

XEdgeAI: A Human-centered Industrial Inspection Framework with Data-centric Explainable Edge AI Approach

Hung Nguyen¹, Loc Nguyen², Hung Cao¹

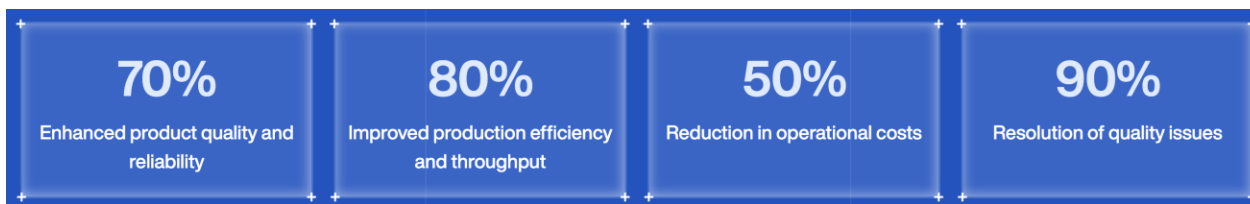
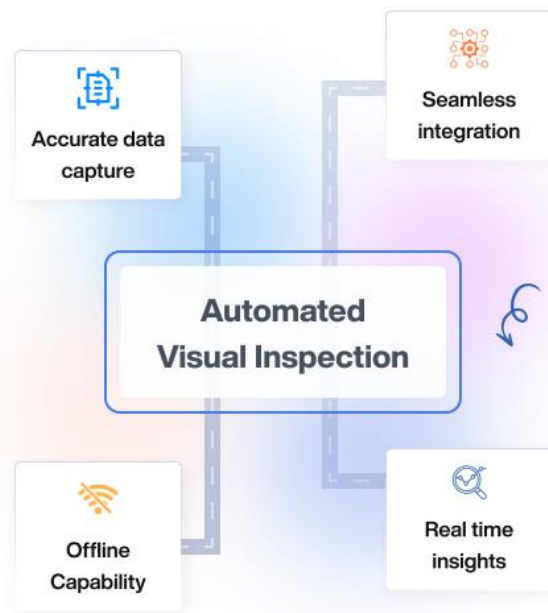
¹Analytics Everywhere Lab, University of New Brunswick, Canada

²Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany

-
- 01** Motivation
 - 02** Methodology
 - 03** Experiment 1: A Comprehensive Evaluation
 - 04** Experiment 2: A Data-Centric Approach
 - 05** Conclusion

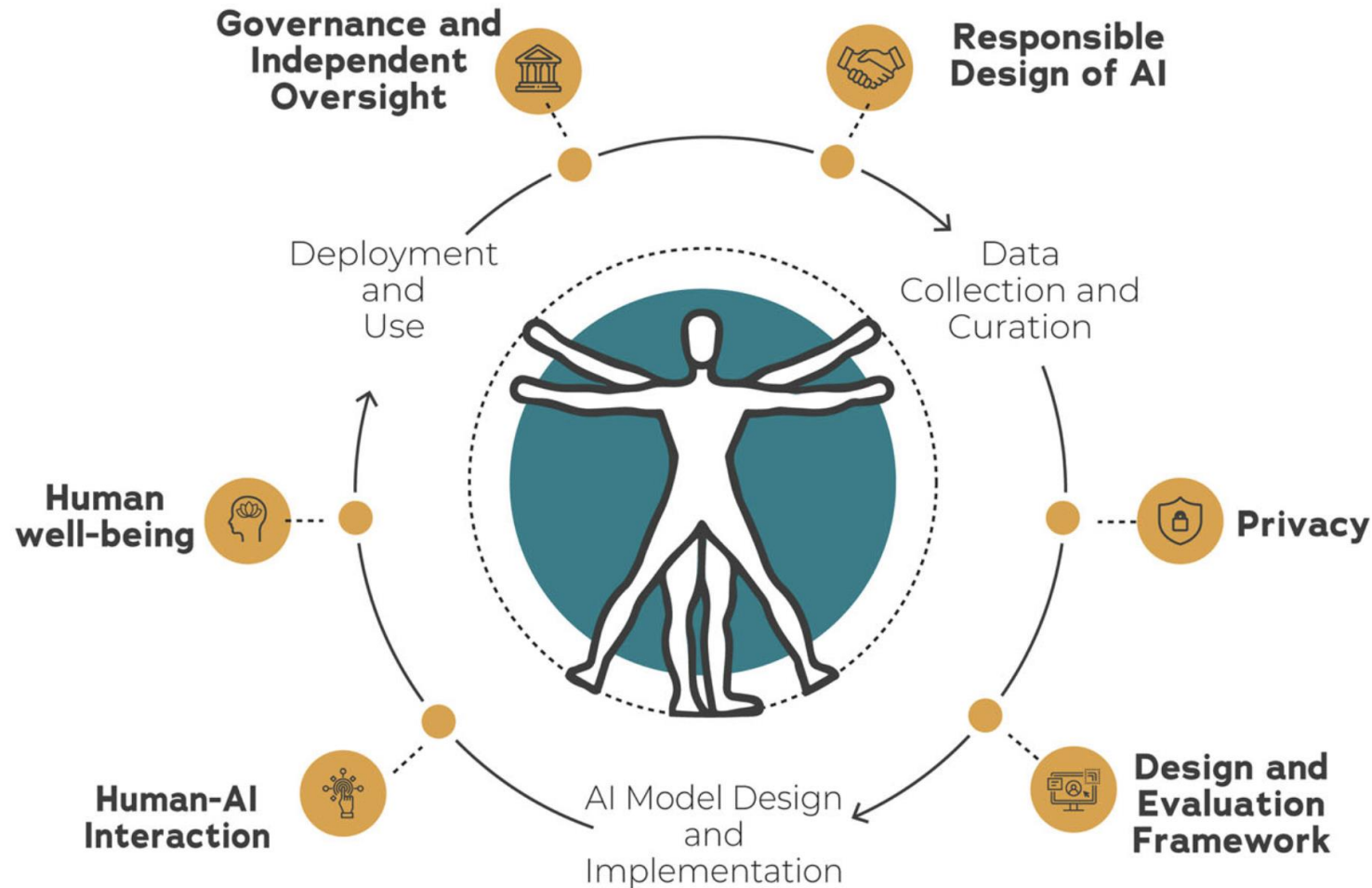
Motivation

Industrial Visual Quality Inspection



<https://www.scanflow.ai/automated-visual-inspection-solution/>





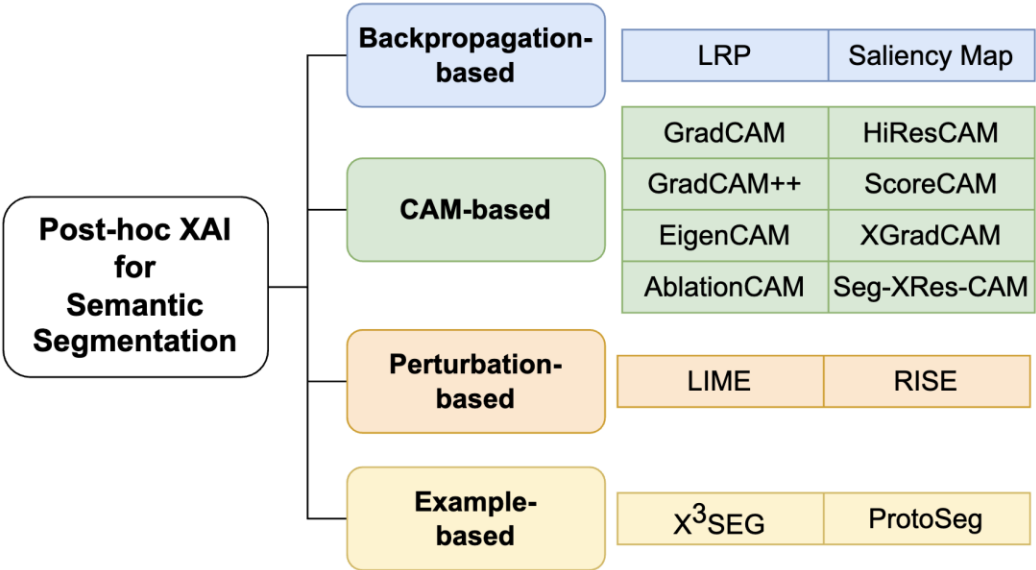
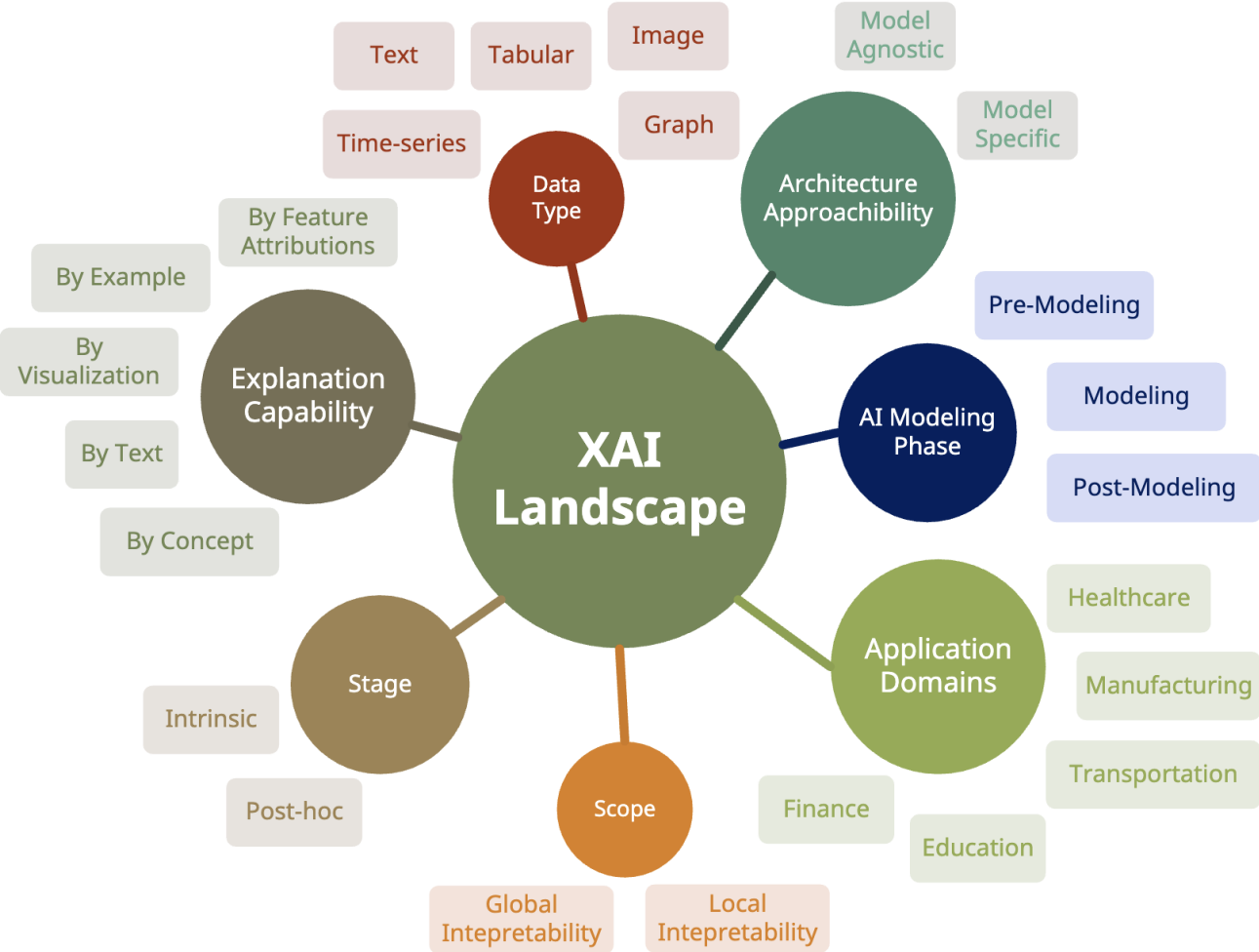


Figure 2: Local Post-hoc XAI Methods for the Semantic Segmentation.

Figure 1: The XAI landscape, categorizing XAI techniques based on explanation scope, data type, modeling phase, architecture, and application domains.

*“An Edge Human-centered Explainable AI framework for
Industrial Visual Quality Inspection Systems”*

Methodology

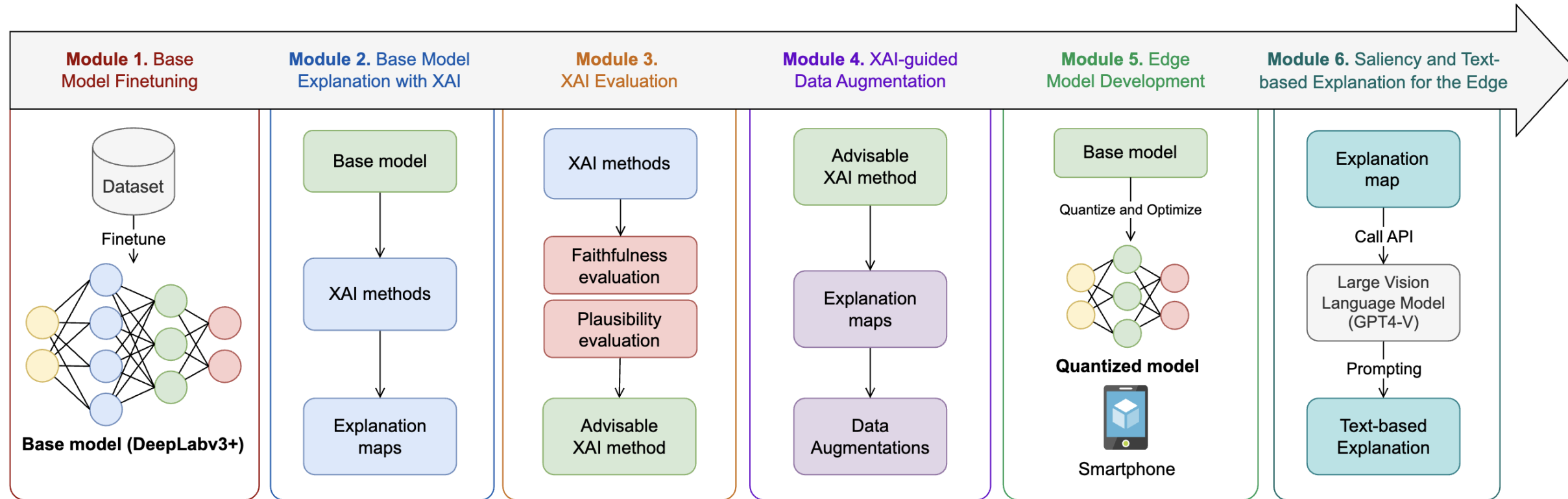


Figure 3: The methodology of the XAI-integrated Visual Quality Inspection framework integrated with XAI methods with 6 building modules: (1) Base Model Finetuning with a provided visual quality dataset, (2) Base Model Explanation with XAI, (3) XAI Evaluation, (4) XAI-guided Data Augmentation to improve the base model performance, (5) Edge Model Development on mobile devices and (6) Saliency and textual Explanation for the Edge. The end-users interact with the framework via a mobile application, while the domain experts can interact via a web application.

Implementation

- Module 1: Base Model Finetuning
- Module 2: Base Model Explanation with XAI
- Module 3: XAI Evaluation
- Module 4: XAI-guided Data Augmentation
- Module 5: Edge Model Development
- Module 6: Saliency and Textual Explanation for the Edge

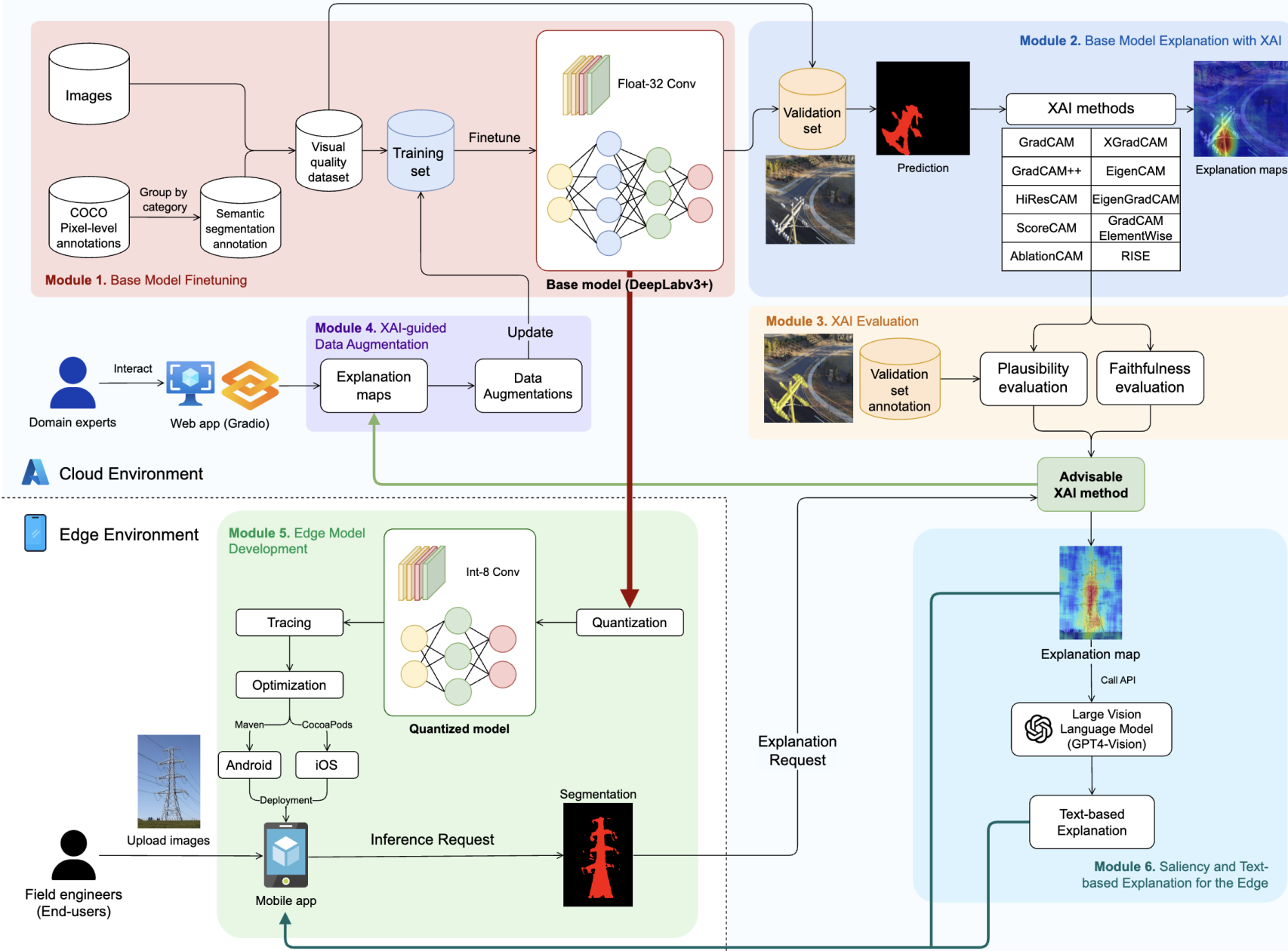


Figure 4: The implementation details of our proposed XAI-integrated Visual Quality Inspection framework.

Experiment 1 – A Comprehensive Evaluation

Module 1 – Base Model Finetuning

DeepLabv3+

DeepLabv3+ is an advanced semantic segmentation architecture combining an Atrous Spatial Pyramid Pooling (ASPP) encoder for capturing multi-scale contextual information [1].

Backbone: ResNet50, ResNet101, MobileNetv2.

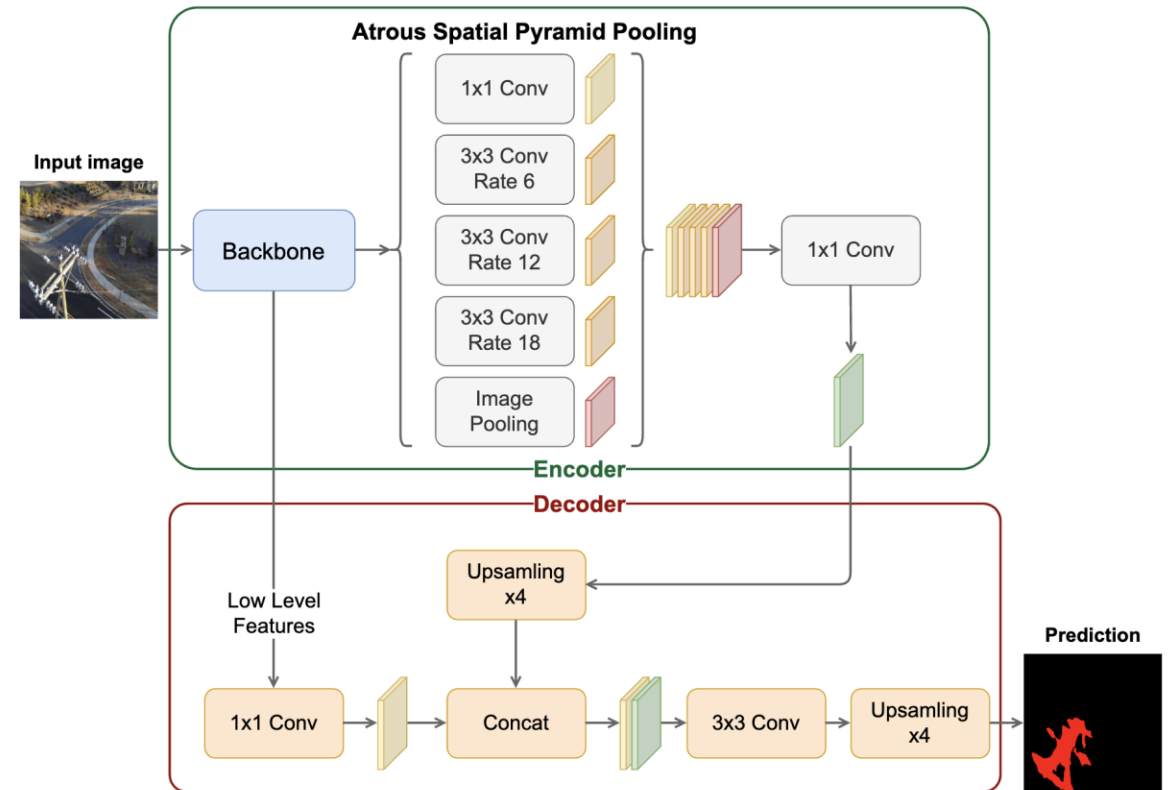


Figure 5: The architecture of the DeepLabv3Plus model, featuring an encoder with Atrous Spatial Pyramid Pooling and a decoder for upsampling and refining segmentation outputs.

Module 1 – Base Model Finetuning

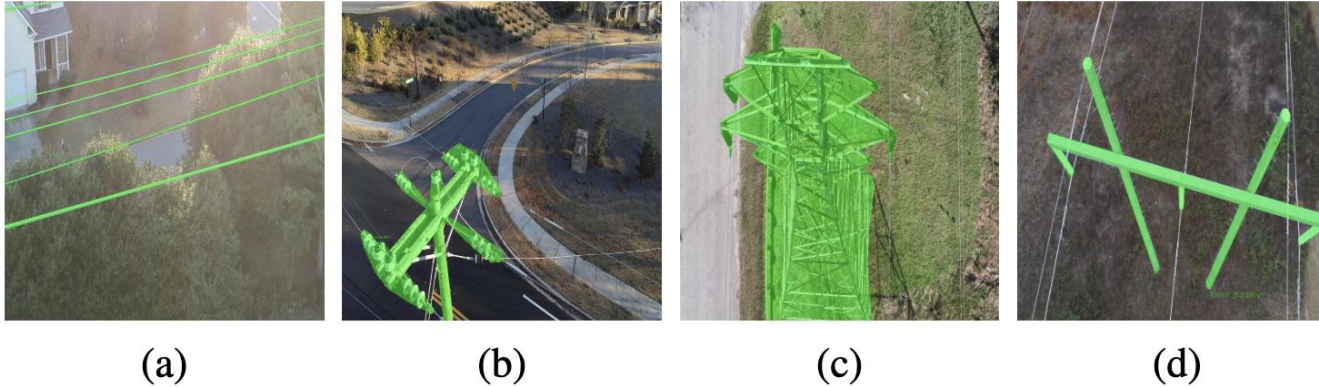


Figure 8: Samples from the TTPLA dataset represent the main objects of categories in the green masks (a) cable, (b) tower_wooden, (c) tower_lattice, (d) tower_tucohy.

Model	#Params	Size(MB)	cable	tower_wooden	tower_lattice	tower_tucohy	mIoU
DLv3P-MobileNetv2-B	4.37M	16.71	53.94	80.11	88.19	86.49	77.18
DLv3P-MobileNetv2-E	4.37M	16.71	54.37	86.49	80.98	88.78	77.82
DLv3P-MobileNetv2-M	3.51M	13.39	48.74	84.54	78.55	86.12	75.48
DLv3P-ResNet50-B	26.67M	101.76	56.66	92.31	93.18	90.63	83.20
DLv3P-ResNet50-E	26.67M	101.76	57.42	92.97	91.34	93.87	83.90
DLv3P-ResNet50-M	21.36M	81.48	53.67	90.39	88.80	91.26	81.53
DLv3P-ResNet101-B	45.66M	174.21	57.22	95.23	96.02	91.42	84.97
DLv3P-ResNet101-E	45.66M	174.21	58.36	96.54	92.89	97.61	86.35
DLv3P-ResNet101-M	36.57M	139.52	54.43	94.07	90.22	95.09	83.95

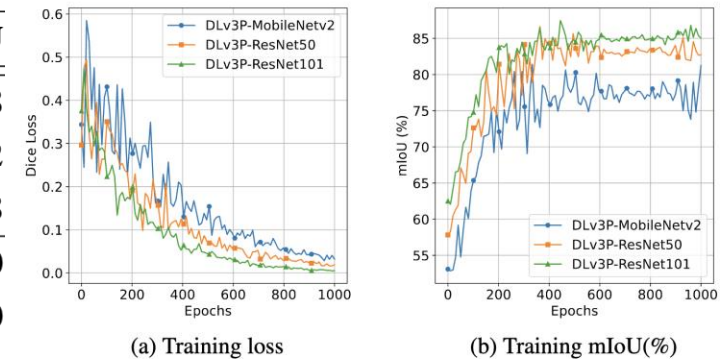


Figure 9: The training loss in Dice loss and accuracy in mIoU(%) of three base DeepLabv3Plus models over 1000 epochs.

Table 2: Accuracy comparison of DeepLabv3Plus (DLv3P) variants with different backbones (MobileNetv2, ResNet50, ResNet101) in terms of average IoU (%) for each category and mIoU (%). The models are evaluated at different stages: base (B), enhanced (E), and mobile (M) on the TTPLA validation set.

IoU: measures the overlap between the saliency map and the ground truth annotation

$$\text{IoU} = \frac{\text{Area}(L^c \cap \text{GT})}{\text{Area}(L^c \cup \text{GT})}$$

BBox: measures the overlap between the bounding box of the saliency map and the ground truth bounding box

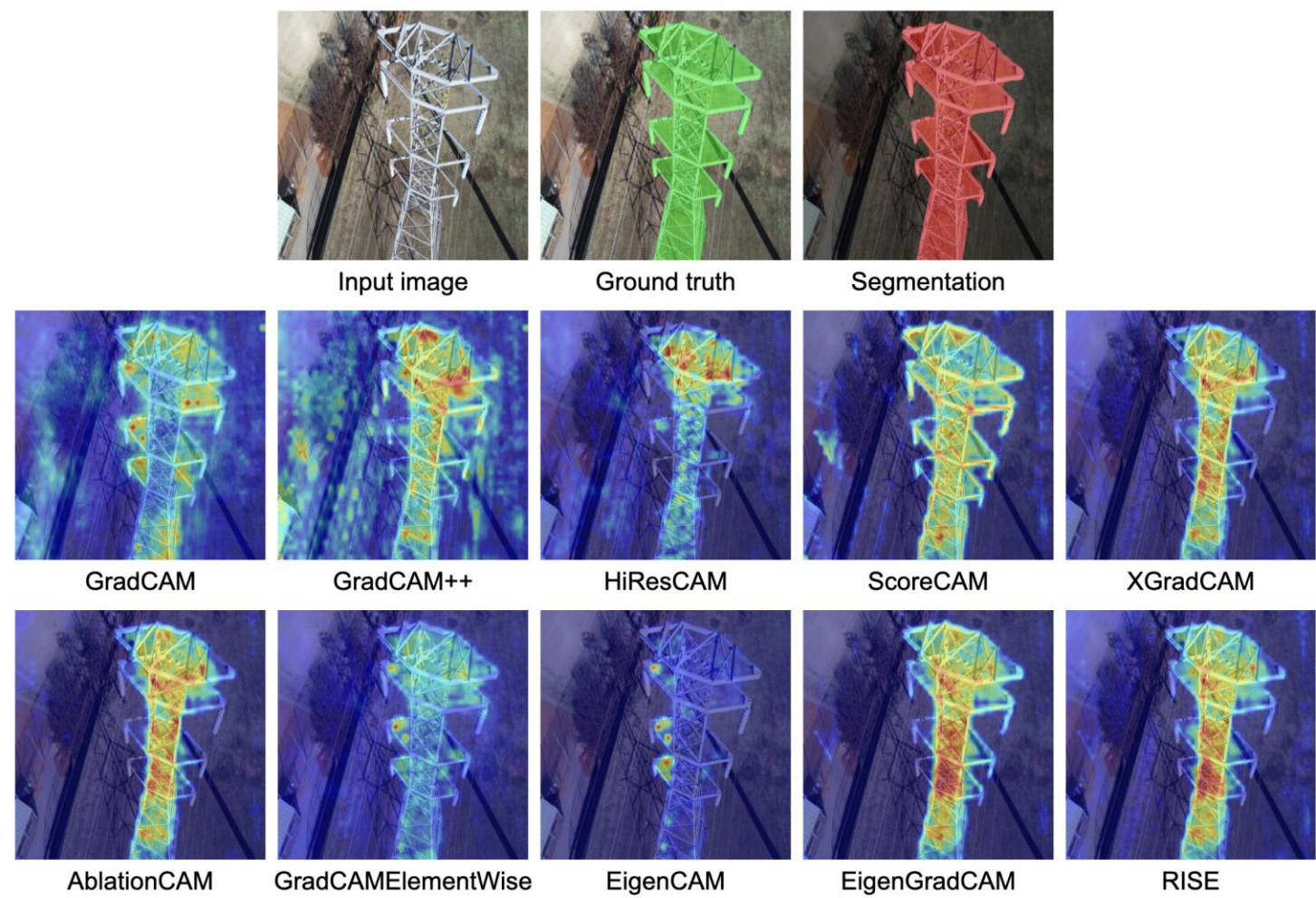
$$\text{Bbox} = \frac{\text{Area}(\text{BBox}(L^c) \cap \text{BBox}(\text{GT}))}{\text{Area}(\text{BBox}(L^c) \cup \text{BBox}(\text{GT}))}$$

Deletion: measures the accuracy of saliency areas by removing pixels from the input image in order of saliency, from large to small.

$$d_{\text{Del}} = \text{AreaUnderCurve}(h_i \text{ vs. } i/n, \forall i = 0, \dots, n)$$

Insertion: measures the comprehensiveness of the saliency area by removing all pixels from the input image and recovering them in order of saliency, from large to small.

$$d_{\text{Ins}} = \text{AreaUnderCurve}(h_i \text{ vs. } i/n, \forall i = 0, \dots, n)$$



Method	EPBG↑	BBox↑	IoU↑	Del↓	Ins↑
GradCAM	50.49	48.39	47.94	0.521	0.527
GradCAM++	58.13	52.24	53.22	0.517	0.547
HiResCAM	60.81	41.69	52.19	0.501	0.559
ScoreCAM	54.01	43.95	51.94	<u>0.434</u>	<u>0.609</u>
XGradCAM	57.94	47.81	53.09	<u>0.594</u>	<u>0.551</u>
AblationCAM	61.03	51.39	54.73	0.498	0.589
GradCAMElementWise	40.95	35.91	39.25	0.859	0.415
EigenCAM	51.49	42.54	51.03	0.545	0.491
EigenGradCAM	64.11	<u>62.24</u>	60.93	0.520	0.534
RISE	<u>62.42</u>	63.52	<u>56.13</u>	0.123	0.691

Table 3: The quantitative result of XAI methods on the validation set \mathcal{D}_{val} on different metrics such as EPBG(%), BBox(%), IoU(%), Drop, Increase, and running time in seconds. For each metric, the arrow \uparrow / \downarrow indicates higher/lower scores are better. The best is in **bold**, and the second best is in underline.

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%.

Module 4 – XAI-guided Data Augmentation

Web UI for Domain Experts

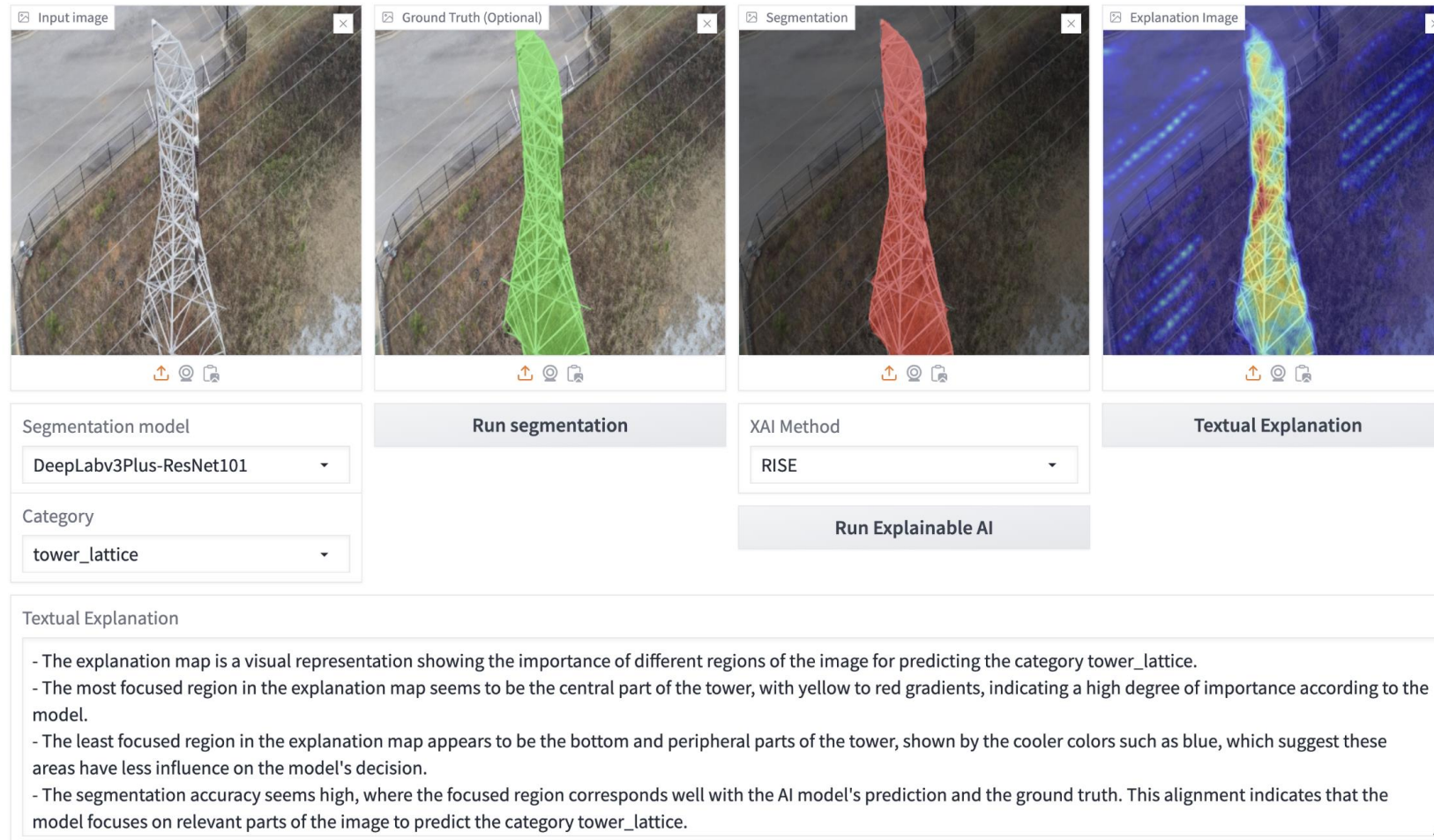


Figure 6: Web-based user interface for domain experts to monitor predictions, explanations, and textual explanations. The input image, segmentation output, saliency map, and textual explanation are displayed, allowing experts to assess the model's performance and provide feedback for annotation augmentation.

Module 4 – XAI-guided Data Augmentation

Data Augmentation Suggestions

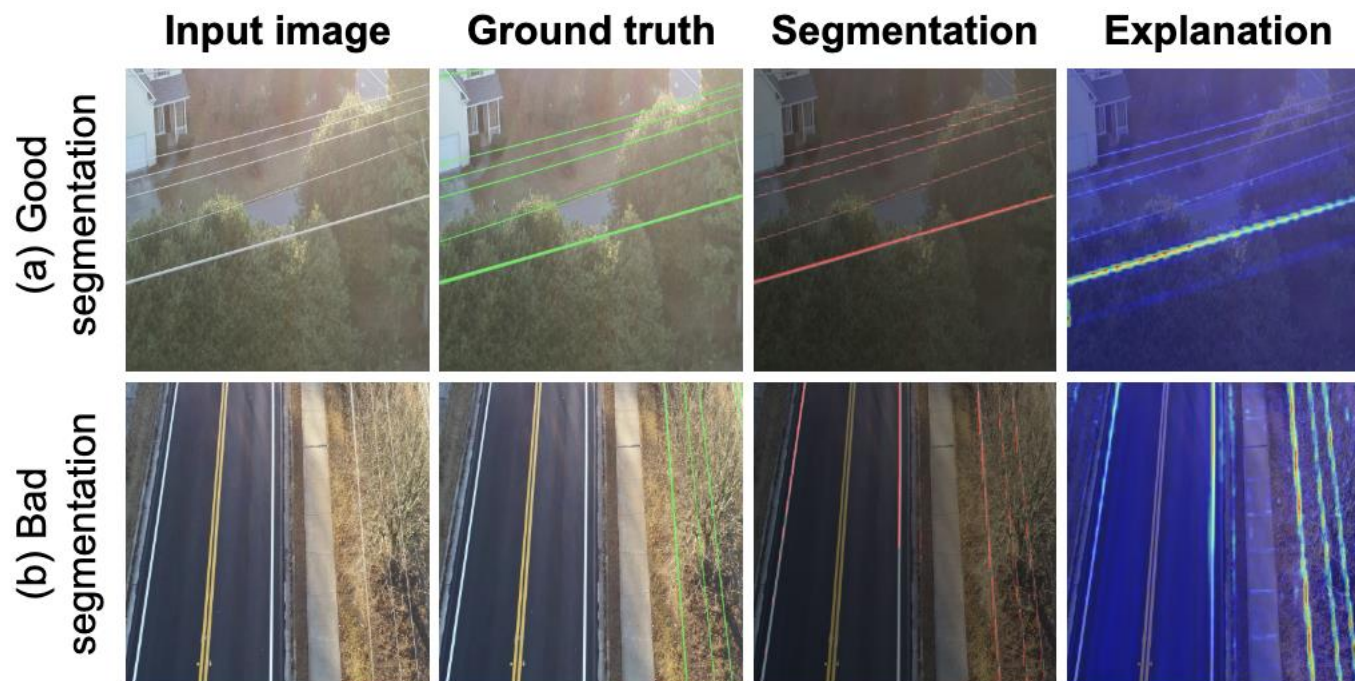
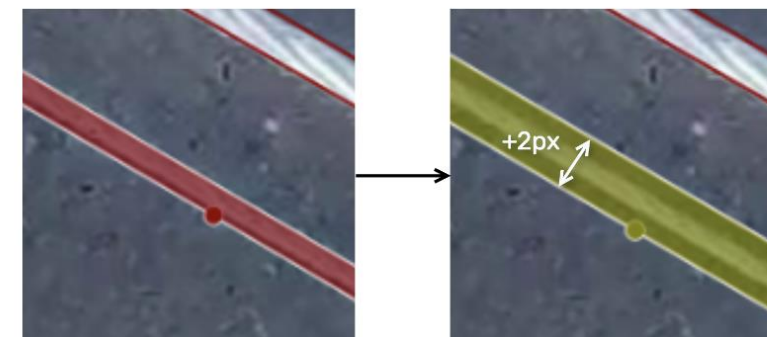
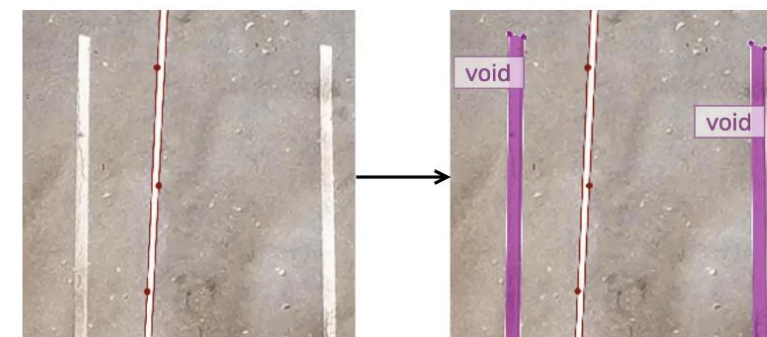


Figure 11: List of input images, ground truth, segmentation of the base DeepLabv3Plus-ResNet101-B model for the “cable” inspection for two segmentation cases: (a) Good segmentation and (b) Bad segmentation.



(a) Annotation Enlargement



(b) Adding Annotations for Perplexed Objects

Figure 12: Annotation augmentation approaches: (a) Annotation enlargement where the annotation size for thin objects like cables is increased by 2 pixels on both sides. (b) Adding annotations for perplexed objects like the road surface marks to guide the model in differentiating between white cables and perplexed objects.

Module 5 – Edge Model Development

Algorithm

Algorithm 1: Model Quantization, Pruning, and Optimization for Mobile

Input: base_model Θ : a base model

Input: input_tensor: an input tensor

Output: mobile_model θ : an optimized, quantized, and pruned model

```
/* Disable batch normalization */
1 for module in base_model.modules() do
2     if isinstance(module, nn.BatchNorm2d) then
3         | module.eval()
4     end
5 end
/* Apply dynamic quantization */
6 base_model =
    torch.quantization.quantize_dynamic(base_model,
    {nn.Conv2d, nn.Linear, nn.ReLU, nn.BatchNorm2d,
    nn.AdaptiveAvgPool2d}, dtype=torch.qint8)
/* Forward pass to simulate inference */
7 base_model(input_tensor)
```

```
/* Apply pruning */
8 for name, module in base_model.named_modules() do
9     if isinstance(module, nn.Conv2d) then
10         | prune.ln_structured(module, name='weight',
11                               | amount=0.1, n=1, dim=0)
12     end
/* Remove pruning reparameterizations */
13 for name, module in base_model.named_modules() do
14     if isinstance(module, nn.Conv2d) then
15         | prune.remove(module, 'weight')
16     end
17 end
/* Trace the model */
18 traced_model = torch.jit.trace(base_model, input_tensor)
/* Optimize the traced model for mobile */
19 mobile_model = optimize_for_mobile(traced_model)
20 return mobile_model
```

Module 5 – Edge Model Development

Results

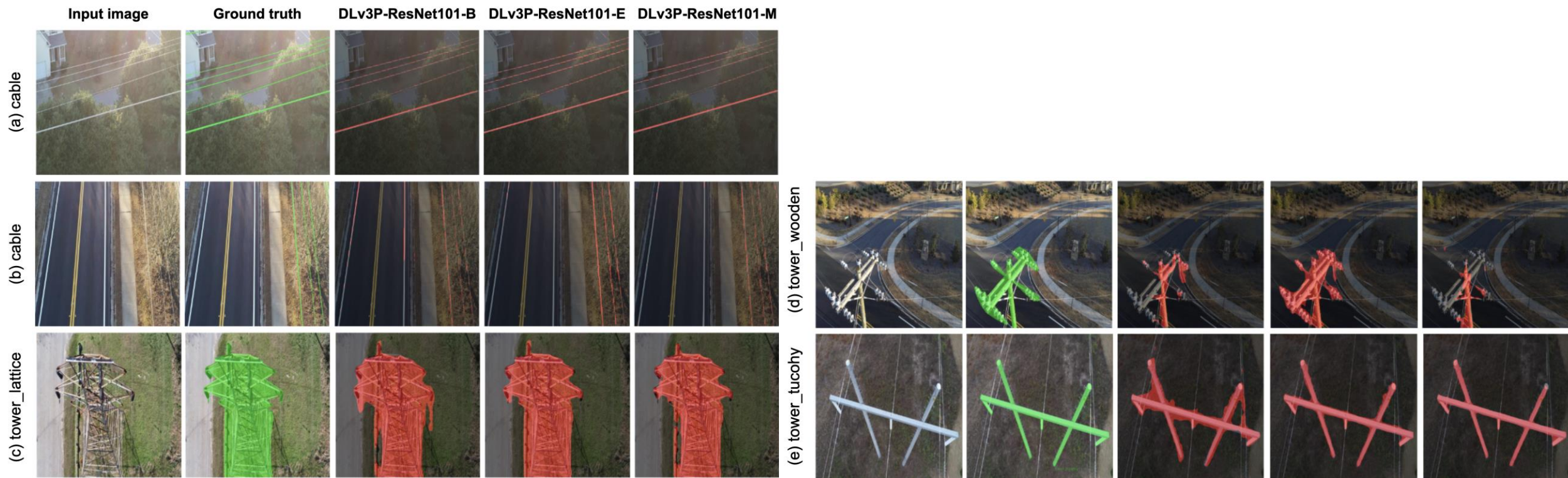


Figure 13: Qualitative results of DeepLabv3Plus (DLv3P)-ResNet101 on four categories of the TTPLA dataset. The models are evaluated at different stages: base (B), enhanced (E), and mobile (M).

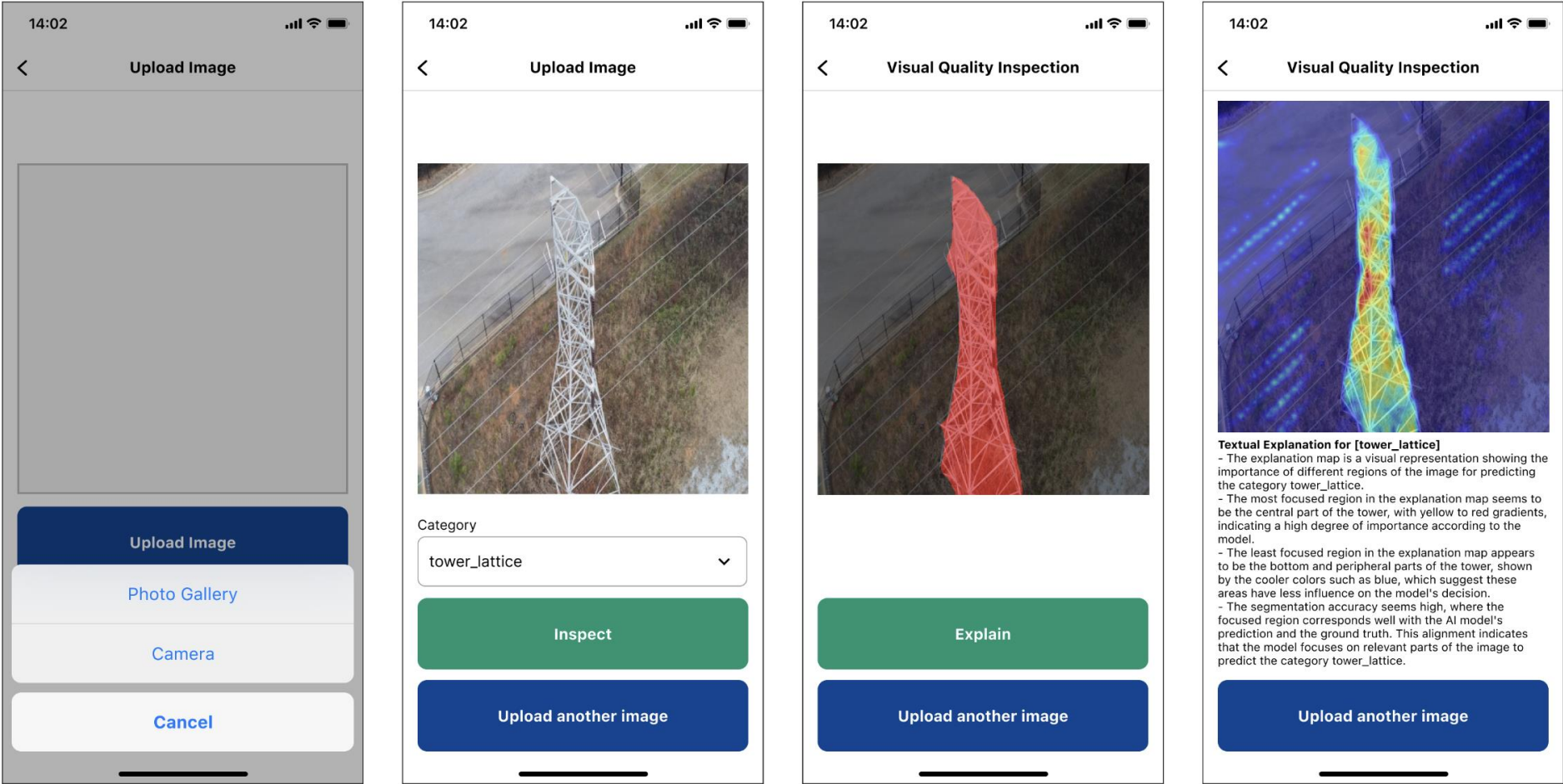


Figure 7: The iOS user interface of the mobile application for end-users, designed for iPhone 11 Pro. Users can upload an image, initiate visual quality inspection for a particular category, view the segmentation results along with explanations, and have the option to upload another image.

Module 6 – Saliency and Textual Explanation for the Edge

Prompt Template

TEMPLATE 1: PROMPT TEMPLATE FOR THE GPT-4 VISION

System Message: You are an Explainable AI expert for semantic segmentation models.

The first image is the original image. The second image is the ground truth of category {category} in the original image. The third image is the prediction of an AI model for category {category} in the original image. The fourth image is the explanation map of the category {category} in the original image. Think step by step to understand how the explanation map and prediction align with the ground truth. First, capture the image context in the original image. Secondly, identify which parts belong to category {category} in the ground truth. Then, you look at the explanation map to see the saliency map for the segmentation mask of the category {category}. Your task is to check if the focused region in the explanation map supports the prediction for category {category}. Your final answer must be concise, simple, and separated by bullet points. First, briefly describe the explanation map. Secondly, describe the most focused region of category {category} in the explanation. Thirdly, describe the least focused region of category {category} in the explanation. Fourthly, assess the localization quality if the focused region aligns with the prediction and ground truth for category {category}.

User Message:

image_url: "data:image/jpeg;base64,{input_image}"

image_url: "data:image/jpeg;base64,{ground_truth_image}"

image_url: "data:image/jpeg;base64,{segmentation_image}"

image_url: "data:image/jpeg;base64,{explanation_image}"

Module 6 – Saliency and Textual Explanation for the Edge

Results

TEMPLATE 2: TEXTUAL EXPLANATIONS OF THE TTPLA DATASET

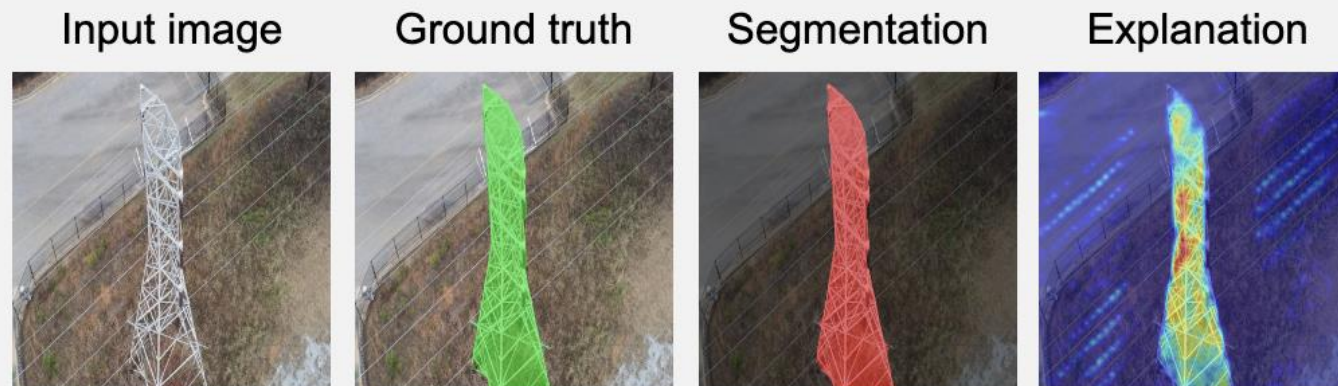


Figure 14: Good segmentation - Category: tower_lattice

Model Output:

- The explanation map is a visual representation showing the importance of different regions of the image for predicting the category tower_lattice.
- The most focused region in the explanation map seems to be the central part of the tower, with yellow to red gradients, indicating a high degree of importance according to the model.
- The least focused region in the explanation map appears to be the bottom and peripheral parts of the tower, shown by the cooler colors such as blue, which suggest these areas have less influence on the model's decision.
- The segmentation accuracy seems high, where the focused region corresponds well with the AI model's prediction and the ground truth. This alignment indicates that the model focuses on relevant parts of the image to predict the category tower_lattice.

Module 6 – Saliency and Textual Explanation for the Edge

Results

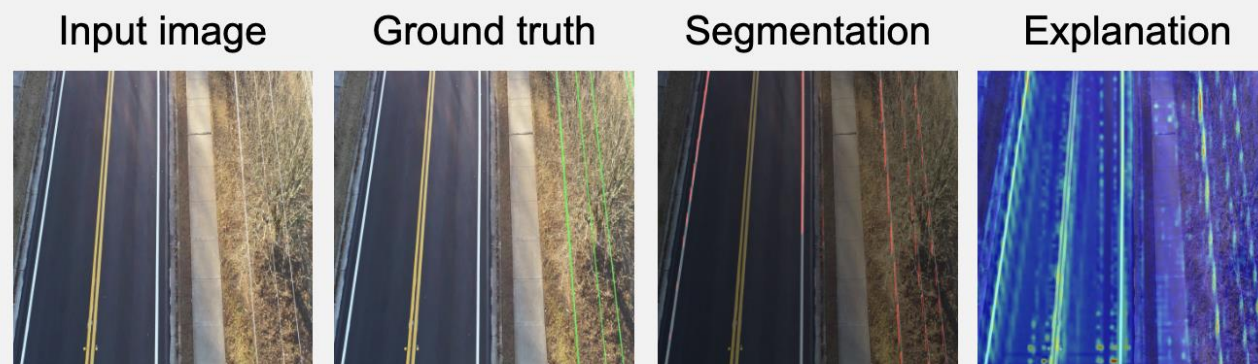


Figure 15: Bad segmentation - Category: cable

Model Output:

- The explanation map is a visual representation that indicates which areas of the image the AI model deemed relevant when determining where the category “cable” is present.
- The most focused region of the cable in the explanation map is evident along the linear features within the road, which correlate with the yellow lane markings and the road’s edges.
- The least focused region in the explanation appears to be the areas on the pavement free from markings and the natural vegetation areas on the right side of the image.
- The focused region in the explanation map does not perfectly align with the ground truth for the cable category. The prediction appears to have false positives along the lane markings. It does not accurately distinguish between the road markings and actual cables, whereas the ground truth demarcates straight lines purported to be cables adjacent to the road on both sides. Therefore, there is a discrepancy between the focused regions in the explanation map and the true location of cables, as shown in the ground truth image.

Experiment 2 – A Data-Centric Approach

Substation Equipment Dataset

Data & Segmentation Results

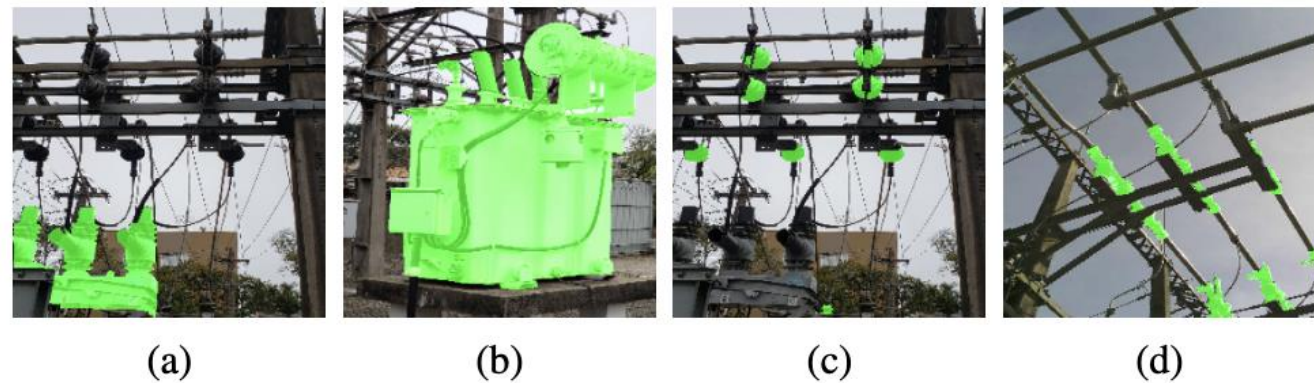


Figure 16: Samples from the Substation Equipment dataset represent the main objects of categories in the green masks (a) recloser, (b) power_transformer, (c) porcelain_pin_insulator, (d) tripolar_disconnect_switch

Model	B	E	M
breaker	92.93	94.51	92.51
closed_blade_disconnect_switch	49.14	50.26	48.02
closed_tandem_disconnect_switch	83.11	84.34	82.04
current_transformer	62.23	63.25	61.07
fuse_disconnect_switch	90.87	92.77	90.03
glass_disc_insulator	53.55	55.42	53.08
lightning_arrester	67.87	69.28	67.41
muffle	88.27	90.06	87.92
open_blade_disconnect_switch	93.34	95.18	93.06
open_tandem_disconnect_switch	83.85	85.54	83.51
porcelain_pin_insulator	30.11	32.03	30.05
potential_transformer	78.54	80.72	78.01
power_transformer	85.73	87.79	85.50
recloser	68.11	70.40	67.50
tripolar_disconnect_switch	83.22	85.24	82.79
mIoU	73.45	75.79	72.58

Table 4: Accuracy comparison of three DeepLabv3Plus models with the ResNet101 backbone at different stages: base (B), enhanced (E), and mobile (M) in terms of average IoU (%) for each category and mIoU (%) on the Substation equipment validation set.

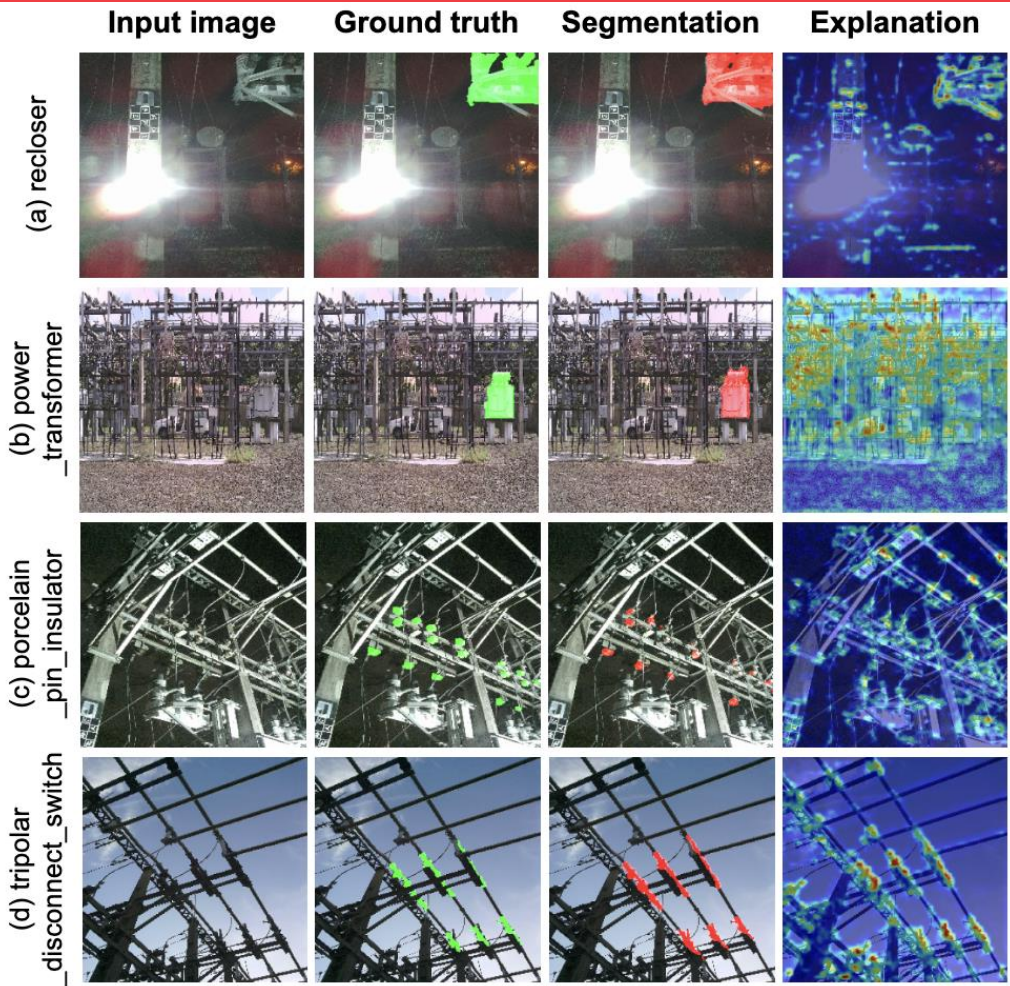


Figure 17: List of input images, ground truth, segmentation of the base DeepLabv3Plus-ResNet101-B model, and the corresponding explanations for four categories of the Substation equipment dataset.

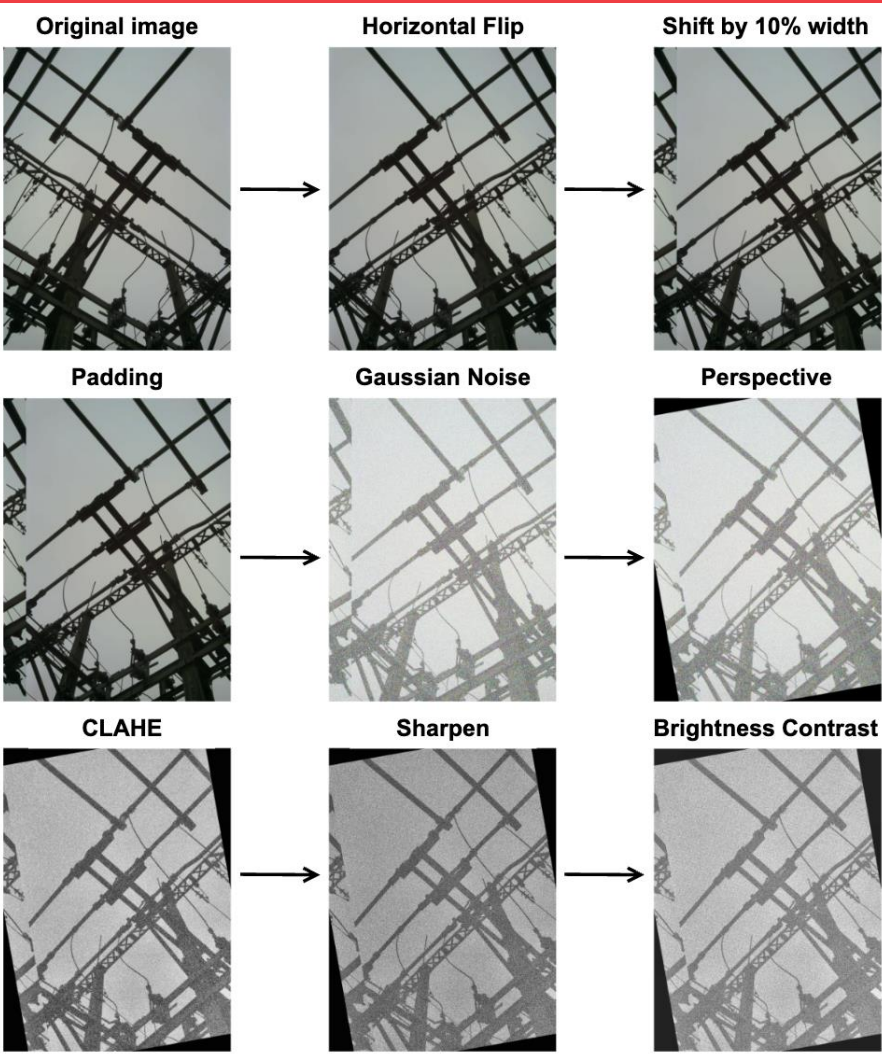


Figure 18: Augmentation techniques on the training set of the Substation equipment dataset.

TEMPLATE 3: TEXTUAL EXPLANATIONS OF THE SUBSTATION EQUIPMENT DATASET

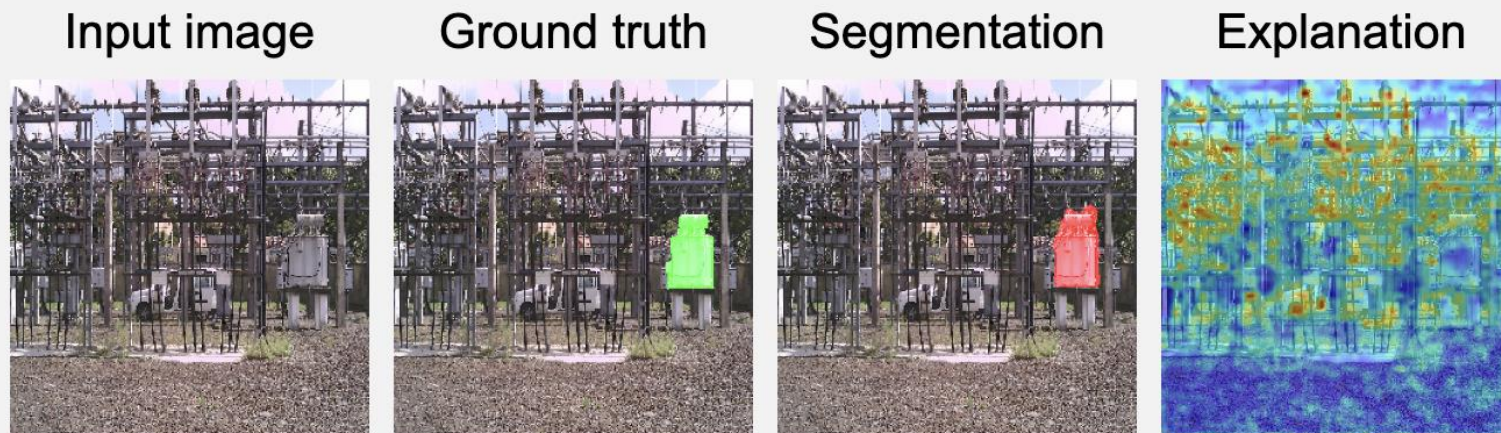


Figure 19: Good segmentation - Category: power_transformer

Model Output:

- The explanation map highlights several regions of importance, with some focus around the power transformer but also attention on other structures.
- The central structures and parts of the power transformer have significant focus in the explanation map.
- The ground and less critical background structures have the least focus.
- The focused regions in the explanation map partially support the prediction for the power transformer. The prediction is accurate, but the explanation map indicates that the model's focus is distributed, suggesting it uses broader contextual cues.

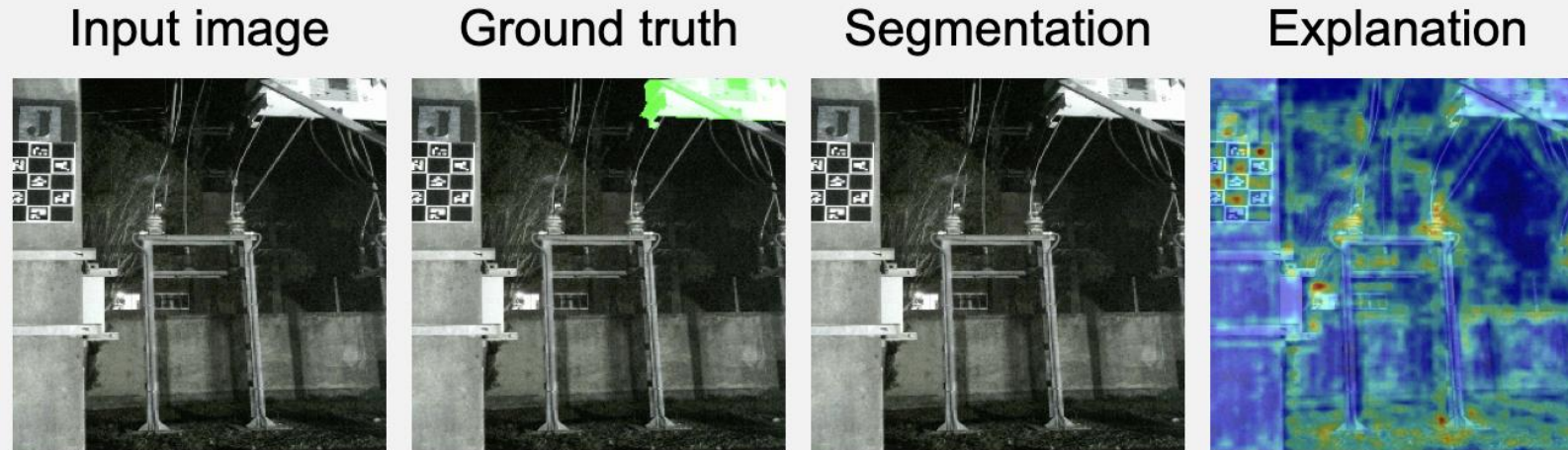


Figure 20: Bad segmentation - Category: recloser

Model Output:

- The explanation map highlights regions of importance, with significant focus on the left-side components and connectors, rather than the recloser.
- The labeled components and connectors on the left are the most focused regions in the explanation map.
- The background and some parts of the ground and structure are the least focused regions.
- The focused regions in the explanation map do not support the prediction for the recloser. The model failed to segment the recloser, indicating a significant issue with the model's ability to recognize and segment this category.

Conclusion

Limitations	Future Works
XAI-guided annotation augmentation process relies on domain experts' expertise and manual effort	Automate the annotation refinement process based on XAI insights
Generating textual explanations on edge devices may introduce latency and computational overhead	Explore more lightweight language models or develop techniques designed for edge computing
Other visual inspection domains in edge computing scenarios remains to be explored	Investigate the generalizability of the approach across visual quality inspection tasks

-
- [1] Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., & Adam, H. (2018). *Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation*. ArXiv.org. <https://arxiv.org/abs/1802.02611>
- [2] Petsiuk, V., Das, A., & Saenko, K. (2018). *RISE: Randomized Input Sampling for Explanation of Black-box Models*. ArXiv.org. <https://arxiv.org/abs/1806.07421>
- [3] Hung, T., Clement, T., Nguyen, Kemmerzell, N., Truong, V. B., Khang, T., Abdelaal, M., & Cao, H. (2024). *LangXAI: Integrating Large Vision Models for Generating Textual Explanations to Enhance Explainability in Visual Perception Tasks*. ArXiv.org. <https://arxiv.org/abs/2402.12525>

Thank you for your listening