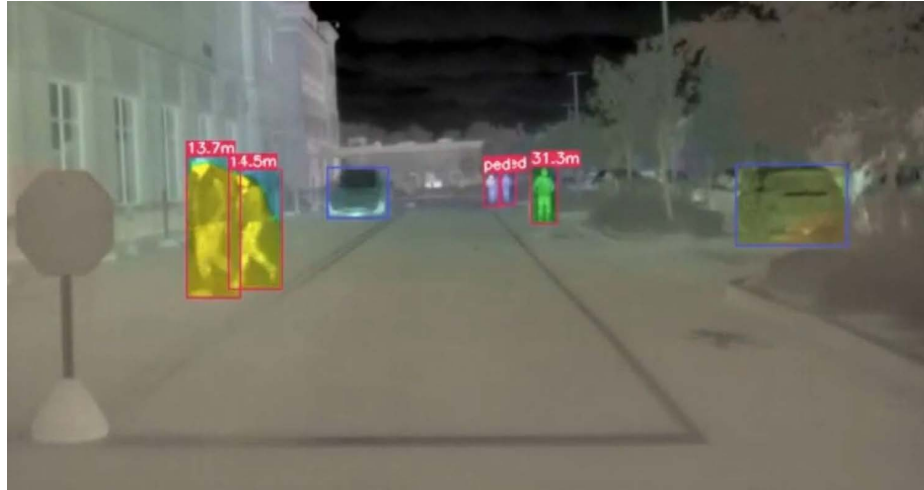


Figure 7 Labeled pedestrians and detected cars

OBJECT FUSION

For the safe deployment of ADAS or fully autonomous vehicles comprehensive identification of all the object types in Figure 7 is essential. While many of these objects can be classified based on only their visible images in the daytime, thermal data can provide the additional information needed to sort activated objects (human or machine) from the static background. For instance, is a truck at the curb preparing to pull out? Thermal imaging can help determine if the engine is running.

Whether for ADAS or autonomous vehicles, continuous operation is essential. The system described here is always on, never blacked out while recalibration is underway. Not only does this avoid interruptions in the thermal data, but it also avoids delays resulting when the thermal data fusion with color images must be re-established. CNNs are always on, operating more efficiently when their input data is uninterrupted. Real-time systems using CNNs should support this operation.

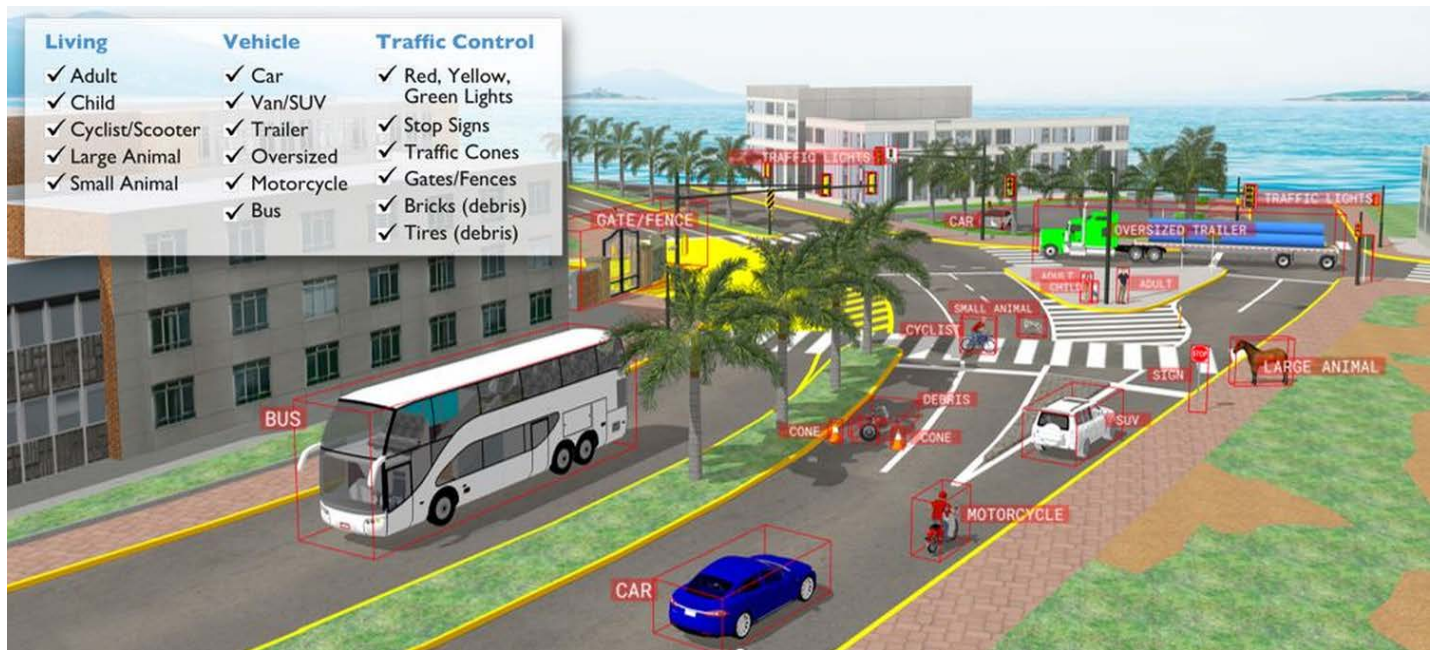


Figure 8 A traffic scenario

REFERENCES

[1] U.S. Army Night Vision and Electronic Sensors Directorate, "Night Vision Thermal Imaging Systems Performance Model (Revision 5)", AMSEL-RD-NV-MS-SPMD, Fort Belvoir, VA, March 2001

[2] Tarnum Java SRL, "Epoch in Neural Networks", <https://www.baeldung.com/cs/epoch-neural-networks>, accessed 27 June 2022

[3] Mahdi, S., Miangoleh, H., Dille, S., et. al., "Boosting Monocular Depth Estimation Models to High-Resolution via Content-Adaptive Multi-Resolution Merging", In Proc CVPR 2021, presentation at <https://youtu.be/IDeI17pHIqo>



YouTube LINKS

CONVOLUTIONAL NEURAL NETWORKS

How Do They Work: <https://youtu.be/QzY57FaENXg>

Filters in Operation: <https://youtu.be/f0t-OCG79-U>

Why Do They Learn: <https://youtu.be/OQczhVg5Hal>

How Do They Remember: <https://youtu.be/piF6D6CQxUw>

How to Find Objects Fast: <https://youtu.be/NM6lrxyObxs>

Training/Validation/Test Data Sets: <https://youtu.be/Zi-OrIM4RDs>

DEMONSTRATIONS OF OWL AI TECHNOLOGIES

April 2022 Pedestrian Demonstration:

<https://youtu.be/JaaTngahlms>

January 2022 Pedestrian Comparison with Visible:

https://youtu.be/TmfzYcGRH_Y

November 2021 Pedestrian and Automobile Classification:

<https://youtu.be/BMGLgnxNI6M>

Example videos of our **THERMAL RANGER** in action can be found on our [YouTube Channel](#) at this link >>>



“OWL’S THERMAL RANGER is unique as it delivers rich detail and 3D response day or night.”



Rochester:
470 WillowBrook Office Park
Building 400, 2nd Floor
Fairport, NY 14450

www.owlai.us



01/03/23