Applied Soft Computing 121 (2022) 108793

Contents lists available at ScienceDirect

Applied Soft Computing

journal homepage: www.elsevier.com/locate/asoc

Exploring thermal images for object detection in underexposure regions for autonomous driving

Farzeen Munir [a], Shoaib Azam [a], Muhammd Aasim Rafique [a], Ahmad Muqeem Sheri [b], Moongu Jeon [a,*], Witold Pedrycz [c,d,e]

[a] School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology, Gwangju, South Korea
[b] Department of Computer Software Engineering, National University of Sciences and Technology (NUST), Islamabad, Pakistan
[c] Department of Electrical and Computer Engineering, University of Alberta, Edmonton, AB T6R 2V4, Canada
[d] Department of Electrical and Computer Engineering, Faculty of Engineering, King Abdulaziz University, Jeddah 21589, Saudi Arabia
[e] Systems Research Institute, Polish Academy of Sciences, Warsaw 01-447, Poland

ARTICLE INFO

Article history:
Received 3 May 2021
Received in revised form 22 December 2021
Accepted 24 March 2022
Available online 1 April 2022

Keywords:
Thermal Object detection
Domain adaptation
Style transfer

ABSTRACT

Underexposure regions are vital in constructing a complete perception of the surrounding environment for safe autonomous driving. The availability of thermal cameras has provided an essential alternative to explore regions where other optical sensors lack in capturing interpretable signals. A thermal camera captures an image using the heat difference emitted by objects in the infrared spectrum, and object detection in thermal images becomes effective for autonomous driving in challenging conditions. Although object detection in the visible spectrum domain has matured, thermal object detection lacks effectiveness. A significant challenge is the scarcity of labeled data for the thermal domain, which is essential for SOTA artificial intelligence techniques. This work proposes a domain adaptation framework that employs a style transfer technique for transfer learning from visible spectrum images to thermal images. The framework uses a generative adversarial network (GAN) to transfer the low-level features from the visible spectrum domain to the thermal domain through style consistency. The efficacy of the proposed object detection method in thermal images is evident from the improved results when using styled images from publicly available thermal image datasets (FLIR ADAS and KAIST Multi-Spectral).

1. Introduction

Object detection, as one of the elemental component of the perception system, has a wide range of applications ranging from medical to autonomous driving. For autonomous driving, the perception of the environment plays a pivotal role in determining safety of autonomous driving. Environmental perception is generally defined as awareness of or knowledge about the surroundings, and the understanding of the situation by the visual perception [1]. Furthermore, since autonomous driving has to offer broader access to mobility, the safety standards as instructed by SOTIF (Safety of the intended functionality)[1] perception system constitute of the object detection must reflect the safe and secure course of action for the autonomous driving.

The sensors commonly used for perception in autonomous driving include Lidar, RGB cameras, and radar. Object detection using these sensor modalities provides the perception for autonomous driving, but in contrast, each of these sensor modalities has its drawbacks. Lidar gives a sparse 3D representation of the environment, but small objects like pedestrians and cyclists are hard to detect at a considerable distance. Similarly, the RGB camera performs poorly in unfavorable illumination conditions such as low lighting, sun glare, and glare from the vehicle's headlight. Radar has a low spatial resolution to detect pedestrians accurately. There exists a performance gap in object detection for adverse lighting conditions [2]. The inclusion of a thermal camera in the sensor's suite provides a way to fill the blind spots in environmental perception. The thermal camera is robust against illumination variation and has the advantage of being deployed day and night. Object detection and classification are indispensable for visual perception, which provides a basis for computing perception in autonomous driving.

Object detection in visible spectrum (RGB) domain is considered sufficient for conventional AI applications, and has resulted in deep neural network models for robust object detection [3] [4] [5]. However, object detection accuracy in thermal images has not yet attained state-of-the-art results compared to visible

* Corresponding author.
E-mail address: mgjeon@gist.ac.kr (M. Jeon).

[1] https://newsroom.intel.com/wp-content/uploads/sites/11/2019/07/Intel-Safety-First-for-Automated-Driving.pdf

https://doi.org/10.1016/j.asoc.2022.108793
1568-4946/

spectrum RGB images. The aforementioned object detection algorithms depend on networks that have been trained on sizable RGB datasets such as ImageNet [6], PASCAL-VOC [7], and MS-COCO [8]. Unfortunately, there exists a scarcity of such large-scale public datasets in the thermal domain. Two primary datasets for urban thermal imagery that are publicly available include the FLIR-ADAS image dataset[2], and KAIST Multi-Spectral dataset [9]. However, the KAIST Multi-Spectral dataset only gives annotations for persons, while the FLIR-ADAS dataset provides annotations for four classes. In order to overcome the absence of the large-scale labeled dataset, here, a domain adoption framework for object detection in the thermal domain is presented.

Currently, numerous approaches for domain adaptation have been introduced, which aim to narrow down the gap between source and target domain. Among many, generative adversarial networks (GAN) [10], and domain adaptation [11] for the feature adaptation are noteworthy. The domain adaptation prospects in data starved thermal images domain which is the motivation of this study. It explores a derivative of closing the gap between visible and infrared spectrum in the context of object detection. Generative models influence domain adaptation; for instance, CycleGAN [12] translates the single instance of the source domain to the target domain without translating the style attributes to the target domain. The low-level visual cues have an implicit impact on the performance of object detection [13]. The delegation of these visual cues in the target domain from the source domain can be beneficial for robust object detection in the target domain.

This work proposes a framework based on domain adaptation for thermal object detection by translating the low-level features adopted from a source domain (RGB) to a target domain (thermal). A multi-style transfer approach is employed in the domain adaptive framework to translate low-level features such as curvatures and edges from the source domain to the target domain. Deep learning-based object detection architectures that rely on classical backbone like VGG [14], ResNet [15] are trained on the multi-style transfer images from scratch for robust object detection in the thermal domain (target domain). Moreover, we have proposed a cross-domain model transfer [3] method for object detection in thermal images supplementing the domain adaptation. The cross-domain model transfer for which the object detection deep neural networks have trained in the source domain (visible spectrum). The trained models, referred to as cross-domain models, are evaluated with multi-style transfer images and without multi-style transfer images in the target domain (infrared spectrum). The proposed techniques are evaluated on FLIR-ADAS, and KAIST Multi-Spectral [9] datasets, and PASCAL-VOC evaluation is used to determine the average mean precision of the detected objects[7]. The major contributions in this work are highlighted below:

1. Fusion of two domains at the data level for the object detection and confirming the hypothesis by extensive experimentation using the available FLIR ADAS and KAIST Multi-Spectral datasets. The underlying thesis is that the style transfer relegates low-frequency features from the source domain to the target domain that form the basis of improved detection accuracy and classification.
2. Improved object detection in the infrared spectrum (thermal images) by exploring the low-level features through style consistency. The proposed object detection framework outperformed existing benchmarks in terms of mean average precision.
3. Cross-domain model transfer paradigm not only enhances the object detection in the infrared spectrum (thermal images) but also provides an alternative yet effective method for labeling the unlabeled dataset.

This work illustrates a novel approach to improve object detection for thermal images is introduced by transferring knowledge through domain adaptation by employing style transfer. This work's primary motivation is to handle the scarcity or non-existence of labeled data, which is an utmost challenge to the research community, and further, the labeling of data is an expensive task.

The paper is organized as follows: Section 2 discusses the related literature. In Section 3, the proposed methodology is discussed. Section 4 focuses on experimentation and analysis of results. Section 5 shows the comparison and discussion about the proposed method. Section 6 concludes the study.

2. Related work

2.1. Object detection

Human vision can identify objects in countless challenging conditions, but it is not a trivial task for autonomous driving. The ultimate goal of object detection in images is to localize and identify all instances of the same object or different objects present in the image.

Significant work is done on person detection in thermal images by considering the temperature difference between hot bodies and cold surroundings. Classical image processing techniques can be used for detection like thresholding is used in [16]. They have formulated the threshold value based on a model which considers different thermal images' characteristics. The Histogram of oriented gradient (HOG) features and local binary patterns (LBP) are used to extract features from thermal images and the features are used to train the Support Vector Machine (SVM) classifiers in [17]. [18] have used HOG features combined with geometric features such as mean and contrast to compute a set of features that are then used to train the SVM classifier. The classical methods lack robust features and accuracy in detecting thermal object detection compared to deep neural networks and are not suitable for the dynamic situation of autonomous drivings. Deep neural networks have gained a reputation in object detection tasks for RGB images and are used for object detection in thermal images[19]. In [20], first, they have trained two separate convolution networks on thermal and RGB images separately. Then, they have proposed four fusion architectures that integrate two convolution networks at different convolution stages. They discover that convolution neural networks train on thermal images and RGB images provide complementary information on discriminating objects in thermal images, thus yielding better performance. Similar work is conducted in [21] where they have proposed fusion architecture to study the benefit of using multispectral data for thermal object detection. [22] have proposed a real-time multispectral pedestrian detector by training You Only Look Once (YOLO) object detector with the input of 3 RGB channels in addition to thermal as to the fourth channel. [23] has proposed a method based on the fusion of thermal and visible domain using target enhanced multi-scale decomposition model. The Laplacian pyramid is used to compute low-frequency features in thermal images and then fuse the information with the visible spectrum to improve the features of the target object, improving the reliability of target recognition and detection.

[2] https://www.flir.in/oem/adas/adas-dataset-form/

[3] The cross-domain model is coined by the cross-domain interoperability where the systems from different domains interact in information exchange, service, or work together to achieve the common goal. The cross-domain model is the knowledge transfer of a model that is trained in one domain and can be used in another domain by implying the feature learned by that model is reused for the other domain.

2.2. Domain adaptation

Typically, neural networks encounter performance degradation when tested upon different datasets due to environmental changes. Furthermore, the dataset is not large enough to train and optimize a network in some cases. Therefore techniques like domain adaptation provide a crucial tool to the research community[24].

The domain adaptation for object detection includes techniques like generating synthetic data or augmentation to real data to train the network. [25] have used publicly available object detection labeled datasets coming from various domains and multiple classes and merged them. For example, the fashion dataset Modanet is merged with the MS-COCO dataset by leveraging Faster-RCNN using domain adaptation. In [26], Faster-RCNN is used to make image and instance-level adaptation. [27] has introduced a two-step method, where they have optimized a detector to low-level features, and then it is developed as a robust classifier for high-level features by enforcing distance minimization between content and style image. [28] has proposed a cross-domain semi-supervised learning structure that takes advantage of pseudo-annotations to learn optimal representations of the target domain. They have utilized the fine-grained domain transfer, progressive confidence-based annotation augmentation, and annotation sampling strategy.

2.3. Transfer learning

In real-world applications, the train and test data do not belong to the same feature space or have similar data distributions, although most machine learning algorithms hold this assumption. In light of the violation of this assumption, most machine learning models need to be rebuilt using new labeled training data [29]. For such task transfer learning helps transfer the knowledge between task domains[30]. [31] has exhibited the transfer learning-based framework for object detection datasets with very few training examples. They have augmented the examples from each class by importing the examples from other classes and transforming them to be more similar to the target class. [32] presents a boosting framework to transfer learning from multiple sources. The brute force transfer of knowledge might transfer weak relationships, which reduces the classifier's performance. The knowledge is borrowed from multiple sources to evade negative transfer. [33] performs a study to examine the efficacy of transfer learning affected by choice of dataset. They have proposed adaptive transfer learning, a simple and effective pre-training technique based on weights computed on the target dataset. [34] solves the fine-grained visual categorization problem using domain adaptive transfer learning. They have fed the neural network additional data by augmenting it through a visual attention mechanism and fine-tuning it on the base network. [35] propose a new technique based on transfer learning to relegate the knowledge from the source task to the target task containing uncertain labels.

2.4. Style transfer

Image Style transfer is a process that renders the image's content from one domain with the style of another image from another domain. [36] has demonstrated the use of feature representation from the convolution neural network for style transfer between two images. They have shown that features obtained from CNN are separable. They manipulate the feature representation between style and content images to generate new and visually meaningful images. [37] have proposed style transfer based on a single object. They have used patch permutation to train a GAN to learn the style and apply it to the content image. [38] has introduced XGAN, consisting of an auto-encoder, which captures the shared features from style and content images in an unsupervised way and along which it learns the translation of style onto the content image. [39] has proposed the CoMatch layer, which learns the second-order statistics of features and then matches them with the style image. Using the CoMatch layer, they have developed the Multi-style Generative Network giving a real-time performance.

There is still a need for improvement in thermal object detection in the context of the aforementioned related literature extending from object detection, transfer learning, style transfer, and domain adaptation. The resurgence of feature extraction without human supervision has greatly improved by the deep neural networks in the visible spectrum RGB domain for the classification, detection, and prediction problems. In addition, the leverage of the proposed approach is to perform domain adaptation for other datasets, like introducing foggy weather in the KITTI dataset [40] or converting day images to night images.

3. Proposed method

This section presents the proposed domain adaptive framework for thermal object detection from visible RGB domain to thermal domain.

3.1. Object detection in thermal images through style consistency (ODSC)

The recent advances in deep learning have revolutionized object detection in the visible RGB image domain. However, there is still room for improvement in the thermal image domain. Object detection focuses on locating and identifying objects of different classes in an image. There could be a single instance of the object from one class or multiple objects from different classes, making the problem of object detection challenging. Deep neural networks as function approximators perform low-level and high-level feature extraction for the classification/prediction problem [13,41], which provides superior features in comparison to handcrafted features. Hence the reason for improved object detection in the visible RGB image. Here, we argue that transferring the low-level features from the source domain (RGB) using domain adaption increases the target domain's (thermal) object detection performance.

The knowledge transfer using the domain adaptation between the thermal image (content images x_c) and visible spectrum (RGB) images (style images x_s), we have adopted the multi-style generative network (MSGNet) for style transfer [39]. Style transfer is considered a technique to reconstruct and synthesize texture based upon the image's semantic content. It provides consistency in content-style interpolation, color preservation, and spatial control. The leverage of translating the specific style from the source to the target domain through the multi-style generative network provides an extra edge over the CycleGAN [5]. The CycleGAN generates one translated image from the source image of a specific style. MSGNet provides the capability to translate multi-style from the source domain to the target domain while closing the gap between the two domains. The network extracts low-level features such as texture and edges from the source domain while keeping the high-level features like location and shape consistent in the target domain. Fig. 2(a) shows the framework for transferring the style from the visible spectrum (RGB) images to thermal images.

The architecture of the MSGNet is shown in Fig. 2(a). MSGNet network takes both the content image x_c and style image x_s as input, while the previously known architectures, like, Neural Style

[37] that takes only the content image and then generates the transferred image. The Generator network (G) is composed of an encoder consisting of the siamese network [42], which shares its network weights with the transformation network through the CoMatch layer. The CoMatch layer matches the second-order feature statistics of content image x_c to the style images x_s. For a given content image and a style image, the activation of the descriptive network at the j^{th} scale $\mathcal{F}^j(x) \in \mathbb{R}^{C_j \times H_j \times W_j}$ represents the content image x_c where C_j, H_j, W_j are the number of feature map channels, the height of feature map and width respectively. The distribution of features in style image x_s is represented using the Gram Matrix $\mathcal{G}(\mathcal{F}^j(x)) \in \mathbb{R}^{C_j \times C_j}$ given by Eq.(1)

$$\mathcal{G}(\mathcal{F}^j(x)) = \Phi(\mathcal{F}^j(x))\Phi(\mathcal{F}^j(x))^T \,, \tag{1}$$

where Φ is a reshaping function in Gram Matrix $\mathcal{G}$ for zero-centered data.

In order to find the desired solution in the CoMatch layer that preserves the semantic content of the source image as well as matches the feature statics of the target style, an iterative approximation approach is adopted by incorporating the computational cost in the training stage, as shown in the Eq.(2)

$$\hat{y}^j = \Phi^{-1}\left[\Phi(\mathcal{F}^j(x_c)^T)W\mathcal{G}(\mathcal{F}^j(x_s))\right]^T \,, \tag{2}$$

where W is a learnable matrix.

The minimization of a weighted combination of the content and style difference between the generator network output and targets for a given pre-trained loss network $\mathcal{F}$. The generator network is given by $G(x_c, x_s)$ and parameterized by W_G, (weights). The learning is done by sampling the content image $x_c \sim X_c$ and style image $x_s \sim X_s$, and estimating the weights, W_G of the generator $G(x_c, x_s)$ to minimize the loss, as shown in Eq.(6)

$$A = \lambda_c \left\| \mathcal{F}_{x_c}(G(x_c, x_s)) - \mathcal{F}_{x_c}(x_c) \right\|_F^2 \,, \tag{3}$$

$$B = \lambda_s \sum_{j=1}^{K} \left\| \mathcal{G}(\mathcal{F}^j((G(x_c, x_s)))) - \mathcal{G}(\mathcal{F}^j(x_s)) \right\|_F^2 \,, \tag{4}$$

$$C = \lambda_{TV} l_{TV}(G(x_c, x_s)) \,, \tag{5}$$

$$\hat{W}_G = argminE_{x_c, x_s}\{A + B + C\} \,, \tag{6}$$

where λ_c and λ_s are the regularization parameters for content and style losses. The content image is considered at scale c and style image is considered at scales $i \in 1, \ldots, K$. The total variational regularization is l_{TV}, which is used for the smoothness of the generated image [43].

The proposed framework for object detection through style consistency is presented in Fig. 2. It illustrates that the network consists of two modules; the first part consists of a multi-style network. It generates the style images by adapting low-level features transformation between the content image consisting of thermal image and style image consisting of the RGB image. As compared to the thermal images, the transferred style images contain low-level features, but the semantic shapes are preserved in these generated images keeping the high-level semantic features consistent. The second module is comprised of the object detector, which inputs styled images from the first module, consequently bridging the domain gap between visible spectrum RGB images and thermal images. We have chosen a notable single-stage Single-shot Multi-box object detector (SSD) for detection architectures.

The single-shot detector (SSD) is a single-stage object detector based on a feed-forward convolutional neural network [4]. SSD architecture consists of a backbone network pre-trained on imageNet data. SSD computes a fixed set of defaults collection of bounding boxes at different scales of each location of the feature map, and for every instance, a probability score to determine the presence of the object contained by the bounding box is defined. A non-maximum suppression step is applied to produce the final detections. SSD during the training matches these default bounding boxes to the ground-truth boxes, the boxes that are matched are called positive examples, and the rest as negative examples. The negative mining is performed to calculate the confidence loss with a ratio of 3:1. The SSD-300 and 512 [4] with backbone VGG16 [14], MobileNet [44] and EfficientNet [45]. are trained on the styled images, which bridge the gap between the visible spectrum RGB images and thermal images. Moreover, Faster-RCNN is also used as an object detector to make a fair comparison. The trained detection network is evaluated on thermal images. The accuracy of testing on thermal images shows the efficacy of object detection.

4. Experimentation and results

4.1. Datasets

In this study, we have used two thermal image datasets. The first is the FLIR-ADAS dataset, and the second one is the KAIST Multi-Spectral dataset [9]. FLIR-ADAS dataset consists of 10228 images with objects annotated using a bounding box as an evaluation measure. The objects are classified into four categories, i.e., car, person, bicycle, and dog. However, the dog category has very few annotations, so it is not considered in this study. The images have a resolution of 640 × 512 and were obtained from FLIR Tau2 Camera. The dataset consists of day and night images, approximately 60% (6136) images are captured during the daytime, and 40% (4092) images are captured during nighttime. The dataset consists of both visible spectrum (RGB images) and thermal images, but annotations are only available for thermal images. The visible spectrum (RGB images) and thermal images are not paired, so the thermal annotations cannot be used with a visible spectrum (RGB images). Thermal images with annotations are only considered in this study. A standard split [4] of the dataset into training and validation data is considered during experimentation. The training dataset consists of 8862 images, and the validation contains 1366 images, as shown in Table 1.

The KAIST Multi-Spectral dataset contains 95000 images from both the visible spectrum (RGB images) and the thermal spectrum, and for each category, the dataset has both daytime and nighttime images. Annotations are only provided for the person class with a given bounding box. The visible spectrum (RGB images) and thermal images are paired, which means annotations for the thermal and the visible spectrum (RGB images) are the same. Images are captured using a FLIR A35 camera with a resolution of 320 × 256. We have applied a standard split [5] of the dataset, using 76000 of the images in the dataset in training and 19000 of the images in the dataset for validation as shown in Table 1.

[4] As given by FLIR ADAS repository

[5] As given by the KIAST repository.

Table 1
FLIR-ADAS and KAIST multi-spectral datasets partition topology for training and testing the proposed network.

Dataset	Total images	Train images	Test images
FLIR-ADAS	10228	8862	1366
KAIST multi-spectral	95000	76000	19000

4.2. Evaluation metric

The efficacy of the proposed method is determined by using the mean average precision (mAP) as an evaluation metric. In computer vision research, object detection is usually classified as a combination of localization and classification problems. The localization deals with the determination of the bounding box coordinates, whereas the classification corresponds to identifying the object labels. The notable object detectors such as Faster-RCNN, SSD, and YOLO have utilized the mAP as an evaluation metric to analyze object detection performance. The concept of mean average precision relies on computing the precision, recall, and intersection over union (IoU) for object detection. Precision for object detection measures the number of correct predictions, whereas the recall corresponds to how well these positive predictions are found. Intersection over Union (IoU) is a measure based on Jaccard Index that evaluates the overlap between two bounding boxes. Precision, recall and IoU are defined as illustrated in Eq.(7), Eq.(8), and Eq.(9), respectively,

$$Precision = \frac{TP}{TP + FP}, \tag{7}$$

$$Recall = \frac{TP}{TP + FN}, \tag{8}$$

$$IoU = \frac{area(B_p \cap B_{gt})}{area(B_p \cup B_{gt})}, \tag{9}$$

where TP, FN, and FP are true-positive, false negative, and false-positive, respectively. B_p and B_{gt} represent the prediction and ground-truth bounding box, respectively. In object detection, precision and recall are calculated using the IoU value for a given IoU threshold. In our experimentation, we have used the Pascal VOC evaluation metric with 50% of IoU threshold ($IoU_{0.5}$). The average precision is evaluated by finding the area under the precision-recall curve as shown in Eq.(10)

$$AP = \int_0^1 p(r)dr. \tag{10}$$

The mean average precision (mAP) is calculated by evaluating the mean AP over all classes and overall IoU thresholds. As we have utilized the Pascal VOC evaluation metric in our proposed work, the AP is calculated for the IoU threshold of 0.5; the mAP score is averaged over all the object classes.

4.3. Object detection in thermal images through style consistency (ODSC)

In contrast to traditional detection methods, deep neural network utilization for the object detection task has enhanced object detectors' performance. Deep learning-based object detectors are classified into two genres: "two-staged detection" and "one-stage detection", where the former focuses on coarse to fine process for object detection while the latter frames the object detection as a complete one-step process.

This work aims to bridge the gap between thermal and RGB domains by incorporating the useful features representation from the RGB domain that helps in the thermal object detection for the autonomous vehicle perception application. To this end, we have utilized the state-of-the-art object detectors from both two-staged and one-staged detectors to evaluate the proposed method.

In the context of a two-stage object detector, Faster-RCNN is the state-of-the-art detector constitutes of a deep convolutional region proposal network and Fast-RCNN detector [3]. The Faster-RCNN object detector includes a feature extractor network, usually composed of deep neural network-based backbone like VGG [14], ResNet [15] pre-trained on imageNet, is used to extract features from the input image. The backbone network is then followed by the region proposal network (RPN), consisting of 3 convolutional layers that generate several bounding boxes known as the region of interest (ROIs). These ROIs are generated by sliding a window at each location of the feature map and simultaneously predicting multiple regions' proposals. The maximum number of possible regions proposals called anchors at each location are pre-determined. These regions proposals are bounding boxes with a higher probability of containing an object. Finally, the Fast-RCNN detection network takes features from the backbone and ROIs from the RPN module and predicts the bounding box and class of objects present in an image. Although the Faster-RCNN breaks the speed bottleneck of Fast-RCNN, it needs improvement in computation to reduce the redundancy at the detection stage.

The Single Shot Multibox Detector (SSD) is the one-stage detector in which the localization and classification are performed in a single forward pass of the network. The SSD object detector is composed feature extractor that serves as a backbone and a multi-box detector for detection. VGG-16 architecture (discarding the fully connected layers) is adopted as a backbone network in SSD for feature extraction. The reason for using the VGG-16 network as a backbone network is its robust performance with simple architecture for image classification tasks and also its useability in transfer learning problems for improving the results. Besides the VGG network, in literature, ResNet and DenseNet are employed as the backbone network for feature extraction yet developing new object detectors on the basis of SSD architecture. Since the focus of this paper is to develop a framework to bridge the gap between thermal and RGB domains for object detection in thermal images, we explicitly follow the original architectural designs of the object detector to determine their usability in the proposed work.

The evaluation of the proposed method is demonstrated using state-of-the-art object detection networks. The object detection networks include Faster-RCNN, SSD-300, and SSD-512. These object detection networks are implemented with different backbone architecture; for instance, ResNet-101 is used as a backbone network in Faster-RCNN; VGG16, MobileNet, and EfficientNet are used with SSD-300; SSD-512 uses VGG16 as backbone architecture. The dataset comprises of FLIR-ADAS and KAIST Multi-Spectral dataset. The FLIR-ADAS dataset is partitioned into training and testing using a standard split, while the KAIST Multi-Spectral dataset is only used in testing the object detection networks. All the networks are implemented in Pytorch, having formulated the data in PASCAL-VOC format. The standard PASCAL-VOC evaluation criteria are used in this study [7].

4.3.1. Baseline

A baseline approach is experimented first for the comparative analysis with the proposed methodology, which involves training and testing of object detection network using thermal images only. In training the Faster-RCNN, ResNet-101 backbone is adapted and trained on the thermal image dataset. The network is trained using Adam optimizer with a learning rate of 10^{-4} and a momentum of 0.9 for total of 15 epochs.

The experimental evaluation with the SSD object detection network constitutes two different architectures, i-e SSD-300 and

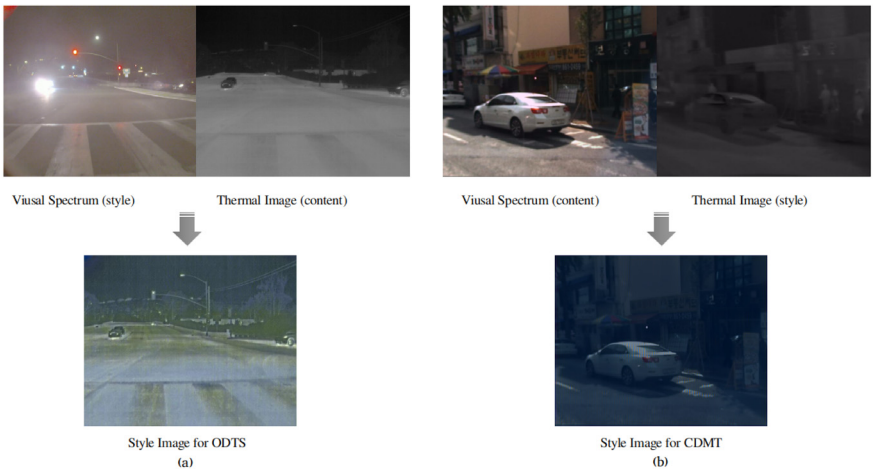

Fig. 1. (a) Object detection in thermal images through style consistency (ODSC). Visible spectrum (RGB image) is treated as a style image whereas, the thermal image is considered as content image. The output shows the enhanced image having low-level features adapt from the visible spectrum. (b) Cross-domain model transfer with style transfer. Style from the thermal image is transferred to the visible spectrum (RGB content image).

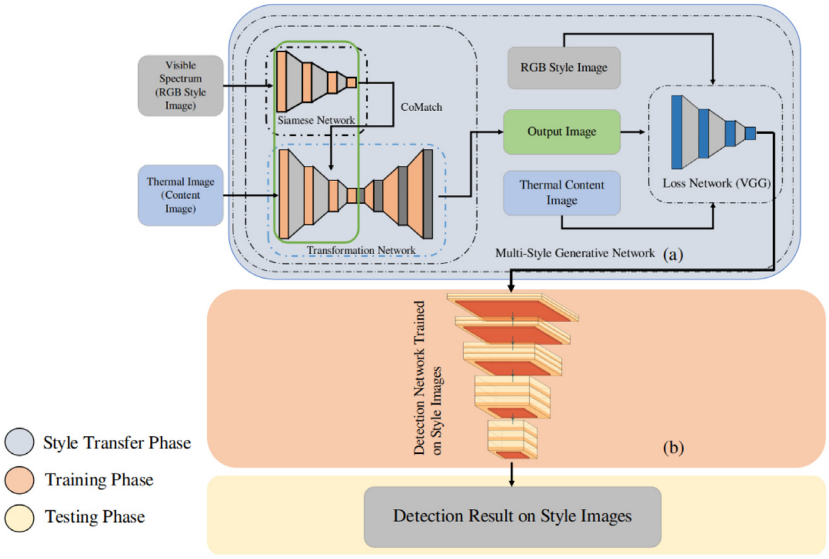

Fig. 2. The proposed model framework for object detection in thermal images through style consistency. (a) Multi-style generative network architecture for generating the style images. Visible spectrum (RGB images) and thermal images are given as style and content image respectively to the network. The siamese network captures the low-level features of style image, which is transferred to the transformation network through the CoMatch layer. A pre-trained loss network is used for MSGNet learning by computing the difference between content and style image with the targets. (b) The detection networks which includes (Faster-RCNN backbone with ResNet-101, SSD-300 with backbone VGG16, MobileNet, and EfficientNet, SSD-512 with VGG16 backbone) are trained on the style images and then tested in the target domain (thermal images) for the object detection.

SSD-512. In the case of training the SSD-300, the backbone networks are trained on the training data. The learning rate for VGG16, MobileNet, and EfficientNet used as the backbone network for SSD-300 are 10^{-4}, 10^{-3}, and 10^{-3}, respectively. For the SSD-512 experimentation, only VGG-16 is used as a backend for training with a learning rate of 10^{-3}. All the networks have used a batch size of 4 on the Nvidia-TITAN-X having 12GB of computational memory.

4.3.2. Experimental configuration of ODSC

In the proposed methodology, the MSGNet is trained with thermal images to serve as a content image, whereas the RGB images correspond to style images, as shown in Fig. 1(a). In training the MSGNet, VGG16 is used as a loss network. The pre-trained weights of the loss network on the ImageNet dataset are employed for training the MSGNet. In a loss network, the balancing weights as referred to in the Eq. (6) are $\lambda_c = 1$ and $\lambda_s = 5$ respectively while the total variational regularization for content and style is $\lambda_{TV} = 10^{-6}$. In the experimental configuration, the size of the style image x_s is iteratively updated, having a size of 256, 512, 768, respectively. The size of the content images is resized to 256×256. The Adam optimizer is used with a learning rate of 10^{-3} in the training configuration. The MSGNet is trained for a total of 100 epochs with a batch of 4 on the Nvidia-TITAN-X.

The trained model of MSGNet results in the generation of style images, as shown in Fig. 1(a). These style images are used in training the object detection networks. The detection networks trained on style images are evaluated on the test data comprise of thermal images. The training configuration of these object detection networks is kept similar as the baseline configuration to make a comparative analysis.

4.3.3. Experimental results

For the evaluation of our experimental configuration, we have tested the baseline and proposed method, on both thermal datasets

Table 2
Quantitative analysis using Baseline configuration for object detection networks.

FLIR ADAS dataset						KAIST multi-spectral dataset
Network architecture	Backbone	car	bicycle	person	Average mAP	person
Faster-RCNN	ResNet-101	0.6799	0.4276	0.548	0.5518	0.5583
SSD-300	VGG-16	0.7561	0.4502	0.6197	0.6087	0.6687
SSD-300	MobileNet-v2	0.4774	0.1943	0.3163	0.3284	0.5998
SSD-300	EfficientNet	0.6809	0.2747	0.4992	0.4849	0.6162
SSD-512	VGG-16	0.8055	0.5399	0.702	0.6825	0.6409

Table 3
Quantitative analysis using Proposed Method (ODSC) configuration.

FLIR ADAS dataset						KAIST multi-spectral dataset
Network architecture	Backbone	car	bicycle	person	Average mAP	person
Faster-RCNN	ResNet-101	0.7190	0.4394	0.6201	**0.5928**	**0.5745**
SSD-300	VGG-16	0.7991	0.4691	0.6253	**0.6312**	**0.7536**
SSD-300	MobileNet-v2	0.5434	0.2798	0.3638	**0.3957**	**0.7465**
SSD-300	EfficientNet	0.7405	0.3512	0.5169	**0.5362**	**0.6770**
SSD-512	VGG-16	0.8233	0.5553	0.7101	**0.6962**	**0.7725**

Table 4
Quantitative analysis of testing object detection networks trained on thermal images and tested on style images.

FLIR ADAS dataset						KAIST multi-spectral dataset
Network architecture	Backbone	car	bicycle	person	Average mAP	person
Faster-RCNN	ResNet-101	0.3030	0.1985	0.2115	0.2377	0.1410
SSD-300	VGG-16	0.6824	0.3286	0.5260	0.5123	0.6137
SSD-300	MobileNet-v2	0.4551	0.1363	0.2899	0.2937	0.4773
SSD-300	EfficientNet	0.3637	0.1193	0.2289	0.2373	0.4449
SSD-512	VGG-16	0.6779	0.3736	0.5538	0.5351	0.4961

(FLIR ADAS and KAIST Multi-Spectral). Table 2 shows the mean average precision (mAP) scores of the baseline configuration for each detection network, i.e., the networks are trained on thermal images and evaluated on thermal images. Table 3 shows that the quantitative results of the proposed method. The best model configuration for the proposed method is (SSD512+VGG16) as shown in experimental results. The mAP score of the best model configuration of the proposed method has a better evaluation score compared to the baseline configuration. The MobileNet-v2 is designed for edge computing devices having a fewer number of parameters in comparison to the VGG network. MobileNet-v2 is small, having low latency and low-power models parameterized to meet the computation constraints of the edge computing devices. The MobileNet-v2, when used as a backbone network with the SSD-VGG object detector, produces low performance on the FLIR dataset due to the fact that MobileNet-v2 has a lower number of parameters in contrast to the SSD-VGG, thus resulting in lower performance. Furthermore, compared to the KAIST Multi-Spectral dataset, the SSD with MobileNet-v2 backbone is only tested on the one class (person) that gives a relatively high mAP score in contrast to the FLIR dataset where it is tested for multi-class. The high mAP score in the case of the KAIST Multi-Spectral dataset is because the mAP score is evaluated on the overall classes present in the dataset. We perform a sanity check by conducting experiment by training network on thermal images and testing them on style images. The detection networks trained on the thermal images tested on the style images show the marginal efficacy, as shown by Table 4. Figs. 4–5 illustrate the qualitative result of object detection in thermal images through style consistency on FLIR ADAS and KAIST Multi-Spectral respectively for all the detection networks.

4.4. Corollary to proposed method: Cross domain model transfer for object detection in thermal images (CDMT)

For the further investigation of the proposed method, a cross-domain model for thermal object detection is designed. The purpose of this study is to analyze the effect of trained RGB detection models on styled and without styled images. It is to be noted that for cross-domain model transfer, the source and target domain are swapped compared to the first part of the proposed work. The reason of this configuration is to analyze the performance of object detectors that are trained on the RGB domain, when applied to thermal domain produce unsatisfactory results because of the fact of domain invariance.

However, if the style from the thermal domain is being employed on the content image of RGB domain, the trained RGB domain object detection networks performance improved since the style transfer bridge the gap between the two domains. Fig. 3 shows the overall framework for cross-domain model transfer object detection in thermal images. The detection networks (Faster-RCNN backbone with ResNet-101, SSD-300 with backbone VGG16, MobileNet, and EfficientNet, SSD-512 with VGG16 backbone) are trained on the visible spectrum (RGB images) and then the trained model is tested on the thermal images. As the detection networks are trained on a different domain, in this case, visible spectrum (RGB) images, the performance of these networks on thermal images will be marginal as can be seen in results. The efficacy of thermal object detection can be increased by using the style consistency. The MSGNet is trained with RGB images as the content image, and the style is borrowed from the thermal images. The style transferred images are then passed to the same detection networks that are trained earlier on the visible spectrum (RGB) images, which improves the object detection in thermal style images. This cross-domain model transfer can be applied as a weak object detection module for the unlabeled dataset, as in our case for thermal images.

4.4.1. Experimental configuration of CDMT

The cross-domain model evaluation employs the training of object detectors on the visible spectrum (RGB images). The KAIST Multi-Spectral dataset is used in this experiment, considering that the labels are available for both domains. The object detection networks incorporated in this study include Faster-RCNN, SSD-300, and SSD-512. The network model configuration is similar to ODSC. The Faster-RCNN is backend with ResNet-101 backbone.

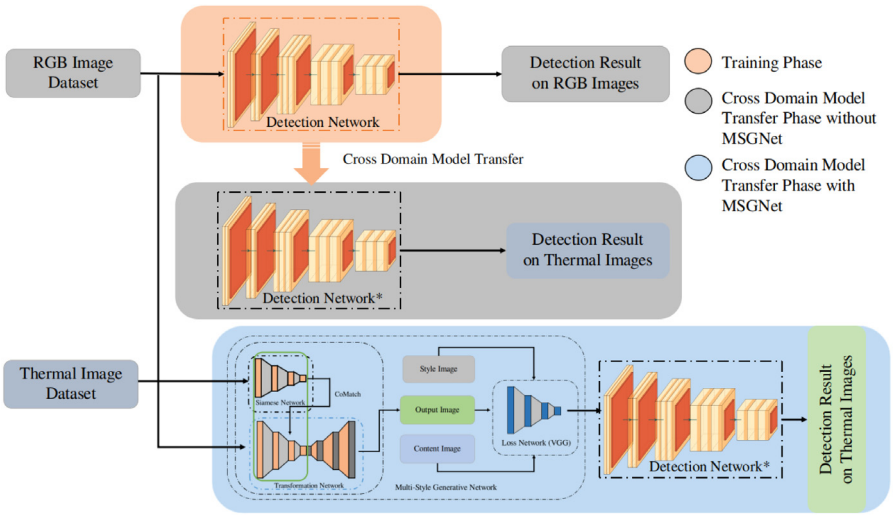

Fig. 3. An overview of the cross-domain model transfer method. The detection networks are trained using the visible spectrum (RGB images). Afterward, these trained models are tested by implying the cross-model transfer with style transfer using MSGNet and also without style transfer. (Detection Network*) implies that the same detection networks are used for testing in the target domain.

Table 5
Quantitative analysis of Cross Domain Model Transfer (CDMT).

KAIST multi-spectral dataset			
	Domain	CDMT without style transfer	CDMT with style transfer
Network architecture	Backbone	person	person
Faster-RCNN	ResNet-101	0.5354	**0.7254**
SSD-300	VGG-16	0.6098	**0.7598**
SSD-300	MobileNet-v2	0.2512	**0.7012**
SSD-300	EfficientNet	0.1995	**0.5495**
SSD-512	VGG-16	0.6202	**0.7702**

The SSD-300 network is experimented with VGG16, MobileNet, and EfficientNet backbone. Furthermore, SSD-512 is backend with VGG16 architecture. The learning rate for training all detection networks is 10^{-3} except for the SSD-300 with EfficientNet backbone, which is tested with 10^{-4}. The batch size is 4 for all the aforementioned detection networks.

Similar to the ODSC, MSGNet is used to generate styled images, as shown by Fig. 1(b). In this case, the content images consist of the visible domain (RGB images), and the style is transferred from thermal images, which signifies that the style transfer between the content image (RGB images) and style image (thermal images) increase the object detection efficacy. The hyper-parameters for the MSGNet are kept the same as described in the experimental configuration of object detection in thermal images through style consistency. The detection networks are then tested on these generated styled images.

4.4.2. Experimental results

The method's assessment is investigated by evaluating the trained network on the styled images and non-styled images (thermal images). Table 5 shows the quantitative results of cross-domain model transfer. The quantitative results show that using the cross-domain model transfer with style transfer increases the object detection efficacy compared to cross-domain model transfer without style transfer. In addition to that, the method of using cross-domain model transfer will overcome the gap of annotating the unlabeled dataset and assists as a weak detector for the unlabeled dataset. The qualitative evaluation of using style transfer for CDMT is shown in Fig. 7 for all the detection networks.

5. Comparison with other state-of-the-art methods

For the efficacy of the proposed methodology, an extensive analysis is conducted using state-of-the-art methods. ACF+T+THOG [9] conducted a study to analyze the effect of thermal channels on aggregated channel features. They used an ensemble of classical techniques to extract features from thermal and RGB color channels, including fast feature pyramids, HOG features, normalized gradients, and histograms of oriented gradients. These hand-crafted features are utilized for pedestrians detection in multi-spectral images. Our proposed method relies upon a deep neural network to obtain features and style transfer to bridge the RGB domain and thermal gap. Moreover, tY model [49] have used a deep neural network to perform person detection using thermal images. However, it uses YOLOv3 architecture and only thermal image as the input for person detection, and they did not perform any fusion or domain transfer during training of the YOLOv3 network. Nevertheless, PiCA-Net [47] and R^3Net[47] utilizes a fusion network that uses saliency maps acquired from thermal images fused with RGB images to extract better features. In addition, they have trained the Faster-RCNN for pedestrians detection and fine-tuned it on extracted feature maps.

Furthermore, for object detection in the thermal domain, Intel [48] uses Faster-RCNN as an object detector to perform transfer learning and concludes from the experiments that a network trained on RGB images and tested on thermal images fail to transfer knowledge, which is a similar conclusion obtained in this paper. MMTOD—UNIT [46] has used CycleGAN to generate the thermal images from RGB images, to remove the dependency of pairing the RGB and thermal images in the dataset. They have used a variant of Faster-RCNN, which used both the thermal and RGB images to detect objects. Using CycleGAN to generate images

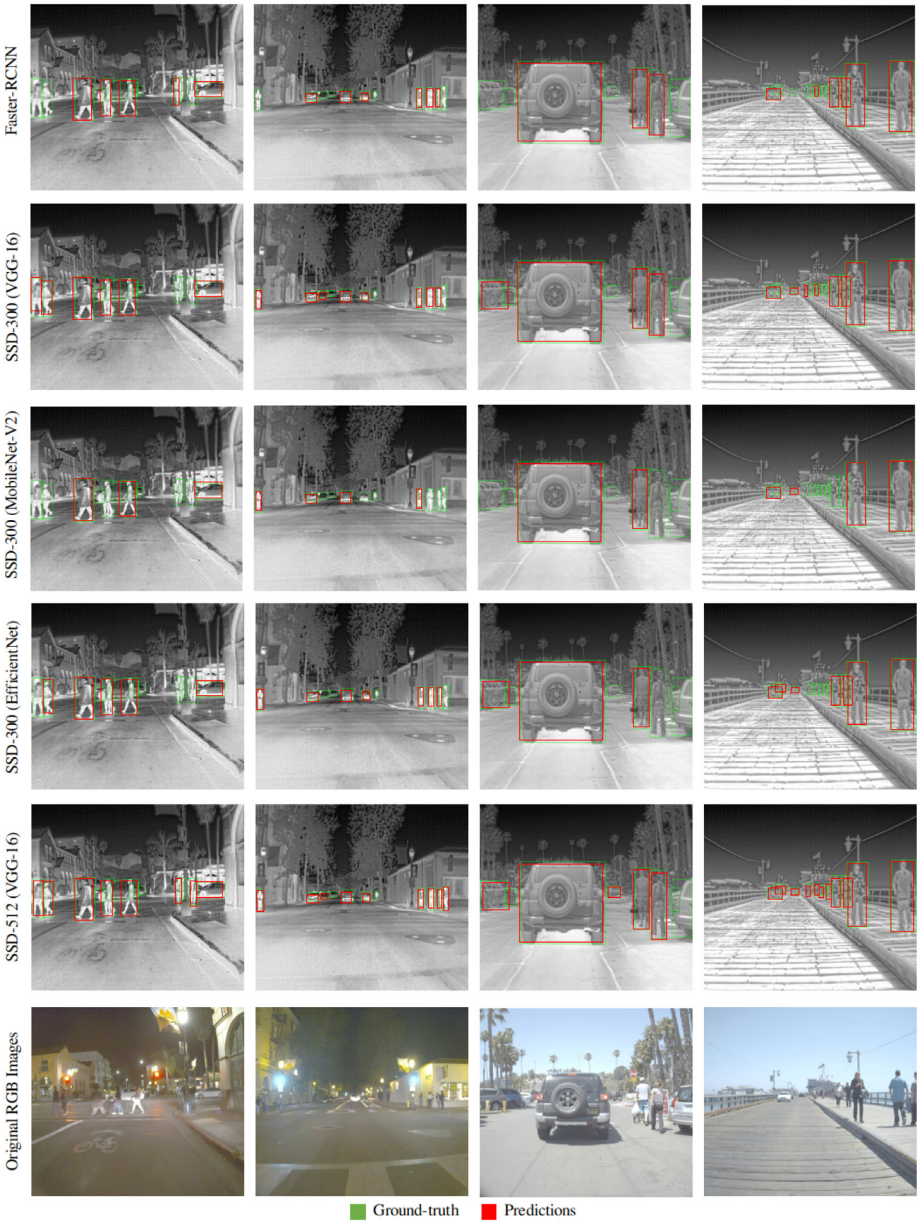

Fig. 4. Illustrates the qualitative results of object detection in thermal images through style consistency. The object detection results of all the detection networks are illustrated along with ground-truth and predictions on FLIR ADAS dataset. The second last row shows the qualitative results of best model configuration (SSD512+VGG16).

introduce negative artifacts in images and distorts the actual features of the image in comparison to style transfer which keeps higher-level feature consistent.

In this work, we have designed a framework to cater to the problem of thermal object detection by transferring the knowledge in the form of low features from the visible-spectrum RGB domain to the thermal infrared domain by utilizing style consistency. For this purpose, the knowledge transfer is done at the image level, and the detection in the thermal domain is carried out by training the deep neural networks object detection networks. In contrast to the state-of-the-art methods, the proposed method's efficacy is quantified by utilizing the mean average precision (mAP) as the evaluation metric. In the computer vision literature, the mAP is extensively used as an evaluation metric to illustrate the efficacy of the object detection network. To this end, in the proposed work, the mAP score is evaluated to make a fair comparison with the state-of-the-art methods. Table 6 shows a comparison between the proposed methods (ODSC and CDMT) and state-of-the-art methods in terms of mAP scores. In our comparison with the state-of-the-art methods, only those methods are considered that incorporate the standard PASCAL-VOC evaluation criteria for both FLIR and KAIST Multi-Spectral datasets.

In evaluating the proposed method (ODSC) in contrast to the state-of-the-art methods, we have incorporated both fusion-based and single modality-based methods dealing with thermal object detection. Since Table 6 illustrates the mAP score between the proposed (ODSC) and state-of-the-art methods, the proposed method (ODSC) achieved the overall mAP score of 0.6962 in comparison to 0.5856 of MMTOD-UNIT [46] in terms of the fusion of visible spectrum RGB and thermal infrared domain for FLIR ADAS dataset. The corresponding mAP score for MMTOD-UNIT [46] on the KAIST Multi-Spectral dataset is unavailable. Furthermore, MMTOD-CG [46] model is evaluated on both FLIR ADAS and KAIST Multi-Spectral datasets achieving 0.5711 and 0.5226 mAP score, respectively. In contrast, the proposed method (ODSC)

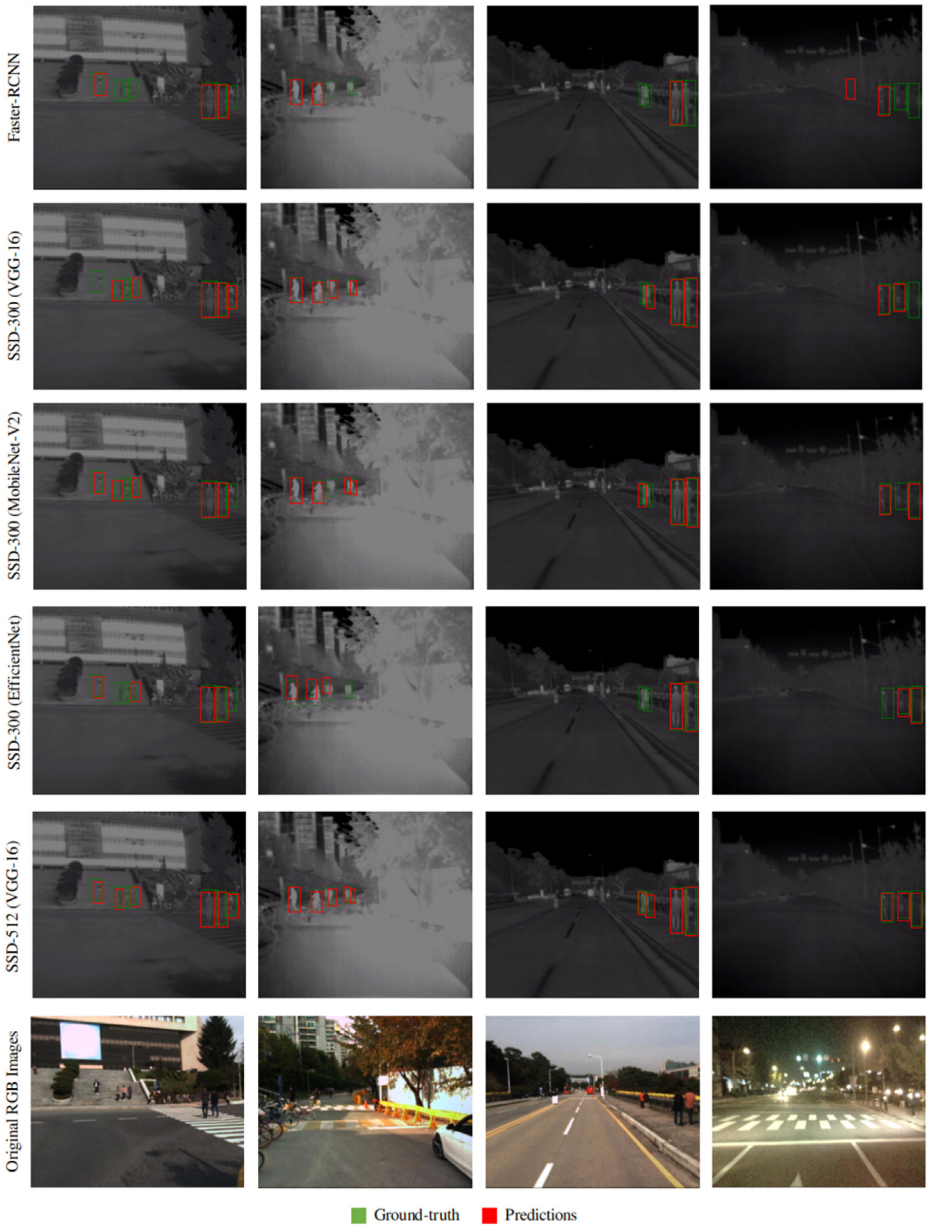

Fig. 5. Illustrates the qualitative results of object detection in thermal images through style consistency. The object detection results of all the detection networks are illustrated along with ground-truth and predictions on KAIST Multi-Spectral dataset. The second last row shows the qualitative results of best model configuration (SSD512+VGG16).

has achieved the mAP score of 0.6962 and 0.7725, respectively, for both FLIR ADAS and KAIST Multi-Spectral datasets. Similarly, the models PiCA-Net [47], R^3Net[47] and ACF+T+THOG [9] are only evaluated on the KAIST Multi-Spectral dataset. They have achieved the mAP score of 0.658, 0.7085 and 0.7139 respectively, in contrast to the proposed method (ODSC) mAP score on the KAIST Multi-Spectral dataset of 0.7725. The quantitative comparison between the proposed method (ODSC) is not limited to the fusion methods only; we have also evaluated the proposed method (ODSC) with the deep neural networks method that utilizes thermal images as a single modality. In this context, Intel [48] Faster-RCNN model and tY model [49] are evaluated on the FLIR ADAS and KAIST Multi-Spectral datasets respectively. The former gives the mAP score of 0.3157 in contrast to 0.6962 of the proposed method (ODSC) for the FLIR ADAS dataset. The latter model has achieved the mAP score of 0.630 compared to 0.7725 of the proposed method (ODSC) for the KAIST Multi-Spectral dataset.

For the proposed method corollary (CDMT) efficacy, only the KAIST Multi-Spectral dataset is utilized to compare the proposed method (CDMT) with the state-of-the-art methods quantitatively. In this context, the proposed method has achieved the mAP score of 0.7702 in comparison to the best state-of-the-art method that is 0.7139 of ACF+T+THOG [9]. It is to be noted that the evaluation for the KAIST Multi-Spectral dataset is only performed for the person class because no other class data is available for this dataset. In addition, the proposed method corollary (CDMT) is further evaluated in the context of performing as a weak labeler for the unlabeled dataset. For this purpose, we have utilized the i3 systems TE-EQ1 / TE-EV1[6] thermal camera for collecting the unlabeled dataset. The weak label annotations using the cross-domain model transfer (CDMT) are illustrated in Fig. 6. To evaluate the efficacy of the proposed CDMT, we have presented

[6] http://i3system.com/uncooled-detector/te-eq1/?lang=en

Table 6
Comparison of our proposed methods (ODSC and CDMT) with state-of-the-art methods. (*) represent average (day+night) mean Average Precision score. (-) indicates that the respective algorithm is not tested on the specified dataset.

	Dataset	FLIR ADAS				KAIST multi-spectral
	Method	car	bicycle	person	mAP	person (mAP)
	MMTOD-UNIT [46]	0.7042	0.4581	0.5945	0.5856	–
	MMTOD-CG [46]	0.6985	0.4396	0.5751	0.5711	0.5226
	PiCA-Net [47]	–	–	–	-	0.658*
	R^3Net [47]	–	–	–	-	0.7085*
	Intel [48]	0.571	0.1312	0.245	0.3157	–
	tY model [49]	–	–	–	–	0.630
	ACF+T+THOG [9]	–	–	–	–	0.7139
Ours (ODSC)	Faster-RCNN+ResNet101	0.7190	0.4394	0.6201	0.5928	0.5345
	SSD300 +VGG16	0.7991	0.4691	0.6253	0.6312	0.7536
	SSD300+ Mobilenet V2	0.5434	0.2798	0.3638	0.3957	0.7465
	SSD300+ EfficientNet	0.7405	0.3512	0.5169	0.5362	0.6770
	SSD512+VGG16	**0.8233**	**0.5553**	**0.7101**	**0.6962**	**0.7725**
Ours (CDMT)	Faster-RCNN+ResNet101	–	–	–	–	0.7254
	SSD300 +VGG16	–	–	–	–	0.7598
	SSD300+ Mobilenet V2	–	–	–	–	0.7012
	SSD300+ EfficientNet	–	–	–	–	0.5495
	SSD512+VGG16	–	–	–	–	**0.7702**

Fig. 6. The illustration of weak label annotation using the cross-domain model transfer performed on our collected unlabeled dataset.

the true positive (TP), false positive (FP), and false-negative (FN) detections. The proposed CDMT experimental analysis shows the accuracy of 67.36% on the whole unlabeled dataset. Furthermore, the inference frame rate for the detection neural network used in the proposed method is illustrated in Table 7. The number of frames per second is calculated on the Nvidia-TITAN-X having 12GB of memory.

6. Conclusion

This study proposes a domain adaptation framework for object detection in underexposure regions for autonomous driving. The framework uses domain adaptation from visible domain to thermal domain through style consistency and utilizes MSGNet to transfer low-level features from the source domain to the target domain, keeping high-level semantic features intact. The proposed method outperforms the existing benchmark for object detection in thermal images. Moreover, the effectiveness of style transfer is strengthened by using a cross-domain model transfer between visible and thermal domains.

Table 7
The inference frame per second evaluation of deep neural networks models used in the proposed work.

Network architecture	Frame per second
Faster-RCNN -ResNet101 backbone	9
SSD300 -VGG16 backbone	31
SSD300 -MobileNetv2 backbone	23
SSD300 -EffecientNet backbone	16
SSD512 -VGG16 backbone	11

The application of the proposed framework is found in autonomous driving under low lighting conditions. Object detection is integral to the core of perception, and failure to detect an object compromises the safety of autonomous driving. Thermal images provide additional meaningful data from the surroundings, and the proposed framework improves object detection results in thermal images, consequently improving the safety of autonomous driving. We aim to integrate lane detection and segmentation into the proposed framework using thermal images in future work.

CRediT authorship contribution statement

Farzeen Munir: Conceptualization, Methodology, Software, Validation, Formal analysis, Writing – original draft, Visualization. **Shoaib Azam:** Conceptualization, Methodology, Software, Validation, Formal analysis, Writing – original draft, Visualization. **Muhammd Aasim Rafique:** Investigation, Review. **Ahmad Muqeem Sheri:** Investigation, Review. **Moongu Jeon:** Resources, Supervision, Writing review. **Witold Pedrycz:** Resources, Supervision, Writing review.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

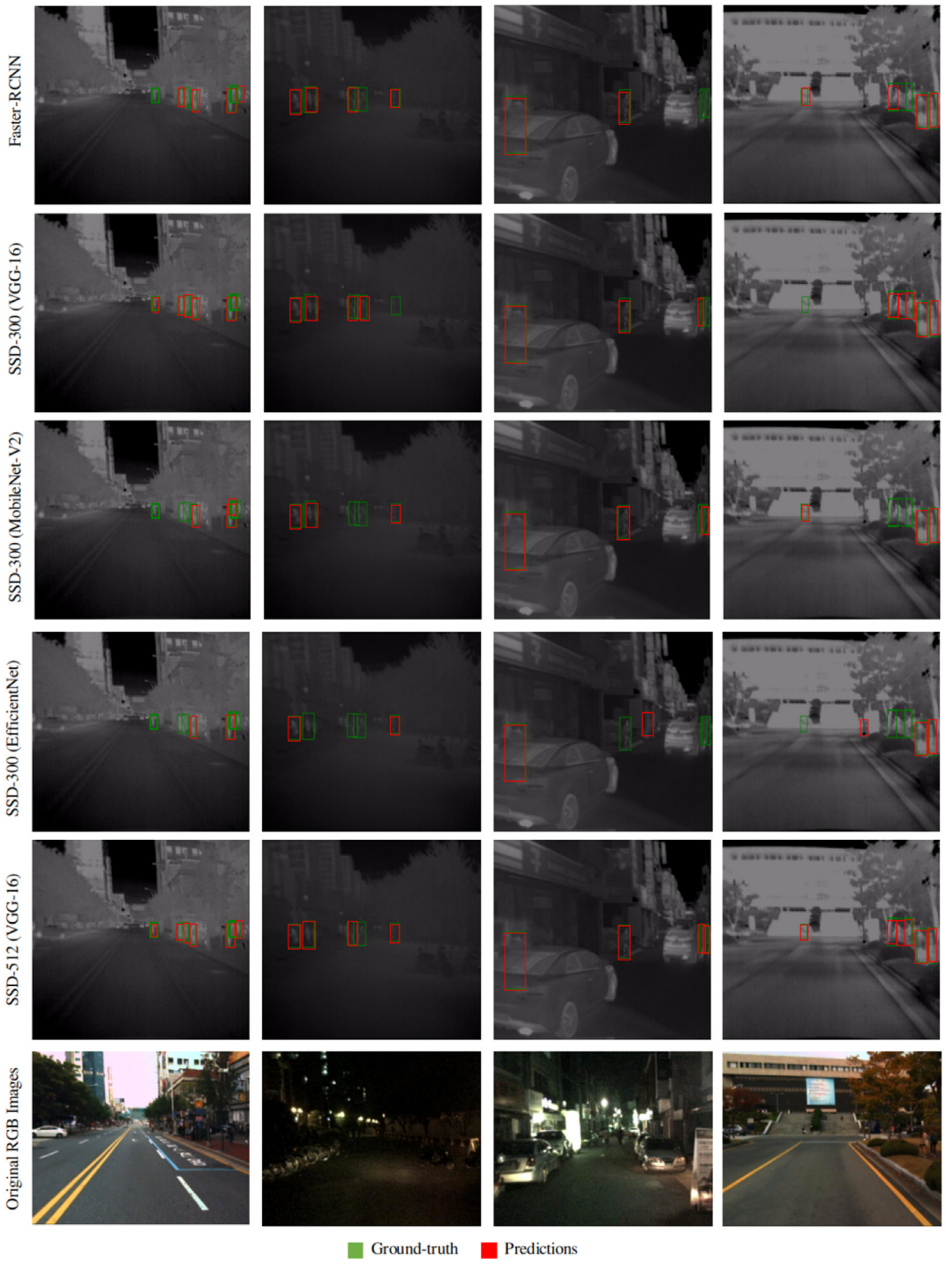

Fig. 7. Object detection results using cross domain model transfer is illustrated. The ground-truth and predictions results of all the detection network are shown. The second last row shows the qualitative results of best model configuration (SSD512+VGG16).

Acknowledgments

This work was partly supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (MSIT) (No. 2014-3-00077), Korea Creative Content Agency (KOCCA) grant funded by the Korean government (MCST) (No. R2020070004), and GIST-MIT Research Collaboration grant funded by the GIST in 2022.

All authors approved the version of the manuscript to be published.

References

[1] E.H. Zube, Environmental perception, in: Encyclopedia of Earth Science, Springer, New York, NY, 1999, pp. 214–216.

[2] J.Van. Brummelen, M. O'Brien, D. Gruyer, H. Najjaran, Autonomous vehicle perception: The technology of today and tomorrow, Transp. Res. C 89 (2018) 384–406.

[3] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, in: Advances in Neural Information Processing Systems, 2015, pp. 91–99.

[4] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.Y. Fu, A.C. Berg, Ssd: single shot multibox detector, in: European Conference on Computer Vision, Springer, Cham, 2016, pp. 21–37.

[5] J. Redmon, A. Farhadi, Yolov3: An incremental improvement, 2018, arXiv preprint arXiv:1804.02767.

[6] J. Deng, W. Dong, R. Socher, L.J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, Ieee, 2009, pp. 248–255.

[7] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, A.C. Berg, Imagenet large scale visual recognition challenge, Int. J. Comput. Vis. 115 (3) (2015) 211–252.

[8] X. Chen, H. Fang, T.Y. Lin, R. Vedantam, S. Gupta, P. Dollár, C.L. Zitnick, Microsoft coco captions: Data collection and evaluation server, 2015, arXiv preprint arXiv:1504.00325.

[9] S. Hwang, J. Park, N. Kim, Y. Choi, I.So. Kweon, Multispectral pedestrian detection: Benchmark dataset and baseline, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1037–1045.

[10] I. Goodfellow, NIPS 2016 tutorial: Generative adversarial networks, 2016, arXiv preprint arXiv:1701.00160.
[11] E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, T. Darrell, Deep domain confusion: Maximizing for domain invariance, 2014, arXiv preprint arXiv:1412.3474.
[12] C. Chu, A. Zhmoginov, M. Sandler, Cyclegan, a master of steganography, 2017, arXiv preprint arXiv:1712.02950.
[13] Z.Q. Zhao, P. Zheng, S.T. Xu, X. Wu, Object detection with deep learning: A review, IEEE Trans. Neural Netw. Learn. Syst. 30 (11) (2019) 3212–3232.
[14] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, 2014, arXiv preprint arXiv:1409.1556.
[15] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[16] M.P. SManda, H.S. Kim, A fast image thresholding algorithm for infrared images based on histogram approximation and circuit theory, Algorithms 13 (9) (2020) 207.
[17] J. Baek, S. Hong, J. Kim, E. Kim, Efficient pedestrian detection at nighttime using a thermal camera, Sensors 17 (8) (2017) 1850.
[18] W. Li, D. Zheng, T. Zhao, M. Yang, An effective approach to pedestrian detection in thermal imagery, IEEE, 2012, pp. 325–329,
[19] Y. Ji, H. Zhang, Z. Zhang, M. Liu, CNN-based encoder–decoder networks for salient object detection: A comprehensive review and recent advances, Inform. Sci. 546 (2021) 835–857.
[20] J. Liu, S. Zhang, S. Wang, D.N. Metaxas, Multispectral deep neural networks for pedestrian detection, 2016, arXiv preprint arXiv:1611.02644.
[21] J. Wagner, V. Fischer, M. Herman, S. Behnke, Multispectral pedestrian detection using deep fusion convolutional neural networks, in: ESANN, 2016.
[22] M. Vandersteegen, K.Van. Beeck, T. Goedemé, Real-time multispectral pedestrian detection with a single-pass deep neural network, in: International Conference Image Analysis and Recognition, Springer, Cham, 2018, pp. 419–426.
[23] J. Chen, X. Li, L. Luo, X. Mei, J. Ma, Infrared and visible image fusion based on target-enhanced multiscale transform decomposition, Inform. Sci. 508 (2020) 64–78.
[24] Zhang Changchun, Qingjie Zhao, Yu Wang, Transferable attention networks for adversarial domain adaptation, Inform. Sci. 539 (2020) 422–433.
[25] A. Rame, E. Garreau, H. Ben-Younes, C. Ollion, OMNIA faster R-CNN: Detection in the wild through dataset merging and soft distillation, 2018, arXiv preprint arXiv:1812.02611.
[26] Y. Chen, W. Li, C. Sakaridis, D. Dai, L. Van Gool, Domain adaptive faster r-cnn for object detection in the wild, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 3339–3348.
[27] A.L. Rodriguez, K. Mikolajczyk, Domain adaptation for object detection via style consistency, 2019, arXiv preprint arXiv:1911.10033.
[28] F. Yu, D. Wang, Y. Chen, N. Karianakis, P. Yu, D. Lymberopoulos, X. Chen, Unsupervised domain adaptation for object detection via cross-domain semi-supervised learning, 2019, arXiv preprint arXiv:1911.07158.
[29] J. Tang, Y. He, Y. Tian, D. Liu, G. Kou, F.E. Alsaadi, Coupling loss and self-used privileged information guided multi-view transfer learning, Inform. Sci. 551 (2021) 245–269.
[30] S.J. Pan, Q. Yang, A survey on transfer learning, IEEE Trans. Knowl. Data Eng. 22 (10) (2009) 1345–1359.
[31] J.J. Lim, R.R. Salakhutdinov, A. Torralba, Transfer learning by borrowing examples for multiclass object detection, in: Advances in Neural Information Processing Systems, 2011, pp. 118–126.
[32] Y. Yao, G. Doretto, Boosting for transfer learning with multiple sources, in: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE, 2010, pp. 1855–1862.
[33] J. Ngiam, D. Peng, V. Vasudevan, S. Kornblith, Q.V. Le, R. Pang, Domain adaptive transfer learning with specialist models, 2018, arXiv preprint arXiv:1811.07056.
[34] A. Imran, V. Athitsos, Domain adaptive transfer learning on visual attention aware data augmentation for fine-grained visual categorization, 2020, arXiv preprint arXiv:2010.03071.
[35] Y. Xiao, H. Wang, B. Liu, A new transfer learning-based method for label proportions problem, Inform. Sci. 541 (2020) 391–408.
[36] L.A. Gatys, A.S. Ecker, M. Bethge, Image style transfer using convolutional neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2414–2423.
[37] Z. Zheng, J. Liu, P ^{2}-GAN: Efficient style transfer using single style image, 2020, arXiv preprint arXiv:2001.07466.
[38] A. Royer, K. Bousmalis, S. Gouws, F. Bertsch, I. Mosseri, F. Cole, K. Murphy, Xgan: Unsupervised image-to-image translation for many-to-many mappings, in: Domain Adaptation for Visual Understanding, Springer, Cham, 2020, pp. 33–49.
[39] H. Zhang, K. Dana, Multi-style generative network for real-time transfer, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018.
[40] A. Geiger, P. Lenz, R. Urtasun, Are we ready for autonomous driving? the kitti vision benchmark suite, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2012, pp. 3354–3361.
[41] P. Roy, S. Ghosh, S. Bhattacharya, U. Pal, Effects of degradations on deep neural network architectures, 2018, arXiv preprint arXiv:1807.10108.
[42] Q. Guo, W. Feng, C. Zhou, R. Huang, L. Wan, S. Wang, Learning dynamic siamese network for visual object tracking, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 1763–1771.
[43] P. Isola, J.Y. Zhu, T. Zhou, A.A. Efros, Image-to-image translation with conditional adversarial networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1125–1134.
[44] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L.C. Chen, Mobilenetv2: Inverted residuals and linear bottlenecks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 4510–4520.
[45] M. Tan, Q.V. Le, Efficientnet: Rethinking model scaling for convolutional neural networks, 2019, arXiv preprint arXiv:1905.11946.
[46] C. Devaguptapu, N. Akolekar, M.M. Sharma, V.N. Balasubramanian, Borrow from anywhere: Pseudo multi-modal object detection in thermal imagery, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2019.
[47] D. Ghose, S.M. Desai, S. Bhattacharya, D. Chakraborty, M. Fiterau, T. Rahman, Pedestrian detection in thermal images using saliency maps, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2019.
[48] K. Agrawal, A. Subramanian, Enhancing object detection in adverse conditions using thermal imaging, 2019, arXiv preprint arXiv:1909.13551.
[49] M. Krišto, M. Ivasic-Kos, M. Pobar, Thermal object detection in difficult weather conditions using YOLO, IEEE Access 8 (2020) 125459-125476.