EPiC Series in Computing

Volume 81, 2022, Pages 101–110

Proceedings of 11th International Congress on Advanced Applied Informatics

Reliable Night Bear and Boar Detection Based on Training with Pseudo Infrared Images

Keigo Fusaka[1], Yoichi Tomioka[2], Hiroshi Saito[3], and Yukihide Kohira[4]

The University of Aizu, Aizu-Wakamatu City, Fukushima, Japan
{s1270143[1], ytomioka[2], hiroshis[3], kohira[4]}@u-aizu.ac.jp

Abstract

In recent years, accidents and damages caused by wild animals have been serious problems. It has become important to detect wild animals accurately at an early stage. A sufficient number of training infrared images is required to detect wild animals taking various postures at night time using deep learning techniques. In this study, we propose a method to increase appropriate training samples for night wild animal detection using annotated daytime images. We employ a model based on Cycle Generative Adversarial Network (CycleGAN) to be able to generate pseudo infrared images from daytime images. In our experiments, we apply the proposed method to bear and boar detection. The experimental results show that the proposed method achieves significant improvements in bear detection accuracy taking various postures.

1 Introduction

In areas close to mountain forests, damage to crops and accidents involving attacks on people by wild animals such as bears and boars have occurred. Traditional countermeasures are patrol, electric fences, and traps. However, it is difficult for people to monitor large areas frequently. Electric fences can be dangerous for people with pacemakers or defibrillators. It is necessary to understand the tendency of the appearance of wild animals to set traps effectively. In order to investigate the tendency, it is important to recognize wild animals with high accuracy. Because wild animals, such as bears and boars, are often active at night, detecting them with high accuracy at night is necessary.

Camera traps using motion sensors and cameras are widely used to capture wildlife safely and automatically. However, manually analyzing the captured images requires a lot of time and effort. In [1], a method for classifying wildlife in images captured by camera traps using ResNet50 [7] has been proposed. However, this method cannot correctly identify small wild animals in the images. In [10], a wildlife detection method based on R-CNN [5] has been proposed. In [13], a method for efficient detection of bears has been proposed by extracting candidate regions where bears are likely to exist using low-resolution images and classifying each candidate region. However, this method focuses on images taken during the daytime and cannot detect wild animals in images taken at night in infrared light. In order to realize an object detection model that can accurately detect wildlife in infrared images taken at night,

a sufficient number of infrared images of the target animal is required. Especially, bears can stand up and pick up fruits on a branch of trees. Therefore, for accurate bear detection, we need to collect infrared images of bears of various postures. Collecting thermal images of wild animals of various postures is labor-intensive work and time-consuming.

On the other hand, in [2], CycleGAN [15] is applied to artificially generate nighttime vehicle images from daytime vehicle images in order to create a nighttime vehicle detection model. They also report that the generated pseudo-nighttime images can be used in addition to the daytime images to generate a more accurate vehicle detection model.

In this paper, we propose a method to generate pseudo-infrared images from daytime images in order to create an object detection model that can detect bears and boars from nighttime infrared images with high accuracy. This method is based on the method described in [2]. The contributions of this paper are as follows:

1. Reference [2] has proposed a CycleGAN-based method that converts daytime images into nighttime images captured by a normal camera. On the other hand, in this paper, we propose a converter that generates pseudo-infrared images by adding the features of infrared images taken at night to images taken with a normal camera. We demonstrate that the accuracy of detecting bears and boars at night can be greatly improved by training an object detection model using the generated pseudo-infrared images.

2. Unlike the vehicles, which are targets of [2], the bears and boars take various postures. For this reason, generating pseudo-infrared images of bears and boars is a more challenging task, and the quality of the generated pseudo-infrared images varies greatly depending on the input images. Low-quality pseudo images are not suitable for training an object detection model. We demonstrate that object detection accuracy can be improved by using image selection to extract only those suitable for training an object detection model. We also analyze the improvement of detection accuracy posture by posture.

The rest of this paper is organized as follows. Section 2 describes You Only Look Once (YOLO) and Generative Adversarial Network (GAN). In Section 3, we explain the proposed method. In Section 4, we show experimental results. Finally, conclusions and future work are presented.

2 Related Reserch

2.1 Object Detection Methods

You Look Only Once (YOLO) [11] is a kind of object detection method based on convolutional neural networks. It is extended in YOLOv3 [12], YOLOv4 [3], and YOLOv5 [8]. YOLOv5 models employ Path Aggregation Networks (PAN) [9] and Cross Stage Partial Network (CSP-Net) [14] to efficiently extract rich features from input images. Official YOLOv5 models are implemented in the PyTorch framework, while YOLOv1 to YOLOv4 are implemented in the darknet framework. YOLOv5 has four types of models: YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x of different sizes. In this paper, we use the YOLOv5x model for reliable bear and boar detection because YOLOv5x is the largest of the four models, which is designed for more accurate object detection.

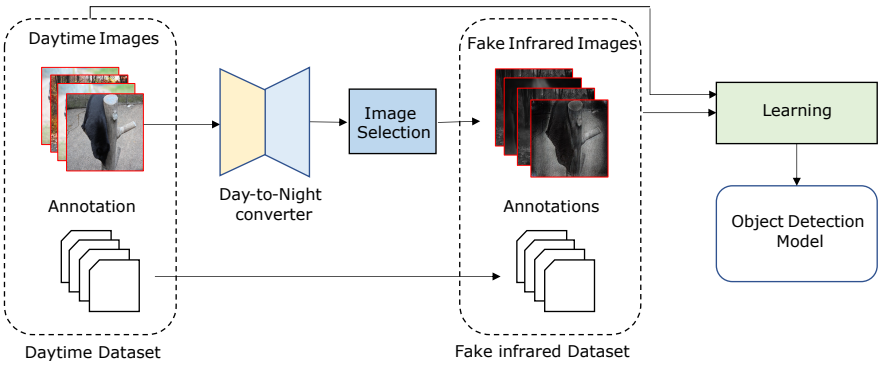

Figure 1: Overview of wildlife detection model generation for night infrared images.

2.2 Generative Adversarial Network

Generative Adversarial Network (GAN) [6] consists of two networks: generator and discriminator. The generator outputs a pseudo image, and the discriminator judges whether the image is real or not. The generator learns to deceive the discriminator while the discriminator learns to identify fake images more accurately. These two networks learn with conflicting goals simultaneously. This is why this network is called a generative adversarial network.

CycleGAN [15] is a method that achieves image transformation by learning the relationship between the domains of two image datasets, instead of learning the relationship between the corresponding pixels of the two images. This allows CycleGAN to acquire an image transformation without a large number of paired training images.

3 Proposed Method

3.1 Overview of the proposed method

Figure 1 shows an overview of the proposed method. The proposed method consists of (1) pseudo-infrared dataset generation, (2) image selection, and (3) wildlife detector training. The input is a set of images taken by normal cameras in the daytime and the corresponding annotation data. In pseudo-infrared dataset generation, we generate nighttime pseudo-infrared images from wildlife images taken with normal cameras during the day. Next, in image selection, we remove images from the generated pseudo-infrared images that are not suitable for training the object detection model. In the generated pseudo-infrared images, the position and orientation of the target object in the image are not transformed. Therefore, the annotation data of the daytime images can be directly used as the annotation data of the generated pseudo-infrared images. In wildlife detector training, the accuracy of nighttime wildlife detection is improved by training the object detection model using pseudo-infrared images in addition to regular daytime camera images. In this paper, we aim to detect bears and boars. Considering the object detection accuracy and training time, we adopt YOLOv5x as the object detection model. Other object detection models can be used as well.

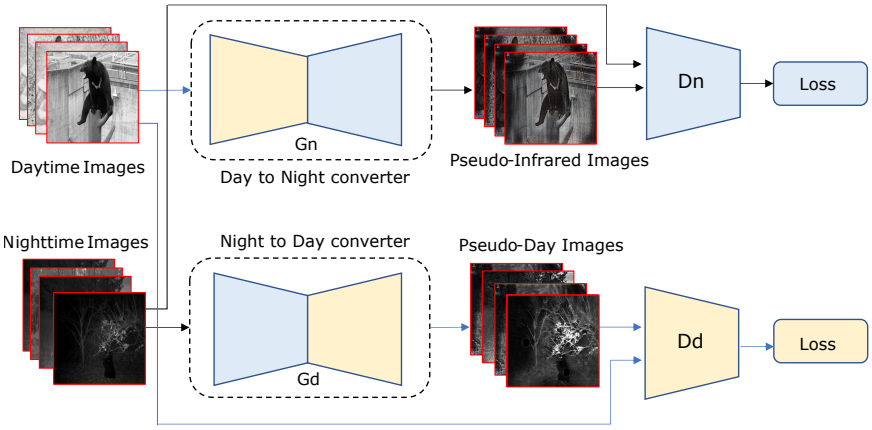

Figure 2: Overview of the CycleGAN Framework.

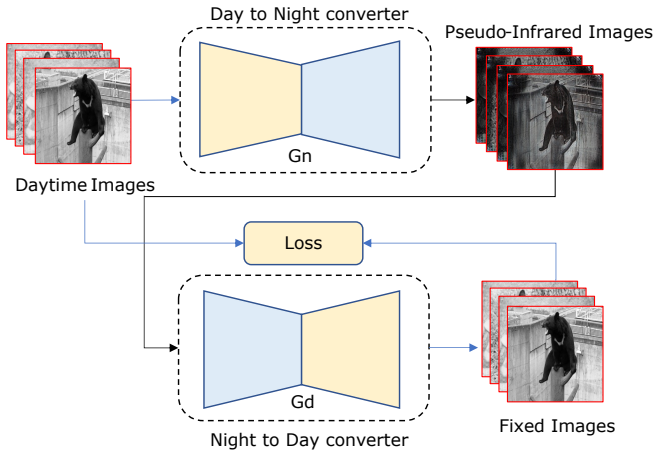

Figure 3: Cycle-consistency constraint.

3.2 Pseudo-infrared dataset generation

In this section, we describe the generation of the pseudo-infrared dataset shown in Figure 1. In this process, pseudo-infrared nighttime images are generated from bear and boar images taken in the daytime. This paper applies two methods to generate the pseudo-infrared images: grayscaling and day-to-night transformation using CycleGAN. In grayscaling, we use the grayscaled image as the input image instead of the pseudo-infrared image. Next, we describe the generation of pseudo-infrared images using CycleGAN.

As shown in Figure 2, CycleGAN can transform an image of one domain into an image of another domain by a converter.

The framework is divided into two parts. One corresponds to the transformation from domains A to B, and the other corresponds to the transformation from domains B to A. The images of domains A and B do not have to be paired images that are the same in time and

space. By using unpaired images, a night-to-day converter is trained based on the unsupervised learning procedure for adversarial generative networks.

In this paper, domain A is the grayscale images of bears and boars taken in the daytime (daytime images). Domain B is the infrared images of the bear and boar taken at night (night images). The day-to-night converter and the night-to-day converter take the independent day and night images, respectively. Each converter transforms the input images into those of the other domain. In the training of the day-to-night converter, the day image is converted into a pseudo-infrared image (false night image) by the current day-to-night converter and then input into the night discriminator. The night discriminator learns to correctly classify the real and fake images, while the day-to-night converter learns to fool the night discriminator. We can achieve a more accurate transformation from the day domain to the night domain by training the day-to-night converter and the night discriminator simultaneously. Similarly, we train the night-to-day converter. To realize more accurate transformations from unpaired images, CycleGAN also imposes a cycle-consistency constraint so that a pseudo night (day) image generated by the day-to-night (night-to-day) converter can be restored to the original image by the night-to-day (day-to-night) converter as shown in Figure 3. The corresponding loss for this constraint can be expressed as $|GD(GN(I_d)) - I_d|$ and $|GN(GD(I_n)) - I_n|$ where GD and GN are day-to-night and night-to-day converters, respectively, I_d and I_n are the real daytime and nighttime images used in training. By using day-to-night converter, we generate pseudo-infrared images from daytime images.

3.3 Image selection

When CycleGAN is used to generate a pseudo-infrared dataset, the quality of a generated pseudo-infrared image depends on the original image. There are pseudo-infrared images with blurred or unclear outlines of bears and boars. Such low-quality pseudo images cause the accuracy degeneration of the object detection model. Therefore, we manually remove the low-quality pseudo-infrared images from the dataset and use the remaining images to train the object detection model.

4 Experimental Results

First, we describe the datasets used to train and evaluate the proposed bear and boar detection models and explain the experimental conditions. Then, the experimental results and discussion are presented.

4.1 Dataset

First, we describe our training dataset. The daytime dataset consists of images taken by ourselves and those from IMAGENET[4]. A daytime dataset, denoted by **D**, consists of 1606 bear images and 496 boar images. A night dataset, denoted by **N**, consists of 542 infrared bear images and 142 infrared boar images taken at night.

Datasets **D** and **N** are manually annotated with two classes, boar and bear. Note that the number of images in dataset **N** is much smaller than that in dataset **D** and that the bear/boar postures and scenes are limited. We generated pseudo-infrared image datasets **GR**, **GA**, and **GA-S** using dataset **D** according to the proposed method. Datasets **D** and **N** were used for training CycleGAN. **GR** is a pseudo-infrared dataset generated by grayscaling. **GA** is a pseudo-infrared dataset generated by CycleGAN from dataset **D**. **GA-S** is a pseudo-infrared

Table 1: The number of images used in each dataset.

	D	**N**	**GR**	**GA**	**GA-S**
Total	2102	684	2102	2102	1964
bear	1606	542	1606	1606	1468
boar	496	142	496	496	496

Table 2: Number of bear images of each posture in the evaluation dataset.

	Front	Back	Side	Sitting	Standing	Lying
Total	30	19	83	18	8	12

dataset generated by image selection from **GA**. The number of images for each training dataset is shown in Table 1.

Next, we describe the images for evaluation. The evaluation dataset consists of 170 infrared bear images and 101 infrared boar images. To evaluate the detection rate of bears taking various postures, we classified the bear images into six categories: Front, Back, Side, Sitting, Standing, and Lying. We show the number of images in each category in Table 2.

4.2 Experimental enrironment and evaluation index

We used PyTorch as our deep learning framework. CycleGAN uses the same architecture as in the literature [15]. The batch size was 16, and the epoch number was 300. Other hyperparameters were used for default values. We used YOLOv5 [8] as the object detection method and YOLOv5x as the model.

The objective of the proposed system is to accurately detect bears and boars. The mean Average Precision (mAP) was adopted to evaluate the quality of the detector. The Average Precision (AP) is defined as the area under the precision-recall curve of a certain object class. The mAP is the average of APs for all classes.

4.3 Pseudo infrared image generation

Figure 4 shows examples of the original images, pseudo-infrared images generated by grayscaling and CycleGAN. Figure 5 shows examples of real infrared images. Compared with the pseudo-infrared image generated by grayscaling, the pseudo-infrared image generated by CycleGAN has two features that are closer to real infrared images. First, infrared images taken at night tend to have shadows in the background where infrared light does not reach. Second, there is a relatively large noise in the images. Although the pseudo-infrared image generated by CycleGAN is unnatural compared with real infrared images shown in Figure 5, they are useful for training an object detection model, as shown in the next subsection. The generation of higher quality infrared images is in our future task.

4.4 Bear and boar detection

To verify the effectiveness of the proposed method, we evaluated the object detection accuracy of YOLOv5x trained on different combinations of datasets **D**, **N**, **GR**, **GA**, and **GA-S**. Table 3 shows the evaluation results.

Figure 4: Examples of pesudo infrared images. The images of the first row are real daytime bear images. The images of the second and third rows are the corresponding gray-scale images and images generate by CycleGAN, respectively.

Figure 5: Real infrared images.

The mAP of the model trained only on daytime image **D** taken by a normal camera was 28.0%. The mAP of the model trained on only dataset **N** was 48.0%. Because the infrared images of bears used for training are not sufficient, the AP of bears is particularly low. On the other hand, training with the combination of datasets **D** and **N** improved the mAP to 72.6%. Next, we evaluated a model trained on the combination of datasets **D** and **GA-S**, assuming that real nighttime images were not available. This model achieved comparable AP with that of **D**+**N** for bears by using pseudo-infrared images. On the other hand, the AP for boar was improved little. In particular, when the skin of a boar in an infrared image is bright due to infrared light, the model of **D** and **GA-S** cannot detect the boar. To further improve the AP

Table 3: mAP of object detection models trained with different datasets. [%]

	D	**N**	**D+N**	**D+GA-S**	**D+N+GR**	**D+N+GA**	**D+N+GA-S**
Total	28.0	48.0	72.6	37.6	71.7	74.3	75.7
bear	55.7	28.9	77.5	73.9	78.2	80.4	83.2
boar	0.2	67.1	67.7	1.3	65.2	63.7	68.1

Table 4: Detection rate of bears taking various postures. [%]

	D	**N**	**D+N**	**D+GA-S**	**D+N+GR**	**D+N+GA**	**D+N+GA-S**
Front	6.6	46.7	50.0	50.0	50.0	66.7	73.3
Back	5.3	10.5	31.6	52.6	31.6	63.2	47.4
Side	65.5	11.5	73.6	78.2	73.6	78.2	80.5
Sitting	72.2	33.3	94.4	77.8	94.4	94.4	100.0
Standing	12.5	0.0	50.0	37.5	50.0	50.0	50.0
Lying	8.3	0.0	8.3	33.3	8.3	16.6	41.7

for boar, we need to improve the quality of pseudo infrared images of boars. Moreover, the model trained with **D+N+GA-S** improved both bear and boar detection accuracy. Using a small number of real night images with pseudo images is also helpful.

To clarify which of the two types of pseudo images is more suitable, **GR** or **GA**, we compared **D+N+GR** and **D+N+GA**. As a result, **D+N+GA** showed 2.6% increase in mAP over **D+N+GR**. From this result, we can see that **GA** is more suitable for bear and boar detection training than **GR**.

We compared the model trained with **D+N+GA-S**, in which the images generated by CycleGAN with poor quality were removed, with the model trained with **D+N+GA**. The mAP of **D+N+GA-S** was 1.4% higher than that of **D+N+GA**. This result indicates that image selection is effective for training bear and boar detectors.

Moreover, we evaluated the detection rate of the bears taking different postures. The results are summarized in Table 4. The results show that the model trained with **D+N+GA-S** is more robust to the difference of posture compared with the other models. The accuracy was improved by adding pseudo-infrared images of bears taking various postures for training. In Figure 6. we show examples of bear detection by models trained with **D**, **D+N**, and **D+N+GA-S**, respectively. For posture categories of Front, Side, and Sitting, we achieve higher accuracy than Back, Standing, and Lying. In the case of Side, we can clearly see the legs of bears, and the shape of bears is characteristic compared with other postures. Front and Sitting postures make it easy to see the bear's face, which is a characteristic of bears, while the back posture makes it difficult to see the bear's face. Furthermore, the bear's body is assimilated into the black background. Therefore, Front and Sitting can be considered to be significantly more accurate than Back. The number of images of Standing and Lying is relatively small in our training dataset. One of the future tasks is to increase the number of test data for each posture and to create reliable dataset.

On the other hand, there were several cases in which the model trained with **D+N+GA-S** misidentified large black objects as bears such as the dark trunk of tree, shadow between trees, shadow of entrance of large metal tube used as trap. Realization of bear and boar detection model which is robust to such large black objects and background is in our future work.

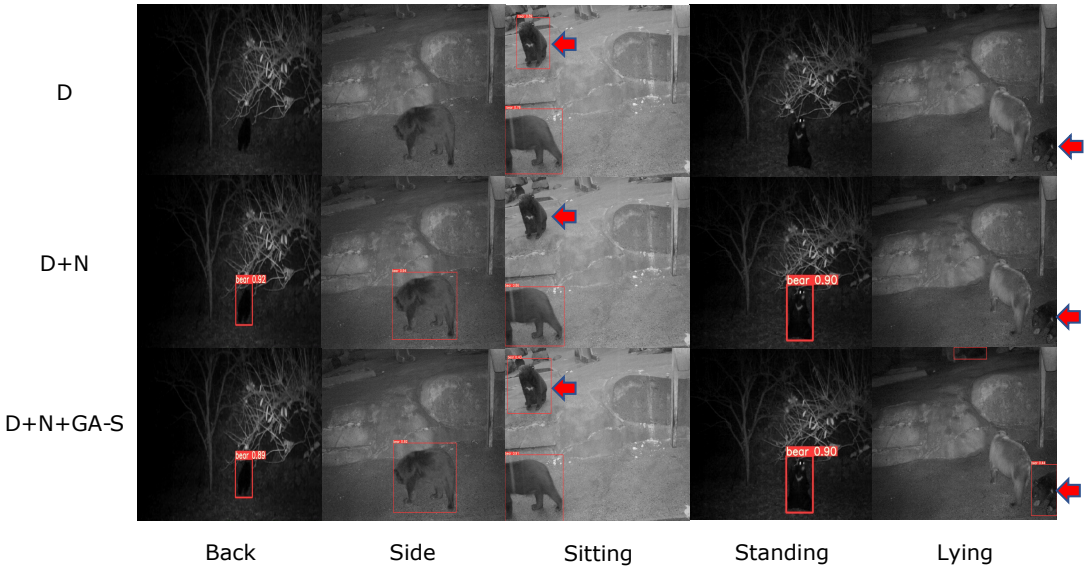

Figure 6: Example of detection results of **D**, **D**+**N**, and **D**+**N**+**GA-S**.

5 Conclusion

In this paper, we have discussed cross-domain (daytime to nighttime) wildlife detection increasing a training dataset without manual annotation process in the target domain (nighttime). We have proposed a day-to-night converter that generates pseudo infrared images from daytime images. We have examined the performance of the proposed method when considering detectors operating in the nighttime domain of infrared images of bears and boars taken in the real world. As a result, the performance has been improved by adding pseudo-infrared images and selecting and excluding images that are not suitable for training. We have demonstrated that the mean Average Precision (mAP) of bear and boar detection has been improved from 18.4% to 75.7% using pseudo images generated by CycleGAN. Moreover, we have demonstrated that our method can be profitable even when the pseudo infrared images are not perfect.

In our future work, we will further improve the accuracy of bear and boar detection. It is important to improve the quality of pseudo infrared images. In this paper, the accuracy of boar detection was not improved significantly. If we can reproduce the features of glowing eyes and bright skins of boars in infrared images, that is helpful to improve the quality of infrared images. Also, we can use multiple GAN-based converters to make pseudo infrared images with various features. Though we limited our target animals to bears and boars in this study, we will make our method applicable to the detection of more wild animals.

References

[1] M. Ando, S. Nakatsuka, H. Aizawa, S. Nakamori, T. Ikeda, J. Moribe, K. Terada, and K. Kato. Recognition of wildlife using deep learning in images taken by camera traps. In *Honyurui Kagaku (Mammalian Science)*. Honyurui Kagaku (Mammalian Science), 2019.

[2] Vinicius F Arruda, Thiago M Paixão, Rodrigo F Berriel, Alberto F De Souza, Claudine Badue, Nicu Sebe, and Thiago Oliveira-Santos. Cross-domain car detection using unsupervised image-to-image translation: From day to night. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2019.

[3] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020.

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[5] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[6] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[8] Glenn Jocher, Alex Stoken, Jirka Borovec, NanoCode012, ChristopherSTAN, Liu Changyu, Laughing, tkianai, Adam Hogan, lorenzomammana, yxNONG, AlexWang1900, Laurentiu Diaconu, Marc, wanghaoyang0106, ml5ah, Doug, Francisco Ingham, Frederik, Guilhen, Hatovix, Jake Poznanski, Jiacong Fang, Lijun Yu, changyu98, Mingyu Wang, Naman Gupta, Osama Akhtar, PetrDvoracek, and Prashant Rai. ultralytics/yolov5: v3.1 - Bug Fixes and Performance Improvements, October 2020.

[9] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8759–8768, 2018.

[10] Mohammed Sadegh Norouzzadeh, Anh Nguyen, Margaret Kosmala, Ali Swanson, Craig Packer, and Jeff Clune. Automatically identifying wild animals in camera trap images with deep learning. *arXiv preprint arXiv:1703.05830*, 1(5), 2017.

[11] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[12] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[13] Masayuki Tokutake, Kaisei Shimura, Yoichi Tomioka, Hiroshi Saito, and Yukihide Kohira. Reliable and efficient bear-presence detection based on region proposal of low-resolution. In *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 2547–2552. IEEE, 2020.

[14] Chien-Yao Wang, Hong-Yuan Mark Liao, Yueh-Hua Wu, Ping-Yang Chen, Jun-Wei Hsieh, and I-Hau Yeh. Cspnet: A new backbone that can enhance learning capability of cnn. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 390–391, 2020.

[15] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.