Adaptive self-supervised vision transformers for multi-sensor automatic target recognition

Sophia Abraham[a], Suya You[b], Jonathan D. Hauenstein[a], and Walter Scheirer[a]

[a]University of Notre Dame, Notre Dame, IN 46556, USA
[b]DEVCOM Army Research Laboratory, Adelphi, MD, USA

ABSTRACT

We introduce a novel framework that combines hierarchical Vision Transformer (ViT) modules with *masked autoencoding* and *contrastive learning* for self-supervised Automatic Target Recognition (ATR). Unlike conventional ViTs, our approach employs a **multi-stage** encoder (each stage embedding lower-level features, then progressively refining them) and a dedicated transformer-based decoder for masked reconstruction. Additionally, we implement a **dynamic hyperparameter tuning** strategy that continuously adjusts the drop path probabilities based on an **attention entropy** metric computed from the last transformer block. This mechanism encourages the model to learn robust, domain-specific representations with minimal manual intervention. While the architecture supports multi-sensor data fusion, we focus on single-modality (EO or IR) data in our experiments due to dataset constraints. Experimental results on the DSIAC dataset demonstrate that our method can adapt to the unique complexities of ATR tasks, balancing performance gains with efficient resource usage. The code for this framework is available on GitHub: https://github.com/sabraha2/Adaptive-self-supervised-vision-transformers-for-multi-sensor-automatic-target-recognition.

Keywords: Self-supervised learning, Transformers, Automatic Target Recognition

1. INTRODUCTION

Automatic Target Recognition (ATR) systems play a critical role in modern defense and security applications, where reliable detection and classification of targets are essential in dynamic and often challenging environments. Traditional ATR approaches have heavily relied on supervised learning, which necessitates large, annotated datasets that are both expensive and time-consuming to acquire.[1] This challenge is further exacerbated when dealing with data from multiple sensor modalities such as electro-optical (EO), radar, LiDAR, multispectral, and hyperspectral imaging. These modalities capture complementary information about targets and their surroundings, yet they also introduce complexities in data fusion and model generalization.[2,3]

In this context, self-supervised learning (SSL) has emerged as a powerful paradigm for learning robust feature representations from unlabeled data. Vision Transformers (ViTs), with their ability to model long-range dependencies and global contextual information, have shown significant promise in various computer vision tasks.[4,5] Motivated by these advances, our work proposes an *adaptive self-supervised Vision Transformer* framework for multi-sensor ATR. Our approach leverages a combination of masked image modeling and contrastive learning to extract meaningful features from vast amounts of unlabeled multi-modal data.[6,7]

A key innovation of our framework is the dynamic hyperparameter tuning mechanism, which adaptively adjusts critical training parameters based on real-time performance metrics such as attention entropy and convergence rates. This adaptive tuning is particularly valuable in ATR applications, where the cost of false negatives is high and operational environments can vary significantly.[8,9] By dynamically modulating parameters like patch size, attention head configurations, and drop path probabilities, our model is able to optimize its architecture and learning process in a domain-specific manner without requiring extensive manual intervention.

Emails: {sabraha2, hauenstein, walter.scheirer}@nd.edu
Email: suya.you.civ@army.mil

To validate our approach, we conduct experiments on datasets representative of real-world ATR scenarios. In particular, we employ the DSIAC dataset, a well-established resource that provides multi-sensor imagery including infrared (IR) and electro-optical (EO) modalities.[10] The DSIAC dataset not only offers a challenging environment for target recognition but also serves as a benchmark for assessing the adaptability and performance of our proposed framework across diverse sensor modalities.[11]

2. BACKGROUND AND RELATED WORK

2.1 Background

Automatic Target Recognition (ATR) is a critical technology in defense and security applications, where rapid and reliable target identification is essential. Early work in this field has surveyed various ATR systems and techniques.[1,8] Traditional ATR methods relied on hand-engineered features and classical machine learning techniques, which often struggled with the complexity of modern sensor data. With the advent of deep learning, convolutional neural networks (CNNs) have been widely adopted for ATR tasks, demonstrating improved performance.[12,13] However, these supervised methods require extensive labeled data and manual hyperparameter tuning,[3,14] which can be particularly challenging when dealing with heterogeneous multi-sensor data.

Modern ATR applications benefit significantly from sensor fusion, where complementary information from different modalities—such as synthetic aperture radar (SAR), infrared, electro-optical, LiDAR, multispectral, and hyperspectral imagery—is combined to enhance detection and classification accuracy.[2,3] For instance, the DSIAC ATR Algorithm Development Image Database[10] has served as an important benchmark for evaluating ATR algorithms, and several studies have used it to compare target detection methods.[11] Additionally, recent surveys have highlighted the growing role of deep-learning-based approaches in SAR ATR.[15]

Self-supervised learning (SSL) has recently emerged as an alternative to traditional supervised methods by leveraging vast amounts of unlabeled data to learn robust representations.[16] SSL techniques, including masked image modeling[7,17] and contrastive learning,[6] have shown promise in reducing the dependency on labeled data. Vision Transformers (ViTs) further enhance these approaches by capturing long-range dependencies and global context.[18,19] Emerging work demonstrates that integrating SSL with ViTs leads to robust feature learning even in challenging environments.[5,20,21]

2.2 Related Work

Supervised ATR Approaches: Early ATR systems predominantly relied on supervised CNN-based methods.[1,12] While these methods achieved competitive performance, they are often constrained by the need for large annotated datasets and extensive hyperparameter tuning.[13,14] Recent work has explored automatic hyperparameter optimization techniques to alleviate some of these challenges.[22,23]

Multi-Sensor Data Fusion: Combining data from multiple sensors has proven beneficial for ATR by capturing diverse information. Several studies have focused on the fusion of modalities, addressing issues such as varying resolutions and noise characteristics.[2,3] The DSIAC dataset, for example, provides a multi-modal platform that has been widely used for both evaluation and comparison of ATR algorithms.[9–11]

Self-Supervised Learning and Vision Transformers: Self-supervised learning has gained attention as a means to overcome the limitations of supervised learning in ATR. Recent surveys offer a comprehensive overview of SSL techniques and their applications in computer vision.[16,24] In particular, masked autoencoders[7,17] and contrastive frameworks[6] have been successfully integrated with Vision Transformers,[18,19] yielding state-of-the-art performance in various vision tasks. These advancements have motivated further research into SSL-based ATR systems, as they offer the potential to learn robust, transferable representations from unlabeled multi-sensor data.[5,20,21]

Adaptive and Dynamic Tuning Techniques: Adaptive hyperparameter tuning is emerging as a promising approach to handle the diverse challenges of ATR, especially in multi-sensor settings. Dynamic methods that adjust learning parameters in response to performance metrics have shown potential in enhancing model robustness and efficiency.[14,23] Such techniques are particularly relevant when the cost of misclassification is high and sensor data characteristics vary significantly across different domains.[3,15]

In summary, while traditional ATR approaches have achieved notable success using supervised deep learning techniques, the need for large labeled datasets and the challenges of multi-sensor fusion have spurred interest in self-supervised learning and adaptive tuning methods. Our work builds upon these advances by proposing an adaptive self-supervised Vision Transformer framework that effectively integrates multi-modal sensor data and leverages dynamic hyperparameter tuning, with a particular focus on evaluation using the DSIAC dataset.[10,11]

3. METHODS

In this section, we provide a detailed description of our *hierarchical* self-supervised Vision Transformer framework for Automatic Target Recognition (ATR). Our approach revolves around three core components: (1) a multi-stage encoder with a transformer-based decoder, (2) a combination of *masked autoencoding* (MAE) and *contrastive* objectives for self-supervision, and (3) an adaptive *dynamic hyperparameter tuning* mechanism guided by an attention-entropy feedback loop.

3.1 Overview

Our proposed Vision Transformer (ViT) framework consists of a **three-stage hierarchical encoder** and a dedicated transformer-based decoder (Figure 1). The encoder progressively processes the input through patch embeddings, generating token sequences that flow through successive transformer blocks. These blocks employ class tokens, apply self-attention mechanisms, and refine the feature representations at each stage.

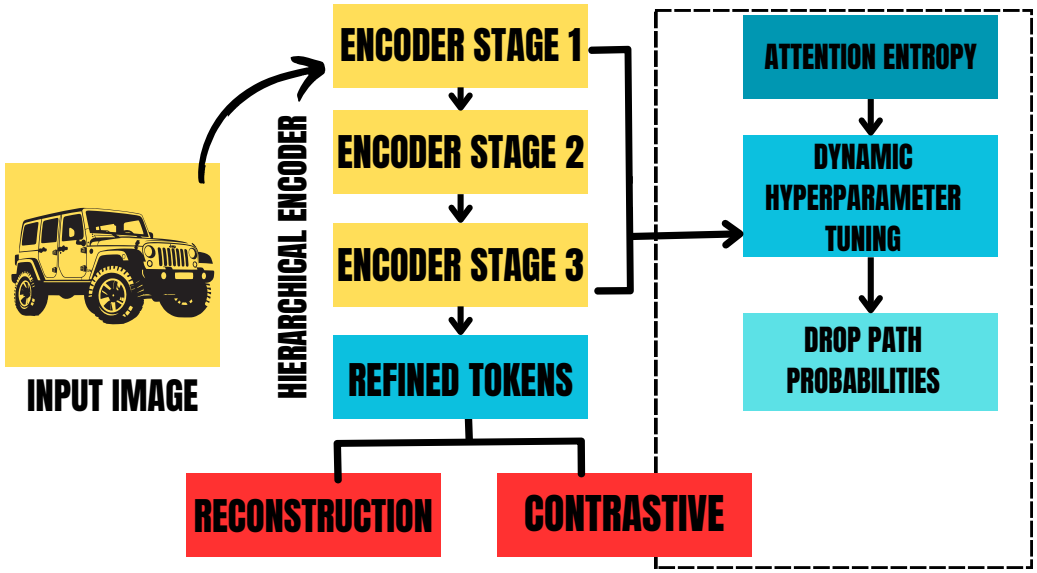

Figure 1. Overview of our adaptive self-supervised Vision Transformer framework for multi-sensor ATR. An input image is passed through a three-stage hierarchical encoder to produce refined token embeddings. These tokens are then used in two complementary self-supervised objectives—masked autoencoding for pixel reconstruction and a contrastive loss for global feature alignment. Simultaneously, we compute the attention entropy from the final stage's self-attention maps and feed it into a proportional controller that dynamically adjusts the model's drop-path probability, providing adaptive regularization throughout training.

During the self-supervised pretraining phase, we optimize two complementary objectives:

1. **Masked Autoencoding (MAE):** A fraction of the input image's pixels (or patches) is randomly masked, and the decoder reconstructs these missing regions. This encourages the model to learn robust features that are resilient to missing data.
2. **Contrastive Embedding:** Two augmented views of the same image are generated, and the model is trained to align their embeddings in the feature space, using a contrastive loss.

These objectives can be used in isolation—*reconstruction-only* or *contrastive-only*—or in combination (*combined*), applying both losses simultaneously. After this self-supervised pretraining, the model is *fine-tuned* for ATR classification using labeled data.

3.2 Hierarchical Transformer Architecture

Our encoder consists of three stages (Stage 1, Stage 2, Stage 3), each with the following components:

- **PatchEmbed:** A module that converts the output features from the previous stage into tokens.
- **Positional Embedding:** A learnable embedding that helps the model maintain spatial awareness across transformer blocks.
- **TransformerBlock:** Each stage contains a series of transformer blocks that apply multi-head self-attention and feedforward layers to refine token representations.

At the final stage, a token sequence is output, with the first token (the class token) being used for classification. Unmasked tokens from this stage are also passed to the decoder for pixel reconstruction in the MAE phase.

Decoder for Masked Reconstruction. The decoder in our framework is a transformer-based network that utilizes multi-head self-attention to process the final encoder tokens. These tokens are passed through several transformer blocks and a linear projection layer to map them to a higher-dimensional space. The decoder then uses a series of transposed convolutions (deconvolutions) to reconstruct the original input image or patches, ensuring the output matches the spatial dimensions of the input. The reconstruction is compared to the masked input using a Mean Squared Error (MSE) loss or a Mean Absolute Error (L1) loss, focusing on the masked regions of the image. Additionally, an optional *perceptual loss* term can be applied, where both the original and reconstructed images are passed through a pretrained VGG16 network (with frozen weights). This perceptual loss helps align high-level features between the original and reconstructed images, encouraging the model to preserve semantic content rather than focusing solely on pixel-level accuracy. The perceptual loss is weighted by a hyperparameter, λ_{perc}, which controls its influence on the total loss. This multi-stage approach, combined with hierarchical transformer processing, allows the decoder to recover fine-grained image details while maintaining a high-level understanding of the input.

3.3 Handling Single-Sensor and Multi-Sensor Data

Although our framework can *stack multiple sensor channels* (e.g., EO + IR in a 4-channel tensor), in this study, we focus on single-modality data due to the available DSIAC images. Specifically, EO images are represented as three-channel tensors, while IR images are one-channel. The input data is normalized, resized, and passed through the hierarchical encoder. While we focus on single-modality data in this work, the architecture is designed to integrate additional sensor channels (e.g., LiDAR or hyperspectral data) in future iterations.

3.4 Dynamic Hyperparameter Tuning via Attention Entropy

A key feature of our framework is the *dynamic* adjustment of the **drop path** probability, which controls the stochastic depth of the model. We continuously monitor the **attention entropy** metric derived from the final block's self-attention matrices. The entropy is calculated as:

$$\text{Entropy} = -\sum_i \alpha_i \log(\alpha_i),$$

where α_i represents the normalized attention weight for each head and token. If the computed entropy deviates from a target value (e.g., 1.0), we update the drop path probability using a proportional control law:

$$\text{new\_drop\_prob} \leftarrow \max\Big(0,\ \min\big(\text{current\_drop\_prob} + \eta \times (\text{target\_entropy} - \text{current\_entropy}),\ \text{max\_drop}\big)\Big),$$

where η is a *tuning factor* (e.g., 0.05), and max_drop is a cap on the drop path probability (e.g., 0.5). This dynamic adjustment occurs regularly (e.g., every 10 mini-batches), allowing the network to *self-regulate* its regularization level based on observed attention patterns.

3.5 Training Protocol

Our training process consists of two phases:

1. **Self-Supervised Pretraining:**
   - *MAE:* Randomly mask a fraction (e.g., 75%) of pixels and reconstruct the missing regions.
   - *Contrastive:* Generate two augmented views of each image and enforce their embeddings to be similar in feature space.
   - *Combined:* Sum both losses (weighted by a hyperparameter) to leverage both local reconstruction fidelity and global feature alignment.

   During this phase, we dynamically adjust the drop path probability based on the attention entropy metric.

2. **Supervised Fine-tuning:**
   - Attach a linear classification head on top of the final stage's class token.
   - Compute the cross-entropy loss for labeled ATR targets.
   - Optionally, continue dynamic drop path tuning or freeze it at the best value discovered during pretraining.

We use the Adam optimizer (or AdamW, depending on the configuration) for training, along with a cosine annealing learning rate scheduler, which gradually reduces the learning rate over the course of training. Additionally, a dynamic learning rate tuner is employed, which adjusts the learning rate further if the validation loss plateaus. This tuner monitors the validation loss during training and reduces the learning rate when there is no improvement for a predefined number of epochs, using a proportional factor. This approach helps to stabilize training and prevent overfitting by allowing the model to fine-tune its parameters more effectively during the later stages of training.

4. EXPERIMENTS

In this section, we provide an overview of the experiments conducted to evaluate our adaptive, self-supervised Vision Transformer (ViT) framework. The primary goal of these experiments was to explore how different hyperparameters affect model performance on **single-sensor** subsets (either EO or IR) from the DSIAC ATR Algorithm Development Image Database.

4.1 Hyperparameter Sweep Setup

We conducted an extensive hyperparameter sweep to assess the impact of various configurations on the performance of our framework. The specific hyperparameters and values can be found in Table 1. Key hyperparameters explored include the SSL mode (Reconstruction, Contrastive, Combined), the mask ratio, the learning rate, and the drop path probabilities. The experiments were designed to explore different configurations across both EO and IR datasets to understand how our model generalizes across different sensor modalities.

For each configuration, we performed self-supervised pretraining with the corresponding SSL mode. The self-supervised phase was followed by fine-tuning the model for ATR classification using labeled data. We varied parameters such as the number of SSL epochs, contrastive weight, target attention entropy, and tuning factor, allowing us to identify the optimal settings for each condition.

Table 1. Hyperparameter Sweep Configuration

Hyperparameter	Values
Modalities	EO, IR
SSL Mode	Reconstruction, Contrastive, Combined
Mask Ratio	0.3, 0.5, 0.75
Lambda (Perceptual Loss Weight)	0.1, 0.5, 1.0
SSL Epochs	100, 200
Fine-Tuning Epochs	50
Batch Size	64
Learning Rate	1e-3, 1e-4
Use Contrastive Objective	True, False
Contrastive Weight	0.1, 0.2
Target Attention Entropy	0.8, 1.0, 1.2
Tuning Factor (Drop Path)	0.03, 0.05, 0.07
Max Drop Path Probability	0.3, 0.5, 0.7

4.2 Dataset and Preprocessing

We used the DSIAC dataset, splitting it into training, validation, and test sets. For EO data, the images were resized to 64×64 pixels and augmented with random cropping, flipping, and color jittering. For IR data, we resized the images to the same resolution and treated them as single-channel tensors. All images were normalized, and batches were fed into the hierarchical transformer encoder for both pretraining and fine-tuning phases.

4.3 Self-Supervised Pretraining

During the self-supervised pretraining phase, we evaluated three different SSL modes: the **Reconstruction-Only (MAE)** approach, where a percentage of the image pixels are masked and the model is trained to reconstruct the masked regions; the **Contrastive-Only** approach, which generates two augmented views of each image and uses a contrastive loss to align their embeddings in feature space; and the **Combined** approach, which combines both MAE and contrastive losses to simultaneously preserve local feature fidelity and ensure global feature alignment.

In all experiments, we dynamically adjusted the drop path probability based on the computed attention entropy. This dynamic regularization mechanism ensures a balance between exploration and convergence during training.

4.4 Supervised Fine-Tuning

Following the self-supervised pretraining, we fine-tuned the model for ATR classification by replacing the SSL heads with a linear classifier on the final class token. Fine-tuning was performed for approximately 50 epochs using labeled ATR data. For scenarios where certain target classes were underrepresented in the training set, we used a weighted cross-entropy loss to address class imbalances.

4.5 Objective and Evaluation Metrics

The primary objective of these experiments was to identify optimal configurations of hyperparameters that maximize the model's performance on ATR tasks. The model's effectiveness was evaluated using standard ATR performance metrics, such as classification accuracy and convergence rates, which we report in the subsequent results section.

In summary, our experimental methodology involved conducting a hyperparameter sweep to assess the impact of various configurations on model performance. We evaluated the model's effectiveness through both self-supervised learning and fine-tuning phases, analyzing how different self-supervised learning (SSL) modes—Reconstruction, Contrastive, and Combined—affect the model's ability to generalize across electro-optical (EO) and infrared (IR) datasets. This approach allowed us to gain insights into the model's performance under different settings and its adaptability to diverse modalities.

5. RESULTS

5.1 Hyperparameter Sweep Analysis

In this section, we present the results of our hyperparameter sweep across both the EO and IR modalities. Table 2 summarizes the aggregated performance metrics across 334 EO runs and 216 IR runs.

Table 2. Overall Aggregated Performance Metrics for EO and IR. Results are computed over 334 EO runs and 216 IR runs.

Modality	Best Acc (%)	Mean Acc (%)	Worst Acc (%)	SD (%)
EO	21.54	14.84	0.00	3.88
IR	19.32	4.41	0.00	6.41

As seen in Table 3, performance across different SSL modes is reported for both EO and IR modalities. For EO, the **combined** and **contrastive** approaches yield similar mean accuracies (17%), with low variability (SD 1.8–1.9%). In contrast, the **reconstruction-only** approach results in a significantly lower mean accuracy (10.54%), but with higher variability (SD = 2.5%). For IR data, the combined and contrastive modes exhibit near-zero performance, while the reconstruction method achieves a mean accuracy of 13.23%.

Table 3. Mean Final Test Accuracy (%) and Standard Deviation by SSL Mode and Modality.

	EO		IR	
SSL Mode	Mean (%)	SD (%)	Mean (%)	SD (%)
Combined	17.34	1.82	0.00	0.00
Contrastive	17.26	1.89	0.00	0.00
Reconstruction	10.54	2.50	13.23	2.34

The results highlight the modality-specific efficacy of our framework. For EO data, the combined and contrastive SSL modes yielded the highest performance, with mean accuracies around 17%. These methods show consistent performance with low variability, indicating that the model effectively leverages the texture-rich nature of EO data.

On the other hand, for IR data, both combined and contrastive methods resulted in near-zero performance, likely because these methods struggled to capture the limited structural information in IR images. The reconstruction-only method, however, achieved a significantly higher accuracy (mean 13.23%), suggesting that IR data benefits more from learning structural cues rather than relying on contrastive alignment. These results underline the need for modality-specific approaches when applying self-supervised learning to multi-sensor ATR tasks.

6. QUALITATIVE RESULTS AND ANALYSIS

In this section, we present qualitative examples from five representative runs. These examples illustrate the behavior of our adaptive self-supervised Vision Transformer across EO and IR modalities and under different training objectives (reconstruction, contrastive, and combined modes). In addition, we discuss the role of the attention maps in revealing the network's focus and highlight that high downstream classification accuracy can be achieved even when the pixel-level reconstructions appear suboptimal.

6.1 Qualitative Reconstructions and Attention Maps

In Figure 2, the IR reconstruction captures large thermal gradients but lacks sharp object boundaries; nonetheless, the attention map correctly highlights the hottest regions around the target silhouette, showing that the model leverages thermal contrast for classification. In Figure 3, the EO reconstruction maintains broad horizon and texture gradients but smooths fine details; its attention map instead focuses on high-contrast texture patches (e.g., vehicle edges), demonstrating that even blurred reconstructions can yield highly discriminative embeddings when guided by the contrastive component. By combining masked autoencoding with view-alignment and tuning the drop-path probability (η=0.05, max = 0.3) to maintain an attention entropy of 1.0, the model learns to prioritize global structure and salient features over exact pixel fidelity—resulting in an 18.46 % final accuracy.

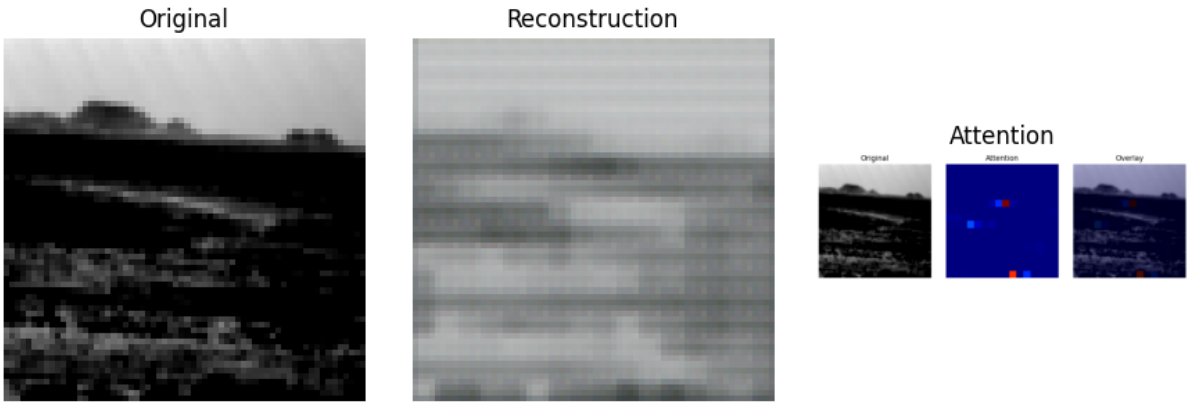

Figure 2. IR sample under Combined SSL. The reconstruction captures coarse thermal gradients but smooths fine edges. The attention overlay (right) highlights hot regions near the target. Test Acc = 18.46%.

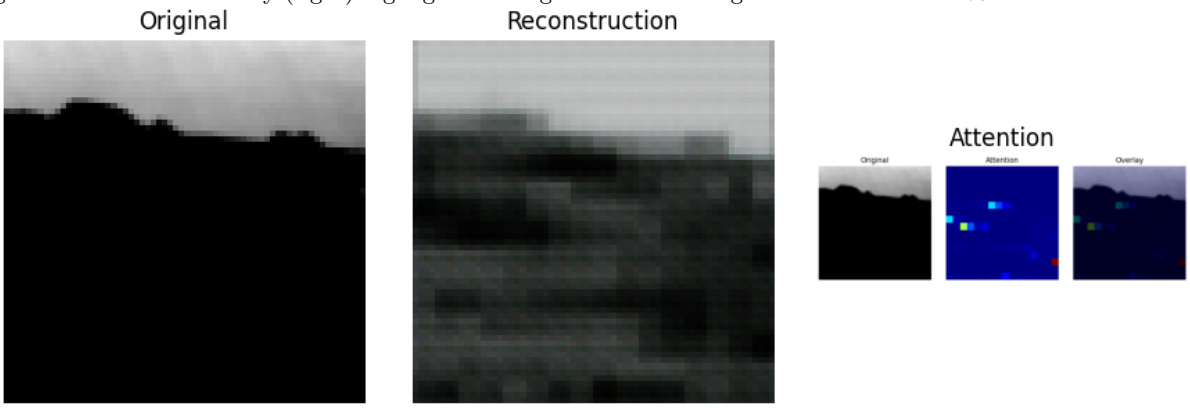

Figure 3. EO sample under Combined SSL. The reconstruction preserves horizon and texture gradients but blurs fine details. The attention overlay highlights high-contrast texture patches. Test Acc = 18.46%.

Figure 4. Qualitative reconstructions and attention overlays for IR and EO test samples under the combined MAE + contrastive self-supervised objective. Both reconstructions show blurring, but the attention maps reveal the model focuses on the most discriminative thermal or texture cues.

6.2 Reconstruction and Contrastive Learning on EO and IR Data

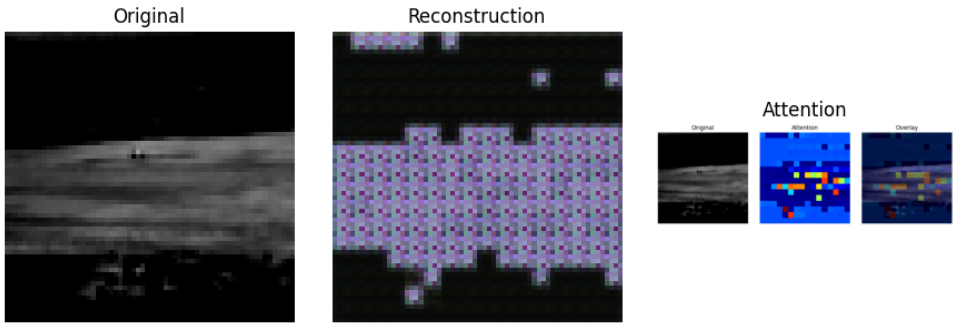

Figure 5. Reconstruction-only mode on EO data. The model maintains structural consistency but lacks fine texture details. Test Acc = 11.11%.

Figure 5 demonstrates reconstruction results in the EO dataset. Although fine texture details are missing, the model captures the general structure. The attention map shows the focus on key target regions, even with imperfect reconstruction quality. This indicates that the model's learned features are discriminative enough for classification, despite the lower visual quality of the reconstruction.

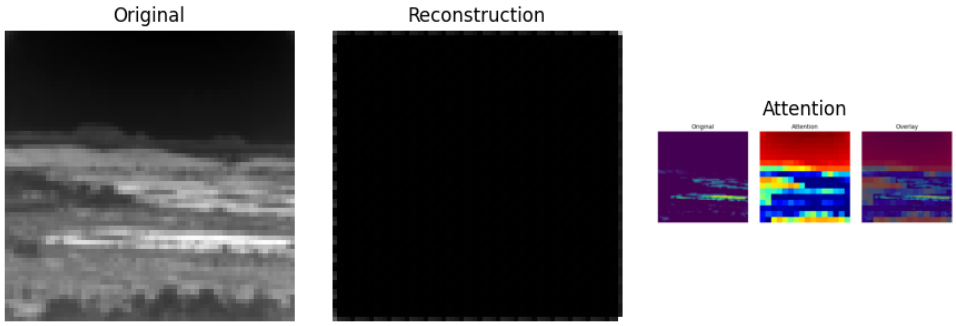

Figure 6. Reconstruction mode on IR data. The model captures the coarse structural cues and the target's shape. Test Acc = 25.00%.

The results in Figure 6 highlight that, for IR data, even with low-texture images, the reconstruction mode successfully recovers the structural features of the target, despite the qualitative representation appearing blank to the human eye. The model's attention map emphasizes the most salient regions, reinforcing the idea that IR recognition benefits from focusing on broader structural cues rather than fine details. Notably, this mode achieved the highest classification accuracy, further supporting its effectiveness for IR-based target recognition.

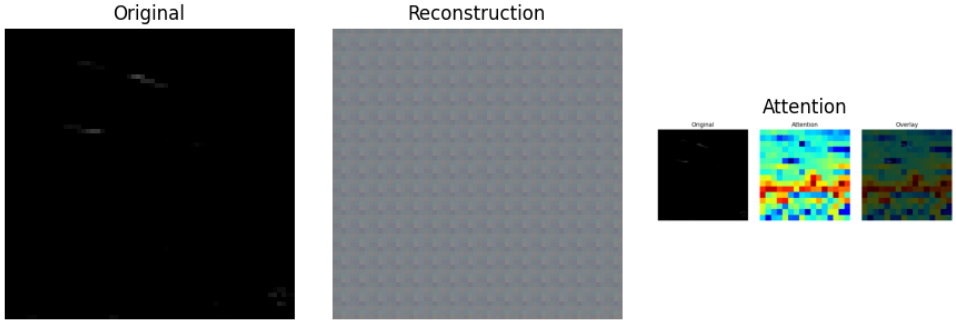

Figure 7. Contrastive-only mode on EO data. The model emphasizes semantic feature alignment, achieving a test accuracy of 13.85%.

In Figure 7, the model was trained in contrastive-only mode on EO data. Rather than focusing solely on reconstruction, the contrastive loss drives the model to learn discriminative features that cluster similar augmented views in feature space. Although the reconstruction quality may appear less perfect compared to the MAE-only approach, the robust feature alignment contributes to a competitive final test accuracy of 13.85%. The attention map shows that the model attends to key regions useful for discrimination, even when overall image reconstruction is poor.

The qualitative analysis confirms several key observations. Reconstruction-only models, as seen in runs with both EO and IR data, preserve the global structure of the target but may miss fine details. Attention maps from these models highlight the most target-critical regions. For IR reconstructions combining reconstruction with contrastive learning proves beneficial, as it emphasizes structural cues that are crucial for accurate recognition. Contrastive learning which focuses on feature alignment across augmented views, contributes to robust classification performance, even when pixel-level reconstructions are less refined.

6.3 Effect of dynamic-tuning hyperparameters on the embedding geometry.

Figure 11 compares t-SNE projections of the last-layer contrastive embeddings under three representative configurations drawn from our hyper-parameter sweep.

- **Run A** (target entropy = 1.0, tuning factor $\eta = 0.07$, max-drop = 0.5, SSL epochs = 100) produces two clearly separated super-clusters connected by a smooth manifold. Individual target classes (indicated

by colour) lie on compact, ordered neighborhoods along this trajectory, suggesting that the aggressive drop-path updates encouraged by the larger η quickly settle into a low-entropy but still expressive feature space.

- **Run B** raises the entropy set-point to 1.2 and lowers η to 0.05 while keeping max-drop = 0.5. The controller therefore spends more iterations near the upper end of the stochastic-depth range, injecting additional regularisation noise. This manifests as a much more dispersed scatter in two-dimensional space: class clusters still form, but they are smaller and inter-class distances shrink, reflecting reduced intra-class cohesion.
- **Run C** extends pre-training to 200 SSL epochs but caps max-drop at 0.3 and drops η further to 0.03. With less regularisation pressure and more training time, the model collapses many classes onto a single elongated arc, indicating over-fitting to a dominant 'direction" in feature space. The silhouette is noticeably flatter than in Run A, corroborating the importance of sufficient stochastic depth during long SSL schedules.

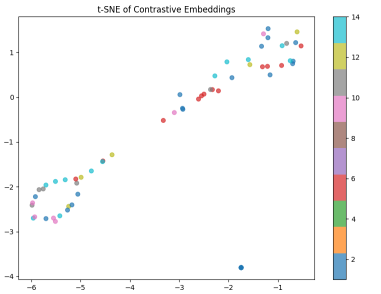

Figure 8. (a) Target entropy 1.0, $\eta =$ 0.07, max-drop 0.5, 100 SSL epochs (Run A).

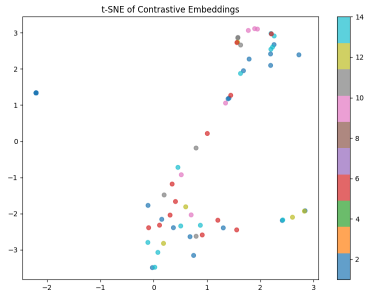

Figure 9. (b) Target entropy 1.2, $\eta =$ 0.05, max-drop 0.5, 100 SSL epochs (Run B).

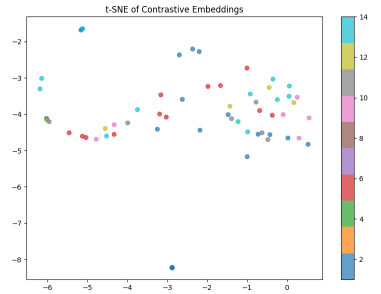

Figure 10. (c) Target entropy 1.0, $\eta =$ 0.03, max-drop 0.3, 200 SSL epochs (Run C).

Figure 11. t-SNE of contrastive embeddings under three dynamic-tuning regimes.

The t-SNE projections across the three configurations reveal distinct behaviors in feature separation based on the dynamic-tuning hyperparameters. In Run A (target entropy = 1.0, η = 0.07, max-drop = 0.5, 100 SSL epochs), the embeddings exhibit clear, well-separated super-clusters, suggesting that the model has successfully learned a discriminative feature space with high intra-class cohesion and well-defined class boundaries. This result aligns with the hypothesis that a moderately high tuning factor encourages sufficient exploration during the self-supervised learning phase, thus facilitating the formation of meaningful clusters in the embedding space. In Run B (target entropy = 1.2, η = 0.05, max-drop = 0.5, 100 SSL epochs), the model's feature space becomes more dispersed, with smaller and less distinct clusters, indicating that the increased entropy set-point and reduced tuning factor may have led to excess regularization. This dilution of feature separation suggests that the model struggled to fine-tune its representations, potentially due to too much exploration, which reduced the compactness of the clusters. Lastly, in Run C (target entropy = 1.0, η = 0.03, max-drop = 0.3, 200 SSL epochs), the extended pre-training duration combined with lower stochastic-depth regularization led to significant overfitting. The embeddings appear collapsed into a singular elongated arc, highlighting that while the model was trained for a longer period, the insufficient regularization during training caused it to converge too early into a dominant feature direction. This collapse reflects a loss of diversity in feature representations, further suggesting the importance of balancing training duration with regularization parameters. These results emphasize that a combination of appropriate entropy set-point, moderate tuning factor, and sufficient stochastic-depth regularization during pre-training is crucial for learning robust, discriminative feature representations that facilitate well-separated class clusters.

6.4 Analysis of Loss Curves for Different Self-Supervised Learning Modes

Figure 15 presents the loss curves for three self-supervised learning (SSL) modes—Reconstruction, Contrastive, and Combined—during training. These plots demonstrate how the model's loss evolves with respect to the number of training steps for each SSL mode.

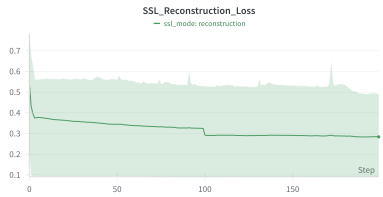

Figure 12. SSL Reconstruction Loss over training steps.

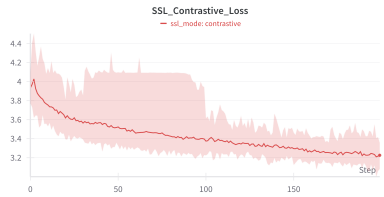

Figure 13. SSL Contrastive Loss over training steps.

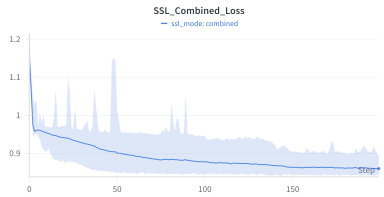

Figure 14. SSL Combined Loss over training steps.

Figure 15. Loss curves for SSL modes during training.

- Reconstruction Mode (Figure 12): The plot for the reconstruction loss shows a sharp decrease at the beginning, followed by a more gradual reduction as the model learns to reconstruct missing image regions. This behavior is expected, as the model first learns to handle large amounts of missing data, stabilizing as the reconstruction task becomes easier over time.
- Contrastive Mode (Figure 13): The contrastive loss shows a similar trend, with a rapid decrease early in training. However, the loss curve in this mode exhibits more noise, reflecting the model's struggle to align multiple augmented views effectively, as indicated by the higher variance in the loss during early iterations.
- Combined Mode (Figure 14): The combined loss, which integrates both reconstruction and contrastive learning objectives, demonstrates a smoother decrease in loss across training steps. The dual-task learning strategy aids the model in gradually stabilizing, suggesting that combining reconstruction and contrastive loss leads to better feature alignment and representation learning.

The smoother loss curves in the combined mode (Figure 14) indicate that this approach benefits from leveraging both objectives, leading to a more stable and efficient learning process, particularly in challenging training scenarios.

7. LIMITATIONS AND FUTURE WORK

While our self-supervised Vision Transformer (ViT) framework for Automatic Target Recognition (ATR) has shown promising results, there are several avenues for improvement. One limitation is the current focus on single-modality data (EO or IR), while the potential benefits of multi-sensor data fusion (e.g., combining EO + IR, LiDAR, or hyperspectral) remain unexplored. Extending our framework to handle multiple modalities could significantly enhance its robustness and accuracy, particularly in complex environments where multiple sensor types provide complementary information.

Another limitation is the sensitivity of the dynamic hyperparameter tuning mechanism to initial settings. Future work will investigate more adaptive tuning strategies, potentially incorporating reinforcement learning or meta-learning to further optimize the model's performance. Additionally, the model's reliance on pretraining with masked autoencoding (MAE) and contrastive learning presents an opportunity for exploring more advanced or hybrid self-supervised methods, such as unsupervised contrastive learning with memory banks or generative pretraining, which could offer further improvements.

The scalability of our approach to larger, more diverse datasets also warrants attention. While we have demonstrated the framework's effectiveness using the DSIAC dataset, real-world ATR tasks often involve a wider variety of targets and environmental conditions. Exploring cross-domain transfer learning and expanding our evaluation to more complex datasets will provide further insights into the generalizability of the model.

From a computational perspective, training large-scale multi-sensor models can be resource-intensive. Future work should focus on optimizing training efficiency through techniques like model pruning, knowledge distillation, or distributed training, which could make the model more practical for deployment in resource-constrained environments.

Finally, there is an exciting opportunity to improve the interpretability of the model. Incorporating explainability techniques, such as attention-based visualization and decision boundary analysis, would not only help validate the model's decisions but also build trust in its predictions—especially for high-stakes applications like defense and security.

8. CONCLUSION

In this paper, we introduced an adaptive self-supervised Vision Transformer (ViT) framework for Automatic Target Recognition (ATR) that integrates hierarchical transformer architectures with masked autoencoding (MAE) and contrastive learning. Our approach also includes a dynamic hyperparameter tuning mechanism that adjusts drop path probabilities based on attention entropy, allowing the model to self-regulate and improve generalization during training.

Experimental results demonstrate the effectiveness of our framework on electro-optical (EO) and infrared (IR) datasets. We found that EO data benefits most from the combined SSL approach (MAE + contrastive learning), while IR data performs better with a reconstruction-focused approach. The qualitative analysis supports these findings, with attention maps showing the model's ability to focus on discriminative features despite imperfect reconstructions.

Overall, this work demonstrates the potential of adaptive self-supervised learning in ATR, particularly for multi-sensor scenarios where traditional supervised methods struggle. Future work will explore the integration of multi-sensor data, advanced regularization techniques, and enhanced attention mechanisms to further improve model robustness and scalability.[‡]

ACKNOWLEDGMENTS

This work was funded by the DEVCOM Army Research Laboratory under the cooperative agreement W911NF-20-2-0218.

REFERENCES

[1] Bhanu, B., "Automatic target recognition: State of the art survey," *IEEE transactions on aerospace and electronic systems* (4), 364–379 (2007).

[2] Liu, Z., Xiao, G., Liu, H., and Wei, H., "Multi-sensor measurement and data fusion," *IEEE Instrumentation & Measurement Magazine* **25**(1), 28–36 (2022).

[3] Blasch, E., Pham, T., Chong, C.-Y., Koch, W., Leung, H., Braines, D., and Abdelzaher, T., "Machine learning/artificial intelligence for sensor data fusion–opportunities and challenges," *IEEE aerospace and electronic systems magazine* **36**(7), 80–93 (2021).

[4] Jaiswal, A., Babu, A. R., Zadeh, M. Z., Banerjee, D., and Makedon, F., "A survey on contrastive self-supervised learning," *Technologies* **9**(1), 2 (2020).

[5] Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., and Joulin, A., "Emerging properties in self-supervised vision transformers," in [*Proceedings of the IEEE/CVF international conference on computer vision*], 9650–9660 (2021).

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

[7] He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R., "Masked autoencoders are scalable vision learners," in [*Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*], 16000–16009 (2022).

[8] Ratches, J. A., "Review of current aided/automatic target acquisition technology for military target acquisition tasks," *Optical Engineering* **50**(7), 072001–072001 (2011).

[9] Low, S., Nina, O., Sappa, A. D., Blasch, E., and Inkawhich, N., "Multi-modal aerial view object classification challenge results-pbvs 2023," in [*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*], 412–421 (2023).

[‡]This manuscript was partially assisted by GPT, an AI language model, which helped in improving the clarity and structure of the writing.

[10] Defense Systems Information Analysis Center, "ATR Algorithm Development Image Database." https://dsiac.org/databases/atr-algorithm-development-image-database/.

[11] Mahalanobis, A. and McIntosh, B., "A comparison of target detection algorithms using dsiac atr algorithm development data set," in [*Automatic Target Recognition XXIX*], **10988**, 47–51, SPIE (2019).

[12] Sakla, W., Konjevod, G., and Mundhenk, T. N., "Deep multi-modal vehicle detection in aerial isr imagery," in [*2017 IEEE winter conference on applications of computer vision (WACV)*], 916–923, IEEE (2017).

[13] Blasch, E., Majumder, U. K., Rovito, T., Zulch, P., and Velten, V. J., "Automatic machine learning for target recognition," in [*Automatic Target Recognition XXIX*], **10988**, 119–130, SPIE (2019).

[14] Neary, P., "Automatic hyperparameter tuning in deep convolutional neural networks using asynchronous reinforcement learning," in [*2018 IEEE international conference on cognitive computing (ICCC)*], 73–77, IEEE (2018).

[15] Li, J., Yu, Z., Yu, L., Cheng, P., Chen, J., and Chi, C., "A comprehensive survey on sar atr in deep-learning era," *Remote Sensing* **15**(5), 1454 (2023).

[16] Rani, V., Nabi, S. T., Kumar, M., Mittal, A., and Kumar, K., "Self-supervised learning: A succinct review," *Archives of Computational Methods in Engineering* **30**(4), 2761–2775 (2023).

[17] Zhang, C., Zhang, C., Song, J., Yi, J. S. K., Zhang, K., and Kweon, I. S., "A survey on masked autoencoder for self-supervised learning in vision and beyond," *arXiv preprint arXiv:2208.00173* (2022).

[18] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929* (2020).

[19] Khan, S., Naseer, M., Hayat, M., Zamir, S. W., Khan, F. S., and Shah, M., "Transformers in vision: A survey," *ACM computing surveys (CSUR)* **54**(10s), 1–41 (2022).

[20] Chen, X., Xie, S., and He, K., "An empirical study of training self-supervised vision transformers," in [*Proceedings of the IEEE/CVF international conference on computer vision*], 9640–9649 (2021).

[21] Atito, S., Awais, M., and Kittler, J., "Sit: Self-supervised vision transformer," *arXiv preprint arXiv:2104.03602* (2021).

[22] Ferreira, L., Pilastri, A., Martins, C. M., Pires, P. M., and Cortez, P., "A comparison of automl tools for machine learning, deep learning and xgboost," in [*2021 International Joint Conference on Neural Networks (IJCNN)*], 1–8, IEEE (2021).

[23] Abraham, S., Kinnison, J., Miksis, Z., Poster, D., You, S., Hauenstein, J. D., and Scheirer, W., "Efficient hyperparameter optimization for atr using homotopy parametrization," in [*Automatic Target Recognition XXXIII*], **12521**, 15–23, SPIE (2023).

[24] Khan, A., Sohail, A., Fiaz, M., Hassan, M., Afridi, T. H., Marwat, S. U., Munir, F., Ali, S., Naseem, H., Zaheer, M. Z., et al., "A survey of the self supervised learning mechanisms for vision transformers," *arXiv preprint arXiv:2408.17059* (2024).