SSTN: Self-Supervised Domain Adaptation Thermal Object Detection for Autonomous Driving

Farzeen Munir, Shoaib Azam and Moongu Jeon
School of Electrical Engineering and Computer Science
Gwangju Institute of Science and Technology
Gwangju, South Korea
Email: (farzeen.munir,shoaibazam, mgjeon)@gist.ac.kr

Abstract—The sensibility and sensitivity of the environment play a decisive role in the safe and secure operation of autonomous vehicles. This perception of the surrounding is way similar to human visual representation. The human's brain perceives the environment by utilizing different sensory channels and develop a view-invariant representation model. Keeping in this context, different exteroceptive sensors are deployed on the autonomous vehicle for perceiving the environment. The most common exteroceptive sensors are camera, Lidar and radar for autonomous vehicle's perception. Despite being these sensors have illustrated their benefit in the visible spectrum domain yet in the adverse weather conditions, for instance, at night, they have limited operation capability, which may lead to fatal accidents. In this work, we explore thermal object detection to model a view-invariant model representation by employing the self-supervised contrastive learning approach. For this purpose, we have proposed a deep neural network Self Supervised Thermal Network (SSTN) for learning the feature embedding to maximize the information between visible and infrared spectrum domain by contrastive learning, and later employing these learned feature representation for the thermal object detection using multi-scale encoder-decoder transformer network. The proposed method is extensively evaluated on the two publicly available datasets: the FLIR-ADAS dataset and the KAIST Multi-Spectral dataset. The experimental results illustrate the efficacy of the proposed method.

Index Terms—Self-supervised learning, Contrastive learning, Thermal object detection
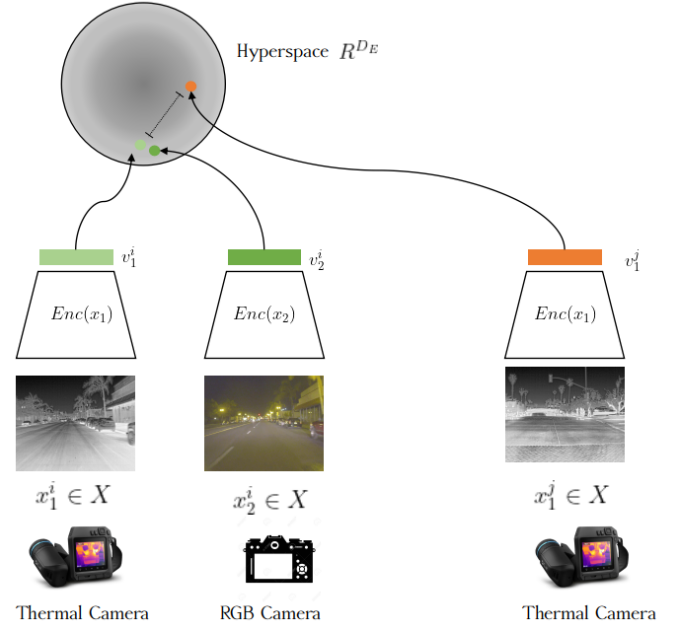
Fig. 1: Given a set of multi-sensor data, self-supervised contrastive learning learn representation to bring same scene together in embedding space (hyperspace), while pushing different scene apart .

I. INTRODUCTION

The recent technological advancements have enabled autonomous vehicles to be deployed on the road, ensuring the safety standard as indicated by SOTIF-ISO/PAS-21448 (Safety of the intended functionality)[1]. Keeping in this context, the perception of the autonomous vehicle has an integral element in defining the environment for the autonomous vehicle [1]. For the environment's perception, the most common exteroceptive sensors include cameras, Lidar, and radar. These sensors have their own merits and demerits; for instance, cameras (visible spectrum) provide an in-depth high resolution of the environment but have an illumination problem and cannot be utilized in night conditions. Lidar uses laser light for modelling the environment and provides the 3D point-cloud data of the environment. Despite providing the 3D environment, Lidar is extremely expensive and has a resolution problem in adverse weather conditions. Radars also enable the autonomous vehicle to identify small objects near the vehicle but have a low resolution at the range. These sensors have limited operating capability in perceiving the environment for the autonomous vehicle at night [2]. The utilization of a thermal camera in the autonomous vehicle's sensor suite provides the necessary solution for the perception of the environment at night yet requires an efficient perception algorithm in term of object detection [3].

Object detection is an essential component that forms the basis for the perception of the autonomous vehicle. Although the use of the deep neural network has illustrated the significant improvement in object detection, yet most of the efforts have focused on the visible spectrum images that include YOLO [4], SSD [5], Faster-RCNN [6], RefineDet [7], M2Det [8]. There is still room for improvement in thermal object

---

[1]https://www.daimler.com/innovation/case/autonomous/safety-first-for-automated-driving-2.htm

detection compared to its counterpart, visible spectrum object detection. Some of the literature work focuses on thermal object detection by employing the domain adaptation techniques to transfer the knowledge from the visible spectrum to the thermal domain for object detection in the thermal domain. In addition, the visible spectrum domain object detectors have also used on the thermal images for object detection. All these approaches for thermal object detection have utilized the supervised learning mechanism for feature representation. The abundance of unlabelled dataset leverage the feature representation by employing the self-supervised learning techniques as a surrogate to supervised learning for the labelled dataset.

Learning good representation without human supervision is a fundamental problem in computer vision. The two main classes that tend to illustrate the possible solution are generative and discriminative. Generative approaches in the form of auto-encoder [9] and generative adversarial network model [10] the pixels in the input space that leads to the computationally expensive representation learning for feature extraction [11] [12]. Besides, the discriminative approaches utilize the objective function for learning the representation by designing a pretext task to train the deep neural network [13]. This framework of learning the representation using the pretext task confide on the heuristics of designing the pretext task that limits the generalization of learned representations. Contrastive learning-based discriminative approaches are employed for learning the representation in the latent space to overcome the heuristics of pretext tasks [14] [15].

The conjecture of this work relies on the hypothesis that the view-invariant representations are encoded by the human brain [16] [17]. Humans view the surrounding environment with different sensory modalities. These sensory modalities are incomplete and noisy, but the prominent factor about the environment, for instance, geometry, semantics, and physics, are shared between these sensory modalities, illustrate the powerful model representation invariant to different views [18]. This study explores this hypothesis for the learning representations that capture information shared between the visible spectrum and thermal domain. For this purpose, we have employed the contrastive learning approach to learn the features embedding in the latent space projecting the distance (normally measured as Euclidean distance) to nearby points for the same scene corresponds to two sensory domain and far apart in the context of different views. Fig. 1 illustrates the pictorial overview of our proposed framework. The proposed neural network SSTN is designed in a two-stage network. The first stage, self-supervised contrastive learning, corresponds to the feature representation by constrative learning and maximizing the mutual information between the two domains and transferring the domain knowledge of visible spectrum to the thermal domain in a self-supervised manner for the object detection in the thermal domain. The later stage illustrates the thermal object detector's design consisting of a multi-scale encoder-decoder transformer network architecture that incorporates the features embedding from the self-supervised contrastive learning stage. The increment of the mean average

score of the proposed method with ResNet101 in comparison to other state-of-art methods is 2.57% for FLIR-ADAS and 2.37% for the KAIST Multi-Spectral dataset.

The main contributions of our work are:

1) The proposed work SSTN network have illustrated the utilization of a self-supervised contrastive learning approach to maximize the information between visible and thermal domain for the object detection in the thermal domain.
2) The contrastive learning approach is incorporated to cater to the scarcity of labelled dataset and has learned the feature representation in a self-supervised manner.
3) In addition, the proposed work, to the best of our knowledge, is the first work to incorporate self-supervised contrastive learning in multi-sensor data for
4) Further, we have extended the feature embedding learned in a self-supervised manner to a multi-scale encoder-decoder transformer network for thermal object detection.
5) We have demonstrated the efficacy of the proposed network (best model) on two public datasets, FLIR-ADAS and KAIST Multi-Spectral dataset, which illustrates the mean average precision score of 2.57% for FLIR-ADAS and 2.37% for KAIST Multi-Spectral dataset for thermal object detection in comparison to other state-of-art methods.

The rest of the paper is organized as follows: Section II gives related work. Section III explains the proposed methodology. The experimentation and results are discussed in Section IV, and section V concludes the paper.

II. RELATED WORK

Deep neural networks have been used as a function approximator for predicting and classifying objects in the visible and infrared spectrum domain. In literature, extensive research has carried out on designing the robust object detector for the visible spectrum domain [4] [8] [6] [5]. For thermal object detection, the research focuses on employing feature engineering by fusing the information. [19] have used visible spectrum domain object detector YOLOv3 for the person detection in the thermal domain and have benchmarked the performance of YOLOv3 with Faster-RCNN, SSD, and Cascade R-CNN. [20] illustrates the thermal object detection problem by augmenting the thermal image frame with the corresponding saliency maps to provide an attention mechanism for pedestrian detection in the thermal domain using Faster-RCNN. The multispectral images (thermal, near-infrared (NIR)) are fused and used to train the YOLOv3 for thermal object detection. Besides deep neural network, classical image processing approaches are also applied for thermal object detection [21]. [22] [23] have applied the HOG and local binary pattern for feature extraction and trained the support vector machine (SVM) classifiers for the object detection in the thermal domain. Much research has also been done by incorporating the domain knowledge transfer between the visible and infrared domain. [24] have employed the generative adversarial approach for modelling
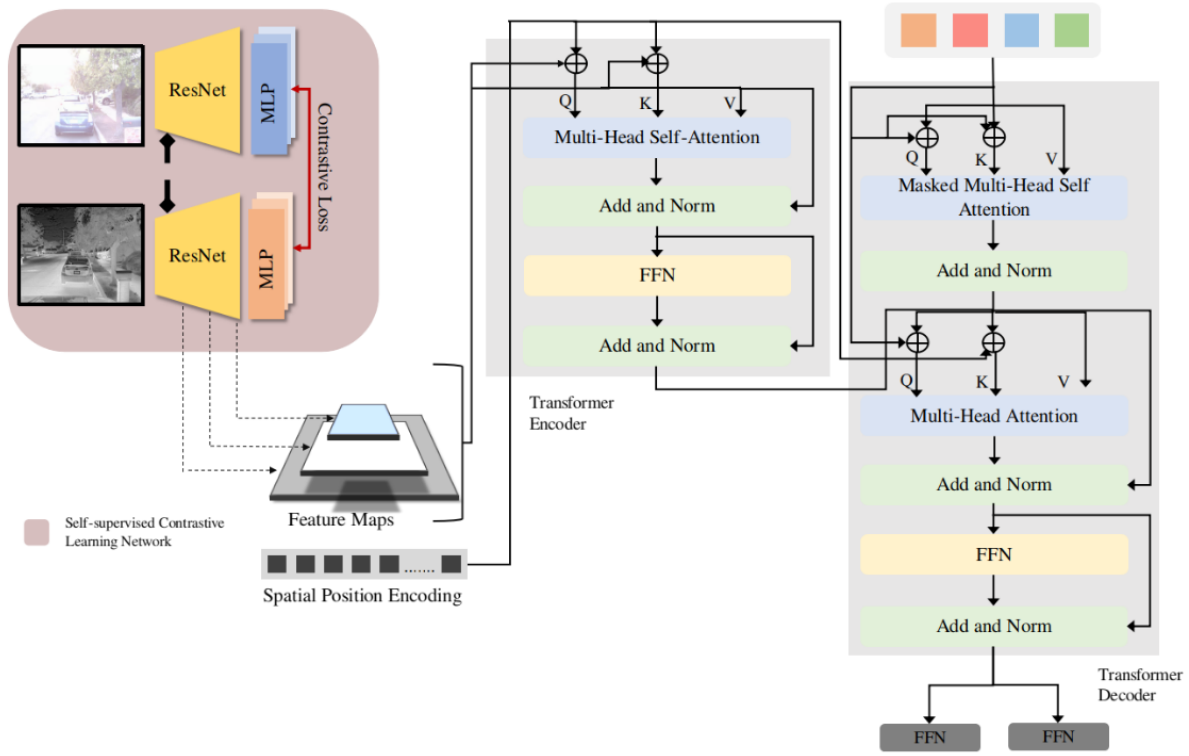
Fig. 2: The overall framework for the proposed method. It includes a self-supervised contrastive learning network and multi-scale feature embedding for the transformer encoder-decoder network for the thermal object detection.

the thermal image from the visible RGB image. This generated thermal image is utilized for training the variant of Faster-RCNN to detect the object in both RGB and thermal images. Similarly, a cross-domain semi-supervised learning framework is proposed for feature representation in the target domain [25]. In addition, researchers have also illustrated the use case of transfer learning between domain using generative adversarial networks [26] [27].

Self-supervised learning provides a way to learn the features embedding without explicit availability of labelled dataset. This representation learning in the context of image domain uses predictive approaches to learn the feature embedding by devising the artificially handcrafted pretext tasks [28] [29] [30] [31]. Although these pretext tasks illustrate the promising feature representations, yet these approaches depend on the heuristics of designing the pretext tasks. Another approach to learn feature embedding is to use the contrastive learning method. [32] have proposed an approach to learn the representation by contrasting the negative and positive pairs. Similarly, to improve the computation efficacy, a method employing the memory bank to store the instance class representation is proposed in this study [33]. This work is further investigated, and the memory bank method is replaced by in-batch negative sampled sampling [34] [35]. In addition, [14] have proposed a simple framework to learn the feature representation through contrastive learning irrespective of memory bank or specialized architectures. This framework explores data augmentation

techniques, a learnable nonlinear transformation layer between representation and contrastive, representation learning with contrastive cross-entropy loss. [36] have explored the contrastive predictive coding and proposed contrastive multiview coding for the representation learning between the different view of the same scene by augmenting them in color channel space. Similarly, a supervised contrastive loss is introduced in this work [37].

III. PROPOSED METHOD

Learning effective visual representation without labelled data is a challenging problem. The goal is to learn representations that capture useful features for deep-learning-based object detection. In this work, we explicitly explore the co-occurrence of information in multi-sensor data that is obtained from the thermal image and RGB image. A two-stage deep neural network, Self-supervised thermal network (SSTN), is proposed, where the first stage consists of pre-training involving self-supervised representation learning by contrastive learning approach. The second stage consists of a multi-scale encoder-decoder transformer network for object detection in the thermal domain based on features learned in the previous stage.

A. Self-supervised Contrastive Learning

The thermal data captures the surrounding information using the heat signature of different objects and exhibits large variations. Thermal images require a robust representation learning

mechanism. Therefore, we have adopted a multi-spectrum self-supervised contrastive learning for feature representation. Fig. 2 shows the self-supervised contrastive learning network. Let x_1 and x_2 represent a set X consisting of the same scene's thermal and RGB image. A stochastic data augmentation module transforms the input X to two correlated images. A neural network encoder maps X to representation vector, $v = Enc(X) \in R^{D_E}$, and is normalized to unit hyperspace in R^{D_E} [37], [14]. Both x_1 and x_2 are separately fed to the same encoder network to obtain a pair of representation vectors. The ResNet network is adopted as an encoder network [38], due to its applicability and common usage. We instantiate a multi-layer perceptron as a projection network that maps v to a vector $w = Proj(v) \in R^{D_E}$ [39]. The multi-layer perceptron consists of a single hidden layer with a size of 2048 and an output layer of size $F_p = 128$. The output vector is also normalized to lie in the unit hypersphere, facilitating the calculation of the inner product in projection space [36].

For the formulation of the contrastive loss, a set of N random pairs of thermal and RGB images $\{X_k\}k = 1....N$, are selected. The corresponding batch size $\{X_l\}l = 1....2N$ for this configuration is $2N$, to be fed for training the network. x'_{2k} and x'_{2k-1} represent thermal and RGB sample pair respectively of X_k. In a multi-spectrum batch, $i \in I \equiv \{1...2N\}$ is an index of the randomly selected thermal sample, and $j(i)$ is an index of its corresponding RGB sample. The Eq.1 gives the self-supervised contrastive learning loss function [36] [40].

$$L^{self} = \sum_{i \in I} L_i^{self} = -\sum_{i \in I} log \frac{exp(w_i \cdot w_{j(i)}/\tau)}{\sum_{b \in B(i)} exp(w_i \cdot w_b/\tau)}, \quad (1)$$

Here, $w_l = proj(Enc(X'_l)) \in R^{D_E}$, $\tau \in R^+$ is a scalar parameter for temperature, $\cdot$ represents the inner product of the two vectors, and $B(i) \equiv I \setminus \{i\}$. i represent index of the anchor and $j(i)$ is called a positive in the batch, while other $2(N - 1)$ indexes ($\{k \in B(i) \setminus \{j(i)\}\}$) are negatives. To clarify each anchor i have one positive pair and $2(N - 1)$ negative pairs.

B. Encoder-Decoder Transformer Network

1) Encoder: A transformer encoder inputs low-resolution feature maps which is obtained from convolution neural network backbone. In this work, a multi-scale feature maps $\{f^l\}_{l=1}^{L-1}$ are obtained from the self-supervised contrastive learning stage, where $f^l \in \mathbb{R}^{C \times H_l \times W_l}$. The multi-scale feature maps $\{f^l\}_{l=1}^{L-1}$, where $L = 4$ are extracted from pre-training stages C_3 to C_5 [2] in the ResNet encoder [38]. The feature maps are transformed by 1×1 convolution $\hat{f} \in \mathbb{R}^{C \times HW}$. The C_5 feature map is obtained by performing 3×3 and stride 2 convolution. All multi-scale feature maps are of $C = 256$ channels. The encoder inputs and outputs the same resolution feature maps. Every transformer encoder layer consists of a multi-scale, multi-head self-attention module [41] [42]. ς_q and ς_k denotes a set of query and key elements.

Let $q \in \varsigma_q$ represent a query element with representation features $z_q \in \mathbb{R}^C$ and $k \in \varsigma_k$ represents a key element with feature embedding $f_k \in \mathbb{R}^C$. The multi-head, multi-scale self-attention features are calculated by

$$(z_q, \{f^l\}_{l=1}^L) = \sum_{m=1} W_m [\sum_{l=1}^L \sum_{k \in \varsigma_k} A_{mlqk} \cdot W'_m f^l], \quad (2)$$

where m represents the number of attention head, l shows the input feature maps levels. $W_m \in \mathbb{R}^{C_v \times C}$ and $W'm \in \mathbb{R}^{C \times C_v}$ represent weight tensor which are learned during the training. ($C_v = C/M$). Attention weight tensor $A_{mlqk} \propto exp\{\frac{z_q^T U_m^T V_m f_k}{\sqrt{C_v}}\}$ is normalized by $\sum_{l=1}^L \sum_{k \in \varsigma_k} A_{mlqk} = 1$, here U_m and V_m are learnable weight tensors. The spatial position encoding are learned and shared among all the attention layers of an encoder for a certain query, key and feature map pair [43].

2) Decoder: The Transformer decoder mimics the encoder framework of sub-layers. The layers consist of multi-scale cross-attention and self-attention modules. Each type of attention module input object queries. The object queries are initially set to zero, and N object queries are learnt positional encodings and encoder memory. In the case of the cross-attention module, object queries derive features from multi-scale feature maps, and output feature maps from the encoder are the key elements. Moreover, in the self-attention module, attention is computed amid object queries. The decoder transforms the N object queries into output embeddings. The output embeddings are then fed to the feed-forward network, which computes the bounding box coordinates and class labels.

3) Feed-Forward Network: The decoder's output embedding is given to the feed-forward network, which consists of 3 layer perceptron network with RELU activation and a hidden dimension of k. A fully connected layer is used as the final layer, which performs the linear projection to predict the output. The output consists of class labels and the bounding box coordinates comprising height, width and the normalized centre coordinates. Moreover, a no-class label is also predicted for the images with no object present in the image.

IV. EXPERIMENTATION AND RESULTS

A. Dataset

In this study, we have used two publicly available datasets i.e. FLIR-ADAS dataset[3], and the KAIST Multi-Spectral dataset [44]. FLIR-ADAS dataset contains 9214 thermal and RGB image pair, and objects are annotated in MS-CoCo format, i.e. a bounding box and class labels. The aforementioned dataset has 3 class labels, car, person and bicycle. The data is collected using the FLIR Tau2 camera and each image has the resolution of 640×512. Both day and night time data is collected. We have used a standard split of the dataset in training and testing as illustrated in Table I.

The KAIST Multi-Spectral dataset has 95000 image pair of thermal and RGB image [44]. However, in this dataset only

---

[2]The convolution blocks as explained in [38]

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

TABLE I: FLIR-ADAS and KAIST Multi-Spectral datasets partition topology for training and testing the proposed SSTN network.

Dataset	Total Images	Train Images	Test Images
FLIR-ADAS	10228	8862	1366
KAIST Multi-Spectral	95000	76000	19000

the person class is annotated. KAIST Multi-Spectral dataset is collected using FLIR A35 camera with a resolution of 340×256 and include day and night time images. We have used the standard split of the dataset for training and testing as illustrated in Table I.

B. Evaluation Metric

In this study, we have used the standard MS COCO evaluation metric [45]. Intersection over union (IoU) is computed between the area covered by the ground-truth bounding box and the area covered by the predicted bounding box, given by Eq.3.

$$IoU = \frac{A_p \cap A_{gt}}{A_p \cup A_{gt}}, \tag{3}$$

The True Positive (TP) is held true for the $IoU > thresh$ and False Positive (FP) is considered when $IoU < thresh$. Based on TP and FP precision and Recall are calculated shown below. Recall and Precision are calculated using Eq.4.

$$Recall = \frac{TP}{TP + FN}, Precision = \frac{TP}{TP + FP}, \tag{4}$$

For each class, Average precision (AP) is computed. AP is the area under the PR curve, which is computed using the Eq.5.

$$AP[class] = \frac{1}{\#thresh} \sum_{IoU \in thresh} AP[class, IoU], \tag{5}$$

MS COCO evaluation metric varies the threshold ($thresh$) from 0.5 to 0.95 with a incremnet of 0.05. In this study $mAP_{IoU} = 0.5$ is considered for all the experiments.

C. Experimentation

This section explains the experimentation details of the proposed network (SSTN) and discusses the results.

1) Self-supervised Contrastive Learning: The self-supervised contrastive network is trained using thermal and RGB image pair. The input images undergo augmentations that include random crop and resize, horizontal flip and colour jitters. For this study, we have only considered ResNet50 and ResNet101 as the encoder for the self-supervised network. The network is trained for 1000 epoch with a batch size of 16, learning rate of 0.05 and temperature of 0.07. The network is implemented using Pytorch deep learning library, and the weights of the network are optimized using the loss function given by Eq.1. All the experimentation is performed on three GPU machine, where each GPU has a memory of 12Gb.

2) Faster-RCNN Baseline: The faster-RCNN with ResNet backbone is employed to develop a baseline for the proposed network [6]. Two types of experiments are performed. First, the network is trained in a supervised manner on thermal image data with no pre-training weights. Second, the network backbone is initialized by the features optimized by the self-supervised contrastive network and finetune on labelled thermal image data. We have adopted Pytorch implementation of Faster-RCNN for this purpose. The input data is augmented with a random scale, randomly crop and resize and horizontally flipped. It is noted that input image augmentation is kept the same for all the experiments conducted with Faster-RCNN and multi-scale encoder-decoder transformer network. The network is trained with a stochastic gradient descent optimizer with a learning rate of 0.02

3) Encoder-Decoder Transformer Network: The transformer encoder-decoder network is trained using an AdamW [46] optimizer with an initial learning rate of $2e^{-4}$, the backbone learning rate of $2e^{-5}$ and weight decay of $1e^-4$. The batch size is set to 2, and the network is trained for 100 epochs. The multi-scale feature maps are extracted from a self-supervised contrastive learning network, as shown in Fig. 2. For the transformer number of encoder and decoder are equal to 6. Absolute position encodings are considered a function of sine and cosine at different frequencies and then concatenated to achieve the final position encoding across the C channels. It is to be noted that in all the experimentation, the hyper-parameters values are determined heuristically.

An optimal bipartite matching scheme is selected for loss calculation inspired by [43]. The proposed method outputs a set of N predictions. Eq.6 defines the bipartite matching loss between a set of ground-truth (y) and a set of the predicted labels $\hat{y}.S_N$ illustrates the permutations of N elements. $L_{match}(y_j, \hat{y}_{\eta(j)})$ corresponds to pair-wise matching of predicted and ground-truth labels. A Hungarian algorithm is used to compute assignments between ground-truth and predictions. The matching cost aggregates both the class labels and the bounding box between the predictions and the ground-truth. Lets element j indexes the ground-truth set $y_j = (c_j, b_j)$, where c_j is the class labels and $b_j \in [0, 1]^4$ is normalized bounding box vector including center coordinates, width and height. The predicted class probability is define by $\hat{p}_{\eta j}(c_j)$ and bounding box is given by $\hat{b}_{\eta j}$. To match the prediction and ground-truth set, direct one-to-one correspondence is found without any duplicates. The Hungarian loss function is illustrated by Eq.7.

$$\tilde{\eta} = \underset{\eta \in S_N}{argmin} \sum_{j}^{N} L_{match}(y_j, \hat{y}_{\eta(j)}), \tag{6}$$

S_N illustrates the permutations of N elements. $L_{match}(y_j, \hat{y}_{\eta(j)})$ corresponds to pair-wise matching of predicted and ground-truth labels. A Hungarian algorithm is used to compute assignments between ground-truth and predictions. The matching cost aggregates both the class labels and the bounding box between the predictions

TABLE II: mAP score on FLIR-ADAS and KAIST Multi-spectral dataset.

Dataset	FLIR-ADAS	KAIST Multi-Spectral
Methods	mAP Score	mAP Score
Faster-RCNN +ResNet50 Backbone	56.70	59.02
Faster-RCNN+ ResNet50 w/ feature activation	62.45	65.61
Faster-RCNN +ResNet101 Backbone	59.03	64.31
Faster-RCNN+ ResNet101 w/ feature activation	64.34	67.52
Encoder-decoder transformer w/ ResNet50	68.77	62.08
Encoder-decoder transformer w/ ResNet101	72.03	68.25
Self-supervised thermal network (SSTN50)	75.19	70.62
Self-supervised thermal network (SSTN101)	**77.57**	**73.22**

TABLE III: mAP score of proposed framework with state-of-the-art methods. (*) represent average (day+night) mAP score. (-) shows that the specific algorithm is not tested on given dataset.

Datasets	FLIR-ADAS	KAIST Multi-Spectral
Methods	mAP Score	mAP Score
MMTOD-UNIT [24]	61.54	-
MMTOD-CG [24]	61.40	-
PiCA-Net [20]	-	65.80*
R^3Net [20]	-	70.85*
Intel [47]	23.70	-
tY [19]	75.00	63.00
(Our) Self-supervised Thermal Network (SSTN101)	**77.57**	**73.22**

and the ground-truth. Lets element j indexes the ground-truth set $y_j = (c_j, b_j)$, where c_j is the class labels and $b_j \in [0,1]^4$ is normalized bounding box vector including center coordinates, width and height. The predicted class probability is define by $\hat{p}_{\eta j}(c_j)$ and bounding box is given by $\hat{b}_{\eta j}$. To match the prediction and ground-truth set, direct one-to-one correspondence is found without any duplicates. The Hungarian loss function is illustrated by Eq.7.

$$L_{Hungarian}(y,\hat{y}) = \sum_{j=1}^{N}[-log\hat{p}_{\eta(j)}(c_j)+ \tag{7}$$
$$\mathbf{1}_{c_j \neq (\phi)}L_{box}(b_j, \hat{b}_{\hat{\eta}}(j))],$$

$\hat{\eta}$ indicate the optimal assignment. The $L_{box}(.)$ computes the score for the bounding box is represented as shown in Eq.8

$$L_{box}(b_j, \hat{b}_{\hat{\eta}}(j)) = \lambda_{iou}L_{CIoU}(b_j, \hat{b}_{\hat{\eta}}(j))+ \tag{8}$$
$$\lambda_{l1}\left\|b_j - \hat{b}_{\hat{\eta}}(j)\right\|,$$

where, l_1 loss with complete IoU loss [38] is used to calculate L_{CIoU}

$$L_{CIoU} = 1 - IoU + \frac{\rho(\mathbf{b}, \boldsymbol{b_{gt}})}{c^2} + \alpha\nu, \tag{9}$$

$1 - IoU + \frac{\rho(\mathbf{b}, \boldsymbol{b_{gt}})}{c^2}$ represents IoU distance with predicted bounding box b and ground-truth bounding box $b_g t$. c is the digonal length of smallest enlosing that covers the two boxes. ρ shows the euclidean distance. α is a positive trade-off parameter and ν illustrates the consistency of aspect ratio given by Eq.10 and Eq.11.

$$\alpha = \frac{\nu}{(1-IoU)+\nu}, \tag{10}$$

$$\nu = \frac{4}{\pi^2}(arctan\frac{w^{gt}}{h^{gt}} - arctan\frac{w}{h})^2, \tag{11}$$

For the training $\lambda_{l1} = 4$ and $\lambda_{iou} = 2$ are used. The object query is set to 300. Moreover, the experiment is also conducting with replacing self-supervised contrastive learning (pre-training stage) with ResNet backbone and trained using no pre-training weights and using the same parameter configuration.

Table II shows the quantitative evaluation of the proposed algorithms. The proposed framework is evaluated on FLIR-ADAS and KAIST Multi-Spectral dataset, using ResNet 50
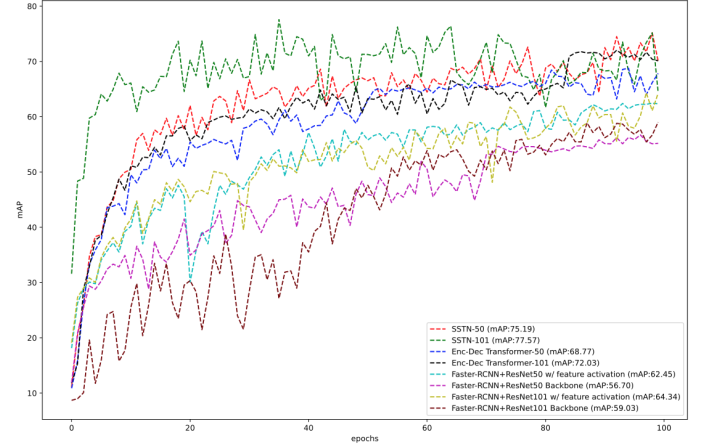

Fig. 3: The illustration of mAP scores for all the models trained and tested on FLIR-ADAS dataset. The SSTN ResNet-101 has the best results.

and ResNet101 as an encoder in the self-supervised contrastive learning stage. ResNet101 enable the self-supervised contrastive network to learn the more reliable representation of the data as the mAP score improve by 2.38% for FLIR-ADAS and 2.6% for KAIST Multi-Spectral dataset as compared to ResNet50. The efficacy of the self-supervised contrastive network is visible from the performance of SSTN with comparison to an encoder-decoder transformer with a ResNet backbone. An increase of 6.42% in mAP score in FLIR-ADAS and 8.54% mAP increase in KAIST Multi-Spectral with ResNet50 and an increase of 5.54% mAP score in FLIR-ADAS and 4.97% mAP improvement with ResNet101. A similar trend is visible in the baseline consisting of the Faster-RCNN network. Using the pre-train weights to finetune on thermal images, improve the accuracy by 5.75% in the FLIR-ADAS dataset and 6.59% in the KAIST Multi-Spectral dataset ResNet50 and 5.31% increase in FLIR-ADAS and 3.21% in KAIST Multi-Spectral with ResNet101. Fig.3 and Fig.4 illustrate the quantitative comparison of mAP score between the proposed best model SSTN-101 and other variants.

Moreover, The proposed algorithm is compared with other-state-of-the art algorithm, as shown in Table III. The Self-supervised thermal network outperforms the existing algorithm by 2.57% for FLIR-ADAS and 2.37% for the KAIST Multi-
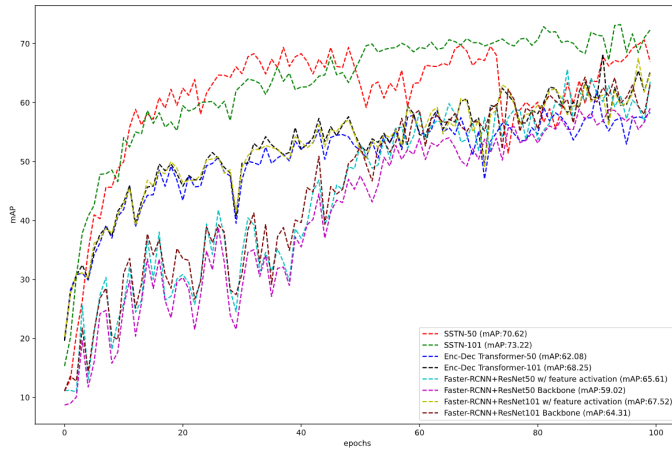
Fig. 4: The illustration of mAP scores for all the models trained and tested on the KAIST Multi-Spectral dataset. The SSTN ResNet-101 has the best results.

Spectral dataset. Fig.5 shows the qualitative results on FLIR-ADAS and KAIST Multi-Spectral dataset.

V. CONCLUSION

This study focuses on thermal object detection since thermal imagery is a fundamental tool for an autonomous vehicle. We have employed the self-supervised technique to learn enhanced feature representation using unlabelled data. A multi-scale encoder-decoder transformer network used these enhanced feature embedding to develop a robust thermal image object detector. The efficacy of a self-supervised thermal network is evaluated on FLIR-ADAS and KAIST Multi-Spectral datasets. The mean average precision of 77.57% is achieved on the FLIR-ADAS dataset and 73.22% is achieved on the KAIST Multi-Spectral dataset. In future work, we aim to fuse Lidar data with thermal image data to improve object detection in adverse environmental condition. Moreover, use the thermal data to understand and classify weather condition to optimize the perception system of the autonomous vehicle.

REFERENCES

[1] F. Rosique, P. J. Navarro, C. Fernández, and A. Padilla, "A systematic review of perception system and simulators for autonomous vehicles research," *Sensors*, vol. 19, no. 3, p. 648, 2019.

[2] J. Van Brummelen, M. O'Brien, D. Gruyer, and H. Najjaran, "Autonomous vehicle perception: The technology of today and tomorrow," *Transportation research part C: emerging technologies*, vol. 89, pp. 384–406, 2018.

[3] F. S. Leira, H. H. Helgesen, T. A. Johansen, and T. I. Fossen, "Object detection, recognition, and tracking from uavs using a thermal camera," *Journal of Field Robotics*, vol. 38, no. 2, pp. 242–267, 2021.

[4] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[5] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European conference on computer vision*. Springer, 2016, pp. 21–37.

[6] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster r-cnn: towards real-time object detection with region proposal networks. corr abs/1506.01497 (2015)," *arXiv preprint arXiv:1506.01497*, 2015.

[7] S. Zhang, L. Wen, X. Bian, Z. Lei, and S. Z. Li, "Single-shot refinement neural network for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4203–4212.

[8] Q. Zhao, T. Sheng, Y. Wang, Z. Tang, Y. Chen, L. Cai, and H. Ling, "M2det: A single-shot object detector based on multi-level feature pyramid network," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 9259–9266.

[9] R. Salakhutdinov and G. Hinton, "Deep boltzmann machines," in *Artificial intelligence and statistics*. PMLR, 2009, pp. 448–455.

[10] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *arXiv preprint arXiv:1406.2661*, 2014.

[11] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[12] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[13] S. Gidaris, P. Singh, and N. Komodakis, "Unsupervised representation learning by predicting image rotations," *arXiv preprint arXiv:1803.07728*, 2018.

[14] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.

[15] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[16] H. E. Den Ouden, P. Kok, and F. P. De Lange, "How prediction errors shape perception, attention, and motivation," *Frontiers in psychology*, vol. 3, p. 548, 2012.

[17] J. Hohwy, *The predictive mind*. Oxford University Press, 2013.

[18] L. Smith and M. Gasser, "The development of embodied cognition: Six lessons from babies," *Artificial life*, vol. 11, no. 1-2, pp. 13–29, 2005.

[19] M. Krišto, M. Ivasic-Kos, and M. Pobar, "Thermal object detection in difficult weather conditions using yolo," *IEEE Access*, vol. 8, pp. 125 459–125 476, 2020.

[20] D. Ghose, S. M. Desai, S. Bhattacharya, D. Chakraborty, M. Fiterau, and T. Rahman, "Pedestrian detection in thermal images using saliency maps," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 0–0.

[21] R. Soundrapandiyan and P. C. Mouli, "Adaptive pedestrian detection in infrared images using background subtraction and local thresholding," *Procedia Computer Science*, vol. 58, pp. 706–713, 2015.

[22] J. Baek, S. Hong, J. Kim, and E. Kim, "Efficient pedestrian detection at nighttime using a thermal camera," *Sensors*, vol. 17, no. 8, p. 1850, 2017.

[23] W. Li, D. Zheng, T. Zhao, and M. Yang, "An effective approach to pedestrian detection in thermal imagery," in *2012 8th International Conference on Natural Computation*. IEEE, 2012, pp. 325–329.

[24] C. Devaguptapu, N. Akolekar, M. M Sharma, and V. N Balasubramanian, "Borrow from anywhere: Pseudo multi-modal object detection in thermal imagery," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 0–0.

[25] F. Yu, D. Wang, Y. Chen, N. Karianakis, P. Yu, D. Lymberopoulos, and X. Chen, "Unsupervised domain adaptation for object detection via cross-domain semi-supervised learning," *arXiv preprint arXiv:1911.07158*, 2019.

[26] Z. Zheng and J. Liu, "P2-gan: Efficient style transfer using single style image," *arXiv preprint arXiv:2001.07466*, 2020.

[27] A. Royer, K. Bousmalis, S. Gouws, F. Bertsch, I. Mosseri, F. Cole, and K. Murphy, "Xgan: Unsupervised image-to-image translation for many-to-many mappings," in *Domain Adaptation for Visual Understanding*. Springer, 2020, pp. 33–49.

[28] C. Doersch, A. Gupta, and A. A. Efros, "Unsupervised visual representation learning by context prediction," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1422–1430.

[29] M. Noroozi and P. Favaro, "Unsupervised learning of visual representations by solving jigsaw puzzles," in *European conference on computer vision*. Springer, 2016, pp. 69–84.

[30] R. Zhang, P. Isola, and A. A. Efros, "Colorful image colorization," in *European conference on computer vision*. Springer, 2016, pp. 649–666.

[31] T. Chen, X. Zhai, M. Ritter, M. Lucic, and N. Houlsby, "Self-supervised gans via auxiliary rotation loss," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12 154–12 163.

[32] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 2. IEEE, 2006, pp. 1735–1742.

[33] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, "Unsupervised feature learning via non-parametric instance discrimination," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3733–3742.

[34] X. Ji, J. F. Henriques, and A. Vedaldi, "Invariant information clustering for unsupervised image classification and segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9865–9874.

[35] M. Ye, X. Zhang, P. C. Yuen, and S.-F. Chang, "Unsupervised embedding learning via invariant and spreading instance feature," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6210–6219.

[36] Y. Tian, D. Krishnan, and P. Isola, "Contrastive multiview coding," *arXiv preprint arXiv:1906.05849*, 2019.

[37] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," *arXiv preprint arXiv:2004.11362*, 2020.

[38] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[39] T. Hastie, R. Tibshirani, and J. Friedman, "The elements of statistical learning. springer series in statistics," in *:.* Springer, 2001.

[40] O. Henaff, "Data-efficient image recognition with contrastive predictive coding," in *International Conference on Machine Learning.* PMLR, 2020, pp. 4182–4192.

[41] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *arXiv preprint arXiv:1706.03762*, 2017.

[42] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable detr: Deformable transformers for end-to-end object detection," *arXiv preprint arXiv:2010.04159*, 2020.

[43] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European Conference on Computer Vision.* Springer, 2020, pp. 213–229.

[44] S. Hwang, J. Park, N. Kim, Y. Choi, and I. So Kweon, "Multispectral pedestrian detection: Benchmark dataset and baseline," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1037–1045.

[45] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision.* Springer, 2014, pp. 740–755.

[46] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[47] K. Agrawal and A. Subramanian, "Enhancing object detection in adverse conditions using thermal imaging," *arXiv preprint arXiv:1909.13551*, 2019.
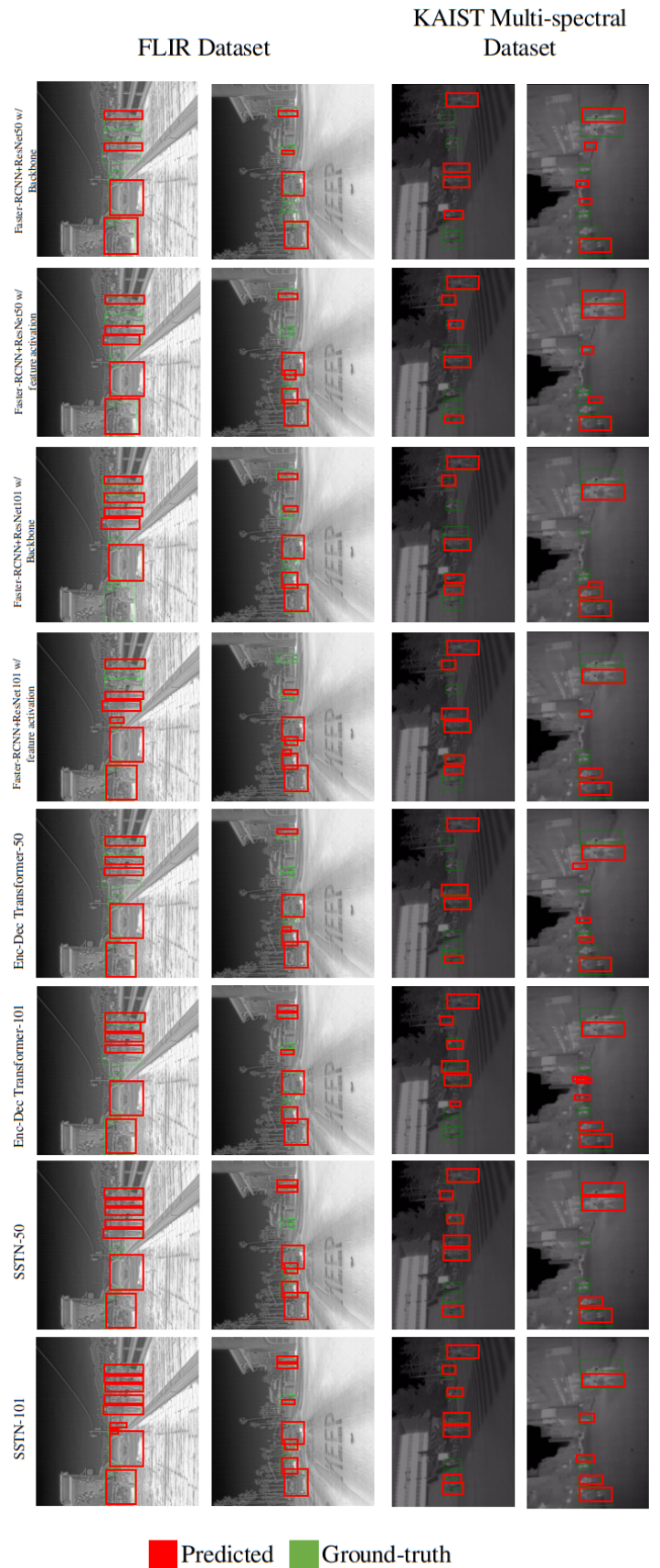
Fig. 5: The qualitative results of proposed thermal object detection using FLIR-ADAS and KAIST Multi-Spectral dataset.(Best viewed in color)