Point-Source Target Detection and Localization in Single-Frame Infrared Imagery

Daniel C. Stumpp[1,2], Andrew J. Byrne[2], and Alan D. George[1]
[1]NSF-SHREC Center, University of Pittsburgh, 4420 Bayard St, Pittsburgh, PA 15213, USA
[2]The MITRE Corporation, 202 Burlington Rd, Bedford, MA 01730, USA
dcs98@pitt.edu, abyrne@mitre.org, and alan.george@pitt.edu

Abstract*—Potential military threats often manifest as dim point-source targets embedded in complex clutter and noise backgrounds, which makes threat detection a significant challenge. A variety of machine-learning architectures have been developed in recent years for performing small-object segmentation in single frames of infrared imagery. Evaluation and comparison of these techniques has been hampered by a lack of reliably labeled data and the use of different evaluation metrics. In this research, we leverage the Air Force Institute of Technology Sensor and Scene Emulation Tool (ASSET) to generate a dataset containing independent frames with unique background and target characteristics. We introduce a standardized method for generating ground-truth segmentation masks for point-source targets that eliminates the risk of manual labeling errors that exist in other small-target segmentation datasets. A local peak signal-to-clutter-and-noise ratio ($pSCNR$) is also introduced and shown to be strongly correlated to probability of detection. Results show that with the use of the generated dataset, existing state-of-the-art small-object segmentation networks can be adapted specifically to the point-source target detection task. A probability of detection (P_d) greater than 80% is consistently achieved while maintaining low false alarm rates. In addition to the task of target detection, we address the problem of target subpixel localization in a single frame. Accurate subpixel localization is important due to the large physical area included in a single pixel. Existing work commonly overlooks this problem or takes the predicted target mask centroid as the subpixel location. In this research, we introduce a transformer-based subpixel localization technique that uses both the predicted target mask and the local pixel intensity to compute an accurate subpixel location. The proposed architecture reduces mean localization error by up to 72% compared to other single-frame methods for target subpixel localization.

TABLE OF CONTENTS

This work was supported in part by NSF Center for Space, High-Performance, and Resilient Computing (SHREC) Industry and Agency Members, and in part by the IUCRC Program of the National Science Foundation under Grant CNS-1738783.

2023 IEEE Aerospace Conference | 978-1-6654-9032-0/23/$31.00 ©2023 IEEE | DOI: 10.1109/AERO55745.2023.10115662

1. INTRODUCTION

Missile early-warning systems rely on electro-optical and infrared sensors for space-based observation of potential threats. In overhead persistent infrared (OPIR) imagery, these threats often manifest as dim point-source targets embedded in complex clutter and noise backgrounds. The characteristics of this problem are similar to the generic infrared small-object detection problem explored for a wide range of applications. These applications, including the OPIR target detection and missile-warning tasks, are often characterized by high-velocity targets, sensor motion, and unpredictable dynamic backgrounds. These characteristics have led to a shift from temporal methods of detection towards single-frame methods for small-object detection in infrared imagery [1]. Early techniques for single-frame target detection relied on mathematical models of target and background characteristics. Recent research, however, has led to a growth in the use of machine-learning (ML) methods, which have been shown to outperform traditional signal processing techniques at the task of small-object detection. These methods approach the small-object detection problem as a binary segmentation task separating target from background. In this research, we explore the application of state-of-the-art small-object segmentation networks to the task of point-source target segmentation and detection in OPIR data.

Evaluation of existing state-of-the-art methods for the task of point-source target segmentation in OPIR data is hampered by the limitations of available datasets. Existing datasets are crafted for the generic small-object segmentation task and contain targets that are too large in addition to backgrounds that are inconsistent with OPIR data. These data limitations have prevented an accurate evaluation of network performance specific to the OPIR target detection and missile-warning tasks. To address these limitations, we generate a dataset specifically designed for this evaluation using the AFIT Sensor and Scene Emulation Tool (ASSET) [2], [3]. A peak signal-to-noise-and-clutter metric is introduced and calculated using the simulated data. This metric is used to evaluate target detection difficulty and the performance of the detection methods tested. Additionally, we introduce a quantitative method for generating ground-truth segmentation masks based on the output of ASSET. This standardization of ground truth improves overall data consistency and allows for accurate target- and pixel-level evaluation of segmentation performance.

The OPIR target detection task requires accurate subpixel localization because the point-source targets are unresolved. Little emphasis is placed on subpixel localization for generic small-object segmentation. However, the characteristics of OPIR sensors are such that individual pixels cover a large physical area, meaning that accurate subpixel localization is required for determining the true position of the target in the infrared frame. In this research, we introduce a transformer-

based method for subpixel localization that uses both raw image data and the predicted segmentation mask to accurately localize the target. This method provides substantial improvement in localization accuracy while adding limited additional complexity to the overall detection and localization pipeline.

2. Related Research

There is a significant amount of related research in the area of single-frame infrared small-object detection. Most methods formulate the problem as a segmentation task and then process clusters to produce final detections. We categorize existing methods into one of two categories: traditional methods or ML methods. Traditional methods typically formulate the segmentation and detection around a mathematical model or filter designed based on prior assumptions about the targets, background, and noise present in a frame. ML methods leverage representative data to train networks for segmentation and detection of targets and are able to learn complex target representations. These ML methods are the focus of the evaluation in Section 5 and have consistently been shown to outperform traditional methods in the research discussed in the following section.

Traditional Methods

Early methods relied on filtering to amplify target signal. Methods such as max-mean filtering [4], top-hat filtering [5], directional high-pass filtering [6], and minimum local Laplacian of Gaussian (min-local-LoG) filtering [7] explored the design of appropriate filters for the task. These methods were sufficient for the detection of bright targets with high signal-to-noise ratios, however, they generally struggled with dim targets and resulted in high numbers of false alarms. The infrared patch image (IPI) model formulates the target segmentation task as an optimization problem where the target and background components of the frame are considered sparse and low rank matrices respectively [8]. While the IPI model exhibited state-of-the-art performance when introduced, it is sensitive to parameter tuning and fails in cases where the underlying assumptions about the target and background characteristics fail.

Many recent traditional methods have relied on the computation of a local contrast measure (LCM) to amplify target signal and suppress background in the infrared frame. The first LCM used for small-object segmentation in a single frame was introduced in [9] and has served as the basis for future methods. The multiscale patch-based contrast measure (MPCM) computes contrast between patches and successfully adapts to various target sizes by performing the measurement at multiple scales [10]. Other methods inspired by LCM include the weighted local difference measure (WLDM) [11], a multiscale LCM with high boost filtering [12], multiscale relative LCM [13], and homogeneity-weighted LCM [14]. Beyond local contrast methods, many other approaches have been proposed, including Gaussian transformation [15], principal component analysis (PCA) [16], and entropy-based window selection [17]. Although traditional methods have improved, they remain vulnerable to cases where their underlying assumptions fail. These failures often occur in cases of dim targets embedded in complex spatial backgrounds as is often the case for point-source targets in OPIR data.

Machine Learning Methods

ML methods have become the dominant approach for the generic small-object segmentation task. They have been shown to consistently outperform the traditional methods discussed when evaluated across a wide range of datasets. There have also been some examples of exploration of ML specific to the target detection task in OPIR data. A 3D convolution and long short-term memory (LSTM) hybrid network was proposed in [18] for missile detection and tracking. However, this approach relied on sequential frames and used a small frame size of only 32×32 pixels [18]. For the reasons outlined in Section 1, we focus in this research on ML methods that operate on a single frame. A variety of ML methods meeting that requirement have been proposed. The following sections present, in order of introduction, the ML architectures that will be explored in this research for the task of point-source target segmentation in OPIR data. These network architectures were originally introduced for the generic small-object segmentation task that includes detection of planes, ships, and other targets, but have promise for performance on point-source segmentation.

Miss Detection vs. False Alarm cGAN—One of the earliest successful ML methods proposed was miss detection vs. false alarm (MDvsFA) [19]. MDvsFA adversarially trained a conditional generative adversarial network (cGAN) with two generators to achieve opposing tasks of minimizing missed detections and minimizing false alarms. Since MDvsFA, other research has explored variations of GAN architectures for infrared target extraction such as in [20].

Asymmetric Context Modulation—The asymmetric context modulation (ACM) module was introduced in [21] to help overcome the inherent challenge of small target features being lost in deep networks. ACM introduces a bottom-up modulation cross-layer feature fusion approach that allows for the combination of low and high-level features [21]. ACM is a module that can be dropped into a host network at cross-layer connections. It was demonstrated in the U-Net [22] and feature pyramid network (FPN) [23] host architectures. The single-frame infrared small target (SIRST) dataset used commonly in other related research is also introduced in [21].

Attentional Local Contrast Network—Expanding on ACM, [24] introduces the attentional local contrast network (ALCNet). This network is unique as it explicitly includes a local contrast prior in the end-to-end network. ALCNet introduces a modified version of MPCM [10] that can be applied to downsampled feature maps and applies it to the network skip connections. ALCNet includes similar bottom-up modulation as introduced in ACM [24]. The authors demonstrate that the inclusion of the local contrast prior allows state-of-the-art performance to be achieved with much smaller networks.

Local Similarity Pyramid Module—The network architecture introduced in [25] computes local similarity across multiple scales to learn rich target representations. These local similarity pyramid modules (LSPM) are combined with a feature aggregation module using channel attention to perform cross-layer fusion. Unlike most other network architectures reviewed, LSPM uses a VGG-16 backbone instead of a ResNet backbone for feature extraction.

Dense Nested Attention Network— Inspired by a stacked U-Net architecture, the dense nested attention network (DNANet) uses multiple U-Net encoder-decoder structures to improve the quality of the resulting segmentation map [26]. Feature maps from each downsampled layer are upsampled

and feature fusion is performed using both channel and spatial attention. Results reported in [26] show that DNANet outperforms ACM, ALCNet, and MDvsFA on the dataset they present called NUDT-SIRST [26]. Importantly, this dataset includes extended targets as well as point-source and spot targets, meaning that the comparison performed is not adequate to predict the performance on a dataset specifically designed for point-source target detection in OPIR data.

Attention-Guided Pyramid Context Network—The attention-guided pyramid context network (AGPCNet) [27] is based on a cross-layer fusion module similar to ACM. Additionally, AGPCNet introduces a context pyramid module that uses the non-local operation to compute global context. This global context is then used as a guide for network attention. Limited comparisons with other state-of-the-art architectures are provided in [27] and the network shows more sensitivity to the architecture configuration than is seen in other networks. However, favorable results are demonstrated in comparison to other methods after performing an architecture parameter search.

Interior Attention-Aware Network— The interior attention-aware network (IAANet) presented in [28] is unique as it first uses a region proposal network to select regions potentially containing a target. A shallow semantic generator then extracts features for each proposed region. These features are then used as the input to a transformer-based attention encoder. The transformer output is then classified to perform the semantic segmentation for the proposed target region [28]. The use of a transformer for generating rich target features was demonstrated to show promise in IAANet and [29], with both proposed networks demonstrating state-of-the-art results.

Other Machine Learning Architectures— The architectures previously outlined will be the focus for evaluation in this research. They were selected based on their varied architectures, reported success on the generic small-object segmentation task, and common use for comparison in other works. Other networks that explore small-target segmentation include PixelGame [30], EAAU-Net [31], ISTDU-Net [32], and ISNet [33]. Each of these networks formulate the small-target detection task slightly differently, however, the use of U-Net-like architectures and cross-layer feature fusion modules is still a common approach throughout.

Subpixel Localization

Subpixel localization of targets is rarely considered in existing research. Many of the ML methods use only pixel-level metrics for evaluation and therefore do not need a continuous subpixel location for evaluating performance [21], [24], [27]. Methods that do compute subpixel location use the predicted target mask centroid [8], [26]. Some methods of learned subpixel object localization have been introduced for microspectroscopy [34], however, to our knowledge, no learned subpixel localization method has been proposed for the task of precisely localizing detected targets in infrared imagery.

3. DATASET GENERATION

The acquisition of realistic and reliably labeled data for training small-target segmentation networks is a significant challenge. The single-frame infrared small target (SIRST) dataset attempted to address this issue by creating a standardized dataset using independent frames of real infrared imagery [21]. SIRST and augmented versions of it have been successfully used to train and evaluate networks including ACM, ALCNet, and DNANet. However, the SIRST dataset still suffers from its use of error-prone hand labeled data and its limited number of samples (only 427 images). Additionally, the SIRST dataset includes scene and target characteristics that are inconsistent with a single application as well as infrared images generated by sensors operating at different wavelengths. To remedy these issues with existing datasets, we introduce a process for generating accurately labeled, OPIR target detection specific datasets for training and evaluation.

ASSET Simulator Configuration

The lack of publicly available real-world data for the OPIR target detection task necessitates the use of simulated data. This research uses the Air Force Institute of Technology Sensor and Scene Emulation Tool (ASSET) to generate target scenarios [2], [3]. ASSET uses physics-based models to generate sensor-realistic scene and target phenomenology. Other research has trained ML models on data generated by ASSET, however, the scope of that research remained limited. Sequences of ASSET-generated frames were used to train the 3D convolution and LSTM in [18], but frame sizes were small at only 32×32 and variation of target and background characteristics was limited. ASSET data was used for unsupervised learning in [35], but with many physical effects such as weather, cloud motion, and sensor artifacts removed from the simulation. To our knowledge, this work represents the most extensive use of ASSET for dataset generation in literature.

The generated dataset frames are size 128×128, are simulated from a geostationary orbit, and contain a single inserted target. Satellite nadir, relative aim-point, target location, and hardware noise are all randomized. The targets are simulated with a uniform distribution of signal-to-noise ratio (SNR) with respect to hardware noise. The value of SNR ranges from 1 to 750. Using these simulation configurations, 700,000 independent frames were generated. A total of 420,000 valid frames remained after removing instances where required simulation output was missing, or the target fell outside of the frame. These frames were used as the base dataset from which subsets could be created based on selection criteria.

Ground Truth Labeling

An accurate method for generation of ground-truth target masks is required because the task of small-target detection is approached as a segmentation problem by most state-of-the-art architectures. A limitation of existing datasets is inaccurate labeling of ground-truth segmentation masks, leading to inconsistency between training samples. To overcome this, a standardized method for generating ground-truth masks using ASSET simulation output is proposed.

Using the point spread function (PSF) and other sensor configuration outputs generated by ASSET, the point response function (PRF) is obtained and then normalized by the total sum of the PRF. A threshold is then applied to the normalized PRF and pixels containing more than the threshold percentage of the PRF are considered part of the target. All other pixels are considered background. The pixel that physically contains the target is always included in the target mask regardless of the PRF threshold applied. Figure 1 shows examples of the ground-truth target shapes as the threshold changes. Note that as the subpixel location of the target changes, the target energy is distributed over surrounding pix-

els differently, which results in different PRFs and generated ground-truth masks.

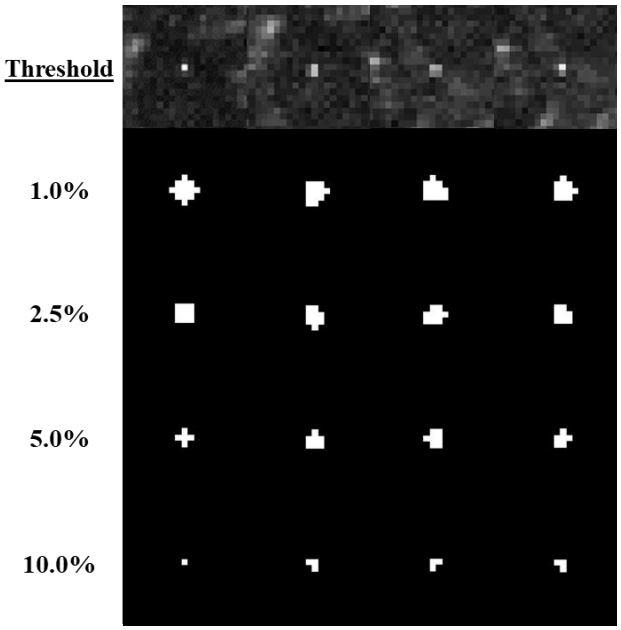


Figure 1. Example of ground-truth masks generated with four different PRF threshold values. The same target is shown with four different subpixel locations to demonstrate how the shape of the target mask changes.

Target Difficulty Metric

The SNR metric used as an input to ASSET for dataset generation only relates the target signal to the hardware noise. It does not account for background variation and clutter. To allow for evaluation of target detection difficulty, a peak signal-to-clutter-and-noise ratio ($pSCNR$) metric is introduced that accounts for spatial variations due to background. $pSCNR$ is calculated as shown in (1), where S_t is target signal inserted by ASSET and σ_b is the background noise and clutter estimate. The background noise and clutter is estimated as the standard deviation of pixel intensity in a local 35×35 region around the target. Pixels within a 5×5 region around the target are excluded to prevent target signal from being included.

$$pSCNR = \frac{max(S_t)}{\sigma_b} \tag{1}$$

This $pSCNR$ metric is used to select frames from the base dataset to be included in the primary benchmark dataset. Targets with a $pSCNR$ less than or equal to 25 are included in the benchmark dataset to be used for training and evaluation of the models being considered. Table 1 shows the statistics for the benchmark dataset being used. The resulting distribution of target $pSCNR$ values is shown in Figure 2. Overall, the dataset is skewed towards more difficult targets with a mean $pSCNR$ of 10.24.

4. SUBPIXEL LOCALIZATION

This section introduces the methods for subpixel localization to be evaluated. The proposed transformer-based architecture is presented in detail. The three other methods to be used for comparison are then described.

Table 1. Statistics for the benchmark dataset.

Frames	$pSCNR$ Range	$\overline{pSCNR}$	PRF threshold
161,000	0–25	10.24	2.5%

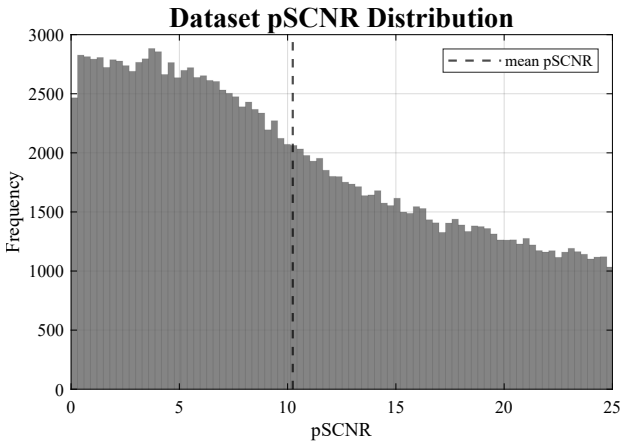


Figure 2. Distribution of $pSCNR$ for all targets in the benchmark evaluation dataset.

Network Architecture

Transformer-based models have become dominant in many vision tasks since the introduction of the vision transformer (ViT) architecture [36]. Transformers have also shown promise for application to the small-target segmentation task [28], [29]. Based on this proven effectiveness for vision and small-target tasks, a transformer-based architecture is proposed to perform subpixel localization after targets are detected.

Figure 3 shows the proposed localization network architecture. The subpixel localization problem is formulated as a local regression on a 5×5 masked input chip generated after detection. Target clusters are first detected and segmented by the segmentation network selected. A 5×5 target chip is then extracted around the centroid of the predicted target cluster. The predicted target mask is then used to remove the information from pixels not identified as part of the target. This is demonstrated by the red masking over the input chip in Figure 3. This masked input chip is then passed into a feature extractor, which first reshapes the input and then appends a learned position embedding. Three three-headed transformer encoders with a hidden dimension of 128 are then used to extract features from the input. Unlike the ViT architecture, we use individual pixels instead of 16×16 patches as the input to the transformer, allowing us to bypass the patch embedding used in ViT architectures. A simple multi-layer perceptron (MLP) is used as a regression head to produce a continuous prediction for the row and column subpixel location of the target. The ground-truth subpixel location produced by the ASSET dataset generation is used for computing loss and evaluation metrics. Mean squared error loss is used for training and Euclidean distance (L2 error) is used for evaluation and comparison of localization performance.

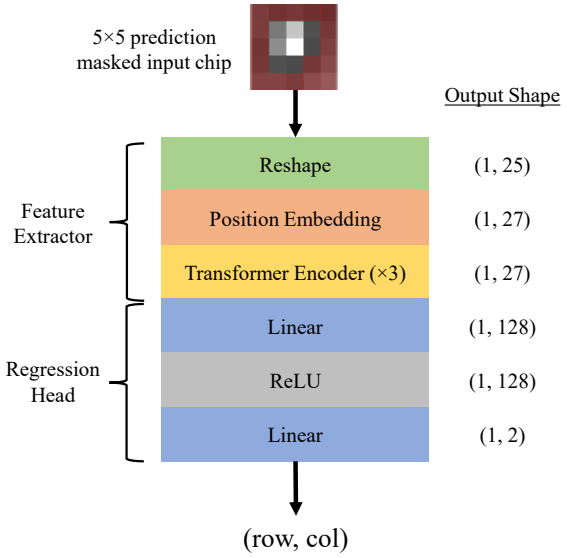


Figure 3. Proposed transformer-based localization architecture. A transformer-based feature extractor and regression head perform a local regression on a masked 5×5 input chip to predict the subpixel location of the target.

Methods for Comparison

This section describes the three methods for subpixel localization that the proposed method will be compared to. These methods rely on traditional processing and do not leverage learned localization. Localization results for each method will be shown in Section 5.

Moments Centroiding— A traditional signal processing method of computing a target's subpixel location is moments centroiding. A moment-weighted centroid of each non-negative pixel in the input chip is computed to estimate the subpixel location of the target. This method requires that spatial background suppression be performed on the input chip for best results. The background is estimated by computing the mean pixel intensity from a local 35×35 region around the input chip. This spatial background estimation is then subtracted from each of the pixels in the input chip before moments centroiding is performed.

Mask Centroiding—The unweighted prediction mask centroid is commonly used in the evaluation of performance for small-target segmentation networks. This method, referred to as mask centroiding, has the advantage of using only the predicted mask and does not use the raw pixel intensities or background suppression. This is the method used for determining the true detections when calculating P_d because it relies only on the predicted mask.

Mask Moment Centroiding—The final method for comparison combines moments centroiding with the information provided by the predicted target mask. Mask moment centroiding first applies the predicted mask to the background suppressed input chip such that all pixels not predicted as part of the target are given a value of zero. Moments centroiding is then applied to the masked input chip. Like moments centroiding, this method requires spatial background suppression to be performed for best results.

5. Experiments and Results

The following sections outline the experiments performed and provide analysis of results. First, the evaluation of the state-of-the-art small-object segmentation networks is presented. Results and analysis of the transformer-based subpixel localization network introduced in Section 4 are then presented.

Segmentation Network Evaluation

This section provides detail and analysis of the segmentation network evaluation performed using state-of-the-art small-object segmentation networks and the benchmark dataset introduced. The evaluation metrics used are discussed and then results and analysis are presented. A thorough comparative analysis is performed to evaluate which network architectures perform best for the task of point-source target segmentation in OPIR data. Although formulated as a segmentation task, these networks also act as detection networks that can then be paired with a subpixel localization method and any other subsequent processing kernels required for target localization and tracking.

Evaluation Metrics—Three metrics are chosen for the evaluation of target segmentation and detection performance. The first metric selected is the normalized intersection over union ($nIoU$) [21]. This is a metric specifically introduced for evaluation of the small-target segmentation task. It is computed as shown in (2) where TP, FP, and FN denote pixel-level true positives, false positives, and false negatives, respectively. N is the total number of samples in the evaluation dataset. Because the benchmark evaluation dataset being used contains one target per frame, $nIoU$ will represent the average intersection over union for a target in the dataset.

$$nIoU = \frac{1}{N}\sum_{i}^{N}\frac{TP[i]}{TP[i] + FP[i] - FN[i]} \tag{2}$$

The use of $nIoU$ provides insight into the pixel-level performance of the networks, however, for the OPIR target detection task the primary concern are target-level evaluations of performance. Probability of detection (P_d) and per-frame false alarm rate (F_a) are used as target-level metrics for evaluation. P_d and F_a are computed as shown in (3) and (4), respectively.

$$P_d = \frac{true\ detections}{total\ targets} \tag{3}$$

$$F_a = \frac{false\ detections}{total\ frames} \tag{4}$$

A predicted target cluster that has a centroid within d_{thresh} of the true target location is considered a true detection. Other predicted clusters are considered false detections. F_a is sometimes reported as a per-pixel false alarm rate, however, for interpretability and to maintain the target-level nature of the metric, we report F_a as a per-frame false alarm rate. This method of computing F_a means that it is dependant on the size of the input frame, but ensures consistent evaluation at the target-level. Results will be labeled as $P_d - d_{thresh}$ and $F_a - d_{thresh}$ to denote the threshold used for the metric calculation. Thresholds of 1 and 4 are used in this evaluation. Note

that the use of target-level P_d and F_a prevent the direct use of a receiver operating characteristic (ROC) curve because merging of predicted target clusters causes the value of F_a not to be strictly increasing as the probability confidence threshold of the network is decreased. For evaluation we will look specifically at the learned P_d-F_a operating point. In practice, the experiments performed in this section show that the trained networks show little sensitivity to changing probability confidence thresholds.

Experimental Setup—All experiments were performed using PyTorch 1.11 and CUDA 11.3. Open-source versions of the models being evaluated were used when available. Networks were trained for a maximum of 50 epochs or a total of 6.5 million training frame iterations. Validation set performance was used for early stopping to ensure best network selection. Identical train, validation, and test splits were used to train all architectures.

Each network architecture and training procedure was kept as originally proposed where possible. The stride of the last convolutional layer in AGPCNet was modified to maintain an appropriate feature map size with 128 $\times$ 128 input frames. The weight initialization method used in ACM was changed to Xavier as done in [26] to enable model convergence and the $maxpool$ layer was removed to ensure appropriate feature map size. Inference times are collected on an NVIDIA Tesla P100.

Network Performance Evaluation—The network benchmark evaluation results are shown in Table 2. The base network is listed as well as the network backbone used. Backbone feature extraction networks used include CAN [37], ResNet [38], VGG-16 [39], and MobileNetV2 [40]. Each network is paired with a backbone it was initially demonstrated with. The addition of MobileNetV2 as a backbone for IAANet is also included. The region proposal network used in IAANet allows the backbone to be easily changed without impacting the rest of the architecture. MobileNetV2 was added because of its common use in edge applications and to increase the variety of backbones evaluated. Overall, results in Table 2 show that AGPCNet variants achieve the best $nIoU$ and P_d-1 results, however, at the expense of a high F_a-1 and high complexity as indicated by the inference times. To determine the overall best performing network, we must consider the network that best balances the often-competing metrics of P_d, F_a, network size, and inference time. Qualitative examples of detection results for each network can be seen in Figure 6.

In addition to the overall probability of detection achieved, the relationship between P_d and the $pSCNR$ of the target is shown in Figure 4. The value of P_d-1 achieved shows a strong dependence on the target $pSCNR$ with the characteristics being similar for all networks that converged. The average Spearman correlation between P_d-1 and $pSCNR$ is 0.834 across all network results. P_d-1 is approximately 1 for $pSCNR > 10$ regardless of the network used but begins to decline rapidly towards zero when $pSCNR < 5$. These results validate the use of $pSCNR$ to quantify target difficulty. AGPCNet can be seen to have a slight advantage over other networks when $pSCNR$ is between 3 and 5, however the overall performance of all networks is similar. P_d is, therefore, insufficient to determine the best network.

Figure 5 shows the calculated P_d and F_a results for each network with two different values of d_{thresh}. This figure provides insight into the trade-off between P_d and F_a. Networks with an operating point located towards the top left of the plot

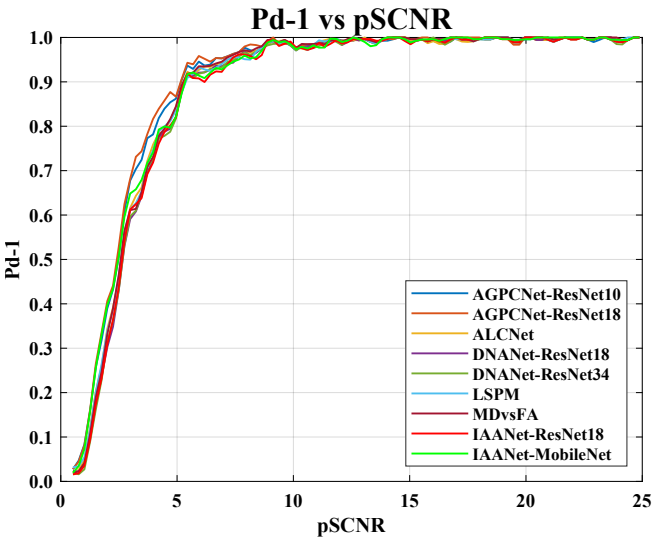


Figure 4. P_d-1 results vs $pSCNR$ on the benchmark evaluation dataset for each of the architectures tested.

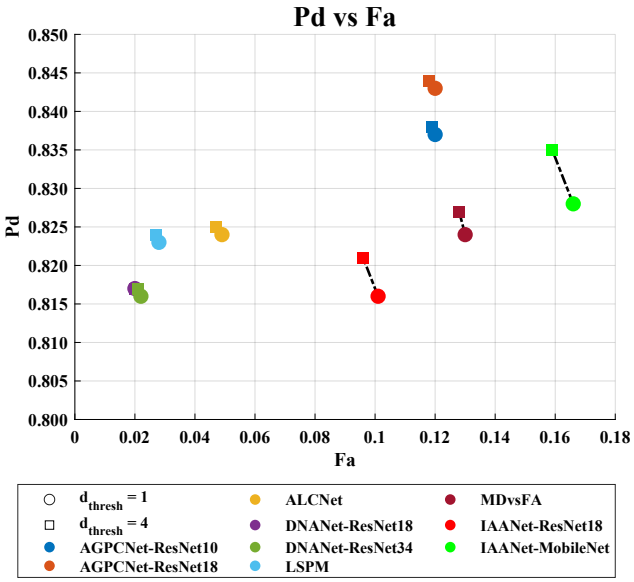


Figure 5. P_d vs F_a plot showing model benchmark performance and shift in performance when applying d_{thresh} values of 1 and 4.

are preferred as they exhibit a high probability of detection and low false alarm rate. The nature of the OPIR target detection task means that a low F_a is important. From Figure 5 it is seen that LSPM and ALCNet both achieve a good trade-off between F_a and P_d. While AGPCNet-ResNet18 achieves the highest overall P_d, it has a significantly higher F_a. Switching networks from LSPM to AGPCNet-ResNet18 would result in a 2.43% increase in P_d-1, but at the cost of a 328.57% increase in F_a-1. This increase in false alarms is unacceptable given the small increase in probability of detection.

Figure 5 also provides insight into the coarse localization ability of the network. The subpixel location of the target for the evaluation of P_d and F_a is found using the centroid of predicted target clusters. The shift in performance when the value of d_{thresh} is changed reflects the accuracy of coarse

Table 2. Benchmark evaluation results for each network architecture considered. Best results are indicated in bold.

Network	Backbone	Parameters (M)	Inference Time $(ms)^{\dagger}$	$nIoU$	$P_d - 1$	$F_a - 1$
MDvsFA [19]	Custom CAN*	3.77	21.23 ± 0.80	0.705	0.824	0.130
ACM-UNet [21]	ResNet-24	1.59	7.82 ± 0.50	–	–	–
ALCNet [24]	ResNet-24	**0.37**	8.63 ± 0.29	0.769	0.824	0.049
LSPM [25]	VGG-16	31.14	$\mathbf{7.77 \pm 0.27}$	0.769	0.823	0.028
DNANet [26]	ResNet-18	4.70	30.37 ± 0.40	0.771	0.817	**0.020**
DNANet[26]	ResNet-34	8.79	50.85 ± 0.98	0.770	0.816	0.022
AGPCNet [27]	ResNet-10	6.08	83.56 ± 1.34	0.785	0.837	0.120
AGPCNet [27]	ResNet-18	12.35	87.47 ± 2.18	**0.791**	**0.843**	0.120
IAANet [28]	ResNet-18	14.05	13.95 ± 2.96	0.760	0.816	0.101
IAANet [28]	MobileNetV2	13.50	19.57 ± 2.48	0.762	0.828	0.166

† Inference time reported as $mean \pm std$ over 5000 inferences on NVIDIA Tesla P100 using PyTorch 1.11 and CUDA 11.3.
* MDvsFA uses a custom context aggregation network (CAN) from [37].
– ACM failed to converge on the selected dataset. Convergence was achieved on easier datasets with higher $pSCNR$.

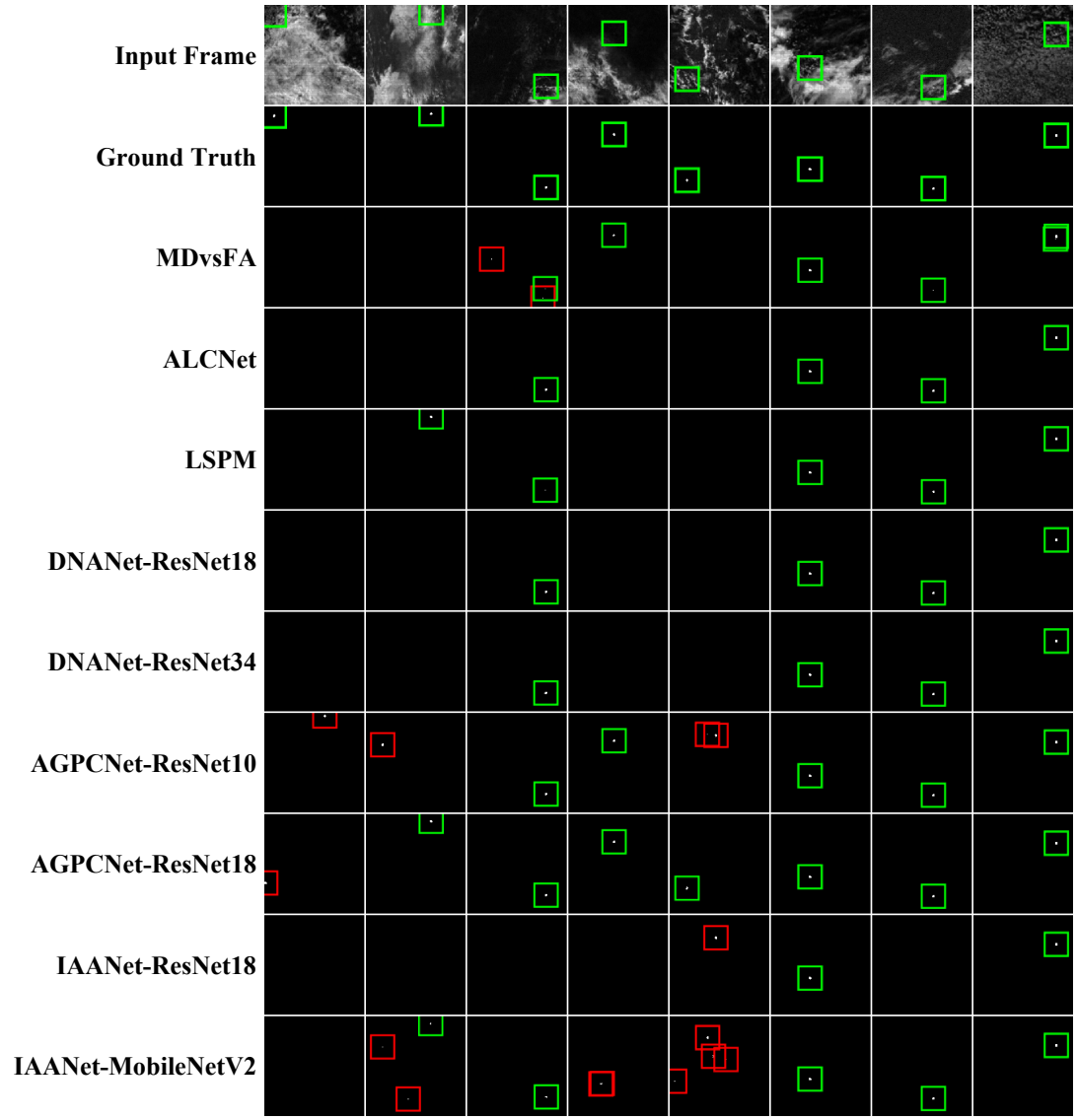


Figure 6. Collection of example results including true detections, false alarms, and missed detections. Low target $pSCNR$ range is selected to show interesting cases. $pSCNR$ values from left to right are 1, 2, 3, 3.5, 4, 4.5, 5, 8.

localization. A small shift is desired as that indicates that when a target is detected, it is done so with small localization error. Most networks tested provide very good coarse localization with almost no shift. IAANet and MDvsFA are two noticeable exceptions. Poor localization in IAANet may be caused by insufficiently detailed target representations being generated by the shallow semantic generator network.

Complexity Comparisons— The nature of missile early-warning systems is such that on-orbit inference may be required. This would necessitate the selection of an architecture that can perform in real-time as OPIR data is acquired. With this potential requirement in mind, the complexity of the network architecture must be considered. Figure 7 and Figure 8 visualize the model size in parameters and the inference latency, respectively. From these plots, it is clear that the ALCNet architecture is the best network when deployment on a resource constrained platform is required. It has the fewest parameters by an order of magnitude and exhibits the second fastest inference time. LSPM is by far the largest network architecture evaluated with 31.14 million parameters compared to the 0.37 million parameters used in ALCNet. However, LSPM should be used in cases where the detection network is not being deployed to a resource constrained device, as it can achieve a lower average number of false alarms.

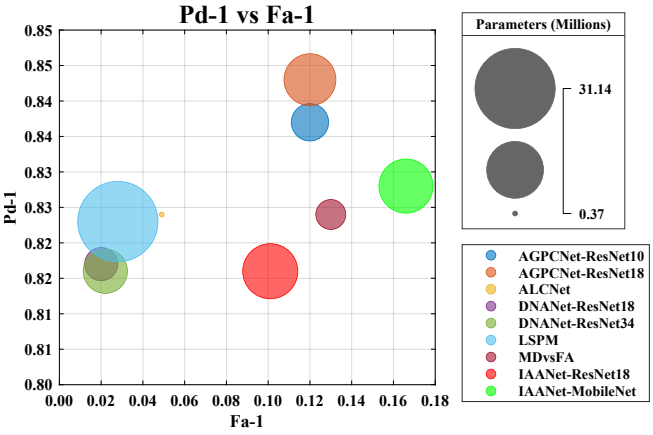


Figure 7. $P_d - 1$ vs $F_a - 1$ bubble plot with model parameters.

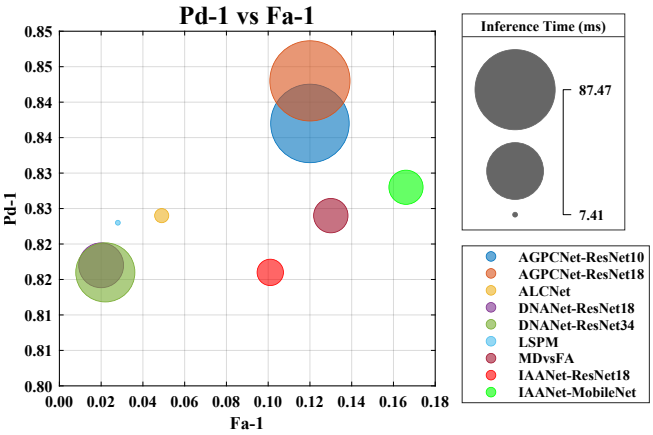


Figure 8. $P_d - 1$ vs $F_a - 1$ bubble plot with mean inference time on NVIDIA Tesla P100.

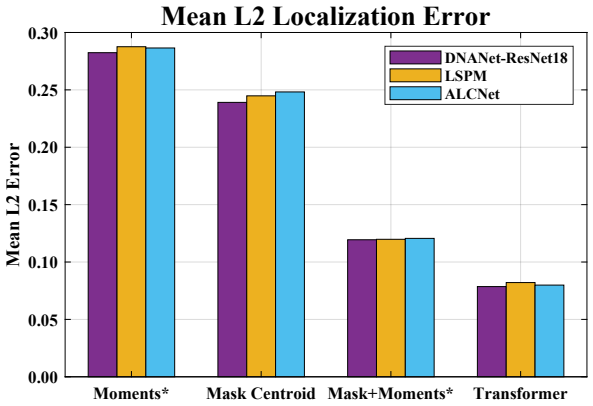


Figure 9. Mean localization error for each of the methods tested using DNANet-ResNet18, LSPM, and ALCNet as detectors. * indicates spatial background suppression is required.

Subpixel Localization Results

The proposed transformer-based localization method was trained and evaluated using ALCNet, LSPM, and DNANet-ResNet18 as the front-end detection and segmentation networks. Different front-end segmentation networks result in different prediction masks and some differences in what targets are and are not detected. Only targets from the benchmark dataset that are detected within a 5×5 region of the ground truth are used to train and test the localization network. The network was trained for 25 epochs using the Adam optimizer and a learning rate of 1×10^{-4}. Early stopping is used to ensure the best network is used for evaluation.

The mean L2 localization error is shown in Figure 9. Results show little dependence on the segmentation network used. This limited variance indicates that each of the three segmentation networks used are able to accurately learn the target shape information such that the localization network can learn a robust target representation and regression model. Using the learned prediction mask with moments centroiding improves results over the baseline moments centroiding method by 58%. This demonstrates that the use of a learned segmentation network to produce the prediction masks can significantly improve subpixel localization. Using the proposed transformer-based localization method provides even further improvement, resulting in a 72% total reduction in mean error. This improvement is achieved while also eliminating the need for background suppression. Figure 10 shows the distribution of localization error when ALCNet is used as the segmentation network for each of the methods tested. The distributions clearly show the superior performance of the transformer-based method, as it has the most heavily skewed distribution and the lowest mean error. Although background suppression and moments centroiding can sometimes perform well, it is susceptible to scenarios where background suppression fails or the initial coarse localization performed by the detection network is poor.

The transformer-based localization network proposed has only 26,231 total parameters and a mean inference time of 2.11 ms. This inference latency would be incurred for every predicted target cluster, however, multiple subpixel localizations could be performed in parallel GPU batches or otherwise parallelized depending on the platform used for deployment.

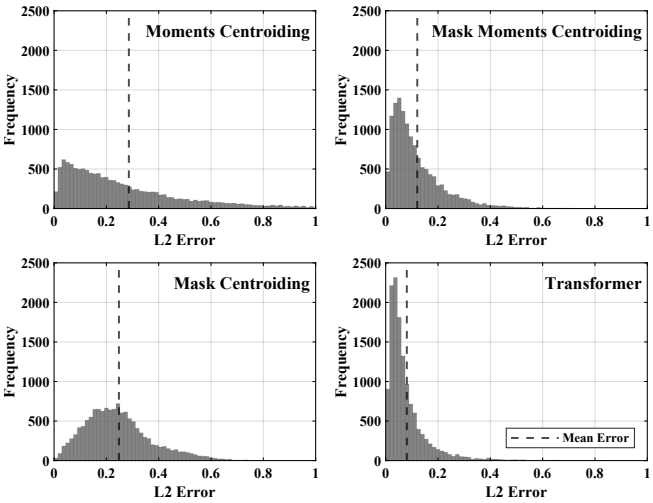


Figure 10. Localization error distributions for the four methods evaluated. ALCNet was used as the detection network for these results.

Upsampling Evaluation— Upsampling of input images has been used in micro-spectroscopy to improve subpixel localization [34]. We perform a comparison of the proposed architecture and two modified architectures that upsample before generating features using the transformer encoder. Both learned (transposed convolution) and unlearned (bilinear interpolation) methods for upsampling are used. The results are shown in Table 3. Learned upsampling is shown to achieve a mean error of 0.0017 pixels less than the error achieved by the proposed method without upsampling. Unlearned bilinear interpolation decreases localization accuracy. The small error reduction from learned upsampling is negligible compared to the 3.99× increase in the number of parameters over the proposed method. Even without upsampling, the transformer-based network is able to perform an accurate local regression on the image patch to find an accurate prediction for the true subpixel target location.

Table 3. Evaluation of different upsampling methods for subpixel localization. ALCNet is used as the detection network for training and evaluation.

Upsampling Method	Parameters	Mean Error
none (proposed)	26,231	0.0807
transposed convolution	104,611	0.0790
bilinear interpolation	104,606	0.0815

6. CONCLUSION

The task of generic single-frame, small-object segmentation in infrared data has been widely explored. Prior research was limited in the evaluation for this task due to the use of datasets with resolved targets and backgrounds inconsistent with OPIR imagery. Additionally, existing datasets relied on hand labeling, making them prone to substantial errors in ground truth. In this research, we have demonstrated that these existing state-of-the-art networks can be applied to the specific task of point-source target segmentation and detection in OPIR imagery. We generated an OPIR-specific dataset with randomized background and target characteristics. Simulated targets with a $pSCNR$ between zero and twenty-five were used for the benchmark evaluation dataset. Training and evaluation of state-of-the-art, small-object segmentation architectures using this dataset enabled detection of point-source targets embedded in complex backgrounds. Nine of ten evaluated networks converged with all results showing $P_d - 1 > 0.80$ and $F_a - 1 < 0.17$. Results show that the probability of detection and target $pSCNR$ have a Spearman coefficient of 0.843, indicating a strong relationship. This strong correlation validates the use of $pSCNR$ as a metric for target detection difficulty. ALCNet and LSPM achieve the most favorable trade-off between the probability of detection and the number of false alarms. Additional analysis in terms of network complexity showed that ALCNet is the best network for implementation on memory constrained embedded devices, while LSPM is the best option otherwise. LSPM inference with 128×128 input frames using a NVIDIA P100 can achieve a frame rate of 128.7 FPS, demonstrating the capability for real-time inference using datacenter-grade GPUs. For larger input sizes, the frame can be tiled, and batch inference can be performed on the 128×128 tiles. Although real-time performance is achieved on a high-performance platform, achieving real-time performance on embedded platforms will require further exploration in future work. The significantly reduced network size of ALCNet shows promise for future deployment to embedded platforms, as its parameters can more easily fit onto edge devices with limited resources.

Unlike applications where the targets are resolved or the imaging distance is short, the OPIR target detection task necessitates accurate subpixel target localization. This processing step is often overlooked or implemented by simply centroiding the predicted target mask. In this research, we have introduced a learned transformer-based method for subpixel localization. This method formulates the task as a local regression on the 5×5 target chip of interest. By leveraging both the predicted target mask and the pixel intensity values of the target chip, the proposed localization network is able to achieve a 72% reduction in the localization error compared to a standard signal processing method of localization. An evaluation is performed to explore the use of upsampling prior to feature extraction from the target chip. The use of upsampling is shown to provide negligible improvement in mean localization error while increasing the size of the network by nearly 4×. The result of this research is a proposed subpixel localization network that has only 26,231 parameters. Inclusion of this network into the overall detection pipeline allows accurate subpixel localization to be performed by leveraging information from both the predicted segmentation mask and the raw OPIR data. The addition of this network to the processing pipeline results in a small latency overhead, which is easily justified by the substantial decrease in localization error achieved.

7. FUTURE WORK

The growing importance of missile warning and target detection using OPIR sensors necessitates continued research in this area. Future work will explore the deployment of detection architectures such as ALCNet and the transformer-based localization network on embedded platforms. Development of a real-time embedded implementation of the detection and localization pipeline proposed will enable future on-board processing of OPIR data with ML algorithms. With this goal in mind, future work will explore the quantization and acceleration of the networks presented. As discussed, ALCNet appears particularly well suited for acceleration on an embedded platform due to its small memory footprint and

low overall complexity. To successfully accelerate ALCNet, the traditional ML kernels must be accelerated, as well as the modified MPCM kernel used in the architecture.

ACKNOWLEDGMENTS

This research was supported by SHREC industry and agency members and by the IUCRC Program of the National Science Foundation under Grant No. CNS-1738783.

This research was also supported by MITRE in collaboration with SHREC. MITRE's participation as an industry member of SHREC is sponsored by the the Tools, Applications, and Processing (TAP) Lab.

Approved for Public Release; Distribution Unlimited 22-3088.

REFERENCES

[1] M. Zhao, W. Li, L. Li, J. Hu, P. Ma, and R. Tao, "Single-frame infrared small-target detection: A survey," *IEEE Geoscience and Remote Sensing Magazine*, 2022.

[2] S. R. Young, B. J. Steward, and K. C. Gross, "Development and validation of the afit scene and sensor emulator for testing (asset)," in *Infrared Imaging Systems: Design, Analysis, Modeling, and Testing XXVIII*, vol. 10178. SPIE, 2017, pp. 56–73.

[3] AFIT Sensor and Scene Emulation Tool (ASSET) [Computer Software]. Air Force Institute of Technology. [Online]. Available: www.afit.edu/CTISR/ASSET

[4] S. D. Deshpande, M. H. Er, R. Venkateswarlu, and P. Chan, "Max-mean and max-median filters for detection of small targets," in *Signal and Data Processing of Small Targets 1999*, vol. 3809. SPIE, 1999, pp. 74–83.

[5] M. Zeng, J. Li, and Z. Peng, "The design of top-hat morphological filter and application to infrared target detection," *Infrared physics & technology*, vol. 48, no. 1, pp. 67–76, 2006.

[6] P. Wang, J. Tian, and C. Q. Gao, "Infrared small target detection using directional highpass filters based on ls-svm," *Electronics letters*, vol. 45, no. 3, pp. 156–158, 2009.

[7] S. Kim, "Min-local-log filter for detecting small targets in cluttered background," *Electronics letters*, vol. 47, no. 2, p. 1, 2011.

[8] C. Gao, D. Meng, Y. Yang, Y. Wang, X. Zhou, and A. G. Hauptmann, "Infrared patch-image model for small target detection in a single image," *IEEE transactions on image processing*, vol. 22, no. 12, pp. 4996–5009, 2013.

[9] C. P. Chen, H. Li, Y. Wei, T. Xia, and Y. Y. Tang, "A local contrast method for small infrared target detection," *IEEE transactions on geoscience and remote sensing*, vol. 52, no. 1, pp. 574–581, 2013.

[10] Y. Wei, X. You, and H. Li, "Multiscale patch-based contrast measure for small infrared target detection," *Pattern Recognition*, vol. 58, pp. 216–226, 2016.

[11] H. Deng, X. Sun, M. Liu, C. Ye, and X. Zhou, "Small infrared target detection based on weighted local difference measure," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 7, pp. 4204–4214, 2016.

[12] Y. Shi, Y. Wei, H. Yao, D. Pan, and G. Xiao, "High-boost-based multiscale local contrast measure for infrared small target detection," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 1, pp. 33–37, 2017.

[13] H. Deng, X. Sun, and X. Zhou, "A multiscale fuzzy metric for detecting small infrared targets against chaotic cloudy/sea-sky backgrounds," *IEEE transactions on cybernetics*, vol. 49, no. 5, pp. 1694–1707, 2018.

[14] P. Du and A. Hamdulla, "Infrared small target detection using homogeneity-weighted local contrast measure," *IEEE Geoscience and Remote Sensing Letters*, vol. 17, no. 3, pp. 514–518, 2019.

[15] C. Yang, J. Ma, S. Qi, J. Tian, S. Zheng, and X. Tian, "Directional support value of gaussian transformation for infrared small target detection," *Applied optics*, vol. 54, no. 9, pp. 2255–2265, 2015.

[16] Y. Dai, Y. Wu, and Y. Song, "Infrared small target and background separation via column-wise weighted robust principal component analysis," *Infrared Physics & Technology*, vol. 77, pp. 421–430, 2016.

[17] H. Deng, X. Sun, M. Liu, C. Ye, and X. Zhou, "Entropy-based window selection for detecting dim and small infrared targets," *Pattern Recognition*, vol. 61, pp. 66–77, 2017.

[18] Y. U. Sinn, K. M. Hopkinson, B. J. Borghetti, and B. J. Steward, "Ir small target detection and prediction with anns trained using asset," in *2019 IEEE Aerospace Conference*. IEEE, 2019, pp. 1–11.

[19] H. Wang, L. Zhou, and L. Wang, "Miss detection vs. false alarm: Adversarial learning for small object segmentation in infrared images," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 8509–8518.

[20] B. Zhao, C. Wang, Q. Fu, and Z. Han, "A novel pattern for infrared small target detection with generative adversarial network," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 5, pp. 4481–4492, 2020.

[21] Y. Dai, Y. Wu, F. Zhou, and K. Barnard, "Asymmetric contextual modulation for infrared small target detection," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 950–959.

[22] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[23] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.

[24] Y. Dai, Y. Wu, F. Zhou, and K. Barnard, "Attentional local contrast networks for infrared small target detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 11, pp. 9813–9824, 2021.

[25] L. Huang, S. Dai, T. Huang, X. Huang, and H. Wang, "Infrared small target segmentation with multiscale feature representation," *Infrared Physics & Technology*, vol. 116, p. 103755, 2021.

[26] B. Li, C. Xiao, L. Wang, Y. Wang, Z. Lin, M. Li, W. An, and Y. Guo, "Dense nested attention network for infrared small target detection," *arXiv preprint arXiv:2106.00487*, 2021.

[27] T. Zhang, S. Cao, T. Pu, and Z. Peng, "Agpcnet: Attention-guided pyramid context networks for infrared small target detection," *arXiv preprint arXiv:2111.03580*, 2021.

[28] K. Wang, S. Du, C. Liu, and Z. Cao, "Interior attention-aware network for infrared small target detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2022.

[29] F. Liu, C. Gao, F. Chen, D. Meng, W. Zuo, and X. Gao, "Infrared small-dim target detection with transformer under complex backgrounds," *arXiv preprint arXiv:2109.14379*, 2021.

[30] H. Zhou, C. Tian, Z. Zhang, C. Li, Y. Xie, and Z. Li, "Pixelgame: Infrared small target segmentation as a nash equilibrium," *arXiv preprint arXiv:2205.13124*, 2022.

[31] X. Tong, B. Sun, J. Wei, Z. Zuo, and S. Su, "Eaau-net: Enhanced asymmetric attention u-net for infrared small target detection," *Remote Sensing*, vol. 13, no. 16, p. 3200, 2021.

[32] Q. Hou, L. Zhang, F. Tan, Y. Xi, H. Zheng, and N. Li, "Istdu-net: Infrared small-target detection u-net," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[33] M. Zhang, R. Zhang, Y. Yang, H. Bai, J. Zhang, and J. Guo, "Isnet: Shape matters for infrared small target detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 877–886.

[34] J. Schroeter, T. Tuytelaars, K. Sidorov, and D. Marshall, "Learning multi-instance sub-pixel point localization," in *Proceedings of the Asian Conference on Computer Vision*, 2020.

[35] J. S. Kent, C. C. Wamsley, D. Flateau, and A. Ferguson, "Unsupervised learning for target tracking and background subtraction in satellite imagery," in *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications III*, vol. 11746. SPIE, 2021, pp. 83–92.

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

[37] Q. Chen, J. Xu, and V. Koltun, "Fast image processing with fully-convolutional networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2497–2506.

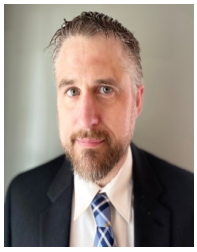
[38] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[39] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[40] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.

BIOGRAPHY

Daniel C. Stumpp *received the M.S. degree in electrical and computer engineering in 2022 from the University of Pittsburgh where he is currently pursuing the Ph.D. degree in electrical and computer engineering. He is a member of the NSF Center for Space, High-Performance, and Resilient Computing (SHREC), where he performs research under the direction of Dr. Alan George. His research interests include high-performance FPGA-based accelerator architectures for embedded platforms, remote sensing applications of machine learning, and novel processing of asynchronous neuromorphic sensor data.*



Andrew J. Byrne *is a principal sensor engineer and group leader for MITRE Labs, a division of MITRE, where he is a member of the Sensors, Electromagnetics, and Electronic Warfare Department. He has 25 years of experience in the design, development, and optimization of sensor systems spanning a wide range of sensing modalities including passive optical sensors in spectral regions from the ultraviolet to the infrared, multi- and hyperspectral sensors, passive radio frequency sensors, monostatic and bistatic microwave radar systems, and high-frequency, over-the-horizon radar systems. His research interests include novel sensing modalities, sensor fusion of disparate sensor data, multiband and ultra-wideband synthetic aperture radar (SAR), and novel algorithms for target detection, clutter mitigation, and colored noise reduction.*

Alan D. George *(Fellow, IEEE) is currently the Department Chair, the Robert Horonjeff Mickle Endowed Chair, and a Professor in electrical and computer engineering (ECE) with the University of Pittsburgh. He is also the Founder and the Director of the NSF Center for Space, High-Performance, and Resilient Computing (SHREC) headquartered at Pittsburgh. SHREC is an Industry/the University Cooperative Research Center (I/UCRC) featuring some 30 academic, industry, and government partners and is considered by many as the leading research center in its field. His research interests include high-performance architectures, applications, networks, services, systems, missions for reconfigurable, parallel, distributed, and dependable computing, from spacecraft to supercomputers. He is a fellow of the IEEE for contributions in reconfigurable and high-performance computing.*