Automatic Target Recognition with Heterogeneous Sensor Platforms

1st Zijing Huang
Department of Agricutral & Biological Engineering
University of Florida
Gainesville, United States
zijing.huang@ufl.edu

***Abstract*—This paper introduces a pioneering algorithm for automatic target recognition, focusing on the role of contrastive learning in enhancing sensor fusion. The core of our research lies in demonstrating how supervised contrastive learning enables a sensor rich in information, such as an RGB camera, to augment the capabilities of other sensors, like an IR camera, particularly in classification tasks. By ensuring similar visibility conditions for both RGB and IR cameras, our approach capitalizes on the detailed data from the RGB sensor to improve the effectiveness of the IR sensor in target recognition. This technique not only boosts the accuracy and representativeness of the recognition system but also underscores the potential of contrastive learning as a pivotal tool in sensor fusion. The subsequent exploration of decision-level fusion through a Bayesian network approach further validates the ability of our algorithm to integrate disparate sensor data effectively. The promising results of our study highlight the algorithm's applicability in diverse sectors, ranging from industrial to military domains, revolutionizing the way multiple sensors collaborate for enhanced target recognition.**

***Index Terms*—Supervised Contrastive Learning, Information Fusion, Bayesian Network**

I. Introduction

Automatic Target Recognition (ATR) systems are utilized in a wide range of fields, such as autonomous driving, military operations, and daily life. ATR falls under the subfield of image processing and understanding and is categorized based on the sensor used. The primary goal of ATR is to increase speed, reliability, and performance by eliminating human intervention in target recognition. Early ATR systems processed analog inputs from infrared systems and used statistical methods for object-based processing and feature extraction. These systems were not programmable and exhibited inconsistent and unstable performance due to limited image data for training. However, the Knowledge-Based ATR system improved pattern recognition performance by optimizing the handling of the local structure while considering the context in a less optimal manner later. Roth's comprehensive survey [1] on primitive machine learning techniques highlights the rapid growth of this learning-based application due to advancements in machine learning and computing resources. The support vector machine (SVM) approach [2] gained popularity in the early 21st century. Researchers have been actively investigating this problem using various machine-learning algorithms for a considerable amount of time.

The field of SAR-ATR systems has recently witnessed significant progress through the adoption of Deep Learning (DL) techniques. Various studies have contributed to this progress. For example, Geng et al. [3] laid the groundwork for implementing DL-based SAR-ATR systems in open fields. Another study by Li et al. [4] focused on ensuring interpretability, recognition accuracy, and robustness against adversarial attacks. Zhang et al. [5] presented a model that addressed the problem of high computation cost and large memory requirements, making it suitable for deployment on resource-limited platforms like microsatellites. Additionally, Zou et al. [6] proposed a deep attention convolutional network to improve SAR target recognition accuracy under speckle corruption. These developments demonstrate the continuing advancement in the SAR-ATR field, with a focus on enhancing performance and efficiency via DL techniques. Although advanced machine learning techniques have demonstrated efficacy in various domains, they do not consider the potential consequences resulting from sensor failures, which is a limitation of these methods [7]. ATR systems require precision, and single-sensor data may not suffice in certain situations, such as low-light conditions where RGB sensors fail. Sensor fusion has gained significant interest from academia, industry, and the military to improve classification performance. However, implementing sensor fusion faces several challenges, including uncertainties in data acquisition from sensors with different modalities, real-time decision-making based on sensor observations, and dynamic changes in sensory observations and the environment [7]. Nonetheless, sensor fusion is similar to the human brain's processing of information from different sensory organs, and handling the data stream is crucial. Therefore, feasible configurations for diverse cases have been proposed to aid in a systematic approach [8].

Image-based fusion involves three key stages: early, middle, and late fusion. In the early fusion stage, pixel-level integration is paramount. This involves combining pixel information. For example, earth observation satellite data can be enhanced this way, as outlined in [9]. The middle stage of fusion then takes over, where the focus shifts to merging processed image features. Here, the data is represented as feature embeddings, and various mathematical transforms are applied to enhance the fusion performance, a process detailed in [10]. Finally, the late fusion stage leverages the probabilistic results derived

from the preceding stages. This stage is instrumental in making final classification decisions, significantly improving the accuracy of target and anomaly detection in satellite imagery, as demonstrated in studies [11], [12]. Each stage, building upon the last, plays a critical role in the comprehensive analysis of satellite data, from initial data integration to refined classification and detection.

To enhance the accuracy and representativeness of feature extraction, we utilize supervised Contrastive Learning (CL) [13] to propose a new image-fusion approach. Contrastive Learning, especially in its supervised form, has gained prominence in the field of deep learning for its effectiveness in learning robust representations. This approach is based on the principle of learning by comparison. It involves training a model to distinguish between similar (positive) and dissimilar (negative) pairs of data [14]. By doing so, the model learns to encode data points in a way that minimizes intra-class variance while maximizing inter-class variance, leading to more discriminative feature representations.

This methodology is particularly advantageous in scenarios with high-dimensional data from diverse sources. It enables models to understand nuanced differences and similarities within the data, which is crucial for tasks such as classification, object detection, and more [15]. Supervised CL extends this idea by using label information to guide the process, ensuring that the learned representations are not only distinct but also relevant to the specific task at hand [13].

Supervised CL is employed to handle the heterogeneous data obtained from different imaging sensors effectively in our work. While RGB cameras provide high-quality visual information under good lighting conditions, their performance diminishes in low-visibility scenarios. On the other hand, IR cameras can capture thermal images in various weather conditions but have limitations with certain materials. By integrating Supervised CL, we aim to develop a classifier that effectively leverages the complementary strengths of both RGB and IR sensors. This approach surpasses the capabilities of a single Convolutional Neural Network (CNN) encoder by creating representations that are robust to the specific weaknesses of each sensor type. As a result, our classifier, enhanced with CL, promises to deliver superior performance, particularly in challenging environmental conditions.

It's also crucial to highlight the unique strengths of Bayesian Networks (BN) in this context. Unlike their traditional counterparts, Bayesian inference provides a framework for understanding how prior beliefs can reverberate in the cortical hierarchy, corrupting sensory evidence and leading to bistable perception [16]. Moreover, the combination of perceptual expectations as prior probabilities with sensory likelihoods through Bayes' rule demonstrates the complex interplay between prior knowledge and sensory information in shaping perception [17]. This approach enables us to surpass the limitations of decision-level fusion and enhance the efficiency of inference in recognizing targets in uncertain conditions. We've done numerical analysis using synthetic data to prove it's efficiency in our problem. The pipeline of our work is shown in Fig. 1

II. Feature-Level Fusion

A. Contrastive Learning

Our contrastive network's architecture is detailed in Fig 2, 3. It is structured with dual encoders—one for infrared (IR) images and another for RGB images. Both encoders leverage CNN to transform raw sensor data into structured embeddings. The two sub-networks have near-identical architectures, with the primary distinction being the size of the input channels. In most Simiese-like networks, identical networks are preferred for weight-sharing purposes. However, in this instance, the objective is to maintain structural consistency. Consequently, all components, except the input, remain unchanged, preserving the inputs in their original form. An alternative approach could involve expanding the IR data to three channels aligning it with RGB data. However, this method has been disregarded to avoid the introduction of extraneous information, thereby retaining the data in its authentic format. All other parameters, including convolutional, pooling, fully connected layers, and classifier parameters, remain consistent across the sub-networks.

The dataset encompasses images from two types of sensors: RGB cameras and IR cameras. Despite their different modalities, both sensors provide image-based data that can be effectively processed using CNNs. Our dataset features three classes of targets: a toy tank, a toy missile launcher, and a null class (nothing exists in the image, and the data is generated by white noise), indicative of the absence of a target. Each class is represented in both RGB and IR formats, ensuring a comprehensive dataset that challenges the model to learn and adapt across modalities. The data is systematically divided into training, testing, and validation sets, as shown in the table below.

TABLE I: Data Distribution Across Different Categories and Sensor Types

	IR			RGB		
	Smerch	T72	Null	Smerch	T72	Null
Train	153	153	153	300	166	358
Test	32	36	34	66	36	77
Validation	34	34	34	63	35	76

The following configuration characterizes the implementation of our model:

- Epochs: 200
- Batch Size: 8
- Weight Decay: 5×10^{-5}
- Learning Rate: 0.001
- Momentum: 0.9

The network employs a Cosine Embedding Loss (*nn.CosineEmbeddingLoss*) from Pytorch [18], optimized using Stochastic Gradient Descent (SGD) [19] with a learning rate scheduler to adjust the rate at specific milestones.

After we have trained the respective encoders using contrastive learning, we freeze the encoders for both IR and

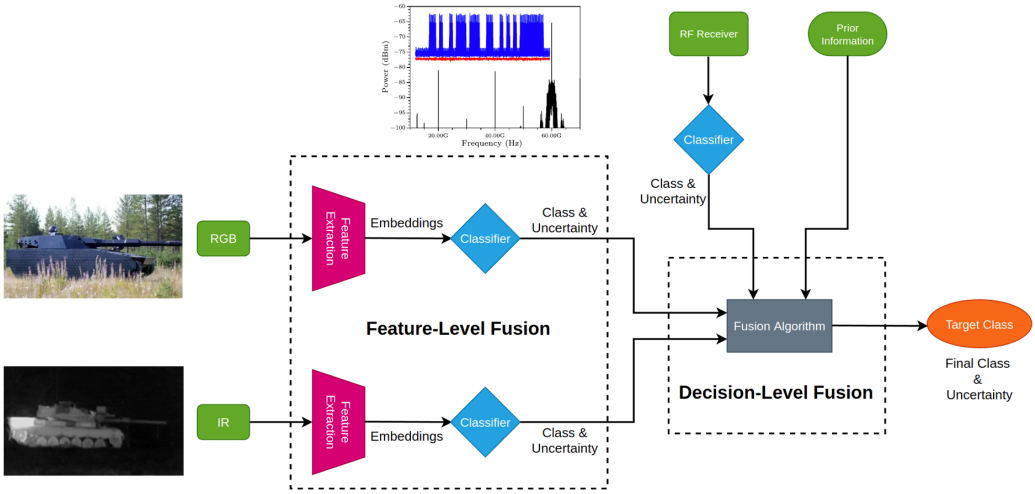


Fig. 1: Illustration of the integrated multi-sensor data fusion and classification pipeline. The system employs feature extraction from both RGB and IR sensors, which capture complementary visual and thermal data, respectively. Embeddings are then generated from each sensor stream and fed into a classifier to obtain class predictions along with uncertainty estimates. Feature-level fusion is applied to combine embeddings before classification, enhancing the representativeness of the features. The classification results from both the individual sensors and the fused features are then processed through a decision-level fusion algorithm. This algorithm incorporates prior information and classifies the target with an associated uncertainty, providing a comprehensive understanding of the target's class in various environmental conditions.

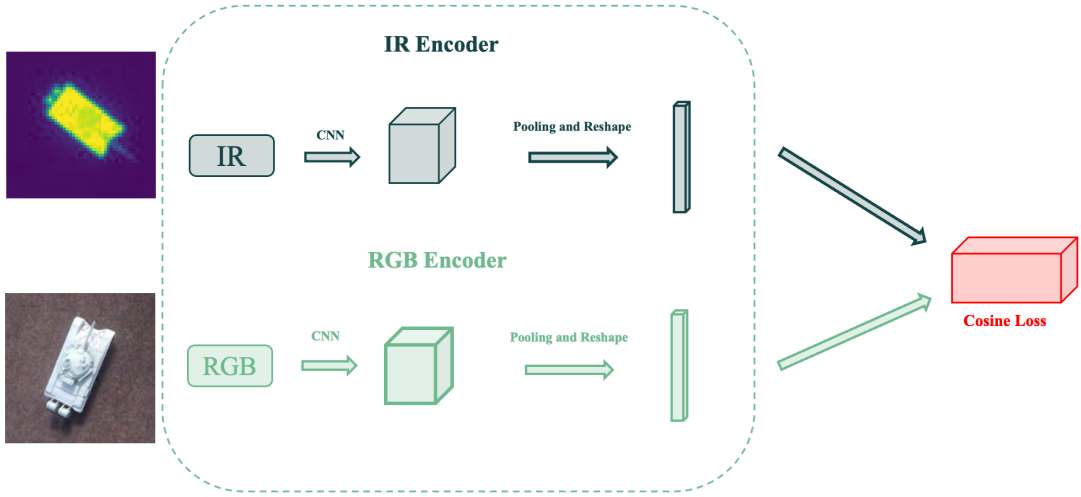


Fig. 2: Architecture of the Contrastive Learning Network for Multi-Sensor Data. The network comprises two parallel encoders: one for infrared (IR) data and another for RGB data. Each encoder consists of a CNN followed by pooling and reshaping operations, which transform the raw sensor data into a one-dimensional embedding. These embeddings are then used to calculate the cosine similarity loss. Not that all architectural components within the network are identical, with the exception of the input channel number.

RGB. We then use these frozen encoders to extract features for training individual classifiers. To investigate the role of contrastive learning in this process, we conducted two comparative experiments. Initially, we train two separate networks for IR and RGB using the same encoder and classifier architectures without employing any contrastive learning techniques, resulting in two distinct classification networks. We compare the outcomes of these networks with classifiers trained using contrastive learning.

Furthermore, we conducted another experiment where we swapped the classifiers trained on different modalities, meaning we applied the classifier trained on IR data to RGB

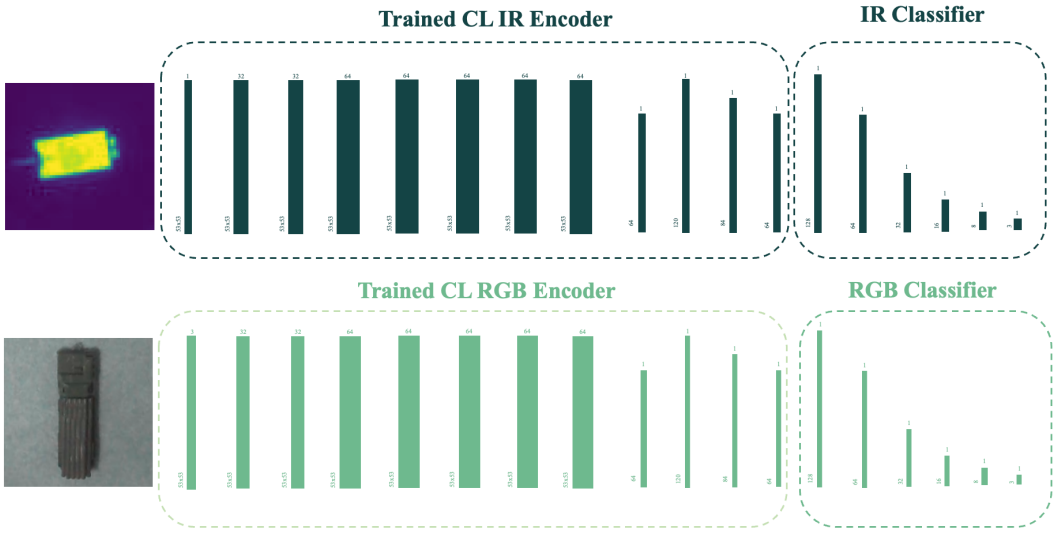


Fig. 3: Classifiers training encoders with contrastive learning

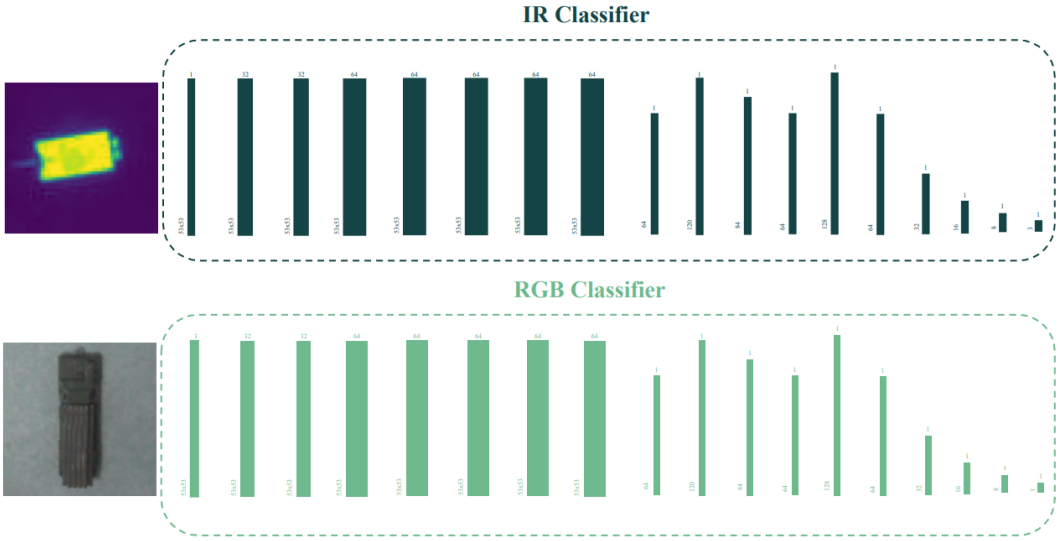


Fig. 4: Classifiers training encoders without contrastive learning

data and vice versa. Similar to the previous experiments, we evaluate the performance with and without the use of contrastive learning. The results of these experiments will be detailed in the following chapter.

III. Decision-Level Fusion

A. Bayesian Network

The creation of the Bayesian Network (BN) involved certain assumptions. Firstly, we neglected time dependency, which means that we did not account for the frequency at which sensors were being sampled. Secondly, the observation was obtained in a specific sequence.

The structure of BN is shown in Fig.1 [20], $Y \in \mathcal{Y}$ is the true target class where $\mathcal{Y} = \{y_{smerch}, y_{t72}, y_{null}\}$, Λ is the sensor type (Λ_{RGB} and Λ_{IR}), and $\hat{Y}$ is the estimated target class from the sensors $\hat{\mathcal{Y}} = \{\hat{y}_{smerch}, \hat{y}_{t72}, \hat{y}_{null}\}$. The classification will be recursively updated for each measurement starting from the predefined prior probability. The conditional probability is necessary for updating, and from the chain rule, we can get the following:

$$p(\hat{Y}, Y, \Lambda) = p(Y|\hat{Y}, \Lambda)p(\Lambda|\hat{Y})p(\hat{Y}) \tag{1}$$

Since the BN is directionally connected:

$$p(\Lambda|\hat{Y}) = p(\Lambda) \tag{2}$$

so that

$$p(\hat{Y}, Y, \Lambda) = p(Y|\hat{Y}, \Lambda)p(\Lambda)p(\hat{Y}) \tag{3}$$

where $p(\Lambda)$ is assumed as uniformly distributed.

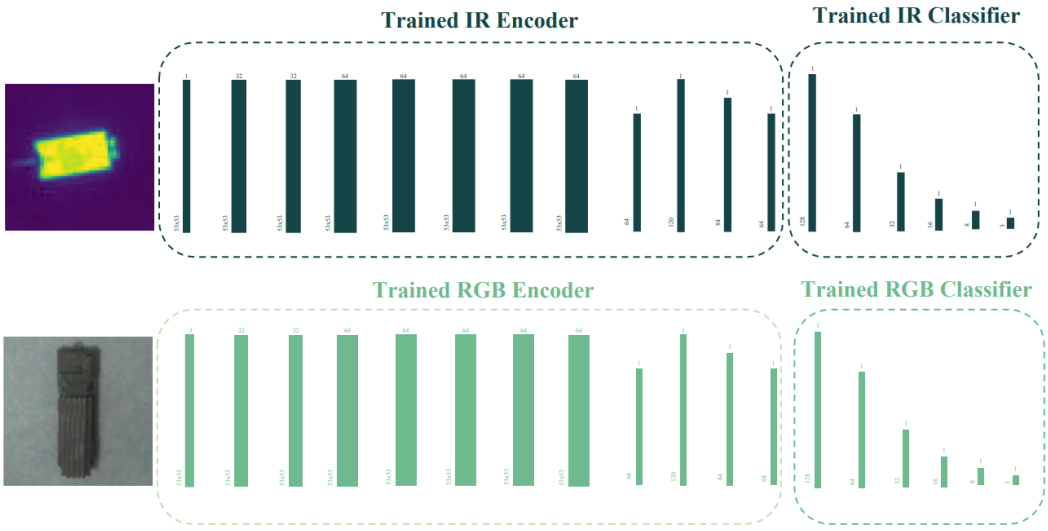


Fig. 5: Classifiers testing encoders without contrastive learning

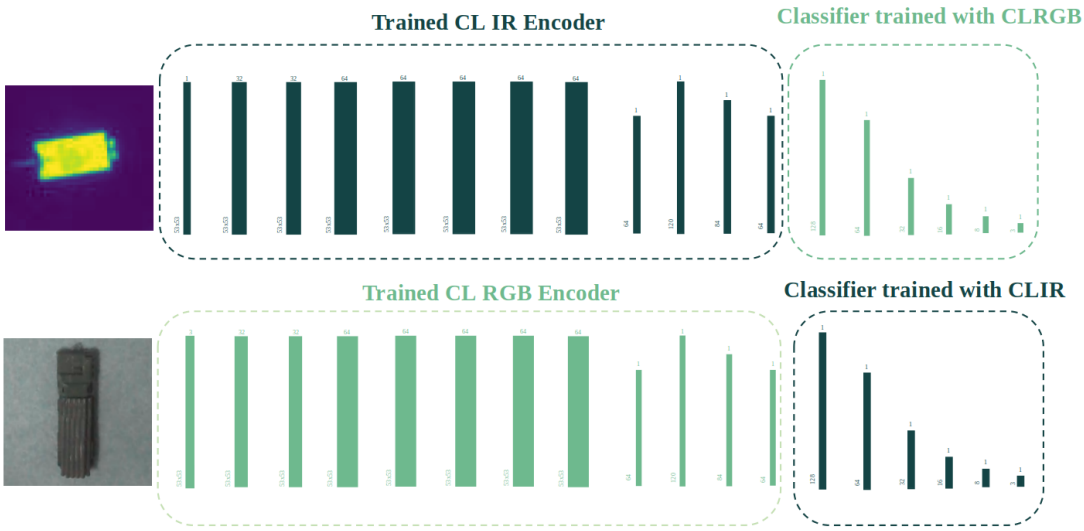


Fig. 6: Cross-modality classifiers testing encoders with contrastive learning

For the recursive update, the time step is denoted by k, and the evidence $e_k = \{\hat{y}_k, \lambda_k\}$. The evidence set up to step k is defined as $M_k = \{e_1, ..., e_k\}$, then the process of iterative update can then be defined as

$$p(Y|M_k) = \frac{p(e_k|Y)p(Y|M_{k-1})}{P(M_k)} \tag{4}$$

where $P(M_k)$ can be eliminated by normalization.

Constrained by the data collection process, we refrained from incorporating upstream classification results obtained through contrastive learning. In order to evaluate the efficacy of our Bayesian network, we employed synthetic data as depicted in Figure 9. In the context of the artifact scenario, the designated true target was defined as "Smerch" as per the reference.

IV. Results

A. Contrastive learning

In Fig. 10a, we see the IR classification without contrastive learning, yielding an average training accuracy of 82% and a validation accuracy of 82%. Conversely, Fig. 10b demonstrates the classification with a classifier trained using an encoder where contrastive learning has been applied, achieving markedly higher accuracies — 95% for training and 97% for validation. This substantial improvement suggests that for IR data, contrastive learning significantly enhances performance.

Comparatively, Fig. 11a and Fig. 11b both report similar accuracies for classifiers with and without contrastive learning (average training accuracy: 78% and average validation accuracy: 76%). Unlike the IR data, this parity suggests that in our contrastive learning training process, IR data may not con-

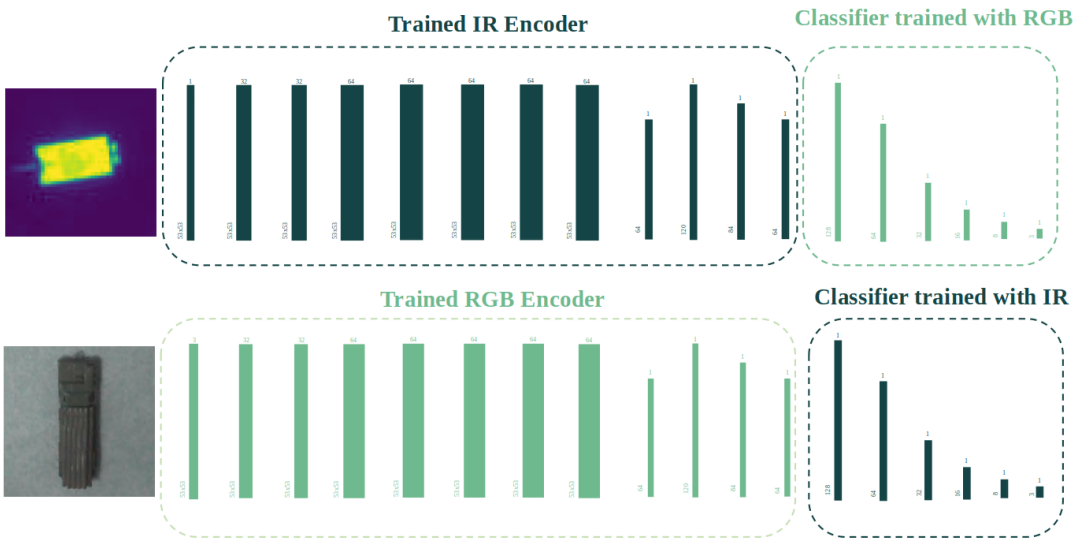


Fig. 7: Cross-modality classifiers testing encoders without contrastive learning

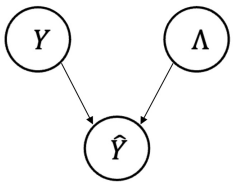


Fig. 8: Baysian Network

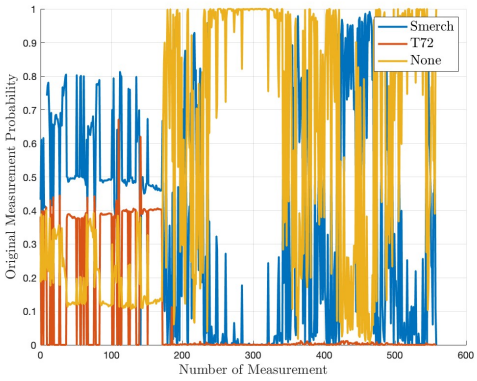
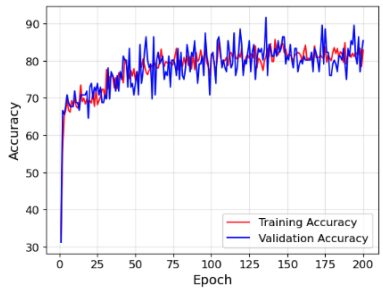
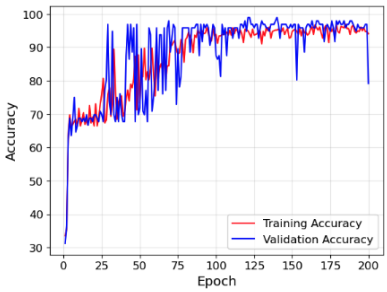


Fig. 9: Synthetic dataset for decision-level fusion with Bayesian Network



(a) IR Single Classification

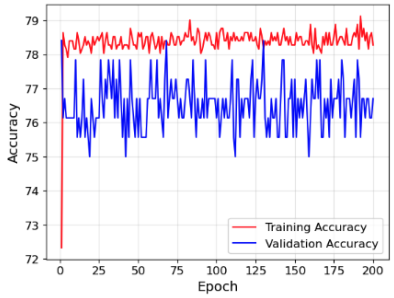


(b) IR CL Classification

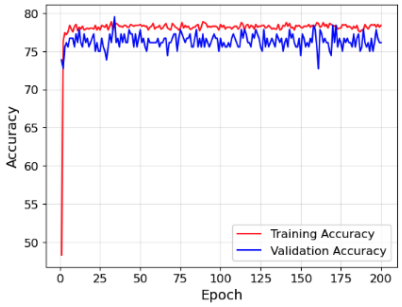
Fig. 10: Infrared Classifications

tribute additional useful information for RGB classification. This discrepancy could be due to the limitations of our dataset, particularly its lack of diverse scenes. In our experimental setup, both IR and RGB effectively visualize the target object. However, the superior performance in IR data classification might stem from the fact that, in identical scenes, RGB provides more detailed information than IR, such as color. Therefore, the previous experiments demonstrate a significant difference in the impact of using contrastive learning for IR data compared to RGB.

Fig. 12 and Fig. 13 display the results we obtained by swapping the classifiers between different modalities for testing. This means we applied the classifier trained on RGB data to IR and vice versa. We then compared the performance on the test set between classifiers from contrastive learning, as shown in Fig. 12a and Fig. 13a, and classifiers not using contrastive learning, as in Fig. 12b and Fig. 13b. The accuracy of both cases gets improved by around 30%. It is evident that the



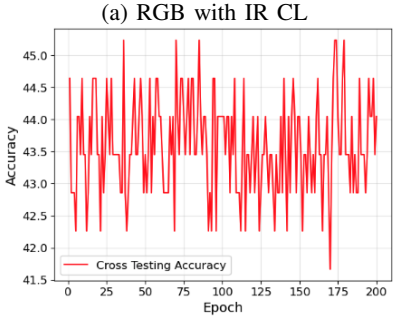
(a) RGB CL Classification



(b) RGB Single Classification

Fig. 11: RGB Classifications

(a) RGB with IR CL

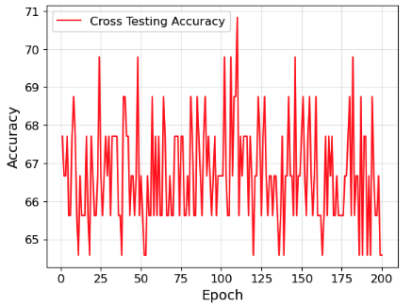


(b) RGB with IR Single

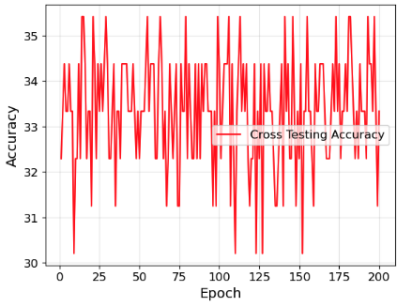
Fig. 13: RGB with IR Images

classifiers using contrastive learning significantly outperform those that do not.

B. Static Bayesian Network



(a) IR with RGB CL



(b) IR with RGB Single

Fig. 12: IR with RGB Images

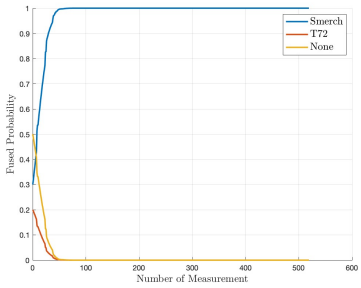


Fig. 14: Real dataset probability fusion with Bayesian Network

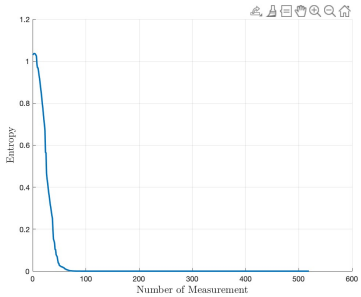


Fig. 15: Real dataset entropy with Bayesian Network

From Fig. 14, we can see the probability of the target being a smerch converging to "Smerch" (the right target class) after around 45 rounds and staying 1 consistently, with the entropy of the system converging to 0 and other confidence dropping to 0, which fits the true label of the scenario. From Fig. 15, we see the entropy of this system. drops as the classification converges to the right class.

V. Conclusion

The integration of contrastive learning within a Siamese neural network architecture represents a significant advancement in the field of machine learning. By using dual encoders for processing both IR and RGB images, our network effectively computes structured embeddings, essential for generating invariant features across sensor modalities. This methodology not only enhances feature extraction but also significantly improves the performance of classification tasks.

Our experiments with different data modalities and classifiers further demonstrate the robustness of the contrastive learning approach. By comparing networks trained with and without contrastive learning techniques, we have provided substantial evidence of its efficacy in enhancing model performance.

Moreover, the development of a Bayesian Network for decision-level fusion, while simplifying assumptions like neglecting time dependency, has added a layer of sophistication to our model. This network, capable of recursive updating based on sensor input, highlights the potential for real-time application in various fields, including autonomous systems and surveillance.

Future work can be extended to different modalities and more complicated decision fusion methods considering temporal information can be included like dynamic Bayesian networks.

References

[1] M. W. Roth, "Survey of neural network technology for automatic target recognition," *IEEE Transactions on neural networks*, vol. 1, no. 1, pp. 28–43, 1990.

[2] Q. Zhao and J. C. Principe, "Support vector machines for sar automatic target recognition," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 37, no. 2, pp. 643–654, 2001.

[3] Z. Geng, Y. Xu, B.-N. Wang, X. Yu, D.-Y. Zhu, and G. Zhang, "Target Recognition in SAR Images by Deep Learning with Training Data Augmentation," *Sensors*, vol. 23, no. 2, p. 941, Jan. 2023.

[4] P. Li, X. Hu, C. Feng, X. Shi, Y. Guo, and W. Feng, "SAR-AD-BagNet: An Interpretable Model for SAR Image Recognition Based on Adversarial Defense," *IEEE Geoscience and Remote Sensing Letters*, vol. 20, pp. 1–5, 2023.

[5] B. Zhang, R. Kannan, V. Prasanna, and C. Busart, "Accurate, Low-latency, Efficient SAR Automatic Target Recognition on FPGA," Jan. 2023.

[6] L. Zou, X. Wang, X. Yu, H. Ren, Y. Zhou, and X. Wang, "Synthetic aperture radar target recognition via deep attention convolutional network assisted by multiscale residual despeckling network," *Journal of Applied Remote Sensing*, vol. 17, no. 1, p. 016502, Jan. 2023.

[7] Yongmian Zhang, Qiang Ji, and C. Looney, "Active information fusion for decision making under uncertainty," in *Proceedings of the Fifth International Conference on Information Fusion. FUSION 2002. (IEEE Cat.No.02EX5997)*, vol. 1. Annapolis, MD, USA: Int. Soc. Inf. Fusion, 2002, pp. 643–650.

[8] N. Kaempchen and K. Dietmayer, "Data synchronization strategies for multi-sensor fusion," in *Proceedings of the IEEE Conference on Intelligent Transportation Systems*, vol. 85, no. 1, 2003, pp. 1–9.

[9] C. Pohl and J. v. Genderen, "Review article multisensor image fusion in remote sensing: concepts, methods and applications," *International Journal of Remote Sensing*, vol. 19, pp. 823–854, 1998.

[10] A. P. James and B. V. Dasarathy, "Medical image fusion: a survey of the state of the art," *Information Fusion*, vol. 19, pp. 4–19, 2014.

[11] B. Peng, W. Li, X. Xie, Q. Du, and K. Liu, "Weighted-fusion-based representation classifiers for hyperspectral imagery," *Remote Sensing*, vol. 7, pp. 14 806–14 826, 2015.

[12] C. Cao, J. Song, R. Q. Su, X. Wu, Z. Wang, and M. Hou, "Structure-constrained deep feature fusion for chronic otitis media and cholesteatoma identification," *Multimedia Tools and Applications*, 2023.

[13] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 18 661–18 673, 2020.

[14] 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.

[15] 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.

[16] P. Leptourgos, V. Bouttier, and R. Jardri, "A functional theory of bistable perception based on dynamical circular inference," *PLOS Computational Biology*, vol. 16, p. e1008480, 2020.

[17] K. Schmack, V. Weilnhammer, J. Heinzle, K. E. Stephan, and P. Sterzer, "Learning what to see in a changing world," *Frontiers in Human Neuroscience*, vol. 10, 2016.

[18] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.

[19] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proceedings of COMPSTAT'2010: 19th International Conference on Computational StatisticsParis France, August 22-27, 2010 Keynote, Invited and Contributed Papers*. Springer, 2010, pp. 177–186.

[20] J. Shin, S. Chang, J. Weaver, J. C. Isaacs, B. Fu, and S. Ferrari, "Informative multiview planning for underwater sensors," *IEEE Journal of Oceanic Engineering*, vol. 47, no. 3, pp. 780–798, 2022.