Canadian AI 2025

Analytics Everywhere Lab AT THE UNIVERSITY OF NEW BRUNSWICK

XGD: Explainable AI-Guided Knowledge Distillation with Feature Refinement for Semantic Segmentation

Hung Nguyen[1], Binh Truong[2], Khanh Nguyen[2], Khang Nguyen[2], Loc Nguyen[4], Francis Palma[3], Hung Cao[1]

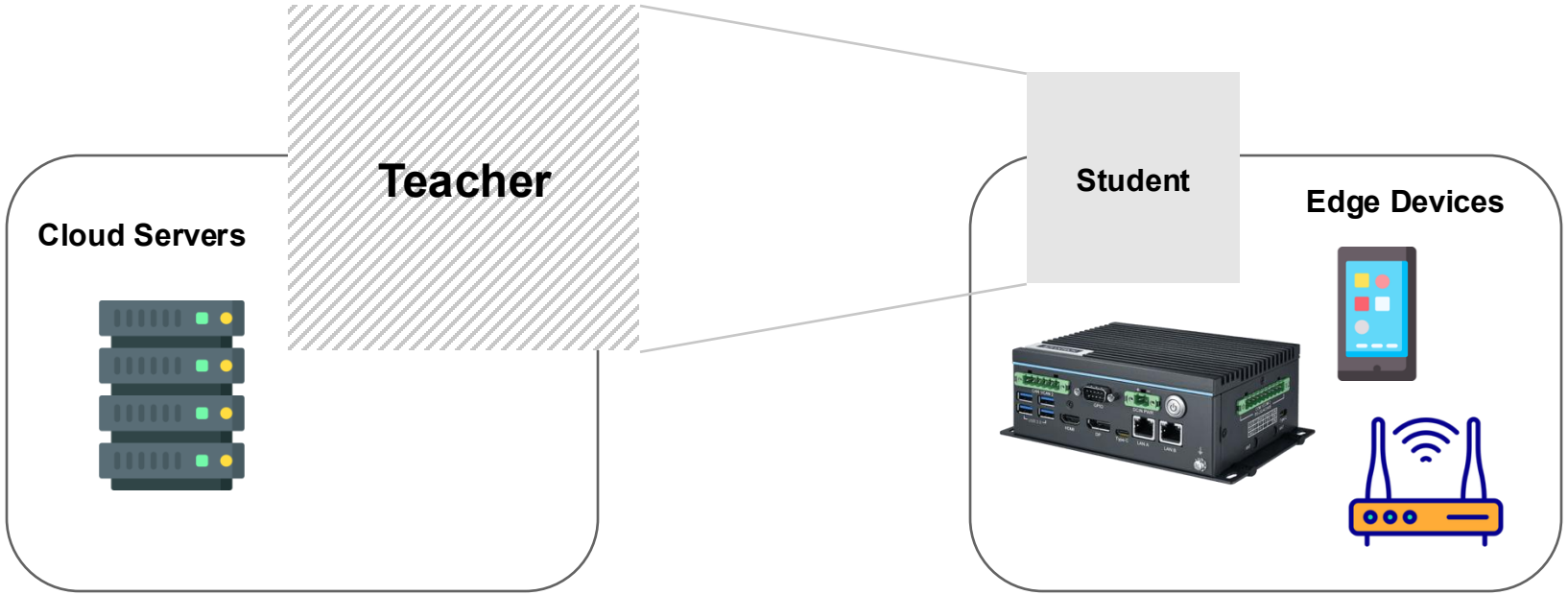
[1]Analytics Everywhere Lab, University of New Brunswick, Canada
[2]Quy Nhon AI, FPT Software, Vietnam
[3]SE+AI Lab, University of New Brunswick, Canada
[4]Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany

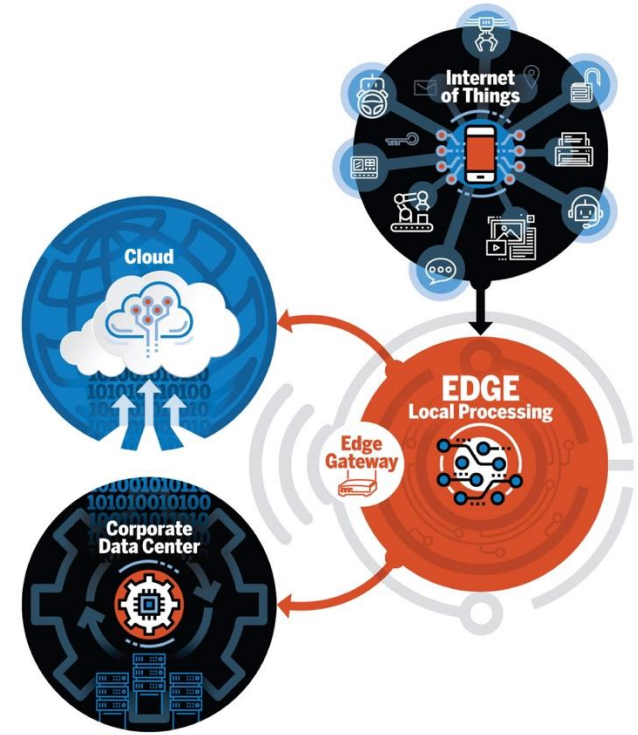
Table of Content

Motivation

Knowledge Distillation?

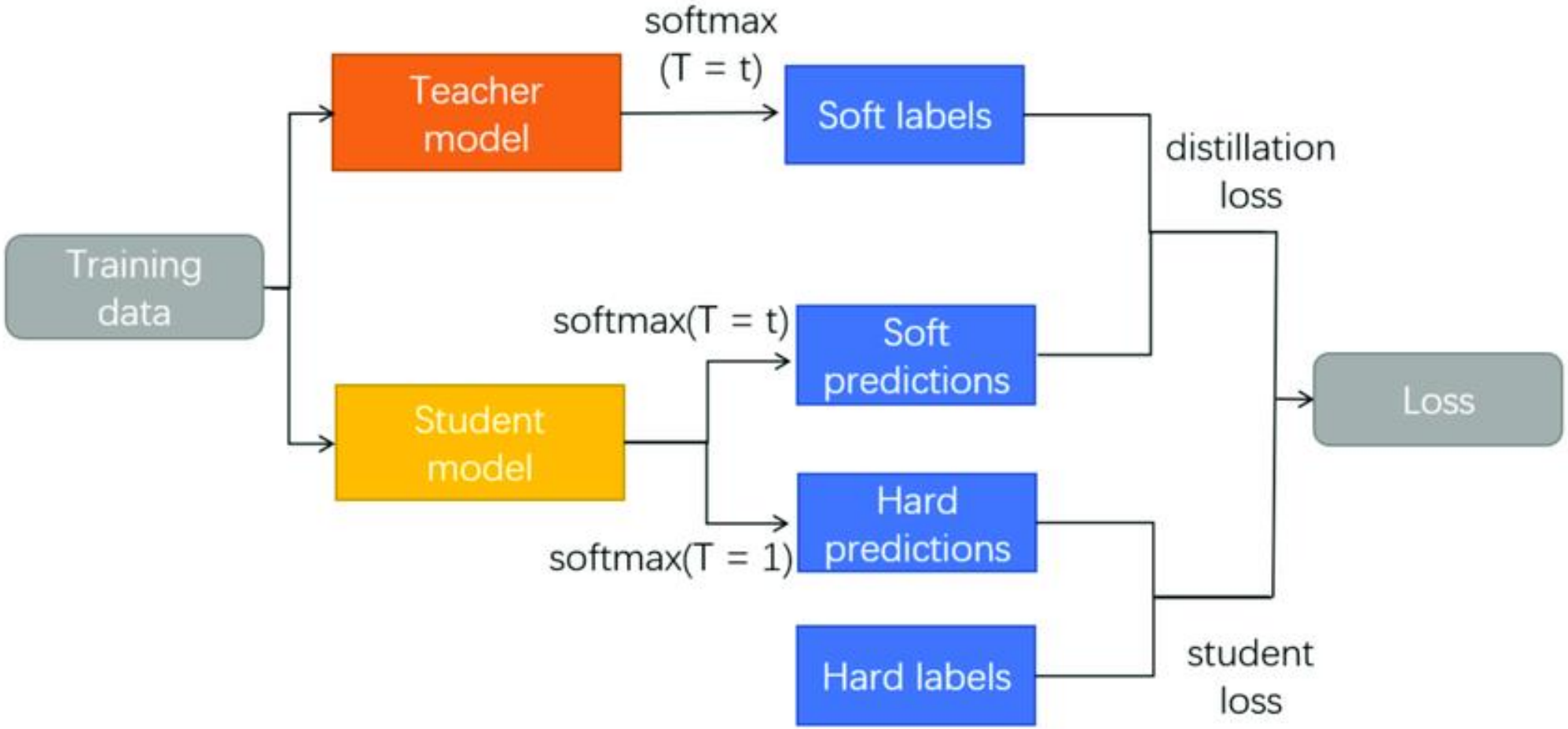
- **Knowledge distillation** transfers knowledge from large, complex "teacher" models to smaller, efficient "student" models by training the student to mimic the teacher's outputs.
- This technique is crucial for deploying AI on resource-constrained edge devices like smartphones and IoT hardware, where large models are too computationally expensive to run.





https://www.orientsoftware.com/blog/edge-computing/

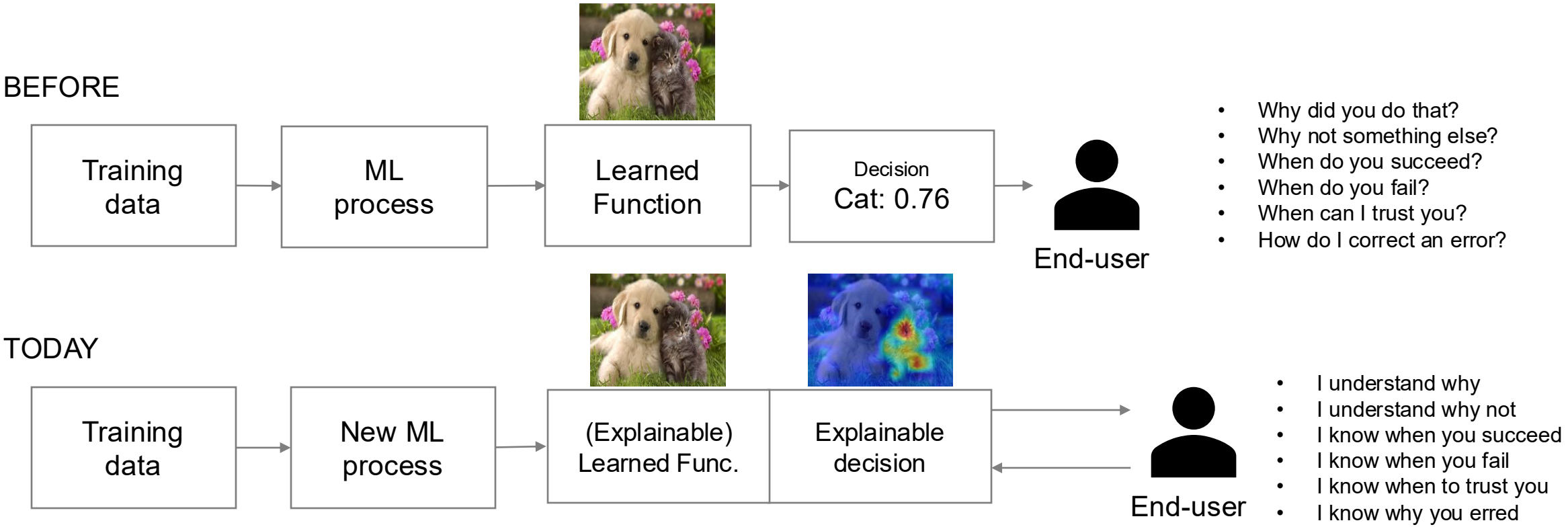
Knowledge Distillation with Soft Labels



Hinton, G. (2015). Distilling the Knowledge in a Neural Network. *arXiv preprint arXiv:1503.02531*.

What is Explainable AI (XAI)?

- XAI is a set of methods that make AI algorithms understandable and transparent to humans.
- Now, XAI explains the reasoning behind AI decisions **(why and when?)**



Going Beyond XAI

XAI reveals crucial insights into AI decision-making processes:

- What features can we use to improve the model performance?
- How can we make explanations more “human-centered” for end-users?

In sensitive contexts like healthcare, the ability to **validate or challenge AI** models through explanations has become a legal requirement.

< >
Montréal Declaration
Responsible AI_
</ >

Government of Canada | Gouvernement du Canada

MENU

Canada.ca › How government works › Policies, directives, standards and guidelines

Directive on Automated Decision-Making

AI Act

GDPR

Semantic Segmentation and Explainable AI

- In our previous work, we discovered a correlation between the model's attention patterns and semantic segmentation performance.
- Models with better semantic segmentation performance exhibit more concentrated heatmaps that focus precisely on target objects.

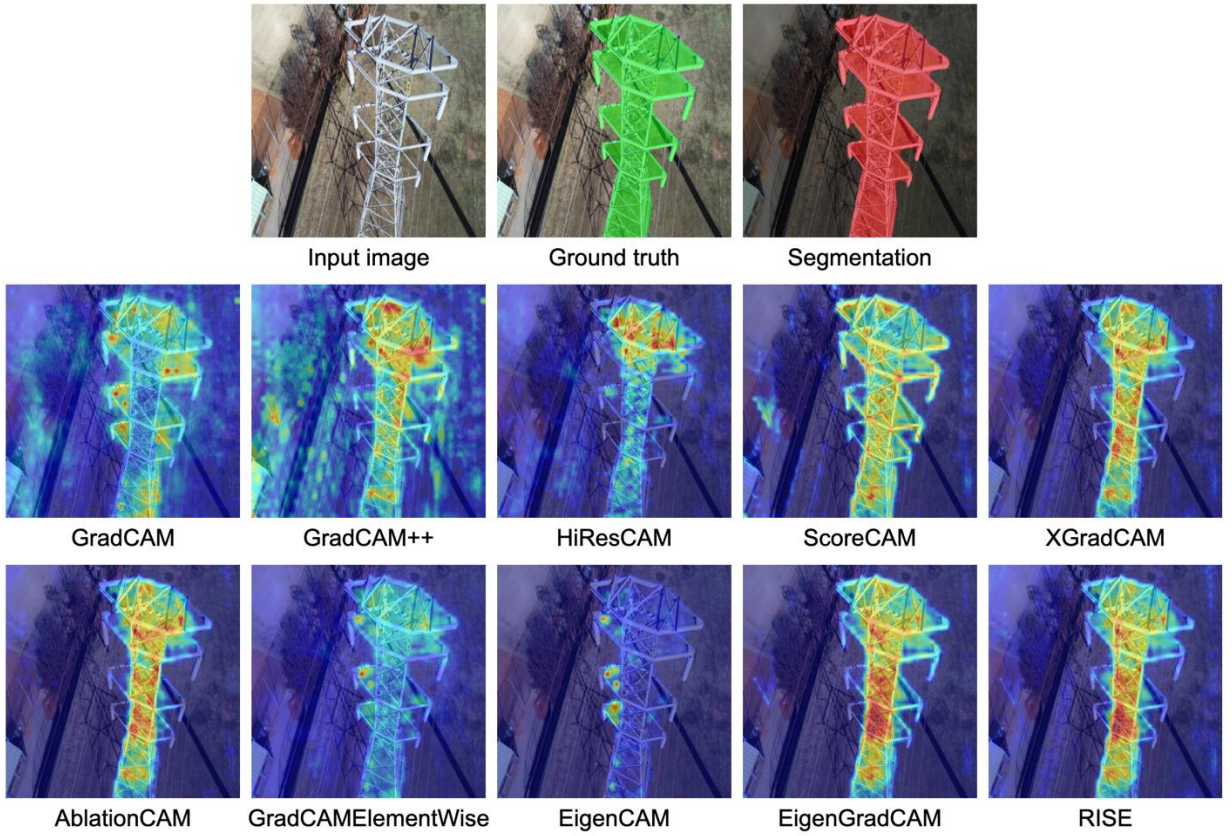


Figure 10: The qualitative evaluation of XAI methods in explaining the base DeepLabv3Plus-ResNet101 model on a validation sample. The category is the tower_lattice. The IoU value between the segmentation and the ground truth is 96.25%.

Nguyen, Hung Truong Thanh, Loc Phuc Truong Nguyen, and Hung Cao. "XEdgeAI: A human-centered industrial inspection framework with data-centric Explainable Edge AI approach." *Information Fusion* 116 (2025): 102782.

Knowledge Distillation via Attention Transfer

"Attention helps the student focus on the important aspects of the teacher's predictions."
(Zhang et al., 2020; Lee et al., 2022)

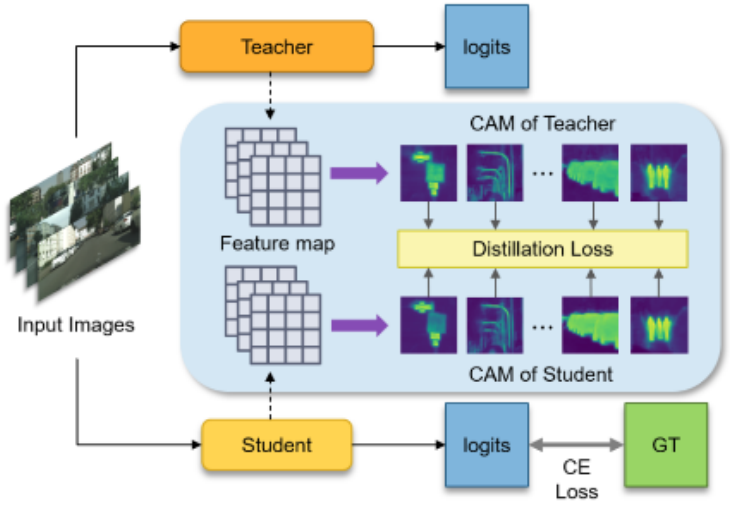


Fig. 1. The overall architecture of Class Attention Transfer. The feature map is used to generate Class Attention Maps. Only the student network is trained with the distillation loss and the task loss.

Cho, Yubin, and Sukju Kang. "Class attention transfer for semantic segmentation." *2022 IEEE 4th International Conference on Artificial Intelligence Circuits and Systems (AICAS)*. IEEE, 2022.

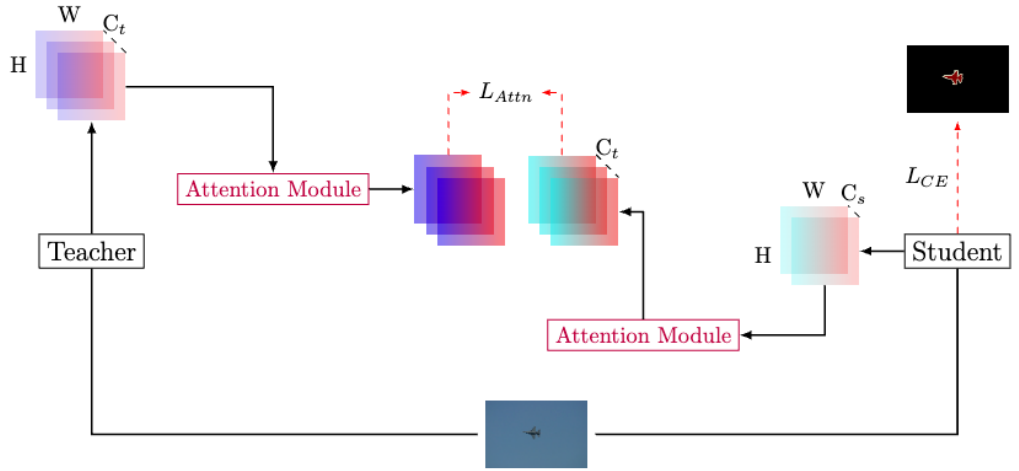


Figure 2: Proposed Attention-guided feature distillation.

Mansourian, A. M., Jalali, A., Ahmadi, R., & Kasaei, S. (2024). Attention-guided Feature Distillation for Semantic Segmentation. *arXiv preprint arXiv:2403.05451*.

Our Contributions

“Attention helps the student focus on the important aspects of the teacher’s predictions.”
(Zhang et al., 2020; Lee et al., 2022)

Semantic segmentation involves the task of classifying each pixel in an input image into one of predefined classes.

Our contributions are:

- Experiment with the soft label and intermediate saliency map (at the feature refinement level) loss using XAI methods (i.e., Grad-CAM) and propose the **XAI-Guided Knowledge Distillation (XGD)** method.
- Apply XGD to **encoder-decoder-based semantic segmentation models** (e.g., DeepLabV3+, PSPNet) on Pascal VOC 2012, Substation, and TTPLA datasets.

Analytics Everywhere Lab
AT THE UNIVERSITY OF NEW BRUNSWICK

Explainable AI-Guided Knowledge Distillation (XGD)

XGD Architecture

3 Loss Components

1. **Semantic Segmentation Loss:** Cross-entropy loss for pixel-wise classification

2. **Pixel-wise Class Probability Distillation:** KL divergence between teacher and student soft probability distributions

3. **XAI-based Saliency Map Distillation:** MSE between GradCAM saliency maps from teacher and student first decoder layers

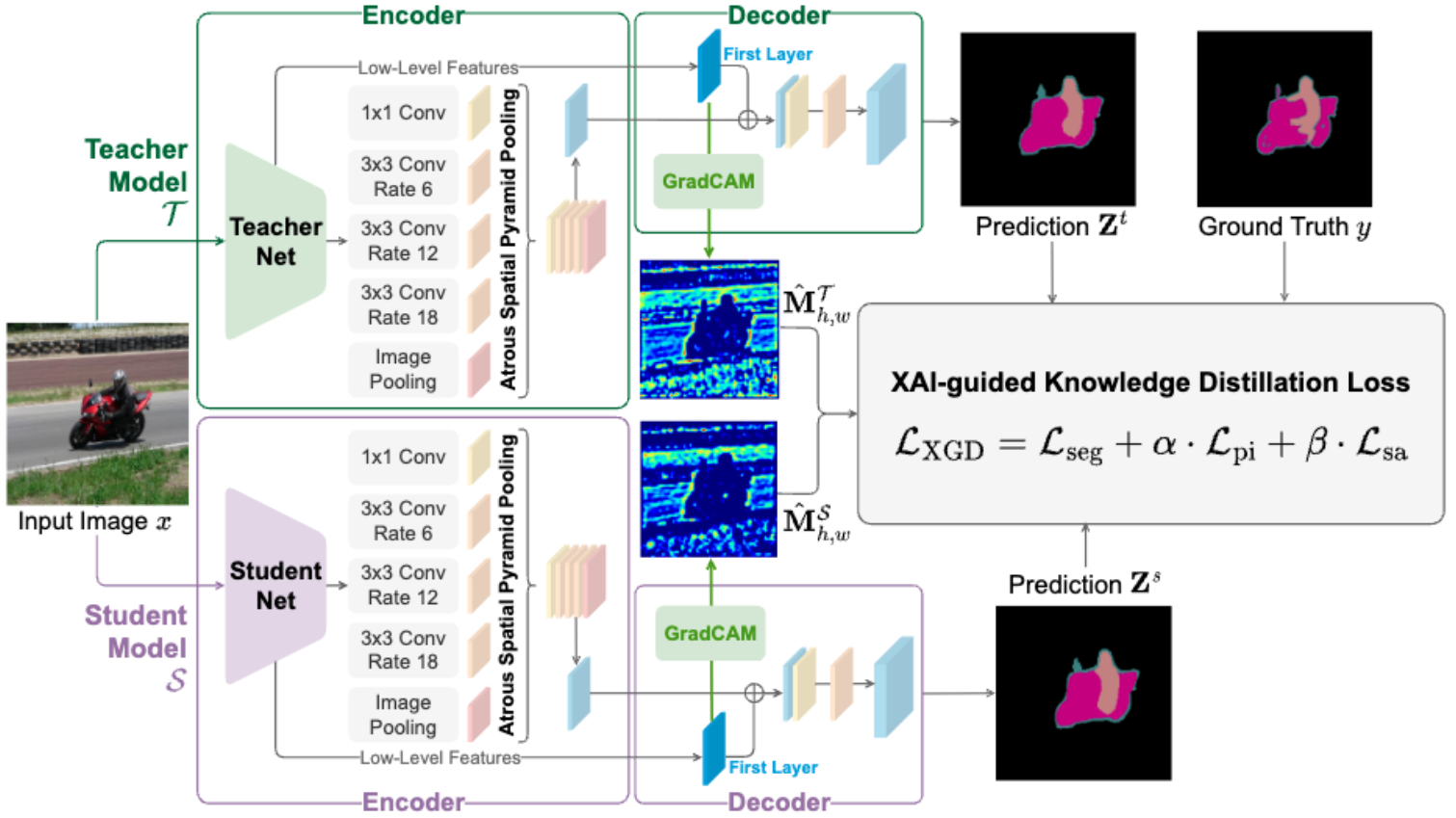


Figure 1. Overview of our proposed XAI-Guided Knowledge Distillation with DeepLabV3+.

Semantic Segmentation Loss

- Primary learning objective for pixel-wise classification to ensure student network learns core segmentation task
- Computes cross-entropy between student predictions and ground truth labels.

$$\mathcal{L}_{\text{seg}} = \frac{1}{H \times W} \sum_{h=1}^{H} \sum_{w=1}^{W} \text{CE}(\mathbf{p}^s_{h,w}, y_{h,w})$$

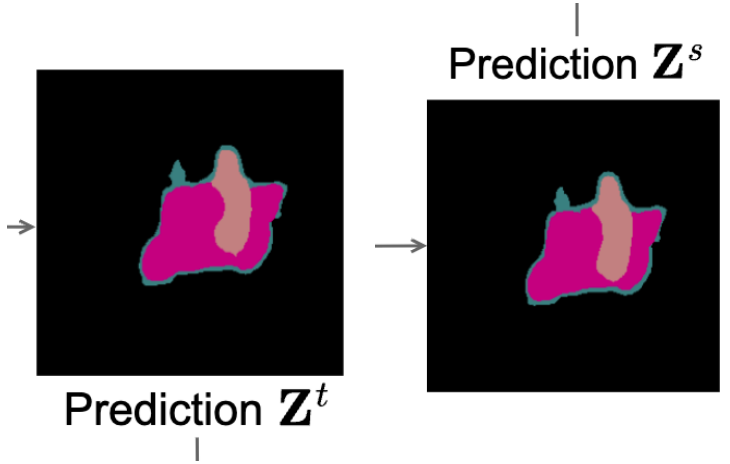
Prediction $\mathbf{Z}^s$

Ground Truth y

Pixel-wise Distillation Loss (Soft Labels)

- **The pixel-wise distillation** enables the student model to capture the teacher's knowledge about inter-class relationships.
- Minimizing the Kullback-Leibler (KL) divergence between the temperature-scaled outputs allows the student to learn the relative probabilities that the teacher assigns to different classes.

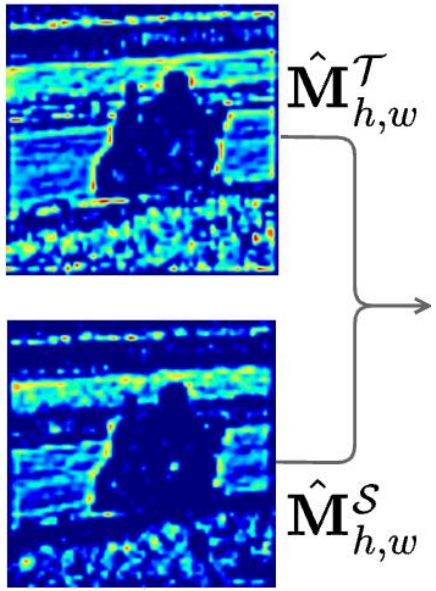
$$\mathcal{L}_{\text{pi}} = \frac{1}{H \cdot W} \sum_{h=1}^{H} \sum_{w=1}^{W} \text{KL}(\mathbf{q}_{h,w}^{\mathcal{T}} \| \mathbf{q}_{h,w}^{\mathcal{S}})$$



Saliency Map Loss

- **The saliency map loss** aligns the focus of the student and teacher models to important regions contributing to the model's prediction.
- **Calculation of Grad-CAM Saliency Maps:**
  - Grad-CAM saliency maps are computed for student and teacher models using the same target layer (i.e., the first decoder layer).
  - The first encoder layer merges high-level semantic features from the encoder with low-level spatial details

$$\mathcal{L}_{\text{sa}} = \frac{1}{H \cdot W} \sum_{h=1}^{H} \sum_{w=1}^{W} \left(\hat{\mathbf{M}}^{\mathcal{T}} - \hat{\mathbf{M}}^{\mathcal{S}} \right)^2$$



Altogether: XGD Distillation Function

The XGD distillation function represents a multi-objective optimization problem, which is defined as follows:

$$\mathcal{L}_{\text{XGD}} = \mathcal{L}_{\text{seg}} + \alpha \cdot \mathcal{L}_{\text{pi}} + \beta \cdot \mathcal{L}_{\text{sa}}$$

where $\alpha \geq 0$ and $\beta \geq 0$ are weighting coefficients for the soft label loss and saliency maps loss, respectively.

Algorithm 1: XAI-Guided Knowledge Distillation (XGD)

Input: Training data $\mathcal{D}$, teacher network $\mathcal{T}$, student network $\mathcal{S}$ with parameters $\theta_{\mathcal{S}}$, coefficients α, β, temperature τ, learning rate η, epochs E.
Output: Optimized student parameters $\theta^*_{\mathcal{S}}$.

```
for epoch = 1, ..., E do
    for (x, y) ∈ D do
        q^T ← softmax(T(x)/τ), q^S ← softmax(S(x; θ_S)/τ)        // Soft probabilities
        L_seg ← CE(q^S, y)                                        // Semantic segmentation loss
        L_pi ← KL(q^T ‖ q^S)                                      // Pixel-wise class probability distillation loss
        M^T, M^S ← GradCAM(T), GradCAM(S)                         // Saliency maps
        L_sa ← MSE(M^T, M^S)                                      // Saliency loss
        L_XGD ← L_seg + α · L_pi + β · L_sa                       // XGD loss
        θ_S ← θ_S − η∇_{θ_S} L_XGD                                // Update the student network parameters
    end
end
return θ_S
```

Experiements and Results

Quantitative Results

† Attention-guided KD

Methods	FLOPs (G)	#Params (M)	TTPLA mIoU(%)			Substation mIoU(%)			Pascal VOC mIoU(%)	
			train	*val*	*test*	*train*	*val*	*test*	*train*	*val*
T: DLV3P-R101	112.90	45.67	80.96	72.67	71.20	81.04	79.90	79.72	89.21	78.33
S: DLV3P-R18	36.88	12.33	75.81	67.61	65.25	75.64	73.69	73.95	83.07	73.39
+PI [11]			76.74	68.67	67.01	76.21	73.27	74.69	83.58	73.87
+SKD [7]			78.12	71.94	67.99	78.24	74.01	75.32	85.12	74.02
+CWD [8]			78.88	72.04	68.84	78.21	74.62	75.94	86.56	74.54
+AT [26]†			76.94	68.94	67.54	76.45	73.54	74.99	83.78	73.99
+CAT [10]†			77.56	71.49	68.24	77.23	74.42	75.12	83.92	73.65
+XGD (Ours)†			**78.95**	**72.23**	**69.82**	**78.34**	**74.94**	**76.01**	**86.74**	**74.99**
S: DLV3P-MBV2	12.30	4.38	74.93	67.32	64.41	75.09	70.81	71.92	82.96	73.31
+PI [11]			75.95	65.94	65.37	76.03	71.76	72.16	83.14	73.54
+SKD [7]			76.83	69.93	67.94	77.83	72.31	73.83	84.94	73.64
+CWD [8]			77.94	71.54	68.49	**78.94**	72.65	74.01	**85.92**	73.96
+AT [26]†			75.98	67.45	66.11	76.34	71.95	72.92	83.84	73.45
+CAT [10]†			76.84	69.67	67.81	77.12	71.42	72.03	84.24	73.54
+XGD (Ours)†			**78.12**	**71.77**	**68.93**	78.03	**72.97**	**74.15**	85.29	**74.01**
S: PSPNet-R18	12.14	11.39	64.25	59.63	59.86	71.11	69.58	71.09	82.04	73.15
+PI [11]			68.29	60.96	60.04	73.65	71.33	72.39	82.19	73.36
+SKD [7]			71.92	62.45	61.93	73.99	71.93	72.83	83.01	73.81
+CWD [8]			72.91	63.13	62.42	74.45	**72.74**	73.31	84.74	74.11
+AT [26]†			69.12	60.57	60.94	73.94	71.84	72.64	82.49	73.45
+CAT [10]†			70.13	61.98	61.23	73.92	71.89	72.01	82.93	73.69
+XGD (Ours)†			**73.11**	**63.76**	**62.59**	**74.98**	72.02	**73.39**	**84.94**	**74.47**

Table 1. Performance comparison with state-of-the-art distillation methods over various student segmentation networks on TTPLA, Substation, and Pascal VOC 2012 datasets. The best results are in **bold**. The second-best results are underlined. The floating-point operations per second (FLOPs) calculation is based on the crop size of 512 × 512.

- Consistent superiority across all three benchmark datasets (TTPLA, Substation, Pascal VOC 2012)
- Effective across various student architectures (DeepLabV3+, PSPNet)
- Works well with different backbones (ResNet18, MobileNetV2)
- Balances global class knowledge with spatial feature refinement

Qualitative Results

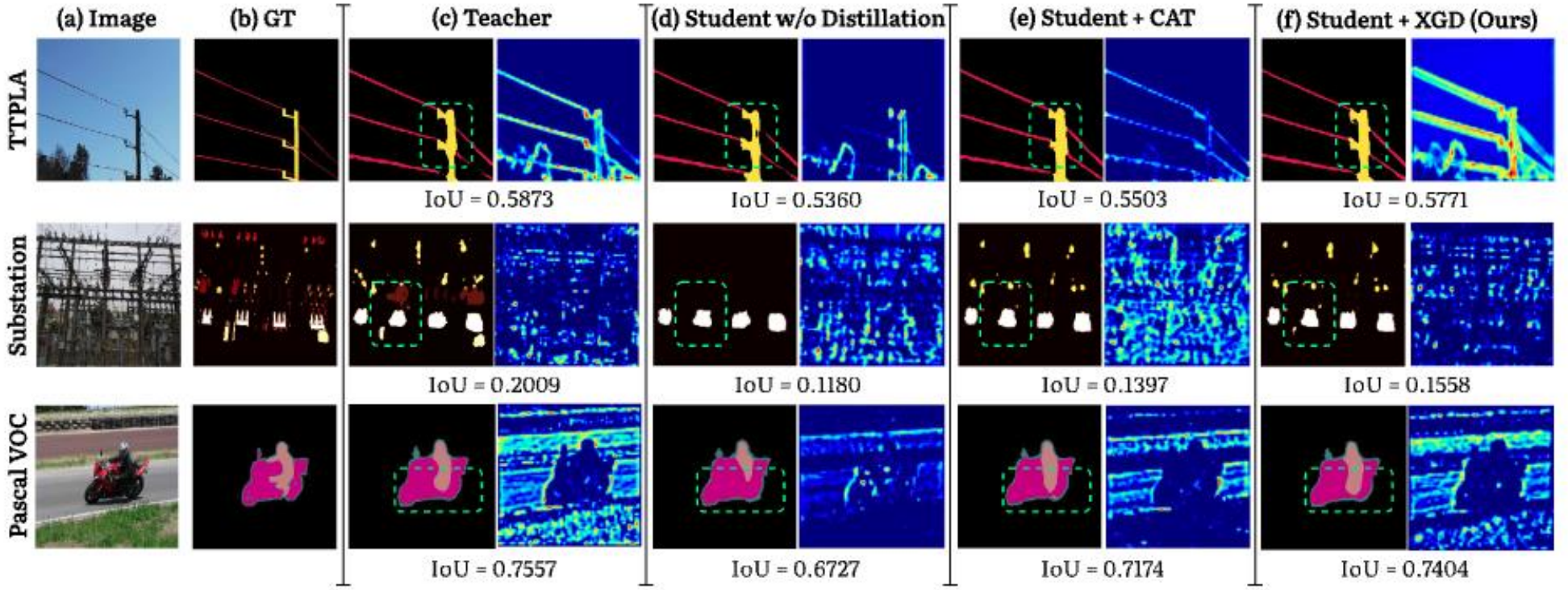


Figure 2. Qualitative segmentation results and saliency maps on TTPLA *test*, Substation *test*, and Pascal VOC 2012 *validation*. (a) raw images, (b) ground-truth (GT), (c) teacher network, (d) original student network without distillation, (e) student network with CAT method, and (f) student network with our XGD method.

Ablation Study: XAI methods

Evaluated XAI methods for generating saliency maps on TTPLA dataset with DeepLabV3+ ResNet18 student network.

- **GradCAM selected as optimal choice**: Provides strong performance without significant computational overhead and effectively guides student network focus on spatially critical regions.
- **EigenGradCAM** achieves the best results (+0.12% over GradCAM), but 7x slower (impractical for large-scale training).

XAI Method	mIoU(%)	Avg. Time(s)
GradCAM [20]	69.82	0.1031
GradCAM++ [17]	68.96	0.1037
HiResCAM [19]	69.74	0.1224
EigenGradCAM [21]	**69.94**	0.7554
LayerCAM [18]	68.83	**0.1021**

Table 3. Ablation study of XAI methods in XAI-based saliency map distillation $\mathcal{L}_{sa}$ on student network (DLV3P-R18) on TTPLA *test*.

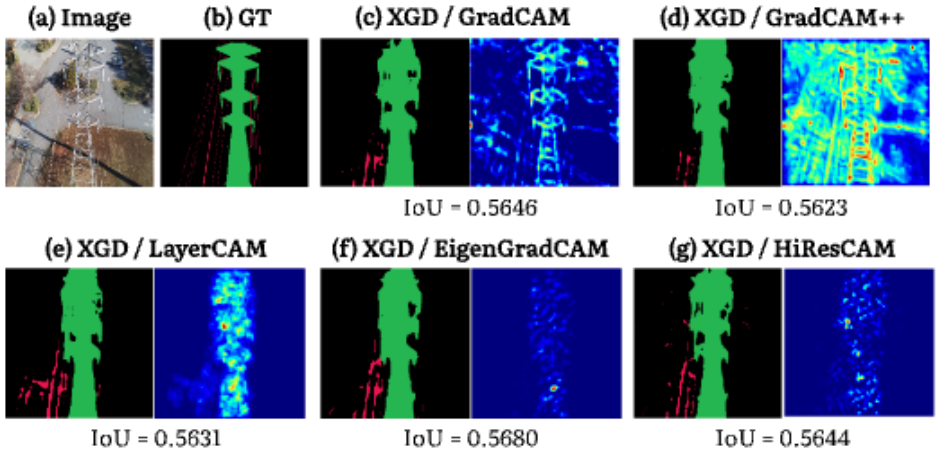


Figure 5. Qualitative segmentation results and saliency maps of student network (DLV3P-R18) with different XAI methods on $\mathcal{L}_{sa}$ on TTPLA *test*.

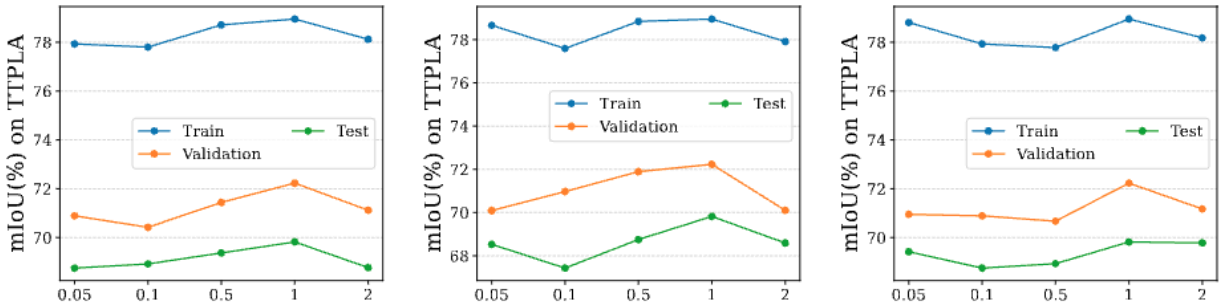
Ablation Study: Hyperparameters

Evaluated impact of weighting coefficients α, β and temperature τ on TTPLA dataset with DeepLabV3+ ResNet18 student network

- Tested range: {0.05, 0.1, 0.5, 1.0, 2.0}
- Optimal combination: **α = 1.0, β = 1.0, τ = 1.0**
- **Equal weighting (α = β = 1.0)** demonstrates balanced importance of pixel-wise class probability alignment and saliency map refinement.
- Effective knowledge transfer requires both global and spatial feature guidance

Loss	Baseline	Distillation		
$\mathcal{L}_{seg}$	✓	✓	✓	✓
$\mathcal{L}_{pi}$	-	✓	-	✓
$\mathcal{L}_{sa}$	-	-	✓	✓
mIoU(%)	65.25	67.01	68.77	69.82

Table 2. Ablation study of distillation loss terms on student network (DLV3P-R18) on TTPLA *test*. Baseline denotes the segmentation loss $\mathcal{L}_{seg}$.

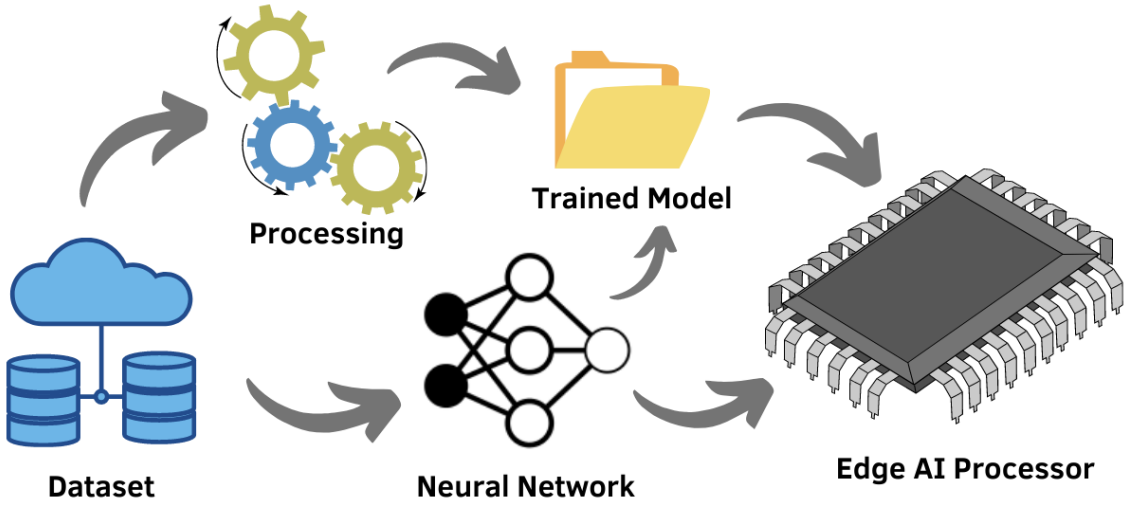


(a) α ($\mathcal{L}_{pi}$ coefficient) (b) β ($\mathcal{L}_{sa}$ coefficient) (c) Temperature τ

Figure 3. The impact of XGD weighting coefficients α, β and temperature τ on the student network (DLV3P-R18) in mIoU(%) on the TTPLA dataset.

Conclusion and Future Works

- Balancing global class knowledge with spatial feature refinement can provide computationally efficient knowledge distillation approach for resource-constrained deployment.
- **Extend to Object Detection:** Apply XGD framework to object detection models.
- **Real-time Optimization:** Further reduce computational overhead for real-time applications.
- Code available at: https://github.com/Analytics-Everywhere-Lab/xaiseg



https://embeddedcomputing.com/technology/iot/edge-computing/edge-ai-is-overtaking-cloud-computing-for-deep-learning-applications

Our mission

We're recruiting MSc. and PhD students!

If you are interested, contact hcao3@unb.ca

Hung Cao, PhD

Assistant Professor, Lab Director
Analytics Everywhere Lab
University Of New Brunswick, Canada
hcao3@unb.ca

Affiliated Faculty

Francis Palma, PhD

Monica Wachowicz, PhD

Trevor Hanson, PhD

Rene Richard, MSc

Students

Hung Nguyen

Asfia Kawnine

Atah Nuh Mih

Alireza Rahimi

Pavi P

...

Krishno Dey

Connor McLenaghan

Ishan Randeniya

Simran Dadhich

Bohdan Savchuk

...