2020 25th International Conference on Pattern Recognition (ICPR)
Milan, Italy, Jan 10-15, 2021

Robust pedestrian detection in thermal imagery using synthesized images

My Kieu*, Lorenzo Berlincioni*, Leonardo Galteri, Marco Bertini, Andrew D. Bagdanov, Alberto Del Bimbo
MICC - Università degli Studi di Firenze
name.surname@unifi.it

***Abstract*—In this paper we propose a method for improving pedestrian detection in the thermal domain using two stages: first, a generative data augmentation approach is used, then a domain adaptation method using generated data adapts an RGB pedestrian detector. Our model, based on the Least-Squares Generative Adversarial Network, is trained to synthesize realistic thermal versions of input RGB images which are then used to augment the limited amount of labeled thermal pedestrian images available for training. We apply our generative data augmentation strategy in order to adapt a pretrained YOLOv3 pedestrian detector to detection in the thermal-only domain. Experimental results demonstrate the effectiveness of our approach: using less than 50% of available real thermal training data, and relying on synthesized data generated by our model in the domain adaptation phase, our detector achieves state-of-the-art results on the KAIST Multispectral Pedestrian Detection Benchmark; even if more real thermal data is available adding GAN generated images to the training data results in improved performance, thus showing that these images act as an effective form of data augmentation. To the best of our knowledge, our detector achieves the best single-modality detection results on KAIST with respect to the state-of-the-art.**

I. Introduction

Pedestrian detection is a core problem in computer vision due to its central role in a broad gamut of practical applications. Application areas such as video surveillance and autonomous driving further require pedestrian detection be 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 [1], [2].

Detectors based on thermal imagery have garnered attention recently as a means to mitigate the sensitivity of visible spectrum imagery to scene-incidental imaging conditions [2], [3], [4]. A growing number of works have also investigated multispectral detectors combining visible and thermal images for robust pedestrian detection [5], [6], [7], [8], [9], [1], [10]. Due to the cost of deploying multiple aligned sensors, multispectral models can have limited applicability in real-world applications. Moreover, and especially important given the recent focus on privacy by the public and national legislative bodies, using visible spectrum sensors does not offer the same privacy-preserving affordances as systems employing only thermal sensors [2].

Thermal-only detectors typically yield lower performance than multispectral detectors since robust pedestrian detection using only thermal data is extremely challenging. 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; for instance the Caltech Pedestrian Dataset [11] has 350,000 annotations in the visible domain, while KAIST Multispectral Pedestrian dataset [12] has $\sim 51,000$ annotations and FLIR ADAS Dataset [13] has $\sim 28,000$. Scaling thermal-only detection to the levels of robustness and accuracy demanded by real-world applications is thus extremely difficult due to this poverty of annotated data.

In this paper we propose to use a generative algorithm to perform data augmentation that can enrich thermal pedestrian datasets for training deep detector architectures. Our approach is based on a Least-Squares Generative Adversarial Network (LSGAN) [14] trained to synthesize thermal pedestrian images from RGB inputs. We investigate the best approaches to exploit these generated images during training, i.e. studying how to mix real thermal images with synthesized ones in order to effectively augment the training set. Experimental results indicate that our trained LSGAN is able to learn to translate RGB pedestrian images to useful thermal versions so that even using $\sim 50\%$ synthetic images results in state-of-the-art pedestrian detection at nighttime and overall day/nighttime. This suggests that the approach can be extended to other domains in which thermal training data is scarce but is possible to effectively exploit the abundance of RGB imagery to adapt it to the thermal domain.

The contributions of this work are:

- we propose a novel generative model based on the Least-Squares Generative Adversarial Network (LSGAN) [14] that is able to synthesize thermal imagery from RGB;
- we propose a mixed real/synthetic training domain adaptation procedure that mixes real thermal imagery with thermal images synthesized from unlabeled RGB pedestrian images using our LSGAN and uses this augmented training set to adapt the YOLOv3 [15] detector;
- we conduct extensive ablation study to probe the effectiveness of our approach and a variety of mixing proportions of real and synthesized imagery; and
- we conduct an extensive set of experiments comparing our approach to the state-of-the-art, and to the best of our knowledge our thermal-only detector outperforms all state-of-the-art single-modality detection approaches on

*Both authors contributed equally to this research.

978-1-7281-8808-9/20/$31.00 ©2020 IEEE

2020 25th International Conference on Pattern Recognition (ICPR) | 978-1-7281-8808-9/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICPR48806.2021.9412764

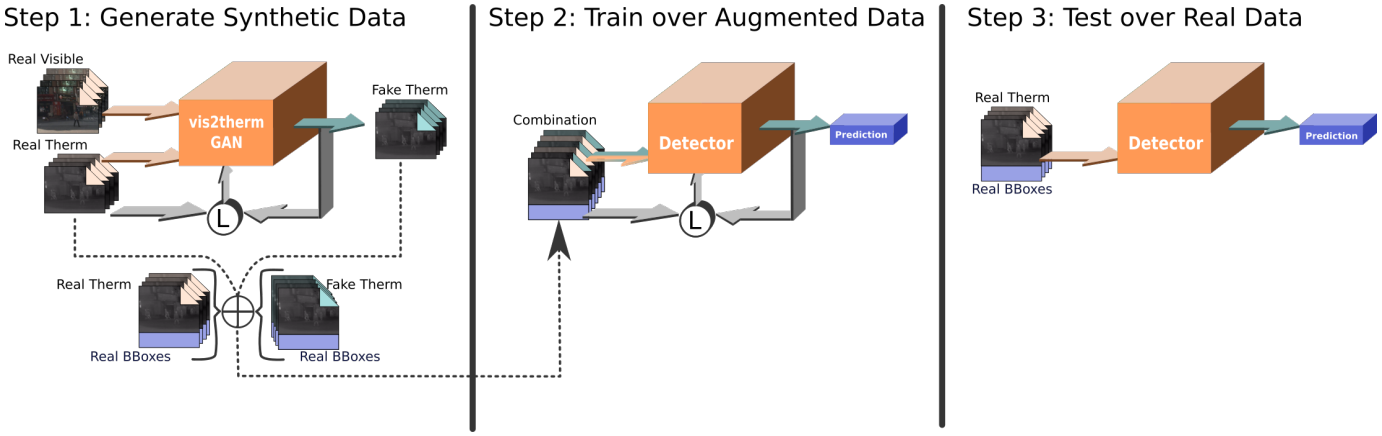


Fig. 1: System overview: the vis2therm GAN generates fake thermal images from visible data; a mixture of real and fake thermal images along with related bounding boxes of objects are used to train an object detector, that is then tested on images from thermal cameras.

the KAIST Multispectral Pedestrian Detection Benchmark [12] by a large margin.

The rest of the paper is organized as follows. In the next section we review the scientific literature related to our proposed approach. In section III we describe our generative model used to synthesize thermal images and our training procedure used to adapt a YOLOv3 pedestrian detector to the thermal domain. We report in section IV on an extensive set of experiments performed to evaluate the effectiveness of thermal pedestrian detection using our approach, and in Section V we conclude with a discussion of our contribution.

II. Related work

The problem of pedestrian detection in thermal imagery has attracted much attention from the research community over the years due to the advantages of thermal cameras in many real-world and critical applications.

A. Pedestrian detection in thermal imagery

Thanks to the reduction of costs and availability of multi-spectral cameras over the past few years, there are numerous recent works exploiting thermal images in combination with visible images for robust pedestrian detections [7], [16], [8], [17], [1], [10], [18], [19], [20], [21], [22], [23]. In contrast, many recent works have investigated pedestrian detection using thermal (IR) imagery only. For example, authors in [24] 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 [25].

The idea of performing several video preprocessing steps to make thermal images look more similar to grayscale images converted from RGB was investigated in [3], who then applied a pretrained and fine-tuned SSD detector. Recently, authors in [26] 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 [2] for pedestrian detection in thermal imagery. We also focus on the thermal-only detection problem. However, our approach is distinct in that we concentrate on domain adaptation via data augmentation during training using synthetic thermal data which is generated by a generative model trained on unlabeled data.

B. Spectrum transfer between visible and thermal

The generation of RGB images from the thermal images has been approached as a grayscale colorization task in several previous works such as [27] where deep multiscale CNNs are used along with classical computer vision post processing techniques over near infrared images. In [28] a CNN is used with a more sophisticated objective function in order to tackle misalignment issues between the two visible and thermal modalities. In [29] instead an encoder-decoder architecture is applied for performing colorization.

Most recent works, however, rely heavily on generative models to perform image-to-image translation between visible and thermal. As defined in [30], the *image-to-image translation* problem is the task of translating one visual representation of a scene into another. Many *domain to domain* translation problems [31], from image denoising [32] to image super-resolution [33], can be cast as image-to-image translation tasks.

Generative Adversarial Networks (GANs), introduced in [34], are one the most significant recent improvements in the field of generative models and have been extensively used for image-to-image translation. The key feature of these models is the competitive min/max game between two networks. GANs have been successfully applied in many computer vision tasks

such as super resolution [35], [36], [37], style transfer[38], image inpainting [39] and domain adaptation[40].

Both [41], [42] use GANs architectures to perform infrared and grayscale colorization. In [41] a DCGAN with one seperate generator per channel is used, while in [42] an improved [38] GAN is proposed. In [43] the authors focused on learning an identity-preserving translation between thermal and visible images of faces. The authors in [44] leverage multiple streams of polarimetric images to synthesize photo-realistic visible images of faces preserving discriminative features. In [45] a multi-image to image generative framework is presented, and one of the proposed settings is infrared and grayscale colorization. Also in [46] the authors used a Cycle-GAN[38] 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 seperate classifier has been studied in multiple previous works such as [47] in which they focus on improving one-shot learning, in [48] where segmentation of medical images is enhanced by GAN augmented data.

In this work we focus on the opposite task: mapping RGB images to the infrared spectrum. The closest related works are [49], [50], [51], [52], as they all employ generative models to translate images from the visible to the thermal spectrum. A modified Cycle-GAN [38] is used in [49], where the performance of drone detection in the thermal spectrum is improved using augmented data coming from a visible to thermal GAN framework, and also in [51], where a pedestrian detector is trained on augmented thermal data. Also in [49] a modified version is proposed which changing the loss with a perceptual texture loss term. In [50], both pix2pix [30] and Cycle-GAN are used to generate thermal images to train an object tracker in the thermal domain; experiments show that images generated with pix2pix are of higher quality, since this approach operates on paired thermal/RGB data.

The authors of [52] present a framework for cross-modality color to thermal person re-identification. The generative model in this work is tasked with the generation of multiple thermal versions of the visible input image, which is then used to match with real thermal gallery set. Here the proposed architecture is a variation of [53], a multimodal image-to-image translation framework composed of multiple networks: cVAE-GAN from [54] and cLR-GAN from [55] which are jointly optimized in a hybrid model in order to cover complementary tasks. One of the major contribution of [53] is the ability to model the distribution of different correct outputs corresponding to the same input.

In our approach we instead rely on a different architecture that combines elements from [14] and [37], as further detailed in Section III-B. The proposed architecture in [37], ESRGAN, focuses on the *super-resolution* problem and improved over the previous state-of-the-art [56] by introducing the Residual-in-Residual Dense Block, removing the Batch-Normalization layers, and changing the perceptual loss term.

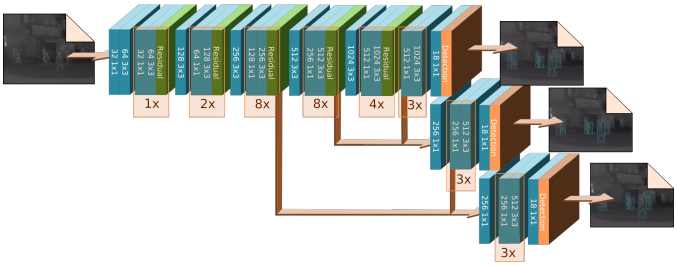


Fig. 2: The YOLOv3 architecture. $k\times$ indicates the repetition of blocks k times.

III. Generative Data Augmentation for Thermal Domain Adaptation

In this section we describe the two main components of our proposed approach. Our thermal pedestrian detector based on YOLOv3 [57] is described in the next section, and our generative model which produces fake thermal images from available RGB images is described in section III-B. An extensive series of experimental results are reported on in section IV-C.

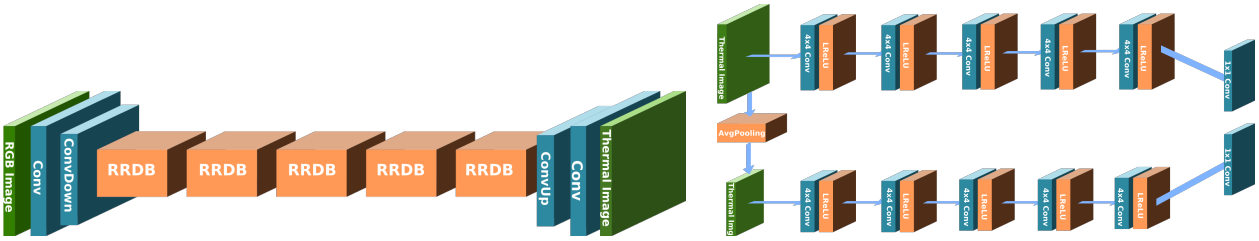
A. Object detection in thermal images

We use YOLOv3 as our base pedestrian detector [57]. Following the Domain Adaptation approach described in [2], we first adapt YOLOv3 in the visible domain by directly fine-tuning it on the visible spectrum images from the KAIST dataset [12]. Then, we use this detector as a starting point for training a thermal detector using a range of mixtures of real and GAN-generated thermal images. Figure 2 illustrates the original YOLOv3 architecture with thermal image as input and the output of the model at three detection scales.

We consider the following training regimes for thermal detectors:

- **Real-Thermal detector**: We directly fine-tune the detector on all available *real thermal images*.
- **Synthesized-Thermal detector**: We directly fine-tune the detector on all the *GAN-generated thermal images (synthesized images)*.
- **Combined-Thermal detector**: We combine all available real images and all the synthesized images into a combined training set and then we fine-tune the detector on it. Note that the number of images in this combined set is double that used for the Real-Thermal and Synthesized-Thermal detectors.
- **Mixed-Thermal detectors**: We mix real images and synthesized images with a proportion varying from 10% to 90%; in total we have 9 mixed sets of images. For example, the mixed set 1 has 10% real images and 90% synthesized images. Note that the number of images used to train these detectors is the same as those used for Real-Thermal and Synthesized-Thermal detectors.

For all experiments we evaluate performance on the KAIST test set of real thermal images.



(a) Generator architecture (b) Discriminator architecture

Fig. 3: Our LSGAN architecture. The prosposed generator (a), composed of multiple Residual in Residual Dense Blocks, and the discriminator structure (b), a multiscale CNN.

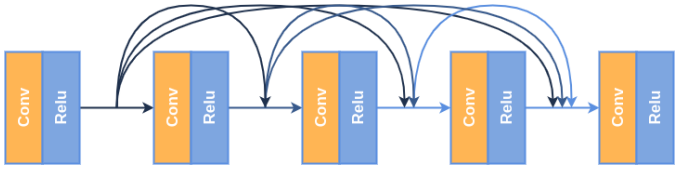


Fig. 4: Dense Blocks. Arcs represent the concatenation between the output of a layer and one of its subsequent layers.

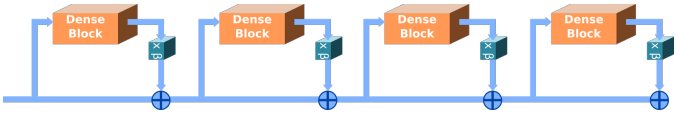


Fig. 5: Residual in Residual Dense Block. The output of Dense Blocks are scaled by β and summed back to their input.

B. Visible to thermal GAN

Our model is an LSGAN trained with both Adversarial and Perceptual losses. The Least Squares GAN (LSGAN) [14], [58] improves on the standard GAN model by changing the loss function from a cross-entropy to a squared distance. It is comparatively more stable and easier to train. The Generator G architecture is built using the Residual in Residual Dense Block (RRDB) as the fundamental unit (see Figure 5). As in [59], we remove the batch normalization layer from the traditional *Conv-BN-LReLU* triplet. After the initial down-sampling convolutions five RDDB blocks are stacked in sequence as shown in Figure 3(a). Each RDDB block is composed of 4 Dense Blocks. Each Dense Block has a growth rate of $k = 32$ and contains five consecutive pairs of convolutional layers followed by a leaky rectified linear unit (LReLU) whose outputs are concatenated as shown in Figure 4.

Dense Blocks. DenseNets, introduced in [60], improve the information flow between layers by adding direct connections between a layer and all subsequent layers. By using this connectivity pattern the l^{th} layer receives the feature maps coming from all the preceding $l-1$ layers as shown in Fig. 4. This dense connection strategy is realized by feeding as input the concatenation of every preceding layer output. DenseNets provide advantages both from a memory consumption and a vanishing gradient standpoint.

Residual in Residual Block. The composition of Residual Networks [61] and DenseNets is the Residual in Residual Dense block (RRDB), as introduced in [37]. A single RRDB is composed of multiple Dense blocks connected in a residual fashion, and is shown in Fig. 5. Finally, the output of the RRDB chain is followed by multiple *upscale-Conv-ReLU* blocks to scale the image back to input size.

Inspired by [62], [63], [64] successful application of multi scale architectures we use a multi-scale discriminator D, shown in Figure 3(b), that makes no use of dense connectivity patterns. It is composed of five convolutional layers, each of them using a 4×4 convolutional kernel with stride 2 and followed by LReLU activation function. The number of feature maps is doubled as depth increases starting from 64. For each of the multiple scales, a single 1×1 convolutional filter is used as final output layer. Finally, the different outputs of every scale is evaluated independently.

Training. We trained the model as a Least Squares Generative Adversarial Network (LSGAN) with a perceptual loss. The discriminator D is trained as a standard LSGAN Discriminator:

$$\begin{aligned} L_{D_{LSGAN}} &= \frac{1}{2}\mathbb{E}_{x\sim p_{data}(x)}[(D(x) - real_{label})^2] \\ &+ \frac{1}{2}\mathbb{E}_{z\sim p(z)}[(D(G(z)) - fake_{label})^2]. \end{aligned}$$

The generator loss is composed of three terms:

$$L_{G_{Adv}} = \frac{1}{2}\mathbb{E}_{z\sim p(z)}[(D(G(z)) - fake_{label})^2] \quad (1)$$

$$L_{G_{MAE}} = |real_{img} - fake_{img}| \quad (2)$$

$$L_{G_{Perceptual}} = (\phi^k(real_{img}) - \phi^k(fake_{img}))^2, \quad (3)$$

which are summed together:

$$L_{G_{LSGAN}} = L_{G_{Adv}} + L_{G_{MAE}} + L_{G_{Perceptual}} \quad (4)$$

Perceptual loss. Perceptual loss functions [65] aim to provide a better measure for similarity compared to metrics such as the PSNR (Peak Signal to Noise Ratio) and SSIM (Structural Similarity Index). They have been shown useful for super-resolution and style-transfer tasks. Our perceptual loss architecture consists of two networks:

- Transformation Network T
- Loss Network ϕ

The Loss Network ϕ is pretrained, usually as a classifier. When training the transformation network T, the loss network ϕ is used as a feature extractor by taking the output of some of its layers. The distance between the target and the generated image in this feature space is used as a loss function for the Transformation Network T. The main motivation behind perceptual loss functions lies on the intuition that computing distances in the high dimensional manifold extracted from a well-trained classifier should result in a better estimate compared to any pixel-space distance measure.

As shown in [66] pixel space metrics can lead to minima that corresponds to blurry results. In this work, since our goal is to detect pedestrians, we use the YOLO detector to drive the generation of the images. The term (3) is a *perceptual loss* defined as the squared distance between the outputs ϕ^k of the k^{th} layer of a pretrained YOLOv3 network for a real and a generated input. We trained the ϕ network on KAIST for a detection task in a thermal setting. We choose the last convolutional layer of YOLOv3 as representation of the input image in the high dimensional space learned by the classifier. Note that the loss network ϕ at this stage acts as a feature extractor and its weights are frozen.

IV. Experimental results

In this section we report on a range of experiments conducted to evaluate the effectiveness of our approach to thermal domain adaptation for pedestrian detection. We first describe the dataset and evaluation metrics used, then in Section IV-B give a qualitative evaluation of the performance of our GAN in generating thermal imagery from RGB input. In Section IV-C we perform an ablative analysis of the use of synthetically generated thermal imagery for data augmentation, and in Section IV-D give a comparison with the state-of-the-art.

A. Dataset and experimental protocol

Dataset. All of our experiments were conducted on the KAIST Multispectral Pedestrian Benchmark dataset [12]. KAIST is a large-scale dataset with well-aligned visible/thermal pairs [46], and it contains videos captured both during the day and at night. KAIST dataset consists of 95,328 image pairs split into 50,172 for training and 45,156 for testing. We follow the standard sampling procedure in [12], [1], [16], we sample every two frames from training videos and exclude heavily occluded and small person instances (< 50 pixels). The final training set contains 7,601 images. The test set contains 2,252 image pairs sampled every 20 frames. For training and testing, we use the improved training annotations from [1] and test annotations from [16].

Performance metrics. As is common practice to compare with the state-of-the-art, we used standard evaluation metrics for object detection, namely miss rate as a function of False Positives Per Image (FPPI), and log-average miss rate for thresholds in the range of $[10^{-2}, 10^{0}]$ with an Intersection over Union (IoU) threshold of 0.5 under the *reasonable* setting [11], [12], [1], [16], [2]. The *reasonable* setting is composed of *day-time*, *night-time*, and *all (both day and night time)* sets of images. Figure 7 shows some example images with our detection results on KAIST dataset.

TABLE I: Ablation study on varying quantities of GAN-generated images. Results are on KAIST in terms of log-average miss rate (lower is better). Best results highlighted in **underlined bold**, second best in **bold**.

	Mixture		**Miss Rate (%)**		
	Real (%)	Synthetic (%)	all	day	night
Synthesized	0	100	45.88	54.37	26.04
Mixed	10	90	44.90	54.24	22.79
	20	80	41.21	51.04	18.92
	30	70	35.32	44.44	16.35
	40	60	34.78	43.45	14.53
	50	50	33.90	41.97	14.64
	60	40	31.50	39.83	12.33
	70	30	32.29	41.68	12.42
	80	20	**25.88**	**33.01**	**11.12**
	90	10	**25.62**	**31.86**	12.92
Real	100	0	28.46	36.32	**11.97**
Combined	all	all	34.29	41.93	16.80

Fine-tuning. All of our detectors were implemented using PyTorch. During fine-tuning to adapt to the thermal domain, at each epoch we set aside 10% of the training images for validation for that epoch. We trained every detector using Stochastic Gradient Descent with the same procedure and hyperparameters: image size 640×512, batch size of 4, We set an initial learning rate of 0.001 if the training set contains 50% or more real images, otherwise we use a learning rate of 0.0001. During fine-tuning, we reduce the learning rate by a factor of 10 every 3 epochs, and training is halted after 10 epochs.

B. GAN results

The GAN framework for the *visible to thermal* transformation was trained on pairs of RGB-LWIR frames from the original training split of the KAIST dataset. In Figure 6 we show some examples detections using the detector trained with 20% sythesized images and 80% real images on two kinds of images. The first row shows detection results on generated images without Perceptual Loss $L_{G_{Perceptual}}$, and the second row gives detection results on generated images by our model trained with $L_{G_{Perceptual}}$. The use of the $L_{G_{Perceptual}}$ seems to result in more true positive (blue boxes) detection results, as well fewer false negative (green boxes).

C. Ablation study

In this section, we report on a series of experiments we conducted to explore the many options available when using GAN generated images (synthesized images) and thermal images (real images) for training the detectors described in Section III-A. Initial experiments with simple augmentation strategies resulted in worse results than the conventional fine-tuning model. Thus, we use the conventional fine-tuning result as a baseline for comparison with various mixing strategies of GAN-generated thermal images. In table I we present results of an ablation study considering all these possibilities. From these results we first note that mixing in a *small* proportion

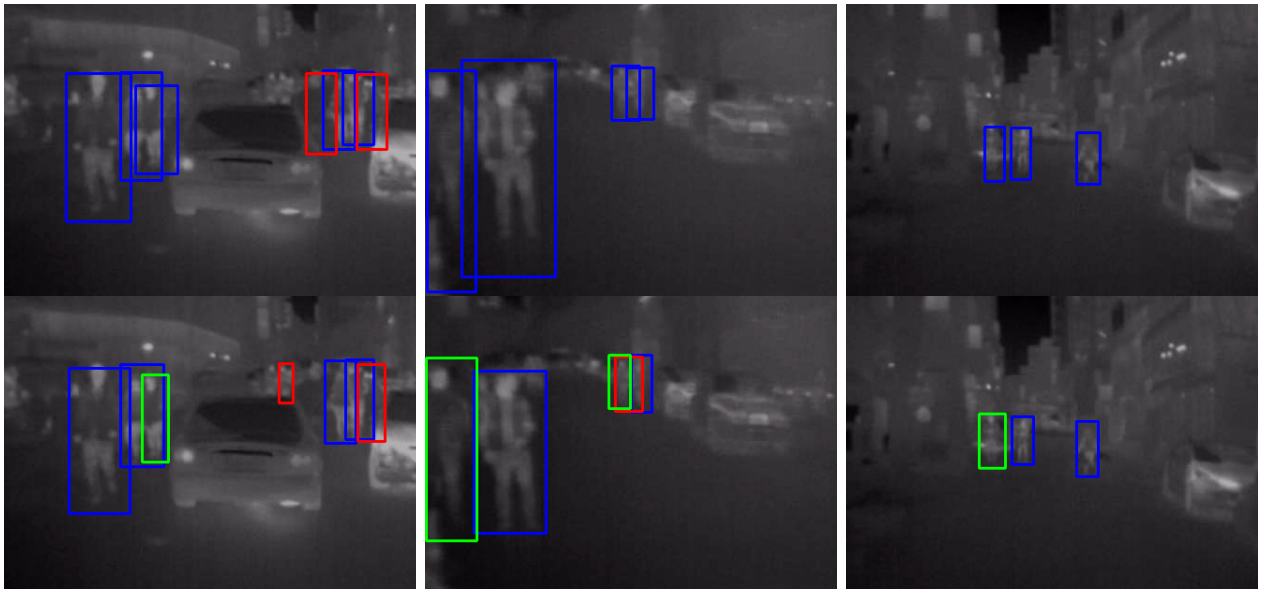


Fig. 6: Example detections using the detector trained with 80% real images and 20% synthesized images. The first row shows detection results with the perceptual loss, while the second row is *without* perceptual loss. Blue boxes are true positive detections, green boxes are false negatives, and red boxes indicate false positives

of synthesized images (**Mixed**) rather than training on a all available real and synthesized images (**Combined**) is generally useful. In fact, the best mixture proportion is 90% real images with 10% percent synthesized images with 25.62% miss rate the "all" setting, and the second best is the **Mixed** of 80% and 20% with 11.12% miss rate in nighttime – an improvement of 5.68% over the **Combined** using all available data. Note that even with fewer than 50% real images our detector achieves results are comparable with state-of-the-art methods. Moreover, observe that mixing more than 50% real images results in improvement over the detector that combining all available real and synthesized images. The result reveals that the small portion of GAN synthesized images is useful for augmentation approach, but it must be consider based on the testing data such as the real test set was conducted on the test phase, thus the **Mixed** and **Real** results are better a little than the **Combined** result.

D. Comparison with the state-of-the-art

Table II compares our results with the state-of-the-art single modality approaches which are mostly trained and tested only on thermal images of KAIST dataset (except the KAIST baseline [12] that is a multispectral method), some other models also used visible images for transfer learning such as [2]. We leveraged unlabeled RGB images of train set for generating synthetic thermal images, then we used this thermal data as augmentation for training; of course, testing was conducted on real thermal images of the test set. Results are compared in terms of log average miss rate (lower score is better). We can see that our approaches obtained the best results with 25.62% of missrate at "all" and 11.12% of missrate at "nighttime" – an improvement of 9.38% over the second state-of-the-art

TABLE II: Comparison with state-of-the-art single-modality approaches on KAIST Thermal in term of log-average miss rate (lower is better). Best results highlighted in **underlined bold**, second best in **bold**.

Detectors		MR all	MR day	MR night
KAIST baseline	[12]	64.76	64.17	63.99
FasterRCNN	[16]	47.59	50.13	40.93
TPIHOG	[4]	-	-	57.38
SSD300	[3]	69.81	-	-
Saliency + KAIST	[25]	-	39.40	40.50
R^3-Net Saliency + KAIST	[25]	-	**30.40**	21.00
VGG16-two-stage	[51]	46.30	53.37	31.63
ResNet101-two-stage	[51]	42.65	49.59	26.70
Bottom-up	[2]	35.20	40.00	20.50
Ours Mixed 40_60		34.78	43.45	14.53
Ours Mixed 80_20		**25.88**	33.01	**11.12**
Ours Mixed 90_10		**25.62**	**31.86**	**12.92**

results. Moreover, our results outperform all existing the state-of-the-art methods by a large margin in both "night-time" and "all". The results of R^3-Net Saliency [25] are a little better than ours in day time due to the advantages of their proposed pixel-level "saliency" annotation set with manually annotated 1,702 images from training and 369 from testing set, and their extraction of deep saliency maps by R^3-Net for augmenting thermal images of both training and testing.

Several different backbones have been used by the methods reported in the table, from VGG16 to Faster RCNN. Our backbone is the conventional YOLOv3 detector, and as fine tuning procedure we followed our previous approach of [2]. The improvements that allowed to surpass the second-best state-of-the-art detector on KAIST (bottom-up [2]) are: 1) the new data annotation as described in section IV-A; 2) the domain adaptation method of [2] and the experimentation with

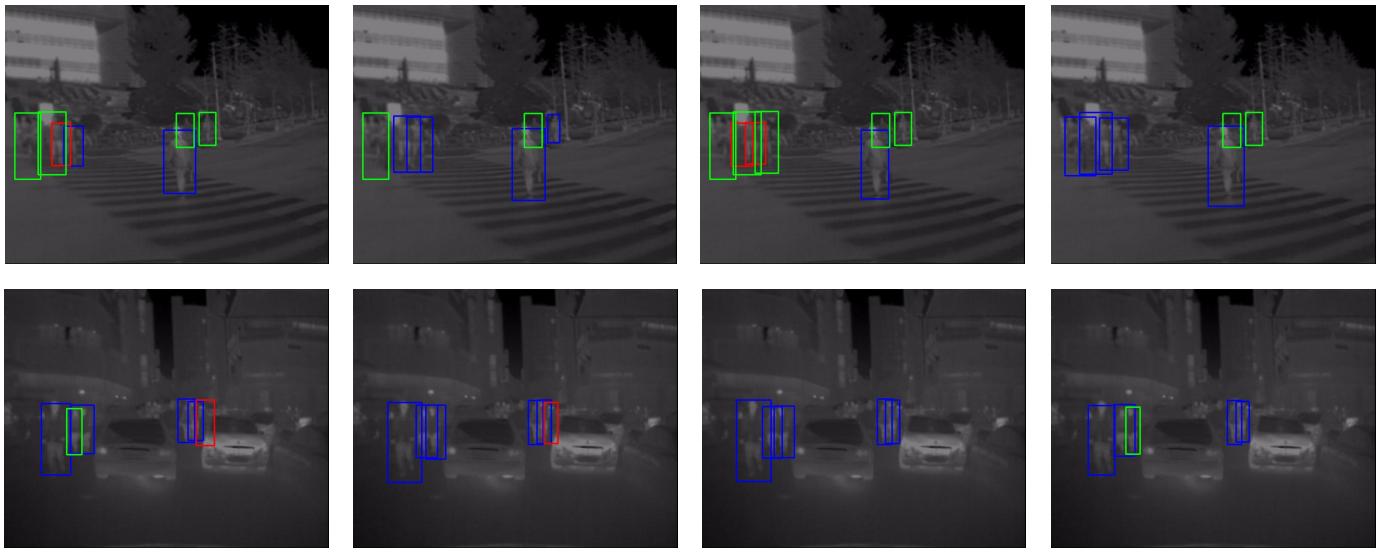


Fig. 7: Examples of KAIST thermal images with detections. The first row is daytime images and the second is nighttime. The first and the second column are detection result on synthetic-only and real-only training, respectively. The third and the last column are combining all and mixed 90% proportion, respectively. Blue boxes are true positive detections, green boxes are false negatives, and red boxes indicate false positives. See section IV-D for detailed analysis.

hyperparameter setting reported in section IV-A. Moreover, with the proposed generated synthesized thermal images with LSGAN and the mixed training procedure, we achieve state-of-the-art performance for both all (day and night) and nighttime.

It is expected that detection in thermal images at nighttime will always be better than daytime results because of the low contrast between pedestrians and background during the day, as noted in [25].

In Figure 7 we show some example detections from four detectors (synthetics, real, combination and mixed90). From these examples we see that the mixed of 90% real images with 10% synthesized images yields more true positive and fewer false positive detections with respect to others. Not surprisingly, **synthesized detector** (the first column) produces a higher number of false positives and missed detections than **real detector** (the second column). The difference is even more pronounced at nighttime (second row of figure 7). The mixed scale 90% real with 10% synthesized images for training (the last columns) makes more true positive and less false positive than the **real detector**.

V. Conclusions

In this paper we proposed a novel GAN architecture, based on LSGAN, to transform visible spectrum images in thermal spectrum ones. We also proposed a novel training procedure that mixes real and synthesized images to adapt the YOLOv3 detector for detection in the thermal domain. Extensive experimental validation shows that our method outperforms state-of-the-art single-modality detectors for pedestrian detection on the KAIST dataset.

Our experiments show that that even using only 50% of available real thermal images it is possible to obtain results that are comparable with state-of-the-art methods trained using 100% real thermal images. This suggests that images generated with our proposed GAN are beneficial and may help to adapt visible spectrum detectors to operate in thermal spectrum in domains suffering from a lack of training data.

Acknowledgement: This research was partially funded by Leonardo.

References

[1] C. Li, D. Song, R. Tong, and M. Tang, "Multispectral pedestrian detection via simultaneous detection and segmentation," in *Proc. of BMVC*, 2018.

[2] M. Kieu, A. D. Bagdanov, M. Bertini, and A. Del Bimbo, "Domain adaptation for privacy-preserving pedestrian detection in thermal imagery," in *Proc. of ICIAP*, 2019.

[3] C. Herrmann, M. Ruf, and J. Beyerer, "CNN-based thermal infrared person detection by domain adaptation," in *Proc. of Autonomous Systems: Sensors, Vehicles, Security, and the Internet of Everything*, 2018.

[4] J. Baek, S. Hong, J. Kim, and E. Kim, "Efficient pedestrian detection at nighttime using a thermal camera," *Sensors*, vol. 17, no. 8, 2017.

[5] Y. Tian, P. Luo, X. Wang, and X. Tang, "Pedestrian detection aided by deep learning semantic tasks," in *Proc. of CVPR*, 2015.

[6] A. Angelova, A. Krizhevsky, V. Vanhoucke, A. Ogale, and D. Ferguson, "Real-time pedestrian detection with deep network cascades," in *Proc. of BMVC*, 2015.

[7] J. Wagner, V. Fischer, M. Herman, and S. Behnke, "Multispectral pedestrian detection using deep fusion convolutional neural networks." in *Proc. of ESANN*, 2016.

[8] D. Konig, M. Adam, C. Jarvers, G. Layher, H. Neumann, and M. Teutsch, "Fully convolutional region proposal networks for multispectral person detection," in *Proc. of CVPR-W*, 2017.

[9] D. Guan, Y. Cao, J. Yang, Y. Cao, and M. Y. Yang, "Fusion of multispectral data through illumination-aware deep neural networks for pedestrian detection," *Information Fusion*, vol. 50, 2019.

[10] C. Li, D. Song, R. Tong, and M. Tang, "Illumination-aware faster R-CNN for robust multispectral pedestrian detection," *Pattern Recognition*, vol. 85, 2019.

[11] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 34, no. 4, 2011.

[12] S. Hwang, J. Park, N. Kim, Y. Choi, and I. Kweon, "Multispectral pedestrian detection: Benchmark dataset and baseline," in *Proc. of CVPR*, 2015.

[13] (2019). [Online]. Available: https://www.flir.com/oem/adas/adas-dataset-form/

[14] X. Mao, Q. Li, H. Xie, R. Y. K. Lau, and Z. Wang, "Multi-class generative adversarial networks with the L2 loss function," *CoRR*, vol. abs/1611.04076, 2016.

[15] J. Redmon and A. Farhadi, "YOLO9000: better, faster, stronger," in *Proc. of CVPR*, 2017.

[16] J. Liu, S. Zhang, S. Wang, and D. N. Metaxas, "Multispectral deep neural networks for pedestrian detection," *arXiv preprint arXiv:1611.02644*, 2016.

[17] D. Xu, W. Ouyang, E. Ricci, X. Wang, and N. Sebe, "Learning cross-modal deep representations for robust pedestrian detection," in *Proc. of CVPR*, 2017.

[18] K. Fritz, D. König, U. Klauck, and M. Teutsch, "Generalization ability of region proposal networks for multispectral person detection," in *Proc. of Automatic Target Recognition XXIX*, vol. 10988, 2019.

[19] L. Zhang, Z. Liu, X. Chen, and X. Yang, "The cross-modality disparity problem in multispectral pedestrian detection," *arXiv preprint arXiv:1901.02645*, 2019.

[20] Y. Cao, D. Guan, Y. Wu, J. Yang, Y. Cao, and M. Y. Yang, "Box-level segmentation supervised deep neural networks for accurate and real-time multispectral pedestrian detection," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 150, 2019.

[21] M. Vandersteegen, K. Van Beeck, and T. Goedemé, "Real-time multispectral pedestrian detection with a single-pass deep neural network," in *Proc. of ICIAR*, 2018.

[22] Y. Lee, T. D. Bui, and J. Shin, "Pedestrian detection based on deep fusion network using feature correlation," in *Proc. of APSIPA ASC*, 2018.

[23] Y. Zheng, I. H. Izzat, and S. Ziaee, "GFD-SSD: Gated fusion double SSD for multispectral pedestrian detection," *arXiv preprint arXiv:1903.06999*, 2019.

[24] V. John, S. Mita, Z. Liu, and B. Qi, "Pedestrian detection in thermal images using adaptive fuzzy c-means clustering and convolutional neural networks," in *Proc. of IAPR MVA*, 2015.

[25] D. Ghose, S. M. Desai, S. Bhattacharya, D. Chakraborty, M. Fiterau, and T. Rahman, "Pedestrian detection in thermal images using saliency maps," in *Proc. of CVPR-W*, 2019.

[26] Y. Cao, T. Zhou, X. Zhu, and Y. Su, "Every feature counts: An improved one-stage detector in thermal imagery," in *Proc. of ICCC*, 2019.

[27] M. Limmer and H. P. A. Lensch, "Infrared colorization using deep convolutional neural networks," in *Proc. of ICMLA*, 2016.

[28] A. Berg, J. Ahlberg, and M. Felsberg, "Generating visible spectrum images from thermal infrared," in *Proc. of CVPR-W*, 2018.

[29] Z. Dong, S.-i. Kamata, and T. P. Breckon, "Infrared image colorization using a s-shape network," in *Proc. of ICIP*, 2018.

[30] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. of CVPR*, 2017.

[31] Y. Choi, M.-J. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "Stargan: Unified generative adversarial networks for multi-domain image-to-image translation," in *Proc. of CVPR*, 2017.

[32] M. Elad and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," *IEEE Transactions on Image Processing (TIP)*, vol. 15, no. 12, 2006.

[33] K. Nasrollahi and T. B. Moeslund, "Super-resolution: a comprehensive survey," *Machine Vision and Applications (MVA)*, vol. 25, no. 6, 2014.

[34] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. of NIPS*, 2014.

[35] L. Galteri, L. Seidenari, M. Bertini, and A. Del Bimbo, "Deep generative adversarial compression artifact removal," in *Proc. of ICCV*, 2017.

[36] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. of CVPR*, 2017.

[37] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. Change Loy, "Esrgan: Enhanced super-resolution generative adversarial networks," in *Proc. of ECCV*, 2018.

[38] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. of ICCV*, 2017.

[39] R. A. Yeh, C. Chen, T. Lim, M. Hasegawa-Johnson, and M. N. Do, "Semantic image inpainting with perceptual and contextual losses," *CoRR*, vol. abs/1607.07539, 2016.

[40] J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. A. Efros, and T. Darrell, "Cycada: Cycle-consistent adversarial domain adaptation," in *Proc. of ICML*, 2018.

[41] P. Suarez, A. Sappa, and B. Vintimilla, "Learning to colorize infrared images," in *Proc. of PAAMS*, 2018.

[42] A. Mehri and A. D. Sappa, "Colorizing near infrared images through a cyclic adversarial approach of unpaired samples," in *Proc. of CVPR-W*, June 2019.

[43] Z. Wang, Z. Chen, and F. Wu, "Thermal to visible facial image translation using generative adversarial network," *IEEE Signal Processing Letters*, vol. 25, no. 8, 2018.

[44] H. Zhang, B. Riggan, S. Hu, N. Short, and V. Patel, "Synthesis of high-quality visible faces from polarimetric thermal faces using generative adversarial networks," *International Journal of Computer Vision (IJCV)*, vol. 127, 2019.

[45] P. Perera, M. Abavisani, and V. Patel, "In2i : Unsupervised multi-image-to-image translation using generative adversarial networks," in *Proc. of ICPR*, 11 2017.

[46] C. Devaguptapu, N. Akolekar, M. M Sharma, and V. N Balasubramanian, "Borrow from anywhere: Pseudo multi-modal object detection in thermal imagery," in *Proc. of CVPR-W*, 2019.

[47] A. Antoniou, A. Storkey, and H. Edwards, "Augmenting image classifiers using data augmentation generative adversarial networks," in *Proc. of ICANN*, 2018.

[48] C. Bowles, L. Chen, R. Guerrero, P. Bentley, R. N. Gunn, A. Hammers, D. A. Dickie, M. del C. Valdés Hernández, J. M. Wardlaw, and D. Rueckert, "Gan augmentation: Augmenting training data using generative adversarial networks," *ArXiv*, vol. abs/1810.10863, 2018.

[49] Y. Wang, Y. Chen, J. Choi, and C. J. Kuo, "Towards visible and thermal drone monitoring with convolutional neural networks," *APSIPA Transactions on Signal and Information Processing*, vol. 8, 2019.

[50] L. Zhang, A. Gonzalez-Garcia, J. van de Weijer, M. Danelljan, and F. S. Khan, "Synthetic data generation for end-to-end thermal infrared tracking," *IEEE Transactions on Image Processing*, vol. 28, no. 4, 2019.

[51] T. Guo, C. P. Huynh, and M. Solh, "Domain-adaptive pedestrian detection in thermal images," in *Proc. of ICIP*, 2019.

[52] V. V. Kniaz, V. A. Knyaz, J. Hladůvka, W. G. Kropatsch, and V. Mizginov, "ThermalGAN: Multimodal color-to-thermal image translation for person re-identification in multispectral dataset," in *Proc. of ECCV-W*, 2019.

[53] J. Zhu, R. Zhang, D. Pathak, T. Darrell, A. A. Efros, O. Wang, and E. Shechtman, "Toward multimodal image-to-image translation," in *Proc. of NIPS*, 2017.

[54] A. B. L. Larsen, S. K. Sønderby, H. Larochelle, and O. Winther, "Autoencoding beyond pixels using a learned similarity metric," in *Proc. of ICML*, 2016.

[55] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "Infogan: Interpretable representation learning by information maximizing generative adversarial nets," in *Proc. of NIPS*, 2016.

[56] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. P. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. of CVPR*, 2017.

[57] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, vol. abs/1804.02767, 2018.

[58] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley, "Least squares generative adversarial networks," in *Proc. of ICCV*, Oct 2017.

[59] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced deep residual networks for single image super-resolution," in *Proc. of CVPR-W*, 2017.

[60] G. Huang, Z. Liu, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. of CVPR*, 2016.

[61] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. of CVPR*, 2016.

[62] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional GANs," in *Proc. of CVPR*, 2018.

[63] I. P. Durugkar, I. Gemp, and S. Mahadevan, "Generative multi-adversarial networks," *CoRR*, vol. abs/1611.01673, 2016.

[64] A. Karnewar and O. Wang, "Msg-gan: Multi-scale gradients for generative adversarial networks," in *Proc. of CVPR*, June 2020.

[65] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. of ECCV*, 2016.

[66] A. Dosovitskiy and T. Brox, "Generating images with perceptual similarity metrics based on deep networks," in *Proc. of NIPS*, 2016.