# Extending Deep Neural Network Trail Navigation for Unmanned Aerial Vehicle Operation within the Forest Canopy

Bruna G. Maciel-Pearson, Patrice Carbonneau, and Toby P. Breckon

Durham University, Durham, UK

**Abstract.** Autonomous flight within a forest canopy represents a key challenge for generalised scene understanding on-board a future Unmanned Aerial Vehicle (UAV) platforms. Here we present an approach for automatic trail navigation within such an unstructured environment that successfully generalises across differing image resolutions - allowing UAV with varying sensor payload capabilities to operate equally in such challenging environmental conditions. Specifically, this work presents an optimised deep neural network architecture, capable of state-of-the-art performance across varying resolution aerial UAV imagery, that improves forest trail detection for UAV guidance even when using significantly low resolution images that are representative of low-cost search and rescue capable UAV platforms.

**Keywords:** deep learning, trail detection, autonomous UAV, unstructured environment

## 1 Introduction

Challenging activities such as Search and Rescue (SaR) missions [14], visual exploration of disaster areas [2,1] aerial reconnaissance and surveillance [13,6], assessment of forest structure or riverscape [4,12,10] have one thing common: an unstructured environment within which a Unarmed Aerial Vehicle (UAV) could be deployed for autonomous navigation. In most scenarios UAV are manually controlled and it is the pilot who defines and navigates the flight. Growing interest in solving this challenge has motivated researchers to investigate the use of Deep Neural Networks (DNN) to identify trail images for UAV navigation. Within such unstructured environments, a trail represents an existing loosely defined navigation pathway (thoroughfare) used by humans and animals that have transited the environment previously. As such, trails tend to facilitate semi-efficient point-to-point transit routes with lesser obstacle occurrence, hence making them key elements of any effective autonomous navigation in these environments. However, in order to train such a DNN for the trail navigation task, a large volume of labelled data is required, which is challenging to obtain due to the nature of the target task in hand (i.e. sub-canopy UAV operation).

Creative ways to address this data colection issue include gathering data with a head-mounted rig [7], a wide-baseline rig [15], flying the UAV into obstacles

[5] and the use of simulated environments [11,9]. Under any such situation, the generalisation of the resulting DNN model remains constrained due to the fact that only one domain have being used in training.

The data gathered from multiple mounted cameras often entails a high level of discrepancy in illumination between images as we can observe from an image triplet retrieved simultaneously during data gathering of the IDSIA dataset (Figure 1). When comparing the three images from the top row (Figure 1), we can observe that the $forward$ camera tends to capture a much clearer view of the trail, with better illumination than the sideways camera. A similar pattern is also observed on the bottom row and although varied illumination conditions may facilitate the distinction from the $left$ and $right$ images to the $center$ it is not an accurate representation of the real environment experienced by an UAV in flight (Figure 1). Additionally, the features characterizing a trail are typically present in the triplet set with the only discrepancy being the extent of sideways vegetation/obstacle that is found in each image ( illustrated by Figure 1). As a result, for a classification problem we observe that any DNN is essentially only learning how to identify where the UAV is positioned within the environment as opposed to finding the position of the trail (Figure 1).


**Fig. 1.** Example from the IDSIA dataset [7] of varied luminance condition often present when using multiple mounted cameras for image data collection within the forest canopy.

In this scenario the steering decision is usually made by identifying wherever the UAV is flying too close to the vegetation/obstacles which are commonly found on the $left$ or $right$ side of the trail. Based on this information the UAV position can be adjusted by calculating the turning angle [7,15], which tends to lead to a new orientation (of the UAV) towards the center of the trail.

By contrast to early work of [7,15,5] that use a multiple camera approach, our work demonstrates that the same trail direction required for automous UAV navigation can be acquired by using imagery gathered by a single forward-facing

camera (Figure 2). This is due the fact that the center of the forward-facing camera usually shows the trail ahead, for a correctly oriented UAV relative to the trail direction (Figure 2 - centre). Additionally, we demonstrate that a trail can be identified in unseen trail examples by training the model with data gathered across varying devices, camera resolutions and forest locations. This not only facilitates more general data gathering but also eliminates the need for synthetic data and augmentation. As result, the same model can be used by UAV with differing sensor payload capabilities.

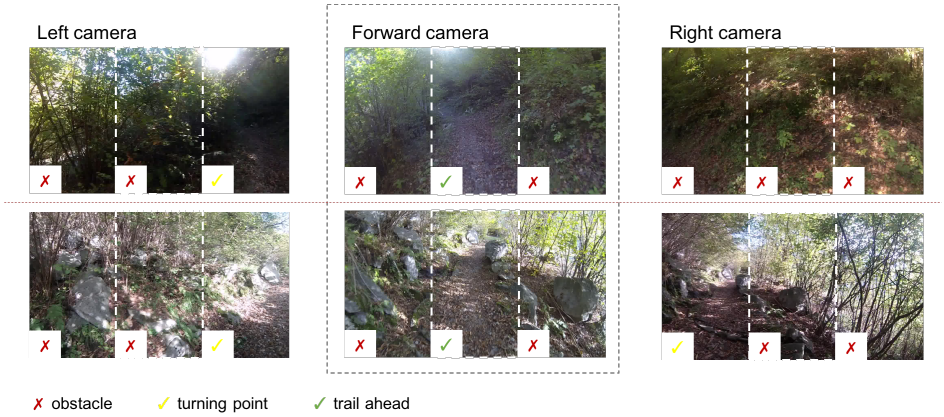


**Fig. 2.** Abstraction of three way image cropping performed on varied camera view (IDSIA dataset [7])

In summary in this work we present a method that both simplifies data gathering and allows real-time labelling of data examples for this trail navigation task, increasing generality in the resulting DNN solution. In addition we present an optmized DNN that learns the position of the trail and a public available dataset (`http://dx.doi.org/10.15128/r1st74cq45z`) gathered localy (Durham/UK) which allows easy reproductibility of this work.

## 2 Related Work

Scene understanding within unstructured environments with varying illumination conditions are critical for autonomous flight within the forest canopy. Significant advancement towards this goal was achieved by [7] which provided a dataset gathered by a human trail walker using a head-mounted rig with three cameras, allowing their proposed DNN architecture to identify the direction of the trail in a given view - $\{left, right, forward\}$. A similar approach is followed by [15] whereby a wide-baseline rig is used, also with three cameras, to gather data which they used to augment the dataset of [7] (denoted: IDSIA dataset).

As a result, the approach presented by [15] is capable of estimating both lateral offset and trail direction. In both cases [7] [15], the authors, follow the common practice of dataset augmentation, via affine image transformations, which adds extra computation without any improved performance guarantees.

Alternatively, synthetic data, from virtual environment models, could potentially replace or at least supplement hard-won real environment data [9,11,18]. However, the significant discrepancy between synthetic data and real-world data often results in models that are trained only on synthetic environment examples not being able to directly transfer this knowledge to real-world operating tasks [5,16,9,3].

Even when training a DNN using only real-world data, it must be noted that the models trained on a limited domain-specific dataset often fail to generalise successfully. In addition, since common DNN architectures require the dataset to be formed from fixed resolution images [8], models commonly fail to generalize across domains. Recent work of [17] investigated the use of reinforcement learning applied in conjunction with Q-learning and adversarial learning frameworks, which could potentially improve the generalisation of the model to different domains. Although encouraging results have been found, which were primarily achieved in a semi-structured environment (roadway), the suitability of this approach in a dynamic and complex environment such as sub forest canopy remains debatable. Instead of only focusing in improving generalisation across domains, our work also investigates the generalisation across varying resolution aerial UAV imagery, by combining a dataset of high-resolution images with a much lower resolution image dataset which better represents UAV with low payload capabilities. Furthermore, we demonstrate a simpler method to data gathering, inspired by the work of [7] that can improve the identification of trails or similar thoroughfares on unstructured environments.

## 3 Approach

Here we are primarily motivated by the three class problem presented by [7] in which an estimation of the trail direction, $\{left, right, forward\}$, is achieved by processing an image triplet of left/right/forward camera views via a DNN. In contrast to [7,15], our approach uses only a single forward facing camera view which is more representative of an operational UAV scenario (i.e a single forward facing camera; minimal size, weight and power). This image view is then itself cropped into $\{left, right, forward\}$ which can be labelled for trail presence/absence (Figure 2).

Using the architecture of [7] (illustrated in Figure 3), we evaluate varying image resolution, the use of additional data augmentation (DA) and activation function (*tanh() / ReLU()*). As illustrated in Figure 3 the network is composed of 10 layers that are subdivided into four convolution and four pooling layers, followed by a fully connected layer and a softmax layer. The result of each convolution layer is fed into a maximum pooling layer. A final *softmax* layer outputs

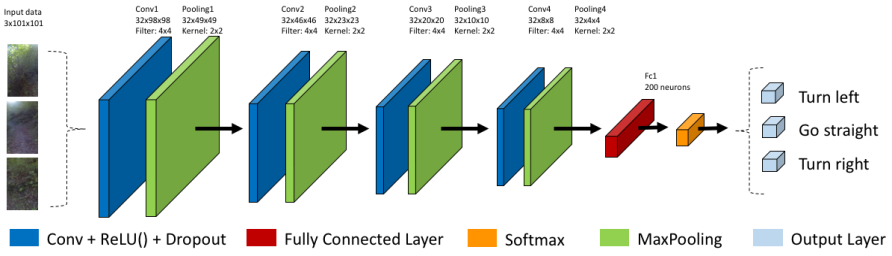


**Fig. 3.** An outline of our DNN architecture - based on [7]

maps to an associated probability for our three class labels $\{left, center, right\}$ from which navigation decisions are then made.

The DNN training uses a gradient descent optimiser, random weight initialisation with zero node biases and is performed over 90 epochs with a 0.05 reduction in learning rate per epoch (decay rate: 0.95). For both training and testing we use the high-resolution ($752 \times 480$) IDSIA dataset (from [7]; denoted H in results of Table 1) and a low-resolution ($320 \times 240$) Urpeth Burn (UB) dataset, gathered locally (County Durham, UK; denoted L in results of Table 1). For training 36,078 high-resolution and 32,017 low-resolution image were used, while for testing 12,252 high-resolution and 5,152 low-resolutions images were used. All DNN training is performed on a Nvidia 1060 GPU.

Further data augmentation (mirror, translation & rotation) was performed on a copy of this original dataset, resulting in a total data set size of 72,135 high-resolution image. For simplicity of reporting, we define $NA$ as non-augmented data obtained results and $DA$ as data augmented obtained results (Table 1).

Our approach allows image labelling to be performed in automatically since the retrieved image, captured facing forward with the trail in the centre, is simply cropped in 3 equal sizes; for each side, a label is associated as follow: $C$ - center (trail), $L$ - left (no trail) and $R$ - right (no trail).

Our DNN model thus learns the signature of a trail associated to class $C$ and non-trail associated to classes $L$ and $R$. These class labels are returned based on the contents of the left, right or center image crop independently of its actual origin from the full sized image. As a result, the presence of class $C$ within the image can be directly correlated to trail presence (Figure 4).

Based on the output certainty from the final *softmax* layer in the DNN for class $C$, the presence of a trail in each of the $\{left, right, center\}$ cropped image regions can be determined facilitating a second labelling for trail presence / absence to be performed. As a result we arrive at a set of six possible classes $\{L-TF, L-NT, R-TF, R-NT, C-TF, C-NT\}$ for our original set of image regions, $\{left, right, centre\}$ with either a Trail Found (TF) or No Trail (NT) label. Within our exploratory formulation, we envisage the navigation decision being taken based on the direction (image region) with the highest level of confidence for trail presence (TF in Figure 4).

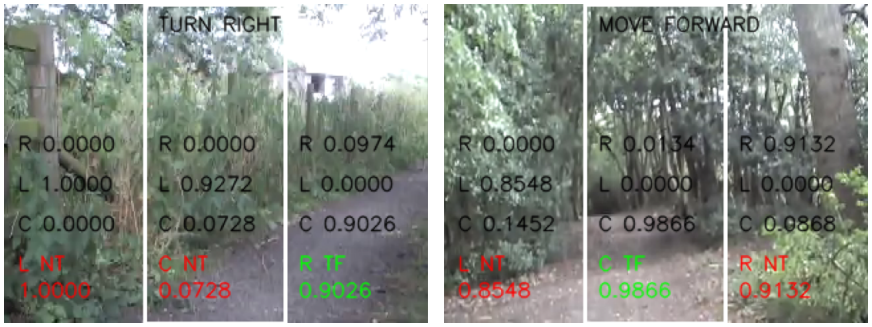


**Fig. 4.** Illustration of level of confidence outputed by the DNN for each image crop.

## 4 Results

Our experimental results are divided into three sets:- (1) image triplet approach of [7] with differing activation functions (*tanh()/ReLU()* - Table 1, upper two divisions), (2) our proposed approach (single forward view image, split into three views - Table 1, middle division, bold) and (3) the impact of high/low/varied image resolutions on performance (Table 1, lower division). Due to the variety of resolution and demographic distribution in the dataset, a 10-fold cross validation was performed across the range of proposed methods (Table 1). The testing dataset (unseen data) was processed by each model, generated during training and its performance can be observed in the blox plot Figure 5, which includes median values and outliers for each fold.

| Approach | DA | Activation Function | Training | Testing | Mean Accuracy($\pm std$) | Precision | | | Recall | | | F1 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | C | R | L | C | R | L | C | R | L |
| Giusti *et al.* [7] | ✓ | *tanh()* | H | H | 0.46 (±0.06) | 0.50 | 0.53 | 0.38 | 0.98 | 0.04 | 0.37 | 0.67 | 0.07 | 0.38 |
| Giusti *et al.* [7] | × | *tanh()* | H | H | 0.73 (±0.08) | 0.97 | 0.72 | 0.62 | 0.62 | 0.77 | 0.81 | 0.76 | 0.74 | 0.70 |
| Giusti *et al.* [7] | ✓ | *ReLU()* | H | H | 0.40 (±0.06) | 0.44 | 0.89 | 0.29 | 0.99 | 0.03 | 0.20 | 0.41 | 0.60 | 0.06 |
| Giusti *et al.* [7] | × | *ReLU()* | H | H | 0.66 (±0.07) | 0.97 | 0.69 | 0.54 | 0.49 | 0.69 | 0.82 | 0.65 | 0.69 | 0.66 |
| **Our Approach** | ✓ | ***ReLU()*** | **H** | **H** | **0.93** (±0.09) | **0.99** | **0.90** | **0.94** | **0.86** | **0.98** | **0.96** | **0.92** | **0.94** | **0.96** |
| **Our Approach** | × | ***ReLU()*** | **H** | **H** | **0.92** (±0.09) | **0.97** | **0.88** | **0.93** | **0.83** | **0.98** | **0.97** | **0.89** | **0.93** | **0.96** |
| High Resolution | × | *ReLU()* | H | L | 0.51 (±0.09) | 0.61 | 0.49 | 0.49 | 0.05 | 0.73 | 0.70 | 0.09 | 0.59 | 0.57 |
| Low Resolution | × | *ReLU()* | L | L | 0.73 (±0.06) | 0.80 | 0.74 | 0.66 | 0.57 | 0.77 | 0.82 | 0.67 | 0.76 | 0.73 |
| Varied Resolutions | × | *ReLU()* | H+L | L | 0.74 (±0.06) | 0.80 | 0.76 | 0.67 | 0.60 | 0.78 | 0.82 | 0.68 | 0.77 | 0.74 |
| **Varied Resolutions** | × | ***ReLU()*** | **H+L** | **H** | **0.93** (±0.02) | **0.97** | **0.91** | **0.91** | **0.84** | **0.97** | **0.98** | **0.90** | **0.94** | **0.94** |

**Table 1.** Results performance for testing in high ($H$) and low ($L$) image dataset combinations, computed for each class ($C$ - Center, $R$ - Right, $L$ - Left)

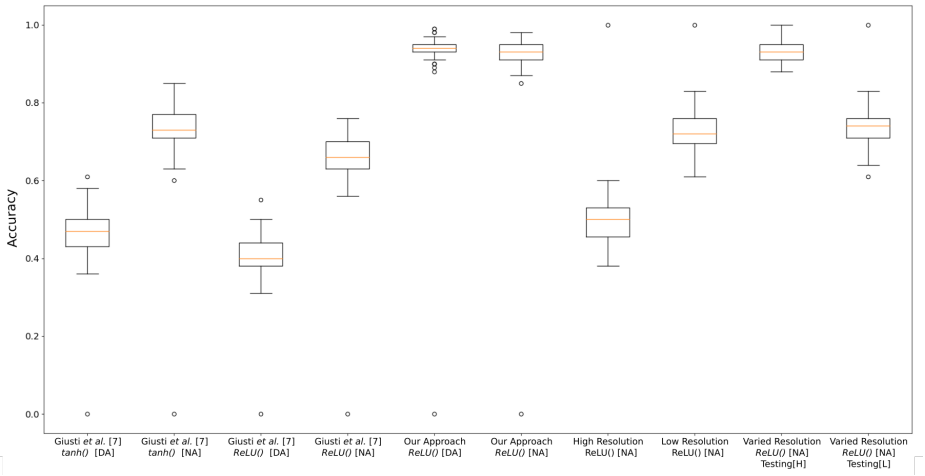


**Fig. 5.** Comparison of performance from the highest accuracy model for each approach, based on 10-fold cross validation.

Overall we see what the use of the *ReLU()* activation outperforms *tanh()* and our approach gives high levels of mean accuracy without the need for data augmentation outperforming the prior reported results in [7] (*in fact no significant improvement was achieved by data augmentation*).

Although our approach fails to generalise when trained with high-resolution images on to low resolution images, it achieves 93% mean accuracy when low-resolution images are added to the training dataset and achieves 73% mean accuracy for training and testing on low-resolution images only (Table 1, lower division).

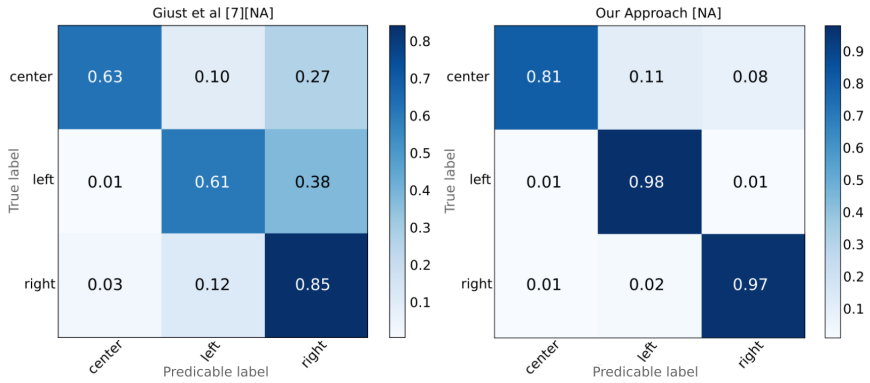


**Fig. 6.** Comparison of performance of Giust et al.[7][NA] (left) versus when Our Approach [NA] (right).

Further analyse of the testing dataset (Table 1) also highlights that augmenting the data, specifically for this scenario, does not improve the classification of the view direction. By looking at the distribution of each approach (Figure 5) we can observe that combining different image resolutions during training is advantageous regardless of the resolution of the testing image.

When analysing the confusion matrix (Figure 6 - left) showing the test results derived from a model trained using Giusti et. al [7] approach, we can observe that since each frame usually contains both trail and vegetation, it is harder for the model to correct distinguish between the classes. In contrast the model trained using our approach (Figure 6 - right) can easily classify each side.

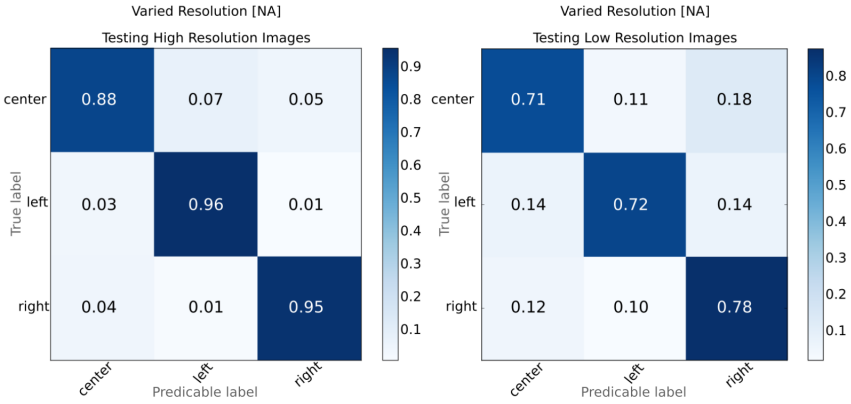


**Fig. 7.** Comparison of how Our Approach [NA] classifies images with low (right) and high (left) resolution.

Regardless of the image resolution in the testing dataset, the same pattern can be observed when comparing the confusion matrix (Figure 7) of a model trained with a dataset containing split images of varied resolution.

These findings can be also observed on the qualitative results presented on Figure 8 and 9, whereby we demonstrate a second labelling for tail presence/absence. Currently we can not quantify the accuracy of this extended labelling without manually checking each one. Due to that the qualitative results shown on Figure 8 and 9 are based on a hand picked selection of the most challenging scenarios from the testing dataset. These scenarios are then evaluated by different models and the level of accuracy is compared accordingly.

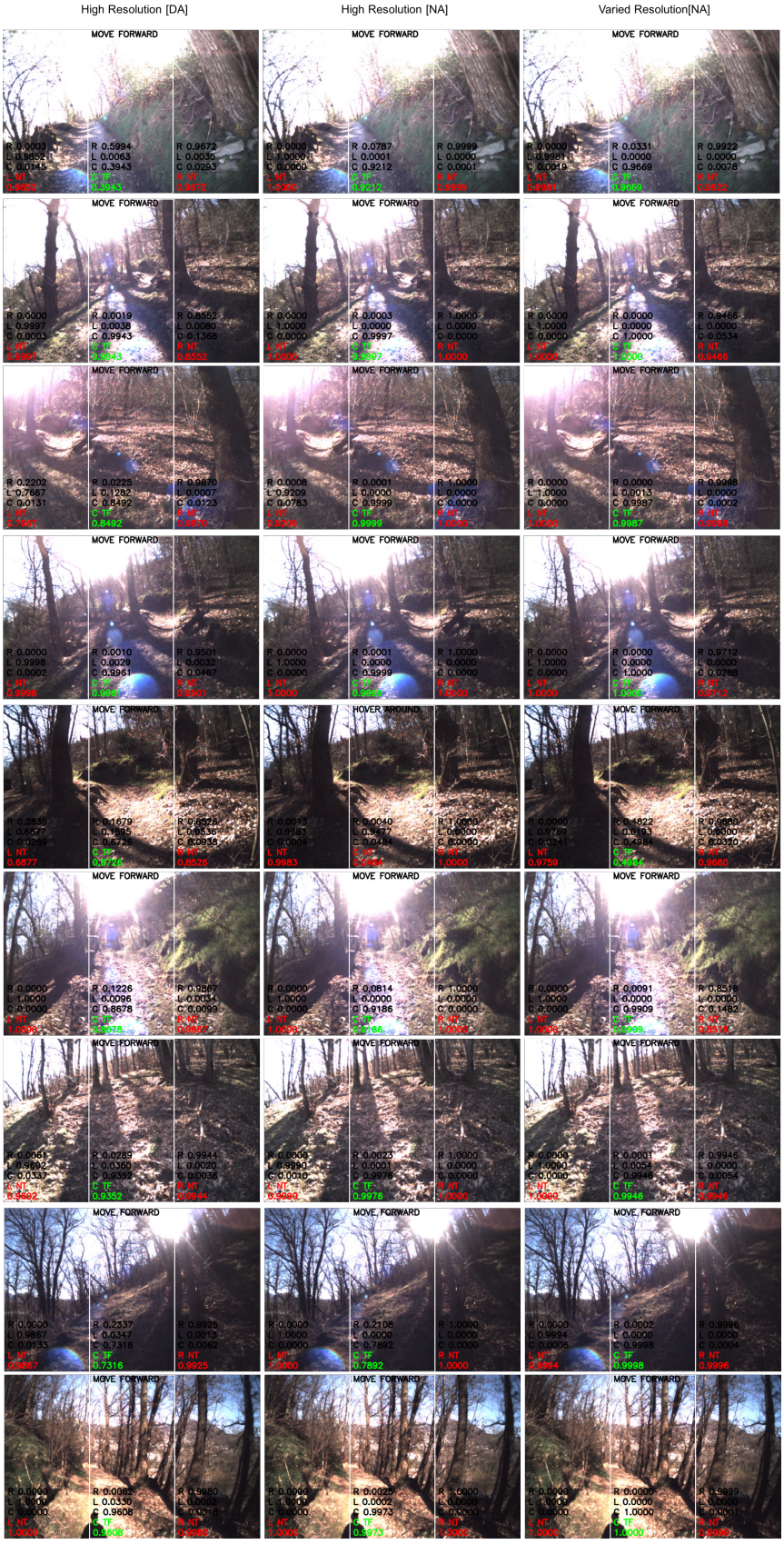


**Fig. 8.** Comparison of performance of different models when tested on high resolution images (IDSIA dataset [7]).

X

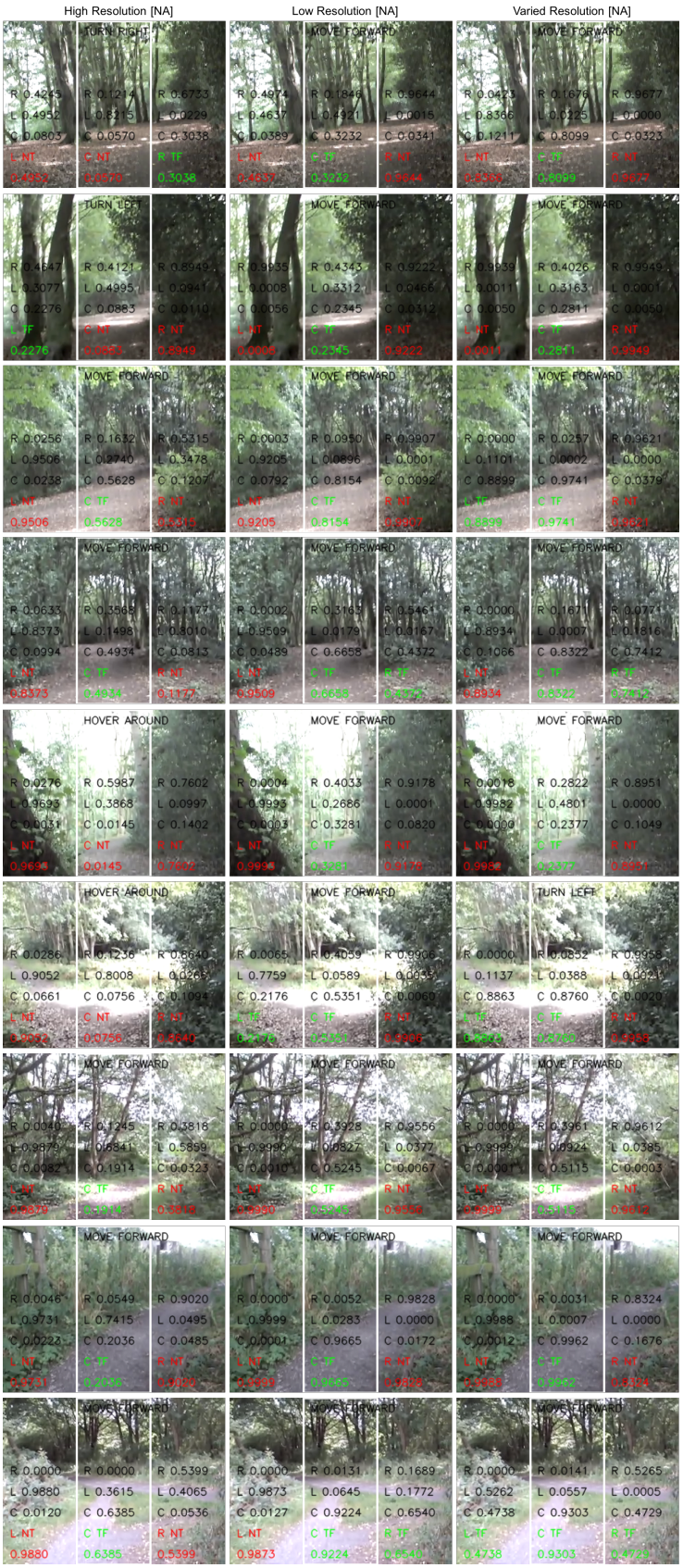


**Fig. 9.** Comparison of the performance of different models when tested on low resolution images.

## 5 Conclusion

In this paper, we present an alternative method to gather and process UAV imagery that improves the level of accuracy for trail navigation under forest canopy based on the use of a single forward facing camera view instead of the multiple camera approach of [7,15]. Our approach also performs well across varying image resolutions and increases the capability of low-cost UAV platforms with limited payload capacity. Future work will include additional aspects of UAV perception and control targeting end-to-end autonomy across this and other challenging operating environments.

## References


1. Ludovic Apvrille, Tullio Tanzi, and Jean-Luc Dugelay. Autonomous drones for assisting rescue services within the context of natural disasters. In *General Assembly and Scientific Symposium*, pages 1–4. IEEE, 2014.
2. Daniel Câmara. Cavalry to the rescue: Drones fleet to help rescuers operations over disasters scenarios. In *Conference on Antenna Measurements & Applications, 2014*, pages 1–4. IEEE, 2014.
3. Paul Christiano, Zain Shah, Igor Mordatch, Jonas Schneider, Trevor Blackwell, Joshua Tobin, Pieter Abbeel, and Wojciech Zaremba. Transfer from simulation to real world through learning deep inverse dynamics model. *arXiv preprint arXiv:1610.03518*, 2016.
4. James T Dietrich. Riverscape mapping with helicopter-based structure-from-motion photogrammetry. *Geomorphology*, 252:144–157, 2016.
5. Dhiraj Gandhi, Lerrel Pinto, and Abhinav Gupta. Learning to fly by crashing. *arXiv preprint arXiv:1704.05588*, 2017.
6. Anna Gaszczak, Toby P Breckon, and Jiwan Han. Real-time people and vehicle detection from UAV imagery. In *Intelligent Robots and Computer Vision XXVIII: Algorithms and Techniques*, volume 7878, page 78780B. International Society for Optics and Photonics, 2011.
7. Alessandro Giusti, Jérôme Guzzi, Dan C Cireşan, Fang-Lin He, Juan P Rodríguez, Flavio Fontana, Matthias Faessler, Christian Forster, Jürgen Schmidhuber, Gianni Di Caro, et al. A machine learning approach to visual perception of forest trails for mobile robots. *IEEE Robotics and Automation Letters*, 1(2):661–667, 2016.
8. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. In *European Conference on Computer Vision*, pages 346–361. Springer, 2014.
9. Klaas Kelchtermans and Tinne Tuytelaars. How hard is it to cross the room?–training (recurrent) neural networks to steer a UAV. *arXiv preprint arXiv:1702.07600*, 2017.
10. Lian Pin Koh and Serge A Wich. Dawn of drone ecology: low-cost autonomous aerial vehicles for conservation. *Tropical Conservation Science*, 5(2):121–132, 2012.
11. Matthias Mueller, Vincent Casser, Neil Smith, and Bernard Ghanem. Teaching UAV to race using UE4Sim. *arXiv preprint arXiv:1708.05884*, 2017.
12. Jaime Paneque-Gálvez, Michael K McCall, Brian M Napoletano, Serge A Wich, and Lian Pin Koh. Small drones for community-based forest monitoring: An assessment of their feasibility and potential in tropical areas. *Forests*, 5(6):1481–1507, 2014.

13. Anuj Puri. A survey of unmanned aerial vehicles (UAV) for traffic surveillance. *Department of computer science and engineering, University of South Florida*, pages 1–29, 2005.
14. Guillaume Rémy, Sidi-Mohammed Senouci, François Jan, and Yvon Gourhant. Sar drones: drones for advanced search and rescue missions. *Journées Nationales des Communications dans les Transports*, 1:1–3, 2013.
15. Nikolai Smolyanskiy, Alexey Kamenev, Jeffrey Smith, and Stan Birchfield. Toward low-flying autonomous MAV trail navigation using deep neural networks for environmental awareness. *IEEE/RSJ Int. Conf. Intell. Robot. Syst. (IROS)*, 2017.
16. Lei Tai and Ming Liu. Deep-learning in mobile robotics-from perception to control systems: A survey on why and why not. *arXiv preprint arXiv:1612.07139*, 2016.
17. Jaeyoon Yoo, Yongjun Hong, and Sungrho Yoon. Autonomous UAV navigation with domain adaptation. *arXiv preprint arXiv:1712.03742*, 2017.
18. Tianhao Zhang, Gregory Kahn, Sergey Levine, and Pieter Abbeel. Learning deep control policies for autonomous aerial vehicles with mpc-guided policy search. In *International Conference on Robotics and Automation, 2016*, pages 528–535. IEEE, 2016.