2021 International Conference on Unmanned Aircraft Systems (ICUAS)
Athens, Greece. June 15-18, 2021

Aerial Thermal Image based Convolutional Neural Networks for Human Detection in SubT Environments

Anton Koval[1], Sina Sharif Mansouri[1], Christoforos Kanellakis[1] and George Nikolakopoulos[1]

Abstract— **This article proposes a novel strategy for detecting humans in harsh Sub-terranean (SubT) environments, with a thermal camera mounted on an aerial platform, based on the AlexNet Convolutional Neural Network (CNN). A transfer learning framework will be utilized for detecting the humans, where the aerial thermal images are fed to the trained network, which binary classifies them image content into two categories: a) human, and b) no human. Moreover, the AlexNet based framework is compared with two related popular CNN approaches as the GoogleNet and the Inception3Net. The efficacy of the proposed scheme has been experimentally evaluated through multiple data-sets, collected from a FLIR thermal camera during flights on an underground mining environment, fully demonstrating the performance and merits of the proposed module.**

I. INTRODUCTION

The continuous development of robotic technologies is constantly increasing the number of real life and mission oriented deployments of these technologies, in a variety of environments and of different complexity. Among the latest major trends in the robotics community is the focus on demonstrating resilient and robust autonomy in Subterranean (SubT) environments, as for example the case of the DARPA Sub-T competition [1]. In such applications, one of the main challenge and real requirement for the robotic autonomy frameworks, is the ability to ensure a safe and robust navigation of the robotic platforms, while operating in a priori known, spatially or completely unknown environments. Except from the fundamental problems of localisation, navigation and mapping, additional feaures towards realistic missions should be added. As such, the missions for search and rescue are gaining lately a massive attention, especially for cases of operations in hostile environments, such as mining tunnels [2], [3]. In such search and rescue missions, a visible light camera is the commonly selected perception sensor. However, the quality of the visual information may significantly degrade, while working at dark environments and as such, there have been presented various lightning strategies to guaranty proper illumination of the environment [4]. Recently, there is a large trend to equip the robots with thermal cameras [5] that have the ability to perceive visual information in the dark environment and as such, increasing the overall environment perception for the autonomy navigation itself but also to introduce new features in the overall mission, as for example the case of search and rescue for humans and survivors [4].

Few works have addressed the task of human detection with a thermal camera [5]. In [6], the authors presented an approach that allows detecting humans in a real-world outdoor environment using a thermal and a visible light camera. However, the performance of this technique depends on the high video processing rate in order to find all the potential objects, while the method was experimentally evaluated only during the day time conditions. Another method to discover humans with a thermal cameras was presented in [7]. The human detection algorithm begins with a static analysis through the classical image processing approach and in the sequel, the dynamic image analysis follows. Finally, the outcome of these two stages is compared in order to achieve better human detection, however, this approach requires an approximate image alignment. The authors in [8] presented a trespasser detection method, which utilized pattern recognition techniques to distinguish humans. The proposed algorithm was experimentally evaluated during the day time conditions and only in an urban environment.

For the general detection of humans, the study [9] addressed the faint detection among elderly people, patients or pregnant women. The proposed faint detection algorithm utilized a thermal camera and could work in both indoor and outdoor conditions. However, it needs to be improved in order to distinguish humans from animals. In the study [10], the authors presented a new algorithm for human detection with a thermal camera. In their approach, the thresholds for generating the binarized image difference, between the input and the background reference images, could be adaptively calculated by using the information obtained from the background image and differential values between the background and input image, based on fuzzy systems. But, similarly to the previous study [9], the animals can be faulty detected as humans. The authors in [11] proposed a human silhouettes extraction from thermal and visible light cameras. The matching method optimization problem was solved with the use of a hierarchical genetic algorithm, where the underlying experiments have indicated good results in day time conditions but the method's accuracy may decrease in the low light. In another study in [12] a multi-spectral pedestrian data-set that contained both thermal and visible light camera images was proposed and evaluated through multi-spectral extension of aggregated channel features. The authors in [13] presented a human detection method for thermal camera that focuses on the combination of the pixel-

*This work has been partially funded by the European Unions Horizon 2020 Research and Innovation Programme under the Grant Agreement No. 869379 illuMINEation.

[1]Robotics and Artificial Intelligence Team, Department of Computer Science, Electrical and Space Engineering, Luleå University of Technology, SE-97187, Sweden {antkov,sinsha,chrkan,geonik}@ltu.se

2021 International Conference on Unmanned Aircraft Systems (ICUAS) | 978-1-6654-1535-4/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICUAS51884.2021.9476853

gradient and body parts processing, also in a three-stage classification process, which has been proposed to reduce the false detection, however, still this method cannot detect groups of overlapped people.

A common disadvantage of all the previous mentioned methods is that none of them was evaluated in harsh environmental conditions, while there are few works that addressed the task of human detection in challenging environments. In [14], the authors' goal was to maximize the situational awareness of the firefighters. To achieve this, a CNN VGG16, coupled with a thermal camera, was used. The test results showed a performance that exceeds a classification probability above 95% in all the classifications when the configuration had 4 layers of convolutional sections. The authors in [15] introduced a human detection algorithm in underground mines and their experiments presented a neural network classifier shows reasonable performance and accuracy.

The main contribution of this work is the implementation of a vision based approach for human detection with an aerial thermal camera. Unlike other studies, the proposed system solves the image classification task through a fine tuning AlexNet [16] with transfer learning [17], adding new classification categories. The proposed method is generic and can be applied to any thermal camera with reasonable resolution for a real-time humans' detection. Additionally, the proposed method was trained through an open access pedestrian data-set and evaluated using a data-set captured in an underground environment. Finally, the proposed framework has been also compared with other two popular CNNs, namely the GoogleNet and the Inceptionv3Net for overall benchmarking purposed.

The rest of the article is structured as it follows. Initially, Section II presents the AlexNet architecture and the transfer learning solution, while Section III presents the collected data-set, the network training and its successful evaluation results, including the comparison studies. Finally, the conclusions are drawn in Section IV.

II. ALEXNET FOR HUMAN DETECTION

In the present article, the overall framework for the visual object recognition utilises a well-known CNN method in [18] called AlexNet [19]. Thus, the section initially provides a brief description of the AlexNet framework, while in the sequel the concept of transfer learning is explained.

A. AlexNet

AlexNet [20] is one of the most widely used CNN methods [21] for image classification. It has in total 60 million parameters, 650.000 neurons, contains 5 convolutional layers and allows to classify images in 1000 different class labels. Therefore, a big data-set is needed for its training and in this article the transfer learning method is chosen due to the limited availability of the thermal imagery data-sets of humans in the SubT environments.

In the presented approach, the AlexNet input is an Red, Green and Blue (RGB) image from a thermal camera with a fixed size of $227 \times 227 \times 3$ pixels, that is subsequently followed with a 11×11 2D convolution layer with an output size of $55 \times 55 \times 96$. The overall proposed AlexNet architecture is presented in Figure 1.

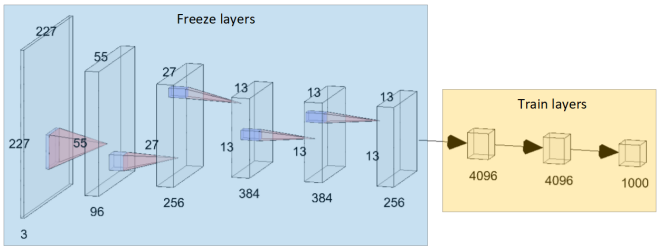

Fig. 1: AlexNet architecture and the transfer learning method.

After this initial stage, there is a 2D Max pooling layer of 3×3 and a $27 \times 27 \times 96$ output, followed by 2-dimensional convolution layers of 5×5 and an output of $27 \times 27 \times 256$. In the continuation, there is also another max pooling layer of size 3×3 and with an output size of $13 \times 13 \times 256$, which runs through 2D convolution layers of 3×3 with an output of the same size. After that, another 3×3 2-dimensional convolution layer with an output size of $13 \times 13 \times 256$, is followed by a 3×3 max pool with an output size of $6 \times 6 \times 256$. This output goes via two fully connected layers and the latter results are fed into a 1000 class label softmax classifier. To sum it up, AlexNet is made of eight layers, five of which are convolutional layers and three of those are fully connected layers. The first two convolutional layers are followed by an Overlapping Max Pooling layer. The remained three convolution layers (third, fourth, and fifth) are directly linked together. Eventually, an Overlapping Max Pooling layer is followed by the last convolution layer (fifth).

Max pooling layers are usually used in CNNs in order to reduce the size of the matrices, while keeping the depth the same. On the other hand, overlapping max pooling uses an adjacent window, which overlaps each other in order to compute the max element from a window each time. It has been proven that this kind of max pooling reduces the top-1 and top-5 error rates [20], an approach that has been followed in this article as well.

One of the main aspects of the AlexNet is the use of the Rectified Linear Unit (ReLU) [22]. The authors [20] proved that by using the ReLU nonlinearity, AlexNet could be trained a lot quicker than using classical activation functions like $sigmoid$ or $tanh$ [23], where their hypothesis have been tested on the CIFAR-10 data-set [24] and the ReLU-AlexNet achieved the same performance (25% training error) with the Tanh-AlexNet in one sixth of the epochs.

B. Transfer Learning

Transfer learning [22], [25] for CNNs is commonly referred to as the method of applying a previously trained CNN in another data-set, where the number of classes to be identified is different from the initial data-set, because it was used in different tasks and with different data-sets. There are two major methods for Transfer Learning, both of which

use the same AlexNet weights on the ImageNet database images [26]. In the first approach, the CNN is considered as a feature extractor, while the last fully connected layer is removed. In the sequel, the features that were extracted from the trained AlexNet can be used to train a classifier similar to [27] in the new data-set. In the second approach, the last connected layer is replaced and later the entire CNN is retrained for the new data-set so that the trained weights are fine tuned.

In the present article, we are proposing the replacing of the last three layers that are configured for 1000 categories (Figure 1) with fully connected layer, softmax layer and classification output layer, of the AlexNet network and fine-tune them to our desired two classes (human, no human). This approach allows accurate classification based on the detected features, while reusing previously trained network layers.

III. EXPERIMENTAL RESULTS

In this Section, the aerial platform utilized in the acquired experimental results will be described, including additional information for the placement of the camera and the overall concept targeting UAVs. Furthermore, the description of the Advanced Driver-Assistance Systems (ADAS) data-set, provided by FLIR, will be presented, which was used for the network training, as well as the experimentally collected data-sets from the Swedish underground tunnels for the CNN evaluation. Finally, the overall training approach will be presented, extended with a discussion on the evaluation results.

A. The Aerial Platform

The evaluation of the proposed method has been performed on data collected from the onboard thermal camera of a UAV navigating along an underground tunnel. The aerial platform is a custom build quadrotor developed at Luleå University of Technology. It carries the AfroFlight Naze 32 Rev6 Flight Controller Unit (FCU), running the ROS-Flight embedded autopilot software [28]. Moreover, it weights 1.5kg and it is powered using a 4-cell 1.5hA LiPo battery which provides around 10 minutes of flight. The processing unit of the platform is the Aaeon UP-Board with the processor Intel Atom x5-Z8350 and 4GB of RAM memory, running Ubuntu 18.04, while all autonomy capabilities have been developed within the Robot Operating System (ROS). The sensor suite of the platform includes, 1) the Prophesse thermal camera Gen3M VGA-CD 1.1 with 70° Field of View (FOV) at 30fps placed in the front part of the vehicle, 2) the PX4Flow optical flow sensor faced towards the ground and is used for velocity estimation on x, y axes, 3) the Lidar Lite v3 single beam lidar facing towards the ground for altitude information and 4) the Rplidar S1, a 2D Time of Flight (TOF) Laser Range Scanner placed on top of the platform main body for extracting 2D information from the surrounding environment used in the autonomy capabilities of the platform. The underground areas are pitch dark, thus the platform is equipped with two 10 W LED light bars placed in the quadrotor arms for additional source of illumination. Figure 2 depicts the aerial platform with the sensor suite.

Fig. 2: The aerial platform before take-off at the underground tunnel

The presented quadrotor is considered an aerial scout resource constrained platform for fast deployment and autonomous navigation in underground complex areas. The thermal camera payload allows to collect information of the visited areas based on their thermal signature and detect humans in the context of search and rescue missions, using the proposed method.

B. FLIR ADAS Thermal Data-set

In the transfer learning approach of the pre-trained AlexNet, it was utilized the ADAS data-set that allows to detect and classify walking and cycling persons, dogs and vehicles in challenging conditions including total darkness, fog, smoke, inclement weather and glare [29]. The data-set was recorded with a FLIR Tau2 camera, operated at 30fps and with a resolution of 640×512 pixels. In our case, Figure 3 depicts an example of the extracted images from the training data-set. The second data-set is collected from Luleå Sweden underground mining tunnels that was collected from an aerial flight with a FLIR Boson 640 camera, a resolution of 640×512 pixels and a frame rate of 60fps. Figure 4 present a snapshot of the aerial platform utilized for gathering the data sets suring a full autonomous mission. In this frame, the thermal camera view is also depicted in the bottom right, while a full video of one of the missions can be reached to the following link https://youtu.be/0gvXjWbPLOA.

Since the thermal camera resolution is significantly lower than the visible light has, the acquired images were not down-sampled or cropped. Table I shows the total number of images that were extracted from the data-set, while Figure 7

Fig. 3: Examples of the extracted images from the FLIR ADAS data-set in the case of a human detection. The images are extracted while the camera approaches to the human.

illustrates several images collected for the human detection in the underground tunnel.

TABLE I: The number of extracted images for each category from the training and the validation data-sets, while the redundant images are excluded.

	human	*no humans*
FLIR ADAS data-set	79	174
Sweden tunnels data-set	169	68

In the continuation, the data-set is categorized manually into two classes of *human* and *no human* for the training and evaluation stages.

C. Training and the Evaluation

The FLIR ADAS data-set was used for training the AlexNet, while the Luleå Sweden underground tunnels data-set is used for the validation of the network. Additionally, the images acquired from the FLIR ADAS data-set were resized to $227 \times 227 \times 3$ pixels. The CNN was trained on a laptop equipped with an Nvidia MX 150 GPU. The training parameters were: mini-batch size of 10, maximum number epochs of 6, initial learning rate of 10^{-4}. As a solving method, the stochastic gradient descent [30] with a momentum optimizer was used. The trained AlexNet network has a 100% accuracy on the training data-set, while the accuracy on the validation data-set was equal to 98.73%, which is a solid result given the size of the dataset. The outcome of the CNN training, with the corresponding accuracy and loss, while training and validation of the data-set, is depicted in Figure 5. For the classification, the loss function was defined as a cross-entropy loss [22], [31].

Furthermore, Figure 6 shows the confusion matrix of the validation data-set, where the rows from the validation data-set correspond to the predicted class and the columns refer to the actual class of the data-set. The diagonal cells show the number and percentage of the trained CNN's proper classifications. For example, 71 images are correctly categorized in the first diagonal into the category *human*, which corresponds to 31.1 percent of the total number of images. Likewise, 157 instances are correctly classified as *no humans*, corresponding to 68.9% of the entire validation data-set. In addition, the off-diagonal cells correspond to wrongly labelled results, leading to a 0.0% error. Additionally, the right gray column indicates the percentages of all the images that are expected to belong to each class that are correctly and faulty categorized. From the other hand, the bottom row displays the percentages of all the examples belonging to each class, which are listed correctly and faulty. The cell located in the plot's bottom right indicates the overall precision. Overall, 100.0% of predictions were right, and 0.0% were incorrect.

Additionally, the obtained results from AlexNet were compared against two other pre-trained networks, namely the GoogleNet [32] and the Inceptionv3Net [33], while the comparison is depicted in Table II. As it can be seen from this Table, the validation accuracy for GoogleNet is lower, while for the Inception3Net it is significantly lower. It is estimated, that such a result comes due to the larger amount of the convolutional layers in these two networks and the fact that GoogleNet and Inception3Net require larger data-sets. The validation loss indicates that the AlexNet classifier has better performance in modelling relationship between network training and validation sets.

TABLE II: The comparison of transfer learning performance between AlexNet, GoogleNet, and Inceptionv3Net.

	AlexNet	GoogleNet	Inceptionv3Net
Training Time [sec]	186	265	2012
Training Accuracy	100%	100%	100%
Validation Accuracy	98.73%	77.64%	29.11%
Training Loss	0.0036	0.008	0.17
Validation Loss	0.041	0.51	1.04

IV. CONCLUSIONS

This work proposed a novel framework based on CNN for detecting humans in SubT environments. The main focus of the developed framework is to provide a generic solution that has a reduced computational cost and a very good performance for detecting humans, while relying only on a thermal camera video stream. The AlexNet CNN has

Fig. 4: UAV performing a full autonomous search and rescue mission, based on an onboard thermal camera video feed https://youtu.be/0gvXjWbPLOA.

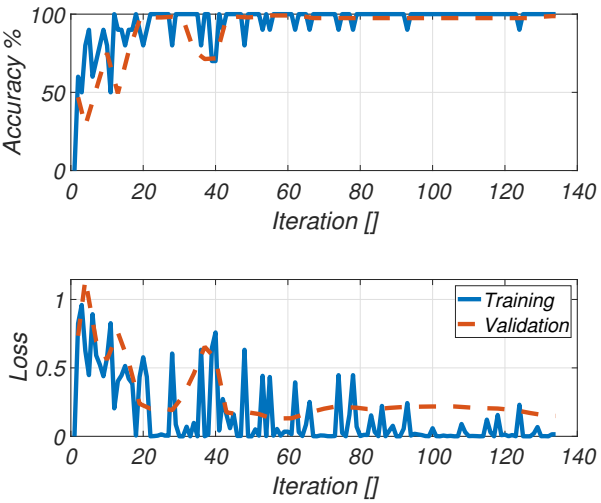

Fig. 5: Accuracy and loss of the AlexNet network on training and validation the data-sets.

Fig. 6: The confusion matrix from the validation data-set.

been trained through a transfer learning approach, mainly for tackling the issue of limited training data-set availability, from real underground environments. The thermal images are fed to the network, which classifies into two categories: human and no human. The method has been validated by using data-sets collected from autonomous flying missions at real underground environments, while the overall efficiency of the proposed approach has been presented and compared with other similar training techniques.

REFERENCES

[1] DARPA. (2020) DARPA Subterranean (SubT) challenge. Accessed: February 2021. [Online]. Available: https://www.darpa.mil/program/darpa-subterranean-challenge

[2] C. Kanellakis, S. S. Mansouri, M. Castaño, P. Karvelis, D. Kominiak, and G. Nikolakopoulos, "Where to look: a collection of methods formav heading correction in underground tunnels," *IET Image Processing*, vol. 14, no. 10, 2020.

[3] T. Tomic, K. Schmid, P. Lutz, A. Domel, M. Kassecker, E. Mair, I. L. Grixa, F. Ruess, M. Suppa, and D. Burschka, "Toward a fully autonomous uav: Research platform for indoor and outdoor urban search and rescue," *IEEE robotics & automation magazine*, vol. 19, no. 3, pp. 46–56, 2012.

[4] C. Kanellakis, S. S. Mansouri, G. Georgoulas, and G. Nikolakopoulos, "Towards Autonomous Surveying of Underground Mine Using MAVs," in *International Conference on Robotics in Alpe-Adria Danube Region*. Springer, 2018, pp. 173–180.

[5] R. Gade and T. B. Moeslund, "Thermal cameras and applications: a survey," *Machine vision and applications*, vol. 25, no. 1, pp. 245–262, 2014.

[6] P. Rudol and P. Doherty, "Human body detection and geolocalization for uav search and rescue missions using color and thermal imagery," in *2008 IEEE aerospace conference*. Ieee, 2008, pp. 1–8.

[7] A. Fernández-Caballero, J. C. Castillo, J. Martínez-Cantos, and

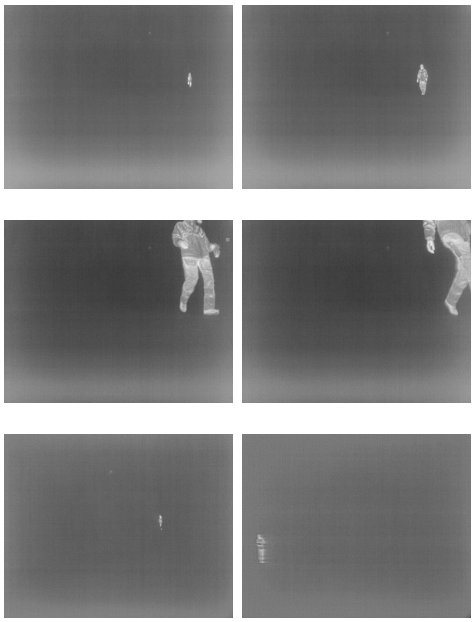

Fig. 7: Examples of acquired images from the Luleå Sweden underground tunnels for the validation data-set.

R. Martínez-Tomás, "Optical flow or image subtraction in human detection from infrared camera on mobile robot," *Robotics and Autonomous Systems*, vol. 58, no. 12, pp. 1273–1281, 2010.

[8] W. K. Wong, Z. Y. Chew, C. K. Loo, and W. S. Lim, "An effective trespasser detection system using thermal camera," in *2010 Second International Conference on Computer Research and Development*. IEEE, 2010, pp. 702–706.

[9] W. K. Wong, H. L. Lim, C. K. Loo, and W. S. Lim, "Home alone faint detection surveillance system using thermal camera," in *2010 Second International Conference on Computer Research and Development*. IEEE, 2010, pp. 747–751.

[10] E. S. Jeon, J. H. Kim, H. G. Hong, G. Batchuluun, and K. R. Park, "Human detection based on the generation of a background image and fuzzy system by using a thermal camera," *Sensors*, vol. 16, no. 4, p. 453, 2016.

[11] J. Han and B. Bhanu, "Fusion of color and infrared video for moving human detection," *Pattern Recognition*, vol. 40, no. 6, pp. 1771–1784, 2007.

[12] S. Hwang, J. Park, N. Kim, Y. Choi, and I. So Kweon, "Multispectral pedestrian detection: Benchmark dataset and baseline," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1037–1045.

[13] S. Budzan, "Human detection in thermal images using low-level features," *Measurement automation monitoring*, vol. 61, 2015.

[14] M. Bhattarai and M. Martínez-Ramón, "Detection and identification of objects and humans in thermal images," *arXiv preprint arXiv:1910.03617*, 2019.

[15] J. Dickens, J. Green, and M. Van Wyk, "Human detection for underground autonomous mine vehicles using thermal imaging," 2011.

[16] S.-H. Wang, S. Xie, X. Chen, D. S. Guttery, C. Tang, J. Sun, and Y.-D. Zhang, "Alcoholism identification based on an alexnet transfer learning model," *Frontiers in Psychiatry*, vol. 10, p. 205, 2019. [Online]. Available: https://www.frontiersin.org/article/10.3389/fpsyt.2019.00205

[17] M. Hussain, J. J. Bird, and D. R. Faria, "A study on cnn transfer learning for image classification," in *Advances in Computational Intelligence Systems*, A. Lotfi, H. Bouchachia, A. Gegov, C. Langensiepen, and M. McGinnity, Eds. Cham: Springer International Publishing, 2019, pp. 191–202.

[18] M. Z. Alom, T. M. Taha, C. Yakopcic, S. Westberg, P. Sidike, M. S. Nasrin, B. C. Van Esesn, A. A. S. Awwal, and V. K. Asari, "The history began from alexnet: A comprehensive survey on deep learning approaches," *arXiv preprint arXiv:1803.01164*, 2018.

[19] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[20] ——, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[21] H. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers, "Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1285–1298, May 2016.

[22] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org.

[23] A. Saxena, "Convolutional neural networks (cnns): An illustrated explanation," *URL https://xrds. acm. org/blog/2016/06/convolutional-neural-networks-cnns-illustrated-explanation/. Last updated*, pp. 06–29, 2016.

[24] A. Krizhevsky, V. Nair, and G. Hinton, "Cifar-10 (canadian institute for advanced research)," 2010. [Online]. Available: http://www.cs.toronto.edu/~kriz/cifar.html

[25] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, Oct 2010.

[26] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *CVPR09*, 2009.

[27] M. A. Hearst, "Support vector machines," *IEEE Intelligent Systems*, vol. 13, no. 4, pp. 18–28, Jul. 1998. [Online]. Available: http://dx.doi.org/10.1109/5254.708428

[28] J. Jackson, G. Ellingson, and T. McLain, "ROSflight: A lightweight, inexpensive MAV research and development tool," in *2016 International Conference on Unmanned Aircraft Systems (ICUAS)*, June 2016, pp. 758–762.

[29] F. S. Inc. Flir thermal dataset. [Online]. Available: https://www.flir.com/oem/adas/adas-dataset-form/

[30] K. P. Murphy, *Machine learning: a probabilistic perspective*. MIT press, 2012.

[31] P. Kim, "Matlab deep learning," in *With Machine Learning, Neural Networks and Artificial Intelligence*. Springer, 2017.

[32] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[33] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.