SINBAD ORIGINS OF CONTEXTUALLY-GUIDED FEATURE LEARNING: SELF-SUPERVISION WITH LOCAL CONTEXT FOR TARGET DETECTION

Olcay Kursun
Department of Computer Science
Auburn University at Montgomery
Montgomery, AL 36117, USA
okursun@aum.edu

Oleg V. Favorov
Joint Department of Biomedical Engineering
University of North Carolina at Chapel Hill
Chapel Hill, NC 27599, USA
favorov@email.unc.edu

Abstract— **The domain of feature learning is replete with self-supervised methodologies that harness local context to decode complex patterns in data. This paper investigates the significance of the SINBAD model within the realm of self-supervised learning. SINBAD, a unique non-convolutional neural network approach, is particularly adept at extracting mutual information and identifying distinct local structures. The underlying theory of SINBAD, anchored in neuroscience, suggests that capturing predictable connections among subsets of features can significantly enhance feature extraction. This principle has been instrumental in the advancement of self-supervised learning as it has given rise to the development of its convolutional extensions that are self-guided by contextual understanding. In this paper, we explore the application of the SINBAD algorithm for automatic target recognition, specifically in detecting military vehicles within complex rural environments. Our methodology includes a preprocessing stage utilizing a linear Support Vector Machine (SVM) to learn and apply a filter. This filter is crucial for excluding image locations that are completely irrelevant, thereby isolating potentially relevant regions for contextual self-supervision. By focusing on these selected areas and avoiding the need to process arbitrary regions, we aim to minimize resource requirements and streamline the self-supervised learning process. In the subsequent phase of our approach, the focus shifts to identifying nonlinear features that provide insights into local contexts. This involves extracting correlated nonlinear functions from nearby, yet distinct and non-overlapping patches. By analyzing these functions, we can discern patterns and relationships within the local context, where each function correlates with its unique, adjacent patch. This sophisticated analysis allows for a deeper understanding of the local area, contributing significantly to the accuracy of the final classification step. Finally, the process culminates with the use of a linear classifier that utilize the SINBAD features to effectively identify and categorize image patches that contain the intended targets. The research was conducted using the TNO-TM Search_2 dataset, which comprises 44 high-resolution images. These images depict cluttered rural landscapes and feature 9 different types of military vehicles. The dataset's complexity and variety provide a comprehensive environment for testing and validating the effectiveness of the SINBAD model in accurately identifying and classifying these vehicles within such challenging and diverse scenes.**

Index Terms— **Self-Supervised Learning, Automatic Target Recognition, Nonlinear Feature Extraction, Contextually-Guided Neural Networks.**

I. INTRODUCTION

This paper delves into the domain of self-supervised feature learning methodologies designed to decode complex patterns in data through the utilization of local context [1, 2, 3, 4]. We focus our investigation on the SINBAD model [5, 6], a meta-learning approach typically implemented as a non-convolutional neural network for extracting mutual information [2] and identifying distinct local structures [3]. This approach has paved the way for convolutional extensions for learning transferrable features [7]. Contrasting with deep CNNs, the SINBAD model draws inspiration from biological neural networks, particularly cortical areas that optimize feature extraction using local contextual information rather than relying on error backpropagation. This approach, grounded in theoretical neuroscience, highlights the importance of spatial and temporal contexts in feature selection. Such contextually selected features are behaviorally beneficial as they capture predictable relationships with other distinct sensory inputs, reflecting structured causal dependencies in the external world.

In this paper, we apply the SINBAD algorithm to the challenge of automatic target recognition, with a specific focus on detecting military vehicles in complex rural environments [8]. Our methodology initiates with a preprocessing phase that employs a linear method (we used linear-SVM [9, 10]) to discard highly irrelevant image locations, and keeping regions that are potentially target-related and suitable for contextual self-supervision. This selective process aims to minimize resource consumption and optimize the self-supervised learning workflow. Progressing to the next phase, the SINBAD neural network is applied to the discover nonlinear features that are predictive of the local contexts. This is achieved by extracting correlated nonlinear functions from nearby, distinct, and non-overlapping patches. Discovering these features in a self-supervised manner improves the generalization capabilities of the subsequent classification phase and leads to higher accuracy. For this final phase, we deploy a linear classifier that utilizes the derived SINBAD features to efficiently identify and categorize patches containing the targets. For our experimental results, we use the TNO-TM Search_2 dataset, consisting of 44 high-resolution images that depict a variety of military vehicles in cluttered rural landscapes [8]. The diversity and complexity of this dataset provide an ideal testing ground for the SINBAD model, demonstrating its effectiveness in accurately identifying and classifying military vehicles in challenging and varied

979-8-3503-1710-7/24/$31.00 ©2024 IEEE

scenes, thus proving its relevance in self-supervised contextually-guided feature extraction [7].

II. RELATED WORK: CONTEXTUALLY-GUIDED SELF-SUPERVISED FEATURE LEARNING

Feature learning is a fundamental task in machine learning and data analysis, aimed at discovering informative representations from high-dimensional data [9]. Effective feature extraction techniques play a crucial role in improving the performance of various machine learning algorithms, including classification, clustering, and dimensionality reduction. Traditional methods, such as Principal Component Analysis (PCA), are limited in their ability to capture nonlinear relationships and preserve local structure [9]. Kernel PCA (KPCA) [9-11], as a nonlinear extension of PCA, leverages kernel functions to map the data into a higher-dimensional feature space where linear PCA can be applied. This enables KPCA to capture intricate nonlinear patterns in the data. However, KPCA does not explicitly take into account the local neighborhood information, which is vital for preserving the local structure and forming a meaningful lower-dimensional representation. Addressing this limitation, manifold learning methods [9], such as Laplacian eigenmaps [12] and t-SNE [13], have gained popularity for their ability to capture the underlying geometry of the data or to visualize/reveal the underlying structure of the data distribution. However, these manifold methods may not fully exploit the power of kernel tricks to handle nonlinearities. Moreover, they often lack the capability for out-of-sample extension, limiting their practical applicability [14].

As a promising alternative, deep learning [1, 4, 7, 15, 16, 17, 18, 19, 20, 21], as a subfield of machine learning, has shown its prowess in many challenging tasks like image recognition, speech recognition, and natural language processing. Deep learning models, especially Convolutional Neural Networks (CNNs), have been effective due to their ability to capture high-level abstractions in data by building complex hierarchies of features [15, 16]. CNNs conduct a series of convolution (Conv) operations, each succeeded by nonlinear transformations typified by sigmoidal or ReLU activation functions. These progressive non-linear operations are instrumental in tuning the neurons of the network to increasingly inferential features. With such inferential and transferrable features, the usefulness of deep learning models extends beyond their impressive accuracy on expansive datasets. CNN features also receive interest because the initial layer features they learn to extract show similarities to the ones derived by biological neurons in the primary visual cortex (V1) [17].

While there are notable parallels between cortical areas and deep CNNs, key distinctions exist. One major difference is that, unlike deep CNNs which rely heavily on error back-propagation from their upper layers to refine initial features, cortical regions utilize self-supervision. This self-supervision is rooted in local contextual information, which is pivotal in optimizing feature tuning. Although this information is locally sourced, it plays a crucial role in guiding feature selection. The widely accepted view in theoretical neuroscience posits that this guidance is derived from the spatial and temporal contexts of the features [2, 3, 7, 22-26]. The behavioral relevance of these contextually-selected features stems from their predictable associations with other distinct features obtained from separate sensory inputs. Such associations reflect the structured causal relationships inherent in the external environment, from which these features originate.

In the CG-CNN framework [7], the concept of context is akin to how a word's meaning in Natural Language Processing depends on its surrounding words. Similarly, in image analysis, the interpretation of a pixel can depend on nearby areas, forming a context. To capture these contextual relationships, CG-CNN introduces contextual groups, each representing a set of training examples with similar patterns. However, simultaneously training on a large number of contextual groups to capture comprehensive contextual regularities in images can be complex. To manage this, CG-CNN employs an iterative training strategy, using an Expectation-Maximization (EM) algorithm, and focusing on a different small subset of contextual groups in each iteration. This approach follows the principles of transfer learning. During each EM iteration, the connections in the Classifier layer are trained in the E-step, holding the Feature Generator's weights constant. Then, in the M-step, the Feature Generator's weights are updated while keeping the newly optimized Classifier connections constant. By iteratively training on different small subsets of contextual groups, CG-CNN creates 'pluripotent' features that can capture various contexts. This training approach provides an efficient method for learning the regularities that define contextual classes by limiting the number of classes in each EM iteration to a manageable level. CG-CNN constructs its multiple views through a self-supervision process, utilizing augmented versions of the input data. By learning a common feature space that maximizes the agreement among these internally generated views, CG-CNN effectively learns the underlying regularities in the data.

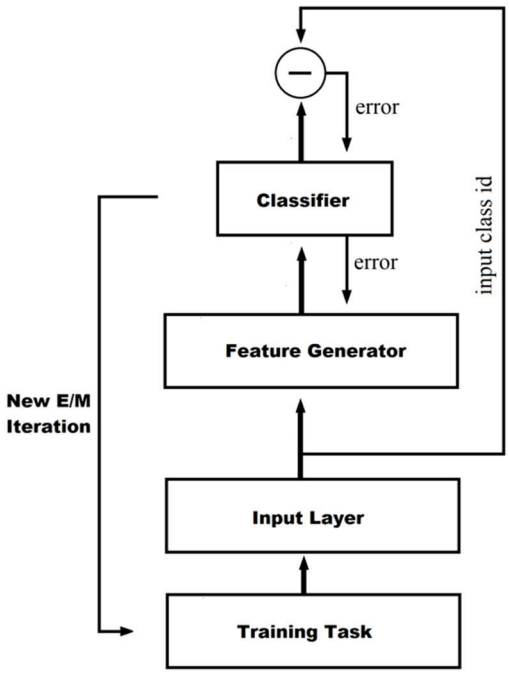

Figure 1. Self-supervised CG-CNN architecture that learns to discriminate auxiliary classes formed by contextual relations.

III. APPLICATION OF THE SINBAD APPROACH FOR AUTOMATIC TARGET RECOGNITION

Our research was conducted using the TNO-TM Search_2 dataset, which includes 44 high-resolution images depicting complex rural landscapes featuring nine distinct military vehicle types [8]. Training was performed on a subset of 24 images, while a separate, randomly selected set of 10 images was reserved for validation of our model's efficacy.

Our methodology employs a layered and iterative strategy akin to the repetitive application of convolution and ReLU layers in CNNs, enhancing performance progressively. This process includes: (1) the initial acquisition of SINBAD features, which mirror the patterns found in our database images, to pinpoint potential target sites; (2) the refinement of SINBAD features, tailored specifically for areas deemed 'suspicious', thereby honing the focus to a more defined set of potential targets; (3) a further iteration to develop an even more advanced set of SINBAD features, applied to this refined set of locations, aiming to further diminish the occurrence of false positives; and this process continues in a similar fashion. Each step in this hierarchical procedure is designed to incrementally improve the accuracy and precision of our target detection, akin to the deepening complexity of a CNN with each additional convolution and activation layer. At each stage, SINBAD features can be developed specifically for those image locations that were considered suspicious by the preceding stages of the analysis. Such specialized SINBAD features will exhibit progressively greater discriminative sensitivity to image details specific to the ‘suspected’ (i.e., containing a vehicle or not yet ruled out) image locations.

The SINBAD network for feature extraction is a constellation of SINBAD cells, each embodying an algorithm that capitalizes on mutual information across varied yet correlated input sets [5, 6, 27, 28]. In our objective of target recognition within natural terrain imagery, SINBAD's self-supervised approach plays a pivotal role. SINBAD's unsupervised learning mechanism identifies various inherent local patterns/redundancies with the assumption that by extracting local dependencies within the imagery, these features can be effectively utilized for distinguishing between different classes, such as the presence or absence of vehicles.

In our study, the determination of whether a specific image segment harbors a target, specifically a vehicle, hinges on the use of a Support Vector Machine (SVM) [9, 10]. Their robustness against overfitting, exceptional generalization capabilities, and rapid convergence mark them as ideal for our purposes. The pivotal aspects of our methodology are depicted in Figure 2. The first SVM (denoted as SVM1) is calibrated to recognize military vehicles within the confines of its limited observational field across various training images. Inevitably, SVM1 is not infallible in its task, it occasionally misidentifies natural terrain as containing a vehicle. The primary function of SVM1 is to conduct a preliminary survey of an entire high-resolution image, pinpointing locations with positive identifications for subsequent scrutiny. This technique enables us to swiftly eliminate 99.65% of the non-relevant image segments. Despite this efficient filtration, we are still confronted with numerous potential target sites. In the next phase, we focus on enhancing SINBAD features exclusively at those junctures flagged as 'questionable' by SVM1. In this instance, a composite of 14 SINBAD cells, termed 'SINBAD Network 1' as seen in Figure 2, was deployed.

Each SINBAD cell is tasked with identifying a unique attribute by learning correlated functions across its dendrites within adjacent, non-overlapping 5x5 pixel fields. This method compels the dendrites to encode the context of their specific 5x5 pixel field in relation to neighboring areas. We designed 14 such SINBAD cells, each contributing to a 14-dimensional 'feature' vector that represents the input field. This vector captures the core characteristics of the image window. These SINBAD-generated features then serve as the input for the secondary SVM (identified as SVM2) in Figure 2. SVM2 is trained to recognize the presence of a vehicle within a 20x20 pixel window. During training, the window is placed only at those image locations that were marked as ‘suspicious’ by SVM1.

IV. EXPERIMENTAL RESULTS

SVM1 misidentifies natural terrain as containing a vehicle, with a 0.35% error rate in our trials. SVM2 greatly reduces the number of False Positives that were made by a factor of 20 without missing any of the real vehicles in the test images. Thus, a sequence of SVM1-SINBAD-SVM2 in our experiments so far was able to detect all the test vehicles while making False Positive mistakes on only 0.015% of the test trials.

Presented in Figure 3 are the results from the initial dual-phase process of vehicular identification conducted on a representative image from the dataset. This particular image was set aside from the training dataset, not contributing to the training of either SVM1 or SINBAD or SVM2, and was solely utilized for the assessment of their trained algorithms. The lower two panels in the figure highlight the specific segments of the image where the SVMs indicated vehicular presence.

The efficacy of stage 1 is evidenced in the third panel of Figure 3, where SVM1 accurately detects a tank's position. Nonetheless, it concurrently generates 155 False Positive signals for locations void of vehicles. Progressing to stage 2, as depicted in the lowest panel, SVM2 successfully eliminates 153 out of the 155 False Positives. It narrows down the potential vehicle-containing locations to eight, of which six validly correspond to different sections of the same tank, resulting in only two erroneous identifications. Consequently, Figure 3 exemplifies the success of our method in refining SINBAD features for areas under scrutiny and leveraging a subsequent SVM trained on these features, significantly diminishing False Positive counts while maintaining the detection of actual targets.

Figure 4 illustrates the detection outcomes across six distinct test images. In each image, three elements are showcased: (1) a segment of the original high-resolution image, (2) the pinpointed squares indicating potential vehicle locations, and (3) an enhanced view where these areas and their context are accentuated, revealing the landscape features SVM2 inaccurately recognized as vehicles. A closer examination reveals that the majority of these misidentified features—such as portions of tree trunks or branches—bear little resemblance to actual vehicles. This observation supports the potential for subsequent SINBAD-SVM stages to discern and classify these

features accurately as non-vehicular. Notably, as referenced in [8], human observers encountered challenges in locating a vehicle within database image 11 (displayed in the top-left panel of Figure 4), with a significant portion (18 out of 62) unable to detect it. In contrast, SVM2 readily identified the vehicle in this image with minimal False Positives. A similar advantage of SVM2 was observed with database image 2 (shown in the top-right panel of Figure 4), another image where human observers demonstrated a high rate of non-detection (16 out of 62 missed it).

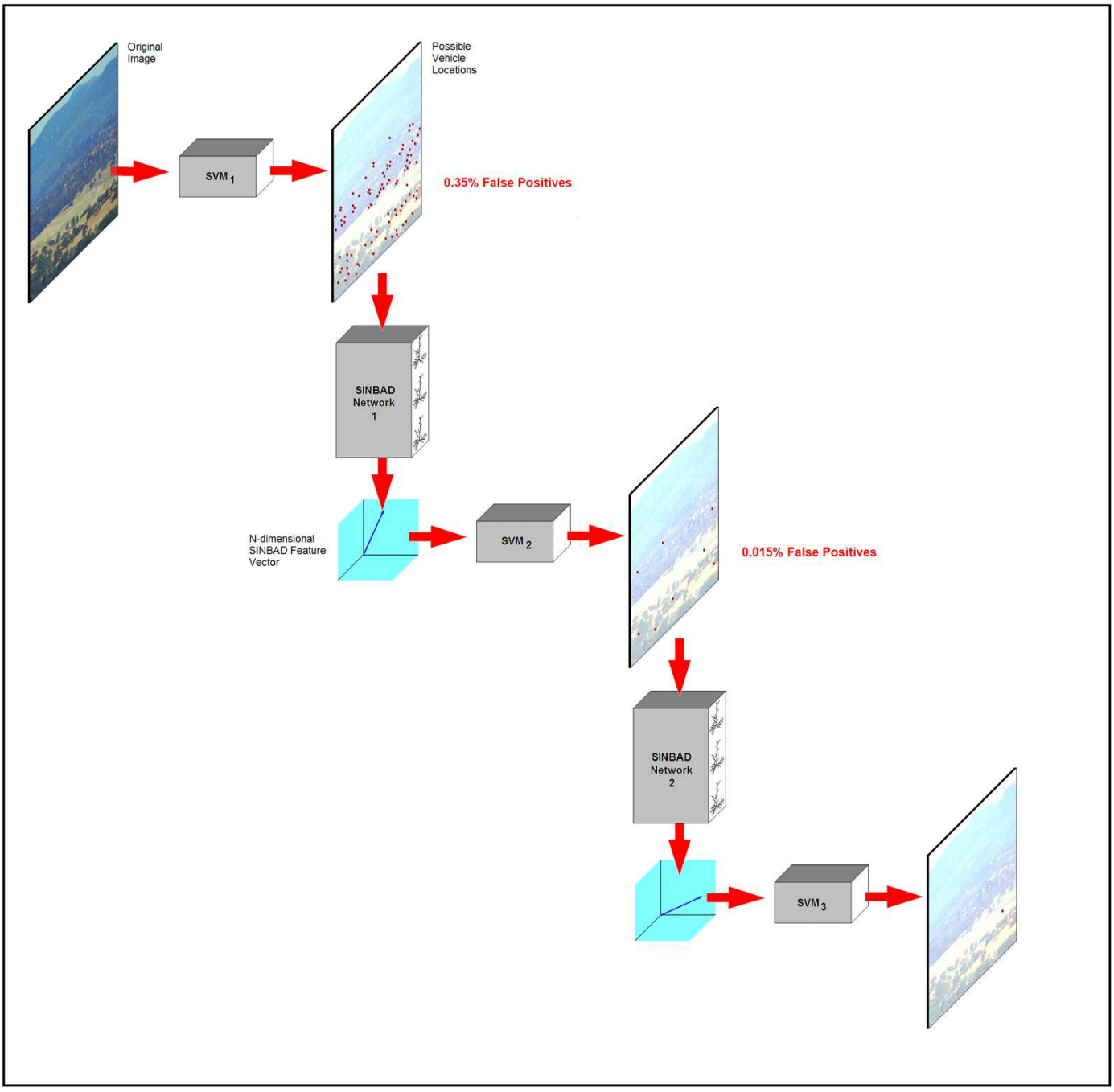

Figure 2. Deep SINBAD network architecture. SVM1-SINBAD-SVM2 portion is demonstrated in this study with favorable target detection results.

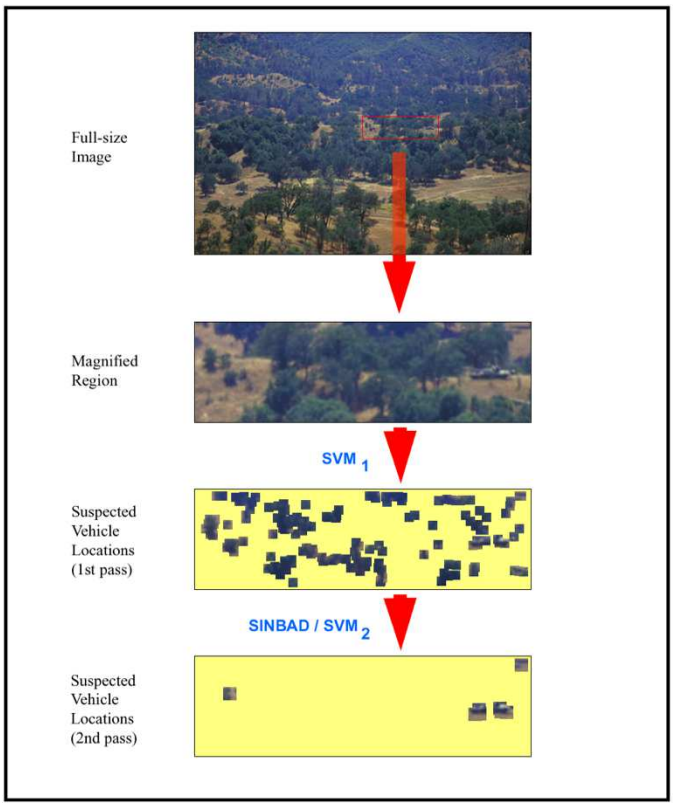

Figure 3. The proposed method eliminates false positives successfully by utilizing the nonlinear SINBAD features.

V. CONCLUSION

In conclusion, our research highlights the effectiveness of local contextual features learned through self-supervision, particularly those derived using the biologically-inspired SINBAD approach, in the context of vehicle detection. Key findings from our experiments include the initial stage with SVM1 accurately identifying a tank location, albeit accompanied by 155 false positives. Remarkably, the SVM2 stage efficiently reduced these to just 2 false positives, while correctly identifying 6 out of 8 potential vehicle locations. This process, centered around the creation and utilization of specialized SINBAD features followed by a refined SVM training, has proven to be highly effective in minimizing false positives without diminishing the ability to detect true targets.

Incorporating advanced techniques like CG-CNN and transfer learning could potentially enhance our system's performance. However, the primary objective of this research was to demonstrate the inherent power of contextual features without resorting to such complex methodologies. Our results clearly indicate that even without these additional layers of complexity, our approach is capable of achieving impressive vehicle detection accuracy.

In addition to exploring these advanced techniques, we are also considering several other strategies to maximize our system's efficiency. This includes optimizing various parameters such as the sizes of viewing windows, SINBAD and SVM parameters, and the number of training samples. Also, image preprocessing techniques like local dynamic range normalization and contrast enhancement could be instrumental in improving the performance, especially given the sub-optimal quality of the original images.

ACKNOWLEDGMENT

This work, was supported, in part, by the National Science Foundation under grant 2003740. We wish to acknowledge the revisiting of insights from SINBAD, as delineated in our previous work, which has been instrumental in establishing a connection between the neuroscientific origins of contextual guidance and the development of CG-CNN.

REFERENCES

[1] Dosovitskiy, A., Springenberg, J. T., Riedmiller, M., & Brox, T. (2014). Discriminative unsupervised feature learning with convolutional neural networks. In Advances in Neural Information Processing Systems, pages 766–774.

[2] Becker, S., Hinton, G. Self-organizing neural network that discovers surfaces in random-dot stereograms. Nature 355, 161–163 (1992).

[3] Phillips, W.A., Kay, J., Smyth, D. (1995) “The discovery of structure by multi-stream networks of local processors with contextual guidance,” Network: Computation in Neural Systems vol. 6, pp. 225-246.

[4] Hjelm, R. D., Fedorov, A., Lavoie-Marchildon, S., Grewal, K., Bachman, P., Trischler, A., Bengio, Y. (2018). "Learning deep representations by mutual information estimation and maximization", ICLR 2019.

[5] Favorov, O.V. and Ryder D. (2004) SINBAD: a neocortical mechanism for discovering environmental variables and regularities hidden in sensory input. *Biological Cybernetics* 90: 191-202.

[6] Ryder, D. and Favorov, O.V. (2001) The new associationism: a neural explanation for the predictive powers of cerebral cortex. *Brain and Mind* 2: 161-194.

[7] Kursun, O., Dinc, S., Favorov, O.V. (2022). Contextually Guided Convolutional Neural Networks for Learning Most Transferable Representations. In Proceedings of the 24th IEEE International Symposium on Multimedia (IEEE-ISM), Naples, Italy, December 2022.

[8] Toet, A., Bijl, P., Kooi, F.L., Valenton, J.M. (1998) A high-resolution image dataset for testing search and detection models. Report TNO-TM-98-A020, TNO Human Factors Research Institute, Soesterberg, The Netherlands.

[9] Alpaydın, E. (2020). Introduction to Machine Learning (4th ed.). MIT Press.

[10] Schölkopf B., Smola A.J. (2002) Learning with Kernels. MIT Press, Cambridge, MA

[11] Schölkopf, B., Smola, A., Müller, K. R. (1998). Nonlinear component analysis as a kernel eigenvalue problem. Neural Computation, 10(5), 1299-1319.

[12] Belkin, M., Niyogi, P. (2003). Laplacian eigenmaps for dimensionality reduction and data representation. Neural Computation, 15(6), 1373-1396.

[13] Van der Maaten, L., Hinton, G. (2008). Visualizing data using t-SNE. Journal of Machine Learning Research, 9(Nov), 2579-2605.

[14] Bengio, Y., Paiement, J. F., Vincent, P., Delalleau, O., Le Roux, N., Ouimet, M. (2004). Out-of-sample extensions for lle, isomap, mds, eigenmaps, and spectral clustering. In Advances in neural information processing systems, pp. 177-184.

[15] Bengio, Y., Goodfellow, I., Courville, A. (2016). Deep Learning. MIT Press.

[16] LeCun, Y., Bengio, Y., Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444.

[17] Yosinski, J., Clune, J., Bengio, Y., Lipson, H. (2014). How transferable are features in deep neural networks?. arXiv preprint arXiv:1411.1792.

[18] Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems; 2012. p. 1097–1105.

[19] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2016. p. 770–778.

[20] Misra, I., van der Maaten, L. (2019). Self-supervised learning of pretext-invariant representations. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4694–4703.

[21] Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., Ng, A. Y. (2011). Multimodal deep learning. In Proceedings of the 28th International Conference on Machine Learning (ICML-11), 689-696.
[22] Becker, S. (1999) Implicit learning in 3d object recognition: the importance of temporal context. Neural Computation 11: 347-374.
[23] Kay, J.W., Phillips, W.A. (2011) "Coherent Infomax as a Computational Goal for Neural Systems," Bull Math Biol, vol. 73, pp. 344–372.
[24] Marblestone, A.H. Wayne, G., Kording, K.P. (2016) "Toward an Integration of Deep Learning and Neuroscience," Front. Comput. Neurosci., vol. 10:94.
[25] Hawkins, J., Ahmad, S., Cui, Y. "A Theory of How Columns in the Neocortex Enable Learning the Structure of the World", Frontiers in Neural Circuits, vol. 11:1-81, 2017.
[26] Clark, A., Thornton, C. (1997) "Trading places: computation, representation, and the limits of uninformed learning," Behavioral and Brain Sciences vol. 20, pp. 57-90.
[27] Kursun, O., and Favorov, O. V. (2004a). SINBAD automation of scientific discovery: from factor analysis to theory synthesis. *Natural Computing* 3: 207-233.
[28] Kursun O. and Favorov O.V. (2004b) What can SVMs teach each other? *Artificial Neural Networks in Engineering* (ANNIE 2004).

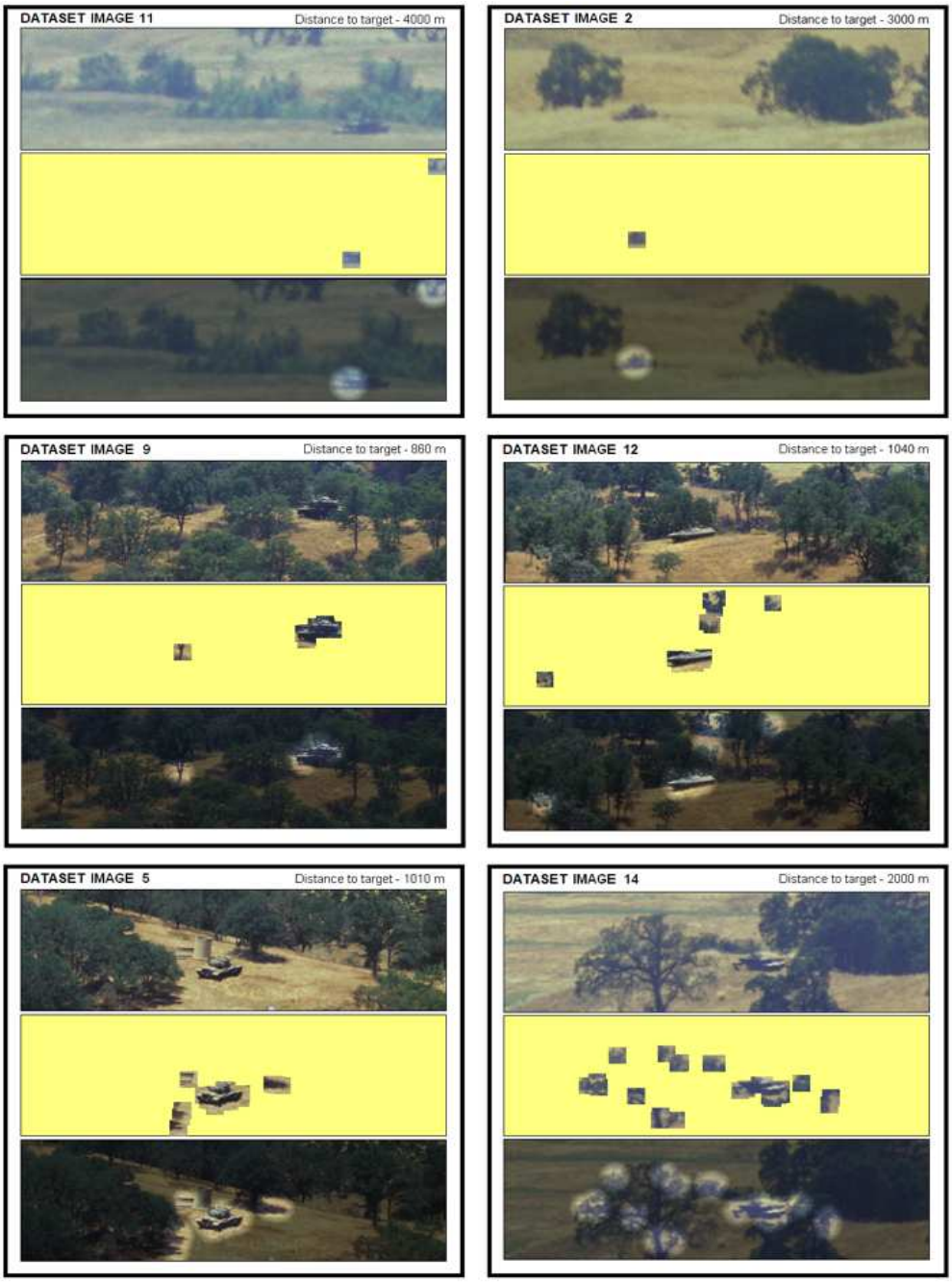

Figure 4. Target detection results of the proposed approach.