HPI Hasso Plattner Institut
IT Systems Engineering | Universität Potsdam

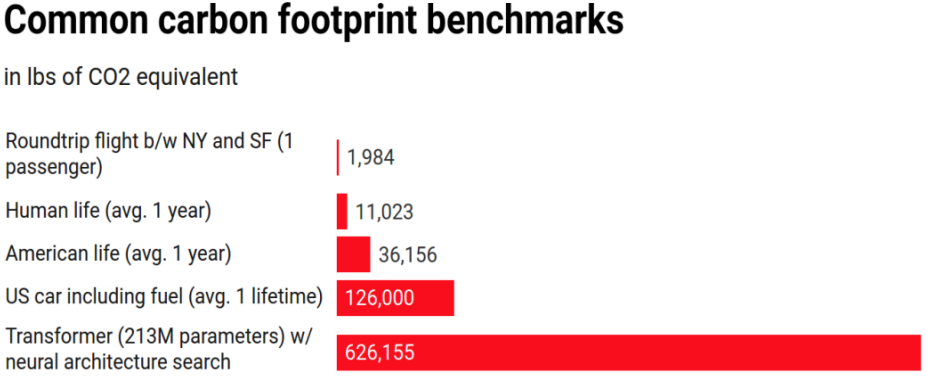
Applied **Edge AI**: Towards Efficient Deep Model Design and Multi-Edge Inference

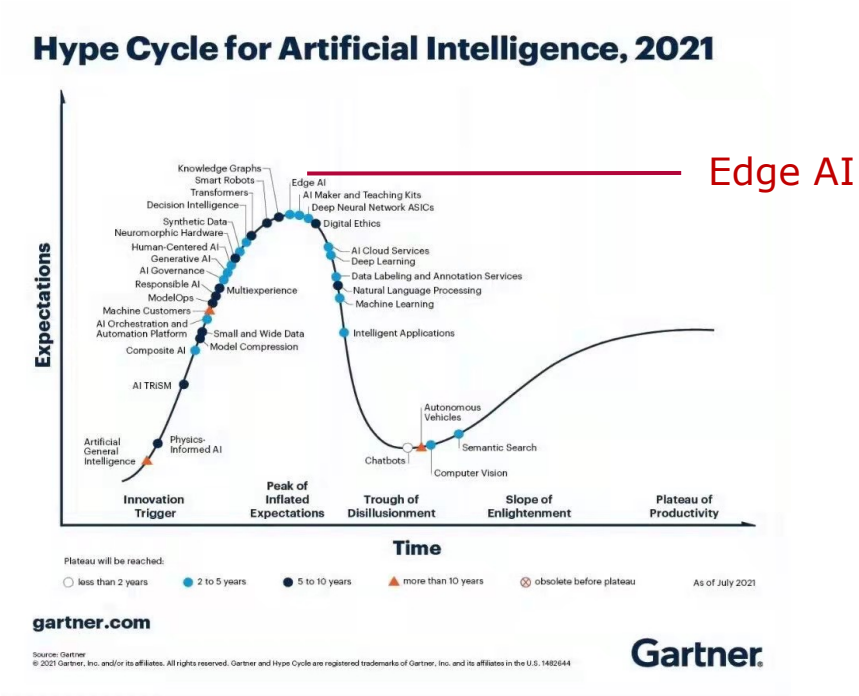
PD Dr. Haojin Yang

Why Edge AI?

- Thanks to Transformer and ViT, the power demand in AI computing has grown significantly.
- The massive carbon emission brought by AI computing cannot be ignored.

OpenAI's GPT-3 (Generative Pre-trained **Transformer,** with **175 billion** parameters): 1287MW, 552 tons [1,2]
- 43 cars or 24 US families / year





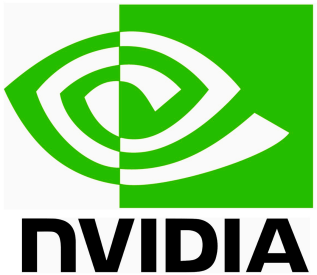
[1] Strubell, Emma, Ananya Ganesh, and Andrew McCallum. "Energy and Policy Considerations for Deep Learning in NLP." In the 57th Annual Meeting of the Association for Computational Linguistics (ACL). Florence, Italy. July 2019

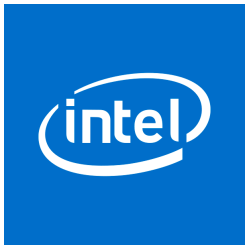
[2] David Patterson et al., "Carbon emissions and large neural network training", Google Research, April 2021

Edge AI Chipset

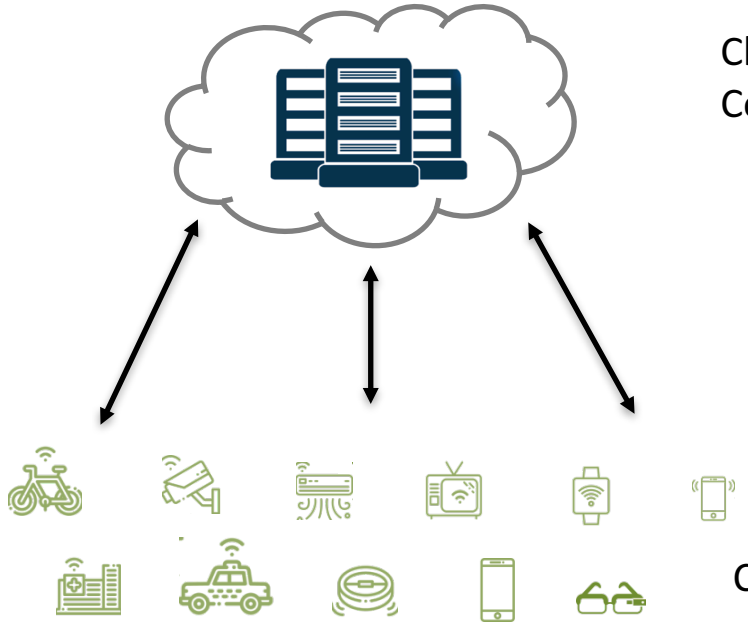
HPI Hasso Plattner Institut
IT Systems Engineering | Universität Potsdam

- According to the report from ABI Research, the revenue of the edge AI chipset market will reach **12.2 billion** U.S. dollars in 2025, while the revenue of the cloud AI chipset market will reach **11.9 billion** U.S. dollars.
- According to Gartner's forecast, by 2020, the number of global IoT devices will exceed 20 billion.
- Major players in the market including *Google, NVIDIA, Intel, Qualcomm, Huawei, Xilinx* etc.





Edge AI



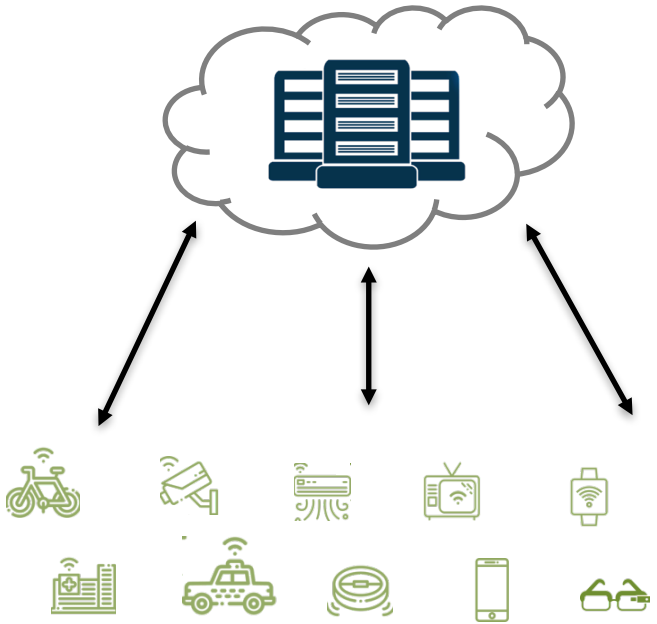
Cloud:
Centralized

Client devices

AI applications

- Limited privacy protection
- Limited interaction
- Limited connections

Edge AI



Big Model

Cloud:
Centralized

small model

Client devices

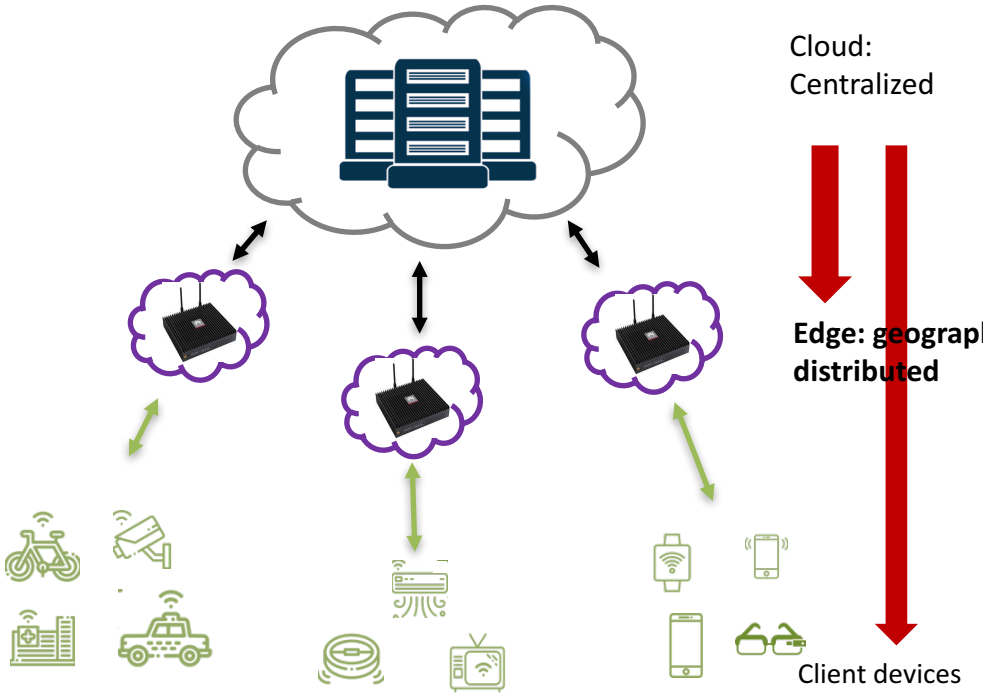
AI applications

- Privacy protection
- Interaction
- Connections

Problem

- Significant accuracy loss
- Increased development complexity
- Hard to meet the needs of massive amounts of mobile applications

Edge AI



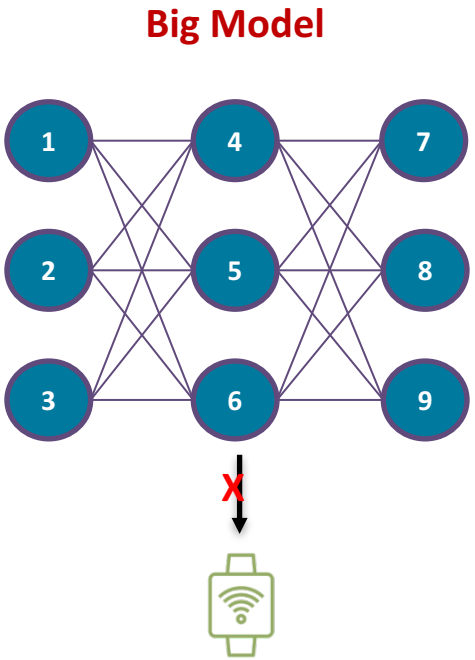
5G Era: Offloading computation

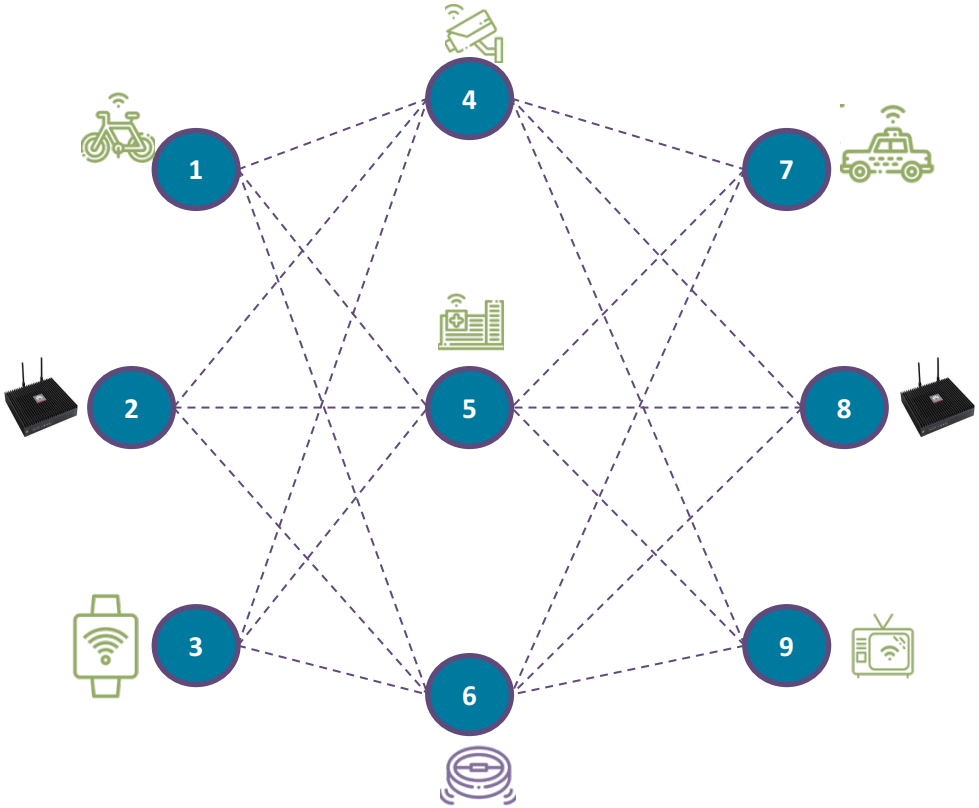
Applications
- Low latency
- Highly interactive
- Massive connections

Characteristics
- Highly efficient
- Decentralized
- Collaborative

Example

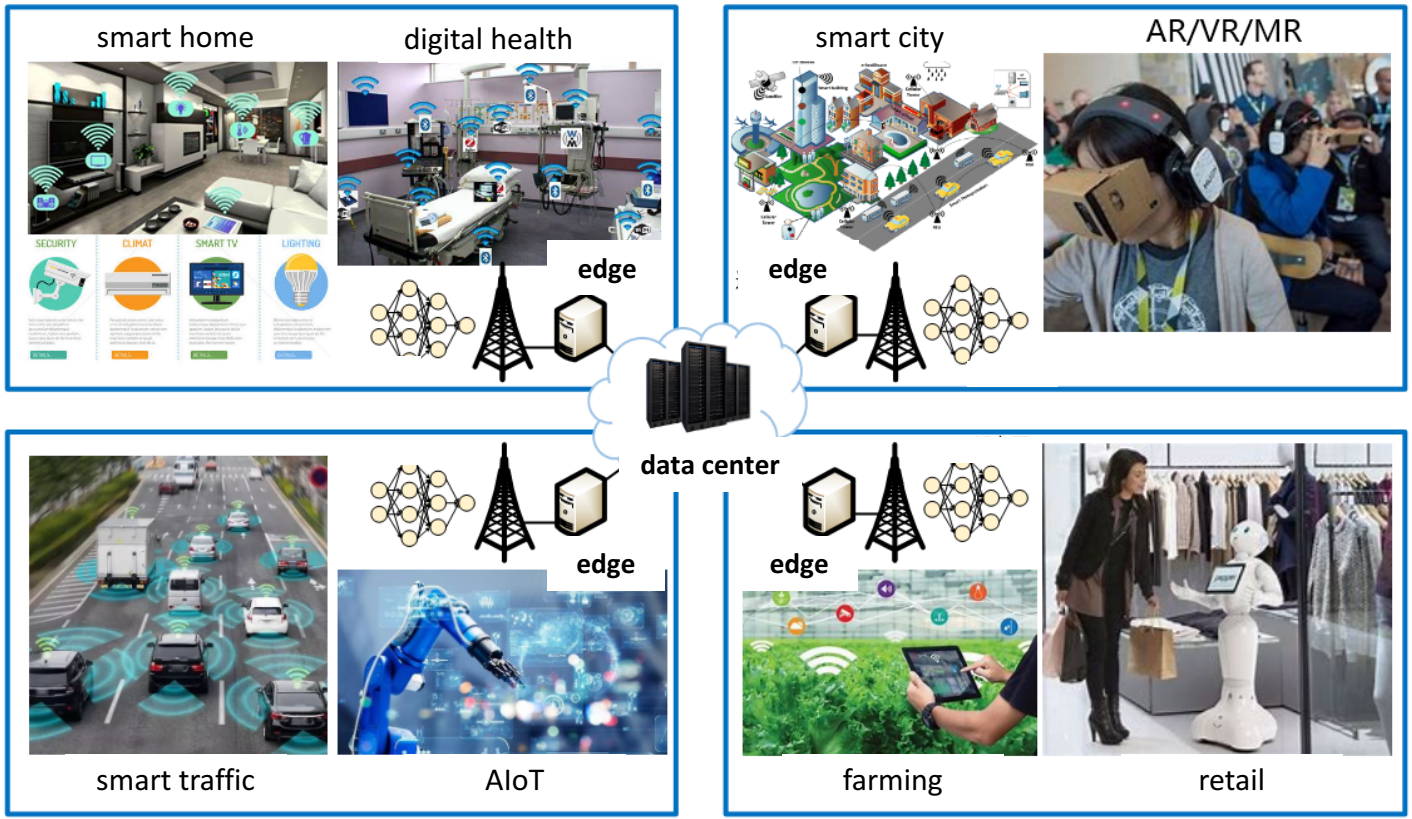
HPI Hasso Plattner Institut
IT Systems Engineering | Universität Potsdam





Edge AI Scenarios

HPI Hasso Plattner Institut
IT Systems Engineering | Universität Potsdam



Edge AI Challenges

HPI Hasso Plattner Institut
IT Systems Engineering | Universität Potsdam

Heterogeneity of Edge Nodes

- Hardware
  - GPU, x86_CPU, Arm_CPU, FPGA, ASIC …
- Operation system
  - Linux, Windows, MacOS, iOS, Android, HarmonyOS …
- AI software frameworks
  - PyTorch, Tensorflow, MXNet, MindSpore, Paddlepaddle, MegEngine …

Privacy and Security

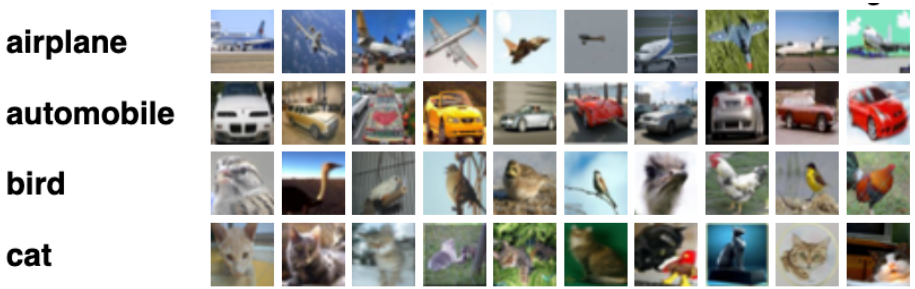
- Privacy protection
- Data silos

Edge AI Challenges

HPI Hasso Plattner Institut
IT Systems Engineering | Universität Potsdam

Heterogeneity of Data

- Heterogeneous data distribution across edge nodes
- Heterogeneous data difficulty across edge nodes
- Non-IID (**I**ndependent and **I**dentically **D**istributed)



https://www.cs.toronto.edu/~kriz/cifar.html

Edge AI Challenges

HPI Hasso Plattner Institut
IT Systems Engineering | Universität Potsdam

Limited Resources

- Computing resources is limited.
- Strict power consumption requirements
- Unbalanced power supply capacity of edge nodes
- Limited storage resources
- Limited network bandwidth and relatively low stability

Unknown Classes and Few-Shot Problem

- The amount of edge data may be very small
- Model cold start often leads to poor results
- Unknown categories cause the supervised learning model to fail.

Our Recent Work

HPI Hasso Plattner Institut
IT Systems Engineering | Universität Potsdam

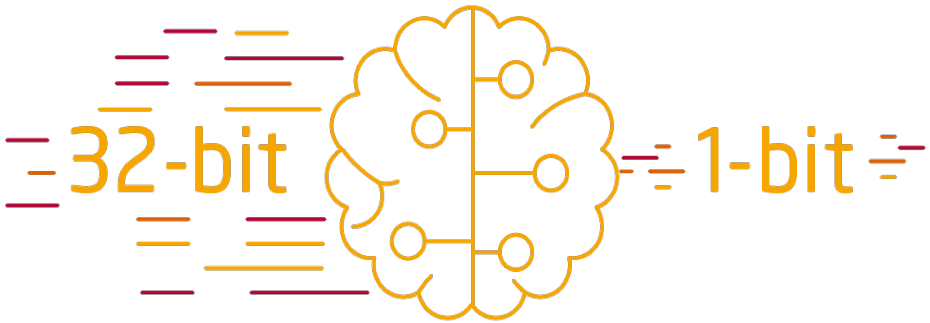
- Efficient binary model: ***BoolNet*** and ***BNext***
- Edge AI framework: *KubeEdge/**Sedna***
- Multi-Edge Inference for Object Re-Identification (ReID)

Deep Learning with **Binary Neural Networks**

HPI Hasso Plattner Institut
IT Systems Engineering | Universität Potsdam

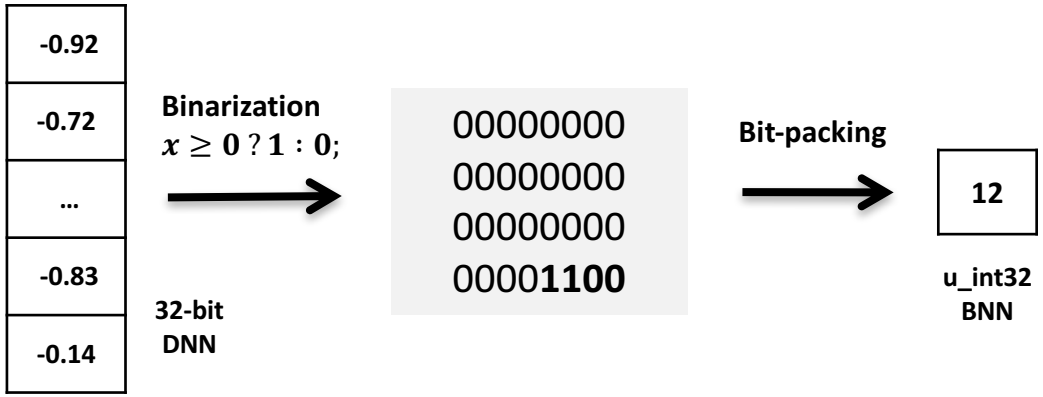
Binary Neural Networks

- State of the art deep neuronal networks are trained and operate on 32-bit models
- Design and Training of deep neuronal networks on binary-level (1-bit) is possible



Binary Neural Networks

- The extreme case **Binary Neural Networks** only use 1-bit information (1 and 0) for weights and inputs instead of 32-bit floating point numbers



- Up to 32x model compression and >58x (theoretical) speedup during inference [1]
- **More than 1000x energy saving on dedicated hardware [2]**

[1] Rastegari, Mohammad, et al. "Xnor-net: Imagenet Classification using Binary Convolutional Neural Networks." European conference on computer vision. Springer, Cham, 2016.
[2] Mishra, Asit, et al. "WRPN: Wide Reduced-Precision Networks." International Conference on Learning Representations. 2018.

Challenges of Binary Neural Networks

HPI Hasso Plattner Institut
IT Systems Engineering | Universität Potsdam

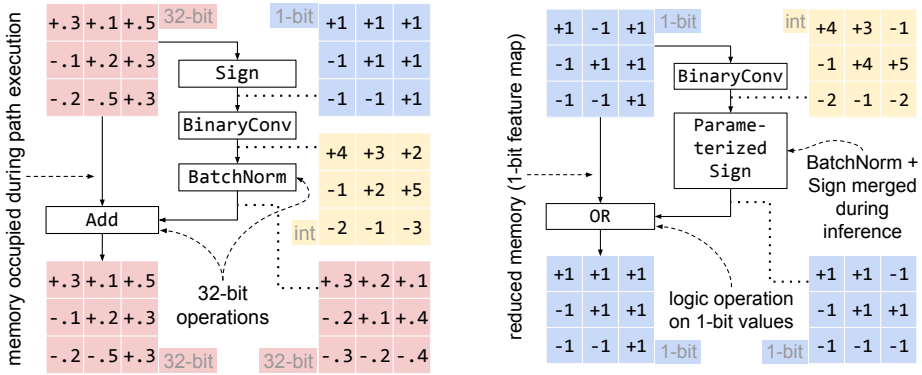
- **Loss of accuracy** compared to 32-bit networks
  - For example, directly binarizing a network trained on ImageNet, leads to an accuracy loss of about 10% [1]
- How do we build BNN's tailor-made **optimizer**?
- Balancing accuracy and **energy consumption** for AI accelerators
- Lack of support for **solid inference acceleration** on heterogeneous hardware

[1] Bethge, Joseph, et al. "BinaryDenseNet: Developing an Architecture for Binary Neural Networks." ICCVW. 2019.

Balancing Accuracy and Energy Consumption

HPI Hasso Plattner Institut
IT Systems Engineering | Universität Potsdam

- BoolNet paper studied the energy consumption of 32-bit layers used in BNNs
- Proposes a novel architecture with minimal 32-bit components for higher efficiency
- Hardware simulation for 5 well-known BNN architectures, energy consumption evaluation



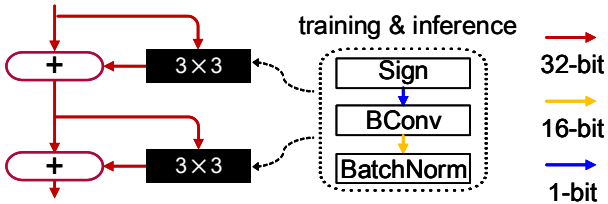
(a) Design in previous work.

(b) BoolNet design.

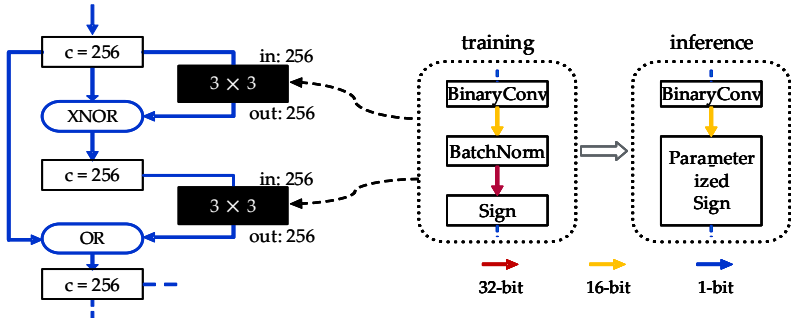
Guo N, Joseph B, et al. "BoolNet: Streamlining Binary Neural Networks Using Binary Feature Maps." arXiv preprint arXiv:2106.06991 (2021).

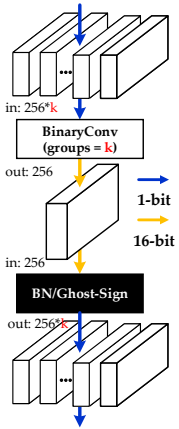
Balancing Accuracy and Energy Consumption

HPI Hasso Plattner Institut
IT Systems Engineering | Universität Potsdam

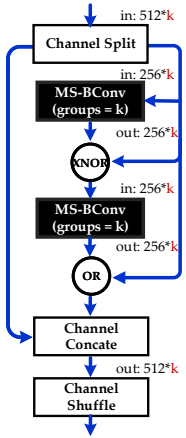


(a) Typical binary basic block.

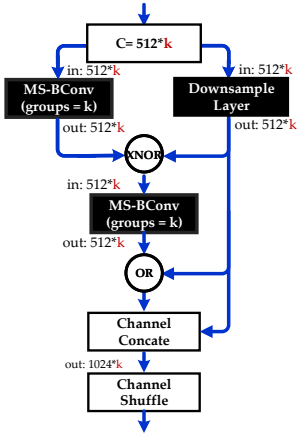




(a) Multi-Slices binary convolution



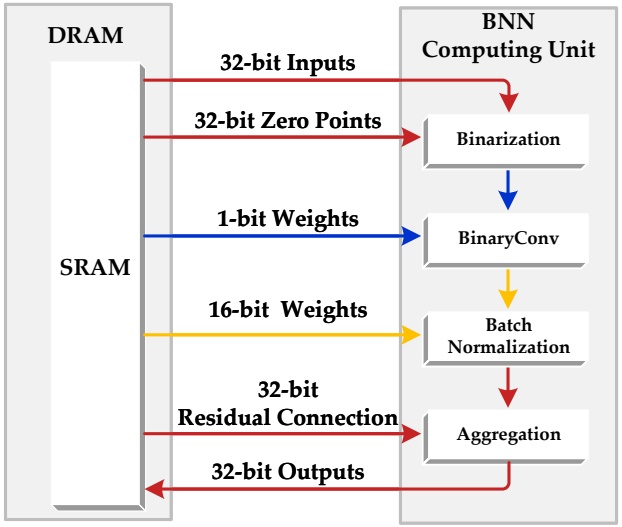
(b) BoolNet basic block



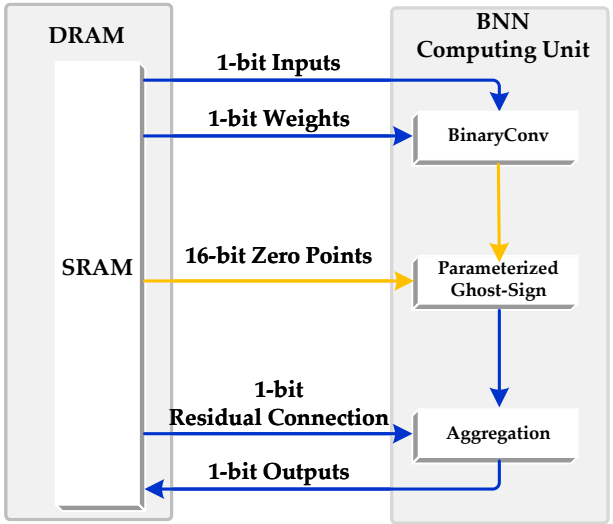
(c) BoolNet downsample block

Balancing Accuracy and Energy Consumption

HPI Hasso Plattner Institut
IT Systems Engineering | Universität Potsdam



(a) Bi-RealNet Data Flow on Hardware



(b) BoolNet Data Flow on Hardware

Figure 8: Hardware data flow comparison between Bi-RealNet and BoolNet.

Simulation Details

HPI Hasso Plattner Institut
IT Systems Engineering | Universität Potsdam

- Hardware simulation for *XNOR-Net, BiRealNet, ReActNet, our BaseNet* and *BoolNet*
- The power and area of computing circuits are given by Design Compiler (DC) with a TSMC 65nm process and 1GHz clock frequency.
- Design and implementation refer to *Conti et al. (2018)* and *Zhang et al. (2021).*
- Aligned design of architecture, data stream, the parallelism of computing units, and total on-chip cache (192KB for feature maps and 288KB for weights)
- For all accelerators: The parallelism of binary convolution is 64x64, while the parallelism of other units is 64.
- The performance depends on the parallelism and is bounded by the convolution. Each accelerator has the same peak performance for convolution, i.e., 4096 GOPs/s

Energy Consumption Evaluation

HPI Hasso Plattner Institut
IT Systems Engineering | Universität Potsdam

- DC provides static power (P_S) and dynamic power (P_D)
- For each layer, according to the amount (A)of each operation and the circuit parallelism (P_a), the number of cycles is: $C_n = A/P_a$ and energy consumption is

$$E_C = C_n * (P_S + P_D)/10^{-9}$$

- For operations with fewer cycles, E estimated by static power:

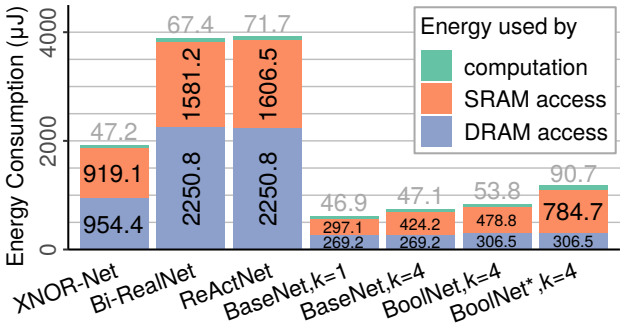
$$E_S = (C_n^{max} - C_n) * P_S/10^{-9}$$

$$E_{total} = E_C + E_S$$

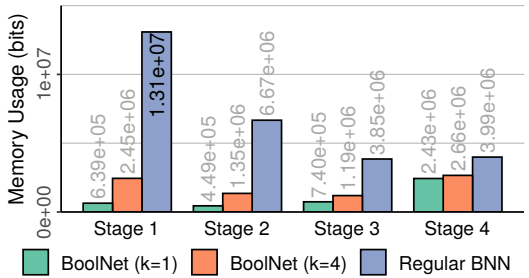
- We evaluate the energy consumption of on-chip SRAM access and off-chip DRAM access by using CACTI 6.5 (CACTI) and the power calculator of DDR provided by Micron (Micron).

Balancing Accuracy and Energy Consumption

HPI Hasso Plattner Institut
IT Systems Engineering | Universität Potsdam



(b) Energy consumption regarding computations and access to DRAM/SRAM.



(b) Memory usage comparison between blocks of different stages.

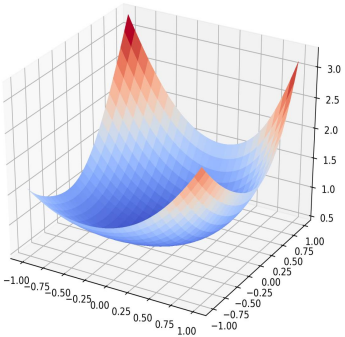
Methods	Bitwidth (W/A/F)	Energy Consumption	Top-1 Acc.	OPs ($\cdot 10^8$)
ReActNet (Bi-Real)†	1/1/32	3.93mJ	65.9%	1.63
Bi-RealNet	1/1/32	3.90mJ	56.4%	1.63
XNOR-Net	1/1/32	1.92mJ	51.2%	1.59
BoolNet, k=4 (ours)	1/1/4	1.33mJ	63.0%	1.64
BaseNet, k=4 (ours)	1/1/4	0.83mJ	58.2%	1.54
BaseNet, k=1 (ours)	1/1/1	**0.70mJ**	53.3%	1.51

Operation	Power (mw)	Area (um^2)	Operation	Power (mw)	Area (um^2)
BConv	108.8	131737	Int8 Conv(1/8)	504	836269
-	-	-	Int Agg	43.5	53238
16-bit Sign	1.4	7956	32-bit Sign	3.3	13548
32-bit RPReLU	137.6	310671	Int8 BN	50.1	274606

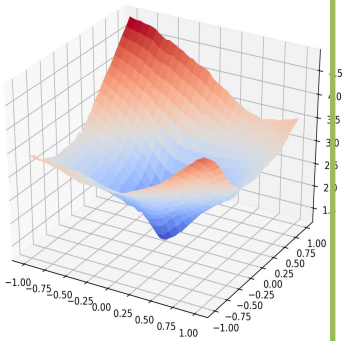
(a) Energy consumption per unit operation and circuit area of commonly used components.

Guo N, Joseph B, et al. "BoolNet: Streamlining Binary Neural Networks Using Binary Feature Maps." arXiv preprint arXiv:2106.06991 (2021).

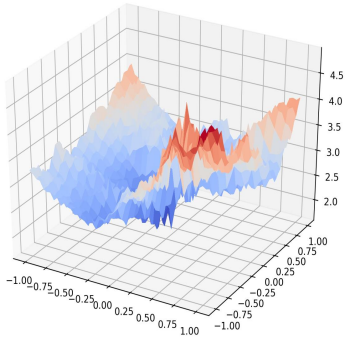
BNext - Loss Landscape Visualization



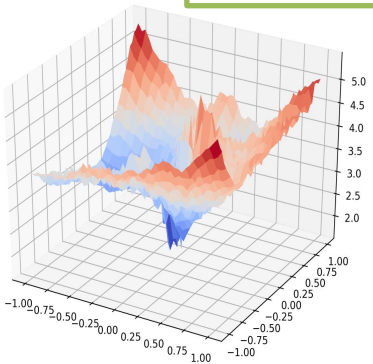
(a) ResNet-18



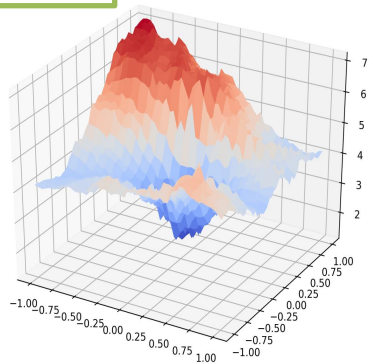
(b) BinaryConnectNet-18



(c) BinaryNet-18



(d) BiRealNet-18



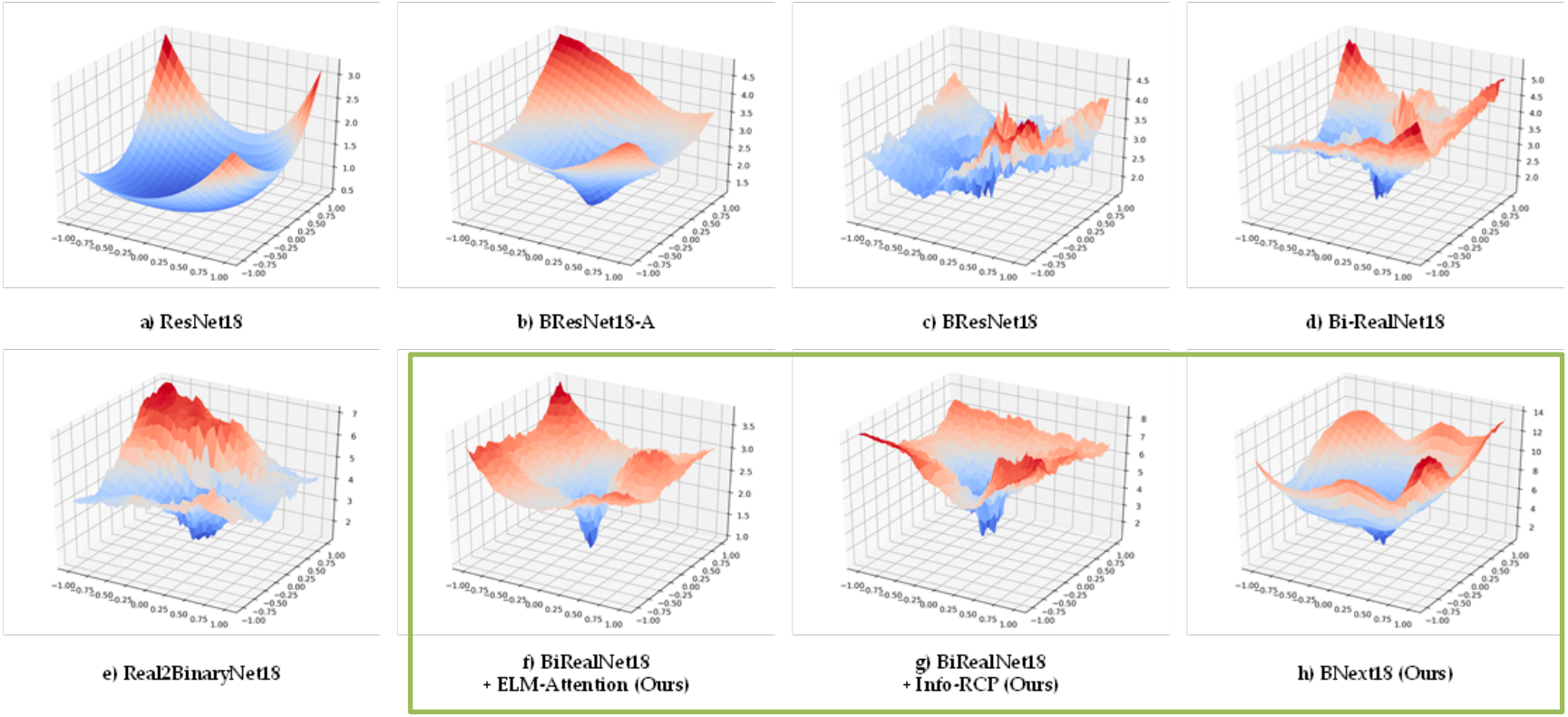
(e) Real2BinaryNet-18

HPI Hasso Plattner Institut
IT Systems Engineering | Universität Potsdam

BNext

A Novel Architecture with better Loss Landscape

HPI Hasso Plattner Institut
IT Systems Engineering | Universität Potsdam



a) ResNet18

b) BResNet18-A

c) BResNet18

d) Bi-RealNet18

e) Real2BinaryNet18

f) BiRealNet18 + ELM-Attention (Ours)

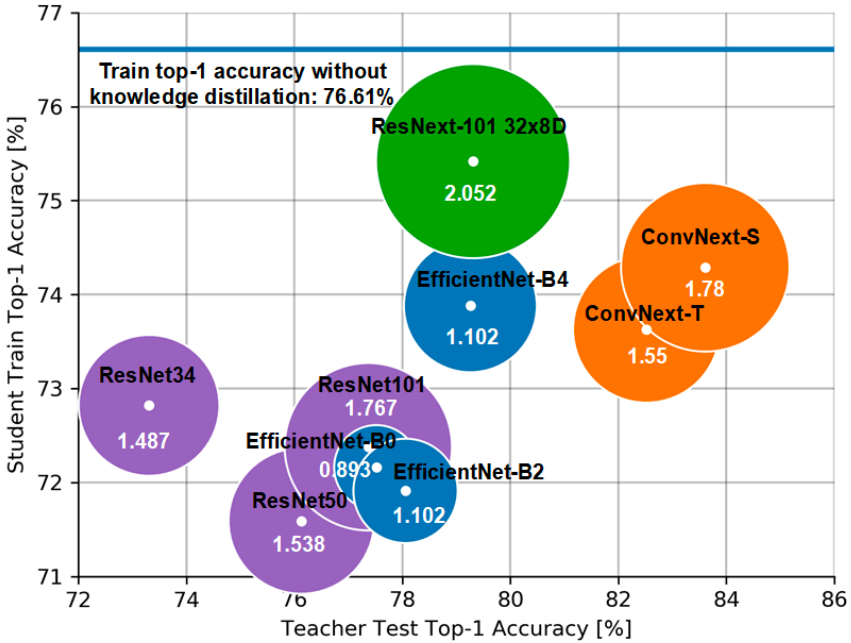
g) BiRealNet18 + Info-RCP (Ours)

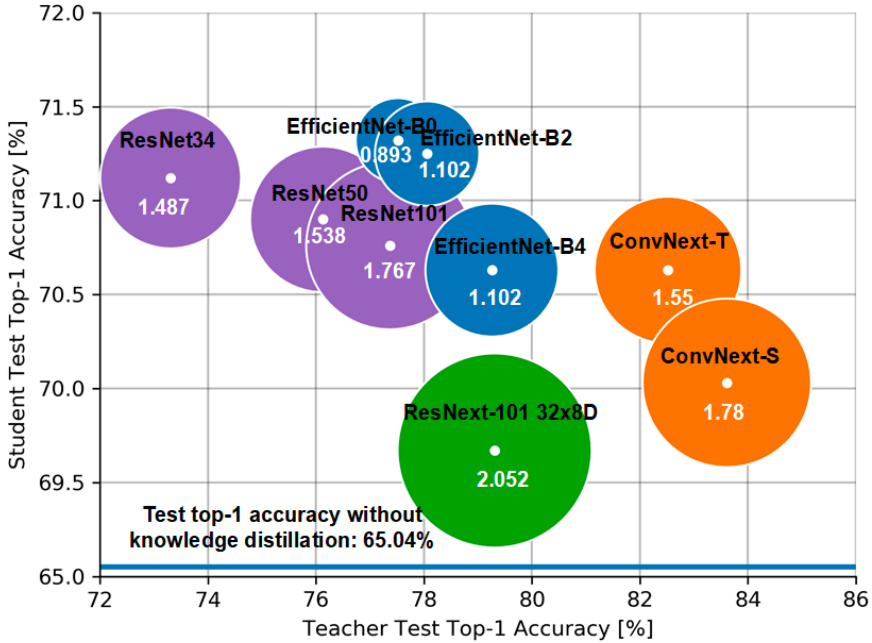
h) BNext18 (Ours)

BNext

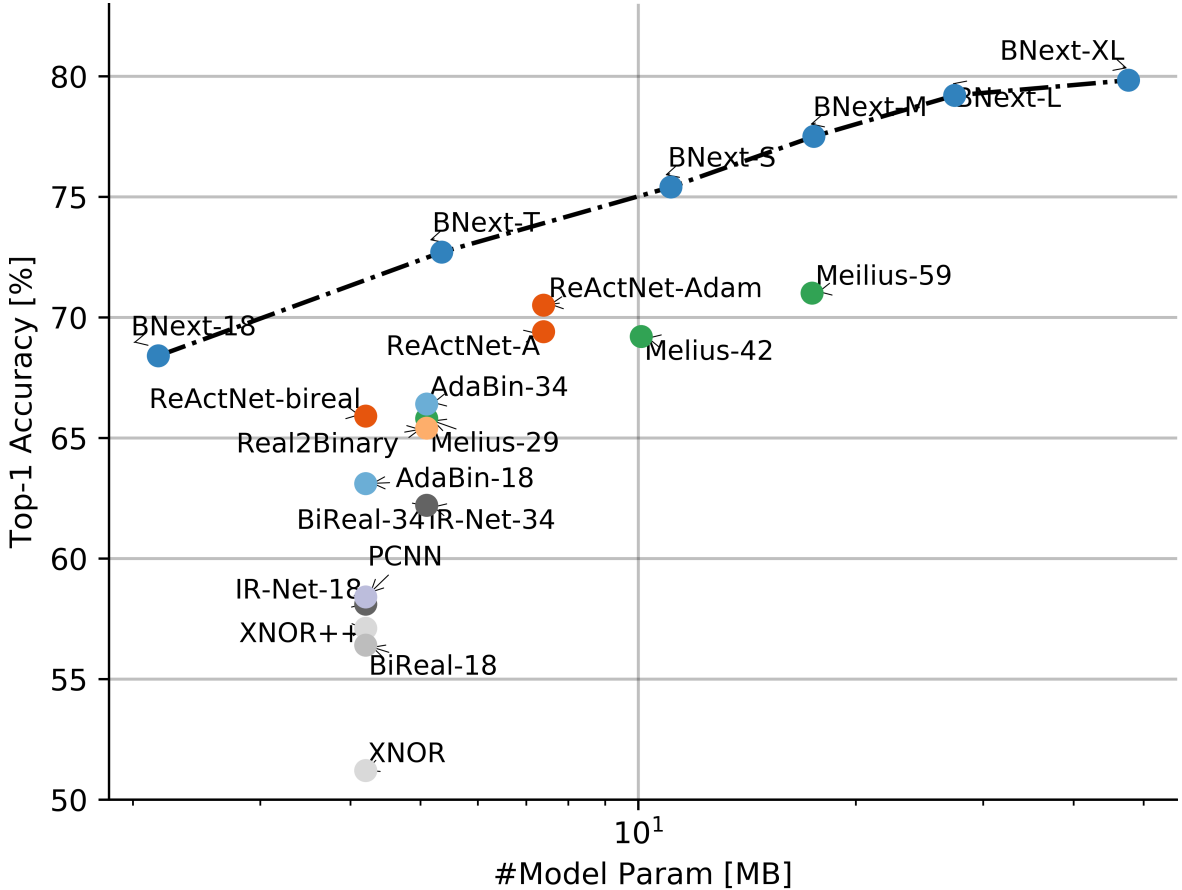
Curriculum Knowledge Distillation

HPI Hasso Plattner Institut
IT Systems Engineering | Universität Potsdam





BNext



Sedna - AI toolkit over KubeEdge

HPI Hasso Plattner Institut
IT Systems Engineering | Universität Potsdam

- An open-source toolkit incubated in KubeEdge SIG AI
- Cloud-Edge synergy capabilities provided by KubeEdge
- Actively developed community


TensorFlow

PyTorch

PaddlePaddle

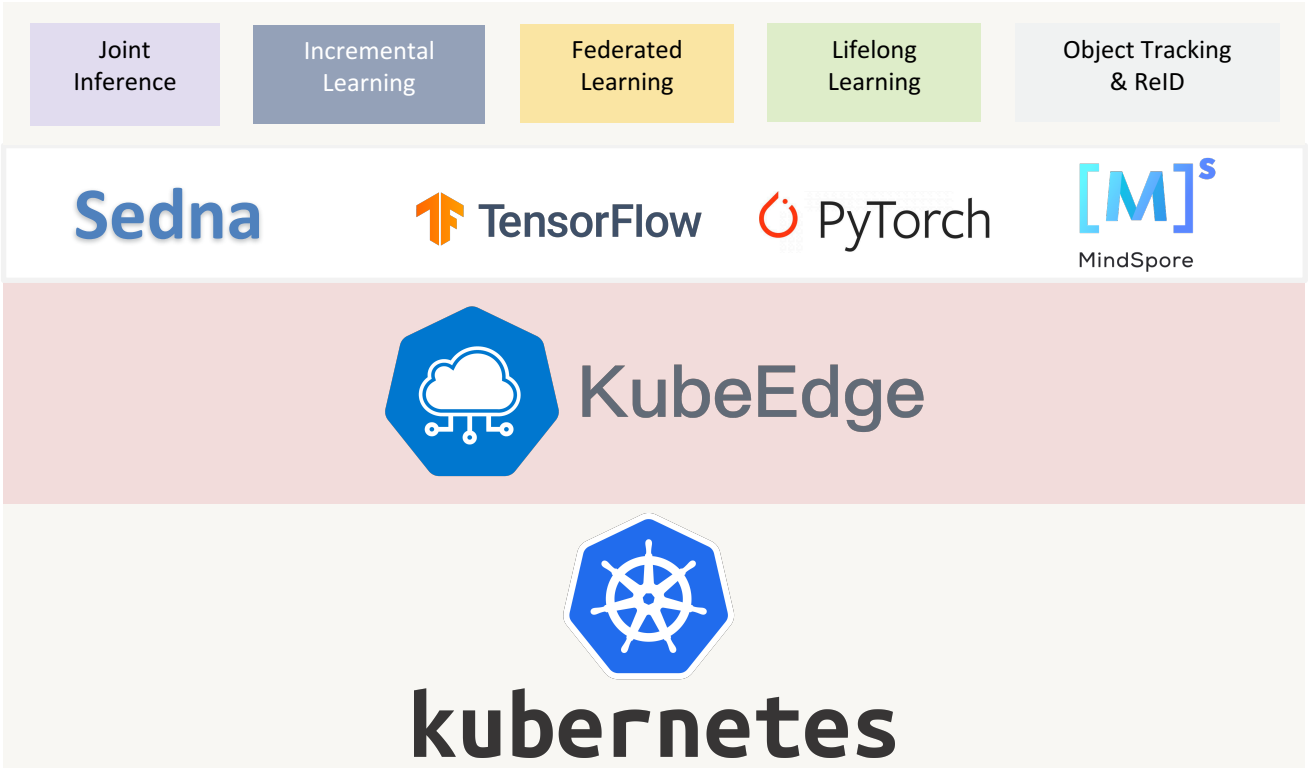
MindSpore

KubeEdge

kubernetes

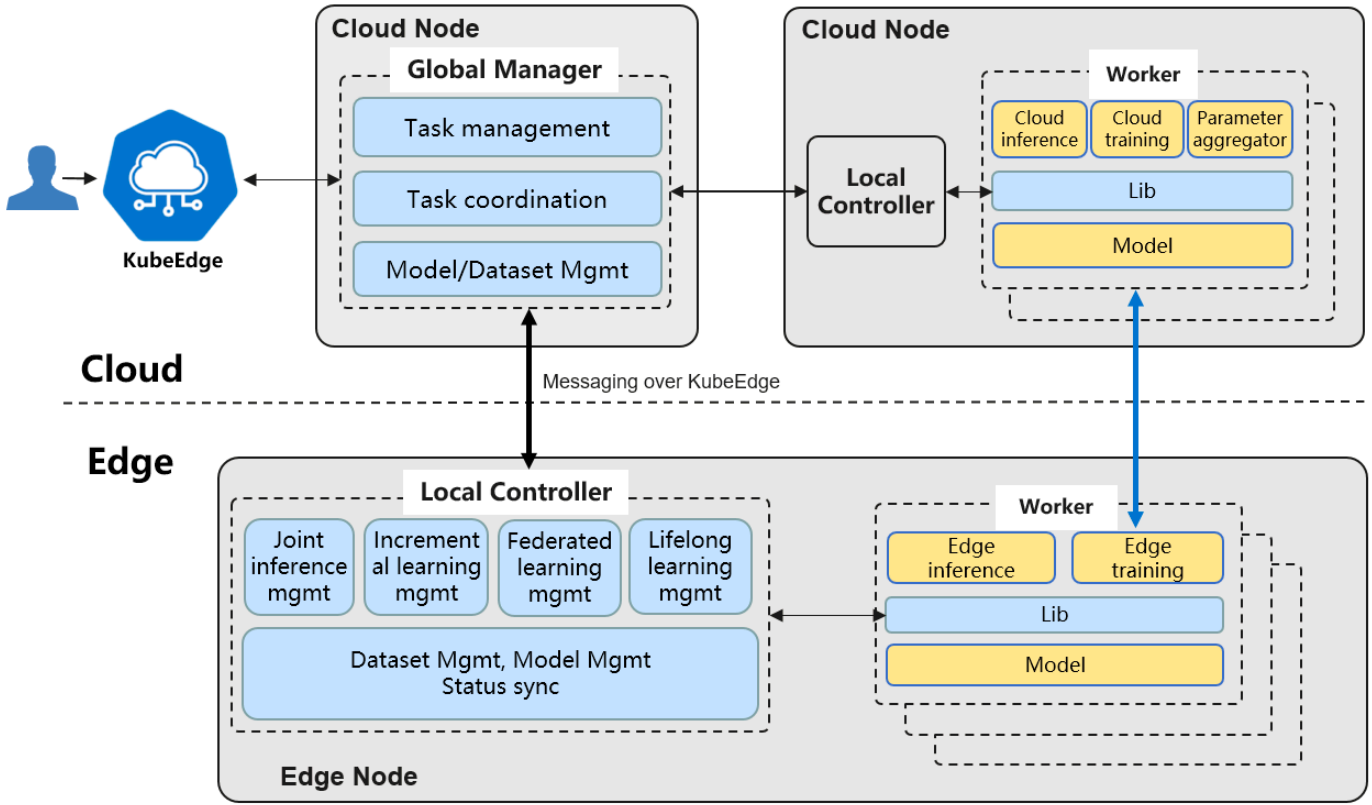
Sedna - Infrastructure

HPI Hasso Plattner Institut
IT Systems Engineering | Universität Potsdam



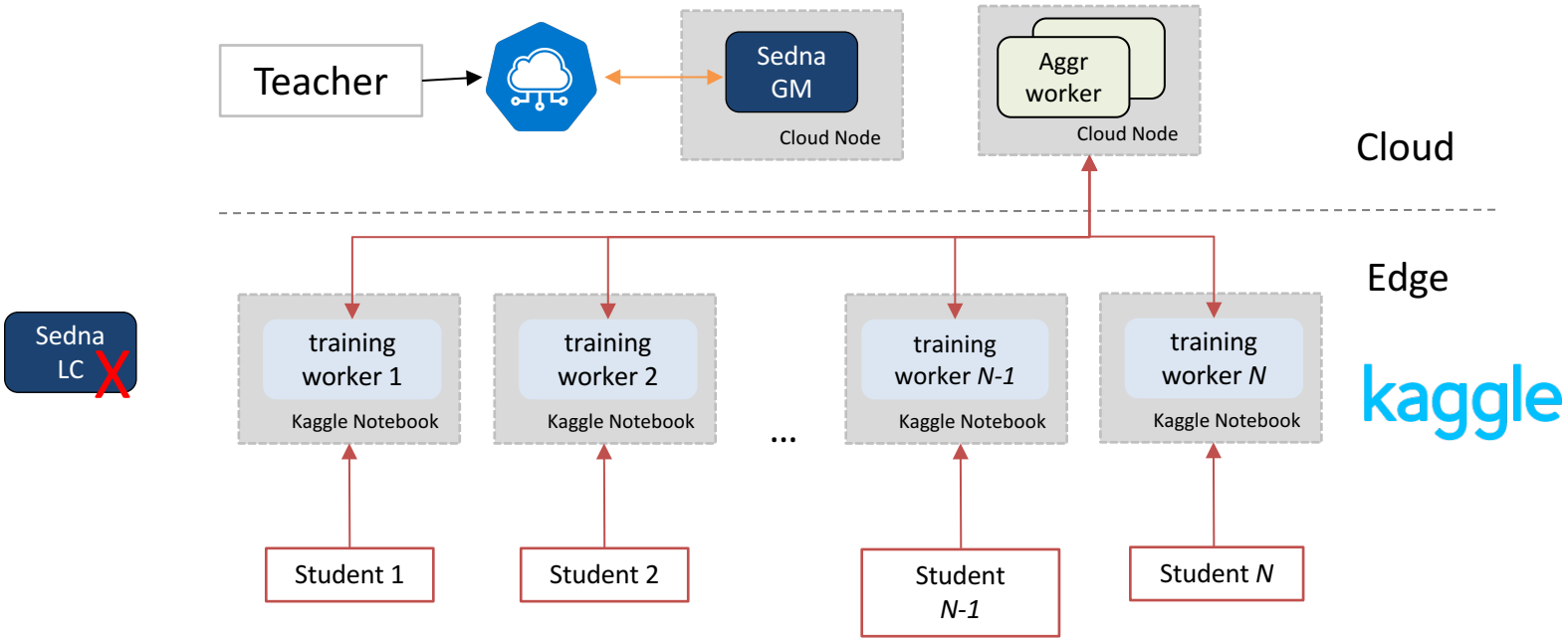
Sedna - Modules

HPI Hasso Plattner Institut
IT Systems Engineering | Universität Potsdam



Federated Learning Practical Assignment

HPI Hasso Plattner Institut
IT Systems Engineering | Universität Potsdam



https://www.kaggle.com/code/jopyth/edge-ai-w5-federated-learning/notebook

https://open.hpi.de/courses/edgeai2022

Edge AI Use Case
Multi-Edge Inference for Object Re-Identification (ReID)

HPI Hasso Plattner Institut
IT Systems Engineering | Universität Potsdam

Smart Port

Monitor the location of **shipment containers** in a smart port to boost logistic operations and allow quick retrieval of potentially lost goods.

When a shipment container enters a monitored area, it's **detected**, **recognized** and **localized.**

Smart Campus

Locate **pedestrians** in a smart campus for occupancy monitoring and to offer personalized services.

When a pedestrian enters a monitored area, it's **detected**, **recognized** and **localized.**

HUAWEI

Edge AI Use Case
Multi-Edge Inference for Object Re-Identification (ReID)

Hasso Plattner Institut
IT Systems Engineering | Universität Potsdam

Our Research
We focus on solving the challenges of managing, deploying, and monitor AI applications on an heterogenous infrastructure. To do so, we build on top of well known components such as Kubernetes, Kubeedge, and Sedna.

Edge-Cloud Synergy
We optimize the use of edge and cloud resources to support flexible deployment of AI workloads based on user requirements.

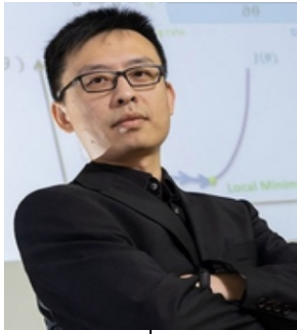
AI Applications
Supports different classes of AI applications such as object detection, classification, re-identification (ReID).

Customizable
Fully customizable by offering ML practitioners the ability to add new models to the framework and creating new ML workloads on the fly, from scratch.

Open Source
Based on fully open-source projects, our code is aswell fully available on GitHub.

Vittorio Cozzolino
PhD, Senior Software Engineer
Huawei MRC

Zi Yang
PhD Candidate
Huawei MRC



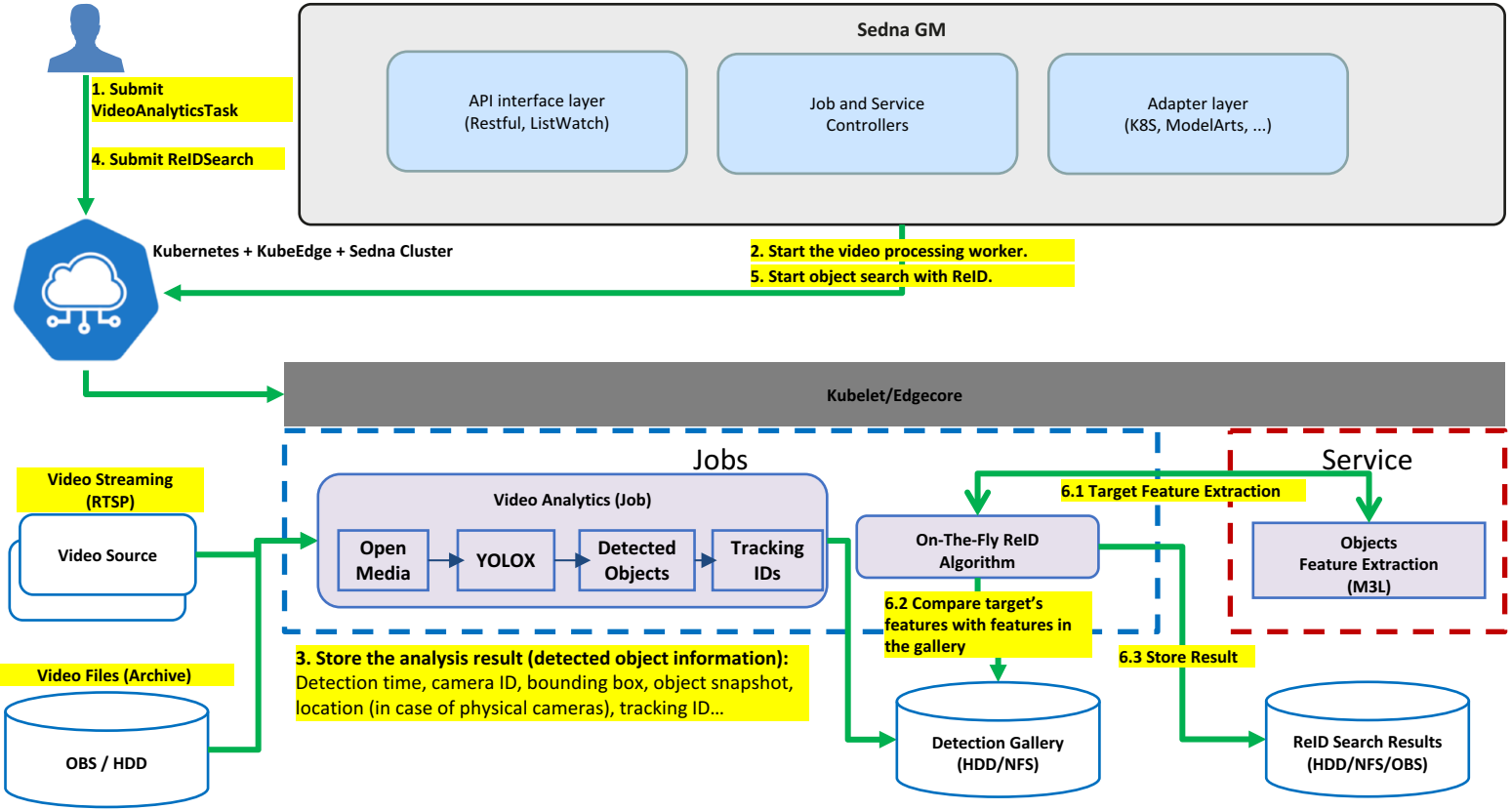
Haojin Yang
PhD, Scientific Advisor
Hasso Plattner Institute

Soumajit Majumder
PhD, Senior AI Researcher
Huawei MRC

Jorge Cardoso
PhD, Chief Architect
Huawei MRC

Multi-Edge ReID - System Design

Hasso Plattner Institut
IT Systems Engineering | Universität Potsdam



Multi-Edge ReID – AI Algorithms

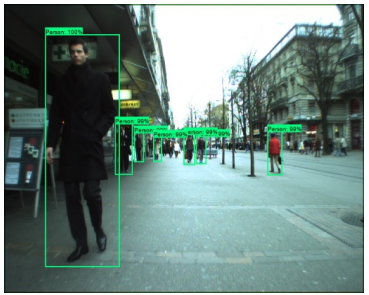
HPI Hasso Plattner Institut
IT Systems Engineering | Universität Potsdam

Object Detection and Tracking

Task of localizing and tracking (multiple) object across a video sequence.

- Domain generalization ability – No data necessary from user/customer end.
- Low latency. Detection and tracking ~45 fps.
- State-of-the-art performance on benchmark datasets.

Vehicle Detection



Pedestrian Detection

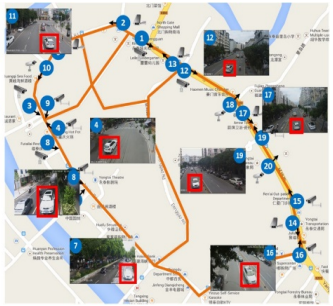
Object Re-identification (ReID)

Task of identifying the same object across multiple viewpoints.

- Discriminative feature extraction across varying viewpoints.
- High Recall ~93%.
- State-of-the-art performance among Domain Generalized ReID methods.
- On-the-fly reidentification without the need to create a features gallery in advance.

Probe vehicle image

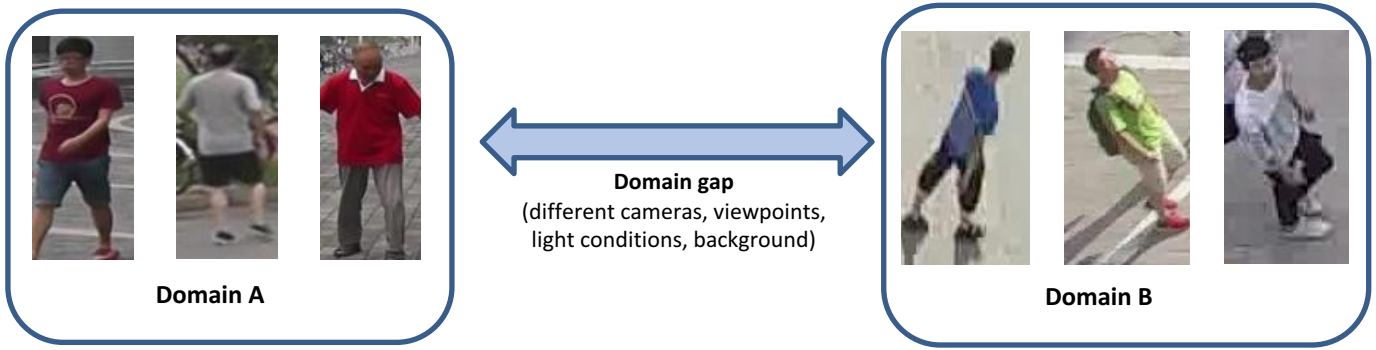
Vehicle Re-Identification



Vehicle Re-identification

DG ReID

Hasso Plattner Institut
IT Systems Engineering | Universität Potsdam



Model	Unseen test sets	Avg. mAP	Avg. Rank@1
SOTA Classic ReID model [1]	GRID, prid-2011, iLIDS, VIPeR	39.4%	30.5%
SOTA domain generalizable model (M3L) [2]	GRID, prid-2011, iLIDS, VIPeR	61.7% (+22.3%)	52.2% (+21.7%)
Our improved DG model	GRID, prid-2011, iLIDS, VIPeR	**77.4%** (+15.7%)	**70.0%** (+17.8%)

Performance on open source datasets

- Apart from evaluating our models on open-source datasets, we also tested them in **complex** and **crowded real-world** scenarios, e.g., warehouse, canteen, lobby, office. The accuracy in most scenarios can even reach **90%**.
- Because of good performance, DG ReID models are being **deployed** to **real-world** applications, e.g., COVID epidemiologic survey, smart ports

Performance in real world applications

Example

HPI Hasso Plattner Institut
IT Systems Engineering | Universität Potsdam

Query Image

ReID Results

Key Contributions

- A new computing model to support execution of distributed AI pipelines involving multiple edge nodes with different responsibilities.
- A new example using our compute model combined with AI to realize distributed object ReID.
- A complete tutorial to let users try out this new feature (https://github.com/kubeedge/sedna/blob/main/examples/multiedgeinference/pedestrian_tracking/README.md)

Video from: https://www.pexels.com/video/two-people-standing-by-the-pedestrian-crossing-while-talking-7263169/

Address:

Hasso-Plattner-Institut für Digital Engineering gGmbH, Prof.-Dr.-Helmert-Str. 2-3 D-14482 Potsdam, Germany

Email: **haojin.yang@hpi.de**

Web: