Partially Fake it Till you Make It:
Mixing Real and Fake Thermal Images for
Improved Object Detection

Francesco Bongini, Lorenzo Berlincioni*
Marco Bertini, Alberto Del Bimbo
MICC - Università degli Studi di Firenze
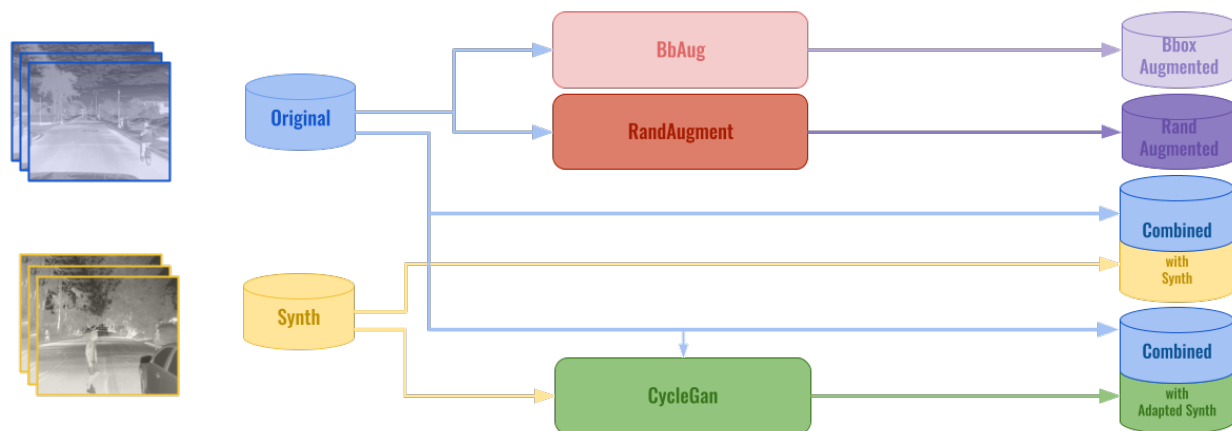Firenze, Italy
[name.surname]@unifi.it

Figure 1: General overview of our data augmentation pipelines; Original and Synthetic datasets get augmented and combined in different ways to provide different training sets, see Tab.1 for experimental results.

ABSTRACT

In this paper we propose a novel data augmentation approach for visual content domains that have scarce training datasets, compositing synthetic 3D objects within real scenes. We show the performance of the proposed system in the context of object detection in thermal videos, a domain where *i)* training datasets are very limited compared to visible spectrum datasets and *ii)* creating full realistic synthetic scenes is extremely cumbersome and expensive due to the difficulty in modeling the thermal properties of the materials of the scene. We compare different augmentation strategies, including state of the art approaches obtained through RL techniques, the injection of simulated data and the employment of a generative model, and study how to best combine our proposed augmentation with these other techniques. Experimental results demonstrate the

effectiveness of our approach, and our single-modality detector achieves state-of-the-art results on the FLIR ADAS dataset.

CCS CONCEPTS

• **Computing methodologies** → **Neural networks**; **Learning from critiques**; **Object detection**; *Image representations.*

KEYWORDS

computer graphics, domain adaptation, pedestrian detection, thermal videos, ADAS, GAN

1 INTRODUCTION

Object detection is a core problem for the perception capabilities of an autonomous vehicle, the identification of its surroundings and of nearby objects is essential to ensure a safe deployment of autonomous cars on the road. In autonomous driving the object detection task is required to be particularly robust across a range of illumination and environmental conditions, including daytime,

---

nighttime, rain, fog, etc. In such conditions, detectors based solely on visible spectrum imagery can easily fail [36, 42], as shown in Fig. 2.

The use of thermal detectors has recently increased as a mean to mitigate the sensitivity of visible spectrum imagery to scene-incidental imaging conditions [4, 29, 36]. A growing number of works have also investigated multispectral detectors combining visible and thermal images for robust pedestrian detection [2, 9, 26, 39, 42, 43, 61, 64].

Figure 2: Top row. *Left)* **The lights of the car headlights dazzle the RGB camera;** *right)* **cars and persons can be recognized in the same view obtained through a thermal camera. Bottom row.** *Left)* **Lack of lighting may cause dangerous situations, no pedestrian is observable in the scene.** *Right)* **the thermal camera does not depend from the intensity of the light and allows to detect the person.**

The cost of multiple aligned sensors however is a limiting factor due to cost, camera size or availability of lighting to operate the visible spectrum camera, and multispectral models can have limited applicability in real-world applications. Moreover, using visible spectrum sensors does not offer the same privacy-preserving affordances as systems employing only thermal sensors [36].

When compared to multispectral detectors, thermal-only ones typically reach lower performances, and besides the lack of the information coming from the visible spectrum there are other challenges for robust object detection using only thermal data. A key performance-limiting factor is the relative lack of annotated thermal imagery available for training state-of-the-art models. Thermal pedestrian datasets are few, and – compared to visible-spectrum datasets – have orders of magnitude fewer annotated instances. As an example the Caltech Pedestrian Dataset [18] has 350,000 annotations in the visible domain, while in thermal domain KAIST Multispectral Pedestrian dataset [33] has ∼ 51, 000 annotations and FLIR ADAS Dataset [1] has ∼ 28, 000. Bringing thermal-only detection to the robustness and accuracy demanded by real-world applications is thus extremely difficult due to the scarcity of annotated data.

Creating realistic training data from 3D models has become a common practice in several domains, e.g. for Advanced Driver Assistance Systems (ADAS). An example is the CARLA simulator [19] that can create RGB, LiDAR, depth, etc. views of a simulated scene. However, thermal imagery is not yet automatically generated, since the 3D engine should account for a very large number of simulation parameters associated to different materials and heat sources.

In this work we propose to simplify the creation of training data adding only a subset of objects of interest to a real scene, thus reducing the cost of modeling a whole environment. In order to improve the quality and visual likelihood of the objects we use a GAN, to further adapt the appearance of the 3D models and to better simulate the output of the shaders used in the 3D animation engine.

The contributions of this paper are:

- **Compositing of 3d fake thermal objects in a real thermal scene**: to the best of our knowledge we are the first to propose this approach, addressing the thermal spectrum domain, in which creating full virtual scenes is not practical. We further propose to use a GAN-based approach to adapt the appearance of the composited objects in the scene.
- **Extensive comparison of data augmentation techniques**: we provide an extensive comparison of recent augmentation techniques that have been originally proposed and evaluated on RGB datasets, like RandAugment [15] and BBAug [70], on thermal data; we evaluate their combinations, either between them and with our 3D compositing approach.
- **State-of-the-art results**: we compare the best thermal augmentations, either alone or in combination, with a large number of multispectral and thermal object detectors. Our proposed augmentation technique outperform all the thermal-only approaches and all the multispectral methods but one.

2 RELATED WORK

2.1 Object detection in thermal imagery

Thanks to the reduction of costs and availability of multispectral cameras over the past few years, there are numerous recent works exploiting thermal images in combination with visible images for robust pedestrian detections [10, 20, 39, 41–43, 46, 62, 64, 65, 67, 68]. One of the key issues with thermal imagery is the image resolution and quality is typically far lower compared to RGB images, and the application of RGB detectors on this low-quality data yields lower performances. More recently works as MMTOD-UNIT [16] use a multi-spectral detector that aims to borrow the knowledge from the data-rich domains such as visual (RGB) without the explicit need for a paired multimodal dataset. In fact, the model introduces pseudo-visible images generated from the thermal spectrum ones using the CycleGAN [69], promoting more information to the training set. The multi-modal Faster-RCNN detector is then trained. Another recent multi spectral detector is CFRM_3 [66] that instead cyclically fuse and refine more spectra. The idea is to use a novel cycle fuse-and-refine module to predict the segmentation mask features of both visible and thermal spectrum. This allows to complementary get important features from different spectra. Note that this method uses only aligned pictures so this is not directly comparable with

the other models, although multispectral cameras typically do not provide aligned frames due to differences in the focal length of the two lenses. In particular, 4,129 well-aligned image pairs have been used for training and 1,013 image pairs for test.

In contrast, many recent works have investigated pedestrian detection in the thermal (IR) domain only. For example, authors in [34] used Adaptive fuzzy C-means for IR image segmentation and a CNN for pedestrian detection. In [4] the authors proposed a combination of Thermal Position Intensity Histogram of Oriented Gradients (TPIHOG) and the additive kernel SVM (AKSVM) for nighttime-only detection in thermal imagery. Thermal images augmented with saliency maps, used as attention mechanism, have been used in [24].

The idea of performing several video preprocessing steps to make thermal images look more similar to grayscale images converted from RGB was investigated in [29], who then applied a pretrained and fine-tuned SSD detector. Recently, authors in [11] designed dual-pass fusion block (DFB) and channel-wise enhance module (CEM) to improve the one-stage detector RefineDet, and proposed their ThermalDet detector for pedestrian detection in thermal imagery. Another recent single-modality work was the Bottom-up Domain Adaptation approach proposed in [35, 36] for pedestrian detection in thermal imagery. Task-conditioned training has been recently proposed in [37], adding the auxiliary task of classifying night and day thermal images and obtaining state-of-the-art results. Also in this work, we focus on the thermal-only detection problem.

2.2 Data augmentation from 3D models

Creation of virtual images, typically using 3D graphic engines, to improve object detection and pose estimation in visual spectrum domain has received a lot of attention from the research community. In the autonomous driving field in particular countless synthethic datasets and virtual environments [21, 49, 56, 57] are used, especially for reinforcement learning models. The advantages of using a virtual environment are multiple such as the ease of collection, the control over the generated data and the possibility to extract information from multiple, perfectly aligned, virtual sensors i.e. in a 3d videogame the RGB image, its segmentation map and depth map *come for free* from the render program. Specific tools have been developed to ease the creation of these scenarios, like CARLA [19], that is specialized for the creation of scenarios for ADAS applications, and VIVID [40], that can be used for indoor navigation, action recognition and event detection, and provides an advanced human skeleton system to simulate complex human actions. Most of the effort for realism in these tools, as well as in the industry of 3d videogame engines, is focused on the visible spectrum, and the simulated thermal views of a scene usually lack the nuances of its visible counterpart.

Because of this, works that have explored the use of tools to create augmented datasets with 3D data have addressed only the visible spectrum context. Some works have used 3D data to estimate hand [54] or body poses [13, 63]. 3D avatars have been recently used to enhance human action recognition in [5, 48]. Object detectors have been trained using 3D syntethic data in [7] and [17]. The former considered classes like person, vehicles and animals

in outdoor city scenes, evaluating the trained detector on surveillance videos. The latter investigates the effectiveness of rendering engines in generating realistic scenes for scenarios where no or insufficient annotated data is available, considering the detection of protective equipment in construction sites.

To the best of our knowledge we are the first to address compositing of 3D objects within real scenes in the thermal domain.

2.3 Data augmentation from synthetic images

The exploiting of synthetic data from a simulator to obtain a larger, more diverse, dataset has been explored in many works. This presents a domain adaptation challenge in which the source domain that needs to be adapted is partially controllable. One approach to this problem focuses on the model i.e. by adding to it a domain classifier and using gradient reversal to close the gap between the two distributions [22] or by freezing the feature extraction section of the model when training it on synthetic data [35]. Another technique works on closing the domain shift by having a shared representation space [31, 47, 58]. Another possible approach instead separately preprocesses the input data using a learned generative model, trained to perform a translation between the two domains. [6, 30, 32, 59]. Also in [16] the authors used a CycleGAN [69] for image-to-image translation of thermal to pseudo-RGB data. The use of these frameworks to perform data augmentation in order to improve the performance of a separate classifier has been studied in multiple previous works such as [3] in which they focus on improving one-shot learning, in [8] where segmentation of medical images is enhanced by GAN augmented data. In [55] synthetic data coming from a simulator is adapted and used to train an RL agent for autonomous driving. Similar to our approach are [30] in which a much more elaborate cycle consistent framework is developed to perform domain adaptation and [50] that improves the CycleGAN approach by adding per instance masks. In [38], a GAN is used to produce fake thermal images to increase the training data; the authors study the best approach on how to mix these additional data to real thermal images, showing how adding 10%-20% percent of fake images improves the performance of the object detection.

2.4 Data augmentation from RGB images

AutoAugment [14] is a augmentation framework for vision models that casts the search of parameters for data augmentation as an optimization problem and solve it using reinforcement learning. RandAugment [15] improves on AutoAugment [14] by both considerably reducing the parameters search space from 10^{32} to 10^2 and matching or exceeding performances of [14]. Both this models are tailored to image classification tasks, more recently a number of works focused on data augmentation specifically developed for detection problems, such as [53] with a neural rendering approach and [70] using a similar technique to [15] the authors train a RNN model in a RL setting to learn optimal augmentation policies including in this transformations translating, zooming or distorting both the image and the bounding boxes.

3 THE PROPOSED METHOD

3.1 Mixing real and fake thermal imagery

The Unity game engine [60] is a powerful framework used to develop modern 3D games. With the help of a 3D artist we exploit this tool to get synthetic data by rendering sequences of realistic 3D characters and cars superimposed over a real thermal spectrum image as background. Fig. 3 shows an example of the process: a real scene from FLIR dataset is used to composite and animate 3D objects, creating a virtual video sequence. The game engine allowed the use of a raw *thermal* shader that simulates the thermal signature of different objects. The 3D artist needs to concentrate only on creating the raw thermal appearance of these objects, without need to design the whole scene. This would have required to compute the thermal properties of all the objects in the background.

Figure 3: *Left)* **Original thermal image from FLIR dataset [1],** *Right)* **Image with composited fake thermal objects.**

We automatically annotated the objects and collected over 10K samples used to augment the training set. In order to better analyze the synthetic data effect on the training we produced different sets of simulated scenes changing parameters such as the source of the background and the number of instances of the different detected categories (Person, Car, Bicycle).

3.2 Improving fake thermal imagery appearance through GAN

We use a Generative Adversarial Network to improve the realism of the data coming from the simulator. In our approach we rely on an LSGAN objective trained in a cycle-consistent regime as in [69]. In order to achieve this we trained the generative model in a cycle-consistent setting between the *source* domain, the synthetic data, and the *target* domain, the data from FLIR dataset. The model architecture is taken from [38] as it proved to be successful in the thermal images augmentation setting and it is built using the Residual in Residual Dense Block (RRDB) as the fundamental unit. As in [44], we remove the batch normalization layer from the traditional *Conv-BN-LReLU* triplet. Similarly to [50] the model is fed both the input image and the segmentation mask (see Fig. 4) as to provide guidance to the model for the different thermal signatures of each category. Fig. 5 provides a sample of the synthetic to real thermal translation; it is interesting to note how the GAN has created the thermal reflection of the heat of the engine on the concrete of the road, that was not part of the 3D model. This physical property has been learned by the GAN from the FLIR thermal dataset. Differently from [50] we do not want our segmentation mask to be modified

by the generative model as this would change its bounding box and be detrimental to the detector performances.
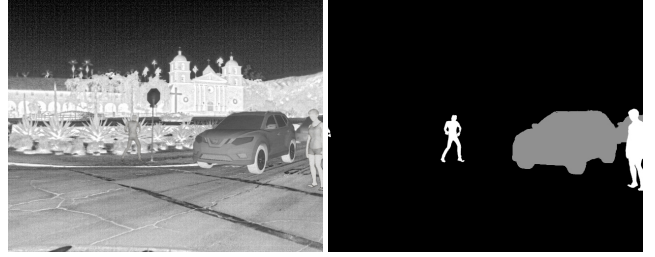

Figure 4: GAN Input data sample: *left)* **composited 3D fake thermal objects in a scene;** *right)* **segmentation masks of the 3D objects** (img, seg)
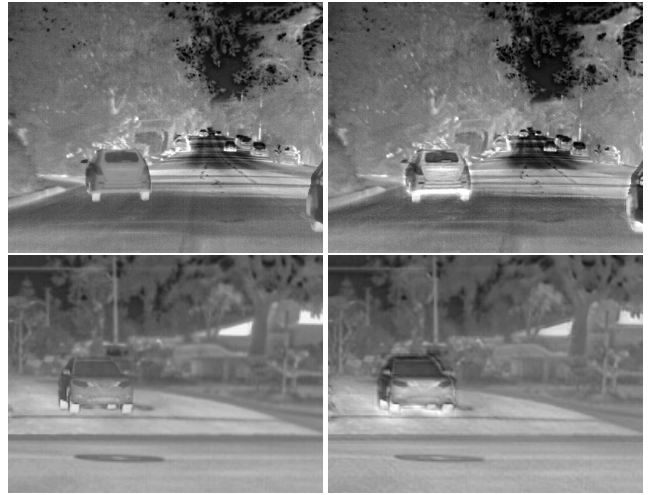

Figure 5: GAN translation synth to real sample (zoomed in and cropped). *Left:* **Original synthethic image from the simulator;** *Right:* **GAN translated version. The GAN learns to transfer the heat emitted from the bottom of a car, a feature missing from the simulator.**

3.3 Other data augmentations

We have also evaluated the applicability of data augmentation techniques developed for RGB images in the thermal spectrum.

3.3.1 RandAugment. We searched the best combinations and strengths of augmentations over a subset of the FLIR training dataset composed of 5,000 images. However, we noticed that some of the transformations used are not suitable for the thermal domain, as shown in Fig. 6.

3.3.2 BBox Augmentation. While data augmentation methods like RandAug are often used for image classification models, Boundary-Box Augmentation [70] (BBAug) focuses on object detection. In the FLIR dataset the annotations refer to objects like bikes, pedestrians and cars; the bounding box annotations of these objects open up the
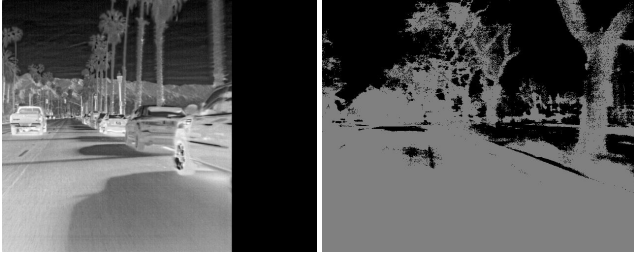
Figure 6: Examples of augmentations obtained with RandAugment [15]. *Left)* translateX and autocontrast have been applied to a picture of the FLIR dataset; *Right)* in this case the pedestrians, bicycles and cars are not detectable at the all due to the solarization transformation.

possibility to modify the content of the boxes. The transformations include, for example, to keep the same image while translating or zooming the content of the boxes. Starting from an original picture of the FLIR data set in figure 7, the content of the bounding box in the middle is modified by this technique.

BBAug [70] is based on the concept of policy, an unordered set of K sub-policies. Each sub-policy is a set of transformations to be applied to the image, similarly to the AutoAugment method originally proposed in [14]. The authors provide the best sub-policies using a Reinforcement Learning search space algorithm. The models are trained on COCO data set with ResNet-50 backbone [28] and RetinaNet detector [45]. For each picture we want to augment, a sub-policy is randomly selected. This method allows us to skip the very expensive search space phase.

Figure 7: Left) An image from FLIR dataset; Right) BBox Augmented version. The content of the box's car in the middle is translated downwards and a cutout square has been added.

3.4 Object detector

We used the layer-wise thermal adaptation of YOLOv3 originally proposed in [36] as object detector, as a baseline for the experiments and to evaluate the data augmentation technique proposed in this paper.

4 EXPERIMENTAL RESULTS

4.1 Dataset and experimental setup

All of our experiments were conducted on the FLIR ADAS dataset [1]. The dataset consists of over 10K images, collected using the

FLIR Tau2 camera, both during the day (60%) and the night (40%) driving on Santa Barbara, CA with clear to overcast weather. We used the three most common classes offered by the FLIR-ADAS dataset, composed by Person (28,151), Car (46,692), and Bicycle (4,457), as this is the commonly used experimental setup and allows us to compare the proposed method with other state-of-the-art approaches.

The training set contains 8,862 frames, while the test set contains 1,366 images taken on different streets from the training set ones. Several detectors have been compared using the mAP score. For the GAN training we used a subset of the FLIR-ADAS data from [66] which provides pixel-aligned couples of RGB and thermal frames. We then used the RGB frames to extract segmentation masks for pedestrians and vehicles from the scene, using a pretrained on MSCOCO MaskRCNN [27] from Detectron2 [25].

We devise and test multiple augmentation strategies by combining the different source of data augmentation previously discussed. From here on we will refer to BBox Augmentation [70] as "BBAug" and to we will keep the name RandAugment for the method of [15].

As previously noted, we added synthetic data to the training set. This data is split into multiple sets in order to perform ablation studies:

- $Synth_a$: Composed of pedestrians walking on a railroad scene; No cars and no roads. About 4K samples.
- $Synth_b$: Composed of both cars and pedestrians over FLIR-ADAS training set scenes. About 10K samples.
- $Synth_c$: Composed of both cars and pedestrian over a railroad scene; no roads. About 4K samples.

We will use the following notation: $GAN_x \rightarrow {}_y$ to refer to an experiment in which a generative model is trained over the subset $Synth_x$, performs inference over $Synth_y$ and finally this *translated* $Synth_y$ data is added to the detector training set. We also tested the generative model from [38] which performs RGB → Thermal translation, and refer to it as $GAN_{rgb} \rightarrow {}_{th}$.

4.2 Ablation studies

In the first experiment, we investigated several data augmentation strategies to improve the performances of our detector. In Table 1 we compare the results of different augmentation techniques and the baseline from [36] in terms of mean average precision (mAP) for all the detected classes. In the second and third experiment, we tested RandAugment [15] and BBAug [70], by picking the best policy for each (see Tab. 2) and the best amount of augmented data (see. Tab. 3).

Analyzing the results of Tab. 1 we can observe that BBAug [70] performed better than Randaugment, bringing the base detector mAP to 75%. Employing Randaugment [15] instead does not show much improvement, probably due to the fact that its transformations are not developed for object detection as shown in Fig.6. The combination of the aforementioned does note give better results than BBAug [70] alone.

The augmentation strategy using the generative model from [38] provides some improvements compared to the baseline method but only by a 0.3% mAP.

The injection of synthetic data, through compositing of fake 3D objects in a thermal scene, instead, both directly from the simulator (Synth$_a$, Synth$_b$, Synth$_c$) and after the domain translation (GAN$_b \rightarrow _a$, GAN$_b \rightarrow _c$), raises the detection performances across all the classes. The best pipeline we found working uses the translated synthetic data from our Synth$_a$ subset, reaching a 76.4% mAP. Also using the same subset followed by BBAug [70] improved the detector mAP performances to 75.6% which is the second best result. The combination of these two last techniques does not yield better results. In general, the proposed idea of compositing fake 3D thermal objects in a thermal scene is an effective augmentation technique. This is further improved adding BBAug or GAN-based augmentation that improves the appearance of the 3D objects. Instead, combining both GAN and BBAug does not improve the results.

Table 1: Comparison of different augmentation strategies in terms of mAP. Best in bold, second best underlined.

Augmentation Strategy	Person	Bicycle	Car	mAP
None (Baseline) [36]	75.6	57.4	86.5	73.2
RandAugment [15]	74.4	60.2	85.4	73.3
BBAug [70]	79.4	58.4	87.2	75.0
RandAugment [15] + BBAug [70]	74.6	61.6	86.0	73.9
GAN$_{rgb} \rightarrow _{th}$ [38]	77.1	56.9	86.4	73.5
Synth$_a$	79.3	60.1	86.2	75.2
Synth$_a$ + BBAug [70]	77.1	64.2	85.6	<u>75.6</u>
Synth$_b$	77.2	60.0	86.2	74.5
Synth$_c$	75.9	60.5	85.2	73.9
GAN$_b \rightarrow _c$	77.6	59.1	85.6	74.1
GAN$_b \rightarrow _a$	78.5	64.0	86.9	**76.4**
GAN$_b \rightarrow _a$ + BBAug [70]	74.7	64.1	86.5	75.1

As previously explained, Boundary Box augmentation [70] introduces 4 policies [1] each composed of subpolicies (i.e. TranslateX $p = 0.6$, $m = 4$, Equalize $p = 0.8$, $m = 10$), which are defined as 2 possible transformations that can be applied to an image. We tested all of them and present them in Tab. 2, concluding that the policy v0 is preferable in FLIR-ADAS. All the experiments in Tab. 1, Tab. 3, Tab. 4 use this policy.

Table 2: mAP for the 4 policies of the BBAug [70] over a subset of the synthetic data.

Policies	v0	v1	v2	v3
Person	77.1	73.4	72.2	72.9
Bicycle	64.2	53.7	54.3	54.5
Car	85.6	84.9	82.6	84.6
mAP	**75.6**	70.7	69.7	70.7

We have also explored the effect of different quantities of augmented data w.r.t the original train set size. In Table 3 a non exhaustive analysis of the percentage of fake w.r.t real suggests, as previously noted in [38], that the optimal amount of generated data to add is between 10% and 20%, for the tested strategies increasing the amount of augmented data becomes detrimental to learning.

---

[1] https://github.com/tensorflow/tpu/blob/master/models/official/detection/utils/autoaugment_utils.py#L15

Table 3: Ablation study on varying quantities of augmented images. The percentage indicates the relative size of the added augmented data.

Technique	@10%	@20%	@50%
BBAug [70]	74.8	75.0	74.4
RandAugment [15]	72.9	73.3	66.5
RandAugment [15] + BBAug [70]	73.9	73.7	73.6
GAN$_{rgb} \rightarrow _{th}$ [38]	73.5	72.6	72.7
Synth$_a$	74.2	74.1	74.0

4.3 Comparison with state-of-the-art approaches

In this experiment we compare the best augmentation techniques evaluated in the previous section, comprising our proposed composition of fake thermal 3D objects with and without GAN processing, with state-of-the-art approaches. All the following experiments are evaluated in terms of Mean Average Precision (mAP).

Table 4: Comparison of the proposed method with state-of-the-art multispectral and thermal-only methods. Best in bold, second best underlined.

Technique	Spectra	Person	Bicycle	Car	mAP
MMTOD-UNIT [16]	Multi	64.5	49.4	70.8	61.5
ABiFN [12]	Multi	66.1	48.5	71.8	62.1
SSD512 VGG16 [52]	Multi	71.0	55.5	82.3	69.6
CFRM_3 [66]	Multi	74.5	57.8	84.9	72.4
SSTN101 [51]	Multi	-	-	-	77.6
YoloV3 Transfer[23]	Thermal	33.2	34.5	55.4	41.0
SSD VGG-16 Transfer [23]	Thermal	61.9	46.1	85.1	64.4
Layer-Wise [36]	Thermal	75.6	57.4	86.5	73.2
RefineDetect512 [1]	Thermal	79.4	58.0	85.6	74.3
ThermalDet [11]	Thermal	78.2	60.0	85.2	74.6
BBAug [70]	Thermal	79.4	58.4	87.2	75.0
Synth$_a$	Thermal	79.3	60.1	86.2	75.2
Synth$_a$ + BBAug[70]	Thermal	77.1	64.2	85.6	<u>75.6</u>
GAN$_b \rightarrow _a$	Thermal	78.5	64.0	86.9	**76.4**

Table 4 compares our results with the state-of-the-art single and multi-spectral approaches. We distinguish multi-spectral detectors, i.e. models which at test time detect objects using both visible and thermal spectrum images, and thermal only detectors, i.e. models that use only the thermal spectrum images. Models which used visible images for transfer learning only during training fall under the *Thermal* category.

YoloV3 Transfer[23] and SSD VGG-16 Transfer [23] are both pretrained models on Coco and ImageNet data sets and tested on FLIR-ADAS. As shown in the Tab.4, our thermal only detector using different augmentations outperform all but one of the multi-spectral detectors. Only Self-Supervised Thermal Network (SSTN101) [51] got better results with 77.57% mAP. Instead, considering thermal-only detectors, our proposed method outperforms all the competing approaches.

Fig. 8 compares the detections obtained using the layer-wise YOLOv3 baseline presented in [35], the results obtained adding the BBAug augmentation technique [70] and the results obtained using our proposed approach, i.e. adding 3D models and improving their

appearance with a GAN. The examples on row 2 show that the kid crossing the street was not detected by the method proposed in [35], but both augmentation approaches are now capable of correctly detecting him. Row 3 shows how adding 3D models, as proposed in our method, help to distinguish persons from bicycles, that are not detected with the two other approaches. Another example of correct detection of a bicycle, is shown in row 4, were both augmentations are able to recognize it.

5 CONCLUSIONS

In this paper we compared several data augmentation strategies to improve a YOLOv3 detector working in a thermal-only domain. Given the challenges of the task, such as its data scarcity, we presented an augmentation pipeline that combines simulated data within a real scene, followed by domain adaptation using a generative model. This approach is particularly suitable for domains, such as thermal imagery, where the creation of completely synthetic scenes is unfeasible or extremely expensive due to the difficulty of modeling all the physical properties of the entities of the scene. Our best combination of strategies reaches states of the art performance in the thermal only domain, and reaches similar or better results with respect to multi-spectral detectors.

ACKNOWLEDGMENTS

REFERENCES

[1] 2019. https://www.flir.com/oem/adas/adas-dataset-form/
[2] Anelia Angelova, Alex Krizhevsky, Vincent Vanhoucke, Abhijit Ogale, and Dave Ferguson. 2015. Real-time pedestrian detection with deep network cascades. In *Proc. of British Machine Vision Conference (BMVC)*.
[3] Antreas Antoniou, Amos Storkey, and Harrison Edwards. 2018. Augmenting Image Classifiers Using Data Augmentation Generative Adversarial Networks. In *Proc. of Artificial Neural Networks and Machine Learning (ICANN)*. 594–603.
[4] Jeonghyun Baek, Sungjun Hong, Jisu Kim, and Euntai Kim. 2017. Efficient pedestrian detection at nighttime using a thermal camera. *Sensors* 17, 8 (2017), 1850.
[5] Mohamad Ballout, Mohammad Tuqan, Daniel Asmar, Elie Shammas, and George Sakr. 2020. The benefits of synthetic data for action categorization. In *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–8.
[6] R. Barth, J. Hemming, and E.J. Van Henten. 2020. Optimising realism of synthetic images using cycle generative adversarial networks for improved part segmentation. *Computers and Electronics in Agriculture* 173 (2020), 105378. https://doi.org/10.1016/j.compag.2020.105378
[7] Erik Bochinski, Volker Eiselein, and Tomas Sikora. 2016. Training a convolutional neural network for multi-class object detection using solely virtual world data. In *Proc. of IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. IEEE, 278–285.
[8] Christopher Bowles, Liang Chen, Ricardo Guerrero, Paul Bentley, Roger N. Gunn, Alexander Hammers, David Alexander Dickie, Maria del C. Valdés Hernández, Joanna M. Wardlaw, and Daniel Rueckert. 2018. GAN Augmentation: Augmenting Training Data using Generative Adversarial Networks. *ArXiv* abs/1810.10863 (2018).
[9] Garrick Brazil, Xi Yin, and Xiaoming Liu. 2017. Illuminating pedestrians via simultaneous detection & segmentation. In *Proc. of IEEE International Conference on Computer Vision (ICCV)*.
[10] Yanpeng Cao, Dayan Guan, Yulun Wu, Jiangxin Yang, Yanlong Cao, and Michael Ying Yang. 2019. Box-level segmentation supervised deep neural networks for accurate and real-time multispectral pedestrian detection. *ISPRS Journal of Photogrammetry and Remote Sensing* 150 (2019), 70–79.

[11] Y. Cao, T. Zhou, X. Zhu, and Y. Su. 2019. Every Feature Counts: An Improved One-Stage Detector in Thermal Imagery. In *Proc. of IEEE International Conference on Computer and Communications (ICCC)*. 1965–1969.
[12] A Sai Charan, M Jitesh, M Chowdhury, and H Venkataraman. 2020. ABiFN: Attention-based bi-modal fusion network for object detection at night time. *Electronics Letters* 56, 24 (2020), 1309–1311.
[13] Wenzheng Chen, Huan Wang, Yangyan Li, Hao Su, Zhenhua Wang, Changhe Tu, Dani Lischinski, Daniel Cohen-Or, and Baoquan Chen. 2016. Synthesizing training images for boosting human 3d pose estimation. In *Proc. of International Conference on 3D Vision (3DV)*. IEEE, 479–488.
[14] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. 2019. Autoaugment: Learning augmentation strategies from data. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 113–123.
[15] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. 2020. Randaugment: Practical automated data augmentation with a reduced search space. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPR-W)*.
[16] Chaitanya Devaguptapu, Ninad Akolekar, Manuj M Sharma, and Vineeth N Balasubramanian. 2019. Borrow from Anywhere: Pseudo Multi-modal Object Detection in Thermal Imagery. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPR-W)*.
[17] Marco Di Benedetto, Fabio Carrara, Enrico Meloni, Giuseppe Amato, Fabrizio Falchi, and Claudio Gennaro. 2020. Learning accurate personal protective equipment detection from virtual worlds. *Multimedia Tools and Applications* (2020). https://doi.org/10.1007/s11042-020-09597-9
[18] Piotr Dollar, Christian Wojek, Bernt Schiele, and Pietro Perona. 2011. Pedestrian detection: An evaluation of the state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 34, 4 (2011), 743–761.
[19] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. 2017. CARLA: An Open Urban Driving Simulator. In *Proc. of the 1st Annual Conference on Robot Learning*. 1–16.
[20] Kevin Fritz, Daniel König, Ulrich Klauck, and Michael Teutsch. 2019. Generalization ability of region proposal networks for multispectral person detection. In *Proc. of Automatic Target Recognition XXIX*, Vol. 10988.
[21] Adrien Gaidon, Qiao Wang, Yohann Cabon, and Eleonora Vig. 2016. Virtual Worlds as Proxy for Multi-Object Tracking Analysis. *CoRR* abs/1605.06457 (2016). arXiv:1605.06457 http://arxiv.org/abs/1605.06457
[22] Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In *Proc. of International Conference on Machine Learning (ICML)*. PMLR, 1180–1189.
[23] Yona Falinie A Gaus, Neelanjan Bhowmik, Brian KS Isaac-Medina, and Toby P Breckon. 2020. Visible to infrared transfer learning as a paradigm for accessible real-time object detection and classification in infrared imagery. In *Proc. of Counterterrorism, Crime Fighting, Forensics, and Surveillance Technologies IV*, Vol. 11542. International Society for Optics and Photonics, 1154205.
[24] Debasmita Ghose, Shasvat M Desai, Sneha Bhattacharya, Deep Chakraborty, Madalina Fiterau, and Tauhidur Rahman. 2019. Pedestrian Detection in Thermal Images using Saliency Maps. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPR-W)*.
[25] Ross Girshick, Ilija Radosavovic, Georgia Gkioxari, Piotr Dollár, and Kaiming He. 2018. Detectron. https://github.com/facebookresearch/detectron.
[26] Dayan Guan, Yanpeng Cao, Jiangxin Yang, Yanlong Cao, and Michael Ying Yang. 2019. Fusion of multispectral data through illumination-aware deep neural networks for pedestrian detection. *Information Fusion* 50 (2019), 148–157.
[27] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask R-CNN. In *Proc. of IEEE International Conference on Computer Vision (ICCV)*. 2961–2969.
[28] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 770–778.
[29] Christian Herrmann, Miriam Ruf, and Jürgen Beyerer. 2018. CNN-based thermal infrared person detection by domain adaptation. In *Proc. of Autonomous Systems: Sensors, Vehicles, Security, and the Internet of Everything*.
[30] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei A. Efros, and Trevor Darrell. 2018. CyCADA: Cycle-Consistent Adversarial Domain Adaptation. In *Proc. of International Conference on Machine Learning (ICML)*.
[31] Weixiang Hong, Zhenzhen Wang, Ming Yang, and Junsong Yuan. 2018. Conditional generative adversarial network for structured domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1335–1344.
[32] Sheng-Wei Huang, Che-Tsung Lin, Shu-Ping Chen, Yen-Yi Wu, Po-Hao Hsu, and Shang-Hong Lai. 2018. Auggan: Cross domain adaptation with gan-based data augmentation. In *Proc. of European Conference on Computer Vision (ECCV)*. 718–731.
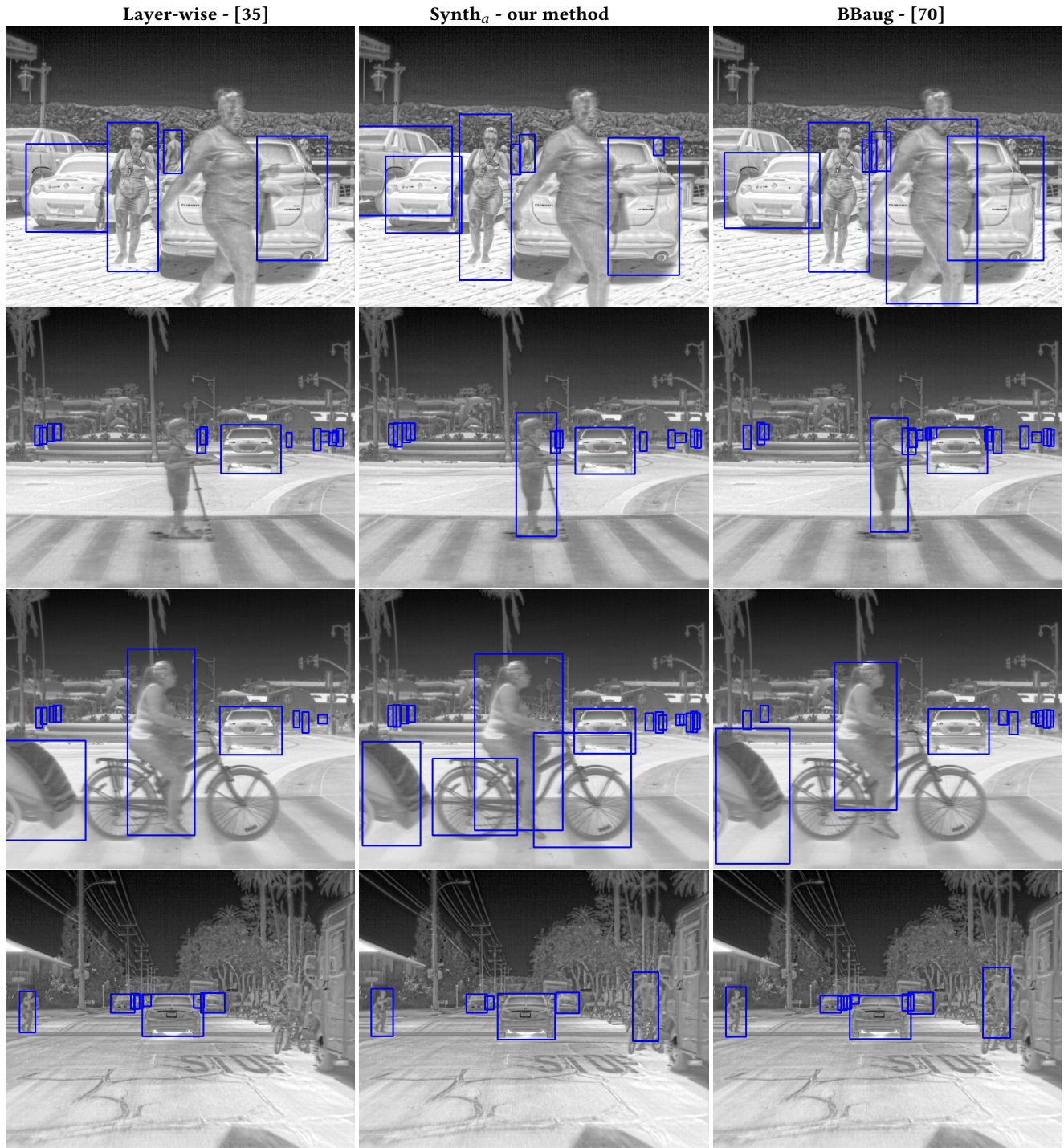
Layer-wise - [35] Synth$_a$ - our method BBaug - [70]

Figure 8: Examples of detections: *left)* **Layer-wise adaptation [35],** *middle)* **out method,** *right)* **BBaug augmentation. In general, augmentations help to detect more persons, even in very critical cases like the kid that is crossing the street on row 2. Our propose method helps the network to better distinguish bicycle from persons, as shown on row 3.**

[33] Soonmin Hwang, Jaesik Park, Namil Kim, Yukyung Choi, and Inso Kweon. 2015. Multispectral Pedestrian Detection: Benchmark Dataset and Baseline. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

[34] Vijay John, Seiichi Mita, Zheng Liu, and Bin Qi. 2015. Pedestrian detection in thermal images using adaptive fuzzy C-means clustering and convolutional neural networks. In *Proc. of IAPR International Conference on Machine Vision Applications (MVA).* 246–249.

[35] My Kieu, Andrew D Bagdanov, and Marco Bertini. 2020. Bottom-up and Layer-wise Domain Adaptation for Pedestrian Detection in Thermal Images. *ACM Transactions on Multimedia Computing Communications and Applications (ACM TOMM)* (2020).

[36] My Kieu, Andrew D Bagdanov, Marco Bertini, and Alberto Del Bimbo. 2019. Domain Adaptation for Privacy-Preserving Pedestrian Detection in Thermal Imagery. In *Proc. of International Conference on Image Analysis and Processing (ICIAP)*.

[37] My Kieu, Andrew D. Bagdanov, Marco Bertini, and Alberto Del Bimbo. 2020. Task-conditioned Domain Adaptation for Pedestrian Detection in Thermal Imagery. In *Proc. of European Conference on Computer Vision (ECCV)*. 546–562. https://doi.org/10.1007/978-3-030-58542-6_33

[38] M. Kieu, L. Berlincioni, L. Galteri, M. Bertini, A. D. Bagdanov, and A. del Bimbo. 2021. Robust pedestrian detection in thermal imagery using synthesized images. In *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE Computer Society, 8804–8811.

[39] Daniel Konig, Michael Adam, Christian Jarvers, Georg Layher, Heiko Neumann, and Michael Teutsch. 2017. Fully convolutional region proposal networks for multispectral person detection. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPR-W)*. https://doi.org/10.1109/CVPRW.2017.36

[40] Kuan-Ting Lai, Chia-Chih Lin, Chun-Yao Kang, Mei-Enn Liao, and Ming-Syan Chen. 2018. VIVID: Virtual Environment for Visual Deep Learning. In *Proc. of ACM Multimedia* (Seoul, Republic of Korea) (*MM '18*). Association for Computing Machinery, New York, NY, USA, 1356–1359. https://doi.org/10.1145/3240508.3243653

[41] Yongwoo Lee, Toan Duc Bui, and Jitae Shin. 2018. Pedestrian Detection based on Deep Fusion Network using Feature Correlation. In *Proc. of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*.

[42] Chengyang Li, Dan Song, Ruofeng Tong, and Min Tang. 2018. Multispectral pedestrian detection via simultaneous detection and segmentation. In *Proc. of British Machine Vision Conference (BMVC)*.

[43] Chengyang Li, Dan Song, Ruofeng Tong, and Min Tang. 2019. Illumination-aware Faster R-CNN for robust multispectral pedestrian detection. *Pattern Recognition* 85 (2019), 161–171.

[44] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. 2017. Enhanced Deep Residual Networks for Single Image Super-Resolution. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPR-W)*.

[45] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proc. of IEEE International Conference on Computer Vision (ICCV)*. 2980–2988.

[46] Jingjing Liu, Shaoting Zhang, Shu Wang, and Dimitris N Metaxas. 2016. Multispectral deep neural networks for pedestrian detection. *arXiv preprint arXiv:1611.02644* (2016).

[47] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I. Jordan. 2017. Deep Transfer Learning with Joint Adaptation Networks. In *Proceedings of the 34th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 70)*, Doina Precup and Yee Whye Teh (Eds.). PMLR, 2208–2217. http://proceedings.mlr.press/v70/long17a.html

[48] Dennis Ludl, Thomas Gulde, and Cristóbal Curio. 2020. Enhancing Data-Driven Algorithms for Human Pose Estimation and Action Recognition Through Simulation. *IEEE Transactions on Intelligent Transportation Systems* 21, 9 (2020), 3990–3999.

[49] Mark Martinez, Chawin Sitawarin, Kevin Finch, Lennart Meincke, Alex Yablonski, and Alain Kornhauser. 2017. Beyond grand theft auto V for training, testing and enhancing deep learning in self driving cars. *arXiv preprint arXiv:1712.01397* (2017).

[50] Sangwoo Mo, Minsu Cho, and Jinwoo Shin. 2019. InstaGAN: Instance-aware Image-to-Image Translation. arXiv:1812.10889 [cs.LG]

[51] Farzeen Munir, Shoaib Azam, and Moongu Jeon. 2021. SSTN: Self-Supervised Domain Adaptation Thermal Object Detection for Autonomous Driving. *arXiv preprint arXiv:2103.03150* (2021).

[52] Farzeen Munir, Shoaib Azam, Muhammad Aasim Rafique, Ahmad Muqeem Sheri, and Moongu Jeon. 2020. Thermal Object Detection using Domain Adaptation through Style Consistency. *arXiv preprint arXiv:2006.00821* (2020).

[53] Guanghan Ning, Guang Chen, Chaowei Tan, Si Luo, Liefeng Bo, and Heng Huang. 2021. Data Augmentation for Object Detection via Differentiable Neural Rendering. *arXiv preprint arXiv:2103.02852* (2021).

[54] Markus Oberweger, Gernot Riegler, Paul Wohlhart, and Vincent Lepetit. 2016. Efficiently creating 3d training data for fine hand pose estimation. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 4957–4965.

[55] Xinlei Pan, Yurong You, Ziyan Wang, and Cewu Lu. 2017. Virtual to real reinforcement learning for autonomous driving. *arXiv preprint arXiv:1704.03952* (2017).

[56] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. 2016. Playing for data: Ground truth from computer games. In *Proc. of European Conference on Computer Vision (ECCV)*. Springer, 102–118.

[57] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. M. Lopez. 2016. The SYNTHIA Dataset: A Large Collection of Synthetic Images for Semantic Segmentation of Urban Scenes. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 3234–3243. https://doi.org/10.1109/CVPR.2016.352

[58] Swami Sankaranarayanan, Yogesh Balaji, Arpit Jain, Ser Nam Lim, and Rama Chellappa. 2018. Learning from synthetic data: Addressing domain shift for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3752–3761.

[59] Ashish Shrivastava, Tomas Pfister, Oncel Tuzel, Joshua Susskind, Wenda Wang, and Russell Webb. 2017. Learning from simulated and unsupervised images through adversarial training. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2107–2116.

[60] "Unity Technologies". [n.d.]. Unity3D,Unity Development Platform. Web page. https://unity.com/ Accessed: April 14, 2021.

[61] Yonglong Tian, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Pedestrian detection aided by deep learning semantic tasks. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[62] Maarten Vandersteegen, Kristof Van Beeck, and Toon Goedemé. 2018. Real-time multispectral pedestrian detection with a single-pass deep neural network. In *Proc. of International Conference Image Analysis and Recognition (ICIAR)*.

[63] Gul Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J. Black, Ivan Laptev, and Cordelia Schmid. 2017. Learning From Synthetic Humans. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[64] Jörg Wagner, Volker Fischer, Michael Herman, and Sven Behnke. 2016. Multispectral Pedestrian Detection using Deep Fusion Convolutional Neural Networks.. In *Proc. of European Symposium on Artificial Neural Networks (ESANN)*.

[65] Dan Xu, Wanli Ouyang, Elisa Ricci, Xiaogang Wang, and Nicu Sebe. 2017. Learning cross-modal deep representations for robust pedestrian detection. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[66] Heng Zhang, Elisa Fromont, Sébastien Lefèvre, and Bruno Avignon. 2020. Multispectral Fusion for Object Detection with Cyclic Fuse-and-Refine Blocks. In *Proc. of IEEE International Conference on Image Processing (ICIP)*. IEEE, 276–280.

[67] Lu Zhang, Zhiyong Liu, Xiangyu Chen, and Xu Yang. 2019. The Cross-Modality Disparity Problem in Multispectral Pedestrian Detection. *arXiv preprint arXiv:1901.02645* (2019).

[68] Yang Zheng, Izzat H Izzat, and Shahrzad Ziaee. 2019. GFD-SSD: Gated Fusion Double SSD for Multispectral Pedestrian Detection. *arXiv preprint arXiv:1903.06999* (2019).

[69] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. In *Proc. of IEEE International Conference on Computer Vision (ICCV)*.

[70] Barret Zoph, Ekin D Cubuk, Golnaz Ghiasi, Tsung-Yi Lin, Jonathon Shlens, and Quoc V Le. 2020. Learning data augmentation strategies for object detection. In *Proc. of European Conference on Computer Vision (ECCV)*. Springer, 566–583.