Scientific Reports

https://doi.org/10.1038/s41598-026-36445-x

Article in Press

Few-shot cross-episode adaptive memory for metal surface defect semantic segmentation

Received: 16 July 2025

Accepted: 13 January 2026

Published online: 18 January 2026

Cite this article as: Zhang J., Ding H., Peng M. *et al.* Few-shot cross-episode adaptive memory for metal surface defect semantic segmentation. *Sci Rep* (2026). https://doi.org/10.1038/s41598-026-36445-x

Jiyan Zhang, Hanze Ding, Ming Peng, Shuzhen Tu, Guiping Chen & Yanfang Liu

We are providing an unedited version of this manuscript to give early access to its findings. Before final publication, the manuscript will undergo further editing. Please note there may be errors present which affect the content, and all legal disclaimers apply.

If this paper is publishing under a Transparent Peer Review model then Peer Review reports will publish with the final article.

Few-shot Cross-episode Adaptive Memory for Metal Surface Defect Semantic Segmentation

Jiyan Zhang[1], **Hanze Ding**[1], **Ming Peng**[1], **Shuzhen Tu**[1], **Guiping Chen**[2,*], **and Yanfang Liu**[1,*]

[1]College of Mathematics and Information Engineering, Longyan University, Longyan, 364012, China
[2]Longyan Tobacco Industrial Co. Ltd., Longyan, 364000, China
[*]chenguiping_ly@163.com; liuyanfang003@163.com

ABSTRACT

Few-shot semantic segmentation has gained significant attention in metal surface defect detection due to its ability to segment unseen object classes with only a few annotated defect samples. Previous methods constrained to single-episode training suffer from limited adaptability in semantic description of defect regions and coarse segmentation granularity. In this paper, we propose an episode-adaptive memory network (EAMNet) that specifically addresses subtle variances between episodes during training. The episode adaptive memory unit (EAMU) leverages an adaptive factor to model semantic dependencies across different episodes. The context adaptation module (CAM) aggregates hierarchical features of support-query pairs for fine-grained segmentation. The proposed global response mask average pooling (GRMAP) introduces a global response normalization to obtain fine-grained cues directly from the support prototype. We also introduce an attention distillation (AD), which leverages fine-grained semantic attention correspondence to process defect region cues and stabilize the cross-episode adaptation in EAMU. Extensive experiments demonstrate that our approach establishes new state-of-the-art performance on both Surface Defect-4^i and FSSD-12 datasets.

Introduction

Metal surface defect detection[1–3] is an essential quality control process on assembly lines. Different from conventional classification[4–7] and anchor box detection methods[8], semantic segmentation attracted extensive attention recently for its specific location description[9,10]. However, the generalization capability of most metal surface defect segmentation networks is constrained by their dependency on large annotated datasets and poor transferability to unseen classes. Data collection and manual annotation consume substantial time and resources, particularly for dense prediction tasks.

Few-shot semantic segmentation (FSS)[11,12] emerges as a promising approach to address these challenges. Specifically, FSS is formulated as a series of episodes, each containing a support set and a query set. In contrast to supervised semantic segmentation[13], which only segments the same classes presented in the training set, FSS utilizes a few annotated support samples to segment unseen classes in the query set. Therefore, the core of FSS centers on how to best utilize the information in the support set. However, the limited support knowledge available in a single episode compromises the model's adaptability to various defect types and often yields coarse segmentation results. As illustrated in Figure 1(1), traditional methods typically conduct feature extraction and fusion operations confined within a single episode to generate pixel-level mask. When combined with well-designed feature processing, these methods can deliver competitive performance. Nevertheless, the form of single episode hinders the models from learning the category distribution and semantic characteristics of the support set, thus limiting their ability to fully capture the defect patterns. This constraint fundamentally causes both limitations by inducing overfitting to episodic features like the brightness of a scratch. Consequently, models lack semantic adaptability and produce coarse segmentation, as they cannot discern essential defect characteristics.

In this paper, we propose an episode adaptive memory network (EAMNet) to overcome the above mentioned drawbacks. Specifically, as shown in Figure 1(2), we design an episode adaptive memory unit (EAMU) to explore the relations between episodes and conduct the cross-episode interaction. The EAMU generates an adaptive factor to model the semantic dependencies across episodes for defect regions. Further, we propose context adaption module (CAM) and global response mask average pooling (GRMAP) to segment fine-grained defect regions. Considering the effectiveness of deep level features in capturing category distribution and semantic patterns, CAM employs the deep level features of the support-query pair within a single episode to enable fine-grained prediction. Simultaneously, GRMAP further enhance the effectiveness of the support prototype through regulating the state of mask average pooling via global response normalization[14]. The limited discriminability of early-stage features during training poses a significant challenge to these modules, particularly the EAMU, which relies on high-quality multi-level features. To overcome this bottleneck, we present an attention distillation (AD). The AD exploits attention correspondence at fine-grained resolutions, meticulously extracting and transferring key semantic information across

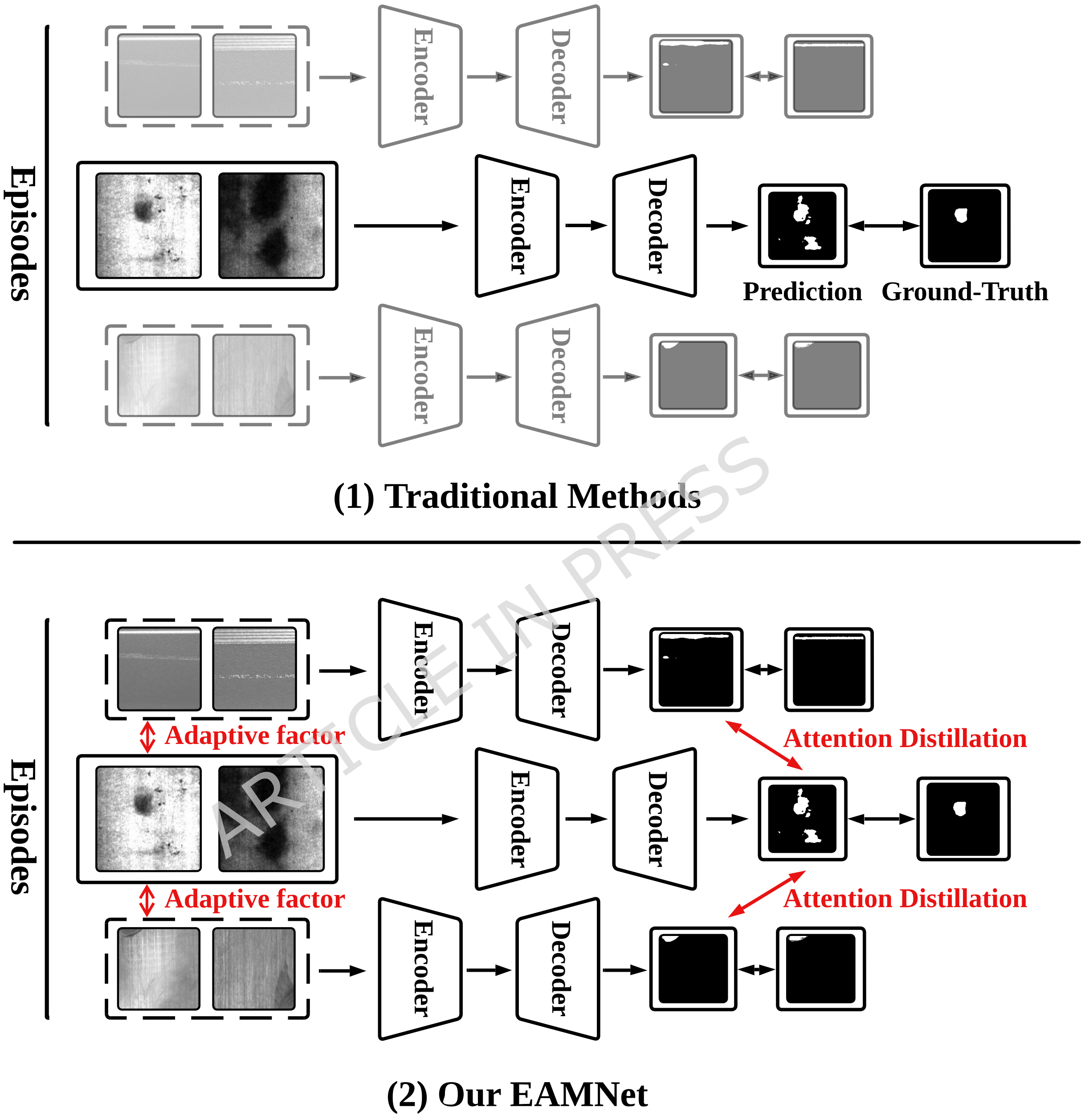

Figure 1. Comparison between (1) Traditional Methods and (2) Our EAMNet. (1) Traditional methods extract semantic knowledge from a single episode, which is not able to learn the holistic semantic patterns. (2) Our EAMNet utilizes an adaptive factor and an attention distillation to capture the semantic knowledge across episodes.

soft-target from attention map. This mechanism empowers the model to analyze the semantic cues across episodes rapidly and adapt to diverse scenarios effectively.

Our contributions can be summarized as follows:

1) An EAMU generates a cross-episode adaptive factor to exploit the semantic dependence across episodes for metal surface defect regions.

2) We propose a CAM and a GRMAP leveraging hierarchical features and global response normalization respectively to accomplish fine-grained segmentation.

3) To accelerate cross-episode semantic learning and enhance generalization, we introduce an AD that transfers fine-grained semantic attention from high-resolution features.

4) Our EMANet sets new state-of-the-art results on both standard benchmarks: Surface Defect-4^i and FSSD-12.

Related Work

Metal Surface Defect Detection

In metal surface defect detection, image classification, object detection, and semantic segmentation represent the fundamental computer vision approaches. Notably, semantic segmentation-based methodologies, which offer pixel-level precision in prediction outcomes, have recently garnered significant attention. Chao et al.[15] conduct an information augmentation and multiscale feature fusion algorithm based on YOLOv8, which uses the information augmentation network to reduce the information loss during feature downsampling extraction. Several studies employ unsupervised learning methods to address metal surface defect challenges[16,17]. Meanwhile, alternative approaches leverage attention mechanisms[18,19] to generate defect attention maps for enhanced localization. Zhang et al.[20] construct the innovative triple-attention mechanism to enhance the models' ability of expressing defect characteristics. Wei et al.[21] introduce a vision transformer model that combines receptive-field attention convolution and context broadcasting median to handle the high variability and sample imbalance in metal defects. Zhang et al.[22] design a dual-branch local-guided global self-attention network to achieve sufficient attention in local details. Jin et al.[23] construct a human guidance to address the data scarcity and effectively leverage expert knowledge. Zhao et al.[24] propose a cross-supervised contrastive learning domain adaptation network with transformer to solve the differences that steel defects exhibit in appearance and background. The aforementioned detection methods fail to address the critical issue of labeled sample scarcity in metal defect inspection. The unsupervised and semi-supervised approaches offer partial remedies but struggle with reliability. To directly overcome these limitations, our EAMNet introduces a few-shot learning paradigm, which systematically addresses sample scarcity through meta-learning and achieves reliable generalization via explicit prototype-based learning within each episode.

Few-Shot Semantic Segmentation

Few-shot semantic segmentation (FSS) generates dense masks for novel classes using minimal annotated support samples. Extending the prototypical learning framework from[25], contemporary FSS approaches[26,27] construct class-specific prototypes to capture discriminative feature representations. Yang et al.[28] propose a bi-orientated rectification few-shot segmentation network based on fine-grained prototypes to address the limitations of current methods that only extract general target prototype. Chen et al.[29] conduct a mask generation module and an iterative refinement module, respectively addressing the inherent two challenges of locating segmented objects and deriving class-specific features in the absence of support mask and semantic labels. Wang et al.[30] leverage semantic word embedding and query set self-supplementing information to handle the inter-class interference and information loss of generalized few-shot semantic segmentation. While these prototype-based methods remain confined to a single episode, they overlook the valuable semantic information across episodes. To overcome the single-episode limitation of these methods, EAMNet employs an EAMU to model cross-episode correlations, which then enables its CAM, GRMAP, and AD to perform fine-grained segmentation.

Knowledge Distillation

Knowledge distillation (KD)[31] is a widely-adopted technique for model compression. Larger models typically exhibit greater representational capacity, while compact models feature fewer parameters, higher computational efficiency, and lower deployment costs. The core objective of KD is to transfer knowledge from a high-capacity teacher model to a lightweight student model while maintaining acceptable performance levels. The approach in [32] implements a self-distillation framework to perform internal knowledge transfer which can enhance the model accuracy without external supervision. Self-distillation operates by partitioning the model into multiple components and facilitating the transfer of knowledge from deeper layers to shallower ones. Additionally, self-distillation is employed to exploit the intrinsic semantic correlations of models. Li et al.[33] utilize self-distillation from the last batch during model training to maximize the performance. Lu et al.[34] introduce a multi-stage dynamic anchor decoder that aggregation capabilities of the sparse attention mechanism to improve the effectiveness of self-distillation. Peng et al.[35] leverage the hierarchical attention maps to create self-distillation resolution. Motivated by[34,35], we design an attention distillation conducting the self-distillation of attention maps and ground-truth to enhance the adaptability of our model.

Method

Problem Setting

Our implementation adopts the standard few-shot segmentation paradigm[36,37] based on episodic meta-learning. We partition the dataset into training set D_{train} and testing set D_{test}, each containing numerous randomly sampled episodes. Each episode comprises a support set $\mathscr{S}$ and query set $\mathscr{Q}$ from the same class c. During meta-training, models learn on classes C_{train} and are evaluated on disjoint unseen classes C_{test} ($C_{\text{train}} \cap C_{\text{test}} = \varnothing$). The FSS model processes $\mathscr{S} = \left\{ \left(I_S^i, M_S^i \right) \right\}_{i=1}^{K}$ and $\mathscr{Q} = \{ (I_Q, M_Q) \}$ from class $c \in C_{\text{train}}$ to predict query mask $\hat{\boldsymbol{M}}_Q$, where $I_S^i, I_Q \in \mathbb{R}^{3 \times H \times W}$ are RGB images. And $\boldsymbol{M}_S^i, \boldsymbol{M}_Q \in \{0,1\}^{H \times W}$ are binary masks, where $H \times W$ denotes spatial dimensions. During training, both $\boldsymbol{M}_Q$ and $\boldsymbol{M}_S$ are utilized. At test time, only $\boldsymbol{M}_S$ is available as input. Crucially, the meta-trained model directly generalizes to novel categories without test-time optimization, as it leverages support semantic cues to localize query regions of interest.

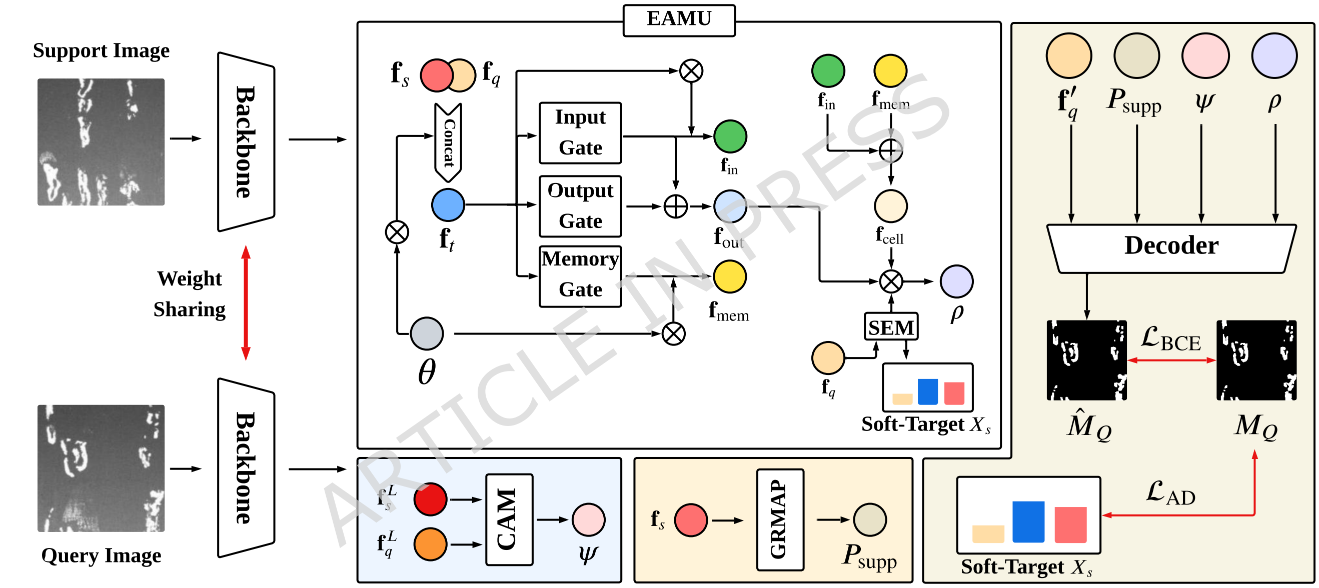

Figure 2. Overview of our Episode Adaptive Memory Network with the EAMU, CAM, GRMAP and AD. Where $\otimes$ denotes matrix multiplication and $\oplus$ represents matrix addition.

Overview

As illustrated in Figure 2, our Episode Adaptive Memory Network (EAMNet) comprises four core components: episode adaptive memory unit (EAMU), context adaptive module (CAM), global response mask average pooling (GRMAP), and attention distillation (AD). More precisely, for the support image I_S and query image I_Q, we employ a shared backbone network to extract L-level features concurrently, where $L = 4$. Subsequently, we introduce the EAMU, which is designed to generate an adaptive factor ρ within the given support-query image pair. This factor, which captures and retains transferable meta-knowledge across different tasks, operates at the cross-episode level Meanwhile, high-resolution features are processed by CAM, which aims to produce context adaptive factor ψ for support-query pairs through capturing category distribution and semantic patterns. CAM functions within a single episode, specializing in fine-grained and spatial context alignment between the current support and query features. To achieve precise prediction, we then employ the GRMAP, which learns to generate the prototype $\boldsymbol{P}_{\text{supp}}$ from support foreground feature by the global response normalization. GRMAP acts as a support enhancer prior to the query adaptation process and focus on producing a robust and globally-aware prototype. We then feed the ρ, ψ, $\boldsymbol{P}_{\text{supp}}$ and the query feature $\mathbf{f}_q$ filtered by the activation function into a decoder to predict the final segmentation mask

$\hat{\boldsymbol{M}}_Q$ for the query image. Then the model can be trained under the supervision of a binary cross-entropy loss $\mathscr{L}_{\text{BCE}}$ computed between the ground-truth mask $\boldsymbol{M}_Q$ and the predicted mask $\hat{\boldsymbol{M}}_Q$. Moreover, we design an extra AD in the end of training stage to enhance the semantic expression at fine-grained resolutions by $\mathscr{L}_{\text{AD}}$. It is crucial to note that AD is an optimization-oriented module distinct from the inference path, designed exclusively to accelerate convergence through knowledge distillation across network stages. Next, we provide a detailed elaboration of each of the previously mentioned modules.

Episode Adaptive Memory Unit

Inspired by SG-One[38], recent FSS models for metal surface defect[39,40] usually leverage the prototypes from the support and query set within a single episode to generate the prediction mask. As the constrained knowledge provided within a single episode, the cross-episode semantic description of defect regions for prototypes is not incorporated into the model. To counter this, the EAMU is employed to derive a cross-episode adaptive factor from the cumulative support-query pairs across episodes by gate units[41], enabling the model to adaptively remember inter-episode semantic dependencies. Different from sequence-modeling memories (e.g., LSTM[41]) or static global memories (e.g., Transformer memory[42]), our EAMU achieves an episode-level memory for meta-learning. Designed for meta-learning, our EAMU operates at an episodic granularity, maintaining a dynamic memory seed that updates only upon episodic task completion. Consequently, the gate units of EAMU consolidate cross-episode experience and guide adaptation, thereby explicitly addressing inter-episode distribution shifts rather than modeling sequences or storing global knowledge. As illustrated in Figure 2, the EAMU receives as input the current episode's support-query feature pair $\mathbf{f}_s, \mathbf{f}_q \in \mathbb{R}^{c\times h\times w}$ where c, h and w denotes the channels, heights and width of the features respectively. Note that the support-query feature pair $\mathbf{f}_s$ and $\mathbf{f}_q$ are generated by the feature pyramid network[43], which conduct the first three level feature maps to obtain the contextual information. Then EAMU employs an architecture comprising an input gate $\mathscr{F}_{\text{in}}$, output gate $\mathscr{F}_{\text{out}}$, and memory gate $\mathscr{F}_{\text{mem}}$ to adaptively retain the semantic details within the current episode, where the $\mathscr{F}_{\text{in}}$, $\mathscr{F}_{\text{out}}$ and $\mathscr{F}_{\text{mem}}$ are comprised of various convolution and activation functions. To perform the cross-episode knowledge, the input gate $\mathscr{F}_{\text{in}}$ combines the current episode features $\mathbf{f}_t \in \mathbb{R}^{2c\times h\times w}$ and the memory seed $\boldsymbol{\theta} \in \mathbb{R}^{h\times w}$. It outputs a condensed feature $\mathbf{f}_{\text{in}} \in \mathbb{R}^{1\times h\times w}$ that encapsulates the novel semantic information learned from the current episode:

$$\mathbf{f}_t = \mathscr{C}\left(\mathbf{f}_q, \mathbf{f}_s\right) \otimes \boldsymbol{\theta}, \tag{1}$$

$$\mathbf{f}_{\text{in}} = \mathscr{F}_{\text{in}}\left(\mathbf{f}_t\right) \otimes \mathbf{f}_t, \tag{2}$$

where $\mathscr{C}$ is the concatenation operation, $\otimes$ represents matrix multiplication and $\boldsymbol{\theta}$ is a trainable parameter. Then, the output of semantic information $\mathbf{f}_{\text{out}} \in \mathbb{R}^{1\times h\times w}$ in the current episode is processed by output gate $\mathscr{F}_{\text{out}}$:

$$\mathbf{f}_{\text{out}} = \mathscr{F}_{\text{out}}\left(\mathbf{f}_t\right) \oplus \mathscr{F}_{\text{in}}\left(\mathbf{f}_t\right), \tag{3}$$

where the $\oplus$ denotes matrix addition.

Having the memory seed $\boldsymbol{\theta}$, we further process the cross-episode semantic description through memory gate $\mathscr{F}_{\text{mem}}$:

$$\mathbf{f}_{\text{mem}} = \mathscr{F}_{\text{mem}}\left(\mathbf{f}_t\right) \otimes \boldsymbol{\theta}, \tag{4}$$

where $\mathbf{f}_{\text{tmem}} \in \mathbb{R}^{1\times h\times w}$ initializes the memory information of defect regions, and $\mathbf{f}_{\text{mem}} \in \mathbb{R}^{1\times h\times w}$ leverages $\boldsymbol{\theta}$ to obtain the holistic cross-episode semantic description. To achieve the corss-episode interaction, we integrate the new feature $\mathbf{f}_{\text{in}}$ with the cross-episode semantic description $\mathbf{f}_{\text{mem}}$ to generate the cell information $\mathbf{f}_{\text{cell}} \in \mathbb{R}^{1\times h\times w}$:

$$\mathbf{f}_{\text{cell}} = \mathbf{f}_{\text{in}} \oplus \mathbf{f}_{\text{mem}}. \tag{5}$$

Next, EAMNet utilizes a squeeze & excitation module (SEM)[44] to further process the fine-grained semantic information from query feature in current episode:

$$\rho = \mathbf{f}_{\text{cell}} \otimes \mathbf{f}_{\text{out}} \otimes \mathscr{F}_{\text{SEM}}\left(\mathbf{f}_q\right), \tag{6}$$

where $\mathscr{F}_{\text{SEM}}$ is the SEM, $\rho \in \mathbb{R}^{1\times h\times w}$ denotes the adaptive factor which contains cross-episode semantic details and cues.

Context Adaptive Module

Given the feature maps of $L = 4$ level processed by backbone, i.e., $\left\{\mathbf{f}_q^l\right\}_{l=1}^{L} \in \mathbb{R}^{lc\times\frac{h}{2^{l+2}}\times\frac{w}{2^{l+2}}}$ and $\left\{\mathbf{f}_s^l\right\}_{l=1}^{L} \in \mathbb{R}^{lc\times\frac{h}{2^{l+2}}\times\frac{w}{2^{l+2}}}$, the contextual category distribution and semantic patterns can be captured through deep level feature maps. Specifically, we calculate the cosine similarity $\boldsymbol{D}_{\cos}^{l} \in \mathbb{R}^{1\times\frac{h}{2^{l+2}}\times\frac{w}{2^{l+2}}}$ of last two-stage support-query pairs $\left\{\mathbf{f}_q^l\right\}_{l=L-1}^{L}$ and $\left\{\mathbf{f}_s^l\right\}_{l=L-1}^{L}$ for fine-grained semantic description, i.e., $\boldsymbol{D}_{\cos}^{L}$ and $\boldsymbol{D}_{\cos}^{L-1}$. As well as the procedure for multiple stage pairs follows analogously:

$$\boldsymbol{D}_{\cos}^{L} = \mathscr{F}_{\mathrm{SIM}}\left(\mathbf{f}_q^L, \mathbf{f}_s^L \otimes \mathscr{R}_{\mathrm{mask}}\left(\boldsymbol{M}_S\right)\right), \tag{7}$$

where $\mathscr{F}_{\mathrm{SIM}}$ calculates the cosine similarity of support-query pair, $\mathscr{R}_{\mathrm{mask}}$ reshapes the support mask $\boldsymbol{M}_S$ to be the same shape as $\mathbf{f}_s^{\mathrm{high}}$. Then we integrate the $\boldsymbol{D}_{\cos}^{L}$ and $\boldsymbol{D}_{\cos}^{L-1}$ to generate the context adaptive factor $\psi \in \mathbb{R}^{2\times h\times w}$ which depicts the contextual semantic pattern of the defect regions in current episode:

$$\psi = \mathscr{R}_{\mathrm{factor}}\left(\lambda\left(\mathscr{C}\left(\boldsymbol{D}_{\cos}^{L}, \mathscr{R}_{\cos}\left(\boldsymbol{D}_{\cos}^{L-1}\right)\right)\right) \otimes \beta\right), \tag{8}$$

where λ is an activation function, β controls the adaptive weight of ψ during downstream and is initialized to 0.4, $\mathscr{R}_{\cos}$ reshapes the $\boldsymbol{D}_{\cos}^{L-1}$ to be the same shape as $\boldsymbol{D}_{\cos}^{L}$ and $\mathscr{R}_{\mathrm{factor}}$ reshapes the input shape to be the $\mathbb{R}^{2\times h\times w}$.

Global Response Mask Average Pooling

Following the general form[35, 38], the key of mask average pooling is the knowledge of support set. However, current methods that take support features as input directly suffer from coarse-grained foreground knowledge. Thus, we introduce the global response normalization[14] to enhance the effectiveness of support knowledge. Specifically, given the c-channel support feature $\mathbf{f}_s = \{\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_i\}_{i=1}^{c}$, we employ a response normalization to compute the feature normalization scores $\mathscr{N} \in \mathbb{R}^{c\times 1\times 1}$:

$$\mathscr{N} = \frac{\|\mathbf{z}_i\|_2}{\sum_{j=1}^{c}\|\mathbf{z}_j\|_2}, \tag{9}$$

where $\|\mathbf{z}_i\|_2$ is the L2-norm of the i-th channel. Subsequently, we calibrate the original input features based on $\mathscr{N}$ to generate the responsive support feature $\mathbf{f}_s' \in \mathbb{R}^{c\times h\times w}$:

$$\mathbf{f}_s' = \left(\mathbf{f}_s \otimes \eta\mathscr{N}\right) + \tau, \tag{10}$$

where η and τ are learnable parameters. For holistic semantic information, we also produce the responsive query feature $\mathbf{f}_q' \in \mathbb{R}^{c\times h\times w}$ through Eqs. (9)-(10) in the downstream work.

Then we apply the mask average pooling to capture the guide prototype $\boldsymbol{P}_{\mathrm{supp}} \in \mathbb{R}^{c\times h\times w}$ from $\mathbf{f}_s'$:

$$\boldsymbol{P}_{\mathrm{supp}} = \mathscr{F}_{\mathrm{AVG}}\left(\mathbf{f}_s' \otimes \boldsymbol{M}_S\right), \tag{11}$$

where $\mathscr{F}_{\mathrm{AVG}}$ is a 2D average pooling operation.

The enhanced query feature $\mathbf{f}_q'$, guide prototype $\boldsymbol{P}_{\mathrm{supp}}$, context adaptive factor ψ and cross-episode adaptive factor ρ are all reshape to the same spatial size and concatenated to a representative prototype $\boldsymbol{P}_{\mathrm{final}} \in \mathbb{R}^{(2c+3)\times h\times w}$:

$$\boldsymbol{P}_{\mathrm{final}} = \mathscr{C}\left(\mathbf{f}_q', \boldsymbol{P}_{\mathrm{supp}}, \psi, \rho\right). \tag{12}$$

Finally, $\boldsymbol{P}_{\mathrm{final}}$ is fed into a decoder to produce segmentation mask $\hat{\boldsymbol{M}}_Q$:

$$\hat{\boldsymbol{M}}_Q = \mathscr{F}_{\mathrm{cls}}\left(\mathscr{F}_{\mathrm{conv}}\left(\boldsymbol{P}_{\mathrm{final}}\right)\right), \tag{13}$$

where $\mathscr{F}_{\mathrm{conv}}$ and $\mathscr{F}_{\mathrm{cls}}$ are two consecutive modules that constitute the decoder.

Attention Distillation

The construction of the cross-episode adaptive factor relies on features from all network stages. However, in early training, the shallow layers yield coarse-grained features due to limited supervision, which slows model convergence and degrades the quality of factor. To address this, we introduce an attention distillation, which refines these early-stage features by transferring knowledge from deeper layers, thereby preserving fine-grained information without compromising cross-episode learning. Moreover, compared to the one-hot labels in ground-truth annotations, the soft targets used in distillation capture inter-category dependencies, thereby providing richer semantic details within individual episodes. Specifically, we apply the $\mathscr{F}_{\text{SEM}}$ to collect the soft-target $\boldsymbol{X}_s$ of query feature $\mathbf{f}_q$:

$$\boldsymbol{X}_s = \mathscr{F}_{\text{softmax}}\left(\mathscr{F}_{\text{SEM}}\left(\mathbf{f}_q\right)\right), \tag{14}$$

where $\mathscr{F}_{\text{softmax}}$ is a softmax layer. Then the KL (Kullback-Leibler) divergence loss is used as supervision from the teacher to student with their softmax output:

$$\mathscr{L}_{\text{AD}} = \boldsymbol{X}_t \log\left(\boldsymbol{X}_t\right) - \boldsymbol{X}_t \log\left(\boldsymbol{X}_s\right), \tag{15}$$

where $\boldsymbol{X}_t$ is the ground-truth $\boldsymbol{M}_Q$.

Finally, the loss function of our model during training can be formulated as:

$$\mathscr{L} = \mathscr{L}_{\text{BCE}} + \alpha \mathscr{L}_{\text{AD}}, \tag{16}$$

where α is a hyperparameter which set to 0.05.

Extend to K-shot Settings

Thus far, our discussion has focused on the one-shot setting, as summarized in Figure 2. To extend EAMNet to the K-shot scenario (where K support images per category are available), we leverage the per-episode adaptive factor. To preserve spatial information and maintain consistency across varying shot counts, we directly concatenate the support features $\mathbf{f}_s^{l} = \mathscr{C}\left(\mathbf{f}_s^{l,1}, \mathbf{f}_s^{l,2}, \ldots, \mathbf{f}_s^{l,K}\right)$ along the channel dimension, along with their corresponding masks.

Experiments

Datasets

Following the setting of[37,45,46], we use metal surface defect datasets, i.e., Surface Defect-4^i[37] and FSSD-12[45], to evaluate EAMNet. Surface Defect-4^i is a general dataset containing 12 various classes of metal surface defect. FSSD-12 is a strip steel surface defect dataset and also contains 12 categories.

For each dataset, we perform 3-fold cross-validation by partitioning all classes equally. Following[37,45], we maintain identical class splits for Surface Defect-4^i and FSSD-12. Two folds serve as training data, with the remaining fold reserved for testing.

Metric and Evaluation

Following established practice, we adopt mean Intersection-over-Union (mIoU) and foreground-background IoU (FB-IoU) as evaluation metrics. For testing, we follow the protocol from[47] to ensure experimental validity. Specifically, each experiment consists of five independent trials using distinct random seeds, with final results representing the average across these trials.

Implementation Details

EAMNet is implemented in PyTorch and trained episodically for 200 epochs on Surface Defect-4^i and FSSD-12. All models are trained on 4 NVIDIA GeForce RTX 3090 GPUs with batch size 2, and tested on a single GPU with batch size 1. The optimizer employed is consistent with[37], and the learning rate is set to 0.0001. To validate our method's backbone-agnostic efficacy, we evaluate it with ResNet-50[48] and VGG-16[49] backbone. During inference, predictions are resized to match original input resolution while preserving ground-truth labels.

Baseline

First, we exclude the EAMU, CAM, GRMAP, and AD from EAMNet. Then, we replace the enhanced query feature $\mathbf{f}_q'$ in the downstream network with the original query feature $\mathbf{f}_q$ to construct the baseline model. The remaining experimental settings are kept consistent with those of EAMNet.

Comparison with State-of-the-Art Methods

Surface Defect-4^i

Table 1 compares mIoU performance on the Surface Defect-4^i dataset between our method and several representative models. The results demonstrate that: (1) EAMNet achieves state-of-the-art (SOTA) performance under both 1-shot and 5-shot settings. Notably, when using the VGG-16 backbone, EAMNet surpasses MAPTNet[46] (the previous SOTA) by significant margins of 4.43% and 5.89% for 1-shot and 5-shot, respectively, which demonstrates that the effective cross-episode adaptation and intra-episode feature processing of EAMNet are key to its superior performance. (2) EAMNet substantially outperforms the

Table 1. Comparison of EAMNet with state-of-the-art methods and semantic networks for metal surface defect FSS on Surface Defect-4^i, evaluated by mIoU and FB-IoU under 1-shot and 5-shot settings using the VGG-16 and ResNet-50 backbone. The **best** and second best results are highlighted accordingly. Improv. (%) represents the percentage improvement in mIoU over the *Baseline.*

Methods	Backbone	1-shot						5-shot					
		Fold-0	Fold-1	Fold-2	Mean	FBIoU	Improv. (%)	Fold-0	Fold-1	Fold-2	Mean	FBIoU	Improv. (%)
HDMNet[35]	VGG16	31.10	28.91	21.26	27.09	52.76	↑4.01	42.33	28.00	25.97	32.10	51.45	↑4.47
DCPNet[36]		28.68	27.45	24.08	26.74	52.22	↑3.65	22.91	27.81	25.59	25.44	51.96	↓2.20
TGRNet(1-normal)[37]		29.78	25.15	24.36	26.43	51.50	↑3.35	37.42	24.66	26.52	29.53	53.27	↑1.90
CPANet[45]		22.03	25.05	24.07	23.72	51.35	↑0.63	30.11	25.95	19.26	25.11	52.21	↓2.53
MAPTNet[46]		39.88	25.98	**26.24**	30.70	50.76	↑7.62	39.93	31.66	29.82	33.80	56.33	↑6.17
PFENet[47]		23.28	19.45	20.48	21.07	51.14	↓2.01	27.94	21.67	25.24	24.95	53.99	↓2.68
Baseline		*28.33*	*24.54*	*16.38*	*23.08*	*49.12*	-	*38.11*	*24.87*	*19.92*	*27.63*	*53.67*	-
EAMNet		**47.01**	**32.91**	25.46	**35.13**	**54.57**	↑**12.04**	**54.90**	**32.64**	**31.53**	**39.69**	**58.28**	↑**12.06**
HDMNet[35]	ResNet50	35.58	**40.79**	27.50	34.62	56.01	↑8.41	38.62	41.11	**32.61**	37.45	56.19	↑9.76
DCPNet[36]		27.19	31.96	24.68	27.94	51.67	↑1.73	42.78	39.35	32.21	38.11	58.77	↑10.43
TGRNet(1-normal)[37]		35.46	32.37	24.75	30.86	53.62	↑4.65	41.61	28.66	27.87	32.71	53.00	↑5.03
CPANet[45]		32.52	29.65	24.66	28.94	51.94	↑2.73	39.36	37.84	27.82	35.01	57.73	↑7.32
MAPTNet[46]		41.27	40.20	22.78	34.75	55.61	↑8.54	46.49	40.85	26.20	37.85	58.22	↑10.16
PFENet[47]		29.45	24.90	16.21	23.52	54.06	↓2.69	33.98	30.07	22.78	28.94	56.92	↑1.26
HMNet[50]		39.49	28.10	23.73	30.44	53.97	↑4.23	44.00	30.19	27.33	33.84	56.74	↑6.15
Baseline		*34.84*	*24.17*	*19.63*	*26.21*	*49.42*	-	*34.55*	*24.53*	*23.98*	*27.69*	*50.66*	-
EAMNet		**44.47**	40.39	**28.38**	**37.75**	**59.25**	↑**11.53**	**51.85**	**41.25**	31.76	**41.62**	**59.23**	↑**13.93**

baseline. For instance, with the VGG-16 backbone, EAMNet achieves 35.13% mIoU compared to the baseline's 23.08%. This improvement stems from the synergistic effect of its core components, with EAMU handling cross-episode adaptation, CAM managing intra-episode refinement, GRMAP responsible for support enhancement, and AD ensuring stable convergence.

EAMNet achieves competitive efficiency with 54.47G FLOPs, indicative of its lightweight and adaptable design for diverse tasks. A detailed comparison of computational costs with TGRNet, CPANet, and MAPTNet is presented in Table 2.

Table 2. Comparison with metal surface defect FSS in computational cost on Surface Defect-4^i. The **best** results are highlighted accordingly.

Methods	mIoU	FLOPs	#Params.
TGRNet	26.43	83.69G	**9.38M**
CPANet	23.72	162.23G	11.98M
MAPTNet	30.70	66.80G	16.80G
EAMNet	**35.13**	**54.47G**	15.79M

FSSD-12

FSSD-12 is an extra validation dataset for the generalization of our model which only contains strip steel surface defect samples. Table 3 compares mIoU and FB-IoU performance on the FSSD-12 dataset. Our EAMNet significantly outperforms recent methods in both 1-shot and 5-shot settings using either VGG-16 or ResNet-50 backbones. With ResNet-50, EAMNet achieves mIoU improvements of 1.92% over MAPTNet[46] (1-shot) and 0.73% over HDMNet[35] (5-shot) by leveraging episode adaptation to exploit latent data correlations for enhanced training and prediction. In addition, EAMNet gains significant improvement over the baseline models. For example, EAMNet with VGG-16 backbone achieves 15.93% and 15.02% mIoU improvement over the baseline model, which proves the superiority of our model in such challenging scenarios.

Qualitative Results

We report some qualitative results generated from several models for metal surface defect detection and our EAMNet on the Surface Defect-4^i benchmarks. Compared with these representative models, EAMNet exhibits the following advantages as shown in Figure 3. (1) EAMNet can more accurately segment the target class, while the previous methods incorrectly segments the seen classes as the target classes (1st to 3rd columns). (2) EAMNet can capture subtle defect details for semantic description of defect regions to address the limited adaptability problem caused by single-episode (4th to 6th columns). (3) EAMNet can provide better fine-grained resolution through the contextual adaptive factor from CAM and the support guidance from GRMAP (7th to 8th columns). (4) Some failure case are exhibited from 9th to 10th columns. These challenges stems from inherent problem in the data, where the model underperforms on subtle defects with low contrast against the background texture. Nevertheless, comparative results show that EAMNet, via its cross-episode adaptive analysis, can capture potential defect cues to coarsely segment the main defect area, outperforming other popular networks in such challenging scenarios.

Table 3. Comparison of EAMNet with state-of-the-art methods and semantic networks for metal surface defect FSS on FSSD-12, evaluated by mIoU and FB-IoU under 1-shot and 5-shot settings using the VGG-16 and ResNet-50 backbone. The **best** and second best results are highlighted accordingly. Improv. (%) represents the percentage improvement in mIoU over the *Baseline.*

Methods	Backbone	1-shot						5-shot					
		Fold-0	Fold-1	Fold-2	Mean	FBIoU	Improv. (%)	Fold-0	Fold-1	Fold-2	Mean	FBIoU	Improv. (%)
HDMNet[35]	VGG16	50.12	49.05	45.51	48.23	66.57	↑6.10	48.26	49.60	46.11	47.99	66.69	↑3.10
DCPNet[36]		53.30	44.52	40.98	46.27	66.01	↑4.14	50.07	48.69	43.07	47.28	67.28	↑2.39
TGRNet(0-normal)[37]		63.74	51.68	49.95	55.12	73.36	↑13.00	**66.16**	60.24	51.07	59.16	74.48	↑14.27
CPANet[45]		50.90	47.39	53.38	50.56	69.26	↑8.43	50.15	37.41	43.39	43.65	64.73	↓1.24
MAPTNet[46]		**65.87**	55.59	51.49	57.65	72.90	↑15.52	62.84	58.93	47.41	56.39	74.88	↑11.50
PFENet[47]		43.65	37.89	36.12	39.22	67.72	↓2.91	44.86	40.66	36.57	40.70	68.92	↓4.19
Baseline		48.40	*38.04*	*39.94*	*42.13*	*63.94*	-	*48.85*	*43.98*	*41.84*	*44.89*	*65.24*	-
EAMNet		63.75	**56.00**	**54.43**	**58.06**	**74.46**	↑**15.93**	65.73	**61.87**	**52.14**	**59.91**	**75.68**	↑**15.02**
HDMNet[35]	ResNet50	60.50	**65.46**	51.42	59.13	74.33	↑12.33	63.03	**68.22**	53.74	61.66	77.33	↑11.78
DCPNet[36]		59.65	63.13	51.84	58.21	74.79	↑11.41	61.68	61.80	52.71	58.73	74.10	↑8.85
TGRNet(0-normal)[37]		61.09	63.24	51.28	58.54	75.20	↑11.74	61.59	65.81	56.27	61.22	76.74	↑11.34
CPANet[45]		54.40	52.59	48.39	51.79	65.15	↑4.99	56.73	55.06	51.92	54.57	71.19	↑4.69
MAPTNet[46]		68.20	58.24	**55.34**	60.59	75.64	↑13.79	63.57	62.52	52.85	59.65	73.47	↑9.76
PFENet[47]		49.00	47.87	41.78	46.22	73.76	↓0.58	50.11	50.98	42.34	47.81	74.77	↓2.07
HMNet[50]		62.15	54.96	50.56	55.89	71.10	↑9.09	60.21	60.02	50.53	56.92	72.12	↑7.04
Baseline		*53.95*	*43.53*	*42.92*	*46.80*	*64.66*	-	*55.84*	*50.24*	*43.57*	*49.88*	*67.74*	-
EAMNet		**68.73**	64.11	54.71	**62.52**	**75.80**	↑**15.72**	**64.35**	65.92	**56.91**	**62.39**	**78.24**	↑**12.51**

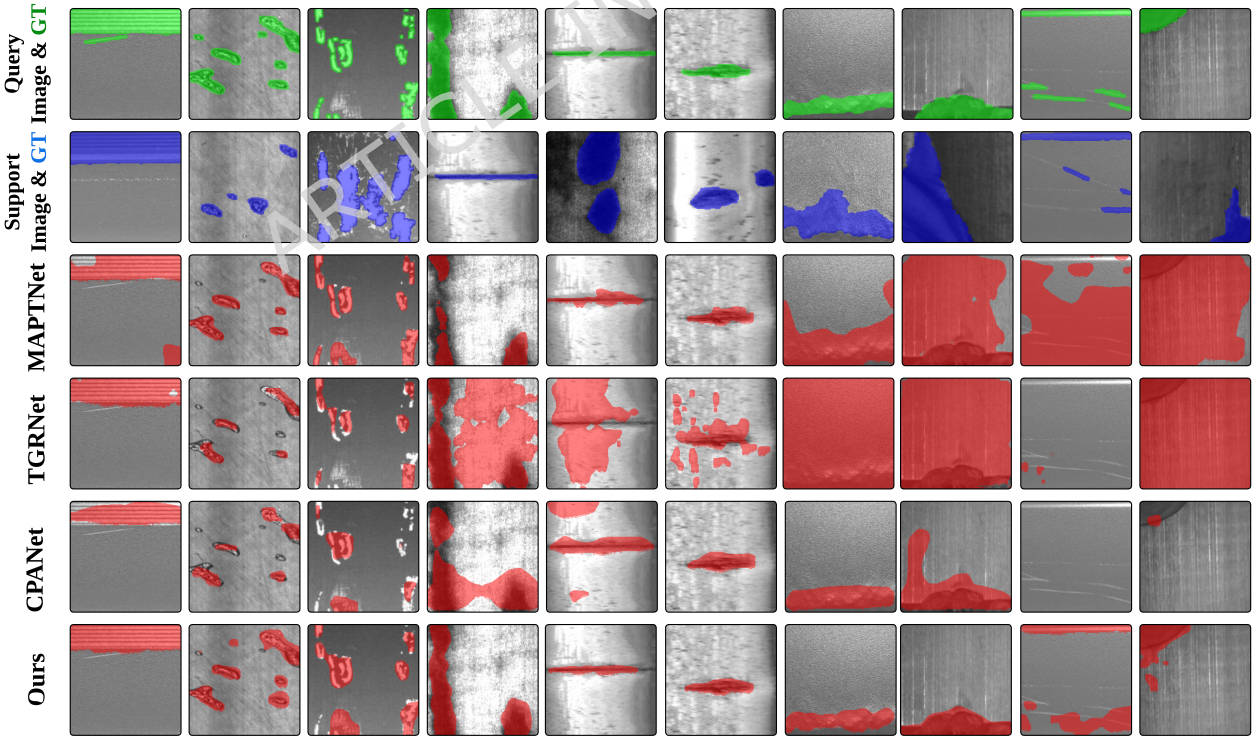

Figure 3. Qualitative results of the MAPTNet, TGRNet, CPANet and proposed EAMNet on Surface Defect-4^{i}.

Ablation Study

We perform ablation studies using the VGG-16 backbone in the 1-shot setting on the Surface Defect-4^i dataset.

Components Analysis

EAMNet comprises four key components: the Episode Adaptive Memory Unit (EAMU) for cross-episode interaction, the Context Adaptation Module (CAM), Global Response Mask Average Pooling (GRMAP), and Attention Distillation (AD). Table 4 summarizes the effectiveness of each component. As the most critical component for facilitating cross-episode interaction, EAMU contributes a significant 5.69% improvement in mIoU. Meanwhile, CAM, GRMAP, and AD are also essential. Collectively, these four modules enable EAMNet to achieve state-of-the-art performance.

Table 4. Ablation studies of each component on Surface Defect-4^i

Components				1-shot				Δ
ρ	ψ	$\boldsymbol{P}_{\text{supp}}$	$\mathscr{L}_{\text{AD}}$	Fold-0	Fold-1	Fold2	Mean	
✓				34.88	30.24	21.19	28.77	↑5.69
✓	✓			35.64	31.57	22.16	29.79	↑6.71
✓	✓	✓		45.77	31.24	24.78	33.93	↑10.85
✓	✓	✓	✓	**47.01**	**32.91**	**25.46**	**35.13**	↑**12.05**
Baseline				28.33	24.54	16.38	23.08	-

β in CAM

To investigate the performance of our contextual adaptive factor within the CAM, we evaluated β values from $\{0.1, 0.2, 0.3, 0.4, 0.5\}$. As shown in Table 5, EAMNet achieves its peak performance with $\beta = 0.4$, and the second-best result with $\beta = 0.5$. Consequently, we set β to 0.4 for all subsequent experiments.

Table 5. 1-shot mIoU of ablation study for hyperparameter β on Surface Defect-4^i

β	1-shot				Δ
	Fold-0	Fold-1	Fold-2	Mean	
0.10	45.30	30.62	24.26	33.40	↑10.31
0.20	45.48	30.92	24.20	33.53	↑10.45
0.30	46.44	31.47	24.40	34.10	↑11.02
0.40	**47.01**	**32.91**	25.46	**35.13**	↑**12.05**
0.50	47.00	32.13	**25.57**	34.90	↑11.82
Baseline	28.33	24.54	16.38	23.08	-

α in loss function

As shown in Table 6, we compare the various hyperparameter α which lies in our loss function to validate the effectiveness of AD. When the $\alpha = 0.05$, we can see this strategy achieves 12.05% mIoU improvement and outperforms other settings.

Table 6. 1-shot mIoU of ablation study for hyperparameter α on Surface Defect-4^i

α	1-shot				Δ
	Fold-0	Fold-1	Fold-2	Mean	
0.01	**47.48**	30.89	25.45	34.61	↑11.53
0.03	47.46	31.51	25.25	34.74	↑11.66
0.05	47.01	32.91	**25.46**	**35.13**	↑**12.05**
0.07	46.81	**33.36**	24.91	35.03	↑11.94
0.09	46.87	32.08	24.85	34.60	↑11.52
Baseline	28.33	24.54	16.38	23.08	-

Conclusions

We propose an episode adaptive memory network (EAMNet) with four major parts (i.e., EAMU, CAM, GRMAP and AD) to the few-shot semantic segmentation for metal surface defect. The EAMU generates an adaptive factor at the cross-episode diagram while alleviating the limited adaptability in semantic description of defect regions within the conventional single-episode training stage. The CAM and GRMAP obtains fine-grained resolution from the contextual adaptive factor and support guidance, as a supplement to the semantic information in a single episode. An attention distillation is designed to store the memory of semantic defect cues between episodes and boost adaptive performance by leveraging semantic attention correspondence. Comprehensive experiments show that EAMNet achieves state-of-the-art performance under all settings.

Data Availability Statement

All data and codes underlying the results of this study are available at the following URL: https://doi.org/10.5281/zenodo.18174740.

Author Contributions

J.Z. and H.D. wrote the main manuscript text, J.Z. and M.P. prepared figures 1-2, H.D., M.P. and S.T. designed the experiments, G.C. and Y.L. provided critical revisions. All authors reviewed the manuscript.

Funding

This work was supported in part by the Fujian Natural Science Foundation under Grants 2023J01978, 2023J01979 and 2024J01855, in part by the Fujian International Cooperation Program in Science and Technology under Grant 2024I0024, in part by the Fujian Regional Development Project under Grant 2025Y3009, and in part by the Doctoral Research Project of Longyan University under Grants LB2023008 and LB2023015.

References

1. Liu, Y., Liu, Y., Guo, X., Ling, X. & Geng, Q. Metal surface defect detection using slf-yolo enhanced yolov8 model. *Sci. Reports* **15**, 11105 (2025).
2. Chan, S. *et al.* Feature optimization-guided high-precision and real-time metal surface defect detection network. *Sci. Reports* **14**, 31941 (2024).
3. Zhou, C. *et al.* Metal surface defect detection based on improved yolov5. *Sci. Reports* **13**, 20803 (2023).
4. Lv, X., Duan, F., Jiang, J.-j., Fu, X. & Gan, L. Deep metallic surface defect detection: The new benchmark and detection network. *Sensors* **20**, 1562 (2020).
5. Wang, W., Han, C., Zhou, T. & Liu, D. Visual recognition with deep nearest centroids. In *the Eleventh International Conference on Learning Representations*, 1–30 (2023).
6. Han, C. *et al.* E^2vpt: An effective and efficient approach for visual prompt tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 17491–17502 (2023).
7. Lu, Y. *et al.* Transflow: Transformer as flow learner. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18063–18073 (2023).
8. Liu, Q., Liu, M., Jonathan, Q. & Shen, W. A real-time anchor-free defect detector with global and local feature enhancement for surface defect detection. *Expert. Syst. with Appl.* **246**, 123199 (2024).
9. Tabernik, D., Šela, S., Skvarč, J. & Skočaj, D. Segmentation-based deep-learning approach for surface-defect detection. *J. Intell. Manuf.* **31**, 759–776 (2020).
10. Guo, B., Wang, Y., Zhen, S., Yu, R. & Su, Z. Speed: Semantic prior and extremely efficient dilated convolution network for real-time metal surface defects detection. *IEEE Transactions on Ind. Informatics* **19**, 11380–11390 (2023).
11. Ma, J., Xie, G.-S., Zhao, F. & Li, Z. Afanet: Adaptive frequency-aware network for weakly-supervised few-shot semantic segmentation. *IEEE Transactions on Multimed.* 1–11 (2025).
12. Wang, C. *et al.* Taylor series-inspired local structure fitting network for few-shot point cloud semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 7527–7535 (2025).

13. Mai, H., Sun, R., Zhang, T. & Wu, F. Rankmatch: Exploring the better consistency regularization for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3391–3401 (2024).

14. Woo, S. *et al.* Convnext v2: Co-designing and scaling convnets with masked autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16133–16142 (2023).

15. Chao, C. *et al.* Iamf-yolo: Metal surface defect detection based on improved yolov8. *IEEE Transactions on Instrumentation Meas.* **74**, 1–17 (2025).

16. Zhang, X., Fang, T., Saniie, J., Bakhtiari, S. & Heifetz, A. Unsupervised learning-enabled pulsed infrared thermographic microscopy of subsurface defects in stainless steel. *Sci. Reports* **14**, 14865 (2024).

17. Panić, B., Borovinšek, M., Vesenjak, M., Oman, S. & Nagode, M. A guide to unsupervised image segmentation of mct-scanned cellular metals with mixture modelling and markov random fields. *Mater. & Des.* **239**, 112750 (2024).

18. Song, Z., Yao, H., Tian, D., Zhan, G. & Gu, Y. Segmentation method of u-net sheet metal engineering drawing based on cbam attention mechanism. *Artif. Intell. for Eng. Des. Analysis Manuf.* **39**, e14 (2025).

19. Zhang, H. *et al.* An efficient model for metal surface defect detection based on attention mechanism and multi-scale feature. *The J. Supercomput.* **81**, 40 (2025).

20. Zhang, L., Li, X., Sun, Y. & Guo, H. Triple-attentions based salient object detector for strip steel surface defects. *Sci. Reports* **15**, 2537 (2025).

21. Wei, H., Zhao, L., Li, R. & Zhang, M. Rfaconv-cbm-vit: enhanced vision transformer for metal surface defect detection. *The J. Supercomput.* **81**, 1–38 (2025).

22. Zhang, G. *et al.* Lggformer: A dual-branch local-guided global self-attention network for surface defect segmentation. *Adv. Eng. Informatics* **64**, 103099 (2025).

23. Jin, Y., Zhang, Y., Shan, D. & Wu, Z. Human-guided zero-shot surface defect semantic segmentation. *IEEE Transactions on Instrumentation Meas.* **74**, 1–13 (2025).

24. Zhao, L., Zhang, Y., Duan, J. & Yu, J. Cross-supervised contrastive learning domain adaptation network for steel defect segmentation. *Adv. Eng. Informatics* **64**, 102964 (2025).

25. Snell, J., Swersky, K. & Zemel, R. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, vol. 30, 4077–4087 (2017).

26. Ye, M. & Zhang, T. Sanet: similarity aggregation and semantic fusion for few-shot semantic segmentation. *Appl. Intell.* **55**, 1–12 (2025).

27. Gong, W. *et al.* Cgnet: Few-shot learning for intracranial hemorrhage segmentation. *Comput. Med. Imaging Graph.* 102505 (2025).

28. Yang, A., Sang, Z., Zhou, Y., Cao, J. & Liu, L. Bi-orientated rectification few-shot segmentation network based on fine-grained prototypes. *Neurocomputing* **620**, 129160 (2025).

29. Chen, S., Yu, Y., Li, Y., Lu, Z. & Zhou, Y. Mask-free iterative refinement network for weakly-supervised few-shot semantic segmentation. *Neurocomputing* **611**, 128600 (2025).

30. Wang, X., Chen, Q. & Yang, Y. Word vector embedding and self-supplementing network for generalized few-shot semantic segmentation. *Neurocomputing* **613**, 128737 (2025).

31. Xiang, Q. *et al.* Dkdm: Data-free knowledge distillation for diffusion models with any architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2955–2965 (2025).

32. Cai, S. *et al.* Diffusion self-distillation for zero-shot customized image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18434–18443 (2025).

33. Li, Q., Wang, Y., Meng, L., Qin, Y. & Tang, B. Sdmc-net: A lightweight and deployable fault diagnosis network using self-distillation and multiscale-depth convolution. *IEEE Transactions on Instrumentation Meas.* **74**, 1–13 (2025).

34. Lu, Z. *et al.* Self-distillation attention for efficient and accurate motion prediction in autonomous driving. *IEEE Transactions on Intell. Transp. Syst.* **26**, 7262–7274 (2025).

35. Peng, B. *et al.* Hierarchical dense correlation distillation for few-shot segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23641–23651 (2023).

36. Lang, C., Tu, B., Cheng, G. & Han, J. Beyond the prototype: divide-and-conquer proxies for few-shot segmentation. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, 1024–1030 (2022).

37. Bao, Y. *et al.* Triplet-graph reasoning network for few-shot metal generic surface defect segmentation. *IEEE Transactions on Instrumentation Meas.* **70**, 1–11 (2021).

38. Zhang, X., Wei, Y., Yang, Y. & Huang, T. S. Sg-one: Similarity guidance network for one-shot semantic segmentation. *IEEE Transactions on Cybern.* **50**, 3855–3865 (2020).

39. Song, K., Feng, H., Cao, T., Cui, W. & Yan, Y. Mfanet: multifeature aggregation network for cross-granularity few-shot seamless steel tubes surface defect segmentation. *IEEE Transactions on Ind. Informatics* **20**, 9725–9735 (2024).

40. Yu, R., Guo, B. & Yang, K. Selective prototype network for few-shot metal surface defect segmentation. *IEEE Transactions on Instrumentation Meas.* **71**, 1–10 (2022).

41. Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural Comput.* **9**, 1735–1780 (1997).

42. Bulatov, A., Kuratov, Y. & Burtsev, M. Recurrent memory transformer. In *Advances in Neural Information Processing Systems*, vol. 35, 11079–11091 (2022).

43. Lin, T.-Y. *et al.* Feature pyramid networks for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2117–2125 (2017).

44. Roy, A. G., Navab, N. & Wachinger, C. Concurrent spatial and channel 'squeeze & excitation' in fully convolutional networks. In *Medical Image Computing and Computer Assisted Intervention*, 421–429 (2018).

45. Feng, H., Song, K., Cui, W., Zhang, Y. & Yan, Y. Cross position aggregation network for few-shot strip steel surface defect segmentation. *IEEE Transactions on Instrumentation Meas.* **72**, 1–10 (2023).

46. Huang, J. *et al.* Multiscale adaptive prototype transformer network for few-shot strip steel surface defect segmentation. *IEEE Transactions on Instrumentation Meas.* **74**, 1–14 (2025).

47. Tian, Z. *et al.* Prior guided feature enrichment network for few-shot segmentation. *IEEE Transactions on Pattern Analysis Mach. Intell.* **44**, 1050–1065 (2022).

48. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 770–778 (2016).

49. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

50. Xu, Q. *et al.* Hybrid mamba for few-shot segmentation. In *Advances in Neural Information Processing Systems*, vol. 37, 73858–73883 (2024).