Multi-modal object Detection in Thermal Imagery using Saliency Map

Jaeyoon Sim Sangjin Lee
Department of Creative IT Engineering, POSTECH
{simjy98, oop523}@postech.ac.kr

Abstract

Thermal images are used to detect the presence of people. Sometimes it doesn't work well because of various conditions when looking for a person in a thermal image. In this case, we can achieve higher accuracy through RGB images together with the thermal image. And we would like to further utilize the characteristics of the thermal image. In this paper, we propose multi-modal object detector trained on natural image domain data to help improve the performance of object in thermal images. And we propose to augment thermal images with their saliency maps to make the most of the features. We propose the use of well-known image-to-image translation frameworks to generate RGB equivalents of a given thermal image and use a multi-modal architecture for object detection in the augment thermal image. We show that our framework has the ability to learn with less data from thermal domain when using our approach and produce some performance.

1. Introduction

The current sensors in self-driving vehicles with level 2 and 3 autonomy lacking thermal imaging do not adequately detect vehicles and pedestrians. Pedestrians are especially at risk after dark and it is really dangerous. Pedestrian detection has already been successfully studied widely in color images and videos using state-of-the-art deep learning techniques for object detection and semantic segmentation. However, thermal images are not successful. But they have the advantages which are very useful in detecting pedestrians in conditions where color images fail, such as night time or under bad lighting conditions. Thermal sensors perform well in such conditions where autonomy level 2 and 3 sensor suite technologies are challenged. Thermal IR cameras are relatively more robust to illumination changes, and can thus be useful for deployment both during the day and night. In addition, they are low cost, nonintrusive and small in size. Consequently, thermal IR cameras have become increasingly popular in applications such as autonomous driving recently, as well as in other many applications. So, detection and classification of objects and pedestrians in thermal imagery are important.

Although object detection has always remained an important problem in computer vision, most of the efforts have focused on detecting humans and objects in RGB imagery. With the development of Deep Convolutional Neural Networks, object detection performance in the RGB domain has been significantly improved using region-based methods, such as R-CNN [1], Fast R-CNN [2] that use selective search, and Faster R-CNN [3] that uses region proposal networks(RPN) to identify regions of interest(ROI). And the aforementioned object detection methods rely on architectures and models that have been trained on large-scale RGB datasets like PASCAL-VOC, and MS-COCO.

Koch and Ulman *et al.* [4] define saliency at a given location by how different this location is from its surroundings in color, orientation, motion, and depth. Salient objects in a scene can be interpreted as being a visual attention mechanism which illuminates pixels belonging to salient objects in a given scene. So, we think that using saliency maps along with thermal images would help to improve the performance of object detectors in thermal imagery. We train detectors on thermal images augmented with their saliency maps generated using deep learning techniques.

In this work, we propose multi-modal framework for object detection in thermal domain, consisting of three branches. The first branch is pre-trained on large-scale RGB datasets such as PASCAL-VOC or MS-COCO and finetuned using a visual RGB input that is obtained using an image-to-image(I2I) translation framework from a given thermal image. The second branch is the standard training process on a relatively smaller thermal dataset which is KAIST Multispectral Pedestrian dataset [5]. And the third branch is the saliency maps from original thermal image by using saliency detectors. This architecture uses high-level features from the RGB domain to improve object detection in the thermal domain. And it also uses pixel-level features from the saliency map which obtained from the original thermal image. So, our framework can borrow from any large-scale RGB dataset available for object detection and additional information from the saliency maps. We can overcome the low performance and problems by using our framework and our experimental results on the recently released KAIST multispectral pedestrian thermal imagery dataset show that the proposed multi-modal framework achieves a high mAP than single-mode fully-supervised detector.

The rest of the paper is organized as follows. Section 2 reviews existing work on object detection methods for saliency detection in images and provides the context for study including a brief overview of early and recent work on applying deep learning for thermal imagery. Section 3 describes our approach and methodology. Section 4 describes the experiments carried out and the results. Finally, Section 5 and Section 6 describes our conclusion and future work.

2. Related Work

Detection and classification of objects in the thermal imagery has been an active area of research in computer vision. There has been amount of work on classifying and detecting people and objects in thermal imagery using standard computer vision and deep learning.

2.1. Thermal detection

Davis *et al*. [6] proposed a two-stage template-based method to detect people in widely varying thermal imagery. To locate the potential person locations, a fast screening procedure is used with a generalized template and AdaBoost ensemble classifier is used to test the hypothesized person locations. Kai *et al*. [7] proposed a local feature-based pedestrian detector on thermal data. They used a combination of multiple cues to find interest points in the images and used SURF [8] as features to describe these points.

Recently, deep learning models enable more descriptive features to be learned. Several methods have been proposed for applying deep learning methods to thermal images. Peng *et al*. [9] proposed a CNN for face identification in near IR images. It is a modification of GoogLeNet and has a more compact structure. Rodger *et al*. [10] developed a CNN trained on short-to-mid range high resolution IR images containing six object classes using and LWIR sensor. This network was successful at classifying other short to mid-range objects in unseen images, although it struggled to generalize to long range targets.

2.2. Saliency detection

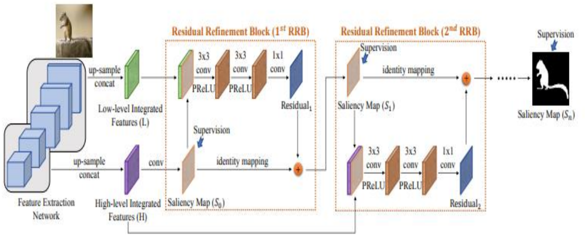

Figure 1. The schematic illustration of R^3-Net.

Salient object detection aims to highlight the most conspicuous object in an image and a substantial number of methods have been developed for it over the past few decades. One of the earliest works on saliency detection inspired by the visual system of primates which shift focus to most conspicuous objects across the visual scene. Traditional saliency detection methods involved using methods like global contrast, local contrast, and other hand crafted features like colour and texture.

Recent methods use CNNs for salient object detection. DHSNet first learns global saliency cues such as global contrast and uses a novel hierarchical convolutional neural network to refine the details of the saliency maps using local context information. Amulet integrates multi-level features at multiple resolutions and learns to predict saliency maps by combining the features at each resolution in a recursive manner. The state-of-the-art networks are PiCA-Net and R^3-Net, which are deep saliency techniques then R^3-net is used in this paper. Figure 1 show the schematic illustration of R^3-Net [11].

2.3. Multi-modal detection

Multi-modal learning refers to a methodology that simultaneously learns multiple datasets with different variable dimensions. Wagner *et al*. [12] applied Aggregated Channel Features and Boosted Decision trees for proposal generation and classified these proposals with a CNN, which fuses Visual and IR information. Choi et al. uses two separate region proposal networks(RPN) for both Visual and IR images and evaluates the proposals generated by both the networks with Support Vector Regression on fused deep features. Konig *et al*. [13] and Liu *et al*. [14] propose a multi-modal framework that combines RGB and thermal information in a Faster-RCNN architecture by posing it as a convolutional network fusion problem.

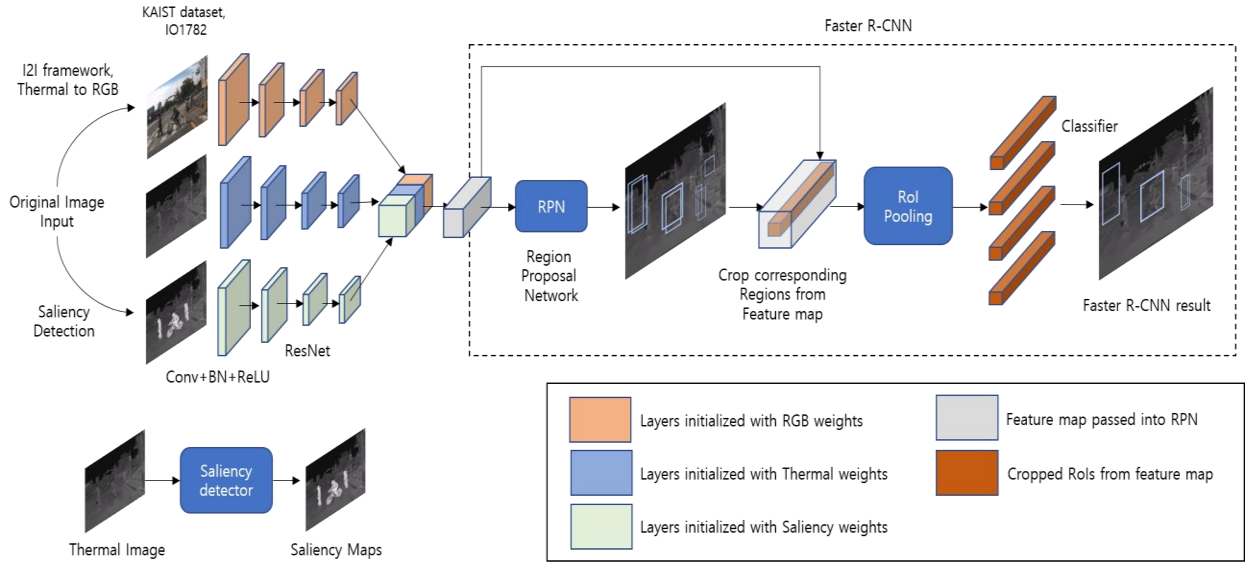

Figure 2. Adaptation of proposed Multi-modal framework for Faster-RCNN

2.4. Image-to-image(I2I) translation

The I2I translation task aims to learn a function to transfer the domain-specific part of a given image to the target domain. In the Pix2Pix [15] method, a paired training dataset is used to train a cGAN in a supervised manner. In the CycleGAN [16], it uses a cycle consistency loss to preserve the key attributes between the input and the translated image. In the ComboGAN [17], a number of methods explore multi domain translation using a single GAN such as StarGAN and GANimation.

Numerous I2I translation problems are inherently multi-modal. The MUNIT and DRIT methods are developed based on partially shared latent space, and use a content encoder as well as a style encoder to decompose the latent space of images into a domain invariant part and a domain specific part, respectively. These methods are able to translate images while preserving the domain invariant properties without supervision.

3. Approach

Our overall proposed methodology multi modal object detection for thermal images by using saliency map is summarized in Figure 2. The key idea of our methodology is to borrow knowledge from data-rich domains such as RGB without the explicit need for a paired multimodal dataset. Additionally, we can use saliency detector to extract special highlighted features from thermal image. We want to achieve this objective by leveraging the success of recent image-to-image(I2I) translation methods such as CycleGAN, UNIT, DiscoGAN to automatically generate a RGB image from a given thermal image. We just use CycleGAN and UNIT because the procedure of DiscoGAN is similar with CycleGAN. And we use R^3-Net to get saliency maps from given thermal image. It is called saliency detector and it can help to improve our performance. And then we totally propose and combine a multimodal Faster R-CNN architecture. Image-to-image translation models aim to learn the visual mapping between a source and target domain. Although there are some supervised image-to-image translation networks, we want to use any unsupervised image-to-image translation network such as CycleGAN or UNIT. These I2I frameworks are effectively translate given thermal image to expected RGB visual image. The combination of saliency maps with thermal images help illuminate the salient parts of the image, while retaining the textural information in the image. After extracting features, the Faster R-CNN network works on the feature map to detect our goal.

Unpaired Image-to-Image Translation: There is a popular unpaired image-to-image translation framework. That is CycleGAN, which wants to learn mapping functions $F: X \rightarrow Y$ and $G: Y \rightarrow X$ where X and Y are source and target domains respectively.

UNIT [18] tackles the unpaired image-to-image translation problem assuming a shared latent space between both the domains. It learns the joint distribution of images in different domains using the marginal distribution in individual domain.

The trained model provides two generators which perform the translation between source and target domains in case of both CycleGAN and UNIT. We use the generator which performs the Thermal-to-RGB translation, RGB as the target domain while training these modes.

Using Saliency Maps for Helping to Detection: We investigate state-of-the art deep saliency networks. We choose to use the R^3-Net among them and this network uses a Residual Refinement Block(RRB) to learn the residuals between the ground truth and the saliency map in a recursive manner. The RRB alternatively utilizes low-level features and high-level features to refine the saliency maps at each recurrent step by adding the previous saliency map to the learned residual.

Multi-modal Object Detection: There are three branches, one for the thermal image, and other for the RGB image and the other for the saliency map from given thermal image. We use an image-to-image(I2I) translation network to make RGB image in this framework. Outputs from these branches are stacked and passed through a 1 x 1 convolution to learn to combine these features appropriately for the given task. The output of this 1 x 1 convolution is directly passed into the rest of the Faster-RCNN network. We use the Region Proposal Network(RPN) loss as used in original Faster-RCNN as follows:

$$L(\{p_i\},\{t_i\}) = \frac{1}{N}\sum_i L(p_i, p_i^*) + \lambda\frac{1}{N}\sum_i p_i^* R(t_i, t_i^*)$$

where i is the index of an anchor, p_i is the predicted probability of anchor i being an object, p_i^* is the ground truth, t_i represents the coordinates of the predicted bounding box, t_i^* represents the ground truth bounding box coordinates, L is the log loss, R is the robust loss function, and λ is a hyper-parameter. We use the same multi-task classification and regression loss as used in Fast-RCNN.

While the use of existing I2I models allow easy adoption of the proposed methodology, the images generated from such I2I frameworks for thermal-to-RGB translation are perceptually far from natural RGB domain images like MS-COCO [19] and PASCAL-VOC [20]. Therefore, during the training phase of this framework, we also update the weights of the I2I generator to learn to combine the RGB and thermal features in a way that helps improve detection.

4. Experiments

4.1. Datasets and Experimental Setup

Datasets: We use the KAIST Multispectral Pedestrian dataset for our experimental studies. The dataset contains around 95,000 8-bit day and night images consisting of only the Person class. These images are collected using a FLIR A35 microbolometer LWIR camera with a resolution of 320 x 256 pixels. The images are then upsampled to 640 x 512 in the dataset.

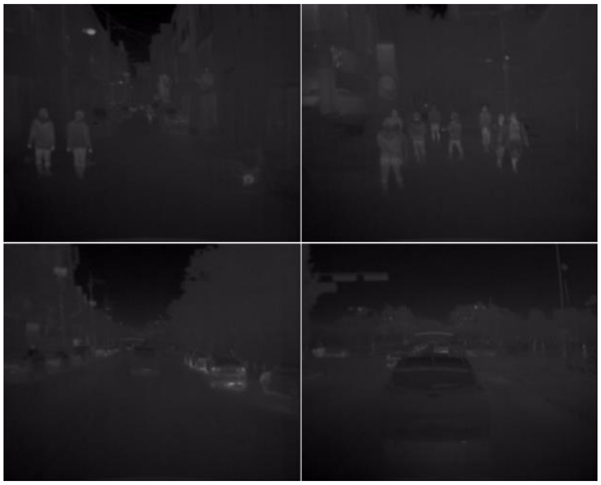

Figure 3. Example Images from KAIST dataset

We use large-scale RGB datasets to improve thermal object detection performance. So we use two popular RGB datasets, PASCAL VOC and MS-COCO. We pre train the object detector on these datasets and use these parameters to initialize the RGB branch of our framework. We finally use the mean Average Precision(mAP) as the performance metric. It is common object detection task.

Baseline: A Faster-RCNN trained in a fully supervised manner on the thermal images from the training set is used as the baseline method for the respective experiments. According to previous studies, we follow the original paper of Faster-RCNN for all the hyperparameters. The KAIST Multispectral Pedestrian datasets provides benchmark test mAP(at IoU of 0.5). We also compare with this result and we will show our framework's performance.

Image-to-Image Translation: We set up and conduct the experiment based on previous studies. We train UNIT and CycleGAN models for KAIST to RGB which uses thermal images from KAIST and RGE images from PASCAL VOC and MS COCO. We use an initial learning rate of 1e-5 for the first 20 epochs, which is decayed to zero over the next 20 epochs. The other hyperparameters of training are as described in original papers.

Deep Saliency Network: We train R^3-Net on thermal images with pixel level annotations. We use an open source implementation and keep the same network architecture as described in the original paper. For training of R^3-Net, we use the implementation of authors and we initialize the parameters of the feature extraction network using weights from the ResNeXt [21] network. We use SGD with learning rate 0.001, momentum 0.9, weight decay 0.0005 and training.

Training our framework Faster-RCNN: We pre-train RGB and thermal detectors with saliency map detector. Our objective is to improve detection in thermal domain, the region proposal network is initialized with weights pre-trained on thermal images. The model is then trained on the same set of images on which the thermal detector was previously pre-trained. We obtain totally three feature maps of 1024 dimension each and these feature maps are stacked back-to-back and passed through a 1 x 1 convolution, which is then passed as input to the region proposal network. RPN then produces the regions of interest that are likely to contain a foreground object. These regions are then cropped out of the feature map and passed into a classification layer which learns to classify he objects in each RoI.

4.2. Results

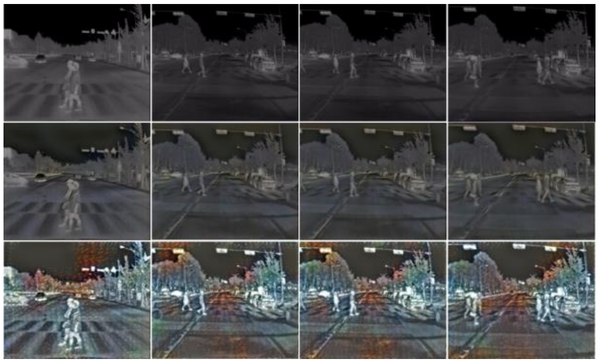

Figure 4. Row 1: Thermal images from KAIST dataset; Row 2: Translation generated using UNIT; Row 3: Translations generated using CycleGAN.

Image-to-image Translation Result: We use UNIT and CycleGAN to translation thermal image to RGB visual image. Figure 4 shows the results of UNIT and CycleGAN trained for thermal to RGB translation. RGB images have both chrominance and luminance information, while thermal images have the luminance part which makes estimating the chrominance for RGB images a difficult task. But by using this method, these generated images add value to the detection methodology.

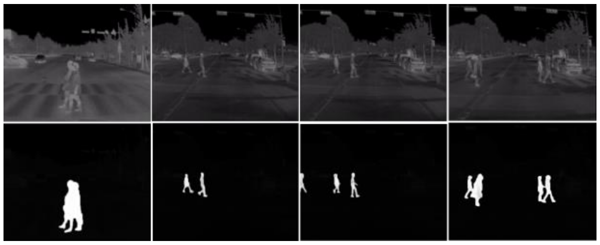

Figure 5. Row 1: Thermal images from KAIST dataset same with Figure 4; Row 2: Result of saliency detection(pixel-level)

Using Thermal Images with Saliency Maps [23] generated from Deep Network Result: We use R^3-Net as a deep network for saliency detection from given thermal images. There are many other networks for saliency map extracting, but recently this network shows the qualitative result and low miss rate. This deep saliency map show that high mean Average Precision (mAP) and low Log Average Miss Rate(LAMR) according to previous study. Figure 5 shows our result of saliency maps generated from deep network. This method illuminates only pedestrians in the scenes, helping the detector identify pedestrians even under difficult lighting conditions. According to known results, deep saliency methods are useful at all times.

Method		**mAP**
Baseline		49.39
Framework	RGB Branch	
MMTOD-UNIT	MSCOCO	54.54
	Pascal VOC	52.40
MMTOD-CycleGAN	MSCOCO	53.56
	Pascal VOC	52.26
Ours-UNIT+R^{3}Net	MSCOCO	54.16
	Pascal VOC	51.30
Ours-CycleGAN+ R^{3}Net	MSCOCO	52.67
	Pascal VOC	52.19

Table 1. Performance comparison baseline on KAIST dataset with MMTOD [22] and Ours.

Thermal Object Detection Result: Table 1 shows the comparison of mAP of our framework against the baseline detector when trained on KAIST datasets respectively. KAIST dataset has only one class, the Person. We observe that in all the experiments, our framework is comparable with previous method such as MMTOD with CycleGAN. During the translation process, the results of UNIT did not show much noticeable change, but CycleGAN changed significantly. However, when we looked at previous studies, we could see that mAPs by UNIT were slightly higher. Then, we checked performance here using our framework. We expected to produce better results using the saliency map, but compared to previous studies, we could confirm that there was no significant difference. Sometimes they show high results, but when tried several times, they were able to see similar mAP. Our architecture also showed several limitations, an increase in the computational cost once the deep saliency network and one more channel was added to the existing work. Furthermore, we have shown inconspicuous results in overall performance improvements. There is a possibility of development when this method is used enough. New tuning of hyperparameters or training in several ways may result in better results. It seems that the structure of the model needs to be slightly modified, and in the future, it may be developed using YOLO or SSD.

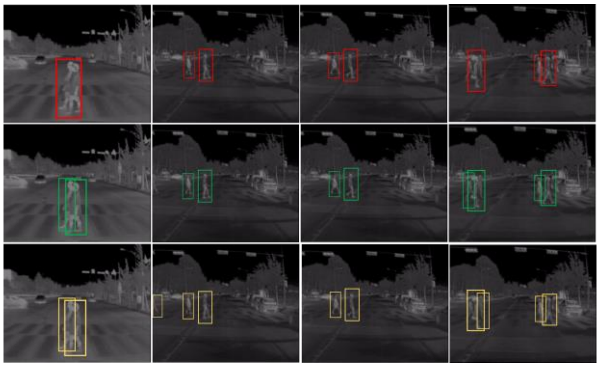

Figure 6. Results of detection on the KAIST. Row 1: Baseline; Row 2: MMTOD; Row3: Ours

5. Conclusion

We propose a novel multi-modal framework to extend and improve upon any Region-CNN-based object detector in the thermal domain by borrowing features from RGB domain and saliency map. In this case, we only need thermal image and we evaluate the performance of our framework applied to a Faster RCNN architecture. Borrowing the features from RGB domain and saliency map showed the potential for development of object detection in thermal imagery.

6. Future Work

In this paper, we augmented thermal images with their saliency maps and used it to extract features. We will be able to utilize different networks to use the saliency map more effectively, in addition to using features in three domains at one, as well as using models differently. Moreover, this saliency technique and multi-modal frame work can develop in other many datasets and environments.

References

[1] R. B. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. CoRR, abs/1311.2524, 2013.

[2] R. B. Girshick. Fast R-CNN. CoRR, abs/1504.08083, 2015.

[3] S. Ren, K. He, R. B. Girshick, and J. Sun. Faster R-CNN: towards real-time object detection with region proposal networks. CoRR, abs/1506.01497, 2015.

[4] C. Koch and S. Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. In Matters of intelligence, 1987.

[5] 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.

[6] J. W. Davis and M. A. Keck. A two-stage template approach to person detection in thermal imagery. In 2005 Seventh IEEE Workshops on Applications of Computer Vision, 2005

[7] K. J¨ungling and M. Arens. Feature based person detection beyond the visible spectrum. 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, pages 30–37, 2009.

[8] H. Bay, T. Tuytelaars, and L. Van Gool. Surf: Speeded up robust features. volume 3951, pages 404–417, 07 2006.

[9] M. Peng, C. Wang, T. Chen, and G. Liu. Nirfacenet: A convolutional neural network for near-infrared face identification. Information, 7(4), 2016.

[10] B. C. I. Rodger and N. Robertson. Classifying objects in lwir imagery via cnns. In In Proc. SPIE: Electro-Optical and Infrared Systems: Technology and Applications XII, 2016.

[11] Z. Deng, X. Hu, L. Zhu, X. Xu, J. Qin, G. Han, and P.-A. Heng. R3Net: Recurrent residual refinement network for saliency detection. In IJCAI, 2018.

[12] J. Wagner, V. Fischer, M. Herman, and S. Behnke. Multispectral pedestrian detection using deep fusion convolutional neural networks. 04 2016.

[13] D. Knig, M. Adam, C. Jarvers, G. Layher, H. Neumann, and M. Teutsch. Fully convolutional region proposal networks for multispectral person detection. In 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pages 243–250, July 2017.

[14] J. Liu, S. Zhang, S. Wang, and D. N. Metaxas. Multispectral deep neural networks for pedestrian detection. CoRR, abs/1611.02644, 2016.

[15] P. Isola, J. Zhu, T. Zhou, A. Efros. Image-to-image translation with conditional adversarial nets, 2017.

[16] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In Computer Vision (ICCV), 2017 IEEE International Conference on, 2017.

[17] A. Anoosheh, E. Agustsson, R. Timofte, L. Gool. ComboGAN: unrestrained scalability for image domain translation. 2017.

[18] M. Liu, T. Breuel, and J. Kautz. Unsupervised image-to-image translation networks. CoRR, abs/1703.00848, 2017.

[19] T. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Doll´ar, and C. L. Zitnick. Microsoft COCO: common objects in context. CoRR, abs/1405.0312, 2014.

[20] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html.

[21] S. Xie, R. Girshick, P. Dollr, Z. Tu, and K. He. Aggregated residual transformations for deep neural networks. arXiv preprint arXiv:1611.05431, 2016.

[22] C. Devaguptapu, N. Akolekar, M. Sharma, V. Balasuburamanin. Borrow from Anywhere: Pseudo multi-modal object detection in thermal imagery. 2019.

[23] D. Ghose, S. Desai, S. Bhattacharya, D. Chakraborty. Pedestrian detection in thermal images using saliency maps, 2019.