Artificial Intelligence and Data Fusion at the Edge

Arslan Munir, *Senior Member, IEEE,* Erik Blasch, *Fellow, IEEE,* Jisu Kwon,
Joonho Kong, *Member, IEEE,* and Alexander Aved, *Senior Member, IEEE*

Abstract*—Artificial intelligence (AI), owing to recent breakthroughs in deep learning, has revolutionized applications and services in almost all technology domains including aerospace. AI and deep learning rely on huge amounts of training data that is mostly generated at the network edge by Internet of things (IoT) devices and sensors. Bringing the sensed data from the edge of a distributed network to a centralized cloud is often infeasible because of the massive data volume, limited network bandwidth, and real-time application constraints. Consequently, there is a desire to push AI frontiers to the network edge towards utilizing the enormous amount of data generated by IoT devices near the data source. The merger of edge computing and AI has engendered a new discipline, that is, *AI at the edge* or *edge intelligence*. To help AI make sense of gigantic data at the network edge, *data fusion* is of paramount significance and goes hand in hand with AI. This article focuses on data fusion and AI at the edge. In this article, we propose a framework for data fusion and AI processing at the edge. We then provide a comparative discussion of different data fusion and AI models and architectures. We discuss multiple levels of fusion and different types of AI, and how different types of AI align with different levels of fusion. We then highlight the benefits of combining data fusion with AI at the edge. The methods of AI and data fusion at the edge detailed in this article are applicable to many application domains including aerospace systems. We evaluate the effectiveness of combined data fusion and AI at the edge using convolutional neural network (CNN) models and multiple hardware platforms suitable for edge computing. Experimental results reveal that combining AI with data fusion can impart a speedup of 9.8× while reducing energy consumption up to 88.5% over AI without data fusion. Furthermore, results demonstrate that data fusion either maintains or improves the accuracy of AI in most cases. For our experiments, data fusion imparts a maximum improvement of 15.8% in accuracy to AI.[1]

***Index Terms*—Edge computing, artificial intelligence, machine learning, data fusion, swarm intelligence, deep neural networks**

I. Introduction and Motivation

Advancements and miniaturization of electronic devices that are not only able to sense and process data but can also communicate with other devices have engendered an era of Internet of things (IoT). These IoT devices are often equipped with a multitude of sensors and generate zillions bytes of data at the network edge. Many of the applications that need to utilize this data have embraced artificial intelligence (AI) and machine learning (ML). However, transferring this gigantic data to the cloud is often infeasible due to limited network bandwidth and real-time constraints of many applications including aerospace systems. Thus, there is a desire to push AI frontiers to the network edge to utilize the enormous amount of data generated by IoT devices nearer to the data source. This desire has led to the merger of *edge computing* (a novel trend in computing that pushes computing power away from the centralized nodes to the logical extreme edges of a network [1]) and AI, resulting in a new discipline — *AI at the edge* or *edge intelligence*. In the edge AI model, AI computations take place either on the user device or somewhere in the network stack beneath the cloud, perhaps on an edge server. According to Steve Roddy, the vice president (VP) of Special Projects in Arm's Machine Learning Group [2]: "The edge is the next stage of the evolution of AI technology because of the physical constraints, the cost constraints, and the practical constraints of running all AI applications in the cloud. It simply doesn't make sense to send all the bits for things like video and audio streaming to the cloud and back down for every situation, every endpoint."

In the age of AI, what is important is not the data alone but what we can do with this data and how to make sense of this data. As more and more data is being collected at the network edge due to the proliferation of IoT and sensing devices, the capability gap to make sense of this gigantic data, whether at the edge or the cloud, in a timely manner is also increasing. For instance, in surveillance applications, the number of traditional sensors (e.g., ground radar, cameras), non-traditional sensors (e.g., IoT), and non-organic airborne platforms (e.g., unmanned aerial systems) has increased the opportunity to detect, track, and identify targets, as well as to counter threats; however, there is a lack of processing capability to do so efficiently and effectively [3]. It is noted that for many applications, much of the collected data is time-sensitive and become useless if not utilized timely. Hence, solutions such as *data fusion* are of paramount significance to enhance the effectiveness and usage of sensed data in a timely manner. Data fusion is defined as the process of combining data from multiple sources to produce more accurate, consistent, and concise information than that provided by any individual data source.

The concept of both data fusion and AI has biological origins. Data fusion is inspired from the capability of advanced biological organisms to assimilate information from multiple senses (e.g., sight, touch, smell, taste) to make better sense of environment and increase their chances of survival. AI

Arslan Munir is with the Department of Computer Science, Kansas State University, Manhattan, Kansas, USA. Arslan Munir's current address is 2162 Engineering Hall, 1701D Platt St, Manhattan, KS, 66506, USA. Erik Blasch is a program officer at the Air Force Office of Scientific Research (AFOSR). Jisu Kwon and Joonho Kong are with the School of Electronic and Electrical Engineering, Kyungpook National University, South Korea. Alex Aved is technical advisor at the Air Force Research Laboratory (AFRL), Information Directorate, Rome, NY. e-mail: {amunir@ksu.edu, erik.blasch.1@us.af.mil, jisu92@knu.ac.kr, joonho.kong@knu.ac.kr, Alexander.Aved@us.af.mil}

[1] This is the author's version of an article that has been published in IEEE Aerospace and Electronic Systems Magazine, vol. 36, no. 7, July 2021. The final version of record is available at: https://doi.org/10.1109/MAES.2020.3043072

is also inspired from the cognition displayed by advanced biological organisms (e.g., humans). Both data fusion and AI are complementary to enable machines to accomplish various intelligent tasks and missions. Whether the AI computations are performed at the cloud or at the network edge, data fusion is needed to provide more concise and consistent data to the AI both at the training and inference phases. This article focuses on AI at the edge and illustrates the benefits of integrating data fusion with AI at the edge.

Our main contributions in this article are as follows:

- We propose a framework for AI and data fusion at the edge.
- We provide a comparative discussion of different data fusion and AI architectures.
- We discuss multiple levels of fusion and different types of AI, and how different types of AI relate to and align with different levels of fusion.
- We highlight and elaborate the advantages of AI and data fusion at the edge including latency, energy efficiency, precision, security, privacy, cost, scalability, and sustainability.
- We present experimental results for latency, energy efficiency, and accuracy for AI and data fusion and AI without data fusion to demonstrate the advantages of AI and data fusion at the edge.

The remainder of this article is organized as follows. Section II presents a framework for data fusion and AI at the edge. Section III provides a comparative discussion of contemporary data fusion and AI architectures. Section IV delineates multiple levels of fusion and different types of AI, and also provides a mapping of different AI and ML types to different fusion levels. Advantages of data fusion and AI at the edge are discussed in Section V. Experimental results are presented in Section VI. Finally, Section VII concludes this article.

II. Framework for Data Fusion and AI at the Edge

In this section, a high-level framework for data fusion and AI/ML processing at the edge is developed. To explain the proposed framework, we provide a brief overview of AI and ML and elaborate on how data fusion enhances AI. Contemporary AI is dominated by ML. Multitude of ML methods, which can be categorized under supervised learning, semi-supervised learning, unsupervised learning, and reinforcement learning, help provide edge intelligence. Deploying an ML model requires first *training* the model and then using the trained model to perform *inference*. Since training is more resource-intensive and time consuming process; in the edge AI model, training is typically done on the cloud and inference is performed at the edge. Although AI training can be performed at edge [4], given the limited resources of many edge devices, AI training will likely continue on powerful cloud-based computers whereas inference will be performed at edge devices. Among the existing ML methods, deep learning provides magnificent performance on various tasks. For example, convolutional neural networks (CNN) have been utilized for image classification, object detection, and recognition. Recurrent neural networks (RNNs) find applications in natural language processing and multi-target tracking. Deep reinforcement learning (DRL) can be leveraged for determining optimal policy/strategy for accomplishing various tasks, such as trajectory optimization of autonomous vehicles. Regardless of any specific AI/ML methods, edge computing provides a means for performance- and energy-efficient execution of AI/ML algorithms.

Fig. 1 depicts our proposed framework for data fusion and AI at the edge. In the proposed framework, data fusion, AI/ML processing, analysis, and decision-making are done at three levels: (i) edge-of-network sensor/IoT nodes, (ii) edge servers or fog nodes, and (iii) cloud servers. AI at both the edge-of-network sensor/IoT devices and edge servers (fog nodes) is referred to as *edge AI*. Edge AI enables computations near the edge of the network and helps in reducing the communication burden on the core network. We distinguish the edge AI done at senor/IoT node level and edge server (fog node) level to highlight the difference in compute capability of the two. In the proposed framework, AI is first performed at the lowest tier of the network edge comprising of senor nodes and IoT devices. These edge-of-network sensor/IoT devices can be sensor nodes sensing particular features (e.g., temperature, humidity, pressure), smart phones, smart vehicles, video and imaging devices (e.g., cameras including night vision imaging cameras, thermal imaging cameras, etc.), and even airborne vehicles, such as unmanned aerial vehicles (UAVs), equipped with different sensors. These edge-of-network IoT devices typically possess limited computation and communication capabilities. Given the increasing proliferation of AI-driven applications, modern IoT devices can be outfitted with various ML accelerators to speed up the execution of ML algorithms in an energy-efficient manner as depicted in Fig. 1. For example, IoT nodes equipped with cameras can execute hardware-accelerated lightweight CNN algorithms, such as MobileNet [5], to perform object classification.

Although modern and futuristic IoT devices can be equipped with AI accelerators [6] as depicted in Fig. 1, many of the contemporary edge-of-network sensors/IoT devices have limited computational capabilities that are not adequate to carry out complex AI tasks. In order to boost the computational and AI capabilities at the network edge, edge servers are installed at the base stations in the edge computing paradigm. Edge servers possess much more resources and computational capability than the edge-of-network sensors/IoT devices. Since edge servers receive data from many edge-of-network devices with diverse application requirements, performance requirements from edge server are also diverse. When AI/ML processing with stringent latency requirements (e.g., for hard real-time systems) is required from edge servers, edge servers are designed with high-performance stationary computing servers with stable power source and high bandwidth network connections. For mobile edge processing, edge servers can be power-/energy-constrained. For example, UAVs, which are typically powered by batteries, can be used as mobile edge servers or devices. These mobile edge devices often employ AI accelerators for fast and energy-efficient AI/ML processing [7] [8]. Moreover, edge servers or fog nodes are more amenable for integration of ML accelerators, such as CNN, RNN, and MLP accelerators as depicted in Fig. 1.

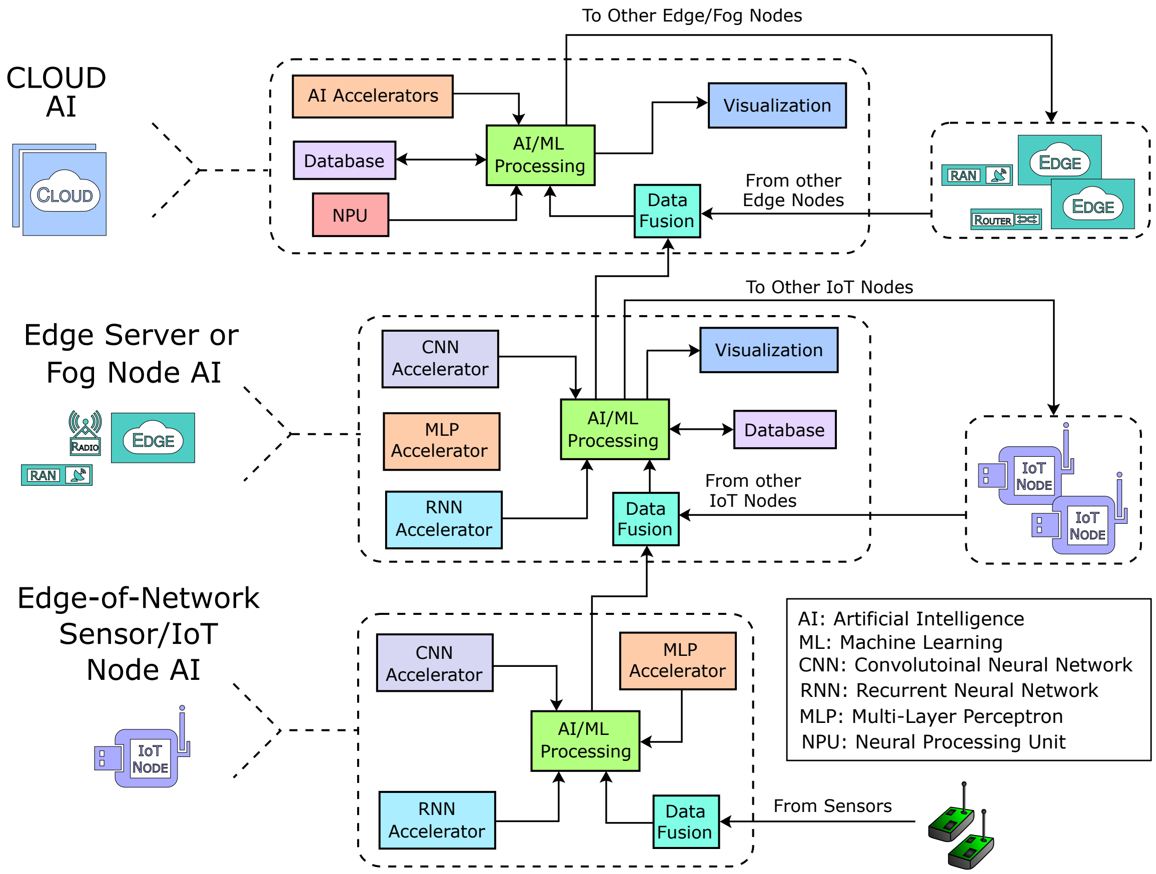

Fig. 1: Framework for data fusion and AI at edge.

Each edge server manages a cluster of edge-of-network sensors/IoT devices in its vicinity. Edge servers provide applications, content, context, services, and storage to edge-of-network IoT devices. For example, edge servers can assist with complex image processing and storage tasks for imaging data acquired from camera sensors at IoT devices. Since edge servers are in close proximity to edge-of-network IoT devices, the edge servers can perform the offloaded tasks from IoT devices with much lower latency and much lower communication overhead as compared to the cloud.

The edge servers in our framework are connected to the top-tier centralized cloud server layer through the core network. The core network consigns locally processed data and information from the edge to the cloud for various purposes such as analytics, archival, and decision-making at a broader scale. At the cloud, high-performance servers are typically used because the cloud needs to service many requests from the edge-of network devices and edge servers. Cloud servers are typically equipped with many high-end central processing units (CPUs) (e.g., Intel Xeon processor [9]) and graphic processing units (GPUs) for database processing and AI/ML processing. For efficient execution of AI tasks, the cloud is often outfitted with a variety of AI accelerators and neural processing units (NPUs) as depicted in Fig. 1. As an example, Google is currently deploying a proprietary AI accelerator called tensor processing unit (TPU) [10] to accelerate the AI/ML workloads in their data centers. The TPU is designed and implemented as an application-specific integrated circuit (ASIC), which is not reconfigurable after the chip fabrication. As another example, Microsoft uses field-programmable gate arrays (FPGAs) for accelerating AI/ML workloads in their cloud servers [11]. Since FPGAs are reconfigurable, the accelerator logic can be flexibly changed depending on users' or service providers' requirements.

In the proposed framework, data fusion plays an important role along with AI as depicted by the data fusion blocks at each hierarchical tier in Fig. 1. Here, we provide a high-level discussion of data fusion at different tiers of the proposed framework whereas the detailed discussion of data fusion that occurs within data fusion blocks in Fig. 1 is provided in Section III. At the lowest tier, data fusion is performed at the IoT node-level to minimize the redundancy in the raw data acquired from sensors. The IoT devices then perform AI on this fused data to help improve accuracy, performance, and energy efficiency in carrying out the AI tasks. The IoT nodes transmit only the sanitized and fused data, and the analytics results on this fused data to the edge servers instead of sending the raw sensor data, which enormously reduces the load on communication network and also conserves the communication/transmission energy at IoT devices. The edge servers then fuse the data received from multiple IoT devices. The edge servers also resolve the topological relationships between sensors and utilize the topological, contextual, and environmental information in data fusion. The edge servers perform AI on this fused data and then report the sanitized and fused data as well as analytics on this fused data to the cloud. The cloud performs the data fusion on the data received from the network edge and then performs AI on this fused data to obtain global analytics and insights.

In the proposed framework, an IoT edge device can computationally offload its data fusion and AI tasks to other edge devices or an edge server. Similarly, edge servers or fog nodes can offload their tasks to cloud. Since many of the AI inference tasks are time-sensitive, computation offloading from an IoT device to another edge-of-network device d_E is advantageous if

$$T_p^{IoT} + T_q^{IoT} > T_p^{d_E} + T_q^{d_E} + T_t^{IoT\text{–}d_E}, \qquad (1)$$

where T_p^{IoT} and T_q^{IoT} represent the average processing latency and average queuing delay, respectively, at the IoT device; $T_p^{d_E}$ and $T_q^{d_E}$ denote the average processing latency and average queuing latency, respectively, at d_E; and $T_t^{IoT\text{–}d_E}$ denotes the transmission latency from an IoT device to another edge-of-network device d_E for sending the data for offloaded computation and receiving the results back from d_E. Similarly, the computation offloading from an IoT device to an edge server S_E is expedient if

$$T_p^{IoT} + T_q^{IoT} > T_p^{S_E} + T_q^{S_E} + T_t^{IoT\text{–}S_E}, \qquad (2)$$

where $T_p^{S_E}$ and $T_q^{S_E}$ denote the average processing latency and average queuing latency, respectively, at S_E; and $T_t^{IoT\text{–}S_E}$ denotes the transmission latency from an IoT device to an edge server S_E for sending the data for offloaded computation and receiving the results back from S_E. The auspiciousness of offloading from an edge server to the cloud can be expressed similar to Eq. (1) and Eq. (2) and is omitted for brevity.

III. Models and Architectures for Data Fusion and AI

There exist many models and architectures for data fusion to address plethora of issues surrounding human factors, AI, and IoT [19]. Table I presents seven major data fusion models and/or architectures and compares them across different metrics, viz., ability to perform (static) data fusion, (dynamic) real-time sensing, capability of operating with human-in-the-loop, applicability to IoT and/or cyber-physical systems (CPS), handling of AI (including big data), and centralized or distributed processing. While Table I highlights traditional data fusion models and architectures with their current capabilities, it is to be noted that with new advances, these architectures can be redefined and enhanced towards recent methods, for example, static to dynamic processing, centralized to distributed processing, human machine teaming, and small data (1/0s) to large data (streaming video).

Typically, the aerospace community focuses on two main data fusion models or architectures. The first one leverages multiple filters in a centralized approach that has been widely extended to distributed methods and applied to object assessment for target tracking and distributed target recognition [20]. A good example of this type of data fusion architecture is the distributed information graph (DIG). The second one brings together big data to a user-defined operating picture (UDOP) in the Joint Directors of Laboratories (JDL)/Data Fusion Information Group (DFIG) model [17]. We note that the UDOP concept extends the common operating picture (COP) to help enable the human operator to supervise the processing, exploitation, and dissemination of information for situation awareness. The UDOP enables rendering and visualization of data analytics services customized to the human operator's needs for efficient command decision-making for a given mission. While the DFIG model supports multiple sensors and distributed users, it still relies on a common location for command and control. Other communities have been looking at fusion architectures that emphasize fusion-focused data analysis such as the data-feature-decision (DFD) [12] model, wireless sensor networks [14], and the more recent efforts from the AI community with the early and late fusion (ELF) methods [15] and generative adversarial networks (GANs). Since 2000, the DFD and the JDL/DFIG have remained similar, while the WSN community has continually utilized the distributed approaches that extend to IoT and CPS. While each of these methods are important for the fusion community, there is a desire for a data fusion architecture that is able to perform distributed command and control amongst a variety of human teams and distributed sensors. Hence, in this section, we elaborate cyber-physical command guided (CPCG) architecture that is able to perform distributed command and control amongst various human teams and distributed sensors and IoT devices. We further briefly discuss DFIG model as it help provide an understanding of fusion levels.

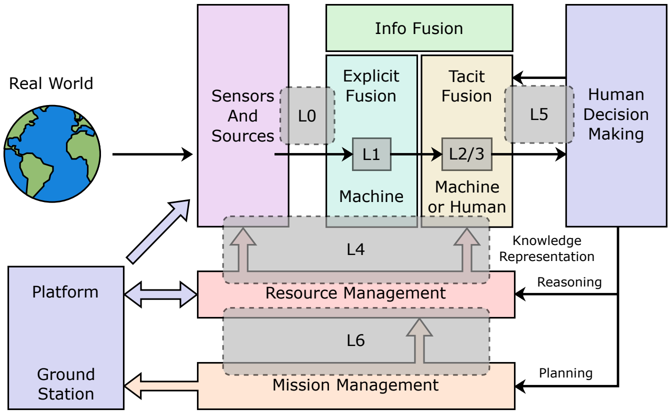

Fig. 2: DFIG model for information fusion.

A. Data Fusion Information Group Model

Fig. 2 depicts JDL/DFIG model. The JDL/DFIG model defines multiple levels of fusion where each level exploits AI developments to support assessment (level 0, 1, 2 and 3) and refinement (level 4, 5, and 6). System management (level 6) incorporates contextual constraints based on mission, objectives, and goals. The DFIG model describes different levels of fusion as:

Level 0 — Data Assessment: Provides estimation and prediction of signal/object observable states based on pixel/signal level data association.

Level 1 — Object Assessment: Provides estimation and prediction of entities/objects based on data association and (both continuous and discrete) state estimation.

Level 2 — Situation Assessment: Provides estimation and prediction of relations between entities/objects.

TABLE I: Comparison of data fusion and AI architectures.
Symbols: ✓– yes; ★ – being used; ■ – been adapted; □ – could be adapted; ✗ – not used.
("yes" indicates that (yes) research supports the directions/usage while "used" implies that it has been implemented in real systems.)

Model and/or Architecture	**Data Fusion (Static)**	**Real-Time (Sensing)**	**Human Centered**	**IoT/ CPS**	**AI (Big Data)**	**Connected (Distributed)**	**Reference**
DFD	✓	★	□	✗	■	□	Dasarathy [12]
DIG	★	✓	■	★	□	✓	Hall et al. [13]
WSN	★	✓	✗	■	□	✓	Chair and Varshney [14]
ELF	★	□	✗	✗	✓	★	Snoek et al. [15]
GAN	★	✗	✗	✗	✓	✗	Goodfellow et al. [16]
JDL/DFIG	✓	★	✓	□	★	★	Blasch et al. [17]
CPCG	✓	✓	✓	✓	✓	✓	Cruise et al. [18]

Level 3 — Impact Assessment: Provides estimation and prediction of impact on planned or estimated actions by the participants.

Level 4 — Process Refinement: An element of *resource management* and provides adaptive data acquisition and processing to furnish estimation of sensing objectives and prediction of impact on planned or estimated actions by the participants.

Level 5 — User Refinement: An element of *knowledge management* and provides adaptive determination of access control and display of information to support decision-making via human-machine interface.

Level 6 — Mission Management: An element of *systems management* and enables spatial-temporal control of assets (e.g., airspace operations), route planning, and goal determination to support decision-making while considering social, economic, and political constraints.

B. Cyber-Physical Command-Guided Architecture

As the name implies, the CPCG architecture combines the elements of the static (DFD) and dynamic (DIG) methods from the late 1990s with the user-focused updates from the JDl/DFIG in the 2000s for "Command Guided" systems, such as swarm of unmanned domain systems (UxS) (where x can be space, air, ground, surface, or undersea). The CPCG architecture seeks to not only utilize the distributed data fusion but also the distributed diffusion of command to cyber-physical elements. Hence, the CPCG architecture leverages the cloud and edge processing to be able to collect data for information fusion (IF), afford consumption and analytics by operator infusion (OI), and then direct needs for control diffusion (CD). The CPCG architecture takes advantage of centralized command with distributed execution by expressing goals and having the contextual agents develop the sensing and action strategy.

The AI agents in CPCG architecture mine data, process and fuse information, and store the results in a distributed space. The AI agents of CPCG architecture are assisted with three different types of data fusion, viz., (i) information fusion, (ii) operator infusion, and (iii) control diffusion. Organization of these three data fusion agents induces different AI architectures as depicted in Fig. 3. These fusion techniques can be characterized based on the data they provide and/or operate on:

(i) Information Fusion (IF): The IF agent mines and processes physical data/information that originates in the external environment. The IF can be assisted with different AI/ML approaches, such as symbolic, probabilistic, connectionist, analogistic, evolutionary, and possibilistic (please refer to Section IV for types of AI and ML).

(ii) Operator Infusion (OI): The OI agent assimilates human-in-the-loop within CPCG architecture for interpreting and assessing processed information/data, specifying mission objectives, interacting with CPCG agents for ML and fusion/diffusion augmentation and refinement with social knowledge, and decision-making. The UDOP enables the human operator to direct his/her decision-making at the highest level of establishing or updating mission objective, or to expand his/her decision-making to involve details of instantaneous coordination among engagement groups within CPCG architecture, or even to decision-making at the level of individual sensors, weapons, actuators, and platform management. This rich human operator access and interaction with CPCG architecture at different levels suggest that the human operator is *infused* into the CPCG architecture.

(iii) Control Diffusion (CD): The CD agent relates to the planning side of AI. The CD dissects or decomposes high-level mission objectives that originate from the human operator into specialized tasks or actions for different engagement capabilities of CPCG architecture. The planning is an AI's effort that generates a sequence of actions based on observations. The planning agent explores the space of all possible actions to select the optimal sequence of actions that meets the mission goals. Thus the planning process diffuses or fans out the high-level mission objectives to the CPCG constituent systems terminating in the lowest-level control signals for individual sensors, actuators, and platforms. The control theory's *duality* between *observation* and *control* is manifested in CPCG architecture as the duality between information fusion and control diffusion. The CD can be assisted with AI techniques such as statistical relational learning and Markov logic networks [3].

The CPCG architecture utilizes the principle of centralized command with decentralized control. The CPCG architecture has four processing types which are described below with reference to Fig. 3.

Top (Planner) — AI-based: The top "operator node" subsumes the three intelligent agents (IF, OI, and CD). The

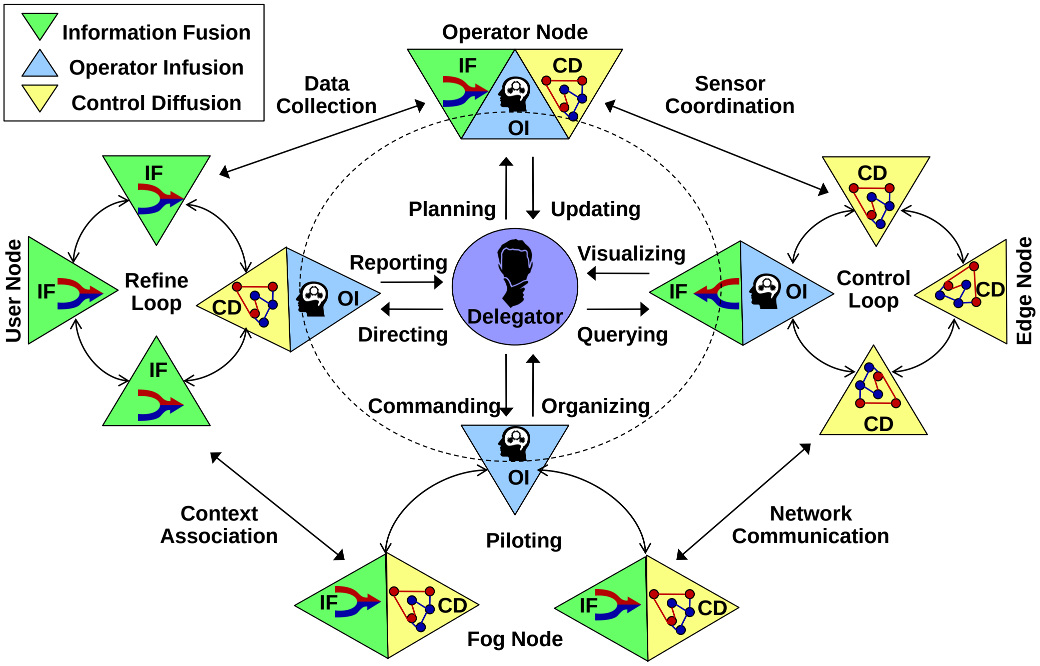

Fig. 3: Cyber-physical command-guided (CPCG) architecture.

commander interacts with these intelligent agents to simultaneously conduct operations and synchronize plans. Since this processing type is most demanding, the IF, OI, and CD leverage a cloud-based approach (Fig. 1) with a UDOP to determine the needs in real-time. For example, if a system is deployed and the scenario has some common patterns and results, the information can be utilized to develop/train semi-supervised AI from labeled (20% from the OI) and unlabeled (80% from the IF) data.

Right (Query-Based) — DFD-based: The commander provides goals to the machine-human agents (IF/OI) from which decisions determine the sensor control loop of what data to collect (CD). The commander queries results from the IF agent that is getting updates from a variety of OI analysis and reports. The OI provides an interface to the human that controls multiple edge units such as IoT devices. If the IoT devices are collecting data from different sources, then the results would be semi-real time as the commander can query the information needed to make decisions, while the IF agent is processing data from multiple updates.

Bottom (Human-Centered) — JDL/DFIG-based: The commander coordinates with other field units and manages other human agents (OI), who then interact with edge node machine agents (IF, CD) sending organized results. In the human-centered approach, there is a direct communication between the users. The OI works with a set of assets (e.g., UAVs) that have on-board processing for IF- and CD-based data collection, which essentially are “edge nodes” as the assets can be either edge devices (e.g., IoT) or edge servers. The results from the machine-based IF-CD are sent to the human via OI interface, which then conveys the salient information to the commander. Since this type is the most distributed approach, it typically takes the longest to execute as the bottleneck is the limited attention of the user-in-the-loop that has to manage the assets as well as provide updates.

Left (Direction-based) — DIG-based: The commander directs human-machine agents (OI/CD) that collect data (IF) to form decisions through a data refinement loop. The commander can provide directions to the OI, which then communicates those to the CD agent. Unlike the query-based (right), here the OI sends commands or directives to the CD agent, which then decides how and when to collect data from edge devices, such as IoT and sensor systems (e.g., cameras) so that the most important information is collected when needed. This type can be represented as a directed graph as the high-level OI directives are imparted as objectives and the CD uses the operating constraints to determine the sensing and actuation needs based on the graph. Since this type is graph-based, it is also machine robust for processing and analysis.

We demonstrate CPCG architecture further through a practical case study in the following.

CASE STUDY: COMMAND-GUIDED SWARM: A practical example of CPCG architecture is command-guided swarm (CGS). A CGS is a multisensor, multi-device (devices can be weapons), multiplatform, single-human-operator system-of-systems (SoS) [18]. The CGS comprises of multiple UxS, where x can be space, air, ground, surface, or undersea, under the mission-oriented tactical coordination of a single human operator. The CGS utilizes advanced AI and human partnering concepts to carry out *fusion of information* originating from the swarm’s multiple sensors and *diffusion of control* out to the swarm’s multifarious platforms, sensors, and devices (weapons). The CGS unifies a collection of semi-autonomous intelligent agents operating in parallel that are neither tightly coupled through a built-in command structure nor completely independent and autonomous. Complex behaviors may emerge from the coordination of semiautonomous agents in CGS, which entails collective intelligence or *swarm intelligence* of the CGS. The swarm intelligence of CGS emerges from the distributed information processing/fusion and engagement of control across multiple AI agents of the CGS. The swarm intelligence leverages active machine learning technologies and human-machine partnership that is enabled by edge computing.

A commander of the CGS may desire to obtain multi-perspective and multimodal observations of an object. Since CGS comprises of different UxS each with different types of edge devices (e.g., visual and infrared sensors), the goal for the swarm is to evolve such that the positions for the viewpoints of multimodal sensors observe an object/target from different perspectives and distances. The multimodal sensors could have an *overlapping* or *orthogonal* viewpoints. The same viewpoints (0 degrees) maximizes data registration whereas orthogonal viewpoints (90 degrees) offer different perspective of the object. The confluence of AI and data (e.g., image) fusion in CGS requires support from models, methods, and control [21]. Models assess theoretical performance of task success based on the range/distance of a device (e.g., camera) to an object. Methods enable empirical performance measurement of multimodal data fusion. Control enables co-ordinated positioning of UxS to obtain multiperspective data. Theoretical models relate object distance to the visual electro-optical and infrared image resolution for object detection and classification. The resolution increases and classification accuracy improves as the distance of the sensor from the object decreases. Theoretical models analyze the probability of success for an object detection and classification task using a single modality (e.g., an electro-optical camera) versus multimodality (e.g., electro-optical and infrared cameras). Once theoretical models indicate benefits of image fusion, AI methods for contextual analysis are then utilized. Context from the scenario includes lighting conditions and position of sensor as a function of range. Context analysis determines which multimodal sensor configuration would yield successful results, that is, whether to use visual, infrared, or visual + infrared camera sensors. Section VI-E demonstrates how multimodal fusion help improve AI precision in CGS.

IV. Alignment of Multi-Level Fusion and AI

Data fusion and AI can be performed at multiple levels utilizing the three hierarchical tiers, viz., edge devices, edge servers, and cloud (Fig. 1). This section discusses different levels of data fusion, characterizes AI into different types and/or stages, and aligns types of AI with different levels of fusion.

A. Multi-Level Fusion

DFIG model defines seven levels of fusion: L0—L6 (Section III-A). Broadly, data fusion can be categorized as *low-level*, *intermediate-level*, or *high-level* depending on the processing stage at which information fusion transpires. In the DFIG model, L0 can be termed as *low-level* data fusion as it combines raw data from multiple processes to produce new raw data. The *intermediate-level* data fusion in the DFIG model encompasses L1 as it provides object assessment based on fusion and extraction of features from raw data. Finally, L2 – L6 (i.e., L2, L3, L4, L5, and L6) in the DFIG model can be construed as *high-level* data fusion as these levels fuse high-level information/features to assess situation and impact, and help refine process, display, and mission management. The low-level information fusion deals with numerical data, such as locations, kinematics, and target attributes, intermediate-level information fusion handles objects/entities, whereas high-level information fusion copes with abstract symbolic information, such as threat, intent, and mission objectives.

Alternatively, levels of fusion can also be characterized as [22]: (i) (sensor, pixel) data fusion, (ii) knowledge/feature fusion, and (iii) decision fusion.

(i) Sensor/Pixel Data Fusion: At the lowest level, raw data produced by sensors and other sources is fused while comprehending the characteristics and relations of the input. This low-level data fusion is also known as *sensor fusion* because the data from different sensors is fused together. In case, sensors are camera sensors (e.g., visual or infrared), sensor fusion is also known as *pixel fusion* because pixels of images obtained from camera sensors are fused together. This sensor/pixel fused data provides an updated representation of data for further processing. Sensor fusion can be implemented as centralized or distributed [23]. In *centralized* fusion architecture, measurements of all sensors are available during the fusion process and a batch method is used to fuse the sensor data. In *distributed* fusion architecture, different sensor measurements are fused with a separate fusion model. Then during the global fusion process, the fusion model information of each sensor is available. The distributed fusion architecture is more scalable with the increasing amount of data as compared to the centralized architecture.

Depending on the sensor configuration, sensor fusion can be classified into three cases: (a) competitive sensor fusion, (b) complementary sensor fusion, and (c) cooperative sensor fusion.

(a) Competitive sensor fusion: In competitive sensor fusion, either data from the sensor of same modality are fused together or the data from the sensors from multiple modalities are first transformed to the same baseline and then fused. Competitive sensor fusion is typically used to reduce noise and uncertainty of the sensor measurements. For example, in case of a surveillance application, multiple competitive camera sensors obtain (homogeneous) images of a target at the same time and fusing those images result in less noisy resultant images that are more suitable for the surveillance or tracking application.

(b) Complementary sensor fusion: In complementary sensor fusion, sensors measure different and distinct parts of the same event and the combination of these (heterogeneous) disparate measurements results in a complete characterization of an event. For example, a complementary set of cameras in a surveillance application can provide an extended picture of the scene which simplifies the subsequent tracking of a target.

(c) Cooperative sensor fusion: In cooperative sensor fusion, a sensor is configured and/or positioned based on the information from other sensors to generate more useful measurements. For cooperative sensor fusion, either sensors can autonomously collaborate to configure each other or some input from human expert can be provided. For example, a tracking task can require adapting the camera angles after observing behavior of a target.

(ii) Knowledge/Feature Fusion: The fused sensor/pixel data provides a basis for feature extraction that develops a model for the underlying data to conceive patterns in the data. The abstraction of fusion components increases with the level of fusion. In the intermediate fusion levels, the data is available in the form of models that represent knowledge from the observed event. The knowledge fusion can be performed at model level or parameter level. In *model fusion*, the knowledge is represented in form of different models, which are fused together. An example of such a model is Gaussian model that provides information about the distribution of data. The mixture of Gaussian distributions produces a Gaussian mixture model (GMM), which describes the distribution of a data set that is more complex than a unimodal Gaussian. The fused model contains more precise knowledge about the overall data distribution. Convolutional neural networks (CNNs) and multiple kernel learning based ensemble methods are other examples of model fusion techniques. In *parameter fusion*, parameters of different models are fused together [22].

(iii) Decision Fusion: At the highest level of fusion, the goal is to improve decision-making and choice of actions. Decisions obtained based on different models can be fused together to obtain better decisions. The decision fusion of multiple models/classifiers can either consist of direct combination of the decisions from the individual models or can select a specific model/classifier for a given input. By observing the impact of a chosen action, the entire fusion process can be adapted for performance improvement and better decision-making.

B. Types and Stages of AI

Since AI research profess to make machines emulate humans, the extent to which an AI system can imitate human capabilities is used as a criterion to define *types of AI*. AI can be classified into four main types based on their functionalities. Type I AI — *reactive machines* belong to the most basic type of AI systems that are purely reactive and do not have the ability to form memories or use past experiences to inform current decisions. These machines can only be utilized for automatically responding to a limited set or combination of inputs. A famous example of a reactive AI machine is IBM's Deep Blue, a supercomputer that beat chess Grandmaster Garry Kasparov in 1997 [24]. Type II AI — *limited memory machines* in addition to having the capability of reactive machines have the ability to learn from historical data to make decisions though this memory is limited and transient. Nearly all existing AI applications (e.g., chatbots, virtual assistants, autonomous vehicles) fall under this AI category. The next two types of AI exist either as a concept or work in progress. Type III AI — *theory of mind* is used to represent a machine (AI agent) that has the ability to form a predictive model of self and others and have the ability to represent and discern the mental states of others, including their emotions, desires, beliefs, and intentions. Theory of mind AI can provide intelligent machines/robots with powerful capabilities, in particular, social intelligence for human-machine interaction [25]. Type IV AI — *self-awareness* is an extension of theory of mind AI and is often regarded as the ultimate objective of all AI research. Self-awareness AI refers to an AI agent that has consciousness and has the ability to form representation of itself and others. Self-aware AI agents know about their internal states and can predict the feelings and actions of others. This type of AI will not only be able to understand and induce emotions in those it interacts with, but also have emotions, needs, beliefs, and likely desires of its own [24]. Although self-aware AI can potentially boost our progress as a civilization tremendously, it can also possibly lead to catastrophe because self-aware AI would have the capability of developing ideas like self-preservation and outmaneuver the human intellect to plot elaborate schemes to take over humanity [24]. Consequently, AI safety has been gaining traction in AI research and non-profit organizations [26].

An alternate system of classification that is more prevalent in AI community is the classification of AI into different *stages*, viz., artificial narrow intelligence, artificial general intelligence, and artificial super intelligence. *Artificial narrow intelligence* (ANI) represents all the existing AI even the most complicated ones including deep learning. ANI refers to those AI systems that can only perform a specific task (e.g., driving, speech recognition) with human-like capabilities. *Artificial general intelligence* (AGI) refers to the capability of AI to learn, perceive, understand and function like humans. AGI will independently build multiple competencies and generalizations across various domains thus massively reducing the time needed for training. AGI will make AI agents just as capable as humans by replicating the multi-functional abilities of humans. *Artificial super intelligence* (ASI) marks the apex of AI research as ASI agents will exceedingly do better at everything than humans because of great memory, processing, analysis, and decision-making. The development of ASI can potentially lead to a scenario referred to as the *singularity* [27].

C. Mapping of AI/ML to Fusion Levels

Table II provides a mapping of AI types to different fusion levels: Type I AI — reactive machines with rules support L0 processing; Type II AI — limited memory utilizes signal processing based methods to provide L1 functions; Type III AI — theory of mind provides representations about the world supporting L2/L3 functions; and Type IV AI — self-awareness supports prediction and interact with L4/L5/L6 analyses. Although in Table II, we have aligned Type III and Type IV AI with data fusion levels L2–L6, this alignment is *idealistic* and imply that these AI types will be able to best perform the capabilities associated with intermediate- and high-level fusion. Currently, all fusion levels are assisted with Type I and Type II AI only and a contemporary (not idealistic) alignment will map Type I and Type II to all fusion levels.

TABLE II: AI aligned with information fusion.

Type of AI	Focus	Objective	Information Fusion Alignment
Type I	Reactive machines	Identify patterns from rules for immediate action	L0 Data assessment
Type II	Limited memory	Estimate response leveraging signal processing	L1 Object assessment
Type III	Theory of mind	Form representations of world and other agents	L2 Situation assessment L3 Impact assessment
Type IV	Self-awareness	Understand self conscious to interact with prediction	L4 Process refinement L5 User refinement L6 Mission refinement

TABLE III: ML aligned with information fusion.

ML Methods	Types of ML						Information
	Symbolic	Probabilistic	Connectionist	Analogistic	Evolutionary	Possibilistic	Fusion
	(Logic)	(Bayes)	(DL)	(SVM)	(GA)	(Fuzzy)	Alignment
Registration, Estimation	✗	✓	✗	✓	✗	✗	L0 Data Assessment
CNN, RNN, LSTM, Estimation	✗	✓	✓	✓	✓	✓	L1 Object Assessment
CNN, Pattern Matching	✓	✓	✓	✗	✗	✓	L2 Situation Assessment
GAN	✗	✓	✓	✗	✓	✗	L3 Impact Assessment
RL, Optimization, Regularization	✗	✓	✓	✗	✗	✓	L4 Process Refinement
Active Learning	✓	✓	✓	✗	✗	✗	L5 User Refinement
RL	✓	✓	✓	✗	✗	✗	L6 Mission Refinement

Since AI is dominated by ML techniques, Table III provides a mapping of ML methods to different fusion levels. It is to be noted that different ML methods can provide different types of AI. For example, deep neural networks (DNNs), either a single DNN or a connection of multiple DNN, can provide different types of AI, ranging from Type I to Type III AI. Table III categorizes ML methods into six types: (i) symbolic, (ii) probabilistic, (iii) connectionist, (iv) analogistic, (v) evolutionary, and (vi) possibilistic.

(i) Symbolic ML: Symbolic ML is based on first-order logic models and truth tables. Symbolic ML has been termed as good old-fashioned AI. It is excellent for implementing machine reasoning but is rigid in the sense that logic statements are either true or false with no possibility of compromise. Failure of logical systems lead to catastrophic failures and thus more sophisticated ML approaches are needed for solving complex problems.

(ii) Probabilistic ML: Probabilistic ML is based on Bayesian statistics, conditional probabilities, and network/graphs of nodes. Probabilistic ML avoids the rigidness of symbolic ML by modeling relationships as conditional probability distributions; however, probabilistic models lack the rich representations and reasoning ability of symbolic ML. The ML for complex systems, such as CGS (discussed in Section III-B as a case study), utilizes a novel hybrid of symbolic and probabilistic ML. This hybrid ML approach, which is referred to as *statistical relational learning* (SRL), combines Bayesian graphical models with first-order logic where logical symbolic representations capture the underlying rich structure of the problem domain, while the probabilistic methods handle the uncertainty in data.

(iii) Connectionist ML: Connectionist ML utilizes computational models inspired by neural architecture of the biological brain. Examples of connectionist ML include artificial neural networks (ANNs) and multilayer ANNs (i.e., deep learning). Table III also list some other ML methods that fall under connectionist ML, such as CNN, RNN, long short-term memory (LSTM), GAN, and reinforcement learning (RL). An RNN is a class of ANNs where connection between nodes constitute a directed graph along a temporal sequence that enables it to exhibit dynamic temporal behavior. RNNs utilize their internal state (memory) to process variable lengths of input sequences, which makes RNNs suitable for tasks such as connected handwriting recognition and speech recognition. LSTM is a type of RNNs that have feedback connections and help overcome the vanishing gradient problem in RNNs. GAN is a machine learning framework based on ANNs that learns to generate new data with the same statistics as the training set. RL is another area of ML that is concerned with determining an optimal set of actions in an environment in order to maximize cumulative reward. RL is typically stated in the form of a Markov decision process (MDP), where dynamic programming provides a solution for the MDP [28]. Deep RL (DRL) combines deep neural networks with reinforcement learning algorithms (e.g., Q-learning) to solve previously unsolvable problems as DRL can learn from raw sensor data and/or images supplied as input. In *active learning*, an ML algorithm can interactively pose queries during the training process, usually in the form of unlabelled data instances to be labelled by a human user. Thus, active learning is an example of human-in-the-loop learning. Connectionist ML often leverages regularization. *Regularization* is a technique which makes slight modifications to the ML algorithm to prevent overfitting to the training data and to make the model more generalizable for different test data sets. Connectionist ML also preprocesses the input data through data registration. *Registration*, often performed for images and thus known as *image registration*, is the process of transforming different images of one scene into the same coordinate system. Registration is required in order to compare or fuse the data obtained from different measurements. For example, images can be taken at different times (multi-temporal registration), by different sensors (multi-modal registration), and/or from different viewpoints [29].

(iv) Analogistic ML: Analogistic ML analyzes data analogies and similarities through distance computations in feature hyperspace. Examples of analogistic ML include support vector machines (SVMs) and nearest neighbors, such as K-nearest neighbor (KNN) algorithm.

(v) Evolutionary ML: Evolutionary ML utilizes computational models inspired by evolutionary competition and survival. Examples of evolutionary ML include genetic algorithms (GAs), genetic programming, and neuroevolution . We note that neuroevolution is similar to genetic programming but the genomes represent ANNs by specifying structure and connection weights.

(vi) Possibilistic ML: Possibilistic ML analyzes ambiguous data using extension of classical logic to represent partial truths. Fuzzy inference systems and possibilistic logic systems are examples of possibilistic ML.

V. Advantages of Data Fusion and AI at the Edge

Data fusion and AI at the edge can provide various advantages in terms of latency, energy, accuracy, security, privacy, cost, scalability, and sustainability as discussed in the following.

Latency: Latency refers to the time spent in the whole AI inference process, including pre-processing, data fusion, model inference, data transmission and post processing [30]. Many edge devices and systems (e.g., surveillance systems, autonomous vehicles, robots) have stringent deadline requirements (in the order of microseconds to milliseconds) and missing those deadlines can result in catastrophes. According to Steve Roddy, the Vice President (VP) of Special Project in Arm's Machine Learning Group [2]: "Applications that people will engage within real-world products such as controlling home devices or providing driver assistance in a car, all of those applications are running on the edge and many will require real-time responses. Any delay from bouncing information to the cloud and back could be a problem." Sri Chandrasekaran, Senior Director of IEEE Standards Association, has also emphasized the importance of speed/latency for AI inference [2]: "We can't overlook the importance of latency. AI at the edge will allow for faster data transfer, which will, in turn, benefit the many industries AI touches, especially industrial IoT and automotive. These industries benefit from AI at the edge because the machines and automobiles must be able to understand many different aspects at once. Sending data to the cloud and back is not only inefficient, but it is also less secure and much slower, ultimately leading to a decrease in productivity and reliability." Research results also verify that data fusion and AI at the edge provides much faster response as compared to sending data to the cloud for fusion and inference [31].

Energy Efficiency: Data fusion and AI at the edge is much more energy-efficient than the data fusion and inference at the cloud because it takes a large amount of power to send data over the air whereas it takes orders of magnitude lesser power to do computations on the device when the data is available on-device. Since many of the edge devices are battery-powered with no energy harvesting system (e.g., solar, thermal), sending data to the cloud for fusion and inference and receiving the results back will expeditiously deplete these devices of the battery power.

Precision: AI precision or accuracy refers to the ratio of the number of input samples that get correct prediction to the total number of input samples [30]. Many edge applications, such as autonomous driving and face authentication, require ultrahigh AI accuracy. Although increase in the number and type of sensors assist in covering large areas, the growing number of sensors have often resulted in an increase in false alarm rates and have compounded the target acquisition process in case of surveillance applications due to the fact that sensors can provide inaccurate, incomplete, or inconsistent data. Data fusion at the edge can help filter the outliers and malicious readings, which results in an improved AI/ML model with better precision than a model that is trained on outliers and malicious data. Furthermore, in many edge applications, AI inference accuracy is also affected by the speed at which an application needs to process the input data. For a video analytics application under a fast feeding rate, some input samples may be skipped due to limited resources of edge devices. In such conditions, data fusion can help improve the accuracy by fusing a few adjacent frames of video and presenting the fused frames to the AI model for inference. The number of frames to be fused will depend on the application and available resources of an edge device. The fused frame will be able to capture the salient information in the video frames and will provide better prediction accuracy instead of skipping the video frames if an edge device could not handle the feed rate.

Security: Security is another advantage that is provided by data fusion and AI at the edge. In case of fusion and AI at the cloud, data needs to travel from edge devices to the cloud and from the cloud to edge devices hundreds of miles over multiple channels including wireless channels and Internet thus exposing the data en route to attackers. The data sent to and received from the cloud can be compromised over the wireless channels, wired channels, intermediate routers, or even the cloud computers itself. Data fusion and AI at the edge devices minimizes the data transfer and thus alleviates the security issues associated with data transfer over multiple channels.

Privacy: Privacy preservation is another advantage of data fusion and AI at the edge. Often the data requiring inference is private and contains sensitive information (e.g., medical records, personal photos, financial reports, target information in defense applications). Sending this private data to the cloud for AI provides no privacy guarantees to the user. Microsoft Research has proposed a homomorphic encryption based solution referred to as CryptoNets [32] that permits a data owner to send their data in encrypted form to the cloud service for inference. Since the cloud does not have access to the encryption key, the user data remain confidential. In CryptoNets, the cloud service is able to provide inference on the encrypted data and return the results to the user in an encrypted form. However, the overhead of homomorphic encryption limits the applicability of this solution on resource-constrained edge devices. Data fusion and AI at the edge enables inference at the edge and thus alleviates the privacy issues associated with sending data to the cloud.

Cost: Data fusion and AI at the edge also provide cost advantages. Due to advancements in semiconductor technology, the cost of system-on-chips (SoCs) is decreasing with increasing capability to perform fusion and AI on these SoCs. These edge SoCs cost much lesser than building the apparatus and infrastructure required to perform fusion and inference in the cloud.

Scalability: The number of edge devices are continuously on the rise approaching 100 billion in near future and producing hundreds of zettabytes of data. If each edge device has to send all the data back to the cloud data center for fusion and AI inference, it will put an enormous pressure on the network bandwidth likely causing the network to collapse (denial of service) as well as will require huge investments on expensive data centers. Data fusion and AI at the edge imparts scalability to an intelligent computing system because majority of the data fusion and AI computations are performed at edge devices and only limited fused and processed data needs to be sent over the network to the cloud.

Sustainability: Data fusion and AI at the edge provide a sustainable solution for emerging smart applications (e.g., autonomous vehicles, smart agriculture, surveillance, swarm intelligence) because increasing advances in semiconductor will continue to make edge devices more powerful to carry out inferences in real-time in a cost-effective manner. Moreover, edge AI will be able to meet AI needs for applications even in communication-denied environments or places where no infrastructure exists for connection to the cloud.

VI. Experimental Results

In this section, we provide experimental results demonstrating latency, energy, and precision advantages of combined data fusion and AI at the edge. The experimental results present latency and energy consumption comparison between two CNN models with and without data fusion. Furthermore, results demonstrate how multimodal fusion help improve the precision of AI for a CGS system.

A. Experimental Setup

We conduct two set of experiments to demonstrate latency, energy, and precision advantages of combined data fusion and AI at the edge. The primary set of experiments use handwritten digit datasets whereas secondary experiments use different types of camera sensors, viz., visual (VI) and medium wavelength infrared (MWIR) outfitted on UAVs.

In our primary set of experiments, we obtain experimental results for MNIST handwritten digit dataset [33]. We clarify that we have chosen this dataset for illustrating the effectiveness of combined data fusion and AI; however, experimental results for other datasets can be obtained similarly. For practical relevance, we note that for social intelligence, often handwritten digits and text need to be analyzed. We have randomly selected 100 handwritten images from the dataset for each digit. We have made 10 sets of each digit image where each set contains 10 handwritten images. We then fuse the images in each set to produce 10 fused images for each digit. Our data fusion technique adds the pixel values at the same location of the ten images in the set. Thus, our primary set of experiments demonstrates competitive pixel data fusion done explicitly and model data fusion (knowledge/feature data fusion) implicitly by the CNN inference.

We consider three use cases for AI and data fusion on multiple hardware platforms suitable for edge computing: (1) data fusion and CNN model execution in Intel Xeon CPU [9] (Xeon_CPU), (2) data fusion and CNN model execution in Nvidia Jetson TX2 graphics processing unit (GPU) (JTX2 [34]) CPU (JTX2_CPU), and (3) data fusion in JTX2 CPU and CNN model execution in JTX2 GPU (JTX2_CPU+JTX2_-GPU). We average the execution time of performing data fusion and CNN inference over ten sets of ten digits. We have run our experiments on Ubuntu 18.04 operating system using CUDA 10.1 as general purpose GPU (GPGPU) framework, and Python 3.6.9 for implementation of data fusion. For implementing AI with data fusion, first the ten images in each set are fused using data fusion in Python and the resulting fused image is then provided as input to the Darknet framework [35] for CNN inference. We use two CNN models: LeNet [36] and AlexNet [37], for our experiments. Regarding the training of CNN models, we utilize the trained weights provided by [38] for LeNet whereas we train the weights of AlexNet ourselves using MNIST training dataset comprising of 60,000 images. To smooth out any inconsistencies in latency due to operating system overhead and variations in environmental parameters, we average the execution time results over ten independent measurements.

B. Average Latency for Combined Data Fusion and AI at the Edge

Table IV shows the speedup of AI (LeNet and AlexNet CNN models) with data fusion over AI without data fusion. Results verify that AI with data fusion provides significant speedups over AI without data fusion. For example, AI with data fusion on Xeon CPU results in a speedup of 2.6× and 9.7× for LeNet and AlexNet, respectively. For JTX2 CPU, AI with data fusion provides a speedup of 3.8× and 9.8× for LeNet and AlexNet, respectively, as compared to AI without data fusion. The data fusion in JTX2 CPU and CNN model execution in JTX2 GPU leads to a speedup of 9.6× and 9.3× for LeNet and AlexNet, respectively.

Since GPUs serve as a decent platform for acceleration of deep learning models, offloading the CNN inference tasks to the JTX2 GPU engenders a speedup of 2.8× to 2.9× for AlexNet as compared to only using the JTX2 CPU. However, experimental results indicate the LeNet execution on GPUs results in lower performance as compared to LeNet execution

TABLE IV: Speedup of AI with data fusion as compared to AI without data fusion.

Edge Computing Platform	**CNN Model**	
	LeNet	**AlexNet**
Xeon_CPU	2.55X	9.72X
JTX2_CPU	3.82X	9.76X
JTX2_CPU + JTX2_GPU	9.57X	9.29X

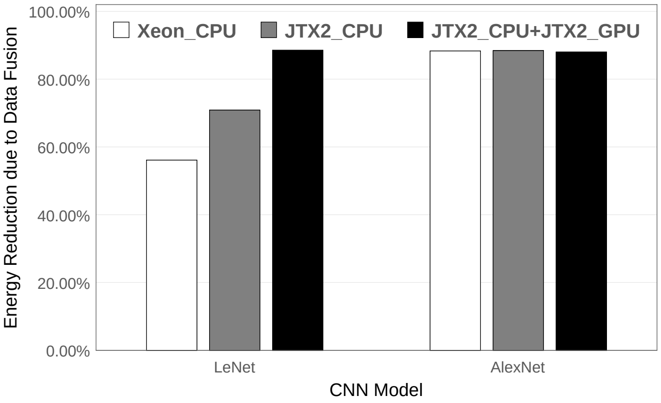

Fig. 4: Energy reduction imparted by data fusion and AI as compared to AI without data fusion.

on CPU. This is due to the small size of LeNet model (i.e., LeNet model consists of only ∼60 thousand parameters), which can be efficiently run on CPU without offloading. For LeNet inference on GPU, the data transfer between CPU, memory, and GPU incurs non-negligible latency overhead that causes higher overall execution time of LeNet inference on GPU as compared to CPU. Conversely, since AlexNet is a large CNN model with ∼60 million parameters, AlexNet inference on GPU provides better performance than CPU because the computation time for large data dominates the data transfer time.

C. Energy Consumption for Combined Data Fusion and AI at the Edge

Data fusion also imparts energy benefits to AI. Fig. 4 depicts the energy reduction furnished by data fusion and AI as compared to AI without data fusion across three edge computing platforms and the two CNN models. Results verify that the data fusion leads to huge energy savings for AI. For Xeon CPU and JTX2 CPU, the energy reductions furnished by data fusion and AI over AI without data fusion are 56.11%–88.32% and 70.88%–88.42%, respectively. For the data fusion and AI execution on GPU along with the CPU in the JTX2 platform, energy reduction is 88.00%–88.52% as compared to AI without data fusion. These energy savings are attained because with data fusion, same amount of input data can be processed much quicker by the AI models as compared to the case without data fusion (Section VI-B). This lower latency of AI with data fusion also translates to lower energy consumption as compared to AI without data fusion.

D. Accuracy for Combined Data Fusion and AI at the Edge

AI with data fusion can obviously result in improved performance as illustrated in Section VI-B because inference is performed on reduced data size. However, there may be a concern regarding accuracy of combined data fusion and AI because the fused data is different from the original raw data. Hence, we measure the impact of data fusion on accuracy using average softmax probability as a metric by which the CNN model predicts a class. For accuracy evaluations, we use Darknet framework [35] with LeNet CNN model. The experimental setup is the same as described in Section VI-A. We generate one image with fusing 10 images by averaging the pixel values, resulting in data size reduction by 90%. For inference with data fusion, we use 100 fused input images while we use 1000 original images in the case of inference without data fusion.

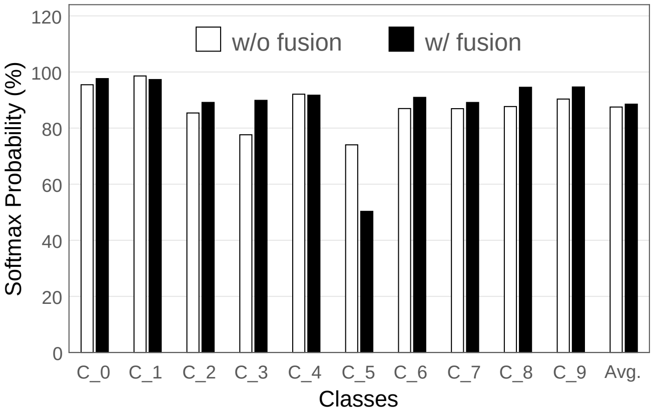

Fig. 5: Accuracy of AI with data fusion as compared to AI without data fusion.

Fig. 5 shows data fusion impact on softmax probability for CNN inference. Our CNN model has ten output classes C_i where i, $\;i \in 0, 1, 2, 3, \ldots, 9$. C_$i$ denotes class i that corresponds to the output for digit i (e.g., C_0 corresponds to digit 0 and C_9 corresponds to digit 9). Results indicate that data fusion does not adversely affect the softmax probability though there is a fluctuation in softmax probability depending on the digit class. Results show that data fusion imparts higher accuracy to AI on average. For example, data fusion results in CNN inference accuracy of 88.53% as compared to the accuracy of 87.51% without data fusion for all digit classes. Results depict that for some classes, data fusion may lead to lower accuracy for AI than without using data fusion. For example, for C_5 in our experiments, softmax probability of CNN inference actually decreases, meaning that the data fusion may lead to inaccurate prediction (or classification). We observe that the softmax probability for C_5 with data fusion decreases by 32% as compared to the softmax probability without data fusion. The reason being that sometime for image classification, data fusion may cause the image to become blurry, which may cause the CNN model to misclassify the input. However, post-processing of fused images and/or applying different data fusion techniques can help improve the accuracy. Results also reveal that for all classes other than C_5, softmax probability with data fusion is either very close to or higher than the softmax probability without data fusion. For example, data fusion provides an improvement of 4.25% for softmax probability as compared

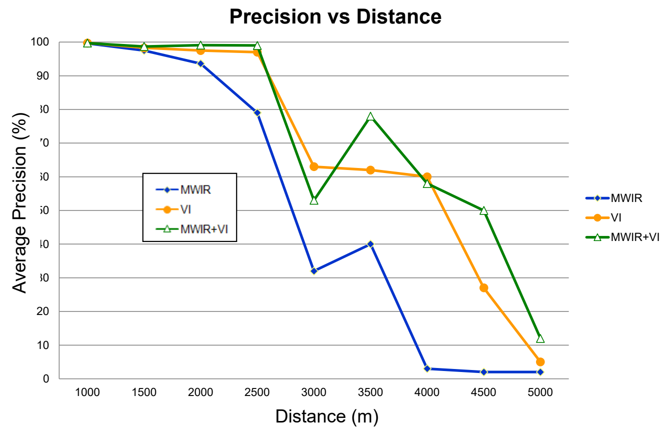

Fig. 6: Precision vs. distance for multimodal fusion and AI.

to the softmax probability without data fusion for all digit classes except C_5. Results show that data fusion imparts maximum improvement in softmax probability for C_3, where the attained improvement is 15.81%. Hence, our results verify that data fusion significantly improves the performance of AI while also improving or maintaining the accuracy of AI in most cases.

E. Multimodal Fusion and AI at the Edge

This experimental result focuses on the CGS case study where a commander of the CGS may desire to obtain multiperspective and multimodal observations of an object using different types of camera sensors, viz., VI and MWIR, outfitted on UAVs. This experiment demonstrates multimodal competitive data fusion and multimodal complementary data fusion. Fig. 6 illustrates how and when multimodal image fusion benefits deep learning based inference from images [18]. Fig. 6 indicates that multimodal image fusion (VI + MWIR) benefits deep learning precision when sensor-to-object distance is greater than 3 km whereas using only visual imagery produces acceptable precision when the sensor-to-object distance is less than 3 km. We clarify that for CGS multimodal sensing, control is asserted through user commands. The control asserted through high-level user commands permeates to different components of CGS through control diffusion, which results in routing of UAVs to positions through which overlapping or orthogonal viewpoints can be obtained as required by the deep learning methods for improving classification precision.

VII. Conclusions

In this article, we have discussed the emerging discipline of data fusion and artificial intelligence (AI) at the edge. As the price of computing continues to fall, edge fusion and intelligence will continue to proliferate. In the foreseeable future, the cloud and edge AI will continue to coexist where the cloud will be mostly used for training of AI whereas more and more inference will be performed at the edge. In this article, we have proposed a hierarchical framework for data fusion and AI at the edge. The article provides a comparative discussion of contemporary data fusion and AI models and architectures with special emphasis on data fusion information group (DFIG) model and cyber-physical command-guided (CPCG) architecture. The article also presents a case study of command-guided swarm (CGS) as a practical application of CPCG architecture. The article aligns the AI techniques to different fusion levels. Finally, the article demonstrates the advantages of AI and data fusion at the edge including latency, energy efficiency, precision, security, privacy, cost, scalability, and sustainability. Experimental results have revealed that combining AI with data fusion can impart a speedup of 9.8× over AI without data fusion. Additionally, data fusion leads to energy savings of up to 88.5% for AI as compared to the AI without data fusion. Furthermore, results have demonstrated that data fusion either maintains or improves the accuracy of AI in most cases. For our experiments, data fusion imparts a maximum improvement of 15.8% in accuracy to AI. Furthermore, experimental results for a CGS case study demonstrate the advantage of multimodal data fusion on precision of AI.

Acknowledgments

This research was supported in part by the Air Force Research Laboratory (AFRL) Information Directorate (RI), through the Air Force Office of Scientific Research (AFOSR) Summer Faculty Fellowship Program®, Contract Numbers FA8750-15-3-6003 and FA9550-15-0001. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the AFRL and AFOSR.

References

[1] A. Munir, P. Kansakar, and S. U. Khan, "IFCIoT: Integrated Fog Cloud IoT: A novel architectural paradigm for the future Internet of Things." *IEEE Consumer Electronics Magazine*, vol. 6, no. 3, pp. 74–82, 2017.

[2] B. Reese, "AI at the Edge: A GigaOm Research Byte," https://gigaom.com/report/ai-at-the-edge-a-gigaom-research-byte/, January 2019.

[3] E. Blasch, R. Cruise, S. Natarajan, A. K. Raz, and T. Kelly, "Control Diffusion of Information Collection for Situation Understanding Using Boosting MLNs," in *Proc. of 21st International Conference on Information Fusion (FUSION)*, Cambridge, UK, July 2018.

[4] B. Gu, J. Kong, A. Munir, and Y. G. Kim, "A Framework for Distributed Deep Neural Network Training with Heterogeneous Computing Platforms," in *Proc. of IEEE International Conference on Parallel and Distributed Systems (ICPADS)*, Tianjin, China, December 2019.

[5] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications," 2017.

[6] K. Lee, J. Kong, and A. Munir, "HW/SW Co-Design of Cost-Efficient CNN Inference for Cognitive IoT," in *Proc. of the International Conference on Intelligent Computing in Data Sciences (ICDS)*. Fez, Morocco: IEEE, October 2020.

[7] Samsung, "What is the NPU in Galaxy and what does it do?" [Online]. Available: https://www.samsung.com/global/galaxy/what-is/npu/

[8] N. Chen, Y. Chen, E. Blasch, H. Ling, Y. You, and X. Ye, "Enabling Smart Urban Surveillance at The Edge," in *IEEE International Conference on Smart Cloud (SmartCloud)*, New York, NY, November 2017, pp. 109–119.

[9] Intel, "Xeon Processor E3-1230 v5." [Online]. Available: https://ark.intel.com/content/www/kr/ko/ark/products/88182/intel-xeon-processor-e3-1230-v5-8m-cache-3-40-ghz.html

[10] N. P. Jouppi, C. Young, N. Patil, D. Patterson, G. Agrawal, R. Bajwa, S. Bates, S. Bhatia, N. Boden, A. Borchers, and et al., "In-Datacenter Performance Analysis of a Tensor Processing Unit," *SIGARCH Comput. Archit. News*, vol. 45, no. 2, pp. 1–12, 2017.

[11] J. Fowers, K. Ovtcharov, M. Papamichael, T. Massengill, M. Liu, D. Lo, S. Alkalay, M. Haselman, L. Adams, M. Ghandi, and et al., "A Configurable Cloud-Scale DNN Processor for Real-Time AI," in *Proceedings of the 45th Annual International Symposium on Computer Architecture*, 2018, pp. 1–14.

[12] B. Dasarathy, "Sensor Fusion Potential Exploitation-Innovative Architectures and Illustrative Applications," *Proceedings of the IEEE*, vol. 85, no. 1, pp. 24–38, January 1997.

[13] D. L. Hall, C.-Y. Chong, J. Llinas, and M. Liggins II, *Distributed Data Fusion for Network-Centric Operations*. CRC Press, 2012.

[14] Z. Chair and P. Varshney, "Optimal Data Fusion in Multiple Sensor Detection Systems," *IEEE Transactions on Aerospace and Electronic Systems*, vol. AES-22, no. 1, pp. 98–101, January 1986.

[15] C. Snoek, M. Worring, and A. W. M. Smeulders, "Early Versus Late Fusion in Semantic Video Analysis," in *Proc. of the 13th ACM International Conference on Multimedia*, Singapore, November 2005, pp. 399–402.

[16] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative Adversarial Nets," in *Proc. of Advances in Neural Information Processing Systems (NIPS)*, Montréal, Canada, December 2014, pp. 2672–2680.

[17] E. P. Blasch, D. A. Lambert, P. Valin, M. M. Kokar, J. Llinas, S. Das, C. Chong, and E. Shahbazian, "High Level Information Fusion (HLIF): Survey of Models, Issues, and Grand Challenges," *IEEE Aerospace and Electronic Systems Magazine*, vol. 27, no. 9, pp. 4–20, September 2012.

[18] R. Cruise, E. Blasch, S. Natarajan, and A. Raz, "Cyber-physical Command Guided Swarm," *Defense Systems Information Analysis Center (DSIAC) Journal*, vol. 5, no. 2, pp. 24–30, 2018.

[19] M. Bedworth and J. O'Brien, "The Omnibus model: a new model of data fusion?" *IEEE Aerospace and Electronic Systems Magazine*, vol. 15, no. 4, pp. 30–36, April 2000.

[20] M. Liggins, C.-Y. Chong, I. Kadar, M. Alford, V. Vannicola, and S. Thomopoulos, "Distributed Fusion Architectures and Algorithms for Target Tracking," *Proceedings of the IEEE*, vol. 85, no. 1, pp. 95–107, January 1997.

[21] E. Blasch, R. Cruise, A. Aved, U. Majumder, and T. Rovito, "Methods of AI for Multimodal Sensing and Action for Complex Situations," *AI Magazine*, vol. 40, no. 4, pp. 50–65, 2019.

[22] S. Beddar-Wiesing and M. Bieshaar, "Multi-Sensor Data and Knowledge Fusion: A Proposal for a Terminology Definition," *arXiv*, January 2020. [Online]. Available: https://arxiv.org/abs/2001.04171

[23] A. Munir, A. Gordon-Ross, and S. Ranka, "Multi-core Embedded Wireless Sensor Networks: Architecture and Applications," *IEEE Transactions on Parallel and Distributed Systems (TPDS)*, vol. 25, no. 6, pp. 1553–1562, June 2014.

[24] N. Joshi, "7 Types Of Artificial Intelligence," https://www.forbes.com/sites/cognitiveworld/2019/06/19/7-types-of-artificial-intelligence/#27e3a666233e, June 2019.

[25] A. F. T. Winfield, "Experiments in Artificial Theory of Mind: From Safety to Story-Telling," *Frontiers in Robotics and AI*, vol. 5, June 2018.

[26] FLI. (2020, August) Future of Life Institute. [Online]. Available: https://futureoflife.org/

[27] J. Strickland, "What's the Technological Singularity?" https://electronics.howstuffworks.com/gadgets/high-tech-gadgets/technological-singularity.htm, August 2020.

[28] A. Munir and A. Gordon-Ross, "An MDP-based Dynamic Optimization Methodology for Wireless Sensor Networks," *IEEE Trans. on Parallel and Distributed Systems (TPDS)*, vol. 23, no. 4, pp. 616—-625, April 2012.

[29] E. Kamoun and J. Joslove, "Image Registration: From SIFT to Deep Learning," https://medium.com/sicara/image-registration-sift-deep-learning-3c794d794b7a, July 2019.

[30] Z. Zhou, X. Chen, E. Li, L. Zeng, K. Luo, and J. Zhang, "Edge Intelligence: Paving the Last Mile of Artificial Intelligence with Edge Computing," *Proceedings of the IEEE*, vol. 107, no. 8, pp. 1738–1762, August 2019.

[31] R. M. Shukla and A. Munir, "An Efficient Computation Offloading Architecture for the Internet of Things (IoT) Devices," in *IEEE Consumer Electronics and Networking Conference (CCNC)*, Las Vegas, NV, January 2017.

[32] N. Dowlin, R. Gilad-Bachrach, K. Laine, K. Lauter, M. Naehrig, and J. Wernsing, "CryptoNets: Applying Neural Networks to Encrypted Data with High Throughput and Accuracy," https://www.microsoft.com/en-us/research/publication/cryptonets-applying-neural-networks-to-encrypted-data-with-high-throughput-and-accuracy/, February 2016.

[33] Y. LeCun, C. Cortes, and C. J. C. Burges, "The MNIST DATABASE of Handwritten Digits." [Online]. Available: http://yann.lecun.com/exdb/mnist

[34] Nvidia, "Jetson TX2." [Online]. Available: https://developer.nvidia.com/embedded/buy/jetson-tx2

[35] J. Redmon, "Darknet: Open Source Neural Networks in C," http://pjreddie.com/darknet/, 2013–2016.

[36] Y. Lecun and L. Bottou and Y. Bengio and P. Haffner, "Gradient-based Learning Applied to Document Recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov 1998.

[37] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Advances in Neural Information Processing Systems 25*, 2012, pp. 1097–1105.

[38] T. Ashitani, "darknet_mnist," https://github.com/ashitani/darknet_mnist, 2020.