I. Opportunity Description

The Defense Advanced Research Projects Agency (DARPA) is issuing a new Artificial Intelligence Exploration (AIE) Opportunity entitled Techniques for Machine Vision Disruption (TMVD), which invites submissions of innovative basic or applied research concepts in the technical domain of disrupting machine vision systems without detailed knowledge of their internal architecture or how they were trained.

This Techniques for Machine Vision Disruption AIE Opportunity is being issued under the DARPA Program Announcement for AIE, DARPA-PA-19-03, which details the AIE Program’s overall intent and provides evaluation and submission instructions in addition to those outlined in this solicitation. To view the original DARPA Program Announcement for AIE, visit beta.SAM.gov (formerly FedBizOpps) under solicitation number DARPA-PA-19-03.

All proposals in response to the technical area(s) described herein will be submitted to the Techniques for Machine Vision Disruption, AIE Opportunity, solicitation number DARPA-PA-19-03-06, and if selected, will result in an award of an Other Transaction (OT) for Prototype Projects, not to exceed $1,000,000. This total award value includes Government funding and performer cost share, if required, or if proposed.

A. Introduction

The utility of machine vision dramatically improved when Convolutional Neural Nets (CNN) were brought to bear, as exemplified by the performance improvements obtained by AlexNet in 2012. Since that time, Artificial Intelligence (AI) computer vision has improved and progressed, achieving superhuman performance with real-time executable codes that can detect, classify and segment within a complicated image.

This revolution in computer vision capabilities was partly catalyzed by the availability of large amounts of labelled training images and readily-available cloud-based computational resources. The CNN paradigm allows a multi-layer network of relatively simple computational nodes or “neurons” to be trained with large numbers of labelled images to create detection or classification machines with a high degree of accuracy; that is, high precision and recall.

Performance can be impressive on both object detection (e.g., “chair” or “not-chair”) and general scene classification (e.g., “beach scene”, “mountain scene”, “scene containing at least one car”). Machine vision techniques allow for automated triage of mass data sets for big data analytics. For example, scraping online databases of images to identify things of interest (people, brands of purchases, density of automobiles, etc.) is now a straightforward task that does not require significant human labor to accomplish.

In the past several years, it has been shown that the fundamental statistical nature inherent in CNN architectures makes them somewhat brittle and easy to disrupt. Making specific alterations to an input image can, under certain circumstances, reduce the accuracy of the machine vision
system by “fooling” the trained classifier.\footnote{Christian Szegedy, et al. “Intriguing properties of neural networks”, arXiv preprint arXiv:1312.6199 (2013)} For example, pixel-based modifications (also called perturbations) add a seemingly random set of pixel values to an input image such that a human observer perceives no difference in the image, but the classifier generates a different output, e.g., it classifies the image as containing a “gibbon” instead of a “panda”.\footnote{Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. “Explaining and harnessing adversarial examples (2014).” arXiv preprint arXiv:1412.6572} Even more surprising is that such attacks can be engineered to be quasi-“universal”, in the sense that an adversarial perturbation generated for a particular image category also significantly degrades the accuracy of images in other categories and the accuracy in other neural net architectures.\footnote{Syed-Mohsen Moosavi-Dezfooli, et al. “Universal adversarial perturbations (2017).” arXiv preprint arXiv:1610:08401v3} However, many of these techniques assume access to the original training corpus and, at least in some cases, the original neural net architecture.

There is a different and somewhat complementary approach that adds a visible element to an input image, such as a “patch” or “graffiti”. These “patch attacks” can have a dramatic impact on performance, in some cases causing the classification machine to “see” a toaster instead of a banana or a 50 mile-per-hour speed limit sign instead of a stop sign.\footnote{Tom B. Brown, et al. “Adversarial patch.” arXiv preprint arXiv:1712.09665v2, May 2018} While these minor image alterations can disrupt many CNNs, they have little ability to fool human observers. Any person who looks at the “altered” image will still classify it correctly, and often won’t perceive a difference from the original image. This inspires questions about the fundamental nature of neural nets as currently implemented. Perhaps the early layers of the network are trained to detect statistically-relevant features that do not reflect “semantic” features that seem so obvious to humans. This in turn may play a part in the kinds of brittleness illustrated by perturbation and patch disruptive attacks. Development and exploration of universal disruption techniques, including the scientific phenomena that enables their success, will enhance our understanding of the inherent nature of neural net architectures and inform more robust approaches.

\section*{B. Objective/Scope}

The goal of the TMVD effort is to develop specific techniques to disrupt neural net-based computer vision technology in situations where neither the original training set nor the specific vision architecture are available. A canonical scenario is the addition of “noise” patterns to images, invisible or almost invisible to human inspection, to prevent the automatic scraping and classification of open-source images by adversarial data mining efforts. In Phase 1, it is permitted to concentrate on a single object class, as long as the resulting test images are of sufficient variety. Further, the desired misclassification of the target class can be any other dissimilar object class (“one to any” misclassification). In other words, the goal of the program/study is to develop image based attacks that whiten the output vector probability, such that a successful attack will have a very high probability that the output class will NOT be the actual class, but without a preference as to what output class is actually reported.

The techniques developed must require no knowledge of the actual images used to train the computer vision system. Further, the techniques developed must require no specific knowledge of the architecture, including the number of input units, the number of hidden layers, the number
of output units and whether they represent a “soft-max” computation, or the specific gradient
descent strategy used by training algorithms during back propagation.

The resulting technology should be as general or “universal” as possible. This means, first, that
the resulting disruption/attack patterns should be effective against a wide variety of images of
the subject at different distances, orientations, natural viewing conditions, degrees of occlusion,
and image resolution. In order to emphasize this robustness, performers should endeavor to
develop techniques that will still perform if the image is rescaled or filtered before processing.
Similarly, it is desired to demonstrate how effective such attacks are when applied to a sequence
of still frames (e.g. video frames) and how effective such techniques would be if video data is
available to scrape and classify. The video problem potentially combines all of the above
problems: changing viewing conditions, distances, occlusion, etc., as well as scaling and filtering
as video images may be compressed. Second, the techniques should generate attack patterns that
cause high probability of misclassification on multiple computer vision neural net architectures,
regardless of the internal structure and training function. These architectures could be standard
convolutional neural nets, schemes involving auto-encoders, recurrent neural networks (RNNs),
long short-term networks (LSTMs), generative adversarial networks (GANs), or other hybrid
architectures.

The Phase 1 Base effort will focus on the development of universal attack algorithms and
demonstrate algorithm success on at least three (3) networks with assessment of extensibility to
additional networks. The current assumption is that the Phase 1 effort will concentrate on attacks
that cause misclassification of a single object class, and will result in misclassifications into
other object classes with little visual or functional similarity to the target class. Phase 2 will be
an Option effort to: (a) analyze the developed algorithms from a signal processing perspective
and identify and characterize the underlying core principles behind their effectiveness; (b)
perform further performance testing of the developed techniques against multiple networks and
training schemes to evaluate the “universality” of the techniques. The performance testing in
Phase 2 must include testing with novel real-world images and not rely solely on standard image
datasets.

C. Structure

Proposals submitted to DARPA-PA-19-03-06 in response to this TMVD AIE Opportunity must
be UNCLASSIFIED and must address two independent and sequential project phases: Phase 1
Core Development (Base Period) and Phase 2 Evaluation and Determination of Core Principles
(Option Period). Notional Periods of Performance for these two Phases are twelve (12) months
for the Phase 1 Base Period and six (6) months for the Phase 2 Option Period. Combined Phase 1
Base and Phase 2 Option efforts for the TMVD AIE Opportunity should not exceed 18 months.
The Phase 1 (Base) award value is limited to $700K. The Phase 2 (Option) award value is
limited to $300K. The total award value for the combined Phase 1 and Phase 2 is limited to
$1,000,000. This total award value includes Government funding and performer cost share, if
required, or if proposed.

Proposals will be evaluated to the extent that the proposed approach addresses the Technical
Area Descriptions identified below. Phase 1 activities will be evaluated to determine feasibility
of the approach and whether to select the Option for Phase 2. Additionally, all proposals
submitted in response to TMVD AIE Opportunity will be evaluated and selected in accordance
with Section 6 of the DARPA AIE Program Announcement, DARPA-PA-19-03, available on

DARPA-PA-19-03-06
beta.SAM.gov here: https://beta.sam.gov/opp/5a5b919dcc337814fe57eb70c147bd72/view

The government may elect to provide TMVD performers access to government-furnished information (GFI) for optional or encouraged use in the program. This GFI may include image collections, software engineering standards, software application programming interfaces (APIs), test harnesses or test methodology documentation, or other such information. GFI may be discussed during contract negotiations and provided following project kickoff.

D. Technical Area Description

Computer vision attacks include so-called perturbation attacks involving the addition of what appears to be random noise to an image, and so-called patch attacks involving the addition of more structured coherent “patches” to the image. Unsurprisingly, there have been robust efforts to explore defenses against these attacks, as well as efforts to understand the causes and principles behind their effectiveness.\textsuperscript{5,6} These efforts hint at the possibility of a more robust, universal attack mechanism that targets the lowest “hidden” layers of the neural net in a general fashion. The hypothesis is that neural nets may be vulnerable to this kind of attack precisely because of the layered network architecture where early layers converge on relatively high-frequency, domain independent features.

TMVD proposals must describe a feasible and convincing strategy for designing and implementing universal adversarial attacks on neural net-based computer vision systems without access to training images or the neural net architecture. The following steps are recommended as part of this strategy, and all need to be executed during the Phase 1 Core Development phase:

- The generation of a comprehensive survey and understanding of the current state-of-the-art in neural net-based computer vision systems, adversarial attack methods and approaches, defense against adversarial attack methods and approaches, and current research directions in all of the above.
- The selection of a militarily-relevant image domain in the broader strategic scheme of operational attacks. Examples include military vehicles in imagery and video, facial recognition in imagery and video, and/or general pattern of life imagery that might have military utility (e.g. video feeds and social media scraping of associated imagery in/around a facility).
- The determination of a data acquisition plan for TMVD that describes how potential training and testing images will be acquired and labeled in sufficient numbers. The use of standard open-source image collections may be sufficient initially, but other sources of images, especially for real-world testing, may be required for full TMVD effectiveness. This plan should include a means for acquiring sufficient images in the militarily-relevant image domain. Automatically generated synthetic images may be included if their realism is high.
- An execution and testing plan for determining the effectiveness of the developed technology. This plan must include the testing of sufficient numbers of images in and out of the military domain over a sufficient breadth of image acquisition conditions and, finally, over a set of at least three (3) diverse neural net architectures. These neural net


DARPA-PA-19-03-06
architectures can include open source architectures and implementations.

In addition, the image domain(s) proposed should meet the following conditions:

- Digital images used in TMVD, both in the militarily-relevant domain and others, must include a wide variety of natural conditions as opposed to just close-up “passport” images of the objects or object class of interest.
- These images must further reflect a variety of distances, poses, resolutions, colorations and occlusion conditions as would be expected in “images in the wild”.
- Images must be visual / electro-optical. Out-of-scope image collections for TMVD would include biological images (e.g., from microscopy), imagery reflecting non-visual spectra and conditions, SAR imagery, and other imagery not considered electro-optical “images in the wild”.

TMVD proposals should include a Phase 2 Option: Evaluation and Determination of Core Principles section that describes a technical plan to evaluate the successful algorithms from a signal processing perspective and identify and characterize the underlying principles behind their effectiveness. The evaluation piece should include:

- A recommended set of additional neural net architectures for evaluation and a mechanism for procuring (e.g., from open sources), training, and testing them using the universal attacks developed during Phase 1.
- A recommended set of additional images or image sources. Phase 2 testing cannot rely solely on standard computer vision image datasets.
- A recommended set of available defensive mitigation strategies and a proposal for implementing a subset of these on a subset of the available neural net architectures.
- A methodology and preferably a mathematical framework for studying and determining the underlying core principles and theory behind the operation of neural nets and the attack mechanisms developed for TMVD. The purpose of this piece is to understand the robustness of the techniques in the face of future algorithmic defenses and suggest new areas for research.

E. Schedule/Milestones

Proposers must address the following Research Project Objectives, metrics, and deliverables, along with fixed payable milestones in their proposals. The task structure must be consistent across the proposed schedule, Task Description Document (TDD), and the Vol. 2 - Price Volume. If selected for award negotiation, the fixed payable milestones provided in Section 9.C of the Vol. 2 – Price Volume will be directly incorporated into Attachment 3 of the OT agreement (“Schedule of Milestones and Payments”). Please see the sample OT for Prototype provided as an attachment to AIE Program Announcement, DARPA-PA-19-03, available on beta.SAM.gov.

For planning and budgetary purposes, proposers should assume a program start date of **July 14, 2020**. Schedules will be synchronized across performers, as required, and monitored/revised as necessary throughout the program. Proposals must include delivery schedules for Phase 1 and Phase 2 that include timelines for preliminary (to facilitate inspection by the Program Manager) and final (to facilitate evaluation) release of deliverables. 

Phase 1 Milestones include:
• **Month 1 After Contract Award (ACA): Project kick-off:** Review of programmatic and technical development plans. Provide a comprehensive description of the strategy for designing and implementing universal adversarial attacks on neural net-based computer vision systems without access to training images or the neural net architecture. Include an overview and plan for generating a comprehensive survey of the state-of-the-art in a timely fashion; further describe how the results of this survey can be incorporated into the overall strategy. Identify a militarily-relevant image class and justify its selection. Provide a credible plan for acquiring labeled images of sufficient quality, quantity, and variety, including those in the militarily-relevant class. Identify and address major challenges (e.g. baseline for comparison, acquisition of training data, dealing with state space complexity, metrics for assessing balance) and provide plan to address each. Provide proposed schedule and programmatic approach, to include roles and responsibilities of project team, key contributors, and any subcontractors. Describe research activities to be performed over the following 3 months.

• **Months 3, 6, and 9 ACA: Interim progress reports:** Provide an interim report describing the progress over the prior 3 months in acquiring a trainable neural net architecture and the images needed to train it. Describe initial design and implementation progress in implementing a universal computer vision neural net attack system. Describe progress towards design and implementation of an evaluation methodology. Provide updates to the overall research plan based on the comprehensive survey of the current state-of-the-art and progress to date. Describe in detail all research exploration activities, including those that “failed”, in order to document the full scope of funded activities and lessons learned. Describe research activities to be performed over the following 3 months.

• **Month 12 ACA: Feasibility study final review:** This milestone should present a compelling case that the universal attack algorithms have solid performance in both a normal context and additional neural nets implemented with at least one defense against adversarial interference. Evidence should be presented that the techniques hold promise for “universality” on other neural net architectures. Describe in detail all research exploration activities, including those that “failed”, in order to document the full scope of funded activities and lessons learned. Review should include discussion of how Phase 1 results will be used to determine and tune Phase 2 strategy. Submit Phase 1 Final Report in preparation for the Phase 2 option. Phase 1 final report should summarize the approach for Phase 2 tasks.

Phase 2 Milestones include:

• **Month 13 ACA: Phase 2 Kick-off.** Review plan and schedule for Phase 2 Option tasks. Present extended evaluation plan, including acquisition and use of multiple neural net architectures reviewed. Describe plan for studying and determining underlying core principles and theory behind the TMVD disruption/attack phenomenon, including the technology or mathematical approach to be used.

• **Month 15 ACA: Interim progress report.** Provide a general review of progress against evaluation activities and definition of core principles, including all research exploration activities.

• **Month 18 ACA: Project final review.** Present summary of all work performed during Phases 1 and 2, with results from Phase 2 evaluation activities. Present review of identified core principles and suggested new research problems and directions. Submit Phase 2 final report.
All proposals must include the following meetings and travel in the proposed schedule and costs:

- To foster collaboration between teams and disseminate program developments, a one-day Principal Investigator (PI) meeting will be held approximately every six months, for a total of three (3) events. For proposal costing purposes, proposers can assume two (2) on the East Coast and one (1) event on the West Coast of the United States.
- Regular teleconference meetings will be scheduled with the Government team for progress reporting, as well as problem identification and mitigation. For proposal costing purposes, proposers can assume one (1) call of one (1) hour each month. Proposers should also anticipate at least one site visit per Phase by the DARPA Program Manager during which they will have the opportunity to demonstrate progress towards agreed-upon milestones.

F. Deliverables

Performers are expected to provide the following deliverables throughout the course of the effort. Refer to descriptions in Section E of this document for additional content details.

<table>
<thead>
<tr>
<th>Item</th>
<th>Item Type</th>
<th>Due Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phase 1 Kick-off Meeting Slides</td>
<td>Slides (PPT, PDF)</td>
<td>1 Month ACA</td>
</tr>
<tr>
<td>Phase 1 Interim Progress Report</td>
<td>Report (PDF, MS Word)</td>
<td>3, 6, 9 Months ACA</td>
</tr>
<tr>
<td>Phase 1 Final Review Slides</td>
<td>Slides (PPT, PDF)</td>
<td>12 Months ACA</td>
</tr>
<tr>
<td>Phase 1 Final Report</td>
<td>Report (PDF, MS Word)</td>
<td>12 Months ACA</td>
</tr>
<tr>
<td>Phase 2 Kick-off Meeting Slides</td>
<td>Slides (PPT, PDF)</td>
<td>13 Months ACA</td>
</tr>
<tr>
<td>Phase 2 Interim Progress Report</td>
<td>Report (PDF, MS Word)</td>
<td>15 Months ACA</td>
</tr>
<tr>
<td>Phase 2 Final Review Slides</td>
<td>Slides (PPT, PDF)</td>
<td>18 Months ACA</td>
</tr>
<tr>
<td>Phase 2 Final Report</td>
<td>Report (PDF, MS Word)</td>
<td>18 Months ACA</td>
</tr>
<tr>
<td>Source code / Software for</td>
<td>Software</td>
<td>12, 18 Months ACA</td>
</tr>
<tr>
<td>developed solutions</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The **Phase 1 Final Report** is expected to include, at a minimum, the following details:

- Results of the comprehensive survey of current state-of-the-art approaches for both adversarial attack methods and defense against such
- A detailed description of the TMVD concept(s) and technology developed within this effort, including:
  - The evolution and intermediate steps in their development
  - The relationship of the developed concept(s) and technique(s) to the current state-of-the-art and recognized research directions;
  - Predictions of each concept’s applicability and performance against targeted image classes, including predicted general applicability and “universality” under different image conditions and on at least three (3) different neural net architectures
- A Description of the specific evaluation strategy used to assess performance of the TMVD concept(s); including:
  - The complete set of labeled and unlabeled images acquired for developing, training, and testing the technology; including the set of militarily-relevant images, as well as any synthetic images generated within the effort and the DARPA-PA-19-03-06
means by which they were generated (anticipated to be an appendix or data file)
  o Descriptions of the neural net-based computer vision systems against which the
    solutions were evaluated
  o Comprehensive results of the evaluation against at least three (3) different
    neural net architectures
  
The **Phase 2 Final Report** is expect to include, at a minimum, the following details:

- Descriptions of the additional neural net architectures used in the Phase 2 evaluation
- Comprehensive results of the evaluation against the additional neural net architectures, including the generalizability or “universality” demonstrated by the evaluation
- Detailed description and documentation of the methodology and framework used to study and determine the underlying principles behind the attack mechanisms and their performance
- Detailed description of potential defensive strategies for preventing or mitigating the results of the attack/disruption technology developed within this effort
- Recommendations for follow-on research

The **Software deliverables** should include files of sufficient quality and completeness to enable the Government customer to build and run (replicate) all implementations developed within this effort. This is anticipated to include, but not be limited to:

- Original source code for developed solution(s). This includes original sources code for any developed synthetic image generation solutions.
- Documentation and execution build instructions
- Description of any third-party libraries and dependencies needed to build and run the developed implementations
- All real & synthetic imagery (acquired or generated) used within this effort, and all labels and metadata associated with the imagery.

### II. Award Information

Selected proposals that are successfully negotiated will result in award of an OT for Prototype Project. Refer to Section 3 of the AIE Program Announcement (DARPA-PA-19-03) for information on awards that may result from proposals submitted in response to this notice. Proposers must review the model OT for Prototype agreement provided as an attachment to the AIE Program Announcement (DARPA-PA-19-03) prior to submitting a proposal. DARPA has provided the model OT in order to expedite the negotiation and award process and ensure DARPA achieves the goal of AIE, which is to enable DARPA to initiate a new investment in less than 90 days from idea inception. The model OT is representative of the terms and conditions that DARPA intends to award for all AIE Opportunities. The task description document, schedule of milestones and payments, and data rights assertions requested under Volumes 1, 2, and 3 of the AIE Opportunity will be included as attachments to the OT agreement upon negotiation and award.

Proposers may suggest edits to the model OT for consideration by DARPA and provide a copy of the model OT with track changes as part of their proposal package. Please note that suggested edits, however, may not be accepted by DARPA. The Government reserves the right to remove a proposal from award consideration should the parties fail to reach agreement on OT award terms.

DARPA-PA-19-03-06
and conditions. If edits to the model OT are not provided as part of the proposal package, DARPA assumes that the proposer has reviewed and accepted the award terms and conditions to which they may have to adhere and the sample OT agreement provided as an attachment, indicating agreement (in principle) with the listed terms and conditions applicable to the specific award instrument.

In order to ensure that DARPA achieves the AIE goal of award within 90 days from the posting date (April 15, 2020) of this announcement, DARPA reserves the right to cease negotiations when an award is not executed by both parties (DARPA and the selected organization) on or before July 13, 2020.

III. Eligibility

Refer to Section 4 of the AIE Program Announcement (DARPA-PA-19-03) for eligibility information.

IV. AIE Opportunity Responses

Responses to the Techniques for Machine Vision Disruption AIE Opportunity must be submitted as full proposals to DARPA-PA-19-03 as described therein. All proposals must be UNCLASSIFIED.

A. Proposal Content and Format

All proposals submitted in response to this notice must comply with the content and format instructions in Section 5 of the AIE Program Announcement (DARPA-PA-19-03). All proposals must use the templates provided as Attachments to the PA and follow the instructions therein.

Information not explicitly requested in the AIE Program Announcement (DARPA-PA-19-03), its attachments, or this notice may not be evaluated.

B. Proposal Submission Instructions

Refer to Section 5 of the AIE Program Announcement (DARPA-PA-19-03) for proposal submission instructions.

C. Proposal Date and Time

Proposals in response to this notice are due no later than 4:00 PM, ET on May 14, 2020. Full proposal packages as described in Section 5 of the AIE Program Announcement (DARPA-PA-19-03) must be submitted per the instructions outlined therein and received by DARPA no later than the above time and date.

Proposals received after this time and date may not be reviewed.

Proposers are warned that the proposal deadline outlined herein is in Eastern Time and will be strictly enforced. When planning a response to this notice, proposers should take into account that some parts of the submission process may take from one business day to one month to complete.
V. Proposal Evaluation and Selection

Proposals will be evaluated and selected in accordance with Section 6 of the AIE Program Announcement (DARPA-PA-19-03). Proposers will be notified of the results of this process as described in Section 7.1 of that same announcement.

VI. Administrative and National Policy Requirements

Section 7.2 of the AIE Program Announcement (DARPA-PA-19-03) provides information on Administrative and National Policy Requirements that may be applicable for proposal submission as well as performance under an award.

VII. Point of Contact Information

Dr. Greg Avicola, Program Manager, DARPA/TTO, PA-19-03-06@darpa.mil.

VIII. Frequently Asked Questions (FAQs)

All technical, contractual, and administrative questions regarding this notice must be emailed to PA-19-03-06@darpa.mil. Emails sent directly to the Program Manager or any other address may result in delayed or no response.

All questions must be in English and must include name, email address, and the telephone number of a point of contact. DARPA will attempt to answer questions publicly in a timely manner; however, questions submitted within seven (7) days of the proposal due date listed herein may not be answered.

DARPA will post a Frequently Asked Questions (FAQ) list under the AIE Opportunity on the DARPA AI Next Campaign webpage at https://www.darpa.mil/work-with-us/ai-next-campaign. The list will be updated on an ongoing basis until one week prior to the proposal due date. In addition to the FAQ specific to this notice, proposers should also review the Program Announcement for AIE General FAQ list on the DARPA Opportunities page under the Program Announcement for AIE (DARPA-PA-19-03).