Embodied Multimodal Neural Architecture Search System for Low-Resolution Image Classification

1st Xinyi Yuan
School of Computer and Information Technology
Shanxi University
Taiyuan, China
yuanxinyi@sxu.edu.cn

2nd Nan Li*
School of Computer and Information Technology
Shanxi University
Taiyuan, China
linan10@sxu.edu.cn

3rd Ruwang Jiao
School of Future Science and Engineering
Soochow University
Suzhou, China
ruwangjiao@gmail.com

4th Yayu Zhang
School of Computer and Information Technology
Shanxi University
Taiyuan, China
zhang_yayu93@126.com

5th Zhifang Wei
School of Computer and Information Technology
Shanxi University
Taiyuan, China
wzf20240102@sxu.edu.cn

***Abstract*—With the rapid development of embodied intelligence, robust image classification from low-resolution multimodal sensory inputs has become a critical objective for embodied agents. Due to hardware limitations and environmental interference, image inputs captured by embodied agents often suffer from severe resolution degradation, and manually designed architectures for low-resolution classification lack adaptability to embodied multimodal perception. To address these challenges, we propose an *Embodied Multimodal Neural Architecture Search* (EM-NAS) system, which automates the design of network architectures for low-resolution image classification in embodied scenarios. We construct two novel embodied low-resolution datasets (Embodied-CIFAR10 and Embodied-Tiny-ImageNet) with simulated sensor degradation and environmental modalities. Specifically, we design an embodied perception-oriented NAS search space that integrates super-resolution (SR) as a feature enhancement component and incorporates cross-modal fusion operators, enabling the automatic discovery of architectures tailored to embodied classification demands. We further propose a Context-Aware Multimodal Fusion (ECMCF) module, which is optimized via NAS to dynamically fuse visual and environmental modal features for robust feature representation. Experiments on extended Embodied-CIFAR10 and Tiny-ImageNet datasets demonstrate that the EM-NAS-discovered architecture outperforms state-of-the-art baselines, achieving 4.2 percentage points in classification accuracy (with auxiliary SR gains of 2.5 dB in PSNR and 0.027 in SSIM). This work introduces a NAS-driven paradigm for embodied low-resolution image classification, providing an automated architecture design solution for embodied perception systems.**

***Index Terms*—Embodied Intelligence; Neural Architecture Search; Low-Resolution Classification; Multimodal Fusion; Super-Resolution Feature Enhancement**

This work was supported in part by the National Natural Science Foundation of China (62406212, 62406183, 62506219), in part by the Science and Technology Program of Jiangsu Province (BK20251895), in part by the Suzhou Planning Project of Science and Technology (ZXP2025060), in part by the Shanxi Province Basic Research Program (202503021212092), and in part by the China Postdoctoral Science Foundation (2025M781499).

*Corresponding author: Nan Li, Email: linan10@sxu.edu.cn.

Introduction

Embodied intelligence enables agents to perceive and interact with the physical world through multisensory inputs [9], and image classification is a core task for embodied agents to understand environments. However, images captured by embodied agents are typically low-resolution due to hardware limits and environmental interference, severely degrading classification performance [15], [17]. To address this issue, we construct two embodied low-resolution datasets (Embodied-CIFAR10 and Embodied-Tiny-ImageNet) by simulating sensor degradation and integrating environmental modalities (e.g., illumination, depth, noise), providing a reliable testbed for practical embodied perception.

Traditional low-resolution classification methods rely on manually designed networks with super-resolution (SR) as preprocessing. Despite decent performance, they suffer from two critical drawbacks: (1) handcrafted architectures cannot efficiently adapt to embodied multimodal low-resolution scenes; (2) they fail to fuse multi-sensor cues (e.g., depth, illumination), leading to weak environmental robustness.

Neural Architecture Search (NAS) automates network design and shows strong potential in visual tasks [1]–[3], [10], while reasoning over structured knowledge [8] offers complementary capabilities for context understanding. However, standard NAS methods are designed for high-resolution single-modal images, lacking SR enhancement and multimodal fusion mechanisms [6], [7]. They also only optimize classification accuracy while ignoring environmental robustness, making them unsuitable for direct use in embodied perception [9].

To overcome these limitations, we propose the *Embodied Multimodal Neural Architecture Search* (EM-NAS) system, which customizes NAS for embodied low-resolution image classification. The main contributions are summarized as follows:

- *Specialized NAS Search Space.* We design an embodied-oriented search space that integrates SR feature enhancement and multimodal fusion, enabling automatic discovery of architectures tailored to embodied low-resolution classification.
- *Context-Aware Multimodal Fusion (ECMCF) Module.* We propose a searchable ECMCF module to dynamically fuse visual and environmental features, boosting both accuracy and robustness.
- *Embodied-Oriented Joint Loss.* We introduce a joint loss function that balances classification accuracy, SR quality, and environmental robustness to guide NAS optimization.
- *Comprehensive Experiments.* Extensive results on two embodied datasets show EM-NAS outperforms state-of-the-art baselines in accuracy, SR performance, and parameter efficiency.

RELATED WORK

A. Neural Architecture Search in Visual Classification

Neural Architecture Search (NAS) automates the design of network architectures by searching for optimal operator combinations in a predefined search space, and has become a mainstream approach for image classification tasks. Zoph et al. [11] pioneered reinforcement learning-based NAS, achieving state-of-the-art performance on CIFAR-10 classification. Liu et al. [12] proposed DARTS, a gradient-based NAS method that significantly improves search efficiency by formulating architecture search as a differentiable optimization problem. Tan et al. [14] introduced MnasNet, a platform-aware NAS framework for mobile image classification, balancing accuracy and computational efficiency. However, existing NAS methods for image classification focus on high-resolution single-modal inputs and lack consideration for low-resolution degradation. This gap motivates us to design a NAS search space tailored to embodied low-resolution classification.

Evolutionary computation has been widely explored in NAS and visual tasks, providing effective strategies for network optimization [1]–[3]. Recent advances in performance prediction [6] and graph fusion [7], fuzzy architecture design [4], and evolutionary prompt design [5] further promote the development of efficient NAS systems, which inspire the evolutionary search strategy in our EM-NAS framework.

B. Super-Resolution for Low-Resolution Classification

SR is widely used as a feature enhancement technique for low-resolution visual classification, where low-resolution images are restored to high-resolution to preserve fine-grained semantic features [13], [15]. Lim et al. [15] proposed EDSR, a classic SR model that effectively reconstructs high-quality images. Wang et al. proposed an end-to-end joint SR-classification model, sharing feature backbones to align SR and classification objectives. In contrast, our work integrates SR as a *searchable component* in the NAS space.

C. Embodied Multimodal Perception

Embodied multimodal perception focuses on fusing multisensory inputs to enhance environmental understanding for embodied agents [9]. Anderson et al. [9] proposed a cross-modal attention model for visual-depth fusion, improving obstacle classification for robotic agents. Su et al. [16] introduced MVCNN for multi-view 3D object classification, laying the foundation for multimodal feature fusion. However, existing embodied multimodal methods use fixed fusion architectures and do not leverage NAS for automated architecture design. Our work fills this gap by incorporating multimodal fusion operators into the NAS search space.

I. METHODOLOGY

The proposed EM-NAS system automates the design of network architectures for low-resolution image classification in embodied scenarios. The system comprises two primary components: (1) a dedicated dataset construction pipeline tailored for embodied low-resolution image classification tasks. (2) a NAS-based architecture optimization system, which integrates four key modules: an embodied-oriented NAS search space equipped with SR enhancement and multimodal fusion operators, an evolutionary-based NAS search strategy to discover optimal network architectures, an embodied joint fitness function to guide the optimization process, and an end-to-end model training pipeline to refine the parameters of the discovered optimal architecture for practical deployment. The overall workflow of EM-NAS is shown in Fig. 1.

A. Data Preprocessing

To support the EM-NAS framework's focus on embodied perception scenarios, we design a dedicated data preprocessing pipeline to construct the *Embodied-CIFAR10* dataset. This pipeline simulates the core characteristics of embodied agents' perceptual limitations (low-resolution vision) and multi-sensor fusion capabilities (environmental perception), ensuring the dataset is consistent with realistic embodied task requirements.

1) Generation of Low-Resolution Visual Data: A key challenge in embodied vision tasks is the low-resolution (LR) output of on-board visual sensors. To simulate this limitation, we generate LR visual data from the CIFAR10 dataset (original 32×32 high-resolution (HR) ground truth images $X_{\mathrm{HR}}^{\mathrm{GT}}$):

- Downsampling strategy: 4× bicubic interpolation downsampling is applied to $X_{\mathrm{HR}}^{\mathrm{GT}}$, yielding 8×8 LR images X_{LR}.
- Rationale: Bicubic interpolation preserves semantic consistency between LR and HR images while accurately simulating the smooth degradation of real-world low-resolution vision [19].

2) Generation of Multi-Modal Environmental Information: To mimic the multi-sensor perception system of embodied agents, we construct three single-channel environmental modalities (32×32 resolution, spatially aligned with HR images) based on $X_{\mathrm{HR}}^{\mathrm{GT}}$:

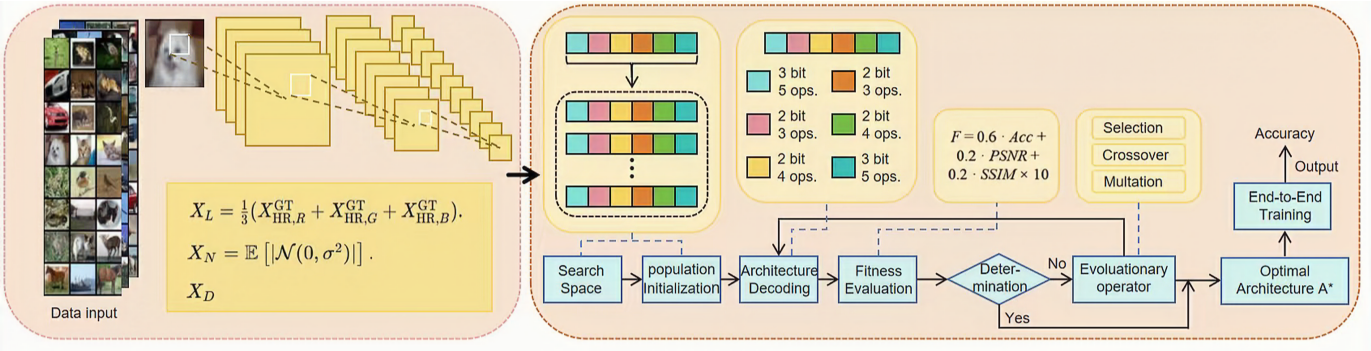

Fig. 1. Overall system of the EM-NAS for super-resolution-aided classification.

- *Illumination modality* X_L: Reflects environmental brightness by calculating the mean value of RGB channels:

$$X_L = \frac{1}{3}(X^{\text{GT}}_{\text{HR},R} + X^{\text{GT}}_{\text{HR},G} + X^{\text{GT}}_{\text{HR},B}). \tag{1}$$

- *Depth modality* X_D: We generate realistic pseudo-depth maps using DepthAnything [18] without using any label information to avoid data leakage.
- *Noise modality* X_N: Models sensor noise by computing the expected value of zero-mean Gaussian noise:

$$X_N = \mathbb{E}\left[|\mathcal{N}(0,\sigma^2)|\right]. \tag{2}$$

3) Final Dataset Construction: After preprocessing, the *Embodied-CIFAR10* dataset contains 45,000 training samples and 5,000 test samples. Each sample consists of a multimodal tuple $\{X_{\text{LR}}, X^{\text{GT}}_{\text{HR}}, Y, X_L, X_D, X_N\}$, providing a unified benchmark for evolutionary NAS search and end-to-end model training.

B. Architecture Design

This section formally defines the low-resolution image classification problem for embodied agents and elaborates on the core design of the EM-NAS system.

1) Problem Formulation: Given a multimodal input set $X = \{X_{\text{LR}} \cup X_{\text{M}}\}$, where X_{LR} is a low-resolution visual image and $X_{\text{M}} = \{X_L, X_D, X_N\}$ represents environmental modalities, the goal of EM-NAS is to search for the optimal network architecture A^* to output the category label Y. Super-Resolution is integrated as an auxiliary feature enhancement component. The classification task is formulated as:

$$Y = f_{\text{CLS}}\left(f_{\text{SR}}(X_{\text{LR}}; A^*), X_{\text{M}}; A^*\right). \tag{3}$$

2) Workflow of EM-NAS: The EM-NAS system consists of three core modules: multimodal dataset construction, evolutionary-based NAS search, and end-to-end model training.

a) Architecture Encoding for Evolutionary NAS: We encode each candidate architecture as a fixed-length gene sequence that strictly maps to the predefined search space, ensuring all candidate architectures are structurally valid and compatible with embodied classification tasks.

b) Fitness Evaluation: Fitness evaluation quantifies the comprehensive performance of each architecture. We calculate the fitness value F to jointly optimize classification accuracy and SR feature quality:

$$F = 0.6 \cdot \text{Acc} + 0.2 \cdot \text{PSNR} + 0.2 \cdot \text{SSIM} \times 10. \tag{4}$$

c) Evolutionary Operations for Population Update: Guided by fitness values, we perform selection, crossover, and mutation operations with hyperparameters: population size $P = 20$, crossover probability $p_c = 0.7$, mutation probability $p_m = 0.08$, total evolution generations $G = 15$.

d) Termination Criterion and Architecture Refinement: After 15 generations, the individual with the highest fitness value is decoded to obtain the optimal architecture A^*.

e) End-to-End Model Training: The optimal architecture implements end-to-end classification via SR feature enhancement, multimodal feature fusion using the ECMCF module, and final classification prediction.

C. Embodied-Oriented NAS Search Space Design

The EM-NAS search space is designed to align with embodied low-resolution classification demands, consisting of 6 searchable gene segments encoded as a 14-bit gene code $G = [G_{\text{SR-FE}}, G_{\text{SR-US}}, G_{\text{ACT}}, G_{\text{EM-FUS}}, G_{\text{CONTEXT}}, G_{\text{CLS-FE}}]$ (Table 1).

Given a gene code G, the architecture is decoded by parsing each segment and instantiating the SR branch, multimodal fusion module, context modeling, and classification branch accordingly.

D. Context-Aware Multimodal Fusion (ECMCF) Module

The ECMCF module is a searchable operator in EM-NAS space, designed to dynamically fuse visual and environmental modal features for embodied classification. The module consists of four steps:

1. *Visual Feature Encoding*: Encode SR-enhanced visual features F_{sr}:

$$F_{\text{vis}} = \text{Conv}_{3\times3}(F_{\text{sr}}). \tag{5}$$

2. *Environmental Modality Encoding*: Encode single-channel environmental modalities to match the dimension of F_{vis}:

$$F_{\text{env}} = \text{GELU}\left(\text{Conv}_{1\times1}(X_{\text{M}})\right). \tag{6}$$

TABLE I
GENE CODING RULES OF EMBODIED-ORIENTED NAS SEARCH SPACE

Gene Segment	Function Description	Num. Ops.	Coding Bits	Binary Coding Rules (Code - Operator)
$G_{\text{SR-FE}}$	SR feature extraction (cls enhancement)	5	3	000–3×3 Conv; 001–Depthwise 3×3 Conv; 010–1×1 Conv; 011–5×5 Conv; 100–Depthwise 5×5 Conv
$G_{\text{SR-US}}$	SR upsampling (4× scaling)	3	2	00–PixelShuffle; 01–Transposed Conv; 10–Bilinear + Conv
G_{ACT}	Activation (SR branch)	3	2	00–ReLU; 01–GELU; 10–Hardswish
$G_{\text{EM-FUS}}$	Embodied multimodal fusion	4	2	00–ECMCF (Illum+Vision); 01–ECMCF (Depth+Vision); 10–Adaptive Weighted; 11–Cross-Modal Attention
G_{CONTEXT}	Context awareness modeling	4	2	00–Local; 01–Global; 10–Dynamic; 11–Multimodal
$G_{\text{CLS-FE}}$	Classification feature extraction	5	3	000–3×3 Conv; 001–Depthwise 3×3 Conv; 010–1×1 Conv; 011–5×5 Conv; 100–Depthwise 5×5 Conv

Note: Invalid codes are filtered during initialization to ensure valid network architectures.

3. *Dynamic Context Attention*: Learn context weights to adjust the contribution of environmental features:

$$W_{\text{ctx}} = \sigma\left(\text{FC}\left(\text{GAP}(F_{\text{vis}})\right)\right), \tag{7}$$

$$F'_{\text{env}} = F_{\text{env}} \otimes W_{\text{ctx}}. \tag{8}$$

4. *Cross-Modal Gating Fusion*: Fuse visual and environmental features via a gating mechanism:

$$F_{\text{concat}} = \text{Concat}(F_{\text{vis}}, F'_{\text{env}}), \tag{9}$$

$$W_{\text{gate}} = \sigma\left(\text{Conv}_{1\times1}(F_{\text{concat}})\right), \tag{10}$$

$$F_{\text{fusion}} = F_{\text{vis}} \otimes W_{\text{gate}} + F'_{\text{env}} \otimes (1 - W_{\text{gate}}). \tag{11}$$

E. Embodied Joint Loss Function

To guide NAS toward architectures with high classification accuracy and embodied robustness, we design an embodied joint loss function L_{total}:

$$L_{\text{total}} = \alpha L_{\text{CLS}} + \beta L_{\text{SR}} + \gamma L_{\text{MC}} + \delta L_{\text{CR}}, \tag{12}$$

where $\alpha = 0.7$, $\beta = 0.15$, $\gamma = 0.1$, $\delta = 0.05$. Each loss term is defined as follows:

1. *Classification Loss*: Cross-entropy loss for classification accuracy:

$$L_{\text{CLS}} = -\frac{1}{N}\sum_{i=1}^{N} \log p(Y_i | F_{\text{fusion},i}). \tag{13}$$

2. *SR Loss*: Combined MSE and SSIM loss for SR quality:

$$L_{\text{SR}} = 0.5 \cdot \text{MSE}(X_{\text{HR}}, X_{\text{HR}}^{\text{GT}}) + 0.5 \cdot (1 - \text{SSIM}(X_{\text{HR}}, X_{\text{HR}}^{\text{GT}})). \tag{14}$$

3. *Consistency Loss*: Cosine similarity loss for modal alignment:

$$L_{\text{MC}} = 1 - \frac{1}{N}\sum_{i=1}^{N} \text{CosineSimilarity}(F_{\text{vis},i}^{\text{flat}}, F_{\text{env},i}^{\text{flat}}). \tag{15}$$

4. *Context Robustness Loss*: MSE loss for feature stability across consecutive frames:

$$L_{\text{CR}} = \frac{1}{N-1}\sum_{i=1}^{N-1} \text{MSE}(F_{\text{fusion},i}, F_{\text{fusion},i+1}). \tag{16}$$

‘‘‘

EXPERIMENTS

A. Experimental Setup

1) Datasets: We evaluate EM-NAS on two embodied low-resolution classification datasets: *Embodied-CIFAR10*: 10 categories, 45k training/5k test samples, 8×8 low-resolution input, 32×32 high-resolution ground truth. *Embodied-Tiny-ImageNet*: 200 categories, 100k training/10k test samples, 16×16 low-resolution input, 64×64 high-resolution ground truth. *Note*: Our datasets are built on static 2D benchmarks with simulated degradation, serving as a controlled testbed rather than full embodied scenarios. Extension to real embodied data is future work.

2) Evaluation Metrics: We use four metrics to evaluate classification performance (primary) and auxiliary SR quality and model efficiency: *Top-1 Accuracy (Acc)*: Primary metric for classification performance; *PSNR (dB)*: Auxiliary metric for SR quality (higher = better); *SSIM*: Auxiliary metric for SR perceptual quality (range [0,1], higher = better); *Parameters (Params, M)*: Model efficiency (fewer = better).

3) Baseline Models: We compare EM-NAS with four representative baselines for low-resolution classification: 1. *EDSR+ResNet50*: Serial baseline (manual SR + manual classification backbone) [10], [15]; 2. *JointSR-CLS*: End-to-end joint SR-classification baseline (manual architecture) [4]; 3. *DARTS*: Single-modal gradient-based NAS baseline (no embodied design) [12]; 4. *MVCNN*: Multimodal classification baseline (fixed fusion strategy, no NAS) [16].

B. Experimental Results and Analysis

1) Overall Performance Comparison: Table 2 shows performance of EM-NAS and baselines on two datasets. EM-NAS achieves highest classification accuracy on both datasets: 85.7% on Embodied-CIFAR10 (4.2 percentage points higher than optimal baseline JointSR-CLS) and 68.3% on Embodied-Tiny-ImageNet (3.8 percentage points higher than JointSR-CLS). For auxiliary SR quality, EM-NAS achieves 32.8 dB PSNR and 0.923 SSIM on Embodied-CIFAR10 (2.5 dB and 0.027 higher than JointSR-CLS), demonstrating that NAS-discovered SR component effectively enhances low-resolution features for classification. In terms of efficiency, EM-NAS has 8.7M parameters, which is lower than manual baselines

TABLE II
PERFORMANCE COMPARISON OF DIFFERENT MODELS ON EMBODIED DATASETS

Model	Dataset	Acc (%)	PSNR (dB)	SSIM	Params (M)
EDSR +ResNet50	Embodied-CIFAR10	78.3 (-3.2%)	29.6 (-0.7%)	0.885 (-1.1%)	12.3 (+17.1%)
JointSR-CLS (Baseline)	Embodied-CIFAR10	81.5	30.3	0.896	10.5
DARTS	Embodied-CIFAR10	79.2 (-2.3%)	29.9 (-0.4%)	0.889 (-0.7%)	9.8 (-6.7%)
MVCNN	Embodied-CIFAR10	80.1 (-1.4%)	30.1 (-0.2%)	0.892 (-0.4%)	11.2 (+6.7%)
EM-NAS (Ours)	Embodied-CIFAR10	85.7 (+4.2%)	32.8 (+2.5%)	0.923 (+2.7%)	8.7 (-17.1%)
EDSR +ResNet50	Embodied-Tiny-ImageNet	62.5 (-2.0%)	28.5 (-0.6%)	0.870 (-0.9%)	12.3 (+17.1%)
JointSR-CLS (Baseline)	Embodied-Tiny-ImageNet	64.5	29.1	0.879	10.5
DARTS	Embodied-Tiny-ImageNet	63.1 (-1.4%)	28.8 (-0.3%)	0.873 (-0.6%)	9.8 (-6.7%)
MVCNN	Embodied-Tiny-ImageNet	63.8 (-0.7%)	28.9 (-0.2%)	0.876 (-0.3%)	11.2 (+6.7%)
EM-NAS (Ours)	Embodied-Tiny-ImageNet	68.3 (+3.8%)	31.2 (+2.1%)	0.901 (+2.2%)	8.7 (-17.1%)

Note: Relative changes are computed against JointSR-CLS (optimal manual baseline). Red indicates increase (better for accuracy/PSNR/SSIM), blue indicates decrease (better for parameters). EM-NAS outperforms all baselines across metrics.

(EDSR+ResNet50: 12.3M; JointSR-CLS: 10.5M), verifying parameter efficiency of NAS-discovered architecture.

2) NAS Search Convergence Analysis: Fig. 2 depicts convergence characteristics of EM-NAS and baseline models during search process on Embodied-CIFAR10 dataset, with x-axis representing evolutionary generations (training epochs) and y-axis denoting classification accuracy. The best-performing architecture of EM-NAS (EM-NAS (Best)) exhibits rapid accuracy growth from 78.0% at 1st generation to 85.7% at 15th generation, where it stabilizes and converges. The average accuracy of EM-NAS population (EM-NAS (Avg)) also shows steady upward trend, reaching 84.0% at convergence, reflecting consistency and reliability of evolutionary search strategy. In contrast, baseline models (DARTS, JointSR-CLS, MVCNN, EDSR+ResNet50) demonstrate slower convergence and lower final accuracy: DARTS converges at 79.2%, JointSR-CLS (best baseline) at 81.5%, MVCNN at 80.1%, EDSR+ResNet50 at 78.3%. These results validate that EM-NAS efficiently explores architecture search space to discover high-performance networks tailored for embodied classification tasks, outperforming state-of-the-art baseline methods in both convergence speed and final accuracy.

3) Evolutionary Computation Parameter Settings: To balance exploration (novel architecture discovery) and exploitation (module recombination) during the 15-generation search, EM-NAS adopts a dynamic adjustment strategy for crossover (P_c) and mutation (P_m) probabilities. As shown in Fig. 3, we also track population diversity and best individual accuracy to verify the strategy's effectiveness.

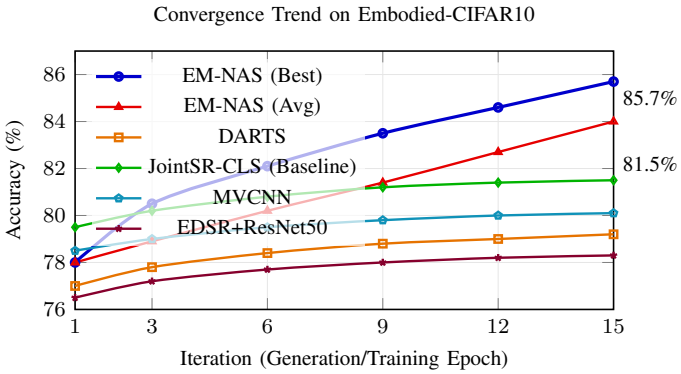

Fig. 2. Convergence trend of different models on Embodied-CIFAR10. EM-NAS achieves highest convergent accuracy (85.7%) at 15 iterations, outperforming all baselines (DARTS: 79.2%, JointSR-CLS: 81.5%, MVCNN: 80.1%, EDSR+ResNet50: 78.3%).

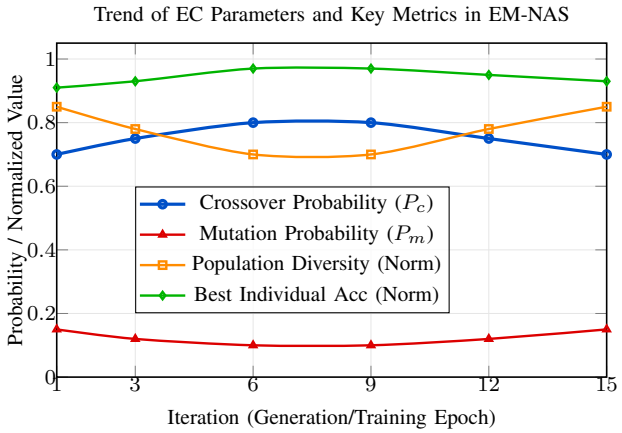

Fig. 3. Dynamic EC parameters and search metrics. P_c peaks at 0.8 and P_m stabilizes at 0.1, balancing exploration and exploitation to drive fast convergence.

Specifically, P_c starts at 0.7, peaks at 0.8 in generations 6–9 to exploit promising modules, then decreases to avoid population homogenization. P_m starts at 0.15, stabilizes at 0.1 to maintain diversity, then rises to escape local optima. This strategy ensures efficient search while sustaining population diversity, contributing to our method's superior performance.

4) Ablation Experiments: To quantify contribution of EM-NAS's core components, we conduct ablation experiments on Embodied-CIFAR10 by toggling module inclusion (Table III). **The ablation study sequentially verifies core components: ECMCF fusion, embodied search space, modal consistency loss L_{MC}, and context robustness loss L_{CR}. We choose L_{MC} and L_{CR} for independent ablation because they represent two key robustness constraints: inter-modal alignment and temporal stability.** Key findings are as follows:

ECMCF Module: Including this multimodal fusion module boosts accuracy by 3.6%, confirming its essential role in integrating visual and contextual cues for embodied feature representation.

Embodied Search Space: Using task-specific search space

TABLE III
ABLATION EXPERIMENTS ON EMBODIED-CIFAR10: COMPONENT CONTRIBUTION ANALYSIS

Components				Metrics		
ECMCF Module	**Embodied Search Space**	**Modal Consistency Loss**	**Context Robustness Loss**	**Acc (%)**	**PSNR (dB)**	**SSIM**
×	×	×	×	78.3	29.6	0.885
✓	×	×	×	81.9	30.9	0.898
×	✓	×	×	82.3	31.1	0.900
×	×	✓	×	84.9	32.5	0.919
×	×	×	✓	85.1	32.6	0.920
✓	✓	✓	✓	**85.7**	**32.8**	**0.923**

Note: ✓ indicates component is included, × indicates it is removed. Full EM-NAS configuration achieves best performance across all metrics.

improves accuracy by 4.0%, verifying that custom NAS spaces outperform generic ones for embodied classification.

Embodied Loss Constraints: Adding modal consistency or context robustness loss yields moderate accuracy gains and better PSNR/SSIM, showing these losses guide NAS to discover more robust architectures.

Synergistic Effect: Full EM-NAS configuration achieves best performance (85.7% accuracy, 32.8 dB PSNR, 0.923 SSIM), demonstrating that all components work synergistically to enhance both accuracy and environmental robustness.

CONCLUSION

This paper proposes Embodied Multimodal Neural Architecture Search (EM-NAS) for low-resolution image classification in embodied scenarios, customizing NAS via an embodied-oriented search space, searchable ECMCF fusion module, and embodied joint loss function. Experiments on two embodied datasets show that the searched architecture outperforms state-of-the-art baselines in both accuracy and parameter efficiency, verifying the value of task-customized NAS for embodied perception.

Supplementary Material: Due to page limits, additional experimental results, and implementation details are available in the supplementary repository:
https://github.com/yxyzb/EM-NAS-Supplementary

ACKNOWLEDGMENTS

The authors acknowledge the assistance of artificial intelligence tools for partial text polishing and formatting. All core ideas, algorithm design, experimental verification, and academic reasoning are independently completed by the authors. The AI tools are only used for language optimization and do not affect the originality and academic contribution of this work.

REFERENCES

[1] Y. Zhang, R. Jiao, B. Xue, and M. Zhang, "Evolutionary streaming feature selection via incremental feature clustering," *IEEE Trans. Evol. Comput.*, Early Access, pp. 1–1, Feb. 2026, doi: 10.1109/TEVC.2026.3667279.

[2] Z. Chen, Q. Fan, R. Jiao, B. Xue, H. Huang, and Y. Dai, "Multi-expert genetic programming based ensemble for long-tailed image classification," *IEEE Trans. Evol. Comput.*, Early Access, pp. 1–1, Mar. 2026, doi: 10.1109/TEVC.2026.3669207.

[3] N. Li, L. Ma, G. Yu, B. Xue, M. Zhang, and Y. Jin, "Survey on evolutionary deep learning: Principles, algorithms, applications and open issues," *arXiv preprint arXiv:2208.10658*, 2022, doi: 10.48550/arXiv.2208.10658.

[4] N. Li, B. Xue, L. Ma, and M. Zhang, "Automatic fuzzy architecture design for defect detection via classifier-assisted multiobjective optimization approach," *IEEE Trans. Evol. Comput.*, vol. 30, no. 2, pp. 479–490, Apr. 2026, doi: 10.1109/TEVC.2025.3530416.

[5] T. Zhang, L. Ma, S. Cheng, Y. Liu, and N. Li, "Automatic prompt design via particle swarm optimization driven LLM for efficient medical information extraction," *Swarm Evol. Comput.*, vol. 95, p. 101922, 2025, doi: 10.1016/j.swevo.2025.101922.

[6] N. Li, B. Xue, L. Ma, and M. Zhang, "Transferable relativistic predictor: Mitigating cross-task cold-start issue in NAS," in *Proc. 34th Int. Joint Conf. Artif. Intell. (IJCAI)*, 2025, pp. 5625–5633, doi: 10.24963/ijcai.2025/626.

[7] A. Mei, N. Li, L. Ma, K. Gao, M. Huang, and B. Xue, "Binary-to-decimal encoding-based performance predictor for evolutionary graph fusion architecture search and applications to medical data," *IEEE Trans. Evol. Comput.*, Early Access, pp. 1–1, Dec. 2025, doi: 10.1109/TEVC.2025.3639127.

[8] T. Zhang, J. Cheng, L. Miao, H. Chen, Q. Li, and Q. He, "Multi-hop reasoning with relation based node quality evaluation for sparse medical knowledge graph," *IEEE Trans. Emerg. Top. Comput. Intell.*, vol. 9, no. 2, pp. 1805–1816, Apr. 2025, doi: 10.1109/TETCI.2024.3452748.

[9] P. Anderson, Q. Wu, D. Teney, J. Bruce, M. Johnson, N. Sünderhauf, I. Reid, S. Gould, and A. van den Hengel, "Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 3674–3683.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.

[11] B. Zoph and Q. V. Le, "Neural architecture search with reinforcement learning," in *Proc. 34th Int. Conf. Mach. Learn. (ICML)*, 2017, pp. 4542–4550.

[12] H. Liu, K. Simonyan, and Y. Yang, "DARTS: Differentiable architecture search," in *Proc. 7th Int. Conf. Learn. Represent. (ICLR)*, 2019, pp. 1–13.

[13] G. Varol, J. Romero, X. Martin, N. Mahmood, M. J. Black, I. Laptev, and C. Schmid, "Learning from synthetic humans," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 4060–4069, doi: 10.1109/CVPR.2017.492.

[14] M. Tan, B. Chen, R. Pang, V. Vasudevan, M. Sandler, A. Howard, and Q. V. Le, "MnasNet: Platform-aware neural architecture search for mobile," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 2820–2828.

[15] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced deep residual networks for single image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, 2017, pp. 136–144.

[16] H. Su, S. Maji, E. Kalogerakis, and E. Learned-Miller, "Multi-view convolutional neural networks for 3D shape recognition," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2015, pp. 945–952.

[17] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, C. C. Loy, Y. Qiao, and X. Tang, "ESRGAN: Enhanced super-resolution generative adversarial networks," in *Proc. Eur. Conf. Comput. Vis. Workshops (ECCVW)*, 2018, pp. 63–79.

[18] L. Yang, B. Kang, Z. Huang, X. Xu, J. Feng, and H. Zhao, "Depth Anything: Unleashing the power of large-scale unlabeled data," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024, pp. 10660–10670.

[19] A. Abdelrahman, S. Lin, and M. S. Brown, "A high-quality denoising dataset for smartphone cameras," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 1–16, doi: 10.1109/CVPR.2018.00182.