Multi-Objective Optimization with Homotopy-Based Strategies for Enhanced Multimodal Automatic Target Recognition Models

Sophia Abraham[a], Steve Cruz[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

Automatic Target Recognition (ATR) often confronts intricate visual scenes, necessitating models capable of discerning subtle nuances. Real-world datasets like the Defense Systems Information Analysis Center (DSIAC) ATR database exhibit unimodal characteristics, hindering performance, and lack contextual information for each frame. To address these limitations, we enrich the DSIAC dataset by algorithmically generating captions and proposing new train/test splits, thereby creating a rich multimodal training landscape. To effectively leverage these captions, we explore the integration of a vision-language model, specifically Contrastive Language-Image Pre-training (CLIP), which combines visual perception with linguistic descriptors. At the core of our methodology lies a homotopy-based multi-objective optimization technique, designed to achieve a harmonious balance between model precision, generalizability, and interpretability. Our framework, developed using PyTorch Lightning and Ray Tune for advanced distributed hyperparameter optimization, enhances models to meet the intricate demands of practical ATR applications. All code and data is available at `https://github.com/sabraha2/ATR-CLIP-Multi-Objective-Homotopy-Optimization`.

Keywords: automatic target recognition, multi-objective optimization, multimodal models, homotopy continuation

1. INTRODUCTION

Over the past decade, deep learning has emerged as the driving force behind advancements in machine learning, revolutionizing tasks ranging from image recognition to natural language processing. As researchers continue to push for model improvements, the practice of transferring state-of-the-art models across domains has become increasingly prevalent. However, this process is far from straightforward and often requires engineering efforts to ensure successful adaptation. A significant challenge lies in the inherent limitations of vision datasets, which typically offer a limited number of visual classes. Consequently, models trained on such datasets may excel only in specific tasks, lacking the versatility demanded by some real-world applications. Recognizing this issue, OpenAI introduced a groundbreaking neural network model named Contrastive Language-Image Pre-training (CLIP).[1] By training on a diverse array of images alongside a broad range of natural language supervision, CLIP transcended these constraints, empowering researchers to address classification requirements without the need for task-specific training data.

Automatic Target Recognition (ATR) stands out as a domain that can greatly benefit from the potential of CLIP. In ATR, systems are tasked with the detection, identification, and classification of objects or targets within sensor data, a crucial function in military applications such as target tracking and surveillance. While conventional vision models may aid in target recognition, they often lack the contextual understanding provided by textual information. On the other hand, CLIP has the unique capability to leverage textual descriptions associated with targets, enabling the model to produce more robust representations that capture both visual and semantic information.

Inspired by promising implications, our research aims to showcase an example in ATR that underscores the innovative capabilities of the CLIP model. ATR extends beyond simple object detection, addressing complex scenarios such as low resolution and long-range targets, often observed in sensors like infrared, thermal, or Synthetic Aperture Radar (SAR). To explore these nuances, we leverage the Defense Systems Information Analysis

Center (DSIAC) ATR Database[2] for its unique characteristics. Given the absence of contextual annotations in the dataset, we generate descriptive captions for each frame. Additionally, we explore optimizing learning models, where our focus lies on a homotopy-based multi-objective framework aimed at strategically balancing essential aspects of CLIP model training: **(1)** minimizing contrastive loss to amplify discriminative accuracy and **(2)** maximizing similarity scores to better synchronize text and image representations. While prior work on YOLOv5[3] demonstrated the efficacy of a single-objective hyperparameter optimization strategy,[4] we chose a more nuanced approach. With these empirical observations, we aspire to fuel advancements in ATR research, ultimately contributing to target recognition and surveillance applications.

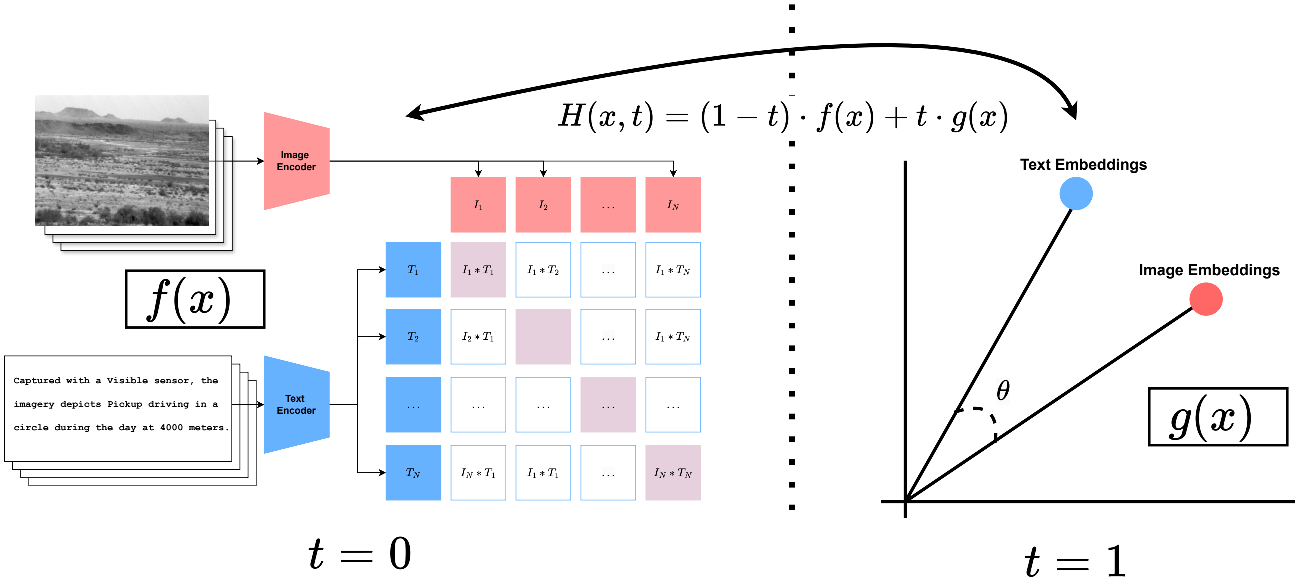


Figure 1: Homotopy-Based Multi-Objective Optimization Framework in CLIP. Our method applies a homotopy-based approach to blend Symmetric Cross-Entropy and Cosine Similarity Losses, controlled by the homotopy parameter t. This parameter evolves over training epochs, adjusting the balance between precise alignment and similarity of multimodal embeddings, to support a cohesive trajectory in learning the shared representation space.

2. BACKGROUND

To our knowledge, the potential advantages of employing a CLIP model in ATR remain unexplored. Unlike classical methods, which often rely on hand-tuned hyperparameters,[5–9] the integration of CLIP introduces a novel paradigm in target recognition. Thus, in this section, we briefly describe recent work on deep learning and hyperparameter optimization.

DeepTarget[10] is a deep learning framework that performs ATR with two VGG16[11] architectures pre-trained on ImageNet.[12] One network detects and localizes potential targets in a scene, which are then used as input to the other network to classify them into their associated target types. The experimentation showed improvements over Faster-RCNN,[13] YOLOv2,[14] and Single Shot MultiBox Detector (SSD)[15] on target detection and recognition. Note that the Comanche (BoeingSikorsky, USA) FLIR dataset used in that work is not publicly available. Chen *et al.*[16] conducted benchmarking studies on three variations of YOLOv2,[14] analyzing their performance under various conditions such as pre-training, target distances, and time of day. Another study by d'Acremont *et al.*[17] proposed a Convolutional Neural Network (CNN) architecture trained on DSIAC[2] and realistic simulated data. Their findings demonstrated superior performance compared to models trained without data augmentation or fine-tuning.

Further investigations into ATR methodologies have explored the application of Faster R-CNN[13] for infrared target detection[18] using DSIAC and even curated *easy* and *hard* test sets for comprehensive evaluation. Meanwhile, VS *et al.*[19] introduced an unsupervised domain adaptive thermal object detection framework, employing

Faster R-CNN initialized with VGG weights[11] for ATR applications on both DSIAC and KAIST.[20] Finally, Poster *et al.*[21] shed light on the challenges posed by limited-size and variability in thermal image datasets. They proposed a novel CNN architecture tailored for small object detection in data-limited settings and introduced a more rigorous evaluation protocol for DSIAC to assess model generalizability. Their model exhibited superior performance, achieving enhanced detection accuracy and minimizing overfitting.

Lastly, we acknowledge the pivotal role of hyperparameter optimization in enhancing model performance. Prior work by Abraham *et al.*[4] focused on single-objective optimization using DSIAC. The HOMOPT model provided an efficient way to search the hyperparameter space of YOLOv5[3] and identify optimal set of parameters. Building upon this, we extend our approach to a multi-objective framework using a CLIP model.

In the domain of ATR, the operational landscape is characterized by a complex array of visual signatures arising from diverse targets and backgrounds. Traditional fine-tuning approaches often drive models towards local optima, adept at distinguishing targets within a narrow subset of conditions but faltering when presented with the full spectrum of operational scenarios. Herein lies the impetus for homotopy-based multi-objective optimization—a methodology that navigates the intricate balance between discrimination and generalization. By employing a homotopy parameter that gradually shifts the focus of the training objective, we push the CLIP model to not only differentiate between targets robustly but also to align these distinctions with human-like, descriptive language cues. This technique incrementally merges the specificity required for accurate target identification with the semantic understanding necessary for flexible ATR applications.

3. METHODS

3.1 Caption Generation

Within our methodology, we prioritize the generation of descriptive captions directly from the structured nomenclature of DSIAC[2] image filenames. This process is fundamental to pairing visual data with linguistically rich annotations, enhancing the training effectiveness of CLIP models.

Initially, we employ regular expressions to accurately extract key identifiers from the filenames. These include sensor codes, scenario specifics, and target classifications. Afterwards, the codes are mapped to descriptive terms utilizing predefined dictionaries, translating them into accessible language that delineates sensor technology (*e.g.*, "Mid-Wave Infrared" vs. "Visible Light") and scenario context (detailing operational conditions such as time of day and distance). This ensures the generation of captions that accurately describe the dynamics within each frame, including the action being performed (*e.g.*, vehicles "navigating in a circular pattern" and humans "engaging in figure-8 movements").

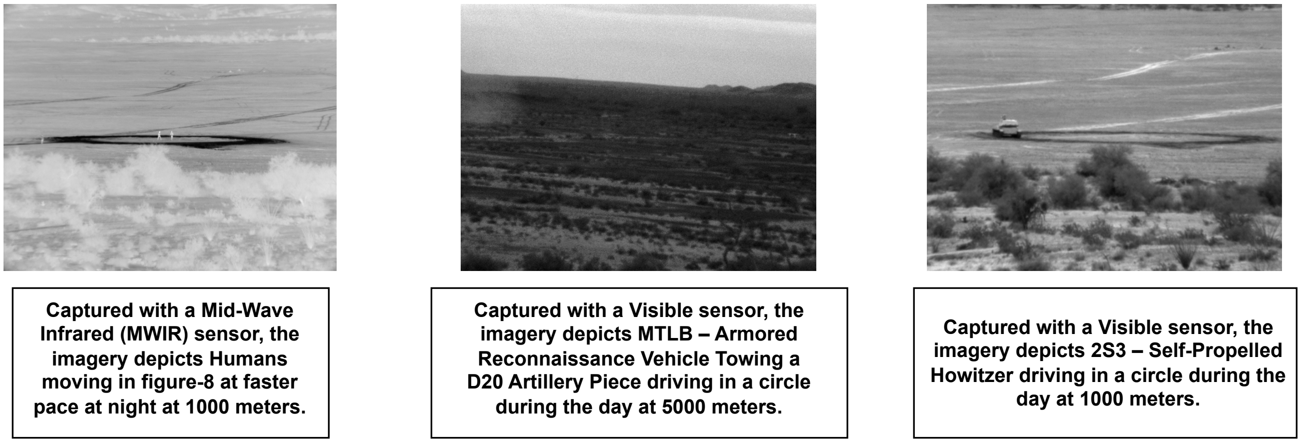
The synthesized captions serve not only to describe the visual content comprehensively but also to embed the imagery within a context that mirrors natural human observation and language use. This nuanced embedding of contextual, environmental, and operational details into the captions significantly enriches the dataset, thereby facilitating a more effective training regime for CLIP models. Examples of images and the generated text can be seen in Figure 2.

3.2 Dataset Splitting for Train and Test

In our study, we crafted the test set to ensure it encapsulates the diversity and complexity inherent in real-world ATR tasks.

We commenced by evaluating the distribution of DSIAC across various scenarios, each characterized by a unique combination of sensor type, time of day, range, and target behavior. To ensure a comprehensive assessment of the model's capabilities, we aimed to include a balanced representation of day and night scenarios, accommodating the fact that no nighttime imagery was collected with the i1co sensor. Hence, for the MWIR sensor (cegr), which possessed nighttime data, we deliberately included scenarios that would test the model's robustness under low-illumination conditions.

Also, scenarios were stratified by range, ensuring that both proximal and distant targets were represented, challenging the model's precision across varying target distances. Similarly, we ensured the inclusion of diverse target types, selecting scenarios that encompassed the full breadth of vehicle classes and human activities,



Captured with a Mid-Wave Infrared (MWIR) sensor, the imagery depicts Humans moving in figure-8 at faster pace at night at 1000 meters.

Captured with a Visible sensor, the imagery depicts MTLB – Armored Reconnaissance Vehicle Towing a D20 Artillery Piece driving in a circle during the day at 5000 meters.

Captured with a Visible sensor, the imagery depicts 2S3 – Self-Propelled Howitzer driving in a circle during the day at 1000 meters.

Figure 2: DSIAC Caption Examples. The resulting captions applied to DSIAC frames from structured image filenames. This captioning approach enhances the understanding of ATR scenarios with CLIP models through visual and contextual information.

including those less frequently observed in the dataset to test the model's recognition capabilities under sparse data conditions.

The selected scenarios for the test set were as follows:

- **Daytime Vehicle Scenarios**: We included scenario codes 2005, 2011, and 2017, providing a broad spectrum of target ranges from 1500 to 4500 meters.
- **Nighttime Vehicle Scenarios**: For the MWIR sensor data, we selected scenario codes 1925 and 1931, which offered visibility into the model's performance at 1500 and 3000 meters, respectively, under nighttime conditions.
- **Human Target Scenarios**: We reserved scenario code 02008 for daytime human targets, ensuring the model was evaluated on its ability to discern human forms at 2000 meters.

Selections were governed by a priority to challenge the model with scenarios that span a diverse array of operational conditions that test its generalization and discriminative abilities. This resulted in a train set of 368,475 image text pairs and a test set of 129,825 image text pairs.

3.3 Homtopy-Based Multi-Objective Optimization

Our method dynamically balances between two critical objectives during the training process: minimizing contrastive loss to enhance discriminative power and maximizing similarity score to improve alignment between the textual and visual representations. We employ a homotopy parameter, t, which evolves from $0-1$ over the course of training, enabling a smooth transition from focusing primarily on contrastive loss to increasingly emphasizing the similarity score. This strategy allows us to navigate the trade-offs between these objectives efficiently, aiming to find an optimal balance that fosters both precise target recognition and robust generalization across diverse scenarios. The optimization process is facilitated by a customized training loop implemented in PyTorch Lightning, which adapts the learning rate and updates the homotopy parameter based on predefined scheduling.

3.4 Experimentatal Setup and Loss Functions

The operational context of ATR imposes unique constraints and demands on machine learning models. To adapt the CLIP models to these requirements, we adopted two distinct fine-tuning strategies: a direct fine-tuning approach and a homotopy-based approach. The former seeks to establish robust image-text pair alignment via a traditional optimization route with a single objective, whereas the latter explores a dynamic balance between contrastive and similarity objectives mentioned previously, which allows for an adaptive learning trajectory. The

```
Algorithm 1 Homotopy-based Optimization for CLIP
Require: $\mathcal{D}$ (dataset), $\mathcal{M}$ (CLIP model), $\alpha_{\text{start}}$ (start learning rate), $\alpha_{\text{end}}$ (end learning rate), $T$ (total training steps)
Ensure: Optimized CLIP model $\mathcal{M}^*$
Initialize model $\mathcal{M}$ with pre-trained weights
Initialize learning rate $\alpha \leftarrow \alpha_{\text{start}}$
Initialize homotopy parameter $t \leftarrow 0$
for $step \leftarrow 1$ to $T$ do
    Sample a batch $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N \sim \mathcal{D}$
    Encode images and texts: $\mathbf{v}_i \leftarrow \mathcal{M}.\text{encode\_image}(\mathbf{x}_i), \mathbf{w}_i \leftarrow \mathcal{M}.\text{encode\_text}(\mathbf{y}_i)$
    Compute symmetric loss: $L_{\text{symmetric}} \leftarrow \frac{1}{2}(F.\mathit{cross\_entropy}(\mathbf{v}, \mathbf{w}) + F.\mathit{cross\_entropy}(\mathbf{w}, \mathbf{v}))$
    Compute similarity score: $S_{\text{similarity}} \leftarrow \text{CosineSimilarity}(\mathbf{v}, \mathbf{w})$
    Compute combined loss: $L \leftarrow (1 - t) \cdot L_{\text{symmetric}} + t \cdot S_{\text{similarity}}$
    Update $\mathcal{M}$ using gradients of $L$
    Update learning rate $\alpha$ and homotopy parameter $t$
    $\alpha \leftarrow \text{LinearSchedule}(\alpha_{\text{start}}, \alpha_{\text{end}}, T, \mathit{step})$
    $t \leftarrow \frac{\mathit{step}}{T}$
end for
$\mathcal{M}^* \leftarrow \mathcal{M}$
return $\mathcal{M}^*$
```

allocation of 32 epochs for the direct fine-tuning approach was to ensure ample opportunity for the model to refine its learning of complex ATR patterns. Meanwhile, the homotopy-based approach required only 16 epochs due to its efficient traversal through the optimization landscape, shifting focus from contrastive discrimination to similarity maximization as training progressed. This dual strategy ensures comprehensive coverage of the learning spectrum, facilitating a nuanced understanding of the ATR domain challenges. Both methods drew upon the foundational pre-training afforded by the train-CLIP library, a versatile tool for CLIP model training created by Cade Gordon,[22] ensuring a robust starting point for our specialized fine-tuning tasks.

3.4.1 Direct Fine-Tuning Approach

Direct fine-tuning employs a bidirectional retrieval framework aimed at optimizing the alignment between image and text representations. This alignment is achieved through a contrastive learning strategy that utilizes symmetric cross-entropy loss, calculated based on similarity scores (logits) derived from dot products of image-text pair embeddings within each batch. The process unfolds as follows:

1. **Logits Calculation:** For each image-text pair within a batch, embeddings are computed and normalized. The similarity scores are then obtained by calculating the dot products of these embeddings. These scores are subsequently scaled by $e^{\text{logit\_scale}}$, which modulates the distribution's sharpness, akin to temperature scaling observed in the NT-Xent[23] loss:

$$\text{logits} = e^{\text{logit\_scale}} \times (\text{image\_embeddings} \cdot \text{text\_embeddings}^T)$$

2. **Symmetric Cross-Entropy Loss Application:** The symmetric cross-entropy loss[24] is applied to the calculated logits in a dual direction—once for images predicting corresponding texts, and once for texts predicting corresponding images. This dual application ensures that the correct matches (identified by the diagonal elements of the logits matrix) effectively guide the loss computation, thereby conceptualizing the task as a classification challenge:

$$L_{\text{sym}} = -\frac{1}{2N}\left(\sum_{\text{images}} \log \frac{e^{\text{logits}_{ii}}}{\sum_j e^{\text{logits}_{ij}}} + \sum_{\text{texts}} \log \frac{e^{\text{logits}_{jj}}}{\sum_i e^{\text{logits}_{ij}}}\right)$$

3. **Loss Averaging:** To ensure balanced penalization for alignment inaccuracies, the losses calculated from both directional predictions are averaged, producing the final loss for the training step.

4. **Learning Rate Scheduling:** We adopt a learning rate schedule that initiates at 5×10^{-4} and methodically decays to 2×10^{-5}, in line with a cosine annealing schedule with warm-up restarts. This pacing is meticulously designed to match the evolving emphasis of the training objectives.

5. **Mini-batching Strategy:** To address computational constraints effectively, training data is partitioned into smaller mini-batches, enabling the practical application of our homotopy-based method across extensive datasets.

This direct fine-tuning approach allows for refined alignment between the visual and textual modalities by focusing on minimizing symmetric cross-entropy loss, thereby enhancing the model's performance in bidirectional retrieval tasks.

3.4.2 Homotopy-Based Training Approach

In this study, we implement a homotopy-based training methodology, building upon the direct tuning strategy, for the CLIP model, balancing between the symmetric cross-entropy loss and similarity maximization to refine the alignment between image and text representations. Central to our approach is the homotopy parameter t, which progressively transitions from 0 to 1 throughout the training epochs, facilitating a seamless shift from focusing on symmetric loss reduction to emphasizing cosine similarity maximization between matching image-text pairs. The combined objective function is formulated as:

$$L_{\text{combined}} = (1 - t) \cdot L_{\text{sym}} + t \cdot L_{\text{similarity}}$$

where L_{sym} denotes the average symmetric cross-entropy loss and $L_{\text{similarity}} = -\frac{1}{N} \sum_{i=1}^{N} \text{sim}(v_i, w_i)$ signifies the negative average cosine similarity across normalized embeddings of corresponding image-text pairs, v_i and w_i. The negative sign preceding the similarity score elucidates our intent to maximize this metric, in stark contrast to minimizing the symmetric loss. This dual-objective strategy ensures a fluid progression throughout training, enabling the model to initially harness distinct features, subsequently fine-tuning these attributes to bolster the congruence between visual and textual modalities.

3.5 Evaluation of Tuned CLIP Models on the DSIAC ATR Test Set

In addition to the standard evaluation on an unseen test, we also report the training dynamics of both methods in terms of the loss and retrieval accuracy. The model's retrieval accuracy, both in the image-to-text and text-to-image directions, is quantified by measuring the proportion of correctly identified matches based on the argmax of the logits. For the test evaluation, batches of images and text pairs were processed by the fine tuned models to generate embeddings, which were then subjected to cosine similarity analysis. This analysis quantitatively measured the alignment between each image-text pair, offering insights into the ability of the models to correctly associate visual content with textual descriptions under various ATR scenarios. Subsequent visualization using t-SNE provided a qualitative assessment of the embedding spaces, revealing the models' discriminative capabilities and their robustness in handling the ATR task.

Particularly of note were the evaluations conducted under low-illumination conditions and at extended ranges, which were critical for assessing the models' performance in operationally challenging scenarios typical of the DSIAC ATR environment. The test set, with 129,825 image-text pairs, served not only as a measure of the models' current capabilities but also as a benchmark for future enhancements. By highlighting the strengths and identifying potential areas for improvement, the evaluation process has laid the groundwork for developing more advanced models that can effectively navigate the complexities of real-world Automatic Target Recognition.

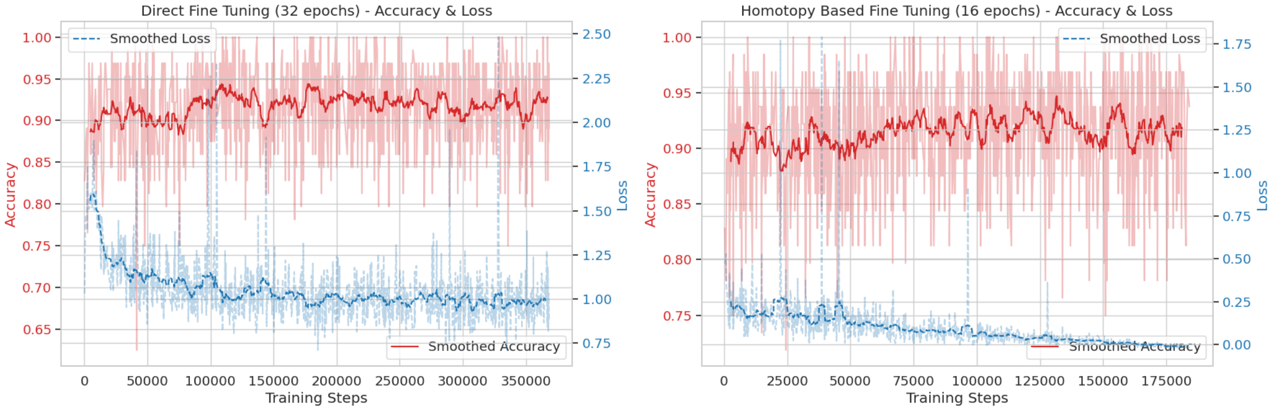


Figure 3: Comparison of Fine-Tuning Approaches on CLIP Model's Training Performance. These plots illustrate a comparison of each of the fine tuning method's training performance. A smoothed line averaging over 20 samples is plotted over the raw trajectories. The left graph presents the direct fine tuning method, illustrating episodic variations in accuracy (red) and loss (blue) across the full span of training iterations, indicative of the model's adaptive response to the fine-tuning process. The right graph shows the multi-objective homotopy method limited to 16 epochs, where a discernible pattern of steady accuracy improvement and consistent loss minimization is observed.

4. RESULTS

4.1 Training Dynamics

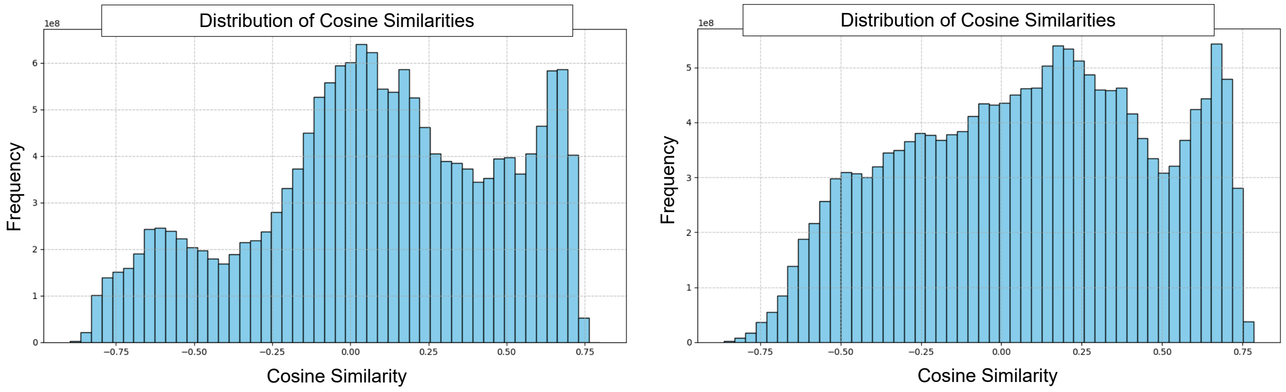
Our comparative analysis of the training dynamics between direct fine-tuning and homotopy-based multi-objective fine-tuning of a CLIP model reveals distinct patterns in both accuracy and loss over training epochs. The direct fine-tuning approach demonstrates a rapid improvement in accuracy in the initial epochs, followed by a plateau, suggesting quick adaptability to the training data. However, the loss trends for direct fine-tuning show fluctuations, indicating potential instability in learning. Conversely, the homotopy-based approach exhibits a steadier, albeit slower, increase in accuracy, coupled with a consistent decrease in loss, suggesting a more stable and gradual learning process.

The training dynamics observed suggest differing efficiencies and effectiveness of the two methods. Direct fine-tuning achieves high accuracy quickly but at the cost of potential overfitting, as indicated by the fluctuating loss. This might suggest that while the model rapidly aligns with the training data's features, it may not generalize well to unseen data. On the other hand, the homotopy-based approach with consistent loss reduction and gradual accuracy improvement indicate a more balanced learning process, possibly leading to better generalization. This method seems to mitigate overfitting more effectively, potentially due to its multi-objective optimization that balances different aspects of the learning process.

4.2 Test Evaluation

The histograms in Figure 4 depict the distribution of cosine similarities between image and text embeddings for two distinct fine-tuning methods applied to a CLIP model, benchmarked on the DSIAC ATR data. Figure 4a presents the distribution resulting from the direct fine tuning. This distribution is characterized by a broad spread of cosine similarities with multiple peaks, suggesting a diverse range of alignment between the image-text pairs. The peaks around the higher similarity values indicate a significant portion of tightly aligned pairs, which is indicative of the model's capability to correlate relevant features across modalities. However, the broader spread towards lower similarities implies the existence of a substantial number of loosely aligned or even misaligned pairs.

In contrast, the homotoy based fine tuning method, as shown in Figure 4b, yields a distinctly different distribution. Here, the histogram exhibits a more pronounced skew towards higher similarity values and a narrower spread, indicative of a more consistently aligned set of embeddings. The presence of a peak towards the higher end of the similarity spectrum and fewer low-similarity outliers suggests that this method fosters a closer semantic connection between the paired embeddings. Notably, the histogram lacks the extended tail towards

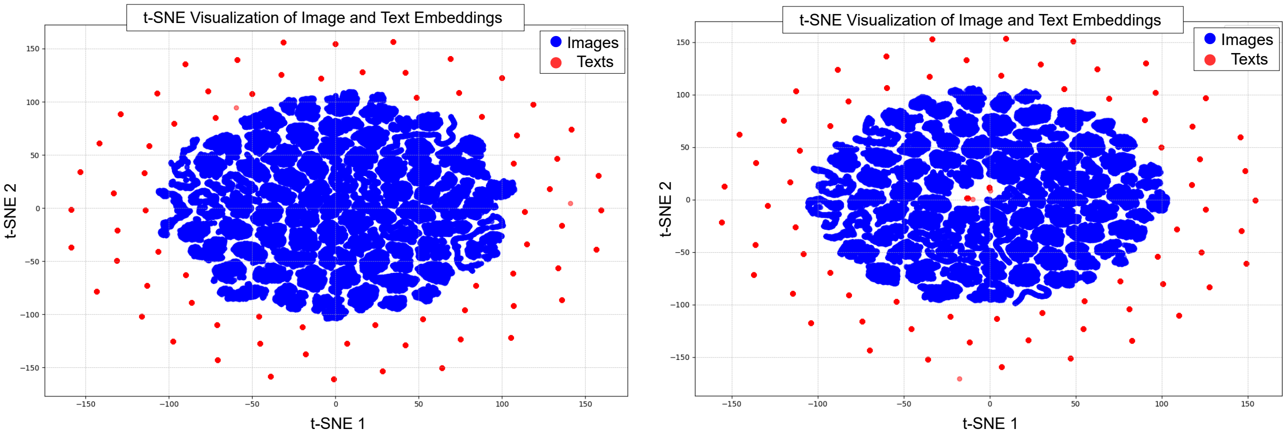


(a) Direct Fine Tuning

(b) Homotopy-Based Fine Tuning

Figure 4: Similarities Distribution. Histograms showcasing the distribution of cosine similarities between image and text embeddings for two fine-tuning approaches on a CLIP model. The left histogram (a) corresponds to direct fine tuning, which demonstrates a varied distribution of similarities, indicative of a strong discrimination capability but possibly less cohesion in image-text alignment. The right histogram (b) corresponds to homotopy-based fine tuning, which shows a shift towards higher similarity scores, illustrating the method's effectiveness in creating more aligned and cohesive image-text representations as the training evolves from contrastive loss towards similarity score maximization

the negative similarity values seen in the direct fine tuning method, implying a reduced incidence of strongly misaligned pairs.



(a) TSNE Plot for Direct Fine Tuning

(b) TSNE Plot for Homotopy-Based Fine Tuning

Figure 5: Distribution of Image and Text Embeddings. The direct fine-tuning plot shows a denser clustering, indicating tight but potentially overfitted correlations. In contrast, the homotopy-based plot reveals a more dispersed distribution, suggesting a flexible representation space that may facilitate better generalization to unseen data. The presence of a central void in the homotopy-based plot (b) could reflect an intentional emphasis on semantic similarity over mere feature matching, a characteristic development of the homotopy approach which gradually shifts from contrastive loss to similarity maximization over the course of training.

Figure 5 provides a comparative visualization of the high-dimensional embedding spaces reduced via t-SNE to contrast directly each of the fine tuning methods. For the direct fine tuning method (Figure 5a), the embeddings are characterized by a more uniform distribution across the t-SNE plot. The image (blue) and text (red) embeddings display a significant degree of overlap, with no clear demarcation between the modalities, implying a more generalist feature capture that may not distinctly differentiate fine-grained semantic nuances.

The homotopy-based multi-objective fine tuning (Figure 5b) presents an embedding space with several notable distinctions. The embeddings appear as denser clusters, signifying a more defined feature space. A remarkable feature of this plot is the central depletion or ‘hole,’ which suggests that as the fine-tuning transitions from a contrastive loss that emphasizes direct image-text pair similarities towards a similarity loss that aligns broader semantic relationships, it fosters a more structured separation between distinct clusters. This could reflect a refined mapping where embeddings are not merely drawn to their exact matches but are also repelled from near-miss pairings, enhancing the model’s discriminative clarity.

5. DISCUSSION

The differential distributions of cosine similarities observed provide insights into their respective impacts on the semantic alignment of the CLIP model’s embeddings. The broader distribution of cosine similarities for the direct fine-tuning approach reflects a less discriminating optimization process, possibly due to a single-objective focus over an extended training period. Conversely, the homotopy-based approach skews towards higher similarity values, suggesting an optimization trajectory that prioritizes the cohesion of semantic associations between images and texts.

The more concentrated distribution of cosine similarities in the homotopy-based approach underscores its potential in creating a more precise and semantically rich embedding space. However, it is imperative to consider the potential overfitting to the training data, particularly given the complex and varied nature of DSIAC. While the homotopy-based method may yield a model adept at capturing the nuanced semantics within a constrained evaluation set, further analysis is required to ascertain its performance across the broader, more diverse spectrum of ATR conditions.

Overall, the findings suggest that the homotopy-based fine-tuning method’s multi-objective optimization produces an embedding space optimized for differentiation, which could be preferable for tasks where the distinction between categories is paramount. Future work should investigate the impact of the central void on retrieval performance and explore the balance between specialization and generalization in the embedding space. Quantitative analyses, such as precision-recall curves and F1 scores for retrieval tasks across various categories, will be essential to validate the insights provided by the t-SNE visualizations and to ascertain the practical implications of each fine-tuning strategy.

6. CONCLUSION

This work sought to enhance the understanding and performance of CLIP models within the specialized domain of Automatic Target Recognition (ATR), leveraging the nuanced and diverse DSIAC ATR Database. Our methodology encompassed the generation of descriptively rich captions from DSIAC image filenames, employing them with corresponding images to train the models through a homotopy-based multi-objective approach. The meticulous curation of the test set, which mirrored the multifaceted reality of ATR scenarios, establishes a rigorous benchmark for model evaluation.

The direct fine-tuning method, applied over an extended period of 32 epochs, equipped the model with a broad semantic understanding, as evidenced by the t-SNE visualization and the distribution of cosine similarities. Meanwhile, the homotopy-based fine-funing, conducted over 16 epochs, demonstrated a discernible shift towards tighter semantic alignment, with the potential for heightened discriminative prowess in ATR tasks. The distinct embedding landscapes revealed by the t-SNE plots and the comparative analysis of cosine similarity distributions provides compelling visual and quantitative testimony to the efficacy of the training methodologies.

Looking forward, the insights gleaned from this research advocate for a nuanced approach to model training, where the choice of methodology aligns with the operational imperatives of ATR systems. The homotopy-based approach, in particular, holds promise for future exploration, especially in its potential to balance contrasting training objectives efficiently. However, recognizing the limited scope of any single study, we advocate for ongoing research to validate these findings across larger datasets and in operational environments.

ACKNOWLEDGMENTS

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

REFERENCES

[1] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al., "Learning Transferable Visual Models From Natural Language Supervision," in [*International Conference on Machine Learning*], 8748–8763, PMLR (2021).

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

[3] Glenn Jocher, "Ultralytics YOLOv5," (2020). https://github.com/ultralytics/yolov5.

[4] 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).

[5] Zhou, Y.-T. and Crawshaw, R. D., "Contrast, Size, And Orientation-Invariant Target Detection In Infrared Imagery," in [*Automatic Object Recognition*], **1471**, 404–411, SPIE (1991).

[6] Gregoris, D. J., Simon, K., Tritchew, S., and Sevigny, L., "Wavelet Transform-Based Filtering For The Enhancement Of Dim Targets In FLIR Images," in [*Wavelet Applications*], **2242**, 573–583, SPIE (1994).

[7] Mahalanobis, A., Muise, R. R., and Stanfill, S. R., "Quadratic correlation filter design methodology for target detection and surveillance applications," *Applied Optics* **43**(27), 5198–5205 (2004).

[8] Musicki, D. and Evans, R., "Clutter Map Information for Data Association and Track Initialization," *IEEE Transactions on Aerospace and Electronic Systems* **40**(2), 387–398 (2004).

[9] Yoon, S. P., Song, T. L., and Kim, T. H., "Automatic Target Recognition and Tracking in Forward-Looking Infrared Image Sequences with a Complex Background," *International Journal of Control, Automation and Systems* **11**(1), 21–32 (2013).

[10] Nasrabadi, N. M., "DeepTarget: An Automatic Target Recognition Using Deep Convolutional Neural Networks," *IEEE Transactions on Aerospace and Electronic Systems* **55**(6), 2687–2697 (2019).

[11] Simonyan, K. and Zisserman, A., "Very Deep Convolutional Networks for Large-Scale Image Recognition," in [*International Conference on Learning Representations*], (2015).

[12] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al., "ImageNet Large Scale Visual Recognition Challenge," *International journal of computer vision* **115**, 211–252 (2015).

[13] Ren, S., He, K., Girshick, R., and Sun, J., "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," *Advances in Neural Information Processing Systems* **28** (2015).

[14] Redmon, J. and Farhadi, A., "YOLO9000: Better, Faster, Stronger," in [*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*], 7263–7271 (2017).

[15] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., and Berg, A. C., "SSD: Single Shot MultiBox Detector," in [*Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*], 21–37, Springer (2016).

[16] Chen, H.-W., Gross, N., Kapadia, R., Cheah, J., and Gharbieh, M., "Advanced Automatic Target Recognition (ATR) with Infrared (IR) Sensors," in [*2021 IEEE Aerospace Conference (50100)*], 1–13, IEEE (2021).

[17] d'Acremont, A., Fablet, R., Baussard, A., and Quin, G., "CNN-Based Target Recognition and Identification for Infrared Imaging in Defense Systems," *Sensors* **19**(9), 2040 (2019).

[18] 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).

[19] VS, V., Poster, D., You, S., Hu, S., and Patel, V. M., "Meta-UDA: Unsupervised Domain Adaptive Thermal Object Detection using Meta-Learning," in [*Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*], 1412–1423 (2022).

[*] We also acknowledge the assistance of ChatGPT, a language model trained by OpenAI, for providing helpful suggestions during the writing of this manuscript.

[20] Hwang, S., Park, J., Kim, N., Choi, Y., and So Kweon, I., "Multispectral Pedestrian Detection: Benchmark Dataset and Baseline," in [*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*], 1037–1045 (2015).

[21] Poster, D., Hu, S., and Nasrabadi, N. M., "Long-Range Thermal Target Detection in Data-Limited Settings Using Restricted Receptive Fields," *Sensors* **23**(18), 7806 (2023).

[22] Gordon, C., "train-clip." https://github.com/Zasder3/train-CLIP (2021).

[23] Sohn, K., "Improved deep metric learning with multi-class n-pair loss objective," *Advances in neural information processing systems* **29** (2016).

[24] Wang, Y., Ma, X., Chen, Z., Luo, Y., Yi, J., and Bailey, J., "Symmetric cross entropy for robust learning with noisy labels," in [*Proceedings of the IEEE/CVF international conference on computer vision*], 322–330 (2019).