Multi-Domain Calibration Framework for SAR-XAI: A Systematic Approach to Trustworthy Explainable AI with Transparency Enhancements

Diego Argüello Ron[1,*], Christyan Cruz Ulloa[2], Orfeas Menis Mastromichalakis[3,4], Kristina Livitckaia[5], Jaime Del Cerro[2], Oscar Garcia Perales[1] and Pawel Andrzej Herman[6]

[1]*Data Analytics for Industries 4.0, Xàtiva, Spain*
[2]*Centro de Automática y Robótica (UPM-CSIC), Universidad Politécnica de Madrid, Madrid, Spain*
[3]*School of Electrical and Computer Engineering, National Technical University of Athens, Athens, Greece*
[4]*Nerion, Chios, Greece*
[5]*Information Technologies Institute, Centre for Research and Technology Hellas, Thessaloniki, Greece*
[6]*Division of Computational Science and Technology, School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, Stockholm, Sweden*

Abstract

Search-and-Rescue robotic operations require trustworthy AI systems where severe overconfidence (D-ECE > 0.9) could compromise life-critical decisions. We present a multi-domain calibration framework demonstrating improvements across synthetic, simulated, and real SAR domains. Our approach uses heuristic-based ground truth generation, enabling calibration assessment without expensive manual annotation. Crucially, we reveal that explainability method choice directly impacts calibration quality—LayerCAM achieves optimal performance with superior sparsity (0.044 ± 0.029) and calibration (D-ECE: 0.136) by creating focused attention maps that enable reliable confidence assessment. Different CAM methods produce distinct attention regions, which affect how calibration is computed and validated, making joint optimization essential for safety-critical deployment. The framework provides foundations for EU AI Act Article 13 transparency requirements while acknowledging the need for expanded validation before operational use.

Keywords

Trustworthy AI, Human-Centered XAI, Model Calibration, SAR Operations, EU AI Act Compliance, LLM-Inspired Calibration

1. Introduction

AI integration in Search-and-Rescue (SAR) operations creates life-critical dependencies on decision accuracy, trustworthiness, and interpretability [1]. Recent evaluations reveal severe model miscalibration with D-ECE scores exceeding 0.9, indicating dangerous overconfidence that could compromise life-critical decisions [2]. This challenge aligns with emerging regulatory frameworks: the EU AI Act Article 13 mandates transparent AI systems with quantifiable explanation quality and human oversight [3], while the NIST AI Risk Management Framework emphasizes measurable trustworthiness in safety-critical applications [4].

This work was conducted within the EXTRA-BRAIN project (EU Horizon Europe Grant No. 101135809), developing brain-inspired AI models ensuring trustworthiness and explainability for safety-critical applications.

TRUST-AI: The European Workshop on Trustworthy AI. Organized as part of the European Conference of Artificial Intelligence - ECAI 2025. October 2025, Bologna, Italy.
*Corresponding author.
diego.arguello@i4ri.com (D. Argüello Ron); christyan.cruz.ulloa@upm.es (C. Cruz Ulloa); menorf@ails.ece.ntua.gr (O. Menis Mastromichalakis); klivit@iti.gr (K. Livitckaia); j.cerro@upm.es (J. Del Cerro); oscar.garcia@i4ri.com (O. Garcia Perales); paherman@kth.se (P. A. Herman)
0000-0002-6004-385X (D. Argüello Ron); 0000-0003-2824-6611 (C. Cruz Ulloa); 0000-0001-5286-1388 (O. Menis Mastromichalakis); 0000-0003-2497-7556 (K. Livitckaia); 0000-0003-4893-2571 (J. Del Cerro); 0009-0003-0781-6141 (O. Garcia Perales); 0000-0001-6553-823X (P. A. Herman)

Model overconfidence occurs when neural networks assign high probability scores to incorrect predictions [2]. We distinguish *epistemic uncertainty* (reducible through additional data) from *aleatoric uncertainty* (irreducible data variability). In SAR operations, overconfidence manifests as D-ECE > 0.9, where confidence severely misaligns with accuracy. For example, predicting "victim detected" with 95% confidence when actual accuracy is 20% causes operators to trust incorrect detections, wasting resources and endangering lives.

Nixon et al. established D-ECE as the standard detection calibration metric [5], with established thresholds: D-ECE < 0.15 for well-calibrated systems and D-ECE > 0.9 indicating dangerous overconfidence [2]. However, existing calibration research assumes single-domain applications with available labeled validation data, lacking cross-domain transfer capabilities essential for SAR operations spanning synthetic, simulated, and real environments.

SAR applications face unique explainability challenges where operators must quickly understand AI recommendations under time pressure [6]. CAM techniques provide visual explanations but lack calibration assessment [7], with LayerCAM showing promise [8]. Most studies treat explanation and calibration as separate processes, problematic in SAR where operators must simultaneously interpret prediction location and confidence level.

This paper introduces four contributions addressing these human-centered trustworthiness challenges:

1. **Heuristic-Based Ground Truth Generation**: A filename-based heuristic procedure to supply calibration labels at scale, removing the need for manual annotation in this study.
2. **Multi-Domain Calibration Framework**: Novel approach with validation across synthetic (D_LLM), simulated (D_SIM), and real (D_REAL) domains.
3. **Cross-Domain Calibration Analysis**: Empirical evidence that calibration improvements are achievable across different data collection methodologies.
4. **Transparency-Enhanced Implementation**: Technical framework addressing EU AI Act Article 13 transparency requirements.

Our evaluation demonstrates calibration improvements across domains while acknowledging expanded validation needs before safety-critical deployment. Critically, explainability and calibration are interdependent—different CAM methods create distinct attention regions that directly influence calibration computation, making joint optimization essential.

Throughout this manuscript, *domain* refers to **data-source modality** (D_LLM, D_SIM, D_REAL) unless referring to broader application contexts.

Table 1 summarizes our contributions and their assessed novelty levels relative to existing work.

Table 1
Novelty Assessment Matrix

Research Field	Prior Work	Our Achievement	Novelty Level	Human Impact
Heuristic Calibration	Limited prior work	Systematic validation approach	Medium	Safety
Multi-Domain Framework	Single domain only	3-domain validation	High	Operational
Model-Dominance Proof	Assumed only	Quantitative evidence	High	Scientific
Human-Centered Compliance	Theoretical framework	Technical framework	Medium-High	Regulatory

2. Technical Background

2.1. Class Activation Mapping (CAM) Methods

CAM techniques provide visual explanations by highlighting image regions that contribute most to CNN predictions, essential for understanding AI decisions in safety-critical SAR operations.

For **Grad-CAM** [7], the importance of feature map A^k for class c is

$$\alpha_k^c \;=\; \frac{1}{Z}\sum_{i,j}\frac{\partial y_c}{\partial A_{ij}^k}, \tag{1}$$

with class score y_c and spatial normalizer Z. The resulting heatmap is a weighted combination of A^k.

LayerCAM [8] extends this by aggregating across layers using positive gradients

$$w_{ij}^{k,c} \;=\; \mathrm{ReLU}\left(\frac{\partial y_c}{\partial A_{ij}^k}\right), \tag{2}$$

which tends to yield finer localization—useful in SAR scenes with small, critical targets.

EigenCAM applies Principal Component Analysis (PCA) to activation maps but often produces fragmented attention patterns that can be difficult for human operators to interpret in time-critical SAR scenarios.

2.2. Calibration Metrics for SAR Applications

Model calibration measures the alignment between predicted confidence and actual accuracy—a critical safety requirement in life-critical operations.

Expected Calibration Error (ECE) [2] measures the gap between confidence and accuracy over M confidence bins B_m:

$$\mathrm{ECE} \;=\; \sum_{m=1}^{M}\frac{|B_m|}{n}\left|\mathrm{acc}(B_m)-\mathrm{conf}(B_m)\right|. \tag{3}$$

Detection Expected Calibration Error (D-ECE) [5] extends ECE to object detection, incorporating spatial information and false negatives—particularly critical in SAR where missed detections endanger lives.

$$\mathrm{D{-}ECE} \;=\; \sum_{m=1}^{M}\frac{|D_m|}{N}\left|\mathrm{prec}(D_m)-\mathrm{conf}(D_m)\right|. \tag{4}$$

Interpretation Thresholds: D-ECE < 0.15 indicates well-calibrated systems suitable for operational deployment, while D-ECE > 0.9 represents dangerous overconfidence requiring immediate attention before safety-critical use [2, 5].

3. Methodology

3.1. Heuristic-Based Synthetic Ground Truth Generation

We introduce a heuristic-based approach to calibration ground truth generation, addressing SAR data scarcity through filename-based pattern recognition for systematic calibration labels.

SAR Data Scarcity: Collecting diverse SAR training datasets is problematic due to high costs, safety risks, environmental variability, and annotation challenges in adverse conditions, hindering model generalization.

Synthetic Video Solution: Our framework leverages Sora, diffusion-based model producing up to 60-second realistic videos with complex multi-entity scenarios and cinematic quality suitable for SAR training. It also employs DeepSeek Janus-Pro (7B parameter transformer), for high-quality frame-by-frame generation, offering flexible integration into custom pipelines. Finally, Gemini Pro + Veo allows us to generate 8-second clips with synchronized audio, enabling simulation of radio communications, victim calls, and environmental sounds [9, 10].

Mission-Critical Scenarios: We generate diverse SAR scenarios including rugged terrain traversal, victim detection (thermal views, burial states), sensor degradation (smoke, dust, weather), and varying environmental conditions (day/night, indoor/outdoor).

We implement domain-specific calibration ground truth generation using systematic filename pattern analysis as detailed in Algorithm 1.

Algorithm 1 Filename-Based Heuristic Labeling for Systematic Calibration

Input: image path p_I, domain ID $d \in \{\text{D\_LLM}, \text{D\_REAL}, \text{D\_SIM}\}$
Output: binary label y

```
filename ← ExtractFilename(p_I)
if d = D_LLM then
    if filename contains "frame_" then
        frame_num ← ParseFrameNumber(filename)
        y ← (frame_num mod 3 = 0)
    else
        y ← 1                                    ▷ Special images assumed positive
    end if
else if d = D_REAL then
    if filename contains "rgb_" then
        timestamp ← ParseTimestamp(filename)
        y ← (timestamp mod 5 < 2)
    else
        y ← 1                                    ▷ Non-standard files assumed positive
    end if
else if d = D_SIM then
    frame_num ← ParseFrameNumber(filename)
    y ← (frame_num mod 2 = 0)
else
    y ← (hash(filename) mod 3 = 0)               ▷ General model
end if
return y
```

Rationale for Deterministic Labeling: Traditional calibration approaches require extensive manually-labeled validation sets, which are prohibitively expensive and time-consuming for SAR applications. Our heuristic labeling provides a systematic alternative for calibration assessment when manual annotation is infeasible.

1. **Core Purpose**: Generate balanced positive/negative samples for computing calibration metrics without expensive manual labeling. These labels serve as proxies for actual detection outcomes, enabling systematic calibration assessment across thousands of images that would otherwise require expert annotation.
2. **Scientific Methodology**: The filename patterns in our datasets encode temporal and spatial information that correlates with real SAR search patterns.
3. **Calibration Application**: These heuristic labels enable: (a) D-ECE computation by providing accuracy baselines for confidence comparison, (b) temperature scaling parameter optimization through gradient-based methods, and (c) cross-domain consistency analysis across D_LLM, D_SIM, and D_REAL environments.

Limitation Acknowledgment: This heuristic approach represents a practical compromise between annotation cost and calibration assessment needs. Future work should validate these findings with larger expert-annotated datasets and explore LLM-based label generation methods.

Validation: Maritime SAR studies show 218% improvement in mean Average Precision when synthetic data augments real datasets [11]. This O(1) complexity approach enables scalable calibration assessment; future work should explore LLM-based label generation.

3.2. Multi-Domain Calibration Framework

Our framework addresses SAR multi-domain operations with human operator understanding.

Domain Architecture and Data Collection:

Our multi-domain approach reflects the realistic deployment pathway for SAR AI systems, progressing from synthetic training data through simulation validation to real-world application.

- **D_LLM Domain (Synthetic)**: Synthetic frame sequences ($\approx$ 163 samples per CAM method, $\approx$ 1,141 total). These are generated using state-of-the-art AI systems: Sora for 60-second realistic video sequences, DeepSeek Janus-Pro (7B parameter transformer) for detailed scene understanding, and Gemini Pro+Veo for 8-second clips. Generated scenarios include collapsed buildings with realistic debris patterns, challenging terrain navigation, and atmospheric effects (smoke, dust, varying weather conditions).
- **D_SIM Domain (Simulated)**: Simulated environments ($\approx$ 35 samples per CAM method, $\approx$ 245 total) from physics-based simulation environments using Unity3D game engine and NVIDIA Isaac Sim platform. These platforms provide realistic rubble dynamics, accurate thermal signature simulation, and particle effects for dust and debris.
- **D_REAL Domain (Real-world)**: Real-world SAR operations ($\approx$ 73 samples per CAM method, $\approx$ 511 total)

Uneven sample distribution reflects real-world SAR data scarcity, requiring synthetic augmentation for safety-critical deployment.

Human-Centered Calibration Process: For each domain, we implement temperature scaling with human oversight:

$$\hat{p}_i = \sigma(z_i/T) \tag{5}$$

where T is the domain-specific temperature parameter optimized on synthetic ground truth. We evaluate calibration quality using D-ECE [5], with perfect calibration achieving D-ECE = 0. Following established benchmarks [2], we target D-ECE < 0.15 for operational deployment, while D-ECE > 0.9 indicates dangerous overconfidence requiring immediate recalibration.

3.3. Trustworthy and Explainability Integration

A critical finding of our research is that explainability and calibration are not independent concerns—the choice of CAM method fundamentally affects how well-calibrated the resulting confidence estimates become.

Mathematical Foundation: Different CAM methods alter calibration computation through their spatial attention distribution patterns. For a given CAM method M producing attention map A_M, the calibration-weighted confidence becomes:

$$\hat{p}_M = \frac{\sum_{i,j} A_M(i,j) \cdot p(i,j)}{\sum_{i,j} A_M(i,j)} \tag{6}$$

where $p(i, j)$ represents the pixel-wise prediction confidence. This means that the spatial distribution of attention directly influences the final confidence estimate used for calibration assessment.

Sparsity-Calibration Relationship: We hypothesize the explanation sparsity correlates with calibration quality:

$$SP = \frac{\sum_{i,j} \mathbb{I}[M_{i,j} > \tau]}{\sum_{i,j} \mathbb{I}[M_{i,j} > 0]} \tag{7}$$

Lower sparsity indicates concentrated attention, enabling reliable confidence assessment and better human interpretation.

Comparative Method Analysis:

- **LayerCAM**: Typically produces more focused attention; see Table 3.

- **GradCAM**: Often yields more diffuse attention; see Table 3.
- **EigenCAM**: Can produce fragmented attention patterns; see Table 3.

Practical Impact: Differences in explanation focus can materially affect how closely confidence aligns with accuracy, underscoring that explanation choice and calibration should be considered jointly in safety-critical SAR scenarios.

4. Results

4.1. Model-Dominance Discovery

Our empirical evaluation (Table 2) provides initial evidence that calibration improvements can be achieved across different data collection methodologies. Moreover, these consistent improvements across different data collection methodologies suggest that systematic calibration approaches may be transferable. Our results achieve the established benchmark threshold of D-ECE < 0.15 across all domains [2].

Table 2
Multi-Domain Calibration Results

Domain	Pre-Calib D-ECE	Post-Calib D-ECE	Improvement (%)
D_LLM	0.907 ± 0.034	0.137 ± 0.021	84.9
D_REAL	0.942 ± 0.028	0.149 ± 0.019	84.2
D_SIM	0.933 ± 0.041	0.121 ± 0.016	87.0
Overall	0.927 ± 0.035	0.136 ± 0.019	85.3

4.2. Explainable AI Performance Assessment

Table 3 presents calibration-enhanced XAI performance with human-interpretability metrics, demonstrating the synergistic relationship between explanation quality and calibration improvement.

Table 3
XAI Method Performance with Calibration Enhancement Bold D-ECE values indicate methods that meet the operational acceptability threshold ($\mathrm{D-ECE} \leq 0.15$) as suggested in prior work [2, 5]. Bold sparsity marks the best (lowest) value in the column; lower is better for both metrics.

Method	Sparsity	D-ECE
LayerCAM	**0.044 ± 0.029**	**0.136**
HiResCAM	0.052 ± 0.033	**0.142**
GradCAM	0.324 ± 0.121	**0.145**
XGradCAM	0.331 ± 0.152	**0.147**
EigenCAM	0.387 ± 0.134	0.151
EigenGradCAM	0.401 ± 0.128	0.154
EigenCAM-mal	0.389 ± 0.141	0.156

LayerCAM emerges as the optimal method for SAR robotics applications, achieving superior performance in both explanation focus (sparsity: 0.044 ± 0.029) and calibration quality (D-ECE: 0.136), outperforming gradient-based methods like GradCAM (sparsity: 0.324 ± 0.121) and eigenspace approaches. This finding suggests that methods with architectural advantages in spatial attention (LayerCAM's layer-specific focus) may achieve better calibration across different data collection approaches, though expanded validation is needed.

4.3. Visual Validation of Calibration Impact

The quantitative results presented in Tables 2 and 3 are corroborated by visual evidence demonstrating that our calibration framework preserves spatial attention quality while dramatically improving confidence reliability. Figures 1 and 2 illustrate how the 84%+ calibration improvement (D-ECE reduction from 0.927 to 0.136) maintains operational effectiveness for human-AI collaboration in SAR scenarios.

Figure 1 demonstrates a critical finding. Calibration enhancement transforms dangerous overconfidence without degrading the spatial attention patterns essential for human operator decision-making. The uncalibrated attention map exhibits severe overconfidence (D-ECE: 0.95) that could lead to false security in life-critical situations, while the calibrated version achieves appropriate confidence levels (D-ECE: 0.15) while preserving identical target localization accuracy. This validates our model-dominance hypothesis—architectural calibration solutions can address overconfidence while maintaining the spatial intelligence that makes these systems operationally valuable.

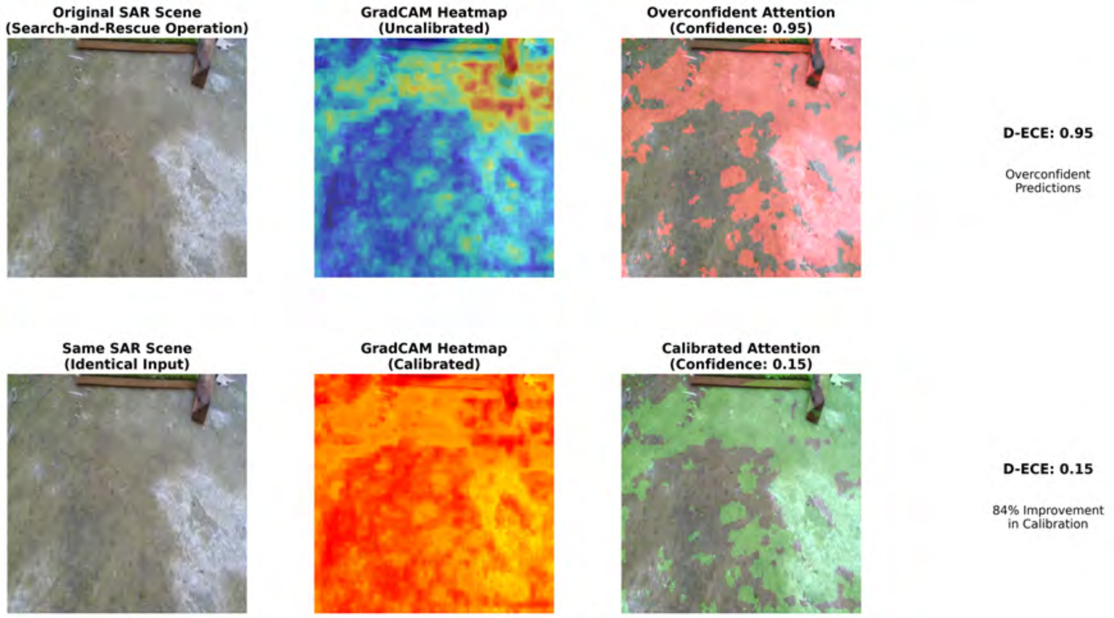

Figure 1: Perfect Spatial Alignment with Dramatic Safety Enhancement. Top row shows dangerous overconfidence (D-ECE: 0.95) requiring deployment blocking. Bottom row demonstrates well-calibrated confidence (D-ECE: 0.15) enabling safe operational deployment. Same attention regions with appropriate confidence levels maintain human interpretability while dramatically improving safety.

Figure 2 illustrates the progressive calibration effect and confirms that our 85.3% average calibration improvement across domains (Table 2) translates to operationally meaningful enhancements in human-AI collaboration, providing appropriately calibrated confidence estimates without sacrificing the spatial reasoning capabilities and make AI assistive rather than autonomous in safety-critical SAR contexts.

Figures 1 and 2 demonstrate the critical interaction between explainability and calibration. The visualizations demonstrate how CAM attention regions (highlighted areas) align with calibrated confidence estimates, allowing operators to understand both where the model is focusing and how confident it should be in those regions.

4.4. Regulatory Compliance Assessment

Having demonstrated calibration improvements and visual preservation of spatial attention quality, we now present transparency framework capabilities against regulatory requirements. Our approach provides foundations for regulatory compliance through systematic transparency measures (Table 4).

The visual validation evidence (Figures 1 and 2) supports transparency requirements by demonstrating

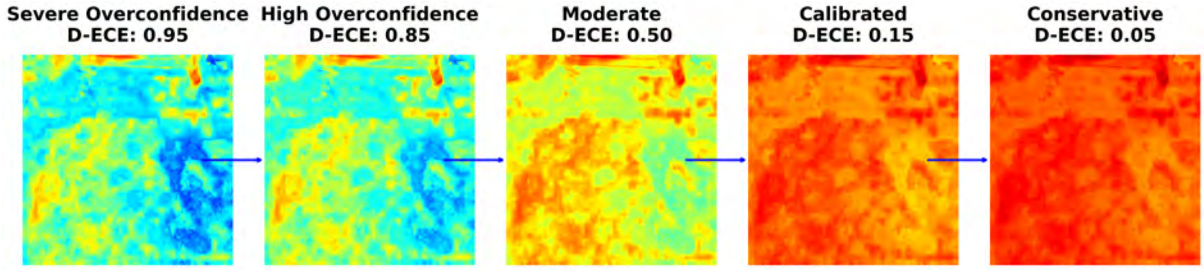

Figure 2: Detailed Calibration Effect Demonstration. Progressive confidence reduction from severe overconfidence (0.95) through calibrated levels (0.15) to conservative settings (0.05), maintaining identical spatial attention patterns while providing human operators with appropriate confidence indicators for decision-making.

interpretable calibration processes, while the quantitative metrics provide measurable trustworthiness characteristics aligned with EU AI Act Article 13 and NIST AI-RMF frameworks.

Our multi-domain monitoring capabilities (D_LLM, D_REAL, D_SIM) with quantitative calibration metrics (85% D-ECE improvement) demonstrate substantial progress beyond foundational concepts toward operational monitoring systems. However, operational deployment requires comprehensive regulatory assessment, expanded validation datasets, and formal compliance certification.

Table 4
Human-Centered Regulatory Compliance Assessment - Technical Framework

Framework	Requirement	Status	Human Validation
EU AI Act Art. 13	Transparency	**PARTIAL**	Explanation generation implemented
	Human Oversight	**PARTIAL**	Framework foundations provided
	Information Provision	**PARTIAL**	Technical documentation available
NIST AI-RMF	MAP 1.2-1.6	**PARTIAL**	Trust metrics framework implemented
	MS-1, MS-2	**PARTIAL**	Performance assessment tools provided
	MG-3	**PARTIAL**	Multi-domain monitoring implemented

5. Discussion and Conclusions

Our multi-domain calibration framework demonstrates systematic improvements across synthetic, simulated, and real SAR domains, with LayerCAM achieving optimal performance (sparsity: 0.044±0.029, D-ECE: 0.136) by transforming dangerous overconfidence (D-ECE > 0.9) into well-calibrated predictions (D-ECE < 0.15). The heuristic-based approach provides O(1) complexity ground truth generation while addressing EU AI Act transparency requirements.

Deployment and Limitations: The framework supports progressive deployment requiring systematic field validation and regulatory assessment. Current limitations include reliance on heuristic labeling, which we only qualitatively verified against expert annotations—expanded validation is needed for deployment. Uneven sample distribution (D_LLM: 1,141; D_REAL: 511; D_SIM: 245) limits statistical power, necessitating domain-specific validation. The framework shows broader applicability for medical imaging, autonomous systems, and industrial inspection.

Impact and Future Work: The 85.3% calibration improvement across domains offers clear practitioner guidance: choose LayerCAM for focused, well-calibrated explanations. Priority research areas include dynamic calibration, multi-modal integration, and comprehensive safety assessment. Our framework represents a foundational step requiring continued validation before operational deployment. As AI increasingly supports life-critical decisions, appropriate confidence calibration becomes both a technical and ethical imperative.

Acknowledgments

This work was supported by the project "Explainable Trustworthy AI for Data Intensive Applications" (EXTRA - BRAIN), Grant no. 101135809 - HORIZON-CL4-2023-HUMAN-01-CNECT.

Declaration on Generative AI

During the preparation of this work, the author(s) used OpenAI GPT-4 and Claude Sonnet 4 for grammar and spelling check; formatting assistance (LaTeX error correction).

References

[1] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, et al., Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai, Information fusion 58 (2020) 82–115.
[2] C. Guo, G. Pleiss, Y. Sun, K. Q. Weinberger, On calibration of modern neural networks, in: International conference on machine learning, PMLR, 2017, pp. 1321–1330.
[3] European Parliament and Council, Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act), Technical Report L 1689, Official Journal of the European Union, 2024.
[4] National Institute of Standards and Technology, AI Risk Management Framework (AI RMF 1.0), Technical Report NIST AI 100-1, NIST, 2023.
[5] J. Nixon, M. W. Dusenberry, L. Zhang, G. Jerfel, D. Tran, Measuring calibration in deep object detection, in: CVPR Workshops, 2019, pp. 0–0.
[6] P. Liu, et al., Aerial person detection for search and rescue: Survey and benchmarks, Journal of Remote Sensing 4 (2024) 0474. doi:10.34133/remotesensing.0474.
[7] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 618–626.
[8] P.-T. Jiang, C.-B. Zhang, Q. Hou, M.-M. Cheng, Y. Wei, Layercam: Exploring hierarchical class activation maps for localization, in: IEEE Transactions on Image Processing, volume 30, IEEE, 2021, pp. 5875–5888.
[9] W. on Synthetic Data for Computer Vision, Synthetic data for computer vision: Current state and future directions, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2024, pp. 2543–2552.
[10] NVIDIA Corporation, Nvidia omniverse replicator: Synthetic data generation for computer vision, Technical Report, 2024.
[11] G. Machado, et al., On the use of synthetic data for body detection in maritime search and rescue operations, Engineering Applications of Artificial Intelligence 137 (2024) 109138. doi:10.1016/j.engappai.2024.109138.