# PFI-RP: on-device AI - TODS for the Automotive industry.

## Proposal submission on FastLane.nsf.gov [DEADLINE: January 12, 2022.]

- □ Project Summary [One (1) page max].
- □ FastLane documentation
  - □ Collaborators and other affiliations
  - □ Current and pending support
  - □ Bio sketch
  - □ Budget
  - □ Data management
  - □ Equipment and facilities
  - □ Sub-award documentation

### Project Description. -Fifteen (15) pages max-

- □ Executive Summary (no more than one page)
- □ From NSF Basic Research to Addressing a Market Opportunity (suggested length: 4 -5 pages)
- □ Technical Challenges and Applied Research Plan (suggested length: 5-7 pages)
- □ Achieving Societal Impact through the Realization of Commercial Potential (2-4 pages)
- □ Project Team (suggested length: 1-2 pages)
- □ Partnerships (suggested length: 1-2 pages)
- □ Training Future Leaders in Innovation and Entrepreneurship (suggested length: 1-2 pages)
- □ Broadening Participation (suggested length: up to 1 page)

## Others

- □ Letter of support from (Maxim Integrated).
- □ Letter of support from (UT entrepreneurship).
- □ Letter of support from (Restaurant).

## Project summary -1 page

# 1 Overview

This PFI-RP project will develop EdgeView, an edge-based computer vision technology that provides an intuitive way for the hospitality industry to understand personnel performance and customer analytics, through an inexpensive and seamless integration of intelligent IoT camera. The hospitality industry experienced fierce pressure to add digital technologies to their physical locations even before the pandemic. The pandemic only made these needs more painfully apparent for restaurants. By exposing the weaknesses and challenges of small and medium restaurants in the US, however, the pandemic has made an excellent case for intelligent Internet of Things (IoT) cameras as an agent to merge physical locations with advanced business intelligence tools. The NSF lineage *Award # 2053272 Collaborative Research: Enabling Intelligent Cameras in Internet-of-Things via a Holistic Platform, Algorithm, and Hardware Co-design*, is the foundation for this project. Building this complex system requires strong collaboration with AI Pow, a Texas company with the edge-AI experience needed to enable this partnership to find a commercialization path for our advances in ML and intelligent IoT cameras by making them accessible for the restaurant industry. This feasibility project's successful outcome will significantly help restaurant managers have valuable employee and customer geospatial analytics to enhance the service in their locations by efficiently and intuitively guiding their workforce.

# 2 Intellectual Merit

State-of-the-art (SOTA) AI-based video processing approaches seek higher accuracy from exponentially larger networks. There is a vast and increasing gap between powerful video intelligence algorithms' prohibitive complexity, and the often constrained resources in IoT devices where those algorithms need to be deployed. Different applications and hardware platforms currently require diverse models that favor different compression schemes, and each application calls for its unique efficiency-accuracy trade-off. Moreover, an efficient end-to-end pipeline, with model compression just being one of its many possible steps, is needed to achieve the edge's practical, efficient video understanding.

The proposed research addresses those research gaps in a synergistic framework: (1) an Adaptive Auto-Compression Engine (AACE) will be designed and presented. AACE can automatically search for the best combination strategy of model compression means, given the specific AI models and the target device and efficiency, in a way that is easy to use and comprehend for end users. (2) the compression will be extended from a single task to multiple tasks to ensure a holistic efficient vision system that possesses multiple functionalities. (3) the partnership will develop two learning-based, low-cost, data-level filtering algorithms (temporal frame skipping, spatial patch dropping) to exploit the crucial data-level sparsity in video and to quickly concentrate the processing power onto temporal-spatially important sub-regions.

# 3 Broader impacts

This PFI-RP project will transform the business analytics market in the hospitality industry, making it far more accessible for small and medium business (SMB) owners to adopt such advanced technologies. The proposed project's successful outcome will enable intelligent IoT cameras that are straightforward to use by the partnering restaurant managers and easy to install inside their restaurants. AI Pow's ongoing partnership with Maxim Integrated will be critical for EdgeView's hardware development. This reduction in complexity and deployment cost will allow the target market segment of this proposal, Mexican and Asian restaurants in Texas, to adopt and benefit from advanced ML tools and embrace the data revolution using AIoT in innovative ways. Therefore, EdgeView can effectively increase the economic competitiveness of minority business owners in the hospitality industry, one of the most valuable in the US.

Similarly, this PFI project will enhance partnerships between academia and industry in the US. It's important to say that while this PFI joint work will apply fundamental on-device AI research to democratize the use of artificial intelligence in Texas and the US, this PFI-RP will extend beyond only developing a product but also implementing synergistic activities to provide a valuable educational experience to the participating students at UT-Austin. The AI Pow team, along with the entrepreneurial office Launchpad at UT-Austin, has the expertise to reach the entrepreneurial education objectives of this PFI proposal towards creating human capital at UT capable of proposing new commercial products from fundamental research funded by the NSF.

# 1 Executive Summary

## 1.1 Societal Need and the Customer

Our target customers are small and medium-sized businesses in the hospitality industry that will benefit from Machine Learning at the edge technology to better understand their employees and customers, reduce waste, provide a better workplace environment, and increase customer satisfaction. In addition, EdgeView will help restaurant owners modernize their physical locations and help them quickly adopt advanced business analytics technologies, *a $13.6 billion market in the US according to IBISWorld* [1], that will allow them to compete and thrive in an AI-first economy.

## 1.2 The Value Proposition

EdgeView is an edge-AI computer vision technology that provides an intuitive way for the hospitality industry to understand personnel performance and customer analytics through an inexpensive and seamless integration of intelligent IoT cameras. Proper employee and customer geospatial analytics aim to enhance service by guiding restaurant's workforce efficiently and intuitively.

## 1.3 The Innovation

Our lineage research enabled Intelligent Cameras in the Internet-of-Things via a Holistic Platform, Algorithm, and Hardware Co-design. This PFI project proposes a practical application for the hospitality industry, which sets two main challenges: intelligent cameras must be straightforward to use by managers and easy to install inside restaurants. Our past results demonstrated that state-of-the-art DNNs could be very efficient, making them practical to deploy in resource-constrained IoT platforms. For this feasibility project, we propose applying the lineage results' multiple capabilities to develop a prototype using the hardware provided by Maxim Integrated, the MAX7800 platform, to detect employees and their interaction with their environment and customers. This ambitious project establishes essential collaboration to produce a marketable technology with enterprise IoT hardware developers and entrepreneurial mentors.

## 1.4 The Partnership

Our research group at the University of Texas at Austin is partnering with AI Pow LLC, a Texas company dedicated to developing industrial solutions for the AI-first economy. This partnership has opened an opportunity for our research group to deploy our state-of-the-art lineage technology in one restaurant location in Texas to collect real data and feedback from a place currently trying to modernize its operations and compete in the digital-first economy. AI Pow has also helped us include in this proposal partners such as Maxim Integrated, which will be vital in co-designing software and an edge-AI embedded system based on the MAX7800 chip to meet the strict target market demands presented throughout this proposal.

## 1.5 Training and Leadership Development in Innovation and Entrepreneurship

UT-Austin's on-campus incubator LaunchPad at UT is the key supporter that will help us test our business hypothesis and develop the entrepreneurial talent in our research group. Notably, they have already helped us find mentors and providing the appropriate education material to set the correct metrics to help us achieve the commercialization of our technology. This partnership entrepreneurial guidance also includes workshops, network events, and related business courses.

## 2 From NSF Basic Research to Addressing a Market Opportunity

As pictured in Fig. 1, this project aims to use our results in on-device AI for intelligent IoT cameras to solve pressing problems for the hospitality industry. Restaurants, in particular, are facing a unique digital transformation driven by the COVID-19 pandemic. As a result, restaurants now have the challenge of adopting digital technologies in their physical stores to understand their customers and better manage their workforce. This project strives to understand customers, employees, and their physical environments and provide clear, actionable items for business owners to serve their customers more efficiently while providing a safe work environment.

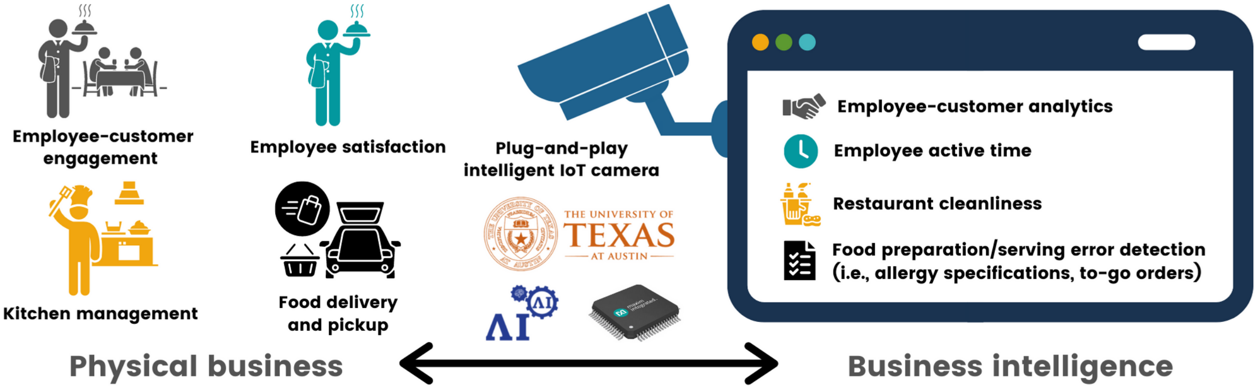

Employee-customer engagement
Employee satisfaction
Plug-and-play intelligent IoT camera
Kitchen management
Food delivery and pickup
THE UNIVERSITY OF TEXAS AT AUSTIN
AI
Employee-customer analytics
Employee active time
Restaurant cleanliness
Food preparation/serving error detection (i.e., allergy specifications, to-go orders)
Physical business
Business intelligence

Figure 1: Representation of intelligent IoT cameras to help the hospitality industry benefit from on-device AI: EdgeView will enable restaurants to understand their customers and employees with smart IoT cameras that are simple to integrate into their physical locations. With the proper hardware, EdgeView will be the digital window to advanced business intelligence for better restaurant management to provide a superior customer experience, better work environment for their employees, a clean space, and safe food.

### 2.1 NSF Lineage

Award # 2053272 Collaborative Research: Enabling Intelligent Cameras in Internet-of-Things via a Holistic Platform, Algorithm, and Hardware Co-design.

### 2.2 Intellectual Merit of the Proposed Product and Broader Impact

**Project Overview:** There has been a tremendous demand for bringing Deep Neural Network (DNN) powered functionality into the Internet of Thing (IoT) devices to enable ubiquitous intelligent "IoT cameras." However, state-of-the-art DNNs have a prohibitive energy cost, making them impractical to be deployed in resource-constrained IoT platforms. This project will develop a novel energy-efficient DNN framework via a systematic integration of platform, hardware, and algorithm co-design innovations. Despite a growing interest in energy-efficient DNNs, existing techniques lack a systematic optimization across the full stack of design abstraction, from systems through algorithms to hardware implementation. The proposed research advocates an innovative, holistic effort towards energy-efficient and adaptive DNN-powered "IoT cameras" by jointly optimizing the platform-, hardware-, and algorithm-level co-design efforts. On the system level, we will address how to automatically generate and adapt DNN models and implementation to meet various "IoT devices" application-specific performance needs and device-specific resource constraints. On the hardware level, we will leverage the observed high sparsity in DNN activations for energy-efficient hardware implementations of both DNN training and inference by using low-cost zero predictors and hence bypass unnecessary computations. On the algorithm level, we will develop innovative factorized sparsity regularization in DNN training and efficient, controllable adaptive inference mechanisms, fully complementing and closely integrating with our hardware innovations.

**Intellectual Merit:** The proposed research highlights a unique combination of machine learning algorithm and hardware design expertise and advocates an innovative, holistic effort towards energy-efficient and adaptive DNN-powered “IoT cameras" by not only combining but jointly optimizing the platform-, hardware-, and algorithm-level co-design efforts. The research outcomes will advance the scientific domain of each level from the system, algorithm to hardware, and a holistic, systematic cross-level methodology for designing energy-efficient intelligent systems.

On the system level, the PIs will address the question of “*how to automatically generate and adapt DNN models and implementation to meet both hospitality-applications performance needs and MAX7800 platform and other device-specific resource constraints?*” This project proposes a morphism-based neural architecture search algorithm that explicitly incorporates real-measured performance into the main objective to efficiently adapt DNN models to meet best the platforms’ resource constraints and the applications’ required performance. Furthermore, the PIs will further develop a technique to automatically generate the optimal layer-wise dataflow (i.e., the algorithm to platform scheduling strategy) when deploying a DNN into a target platform to minimize the energy cost.

On the hardware level, the PIs propose to leverage the observed high sparsity (up to 80%) in DNN activations for energy-efficient hardware implementations of both DNN training and inference by using low-cost zero predictors. As a result, the energy consumption associated with high-cost computations, memory accesses, and data movements will be greatly reduced without degrading the inference accuracy.

On the algorithm level, the PIs will develop innovative regularizations in DNN training, complementing and closely integrated with our hardware innovations. Further, this project proposes an energy-aware dual dynamic inference mechanism for DNN inference to achieve further extra “on-the-fly” energy savings under either “soft” (input-dependent) or “hard” (resource-dependent) constraints adaptively adjusting the model complexity to avoid unnecessary computations.

**Results currently achieved from the project:** In this project, PI Wang’s team has been responsible for the "algorithm-level" effort. In year 1, PI Wang and his collaborators successfully developed the dual dynamic inference (DDI) algorithm [2] for energy-efficient deep network inference. DDI incorporates two input-adaptive inference schemes: i) input-dependent dynamic inference: the model will execute fewer computations for the inference of simpler inputs, and more computations when the inputs are harder; and ii) resource-dependent dynamic inference: the model has to complete its inference and output a good prediction within some (pre-defined or time-varying) certain energy limit, for every sample. The two different mechanisms represent the complementary "soft" and "hard" constraints to save energy in practice, which had not been considered together before. We conduct extensive image classification experiments on CIFAR 10 and ImageNet benchmarks, demonstrating the superior accuracy-resource trade-off and the flexibility of DDI, over existing dynamic inference methods.

In year 2, PI Wang’s team took a deep dive into a special pruning strategy called the lottery ticket hypothesis [3], i.e., deep models contain extremely sparse matching subnetworks capable of training in isolation to full accuracy and transferring to other tasks. In NLP and computer vision, enormous pre-trained models have become the standard starting point for training on a range of downstream tasks, and bridge those models with resource-constrained platforms is of practical significance for powerful edge AI. PI Wang’s pioneering work revealed that such trainable, transferrable subnetworks exist in various gigantic pre-trained models. Taking BERT as one example, for a range of downstream tasks, we indeed find matching subnetworks at 40% to 90% sparsity. Moreover, we find these subnetworks at (pre-trained) initialization, a deviation from prior NLP research where they emerge only after some amount of training [4]. As another example found in computer vision, from all pre-trained weights obtained by ImageNet classification, simCLR and MoCo, we are also consistently able to locate matching subnetworks at 59.04% to 96.48% sparsity that transfer to multiple downstream tasks, whose performance also see no degradation compared to using full pre-trained weights [5]. In year 3 of this project, PI Wang's team investigates hardware-aware pruning and neural architecture search and extends their efforts beyond academic benchmarks to realistic perception tasks on the edge.

**Broader Impacts:** The lineage project advanced the scientific domain of each level, from system and algorithm to hardware, and a holistic, systematic cross-level methodology for designing energy-efficient intelligent systems. Progress on this project has enabled ubiquitous DNN-powered intelligent functions in a significantly increased number of resource-constrained daily-life devices across numerous camera-based IoT applications such as person and object identification, surveillance and security, and action recognition. As camera-based IoT devices penetrate the hospitality industry, enabling DNN-powered intelligence to be pervasive in these devices, the proposed product, EdgeView, will have a tremendous impact on the US economy. We also want to highlight that EdgeView will be implemented in a minority market segment defined as Mexican and Asian restaurants (largest market segment in Texas) in which the PI and Co-PI will develop EdgeView to consider language and cultural needs for our partner B2J-suancai-fish. The goal is to make smart IoT cameras attractive to business owners of such an important market segment.

## 2.3 Market Analysis

Business analytics tools for the hospitality industry are mainly composed of costly tailor-made and complex software systems, leaving an untapped market that demands more intuitive and accessible solutions. According to IBISWorld [1], the business analytics industry will generate $115.2 Billion US dollars in revenue in the US this year. Remarkably, the SMB in the hospitality industry as a whole can benefit highly from advances in hardware and software solutions as it represents more than $4.61 Billion of this market, as shown in Fig. 2, with a reported 13.5% average annual growth rate (AAGR) from 2016 to 2021.

Notably, the retail industry relies heavily on business analytics tools for enterprise resource planning, customer relationship management, and performance management. From previous collaborations with AI Pow, we noticed that business analytics tools allow retailers to lower waste, maintenance costs, reduce employee downtime, and improve customer satisfaction.

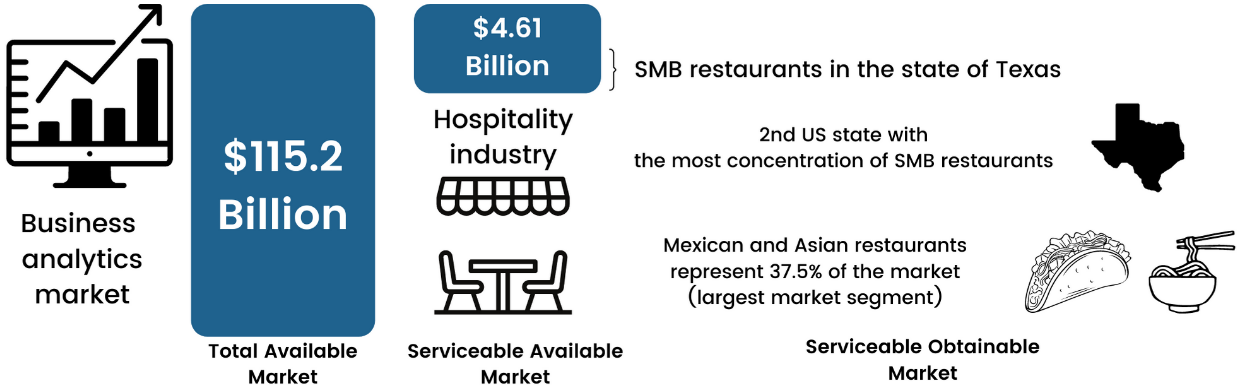

Business analytics market
$115.2 Billion
Total Available Market
$4.61 Billion
Hospitality industry
Serviceable Available Market
SMB restaurants in the state of Texas
2nd US state with the most concentration of SMB restaurants
Mexican and Asian restaurants represent 37.5% of the market (largest market segment)
Serviceable Obtainable Market

Figure 2: TAM, SAM and SOM of business analytics market and hospitality industry segment breakdown.

Our partner B2J-suancai-fish is a single-location and family-operated restaurant that provide food services to patrons who order and are served while seated (i.e., waiter and waitress service) and pay after eating in the city of Houston, TX. Our initial hypothesis is that the technology developed in this project will help SMBs accelerate their transition to a physical-digital restaurant merge. It is also important to mention that our partner is an example in the most dominant restaurants in Texas, which are Mexican and Asian, which represent 37.5% of the total market.

## 2.4 Competitive Technologies

Commercial camera technology for the hospitality industry is dominantly cloud-based and is challenged by two main problems that intelligent edge IoT Cameras can solve. First, since our system runs ML models, installation becomes easier as the setup process is just plug-and-play. A stand-alone functionality has an enormous advantage for business owners that lack the network expertise and data science background to set up complex cloud-based analytics. Second, the ML models produced in our NSF lineage and the embedded systems developed by AI Pow are very low power, which allows the EdgeView to run on batteries tentatively. Efficient power consumption has an enormous advantage because it further reduces installation costs by eliminating wiring. In addition, power consumption is expected to be low because the camera won't transmit a video feed, just relevant information and snapshots, which directly results in a light use of the current restaurant network, further reducing the setup cost of the system we propose.

Few intelligent video cameras have combined connected cameras with intelligence at the edge, such as Qualcomm's Vision Intelligence Platform for smart cities, Presto [6] camera technology for restaurant's drive-through, and Blink camera system for smart home security. As a result, they are able to perform several standard image understanding tasks such as face/body detection, object classification, with affordable resource budgets at the edge platforms, some even battery-powered (e.g., Blink). However, running the more sophisticated and costly video-based applications with low latency and energy costs remains challenging for the edge, calling for more holistic algorithm-data-hardware-task co-designs, as described in Section 3. Also, to our best knowledge, no smart camera product has targeted the hospitality industry.

### 2.5 Intellectual Property

Advanced functions, such as state-of-the-art human and object detection, action recognition on edge devices, will be protected under the Berkeley Software Distribution (BSD) license, which forbidden users to develop their product based on these modules privately. We want to make sure our customers comply with the license agreement's requirements and avoid as much as possible illegal practices from third parties.

Additionally, our partner, AI Pow, has also developed a watermark technology for ML, and we plan to be incorporated in future product releases to protect our IP and claim ownership of the resulting assets of this collaboration. Without such watermark technology from AI Pow, our digital assets can be seriously susceptible to content theft or unauthorized use. The access to ML-specific IP protection technologies further exemplifies the significance of the UT-AI Pow partnership. Finally, trade secrets strategies that are standard in the industry may also be part of our intellectual property strategy, including a Non-Disclosure Agreement (NDA) and a non-compete clause. UT's on-campus accelerator will be an important guide in our IP strategy.

## 3 Technical Challenges and Applied Research Plan

3.1 Outlier Detection for Automotive: Offline training stage. Describe how to deal with the Automotive data, the algorithm, AutoML.

3.2 Model Compression: Distillation, quantilization. Decision tree.

3.3 New Data Collection and Fine-tuning: Collect new data from the car. Human-in-loop (on device, meta-aad). Retrain the model and deploy new model when maintenance

One latest trend of AI is undoubtedly to move into the IoT devices [7], and there has been a tremendously growing interest in bringing ML-powered intelligence into these devices to enable ubiquitous intelligence that promises to transform the quality of human life [8, 9]. Today's large-format, high frame rate, high dynamic range camera sensors produce massive quantities of video data among other sensory modalities, today's large format, high frame rate. The commercially available formats already produce upwards of 30 Gb/sec and seek to triple that to 100 Gb/sec in the coming few years. Meanwhile, state-of-the-art video understanding models, typically deep neural networks (DNNs), require billions of operations per inference and require vast quantities of memory, making them poor candidates for embedded implementation. Furthermore, small traditional DNNs cannot operate with high accuracy on high dimensionality input data because they do not contain enough parameters to train to high accuracy. Instead, current state-of-the-art (SOTA) approaches to AI-based video processing seek higher accuracy from exponentially larger networks for marginal improvements to accuracy.

To reclaim the accuracy and functionality of DNNs in power-constrained sensing platforms, researchers have devoted profound efforts to creating Low Size, Weight, and Power (SWAP) back-end algorithms. For example, many model compression techniques have been developed, such as pruning and quantization, to reduce ML algorithms' complexity while maintaining their accuracy largely. Despite progress being made, there exists a vast, and increasing gap between the prohibitive complexity of powerful ML algorithms and the constrained resources in IoT devices [10]. Specifically, the following research gaps and challenges are identified:

- □ First, different applications and hardware platforms require diverse models that favor different compression schemes. There is unlikely to be a winning-all compression scheme.
- □ Moreover, each application calls for its unique efficiency-accuracy trade-off, for which even experts require tedious trials-and-errors to find the best solution.
- □ Further, even model compression is just one single and often the most idealistic step of the efficient implementation pipeline. For efficiently understanding video in the edge, many more factors such as data redundancy and cross-task model re-using.

### 3.1 Adaptive Model Compression Framework: From Handcrafted to Automated

For model efficiency, we mainly focus on tools from model pruning, quantization and distillation. Pruning removes redundant parameters that are insensitive to performance [11]; quantization reduces number of bits in weights [12]; and distillation trains a small network by supervision of a large model [13]. For hardware efficiency, the community mainly co-design both networks and hardware operators to reach lower power consumption and faster inference [14].

All the above options are handcrafted. Existing toolboxes only provide opportunities for choosing one from a few stand-alone compression techniques (e.g., pruning or quantization), while existing works have demonstrated that combining different compression techniques can achieve better trade-offs between complexity and accuracy [15, 16]. However, it is nontrivial to identify optimal combinations of varying compression methods tailored for the wide variety of IoT applications and accuracy-efficiency trade-offs, which requires significant expertise in ML compression and is not widely practical for users with limited ML compression knowledge. Although neural architecture search (NAS) methods [17] have been adopted to find lightweight architectures, the explosive search cost of NAS by repeatedly modifying and evaluating model architectures in the gigantic search space usually puts it beyond the hands of practitioners. Moreover, the designs of NAS search spaces and search algorithms still rely on domain expertise and handcrafting.

We aim to propose an Adaptive Auto-Compression Engine (**AACE**) that is easy to use and comprehend to search for the best combination strategy of model compression automatically means given the ML models and the target device and efficiency. For the baseline of AACE's Auto-Compression, we will start with vanilla sequential combinations, i.e., we pre-define a set of commonly used pruning ratios and quantization bits and then adopt a grid search method to combine different pruning and quantization compositions.

Although such grid search is more efficient than exhaustively exploring the whole design space, it is still time-consuming and computationally expensive, motivating us to propose a more principled Auto-Compression method via reinforcement learning (RL). Our Auto-Compression can efficiently search for the optimal combination of different compression techniques to maximize the accuracy while satisfying the target efficiency. Our Auto-Compression engine adopts the DDPG [18] agent as in [19] to acquire a set of actions (e.g., layer-wise pruning ratios and quantization bits), given the pretrained model, target device, and hardware efficiency. In particular, we apply the compression strategy sampled from the DDPG agent to train the given model for only a few epochs, starting from the pretrained model, and then use the validation accuracy as the reward to update the DDPG agent. This process is iterated to acquire the converged compression combination that meets the specified accuracy and efficiency. The RL search is described below: **State Space.** The state space is defined as:

$$\mathcal{S} = \{L,\ L_{total},\ type,\ H,\ W,\ C_{in},\ C_{out},\ K,\ S,\ skip,\ E_{target},\ E_{current},\ \beta_{l-1},\ B_{l-1}\}, \tag{1}$$

where $L$ is the layer index, representing the current layer to be optimized, $L_{total}$ is the number of layers, $type$ is the layer type (e.g., 2D convolution, 1D convolution, or fully connection layer), $H$ and $W$ are the input feature map's height and width, respectively, $C_{in}$ and $C_{out}$ are the input and output channel number, respectively, $K$ and $S$ are the convolution kernel size and stride (both set to 1 for fully connected layer), respectively, $skip$ is a binary number indicating if this layer has a skip connection, $E_{target}$ and $E_{current}$ are the target and current hardware efficiency, $\beta_{l-1}$ and $B_{l-1}$ are the pruning ratio and quantization bit-width of the previous layer.

**Action Space.** Auto-Compression of AACE sequentially determines the compression strategy for each layer. Specifically, the action space is defined as:

$$\mathcal{A} = \{\beta_l,\ B_l\}, \tag{2}$$

where $\beta_l \in (0, 1]$ and $B_l \in \{8-\text{bit}, 16-\text{bit}, 32-\text{bit}\}$ are the pruning ratio and quantization bits at the current layer $l$. Auto-Compression prunes and quantizes the current layer using the $L_1 - Norm$ ranked pruning and Tanh-PACT quantization according to the output action.

**Reward.** The reward of Auto-Compression consists of the hardware efficiency and the task accuracy. After generating actions for all the layers in the ML model, given $E_{target}$, $E_{current}$, the original accuracy $Acc_{original}$, and the compressed accuracy $Acc_{current}$, the reward is computed as:

$$\mathcal{R} = \begin{cases} -100, & \text{if } E_{target} < E_{current} \\ Acc_{current} - Acc_{original}, & \text{if } E_{target} \geq E_{current} \end{cases} \tag{3}$$

It is worth noting that Auto-Compression also supports user-customized reward functions to accommodate diverse user scenarios, and hardware efficiency constraint is not a hard constraint, motivating us to explore on-device ML, performance estimators. While real-measured ML execution costs are very useful in developing device-aware compressed/efficient ML algorithms, it is often impractical to incorporate real-device measurement due to its tedious efforts and required knowledge of deploying ML models into devices. As such, various on-device ML performance estimators have been developed, which can be categorized into the following three groups:

□ *Lookup table-based estimators*: The hardware performance of different operators on a target device (e.g., a specific mobile phone) are first collected and then used to construct a lookup table. On-device performance of ML models is then estimated by summing up the performance of all the corresponding operators [20,21].

□ *Regression model-based estimators*: A regression model, of which the input is embedding vectors representing the considered ML model, is trained with ML models and their on-device costs on the target device and then used to predict the performance of unseen models [22,23].

□ *Device specific simulator*: Unlike commercial IoT devices, whose DNN execution details (e.g., dataflow and tilling factors) are often a black-box to general users, ASIC- and FPGA-based accelerators [24–27] often provide the details of their micro-architecture and dataflow, making it possible to obtain their acceleration performance via analytical/cycle-accurate models.

We will develop built-in support for a wide spectrum of IoT devices. As such, our AACE framework distinguishes itself from existing estimators as the most comprehensive estimator supporting more number of devices and having little constraints on the ML model structures.

### 3.2 From Single-Task to Multi-Task Efficient Video Understanding

We need to implement and deploy all sub-algorithms with real-time inference for a holistic vision system that possesses multiple functionalities. For example, it could share backbone weights for different tasks to reduce costs. Moreover, learning numerous related tasks in parallel can benefit the performance of each task. However, most SoTA model compression methods focus merely on a single-task network and a limited number of datasets [28]. Model compression on multi-task networks could be applicable in many cases yet remains under-explored. We will propose an effective multi-task compression method that can combine multiple single-task models into one and compress jointly via utilizing cross-task relationships.

To effectively integrate information across different tasks, it is crucial to design loss functions to capture between-task features carefully. Therefore, we propose a two-step compression procedure. First, we use a relation-based knowledge distillation [29, 30] with multiple teachers [31, 32] to effectively integrate pre-trained single-task networks into one multi-task network, which can be designed to effectively preserve cross-task relationships [28,33]. Second, we prune this integrated network [34–36] on to reduce the model size further, aiming to achieve equal or higher accuracy compared to the old Single-task networks while having a much smaller total size. Taking, for example, two pre-trained single-task networks $M_1$ on task $T_1$ and $M_2$ on task $T_2$, we create a shared-trunk network with two different output branches and construct the following relation-based loss function:

$$L_{Rel}(f_t, f_s) = \sum_{i=1}^{k} L_R(\psi_t(\phi_1(f_{t1i}), \phi_2(f_{t2i})), \psi_s(\phi_1(f_{s1i}), \phi_1(f_{s2i}))), \tag{4}$$

where $k$ represents the number of feature maps we extract relation, $f_{t1i}$ represents the $i$th feature map in teacher network $T_1$, $f_{s1i}$ is the $i$th feature map in the output branch of student network for $T_1$, $\phi(.)$ is the transformation function in case feature maps are not in the same dimensions, $\psi(.)$ is the similarity functions between pairs of feature maps, and $L_R$ is the correlation function between teacher and student feature maps. In this way, the feature-level relation between teacher networks will be preserved to facilitate the training of the multi-task model. The total loss is hereby as follows: $L_{total} = L_{s1} + L_{s2} + \lambda L_{Rel}(f_t, f_s)$, where $L_{s1}$ and $L_{s2}$ represent the task specific loss on student network, and $\lambda$ is the coefficient controlling the feature-level task relation preservation.

Based on this formulation, our longer-term will be to extend AACE to the multi-task scenario as well, which will in principle be straightforward but may practically require a more delicate search on compressing different parts of the network due to various single tasks' trade-offs.

### 3.3 Efficient Video Recognition Pipeline: Dropping Data-Level Redundancy

It is important to note that compressing the back-end AI model is only one of many important steps for practical video-based systems. Using compute capacities available on the intelligent camera allows for lower usage in the cloud. Less interesting parts of the video can quickly be filtered out at the frontend, which dramatically reduces the bandwidth that needs to be provisioned [37].

We propose to exploit this crucial data-level sparsity of video data for dimensionality and computation reduction by quickly zooming in with temporal-spatial saliency at the camera frontend. Specifically, we will first develop a reinforcement-learning-based temporal frame filtering algorithm, together with a bi-directional spatial attention algorithm. Both are targeting simple deployment on even low-resource frontend cameras, with a potential for in-memory or even in-pixel computing. They will guide the backend AI model to concentrate its processing power on only temporal-spatially important sub-regions quickly. The backend model will also provide top-down feedbacks to data sparsification. Note that those techniques will be **plug-and-play** with the model compression components in Section 3.1 and 3.2.

**Temporal Frame Filtering:** we propose a tiny always-on ML model that can filter out unnecessary frames at the earliest stage. The proposed ML model outputs the probability of whether the current frame can be skipped to compute. To accommodate the limited processing power, the model needs to be ultra-light, containing only simple operations. On the other hand, as the filtering largely determines the end-task performances, the model must be flexible enough to incorporate the feedback from the later stage of computing to adjust the policy.

To this end, we propose an RL-based approach, which *can be trained effectively off-line while maintaining simple yet efficient deployment*. Specifically, the module takes inputs from two consecutive frames and their differences with greyscale only. Previous work has shown that 64 channels of $3 \times 3$ convolutions can be computed in real-time on 1080P color images with only 22.62 mW power consumption [38]. Built upon this finding, we propose to use a 2-layer fully-convolution network followed by a global pooling, constrained by the PIP circuit capability.

The model will be off-line trained via RL. Given a video sequence, at each time step $t$, we sample a binary decision (*i.e.* drop or keep) to construct a new sequence of selected frames. And then, we apply a well-trained object tracker on top of it. The tracking performance indicates the frame selection quality, which is used as the main reward. Meanwhile, to encourage sparse selection for data reduction, the number of selected frames will serve as the negative reward. Our policy network will be trained to maximize the expected future reward via policy gradient methods [39]. For safety guarantee, the model will forcibly maintain a minimum frame rate $k$.

The model also accepts task-oriented information from the end-task ML algorithm via a low latency feedback loop to self-adaptively adjust its frame rate. We plan to adapt the seminal top-down attention model of [40], which weights inputs by an attention field. In the case of the image, each attention vector element corresponds to a spatial region of the image. These attention weights consider the task relevance of each region of the input space for the goal. It can be extended to attending spatial-temporal cubes in video data. As the feedback information must be simple, we propose a temperature factor $T$ in the softmax of the selection policy. If $T$ is one, the output distribution will be the same as the original softmax output; the smaller $T$ is, the sharper the distribution is. If the powerful end-task AI model believes that the current in-pixel filtering is too aggressive, it will reduce $T$ to increase the frame rate.

**Spatial Region Focusing:** Based on the first-stage always-on temporal filtering, we further propose the second-stage data reduction, to *attentively focus* on the salient region of a frame to eliminate uninformative pixels. Our ML model will achieve spatial attention and saliency in this objective, which can detect and characterize important spatial signals while monitoring high FPS video streams with low latency. The output will be a localized sub-image for each frame, on which the backend ML model could be focused.

There are two main forms of attention: the top-down attention selects regions of the input (stimulus) space most relevant to the specific task, and the bottom-up one detects salient formation in the environment regardless of the current task demand. The two usually have bi-directional interactions [41]. We propose a bi-directional spatial attention mechanism, that consists of the bottom-up data-driven saliency detection, and the top-down task-driven attention detection. Specifically, each input frame will be divided into $N \times N$ patches, each patch denoted by $x_i$ (full image $x$), where $N$ is chosen during offline training to balance processing granularity and latency. Each patch $x_i$ has an attention weight $e(x_i)$, computed as a weighted combination of the bottom-up attention $g_{\mathrm{bu}}(x_i|x)$ and the top-down attention $f_{td}(\mathbf{h})$ as follows: $e_i(x) = f_{\mathrm{td}}(\mathbf{h}) + \lambda\, g_{\mathrm{bu}}(x_i|x)$. Finally, a continuous region consisting of multiple image regions that has the largest accumulative attention weights will be passed to the later stage of computing.

To enable the computation of the bottom-up attention function $g_{\mathrm{bu}}$ in low-cost hardware, we propose a self-supervision method. We plan to use its immediate surrounding patches for each patch to reconstruct it while measuring the discrepancy. Intuitively, the more significant the discrepancy, the more "surprise" happened in the specific patch, meaning the model should pay

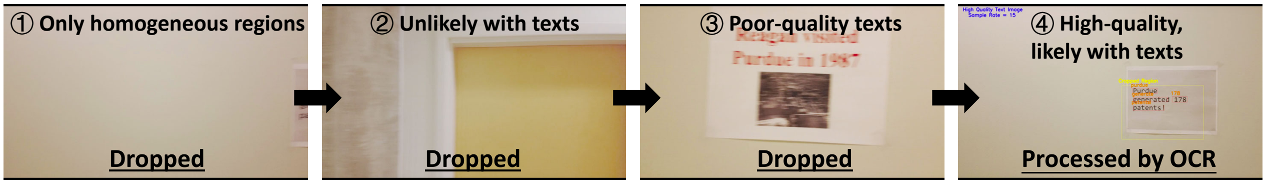


Figure 3: Examples when running the award-winning multi-stage video text spotting system developed by PI Wang's team. It is implemented on Raspberry Pi 3B+, loaded on an UAV flying in a test site. From top left to bottom right demonstrates our efficient video processing hierarchy.

more attention to it. The reconstruction model can be trained offline via a denoising auto-encoder [42]. For the top-down attention, it includes two hyperparameters: **h** and $\lambda$, where **h** is an attention weight vector and $\lambda$ is the adaptive coefficient. Both parameters are computed by the end-task powerful ML model and provided via the feedback loop (similar to temporal). For example, if later stage computing realizes the send-in sub-images contain no objects for many consecutive timestamps, then it will modify **h** and $\lambda$ in such a way that all pixels will be passed in. The combined top-down/bottom-up attention system will improve the system throughput while ensuring it is not be trapped in a bad state (*e.g.* missing important objects for a large number of frames).

### 3.4 Preliminary Work

Our team has a very strong record in efficient object detection [43,44], multi-camera re-identification [45], video object tracking [46], and co-design with hardware [16]. PI Wang's group won the second prize in the prestigious 2020 IEEE low-power computer vision (LPCV) challenge. The task was to recognize texts from UAV captured video using a Raspberry Pi 3B+. The group's winning solution includes a multi-stage pipeline, as demonstrated in Figure 3. It progressively rejects frames containing no object, objects but unlikely texts, or texts but in poor image quality. Eventually, only the frames with clean text regions ($\sim$1/20 - 1/30 of total frames) are fed to the final sophisticated OCR model. The first step was purely based on input smoothness, similar to our proposed *bottom-up attention*. The second and third steps considered early-stage *top-down feedbacks*.

### 3.5 Success Metrics

Our technical metrics include the correct accomplishment of detection of persons and their interaction with the environment. This is measured in a confidence score for the edge device. We will compare our solution with the commercial intelligent cameras available in this market sector, such as solutions developed by Presto [6]. The target is to achieve 20× data bandwidth reduction, 20× model size/storage reduction, 100+× computing latency reduction (measured by FLOPs, and hardware-measured latency), and 20+× energy saving (measured by energy-delay product, and real hardware energy measurement), while not compromising the achievable accuracy on the chosen video-based utility tasks that are of practical relevance to the hospitality industry. The candidate utility tasks include video abnormal event detection, person recognition and tracking, person counting, body pose or gesture recognition, to name a few. We will consider and evaluate the hardware platforms, including the Raspberry Pi 3b and Maxim's accelerated-CNN MAX78000 ASIC platform. Our results will directly correlate to the restaurant's quantitative factors, such as speed of service, location cleanliness, complaints from customer service, and increased traffic. And qualitative factors such as restaurant's image, customer service, and employee satisfaction.

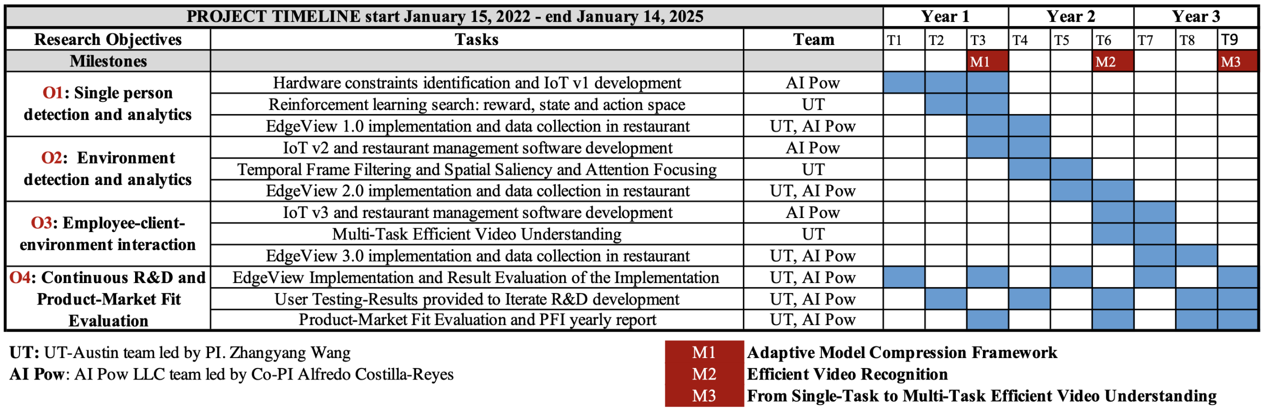
### 3.6 Risk and Mitigation Plan

We have discussed many of the technical challenges and design choices in Sections 3.1 - 3.3. We next enumerate some of these technical risks, along with strategies for risk mitigation. For example, suppose an accuracy degradation is encountered at the beginning of the video recognition pipeline, e.g., data-level dropping in Section 3.4. In that case, we will re-adjust the dropping parameters (*e.g.*, the minimum frame rate $k$ in temporal filtering and the spatial smoothing factor in spatial focusing) to elevate the accuracy while maintaining overall efficiency. The performance could be further improved via jointly hyperparameter training using advanced black-box optimization techniques such as zeroth-order learning [47,48]. As another example, if the degradation comes from the backend AI model, we will alternatively resort to the knowledge distillation or self-supervised learning strategy to abate the accuracy loss of pruned models.

**High barriers to market entry** is an unwanted possibility, and this is because the implementation of edge AI to detect persons and their interaction with the environment and other humans remains unknown due to its case-specific attribute. To mitigate this risk, we will further pair scientific R&D on on-device AI and a build-measure-learn feedback loop to provide corresponding engineering solutions in RetailEdge to address the problems above. Specifically, we will expand our findings into our system by following our previous research on contextual business analytics following a lean methodology approach. We will then advance the end-to-end system by applying our efficient video recognition technologies.

### 3.7 Timeline and Milestones

Our research and development will center on intelligent IoT cameras which our target customers can readily adopt.



| PROJECT TIMELINE start January 15, 2022 - end January 14, 2025 | | | Year 1 | | | Year 2 | | | Year 3 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Research Objectives | Tasks | Team | T1 | T2 | T3 | T4 | T5 | T6 | T7 | T8 | T9 |
| Milestones | | | | | M1 | | | M2 | | | M3 |
| O1: Single person detection and analytics | Hardware constraints identification and IoT v1 development | AI Pow | 🟦 | 🟦 | 🟦 | | | | | | |
| | Reinforcement learning search: reward, state and action space | UT | | 🟦 | 🟦 | | | | | | |
| | EdgeView 1.0 implementation and data collection in restaurant | UT, AI Pow | | | 🟦 | 🟦 | | | | | |
| O2: Environment detection and analytics | IoT v2 and restaurant management software development | AI Pow | | | 🟦 | 🟦 | | | | | |
| | Temporal Frame Filtering and Spatial Saliency and Attention Focusing | UT | | | | 🟦 | 🟦 | | | | |
| | EdgeView 2.0 implementation and data collection in restaurant | UT, AI Pow | | | | | 🟦 | 🟦 | | | |
| O3: Employee-client-environment interaction | IoT v3 and restaurant management software development | AI Pow | | | | | | 🟦 | 🟦 | | |
| | Multi-Task Efficient Video Understanding | UT | | | | | | 🟦 | 🟦 | | |
| | EdgeView 3.0 implementation and data collection in restaurant | UT, AI Pow | | | | | | | 🟦 | 🟦 | |
| O4: Continuous R&D and Product-Market Fit Evaluation | EdgeView Implementation and Result Evaluation of the Implementation | UT, AI Pow | 🟦 | | 🟦 | | 🟦 | | 🟦 | | 🟦 |
| | User Testing-Results provided to Iterate R&D development | UT, AI Pow | | 🟦 | | 🟦 | | 🟦 | | 🟦 | 🟦 |
| | Product-Market Fit Evaluation and PFI yearly report | UT, AI Pow | | | 🟦 | | | 🟦 | | 🟦 | 🟦 |

**UT:** UT-Austin team led by PI. Zhangyang Wang
**AI Pow**: AI Pow LLC team led by Co-PI Alfredo Costilla-Reyes

| | |
|---|---|
| M1 | **Adaptive Model Compression Framework** |
| M2 | **Efficient Video Recognition** |
| M3 | **From Single-Task to Multi-Task Efficient Video Understanding** |

Figure 4: EdgeView project timeline.

**Milestone 1.** EdgeView 1.0 is composed of the first iteration of the IoT hardware and ML software (adaptive model compression framework) to identify a person and obtain metrics such as idle time of employees and customers, restaurant abandonment rates, and employee recess time. The first year will help us distinguish the constrains SMB constraints face in their locations.

**Milestone 2.** EdgeView 2.0 will build on top of the first iteration to also detect other objects in the environment, such as empty table time, food detection for correct to-go order fulfillment to reduce the error reduction for order served (i.e., orders with allergic reaction food-specifications), table and kitchen cleanliness. For this stage, we propose a restaurant management desktop software and upgraded hardware.

**Milestone 3.** EdgeView 3.0 will build on top of two previous iterations to analyze customer-employee-environment engagement. For example, our restaurant partner requested particular food safety features such as use-of-cellphone for employees directly serving or preparing food to ensure a clean kitchen-to-table food service. Another important metric is guiding employees to engage with the client when looking for assistance promptly. We anticipate that EdgeView 3.0 will require a more advanced IoT v3 hardware to provide efficient food preparation and packaging for delivery services as they become increasingly important for the restaurant industry.

# 4 Achieving Societal Impact through Commercial Realization

### 4.1 Commercialization strategy

As we advance EdgeView's product completion through this Partnership For Innovation project, this award will help us decrease our technology's market and technology risk. In addition, the PI and Co-PI team has discussed the following options to move EdgeView technology from fundamental research to the marketplace.

A first option is a use of licensing as our commercialization strategy of choice. Our research group at UT-Austin will specialize in technology development with this strategy while conceding the marketing and sales activities to the AI Pow team to effectively become a licensee. Specifically, AI Pow LLC will need to carry all of the other tasks associated with commercialization, such as marketing, sales, and distribution, and hardware activities such as further embedded engineering, manufacturing, system integration, etcetera. Thus, a licensing commercialization strategy will be beneficial to both parties; on the one hand, the UT-Austin team can focus on research and development while AI Pow can use this strategy to enter new markets or retain market position using a technology advantage without having to make costly investments in the original R&D, skilled personnel, and equipment.

If necessary, specific financing methods to mature a licensable technology include further participation in additional federal programs oriented to encourage fundamental and applied R&D, such as the Small Business Technology Transfer (STTR). In the same way, other companies are welcome to join our PFI efforts to further complete the commercialization of EdgeView, which includes new restaurant groups in different states in the US. Finally, a private contract for application-specific R&D, is another form of financing that could help us further guide the sustainability of the commercialization efforts after implementing the research PFI activities.

An industry partner like AI Pow will help us understand the technical and business functions required to commercialize EdgeView. Particularly, it will be key for the UT-Austin team to understand the pros and cons of manufacturing and marketing EdgeView to a limited number of system integrators in the hospitality industry; it may be more appropriate to establish a strategic alliance with AI Pow in the supply chain. For example, AI Pow could become a site for testing our current and future technologies and the sales channel of such technology. By forming this UT-Austin AI Pow partnership, we want to develop a teaming partner.

Another tentative commercialization strategy that both teams have pondered is an equity investment in the partner company in exchange for the exclusive use of the outputs of this collaboration. This option will require collaboration beyond mainly providing technical expertise. We are aware that the I-Corps training will be helpful for our partnership to find a product-market fit but equally important, to understand better the hospitality industry and recognize the best business model for our joint technical and business efforts.

### 4.2 Assessment plan to gauge the success of the research partnerships and third-party collaboration

The partnership described in this PFI proposal recognizes the benefits of the I-Corps participation to mature the technology commercialization effort of EdgeView. I-Corps will be a pivotal instrument to gauge the success of our research partnership. AI Pow will be directly involved with this process as the industry mentor that can help us reach out to potential customers and understand the market from a realistic and practical perspective. I-Corps aggressively focuses on validating and invalidating our business hypothesis in our search of product-market fit. This iterative process involves strong collaboration between our research group and AI Pow. The weekly customer discovery assessment of this program represents a way to gauge the success of the research partnerships as it provides a systematic way to understand if our value proposition has a market fit through actual data from potential customers and investors. The overall assessment of our product discovery efforts, and thus our planned commercialization strategy will be more evident when preparing for the go-no-go decision day.

Another continuous evaluation system will be through our collaborators at the UT-Austin SEAL incubator. Particularly, they can help us provide unbiased feedback regarding how functional the prototype is and how attractive our business model is to potential investors. SEAL will help us measure if this partnership has been successful for the PI, the Co-PI team, and other players in the following ways.

UT-Austin as a whole will benefit from this collaboration as they establish a project to translate, and possibly license, the technology developed in this campus to the industry. Notably, the PI's research lab will establish a channel to further their fundamental research into deploying more mature technologies to solve real market needs. Realistic and well-articulated fundamental-research objectives should lead to several future grant proposals awarded. Additionally, the PI's research group will train students and postdoctoral researchers with the tools and network to lead future entrepreneurial endeavors. A way to measure this progress is through the programs offered by SEAL, such as the attendance and completion of entrepreneurial courses, workshops, business plan competitions, and related certificates.

AI Pow will deem this collaboration a success by acquiring a technical competitive advantage by implementing state-of-the-art technology developed in this proposal. By partnering with UT-Austin, AI Pow can develop more confidence for their current and future investors and clients while also gaining higher visibility in the market. Austin's vibrant and growing hospitality industry also represents more potential clients to AI Pow, and, therefore, the number of new clients is a metric we want to measure to gauge the success of our collaboration. Our partnership will also measure the PR generated during our collaboration, usually presented in different technology and entrepreneurial venues, such as MIT technology review and Forbes Magazine.

An important metric for the PI and Co-PI relates to the number of future partners introduced by both teams that will effectively grow their joint business network. For example, through AI Pow, the company Maxim Integrated, a NASDAQ-100 company, is now a collaborator of both teams working in state-of-the-art edge technology. The symbiosis of the PI and Co-PI teams has been equally fruitful to AI Pow, which now has access to the talented students and researchers in our research group that will be qualified to join them if the right opportunities arise. Not to mention that EdgeView is well-positioned to access various opportunities for PR, investment, and access to talent since UT-Austin is at the center of one of the most active and emerging entrepreneurial ecosystems in the world.

## 5 Project Team

**PI Zhangyang Wang** is currently an Assistant Professor of Electrical and Computer Engineering at UT Austin. Prof. Wang's team has solidified a leading role in the fields of machine learning and computer vision. His team has made many contributions to automated machine learning (AutoML), robust learning, and efficient learning. In particular, his team actively works on resource-efficient training and inference for deep networks, as well as algorithm-hardware co-design. Their innovative works are widely published in the top venues such as NeurIPS, ICML, ICLR, CVPR, ICCV, ECCV and ISCA. Those works have attracted wide attention, covered by high-visibility technical media such as MIT News and IDG Connect. Recently, his team won the second prize in the prestigious 2020 IEEE low-power computer vision (LPCV) challenge (video track). His research has been supported by many federal, industry and university grants.

**Co-PI Alfredo Costilla-Reyes** is currently leading the technology innovation at AI Pow LLC. He graduated from Entrepreneurship and Technology Commercialization program at Mays Business School and the doctorate program in Electrical Engineering, both from Texas A&M University. Dr. Alfredo has been a recipient of the NSF I-Corps Site, 2017-2018 Kirchner, Silicon Labs and the McFerrin-Entrepreneurship Fellowships, and the prestigious Mexico National Youth Award, presented by the president of Mexico for his contributions in science, technology, and entrepreneurship. Specifically, Dr. Alfredo has led projects regarding embedded software and systems for future agriculture, battery-less wearable consumer electronics, application-specific integrated circuits, and wireless systems for IoT applications. His research and entrepreneurial endeavors have participated in YCombinator's YC120 event, Silicon Valley Bank Trek, McFerrin Center for Entrepreneurship business incubator, and Rice University's OwlSpark accelerator.

**Scott Hoang** is a first-year Ph.D. student in UT Austin, supervised by PI Wang. His research interests include neural architecture search, model compression, and video understanding. He has been a student entrepreneur and an active participant of many tech entrepreneur camps.

# 6 Partnerships

We have established one partnership with AI Pow that has also expanded the third party participation with LaunchPad at UT, the on-campus accelerator and incubator, Maxim integrated, and B2J-suancai-fish restaurant. AI Pow contribution will be critical in our business exploration and co-guiding the research efforts of this PFI proposal towards solving the current needs as the hospitality industry transitions to a mixed physical-digital operation. AI Pow expertise in embedded systems, edge-AI, and commercialization will catalyze and accelerate technology development toward commercialization. We are confident that our joint participation will solve real problems using the seed fundamental research derived from the NSF lineage presented here.

Maxim Integrated has long collaborated with AI Pow. Now we can leverage such a previous relationship to further reduce the risk of this technology by using their hardware expertise in application-specific integrated circuit (ASIC) architecture background for IoT and ML. We are thrilled to be their partners for their premier on-device AI platform, which will give us access to essential resources, such as development platforms, curated data, and technical discussions.

The restaurant B2J-suancai-fish is a great representative example of the target market we believe will benefit the most from EdgeView. This key participant will help us gather information and understand the most urgent problems that business owners in the hospitality industry regularly face in their daily activities. B2J-suancai-fish will facilitate an actual establishment where EdgeView will be installed and the resulting data analyzed. B2J-suancai-fish participation will be vital in attracting other restaurants.

**We have established the following assessment plan that will help gauge the success of our partnership**. While our team at UT-Austin will get the chance to place our state-of-the-art research outputs in research articles (demo papers) and to the market, AI Pow will be able to co-design and produce a prototype with cutting-edge technology, which then can translate into a possible competitive advantage. Our joints metrics are the number of retailers that show interest or have adopted EdgeView during this PFI project; investors interesting in participating in the financing round to commercialize this technology; licenses generated and adopted by AI Pow as well as the revenue generated from EdgeView's licenses.

Maxim Integrated will find this partnership a success by finding in AI Pow a new client that supplies their MAX7800 edge-AI platform. In the same way, our UT-AI Pow partnership will also benefit them through the Demo Papers we plan to produce that will use and showcase their demo boards for the scientific community. Finally, our B2J-suancai-fish partner will benefit significantly by adopting the latest restaurant-intelligence technology over other family-owned restaurants and big-chain restaurants. This PFI partnership will measure success for B2J-suancai-fish in how well EdgeView satisfies their expectations. To achieve this, AI Pow and our team plan to have monthly meetings to analyze the following metrics: employee performance, customer and employee satisfaction, average revenue per employee, waste reduction, to name a few. Finally, we want to stress that B2J-suancai-fish will benefit from using a technology that solves a critical business analytics problem but also considered linguistic and cultural components of their business.

# 7 Training Future Leaders in Innovation and Entrepreneurship

## 7.1 Educational and leadership development plan

**The LaunchPad at UT** is UT Austin's entrepreneurial detonator that enables the most promising emerging startups across the UT-Austin campus and helps them confront their next market-driven milestone. LaunchPad at UT will help our students and faculty entrepreneurs transition out of UT resources and the marketplace. One particular program we plan to participate in is SEAL, a nine-week summer plan that helps graduate students and faculty teams explore a core company hypothesis to decide to accelerate, pivot, or rightfully shut down their venture. This program also includes workshop Sessions, where veteran entrepreneurs and subject matter experts lead workshop sessions and provide unfiltered insider knowledge on building a scalable business. Sessions cover technical topics and human dynamics such as go-to-market strategies and the multifaceted reality of being a founder.

Community is a central entrepreneurial education development part. NSF I-Corps and LaunchPad at UT handpicks experts specific to our PFI project to mentor the PI and Co-PI team throughout the entire program. At the same time, LaunchPad at UT facilitates intentional relationship-building with entrepreneurs and investors from the Austin startup community. In the same way, such an audience is invited to D-Day, where our students will pitch to tentative investors and customers. Our graduate students, professors, and other researchers will provide feedback and coaching to help all members develop their communication and presentation skills. Then, based on individual research and expert feedback, teams determine whether their venture can survive the real world and make a go or no-go decision.

## 7.2 Intellectual merit and broader impacts

The proposed project's educational and leadership activities will enhance the knowledge and readiness of the student for innovation and technology commercialization beyond the usual research experience by allowing them to create a network of more entrepreneurs of similar backgrounds that later on and become business partners. Also, we have planned this proposal to encourage activities for the student to engage with multiple people that will require him to develop skills to convey a complex idea in persuasive terms that investors and potential customers can understand the business potential of their research. Moreover, by completing multiple interviews during I-Corps and the help of Launchpad at UT, the student will create a network of experts in AI, an industry with enormous potential for innovation and further business participation.

With the successful completion of the I-Corps and SEAL programs, the student will be better positioned to understand a systematic process to find a product-market fit that can apply in future projects. Similarly, the student's direct involvement in preparing proposals for technology commercialization purposes will provide opportunities to learn such a process from our experienced research team. In addition, the graduate student will have a chance to learn best practices in proposal preparation, including identifying key research questions, defining objectives, describing the approach and rationale, market analysis, and constructing a work plan, timeline, and budget.

# 8 Broadening Participation

The effects of this project will directly benefit a Chinese restaurant and will be expanded to include Mexican restaurants, which combined represent the majority of the single location full-service restaurants in Texas. Such heritage extends to the PI and Co-PI culture, who have made their mission is to give SMB owners of minority backgrounds access to advanced technology to be competitive in our big-data-driven world.

A goal of our partnership with AI Pow is to produce a product and service that can be affordable for SMBs in the US. We have established a requirement to provide an easy way to install and adopt these technologies for business owners that don't have a team or the engineering background to implement the solutions from scratch. To stimulate the US economy, EdgeView will enable the American hospitality industry to embrace the digital revolution. We plan to do so by providing tools that are affordable but also easy to understand. An easy-to-understand EdgeView means that our business analytics software results should give the business owners actionable items to be more efficient, provide excellent customer service, and have a competitive advantage.

Finally, this PFI project will enhance partnerships between academia and industry in the United States. The UT-AI Pow partnership will follow aligned goals with their complementary skill-sets. Our joint vision is to use on-device AI to level the field for small restaurant owners to effectively democratize the use of artificial intelligence in Texas, the US, and beyond.

# References

[1] Dan Cook. *IBISWorld US Industry (NAICS) Report 51121C. Business Analytics  Enterprise Software Publishing in the US*, March 2021. Retrieved from IBISWorld database.

[2] Yue Wang, Jianghao Shen, Ting-Kuei Hu, Pengfei Xu, Tan Nguyen, Richard Baraniuk, Zhangyang Wang, and Yingyan Lin. Dual dynamic inference: Enabling more efficient, adaptive, and controllable deep inference. *IEEE Journal of Selected Topics in Signal Processing*, 2020.

[3] Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. *arXiv preprint arXiv:1803.03635*, 2018.

[4] Tianlong Chen, Jonathan Frankle, Shiyu Chang, Sijia Liu, Yang Zhang, Zhangyang Wang, and Michael Carbin. The lottery ticket hypothesis for pre-trained bert networks. *NeurIPS*, 2020.

[5] Tianlong Chen, Jonathan Frankle, Shiyu Chang, Sijia Liu, Yang Zhang, Michael Carbin, and Zhangyang Wang. The lottery tickets hypothesis for supervised and self-supervised pretraining in computer vision models. *CVPR*, 2020.

[6] PitchBook Data. *Presto | Private Company Profile*, July 2021. Retrieved from PitchBook database.

[7] Cisco Edge-to-Enterprise IoT Analytics for Electric Utilities. *Cisco Solution Overview*, 2018.

[8] Benjamin Ross. AI at the Edge Enabling a New Generation of Apps, Smart Devices, March 2020.

[9] Massimo Merenda, Carlo Porcaro, and Demetrio Iero. Edge Machine Learning for AI-Enabled IoT Devices: A Review. *Sensors (Basel, Switzerland)*, 20(9), April 2020.

[10] Mi Zhang, Faen Zhang, Nicholas D. Lane, Yuanchao Shu, Xiao Zeng, Biyi Fang, Shen Yan, and Hui Xu. *Deep Learning in the Era of Edge Computing: Challenges and Opportunities*, chapter 3, pages 67–78. John Wiley & Sons, Ltd, 2020.

[11] Zhuang Liu, Mingjie Sun, Tinghui Zhou, Gao Huang, and Trevor Darrell. Rethinking the value of network pruning. *arXiv preprint arXiv:1810.05270*, 2018.

[12] Xiaofan Lin, Cong Zhao, and Wei Pan. Towards accurate binary convolutional neural network. In *NeurIPS*, 2017.

[13] Antonio Polino, Razvan Pascanu, and Dan Alistarh. Model compression via distillation and quantization. *arXiv preprint arXiv:1802.05668*, 2018.

[14] Bichen Wu. Efficient deep neural networks. *arXiv preprint arXiv:1908.08926*, 2019.

[15] Asit Mishra and Debbie Marr. Apprentice: Using knowledge distillation techniques to improve low-precision network accuracy. In *International Conference on Learning Representations*, 2018.

[16] Yang Zhao, Xiaohan Chen, Yue Wang, Chaojian Li, Haoran You, Yonggan Fu, Yuan Xie, Zhangyang Wang, and Yingyan Lin. Smartexchange: Trading higher-cost memory storage/access for lower-cost computation. In *ISCA*, 2020.

[17] Jiahui Yu. *Towards efficient, on-demand and automated deep learning*. PhD thesis, University of Illinois at Urbana-Champaign, 2020.

[18] Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.

[19] Yihui He, Ji Lin, Zhijian Liu, Hanrui Wang, Li-Jia Li, and Song Han. Amc: Automl for model compression and acceleration on mobile devices. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 784–800, 2018.

[20] Han Cai, Ligeng Zhu, and Song Han. Proxylessnas: Direct neural architecture search on target task and hardware. *arXiv preprint arXiv:1812.00332*, 2018.

[21] Bichen Wu, Xiaoliang Dai, Peizhao Zhang, Yanghan Wang, Fei Sun, Yiming Wu, Yuandong Tian, Peter Vajda, Yangqing Jia, and Kurt Keutzer. Fbnet: Hardware-aware efficient convnet design via differentiable neural architecture search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10734–10742, 2019.

[22] Evgeny Ponomarev, Sergey Matveev, and Ivan Oseledets. Leti: Latency estimation tool and investigation of neural networks inference on mobile gpu. *arXiv preprint arXiv:2010.02871*, 2020.

[23] Yunyang Xiong, Hanxiao Liu, Suyog Gupta, Berkin Akin, Gabriel Bender, Pieter-Jan Kindermans, Mingxing Tan, Vikas Singh, and Bo Chen. Mobiledets: Searching for object detection architectures for mobile accelerators. *arXiv preprint arXiv:2004.14525*, 2020.

[24] Yu-Hsin Chen, Joel Emer, and Vivienne Sze. Eyeriss: A spatial architecture for energy-efficient dataflow for convolutional neural networks. *ACM SIGARCH Computer Architecture News*, 44(3):367–379, 2016.

[25] Hardik Sharma, Jongse Park, Naveen Suda, Liangzhen Lai, Benson Chau, Vikas Chandra, and Hadi Esmaeilzadeh. Bit fusion: Bit-level dynamically composable architecture for accelerating deep neural network. In *2018 ACM/IEEE 45th Annual International Symposium on Computer Architecture (ISCA)*, pages 764–775. IEEE, 2018.

[26] Xiaofan Zhang, Junsong Wang, Chao Zhu, Yonghua Lin, Jinjun Xiong, Wen-mei Hwu, and Deming Chen. Dnnbuilder: an automated tool for building high-performance dnn hardware accelerators for fpgas. In *2018 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, pages 1–8. IEEE, 2018.

[27] Pengfei Xu, Xiaofan Zhang, Cong Hao, Yang Zhao, Yongan Zhang, Yue Wang, Chaojian Li, Zetong Guan, Deming Chen, and Yingyan Lin. Autodnnchip: An automated dnn chip predictor and builder for both fpgas and asics. In *The 2020 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays*, pages 40–50, 2020.

[28] Yu Cheng, Duo Wang, Pan Zhou, and Tao Zhang. A survey of model compression and acceleration for deep neural networks. *arXiv preprint arXiv:1710.09282*, 2017.

[29] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[30] Jimmy Ba and Rich Caruana. Do deep nets really need to be deep? In *NeurIPS*, 2014.

[31] Emilio Parisotto, Jimmy Lei Ba, and Ruslan Salakhutdinov. Actor-Mimic: Deep multitask and transfer reinforcement learning. *arXiv preprint arXiv:1511.06342*, 2015.

[32] Andrei A Rusu, Sergio Gomez Colmenarejo, Caglar Gulcehre, Guillaume Desjardins, James Kirkpatrick, Razvan Pascanu, Volodymyr Mnih, Koray Kavukcuoglu, and Raia Hadsell. Policy distillation. *arXiv preprint arXiv:1511.06295*, 2015.

[33] Ini Oguntola, Subby Olubeko, and Christopher Sweeney. SlimNets: An exploration of deep model compression and acceleration. In *HPEC*, 2018.

[34] Song Han, Jeff Pool, John Tran, and William Dally. Learning both weights and connections for efficient neural network. In *NeurIPS*, 2015.

[35] Yann LeCun, John Denker, and Sara Solla. Optimal brain damage. In *NeurIPS*, 1989.

[36] Hao Li, Asim Kadav, Igor Durdanovic, Hanan Samet, and Hans Peter Graf. Pruning filters for efficient convnets. *arXiv preprint arXiv:1608.08710*, 2016.

[37] Ganesh Ananthanarayanan, Paramvir Bahl, Peter Bodík, Krishna Chintalapudi, Matthai Philipose, Lenin Ravindranath, and Sudipta Sinha. Real-time video analytics: The killer app for edge computing. *computer*, 50(10):58–67, 2017.

[38] Ruibing Song, Kejie Huang, Zongsheng Wang, and Haibin Shen. An ultra fast low power convolutional neural network image sensor with pixel-level computing. *arXiv preprint arXiv:2101.03308*, 2021.

[39] Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 1992.

[40] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2015.

[41] Martin Sarter, Ben Givens, and John P Bruno. The cognitive neuroscience of sustained attention: where top-down meets bottom-up. *Brain research reviews*, 2001.

[42] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *ICML*, 2008.

[43] Zhenyu Wu, Karthik Suresh, Priya Narayanan, Hongyu Xu, Heesung Kwon, and Zhangyang Wang. Delving into robust object detection from unmanned aerial vehicles: A deep nuisance disentanglement approach. In *ICCV*, 2019.

[44] Jiahui Yu, Yuning Jiang, Zhangyang Wang, Zhimin Cao, and Thomas Huang. Unitbox: An advanced object detection network. In *ACM MM*, 2016.

[45] Tianlong Chen, Shaojin Ding, Jingyi Xie, Ye Yuan, Wuyang Chen, Yang Yang, Zhou Ren, and Zhangyang Wang. Abd-net: Attentive but diverse person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8351–8361, 2019.

[46] Xinchao Wang, Bin Fan, Shiyu Chang, Zhangyang Wang, Xianming Liu, Dacheng Tao, and Thomas S Huang. Greedy batch-based minimum-cost flows for tracking multiple objects. *IEEE TIP*, 2017.

[47] Sijia Liu, Pin-Yu Chen, Bhavya Kailkhura, Gaoyuan Zhang, Alfred O Hero III, and Pramod K Varshney. A primer on zeroth-order optimization in signal processing and machine learning: Principals, recent advances, and applications. *IEEE Signal Processing Magazine*, 37(5):43–54, 2020.

[48] Xiangyi Chen, Sijia Liu, Kaidi Xu, Xingguo Li, Xue Lin, Mingyi Hong, and David Cox. Zo-adamm: Zeroth-order adaptive momentum method for black-box optimization. *NeurIPS*, 2020.