Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Artifact-Aware Lightweight Dermoscopic Lesion Segmentation with Reproducible Corruption Robustness Evaluation

Haixiao Gong
Guangxi Key Laboratory of Machine Vision and Intelligent Contro
Cuijiao Yin
yincuijiao@yeah.net
Wuzhou Vocational College
Jie He
Guangxi Key Laboratory of Machine Vision and Intelligent Contro
Hui Guo
Guangxi Key Laboratory of Machine Vision and Intelligent Contro
Mei Tang
Wuzhou Vocational College
Linting Meng
Guangxi Key Laboratory of Machine Vision and Intelligent Contro
Chengchen Gong
Qinghai Minzu University

Article

Keywords:

Posted Date: July 15th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10197577/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Scientific Reports on November 14th, 2026. See the published version at https://doi.org/10.1038/s41598-026-74096-0.

Artifact-Aware Lightweight Dermoscopic Lesion Segmentation with Reproducible Corruption Robustness Evaluation

Haixiao Gong[1,2], Cuijiao Yin[3,*], Jie He[1,2], Hui Guo[1], Mei Tang[3], Linting Meng[1,2], and Chengchen Gong[4]

[1]Guangxi Key Laboratory of Machine Vision and Intelligent Control, Wuzhou, 543002, China
[2]School of Electronic Information and Artificial Intelligence, Wuzhou University, Wuzhou, 543000, China
[3]Wuzhou Vocational College, Wuzhou, 543002, China
[4]School of Tourism, Qinghai Minzu University, Xining, 810007, China
[*]Corresponding author. Email: yincuijiao@yeah.net (Cuijiao Yin)

ABSTRACT

Automated dermoscopic lesion segmentation is a key step in computer-aided skin cancer analysis, yet lightweight models for mobile or point-of-care deployment can be unstable when images contain acquisition artifacts such as hair, specular highlights, skin lines, and air bubbles. This paper addresses this problem by proposing an artifact-aware lightweight U-shaped segmentation framework together with a reproducible corruption benchmark for robustness evaluation. The proposed framework first estimates artifact-prone regions through a multi-cue stem that combines Laplacian, bright-region, dark-region, and learned responses. It then uses efficient multi-scale encoder blocks with depthwise separable operations and channel gating to preserve lesion-relevant features under a limited parameter budget. During decoding, a mask-guided skip-fusion module integrates coarse lesion predictions with propagated artifact cues, reducing the direct transfer of contaminated shallow features while retaining boundary information. To evaluate robustness beyond clean ISIC validation images, we construct fixed corrupted subsets from ISIC 2017 and ISIC 2018 by applying four clinically motivated artifact categories at two severity levels, with unchanged ground-truth masks and recorded artifact metadata. Experiments compare clean-only training, direct artifact augmentation, and curriculum-based artifact exposure against representative lightweight baselines. The results show that the proposed framework maintains competitive clean-domain accuracy, improves corrupted-domain robustness, benefits from curriculum training under stronger artifact severity, and transfers better to PH2 without fine-tuning. Ablation studies further verify the complementary effects of artifact-aware input modeling, robust encoder design, and mask-guided skip fusion. Overall, this work provides a lightweight segmentation framework and a reproducible evaluation protocol for studying artifact robustness in dermoscopic image segmentation. Our code is available at: https://github.com/28001736/ARL-UNet.

1 Introduction

Skin cancer is among the most common malignancies worldwide, and early detection is crucial for improving prognosis[1]. Dermoscopy reveals sub-surface lesion structures[2], and accurate lesion segmentation supports classification, border analysis, and longitudinal monitoring[3]. With growing interest in mobile and point-of-care systems, segmentation models must also be computationally lightweight[4].

Recent dermoscopic segmentation studies have therefore focused on lightweight U-shaped networks[5,6], mainly evaluated on ISIC benchmarks[7,8]. Progress has been made through compact architectures[9,10], efficient attention modules[11,12], and refined skip connections[13–16]. However, most evaluations still rely on clean validation images, so clean-domain performance does not necessarily reflect reliability under acquisition artifacts.

This gap directly affects visual feature extraction and boundary localization. Dermoscopic images often contain terminal hair, skin lines, specular highlights, and air bubbles[2,17]. Hair and skin lines introduce thin high-frequency structures that can be confused with lesion contours, while highlights and air bubbles cause local saturation, contrast shifts, and partial occlusion. These degradations are particularly challenging for lightweight models, whose limited capacity makes it harder to separate lesion-relevant patterns from artifact-induced responses. In U-shaped architectures, skip connections may further pass contaminated shallow features into the decoder, leading to boundary distortion or false foreground activation.

Existing methods only partially address this issue. Classical artifact removal pipelines[18–20] are detached from segmentation and may remove diagnostically useful structures. Modern lightweight segmentation networks[11,13,21,22] mainly optimize efficiency and clean-image accuracy, but rarely model artifact-prone regions explicitly. In addition, without a fixed corrupted-

domain benchmark[23], it remains difficult to compare robustness under clinically plausible image degradations.

To address these issues, we propose **ARL-UNet**, an artifact-aware lightweight U-shaped segmentation framework. Its design follows a reusable principle: estimate artifact cues before downsampling, propagate them across scales, and use them conservatively to modulate skip features rather than directly removing image regions. ARL-UNet instantiates this principle with an *Artifact-Aware Stem* for multi-cue artifact estimation, a *Robust Encoder Block* using depthwise separable multi-scale convolution and channel gating[4,11], and a *Mask-guided Robust Skip Fusion* module that combines coarse lesion masks with artifact cues during decoding.

Beyond the network design, we construct a reproducible corrupted dermoscopic benchmark from fixed ISIC validation splits[7,8]. Inspired by ImageNet-C[23], the benchmark applies four dermoscopy-relevant artifact categories at two severity levels while keeping ground-truth masks unchanged. The artifact-generation protocol, severity definitions, artifact masks, and metadata support controlled comparison of artifact-specific degradation.

We also analyze artifact-aware training strategies, including clean-only training, direct artifact augmentation, and curriculum-based artifact exposure[24]. Experiments compare representative lightweight baselines under clean, corrupted, and cross-dataset settings, and report changes in mIoU, HD95, artifact-specific degradation, and PH2 zero-shot transfer.

The main contributions of this work are as follows:

- We propose an **artifact-aware lightweight segmentation framework** that combines early artifact cue estimation, efficient robust encoding, and mask-guided skip fusion. The skip-fusion principle provides a reusable way to reduce the propagation of artifact-contaminated shallow features in U-shaped networks.
- We construct a **reproducible corrupted dermoscopic benchmark** with four clinically relevant artifact categories and two severity levels. The benchmark defines a reusable artifact-generation protocol for evaluating robustness beyond conventional clean ISIC testing.
- We provide a systematic study of **artifact-aware training strategies**, including clean-only training, direct augmentation, and curriculum learning, and evaluate their effects on clean-domain accuracy, corrupted-domain robustness, artifact-specific degradation, and PH2 zero-shot transfer.

The remainder of this paper is organized as follows. Section 2 reviews lightweight medical image segmentation, dermoscopic artifact handling, and robustness evaluation. Section 3 presents the corrupted dermoscopic benchmark and Section 4 introduces the proposed ARL-UNet architecture. Section 5 describes the experimental setup and quantitative comparisons. Section 7 concludes the paper.

2 Related Work

2.1 Medical Image Segmentation

U-Net[5] and its encoder-decoder variants have dominated medical image segmentation, yet their computational cost limits deployment in resource-constrained settings. A primary direction for lightweight design is depthwise separable convolution[4], which reduces parameters and computation without proportional loss in representational capacity; representative works include DSNet[25] and LDMRes-Net[26]. Attention mechanisms such as squeeze-and-excitation gating[11], spatial attention, and inception-style multi-scale convolution have further been integrated to compensate for reduced feature capacity, as exemplified by IS-UNeXt[27] and SLP-Net[28]. Transformer-based models such as Swin-UNet[29] and TransUNet[30] capture long-range dependencies effectively, but their quadratic complexity renders them ill-suited for low-resource deployment. State-space models such as Mamba[31], recently extended to biomedical segmentation[32], offer linear-time sequence modeling as an alternative, though their efficiency advantage in lightweight regimes remains an open question.

2.2 Dermoscopic Artifact Handling

Dermoscopic images routinely contain acquisition-induced artifacts, including terminal hair, skin surface lines, specular highlights, and air bubbles, that occlude lesion boundaries and confuse foreground-background discrimination[33]. Classical preprocessing pipelines address this through morphological operations and inpainting[18], while deep learning approaches such as VAE-based[34] and U-Net-based[35] methods have substantially improved removal quality. However, all such pipelines remain decoupled from the segmentation objective and risk removing diagnostically relevant structures alongside artifacts.

Fewer studies embed artifact robustness directly within the segmentation network. LMSAUnet[36] proposed a lightweight architecture with attention to partially mitigate artifact sensitivity, but lacks explicit artifact detection and a controlled benchmark for reproducible robustness evaluation.

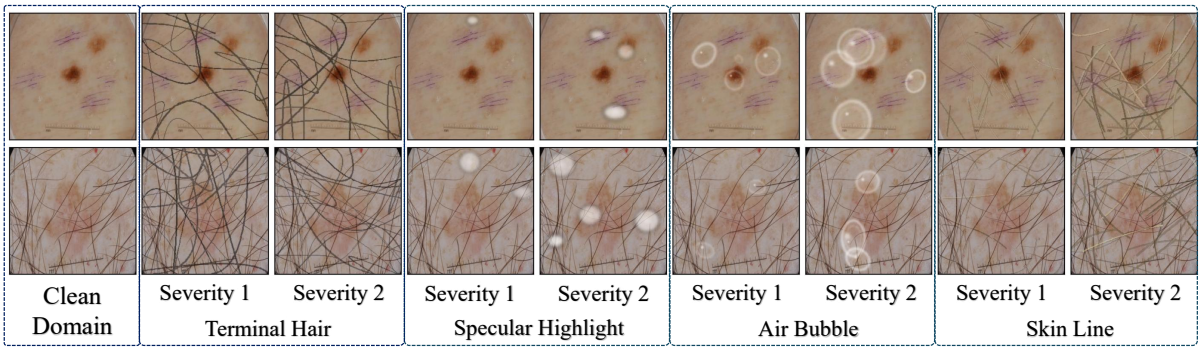


Figure 1. Examples of the corrupted dermoscopic benchmark. Each row shows one clean dermoscopic image and its corrupted counterparts.

2.3 Robust and Efficient Segmentation

Systematic robustness evaluation has gained attention since ImageNet-C[23], which measures model sensitivity to common corruptions across multiple severity levels. This idea has been extended to semantic segmentation[37], revealing that high clean-domain accuracy does not guarantee robustness, and motivating corruption-robust architectures and augmentation strategies[38].

Recent visual computing studies also improve robustness through efficient feature modeling and multi-scale fusion. DCE-Unet[39] combines a lightweight segmentation core with dual multi-scale fusion for real-time sea-sky line detection, while MDF-Polyp[40] integrates spectral-domain enhancement, deformable spatial attention, and Mamba-based state-space modeling for boundary-aware polyp segmentation. In medical image analysis, robustness studies mainly address cross-scanner or cross-site domain shift. Curriculum-based Fourier domain adaptation[41] is effective for scanner-induced shifts, but does not target dermoscopy-specific artifacts such as hair, highlights, air bubbles, and skin lines.

3 Corrupted Dermoscopic Benchmark

3.1 Design Principles

Clean-domain evaluation alone is insufficient to assess whether a segmentation model is suitable for real clinical deployment[42]. Dermoscopic images acquired in routine practice frequently contain acquisition artifacts, which may alter local contrast, introduce spurious high-frequency boundaries, or partially occlude lesion structures[43]. To measure robustness under such conditions, we construct a controlled corrupted dermoscopic benchmark governed by four principles.

- **Reproducible.** Corrupted images are generated offline from fixed validation splits using deterministic random seeds. Each image–artifact–severity combination has a stable derived seed, so the same corrupted sample can be regenerated independently of file ordering or processing order.
- **Paired.** For each clean image, corrupted counterparts are generated while the original lesion mask remains unchanged, enabling direct measurement of performance degradation caused by each artifact type and severity level.
- **Artifact-specific.** Each corrupted subset contains one dominant artifact type rather than mixed corruptions, allowing artifact-specific analysis of model failure modes.
- **Clinically motivated.** The selected artifact types correspond to common dermoscopic acquisition disturbances, rather than generic corruptions such as Gaussian noise or blur.

Formally, given a clean image–mask pair $(\mathbf{x}i, \mathbf{y}i)$, the corrupted image is generated as

$$\tilde{\mathbf{x}}i^{a,s} = \mathscr{C}a,s(\mathbf{x}i;,\theta i^{a,s}), \tag{1}$$

where a denotes the artifact type, $s \in 1,2$ the severity level, $\mathscr{C}a,s$ the artifact generation function, and $\theta i^{a,s}$ the fixed generation parameters. The ground-truth mask $\mathbf{y}i$ remains unchanged. An artifact region mask $\mathbf{m}i^{a,s}$ is also generated to indicate the altered pixels; it is not used as lesion supervision during standard evaluation, but supports benchmark analysis, artifact statistics, and artifact-aware training.

Table 1. Principal generation hyperparameters of the corrupted dermoscopic benchmark for the two severity levels.

Artifact	Parameter	Severity 1	Severity 2
Hair	# Strands	14–25	15–26
	Thickness	1–3	1–3
	Opacity	122–188	115–186
Highlight	# Spots	1–3	2–4
	Radius scale	0.035–0.085	0.045–0.115
	Opacity	76–136	96–168
Air bubble	# Bubbles	2–4	2–5
	Radius	9–28	11–38
	Fill opacity	0.14–0.30	0.18–0.37
	Rim opacity	0.27–0.52	0.34–0.64
Skin line	# Lines	18–29	32–47
	Length scale	0.22–0.55	0.30–0.72
	Thickness	1–2	1–3
	Opacity	76–128	96–160

3.2 Artifact Categories

The benchmark comprises four dermoscopy-relevant artifact categories: Terminal Hair, Specular Highlight, Skin Line, and Air Bubble. The four artifact categories are selected to cover different visual degradation mechanisms. Hair and skin lines mainly test robustness to thin high-frequency structures, highlights test local saturation and intensity distortion, and air bubbles test semi-transparent occlusion and boundary interference. This design allows the benchmark to evaluate not only average robustness, but also artifact-specific failure modes in feature extraction and boundary localization.

Terminal Hair artifacts appear as dark, thin, curvilinear structures crossing the skin or lesion region. Their high-frequency edges can be confused with lesion boundaries, particularly when strands overlap the lesion contour. Hair corruption is synthesized by drawing multiple dark curved strokes with randomized length, orientation, thickness, and opacity.

Specular Highlight artifacts are bright reflection regions caused by illumination, skin moisture, or dermoscopic contact media. They saturate local image intensity and obscure pigment structures. Highlight corruption is generated as smooth bright blobs with randomized location, size, and transparency, evaluating whether a model maintains stable foreground predictions under local saturation.

Air Bubble artifacts appear as semi-transparent circular or elliptical structures with bright rims and local contrast changes, partially covering lesion interiors or boundaries. Unlike Terminal Hair, Air Bubble creates broader low-contrast occlusions rather than thin dark lines. Bubble corruption is generated by placing translucent elliptical regions with rim-like intensity perturbations.

Skin Line artifacts are fine, low-contrast linear patterns arising from natural skin texture and acquisition geometry. Unlike Terminal Hair, they are lighter and weaker, but dense patterns can still interfere with boundary localization. Skin Line corruption is synthesized as shallow curved or elongated structures with randomized color, transparency, and spatial distribution.

3.3 Severity Levels

For each artifact category, two severity levels are defined: Severity=1 (mild-to-moderate) and Severity=2 (stronger but still clinically plausible). This two-level design distinguishes models that are stable only under weak perturbations from those that remain reliable under more challenging acquisition conditions. Unrealistically destructive corruptions are deliberately avoided; the goal is to approximate clinically meaningful artifacts rather than to construct adversarial examples.

Severity scaling is artifact-dependent. For Terminal Hair and Skin Line, higher severity increases the number, length, thickness, and opacity of linear structures. Specular Highlight controls the number, size, and brightness of reflective regions. For Air Bubble, it affects the number, radius, and opacity of bubble structures. The principal generation hyperparameters for each severity level are reported in Table 1.

3.4 Benchmark Implementation

The benchmark is implemented as an offline generation pipeline based on the ISIC 2017[7] and ISIC 2018[8,17] datasets. We use fixed 7:3 train–validation splits, and the validation split is used as the held-out test split for robustness evaluation. For each clean validation image, the pipeline applies every combination of artifact type and severity level, while keeping the original

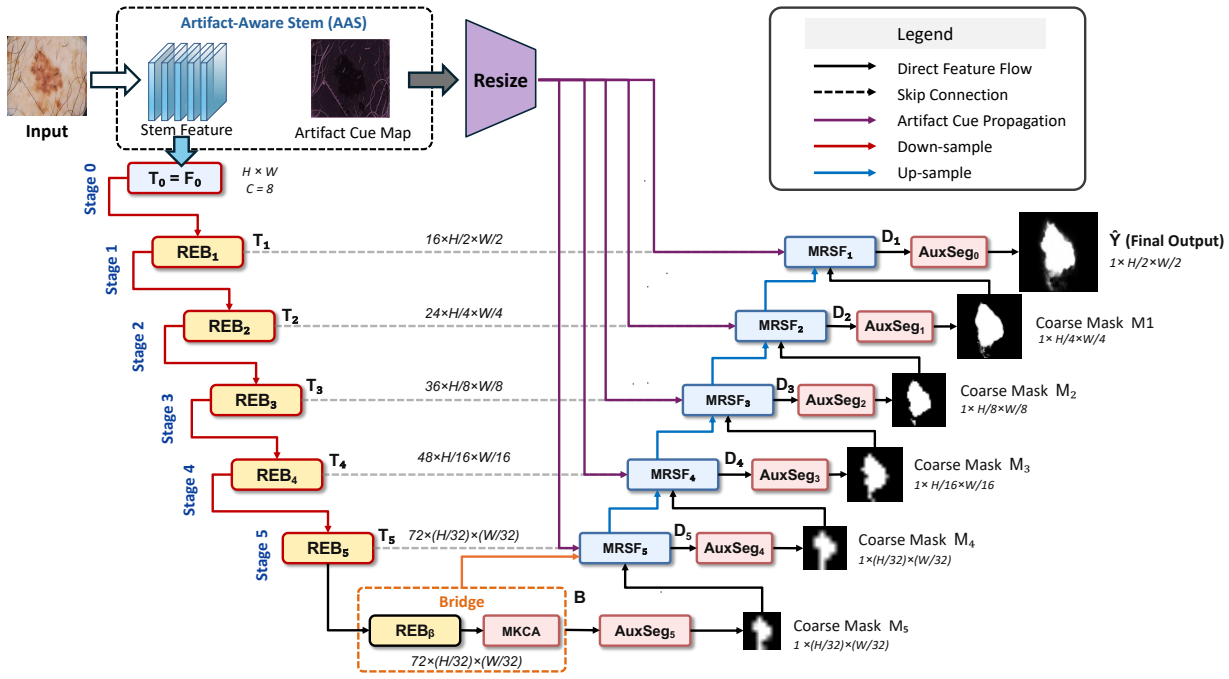


Figure 2. Overall architecture of the proposed ARL-UNet.

lesion mask unchanged. Clean images and masks are stored once and shared across all corrupted subsets to ensure mask consistency and avoid redundant storage. Representative examples are shown in Fig. 1, where each clean image is paired with four artifact categories at two severity levels.

The main generation hyperparameters are reported in Table 1. Integer ranges are sampled uniformly. Highlight radius and Skin Line length are defined relative to the shorter image side $\min(H, W)$, while Air Bubble radius is measured in pixels. For overlay-based artifacts, including Hair, Highlight, and Skin Line, opacity follows the 8-bit alpha range (0–255); for Air Bubble, fill and rim opacities are normalized blending weights in $[0, 1]$. The benchmark uses a fixed base seed of 2026. For each image–artifact–severity triple, a deterministic derived seed is computed and recorded, together with the sampled generation parameters.

For each corrupted image, the pipeline records the image identifier, source image path, corrupted image path, artifact type, severity level, derived seed, artifact area ratio, lesion-overlap ratio, and the sampled generation parameters. The artifact area ratio measures the affected image region, while the lesion-overlap ratio measures how much of the lesion is covered by artifacts. These metadata provide a transparent characterization of the benchmark and support analysis of robustness degradation.

To support reproducibility, we release the benchmark generation scripts, evaluation scripts, fixed split lists, and artifact-mask generation code in GitHub.

4 Methodology

This section presents the proposed ARL-UNet for artifact-robust lightweight dermoscopic lesion segmentation. The central design motivation is that common dermoscopic artifacts and lesion boundaries often share similar low-level visual properties, such as high-frequency edges, local contrast changes, and thin curvilinear structures. A lightweight network must therefore preserve lesion-sensitive spatial details while avoiding the direct propagation of artifact-contaminated shallow features. To address this problem, ARL-UNet introduces three components: an Artifact-Aware Stem (AAS) for input-level artifact cue extraction, a Robust Encoder Block (REB) for efficient multi-scale local representation, and a Mask-guided Robust Skip Fusion (MRSF) for conservative artifact-aware decoder fusion.

Table 2 clarifies the methodological position of ARL-UNet. Existing lightweight variants mainly improve efficiency, attention design, or multi-scale feature fusion, whereas ARL-UNet explicitly models artifact-prone regions and uses them to guide skip fusion. In our model, AAS and MRSF are the artifact-specific components, while REB provides efficient robust feature encoding.

Table 2. High-level comparison between standard U-Net, representative lightweight U-shaped baselines, and ARL-UNet.

Model	Main Design Focus	Lightweight	Artifact Cue	Artifact-Aware Fusion
U-Net[5]	Plain encoder–decoder with standard skip connections.	–	–	–
UNeXt-S[44]	Convolutional–MLP design for fast segmentation.	✓	–	–
MAL-UNet[16]	Multi-attention design for global/local feature modeling.	✓	–	–
EGE-UNet[45]	Group-enhanced lightweight attention.	✓	–	–
UCM-Net[9]	Hybrid CNN–MLP design for highly compact segmentation.	✓	–	–
EMA-UNet[46]	Multi-attention feature enhancement.	✓	–	–
DCE-UNet[39]	Lightweight segmentation with dual-channel attention, edge-aware downsampling, and multi-scale fusion.	✓	–	–
ARL-UNet (Ours)	Artifact-aware lightweight design with explicit artifact modeling and robust skip fusion.	✓	✓	✓

4.1 Overall Architecture

ARL-UNet follows a compact U-shaped encoder–decoder architecture with deep supervision. The architecture is designed according to the degradation pathway of dermoscopic artifacts. Since artifacts first appear as low-level visual disturbances, AAS estimates artifact-prone regions before downsampling. Since lightweight encoders may have limited context modeling ability, REB improves local multi-scale representation with efficient depthwise operations. MRSF filters skip information using both coarse semantic guidance and propagated artifact cues.

As shown in Fig. 2, given an input dermoscopic image $\mathbf{x} \in \mathbb{R}^{3\times H\times W}$, AAS first produces an initial stem feature together with an artifact cue map:

$$(\mathbf{F}_0, \mathbf{A}_0) = \mathrm{AAS}(\mathbf{x}), \tag{2}$$

where $\mathbf{F}_0$ is the stem feature and $\mathbf{A}_0 \in [0,1]^{1\times H\times W}$ denotes the estimated artifact-prone region. The encoder then progressively downsamples the feature map through five stages, each composed of a downsampling operator followed by a REB:

$$\mathbf{T}_i = \mathrm{REB}_i(\mathrm{Down}_i(\mathbf{T}_{i-1})), \quad i = 1,\ldots,5, \tag{3}$$

with $\mathbf{T}_0 = \mathbf{F}_0$. At the bottleneck, $\mathbf{T}_5$ is processed by a REB and a multi-kernel context aggregation module to yield the initial decoder feature:

$$\mathbf{D}_5 = \mathrm{Bridge}(\mathbf{T}_5). \tag{4}$$

The aggregation module applies five parallel depthwise branches with 3×3, 5×5, dilated 3×3, 1×7, and 7×1 kernels, and fuses their concatenation through a 1×1 projection.

During decoding, each encoder skip feature is fused with the corresponding decoder feature under coarse mask guidance and the propagated artifact cue (MRSF):

$$\mathbf{Y}_i = \mathrm{MRSF}_i(\mathbf{D}_i, \mathbf{T}_i, \mathbf{M}_i, \mathbf{A}_0), \quad i = 5,\ldots,1, \tag{5}$$

where $\mathbf{D}_5$ is the bottleneck feature and $\mathbf{D}^1_{j=4}$ is obtained by upsampling the fused output $\mathbf{Y}_{i+1}$ of the previous decoder stage; $\mathbf{M}_i$ is a coarse lesion mask predicted by an auxiliary head from the decoder feature entering stage i, and $\mathbf{A}_0$ is resized to the spatial resolution of the current scale. The final lesion probability map is produced by a prediction head Head, which applies a 1×1 convolution and bilinearly upsamples the result to the input resolution, followed by a sigmoid activation:

$$\hat{\mathbf{y}} = \sigma(\mathrm{Head}(\mathbf{Y}_1)). \tag{6}$$

In our implementation, the base channel configuration is $[8,16,24,36,48,72]$, which keeps the model within a lightweight parameter budget while leaving sufficient capacity for robust feature filtering.

4.2 Artifact-Aware Stem

As shown in Fig. 3 (a), the Artifact-Aware Stem (AAS) is designed to estimate artifact-prone regions before any spatial downsampling. This is important because hair strands, specular highlights, skin surface lines, and bubble rims are primarily

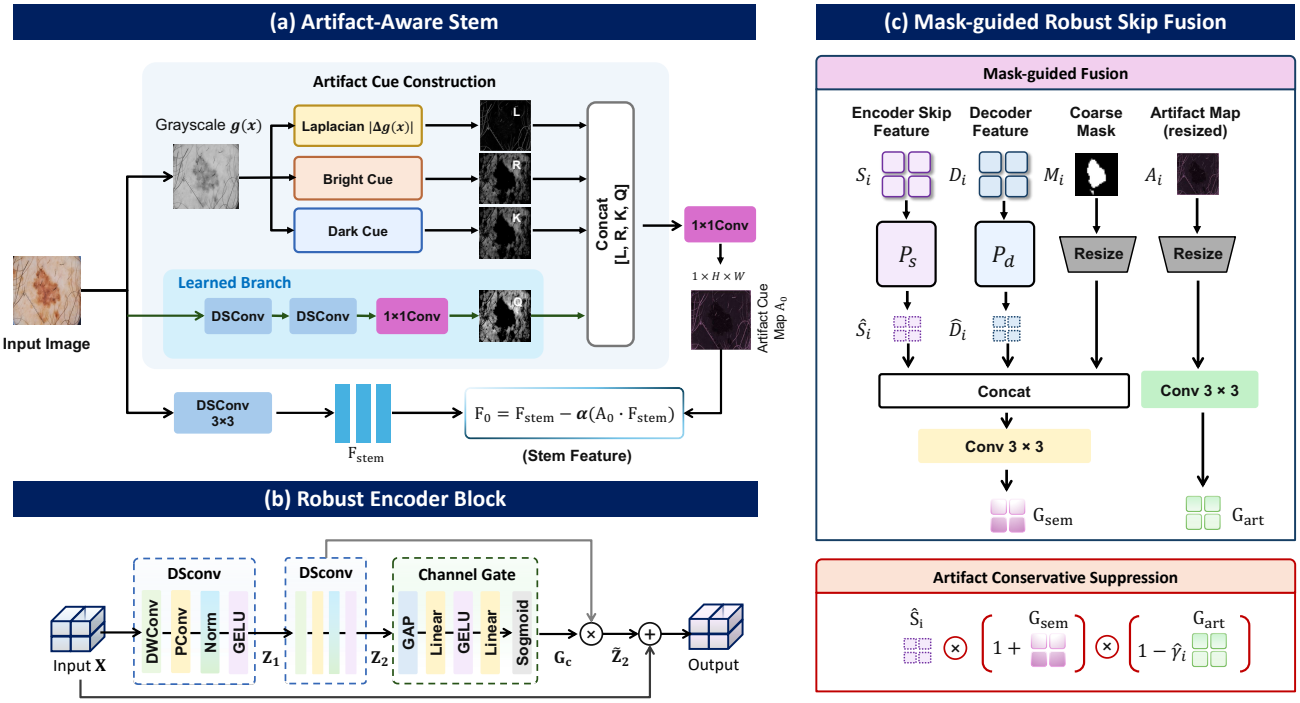


Figure 3. Detailed structures of the proposed ARL-UNet modules.

expressed as shallow local intensity patterns. If these disturbances are not identified at the input stage, they can be encoded as ordinary boundary features and propagated through skip connections.

Let $g(\mathbf{x})$ denote the grayscale input obtained by averaging the channels of $\mathbf{x}$ and rescaling the result to $[0, 1]$ per image. AAS constructs three hand-crafted artifact cues from $g(\mathbf{x})$. First, a Laplacian response captures high-frequency line and edge structures:

$$\mathbf{L} = |\Delta g(\mathbf{x})|. \tag{7}$$

Second, a bright-region cue $\mathbf{R}$ highlights saturated or reflective regions:

$$\mathbf{R} = \mathrm{ReLU}(g(\mathbf{x}) - \tau_h), \tag{8}$$

where $\tau_h = 0.75$ is the high-intensity threshold. Third, a dark-region cue $\mathbf{K}$ captures hair-like or border-like dark structures:

$$\mathbf{K} = \mathrm{ReLU}(\tau_l - g(\mathbf{x})), \tag{9}$$

where $\tau_l = 0.25$ is the low-intensity threshold. In addition to these hand-crafted cues, AAS includes a lightweight learned branch:

$$\mathbf{Q} = \sigma(f_\theta(\mathbf{x})), \tag{10}$$

where f_θ consists of two DSConv layers followed by a 1×1 projection. The learned branch complements the hand-crafted cues by capturing artifact appearances that simple brightness or edge operators cannot fully describe. The four cues are then concatenated and fused into a single artifact map:

$$\mathbf{A}_0 = \sigma\left(\mathrm{Conv}_{1\times1}\left([\mathbf{L}, \mathbf{R}, \mathbf{K}, \mathbf{Q}]\right)\right). \tag{11}$$

In parallel, the stem feature is extracted from the input by a depthwise separable convolution:

$$\mathbf{F}_{\mathrm{stem}} = \mathrm{DSConv}_{3\times3}(\mathbf{x}). \tag{12}$$

The artifact map is then used for residual soft suppression:

$$\mathbf{T}_0 = \mathbf{F}_0 = \mathbf{F}_{\text{stem}} - \alpha\,(\mathbf{A}_0 \otimes \mathbf{F}_{\text{stem}}), \tag{13}$$

where $\otimes$ denotes element-wise multiplication and α is a learnable scalar initialized to zero. This zero-initialized residual design makes AAS behave like a plain stem at the start of training, allowing the model to first learn stable lesion representations and then gradually activate artifact suppression only when beneficial.

4.3 Robust Encoder Block

The Robust Encoder Block (REB) improves local context modeling while keeping the network lightweight. Dermoscopic lesions contain both fine boundary details and larger-scale shape information; a single small convolution is efficient but may be insufficient to capture broader lesion context or to suppress elongated artifacts. REB therefore combines two depthwise separable convolutions with complementary receptive fields, followed by a channel gate. As shown in Fig. 3 (b), given an input feature $\mathbf{X} \in \mathbb{R}^{C_{\text{in}} \times h \times w}$, REB first applies a 3×3 depthwise separable convolution:

$$\mathbf{Z}_1 = \phi\,(\text{PConv}_1(\text{DWConv}_{3\times3}(\mathbf{X}))), \tag{14}$$

where $\text{DWConv}_{3\times3}$ denotes a 3×3 depthwise convolution, PConv_1 a pointwise (1×1) convolution, and $\phi(\cdot)$ a normalization followed by a GELU activation; together these three operations form one depthwise separable convolution. A second depthwise separable convolution then enlarges the effective receptive field:

$$\mathbf{Z}_2 = \phi\,(\text{PConv}_2(\text{DWConv}_{k\times k}(\mathbf{Z}_1))), \tag{15}$$

where $k = 5$ in the regular encoder stages. In the bottleneck stage, this second convolution instead uses a dilated 3×3 depthwise kernel with a dilation rate of 2, which enlarges the receptive field without adding additional kernel parameters.

To adaptively emphasize informative channels, REB applies a squeeze-style channel gate:

$$\mathbf{G}_c = \sigma\,(W_2\,\delta\,(W_1\,\text{GAP}(\mathbf{Z}_2))), \tag{16}$$

where GAP denotes global average pooling, δ is the GELU activation, σ is the sigmoid function, and W_1, W_2 are linear projection weights. The gated feature is then computed as

$$\tilde{\mathbf{Z}}_2 = \mathbf{G}_c \otimes \mathbf{Z}_2. \tag{17}$$

Finally, REB aggregates the result through a residual connection:

$$\text{REB}(\mathbf{X}) = P(\mathbf{X}) + \beta\tilde{\mathbf{Z}}_2, \tag{18}$$

where $P(\cdot)$ is an identity mapping when $C_{\text{in}} = C_{\text{out}}$ and a 1×1 projection for channel alignment otherwise, and β is a learnable residual scale initialized to one. Because the spatial operation is depthwise, a $k \times k$ depthwise separable convolution costs only about $k^2 C_{\text{in}} + C_{\text{in}} C_{\text{out}}$ parameters, substantially fewer than the $k^2 C_{\text{in}} C_{\text{out}}$ of a standard convolution. REB therefore provides multi-scale local modeling without sacrificing the lightweight nature of ARL-UNet.

4.4 Mask-guided Robust Skip Fusion

Skip connections are essential for recovering spatial details in U-shaped segmentation networks, but they also serve as a major pathway through which shallow artifact responses enter the decoder. In dermoscopic images, lesion boundaries and artifacts can both appear as high-frequency structures. Directly adding encoder and decoder features may therefore strengthen hair edges, highlight boundaries, skin-line patterns, or bubble rims. As shown in Fig. 3 (c), Mask-guided Robust Skip Fusion (MRSF) module addresses this issue by jointly using semantic guidance from a coarse lesion mask and artifact guidance from the AAS artifact map.

At decoder scale i, let $\mathbf{S}_i \equiv \mathbf{T}_i$ be the encoder skip feature, $\mathbf{D}_i$ be the decoder feature, $\mathbf{M}_i$ be the coarse lesion mask, and $\mathbf{A}_i$ be the artifact map resized from $\mathbf{A}_0$. The skip and decoder features are first projected to a common channel dimension by 1×1 convolutions:

$$\hat{\mathbf{S}}_i = P_s(\mathbf{S}_i), \qquad \hat{\mathbf{D}}_i = P_d(\mathbf{D}_i). \tag{19}$$

MRSF then computes a semantic gate from the projected skip and decoder features together with the coarse mask:

$$\mathbf{G}_{\text{sem}} = \sigma\left(\text{Conv}\left([\hat{\mathbf{S}}_i, \hat{\mathbf{D}}_i, \mathbf{M}_i]\right)\right). \tag{20}$$

This gate encourages the skip feature to emphasize regions consistent with the lesion hypothesis at the current decoder stage. In parallel, the artifact map $\mathbf{A}_i$ is converted into a channel-wise artifact gate:

$$\mathbf{G}_{\text{art}} = \sigma\left(\text{Conv}(\mathbf{A}_i)\right). \tag{21}$$

Instead of hard-removing artifact regions, ARL-UNet applies a conservative soft suppression:

$$\hat{\gamma}_i = \min(\max(\gamma_i, 0), \gamma_{\max}), \tag{22}$$

$$\mathbf{S}'_i = \hat{\mathbf{S}}_i \otimes (1 + \mathbf{G}_{\text{sem}}) \otimes (1 - \hat{\gamma}_i \mathbf{G}_{\text{art}}), \tag{23}$$

where γ_i is a learnable scalar initialized to zero and $\gamma_{\max} = 0.3$ is the maximum artifact-suppression strength. Here, the enhancement factor $1 + \mathbf{G}_{\text{sem}}$ amplifies skip responses in lesion-consistent regions, whereas the suppression factor $1 - \hat{\gamma}_i \mathbf{G}_{\text{art}}$ attenuates them in artifact-prone regions. This formulation is intentionally conservative: at initialization $\hat{\gamma}_i = 0$, so the module reduces to pure semantic skip enhancement without any artifact suppression. During training, the network learns whether and where artifact suppression should be activated, and the upper bound $\gamma_{\max}$ prevents excessive suppression, which is important because true lesion boundaries may also activate the artifact cue.

The fused output is refined by a Robust Encoder Block:

$$\mathbf{Y}_i = \text{REB}\left(\hat{\mathbf{D}}_i + \mathbf{S}'_i\right). \tag{24}$$

By combining coarse semantic guidance with artifact-aware conservative suppression, MRSF filters skip information more selectively than standard additive fusion, enabling the decoder to recover lesion details while reducing the influence of artifact-contaminated high-frequency features.

4.5 Deep Supervision and Training Objective

ARL-UNet attaches auxiliary prediction heads to multiple decoder stages. These coarse predictions provide additional gradient supervision and serve as semantic guidance in MRSF. Let $\hat{\mathbf{y}}$ be the final prediction and $\{\hat{\mathbf{y}}_j\}_{j=1}^{5}$ be the auxiliary predictions, where $\hat{\mathbf{y}}_j$ is the coarse mask $\mathbf{M}_j$ upsampled to the input resolution. The base segmentation loss is a combination of binary cross-entropy and Dice loss:

$$\mathscr{L}_{\text{seg}}(\hat{\mathbf{y}}, \mathbf{y}) = \mathscr{L}_{\text{BCE}}(\hat{\mathbf{y}}, \mathbf{y}) + \mathscr{L}_{\text{Dice}}(\hat{\mathbf{y}}, \mathbf{y}). \tag{25}$$

The final training objective is

$$\mathscr{L} = \mathscr{L}_{\text{seg}}(\hat{\mathbf{y}}, \mathbf{y}) + \sum_{j=1}^{5} \lambda_j \mathscr{L}_{\text{seg}}(\hat{\mathbf{y}}_j, \mathbf{y}), \tag{26}$$

where λ_j denotes the auxiliary loss weight for decoder stage j. In our implementation, $\lambda_j = (6 - j)/10$ for $j = \{1, \ldots, 5\}$, assigning higher supervision weight to shallower decoder stages where spatial detail recovery is most critical. Unless otherwise specified, the same loss formulation is used for both clean-only training and artifact-aware training variants, so that the effect of the proposed architecture can be separated from that of data augmentation.

5 Experiments

This section evaluates ARL-UNet from four complementary perspectives: (i) clean-domain segmentation accuracy on ISIC 2017[7] and ISIC 2018[8,17] to verify that the artifact-robust design does not compromise standard segmentation performance; (ii) robustness on the proposed corrupted benchmarks under four artifact categories at two severity levels; (iii) zero-shot generalisation on PH2[47]; and (iv) ablation studies isolating the contribution of each architectural component and analysing the effect of artifact-aware training strategies. Qualitative visualizations are additionally provided to support interpretation.

5.1 Datasets

ISIC 2017 and ISIC 2018 serve as the primary training and evaluation datasets, both providing dermoscopic images with binary lesion masks. Their corresponding corrupted benchmarks are constructed following the protocol in Section 3, transforming each clean image into eight corrupted counterparts (four artifact types $\times$ two severity levels) while keeping the lesion masks unchanged. PH2[47] is used exclusively for zero-shot evaluation: models trained on ISIC are tested directly on PH2 without fine-tuning to assess whether the learned representation generalizes to a dataset with different acquisition characteristics.

Table 3. Dataset usage in the experiments.

Dataset	Training	Evaluation		
		Clean	Corrupted	Zero-shot
ISIC 2017	✓	✓	✓	–
ISIC 2018	✓	✓	✓	–
PH2	–	–	–	✓

5.2 Experimental Design

The experimental design is motivated by a central observation: clean-domain performance on ISIC benchmarks alone is insufficient to characterize lightweight dermoscopic segmentation models. A robust model must simultaneously achieve high clean-domain accuracy, minimal degradation under clinically relevant artifacts, and stable generalization to unseen data. To this end, the experiments are organized around three complementary evaluation settings, summarised in Table 3.

- **Clean-domain evaluation.** Models are trained and evaluated on standard ISIC 2017 and ISIC 2018 splits, following the conventional protocol in dermoscopic segmentation literature, to establish baseline accuracy.
- **Corrupted-domain evaluation.** Models are evaluated on our proposed corrupted benchmarks derived from ISIC 2017 and ISIC 2018, covering four clinically relevant artifact types: hair, specular highlights, air bubbles, and skin lines. Each at two severity levels $\{1, 2\}$.
- **Zero-shot generalisation.** Models trained on ISIC are evaluated directly on PH2 without fine-tuning, verifying that robustness transfers beyond the source distribution rather than overfitting the synthetic benchmark.

To disentangle the contributions of architecture and training strategy, we further analyze three training regimes applied to both ARL-UNet and selected lightweight baselines:

- $\mathscr{T}_0$**: Clean-only training.** Standard geometric and color augmentations are applied without synthetic artifact overlays, measuring the intrinsic robustness conferred by the architecture alone.
- $\mathscr{T}_1$**: Direct augmentation.** Synthetic artifacts are randomly injected throughout training, evaluating whether naive artifact exposure improves corrupted-domain robustness.
- $\mathscr{T}_2$**: Curriculum training.** Training begins with a clean-only phase before progressively introducing artifact augmentation, testing whether structured exposure can strengthen robustness while preserving clean-image accuracy.

The experimental protocol supports our central argument step by step. Clean-domain comparison establishes that ARL-UNet is competitive on standard benchmarks. Corrupted-domain and zero-shot evaluation, then verify that this performance is genuinely robust rather than distribution-specific. Module ablation identifies which architectural components drive robustness, while training-regime ablation clarifies the role of artifact augmentation strategies. All models share the same optimizer, input resolution, loss function, and data splits unless otherwise stated.

5.3 Implementation Details

All models are trained under a unified configuration summarised in Table 4. Unless a baseline requires its officially reported setup, the same hyperparameters, preprocessing pipeline, and evaluation scripts are applied across all compared methods to ensure a fair comparison. At test time, all models employ the same test-time augmentation (TTA) strategy comprising horizontal flipping, rotation, and resizing. For ARL-UNet, the encoder channel configuration is $[8, 16, 24, 36, 48, 72]$, and the skip-fusion module defaults to the proposed conservative soft artifact-guided fusion.

5.4 Compared Methods

For clean-domain evaluation, we compare ARL-UNet against eight representative segmentation models: TransFuse[48], Attention-UNet[13], UNet[5], UNeXt-S[44], MAL-UNet[16], EGE-UNet[45], UCM-Net[9], EMA-UNet[46], and DCE-UNet[39]. For robustness evaluation, a subset of lightweight baselines: UNet, UCM-Net, and EMA-UNet are retrained alongside ARL-UNet under three training regimes $\{\mathscr{T}_0, \mathscr{T}_1, \mathscr{T}_2\}$ to enable controlled comparison under corrupted-domain conditions.

Table 4. Unified training configuration applied to all models.

Setting	Value
Input resolution	256×256
Optimizer	AdamW
LR schedule	Cosine annealing
Initial learning rate	1×10^{-3}
Minimum learning rate	1×10^{-6}
Weight decay	1×10^{-2}
Batch size	8
Training epochs	320
Loss function	BCE + Dice

Table 5. Default curriculum schedule used in the main experiments. $\mathscr{A}$={Hair, highlight, air bubble, skin line}.

Stage	Epochs	Probability	Artifacts	Severity
Stage 1	1–160	0	None	–
Stage 2	161–240	0.50	$\mathscr{A}$	1
Stage 3	241–320	0.75	$\mathscr{A}$	1, 2

5.5 Evaluation Metrics

For clean-domain segmentation, we report mean Intersection over Union (mIoU), Dice coefficient, Accuracy (Acc.), Sensitivity (Sen.), Specificity (Spe.), and the 95th-percentile Hausdorff Distance (HD95). Higher values indicate better performance for all metrics except HD95. For corrupted-domain evaluation, each metric is averaged across the four artifact corruption types. Model efficiency is characterized by the number of parameters (Params, in Millions) and floating-point operations (GFLOPs), as reported in Table 6.

5.6 Clean-domain Comparison

We first evaluate ARL-UNet under the standard clean-domain setting on ISIC 2017 and ISIC 2018. All models are trained with clean-only data and evaluated on clean validation/test splits. This experiment verifies whether the robustness-oriented design maintains competitive clean-domain segmentation before considering artifact robustness.

Tables 7 and 8 summarize the results. On ISIC 2017, ARL-UNet achieves the best mIoU, Dice, Accuracy, and HD95. Compared with EMA-UNet, ARL-UNet improves mIoU and Dice by 0.94 and 0.57 percentage points, respectively, while reducing HD95 by 0.74.

In ISIC 2018, ARL-UNet again achieves the best overall performance across the aforementioned metrics. Relative to EMA-UNet, it improves mIoU and Dice by 0.97 and 0.58 percentage points, respectively, and reduces HD95 by 0.51.

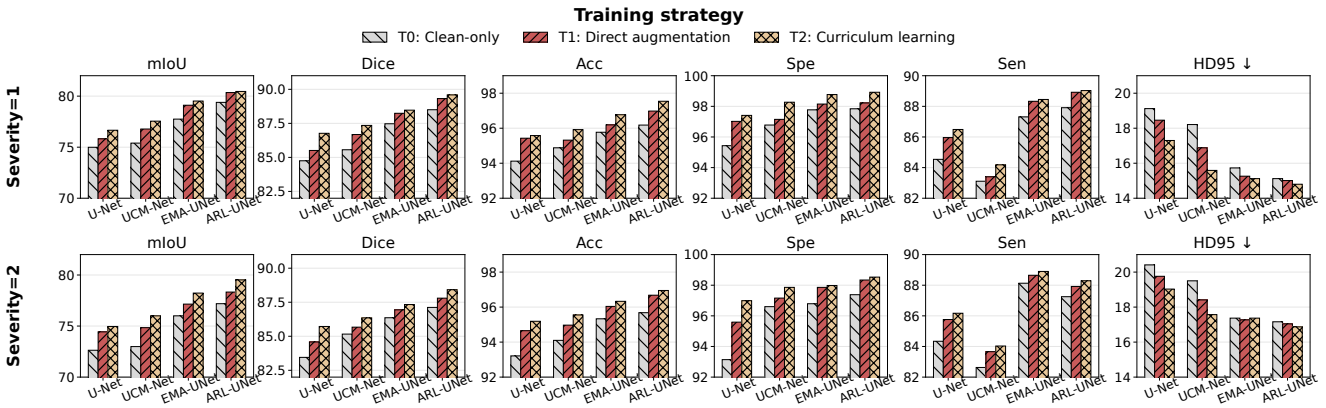


Figure 4. Main robustness comparison on the corrupted ISIC 2017 benchmark.

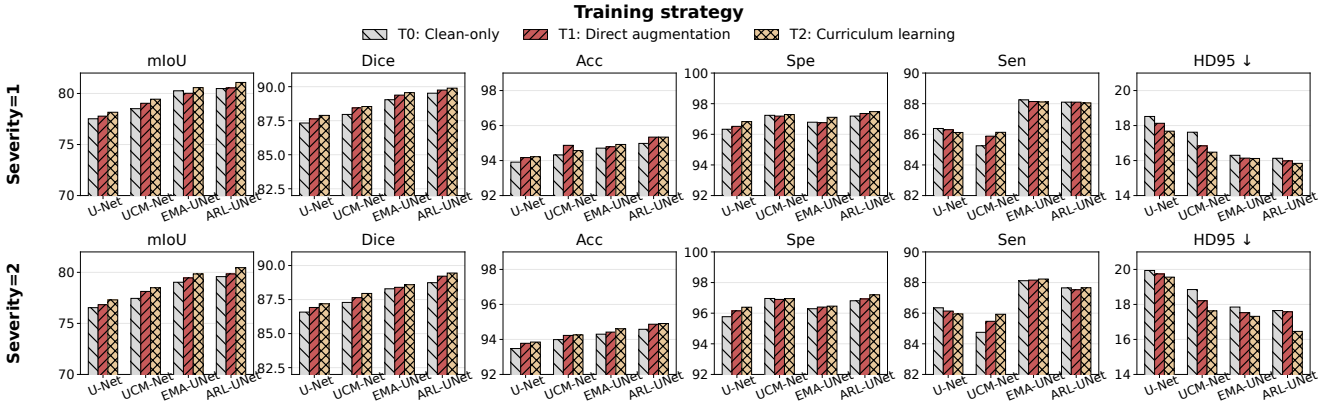


Figure 5. Main robustness comparison on the corrupted ISIC 2018 benchmark.

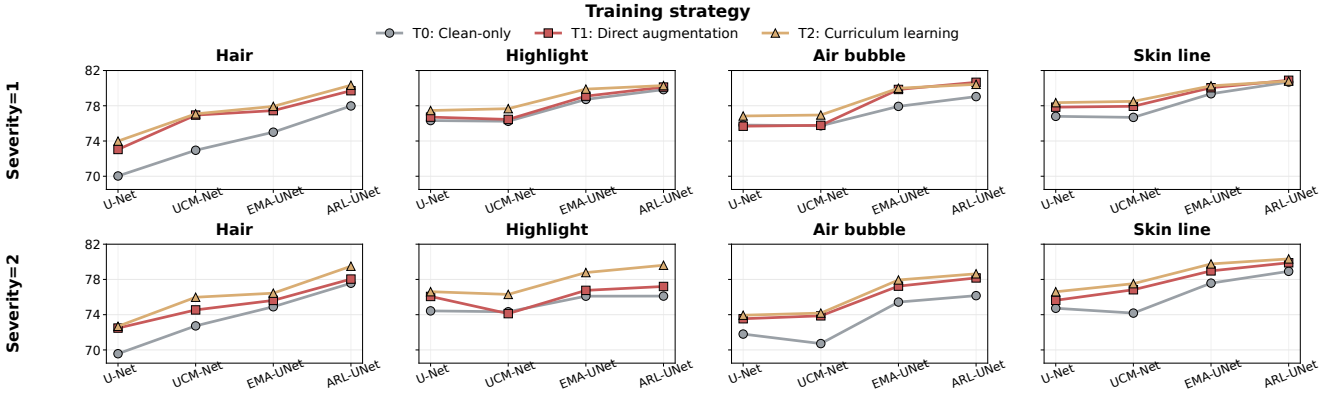


Figure 6. Per-artifact robustness mIoU on the corrupted ISIC 2017 benchmark.

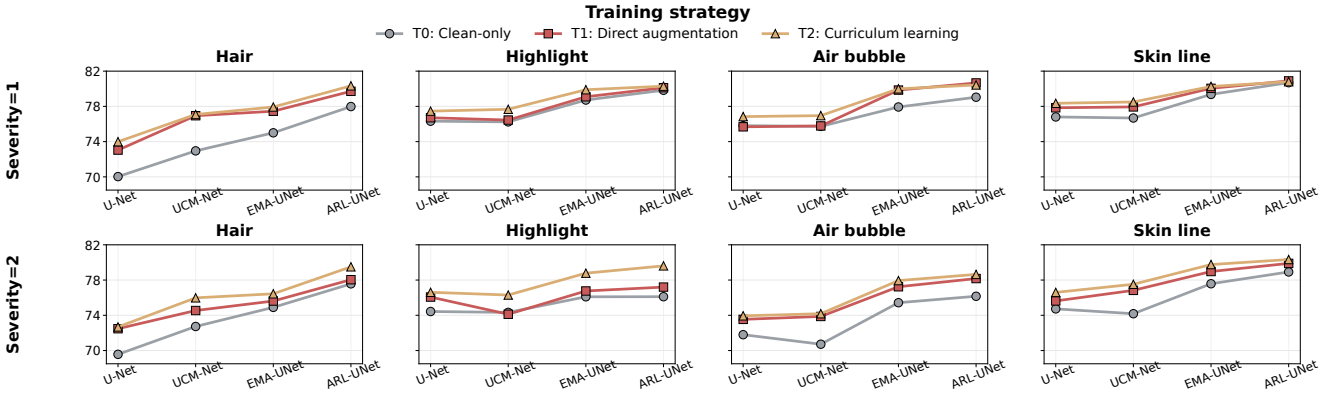


Figure 7. Per-artifact robustness mIoU on the corrupted ISIC 2018 benchmark.

Table 6. Model complexity comparison in terms of parameters (Params) and computational cost (GFLOPs).

Model	Params (M)	GFLOPs
TransFuse	26.270	11.530
Attention-UNet	7.980	14.430
UNet	7.770	13.780
DCE-UNet	0.370	1.341
UNeXt-S	0.300	0.100
MALUNet	0.175	0.083
EGE-UNet	0.053	0.072
UCM-Net	0.048	0.047
EMA-UNet	0.062	0.067
ARL-UNet (Ours)	0.147	0.113

Table 7. Clean-domain comparison on ISIC 2017.† denotes that results for competing methods are taken directly from[46]. The best and second best results are highlighted in red and blue.

Model	mIoU	Dice	Accuracy	Specificity	Sensitivity	HD95
TransFuse†	79.63 ± 0.40	88.66 ± 0.25	96.24 ± 0.17	97.96 ± 0.65	87.71 ± 2.29	14.86 ± 0.41
Attention-UNet†	77.24 ± 0.19	87.16 ± 0.12	95.76 ± 0.07	97.76 ± 0.17	90.47 ± 0.93	15.89 ± 0.56
UNeXt-S†	77.47 ± 0.58	87.30 ± 0.36	95.83 ± 0.13	97.88 ± 0.18	85.65 ± 0.59	16.48 ± 0.32
MAL-UNet†	78.57 ± 0.40	88.00 ± 0.25	96.04 ± 0.11	97.94 ± 0.27	86.63 ± 0.81	15.12 ± 0.20
EGE-UNet†	79.50 ± 0.67	88.58 ± 0.41	96.23 ± 0.13	98.02 ± 0.11	87.31 ± 0.82	14.94 ± 0.45
UNet	76.35 ± 0.25	86.57 ± 0.16	95.57 ± 0.04	97.78 ± 0.10	84.99 ± 0.52	20.17 ± 0.96
UCM-Net	78.40 ± 0.27	87.88 ± 0.18	96.06 ± 0.05	98.32 ± 0.30	85.04 ± 1.42	15.73 ± 0.50
EMA-UNet	80.11 ± 0.22	88.97 ± 0.12	96.37 ± 0.08	98.06 ± 0.26	87.81 ± 0.81	14.42 ± 0.37
DCE-UNet	79.78 ± 0.34	88.96 ± 0.36	95.25 ± 0.24	97.68 ± 0.45	88.54 ± 2.29	15.24 ± 0.49
ARL-UNet (Ours)	81.05 ± 0.19	89.54 ± 0.12	96.52 ± 0.05	98.05 ± 0.19	88.94 ± 0.78	13.68 ± 0.31

5.7 Robustness on Corrupted Dermoscopic Benchmarks

We evaluate artifact robustness on the corrupted ISIC 2017 and ISIC 2018 benchmarks. U-Net, UCM-Net, EMA-UNet, and ARL-UNet are trained under three regimes: clean-only training $\mathcal{T}_0$, direct artifact augmentation $\mathcal{T}_1$, and curriculum learning $\mathcal{T}_2$. This setting tests both zero-shot robustness to artifacts and the effect of artifact-aware training.

Fig. 4 and Fig. 5 summarize the main results. Under $\mathcal{T}_0$, ARL-UNet already improves mIoU over EMA-UNet at both severity levels, with gains of 1.64 and 1.19 points on ISIC 2017, and 0.22 and 0.55 points on ISIC 2018. It also achieves lower HD95 in these settings, indicating better boundary localization without artifact-augmented training.

With direct augmentation $\mathcal{T}_1$, ARL-UNet remains competitive in overlap-based metrics. Compared with EMA-UNet, it improves mIoU by 1.24 and 1.18 points on ISIC 2017 under Severity={1,2}, respectively, and achieves the highest mIoU on ISIC 2018 under both severities. This suggests that the proposed architecture can use artifact-augmented samples more effectively than the competing baselines.

Curriculum learning $\mathcal{T}_2$ gives the largest gains under stronger corruptions. Compared with clean-only training, ARL-UNet improves Severity 2 mIoU by 2.33 points on ISIC 2017 and 0.88 points on ISIC 2018. Under the same curriculum setting, it outperforms EMA-UNet by 1.29 and 0.62 points in Severity 2 mIoU on ISIC 2017 and ISIC 2018, respectively, while also obtaining lower HD95 in most settings.

Fig. 6 and Fig. 7 further report artifact-specific results. On ISIC 2017, ARL-UNet is particularly effective for hair and skin-line corruptions, which introduce boundary-like structures. Under Severity=2 and $\mathcal{T}_0$, it improves over EMA-UNet by 2.67 points for hair and 1.34 points for skin line. On ISIC 2018, ARL-UNet achieves the best mIoU across all four artifact types under $\mathcal{T}_2$ and Severity=2, with clear gains on highlight and skin-line corruptions. Overall, the corrupted-benchmark results show that ARL-UNet provides a better balance between lesion coverage and boundary precision across datasets, severity levels, training regimes, and artifact types.

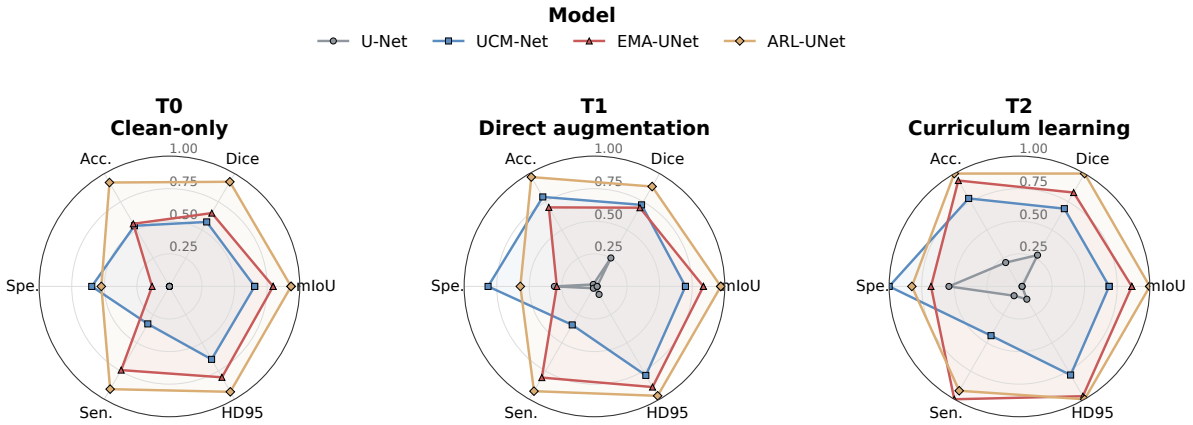


Figure 8. Zero-shot evaluation on PH2. Models are trained on ISIC 2017 and directly tested on PH2 without fine-tuning. (All axes are min-max normalized; HD95 is inverted so that larger is better.)

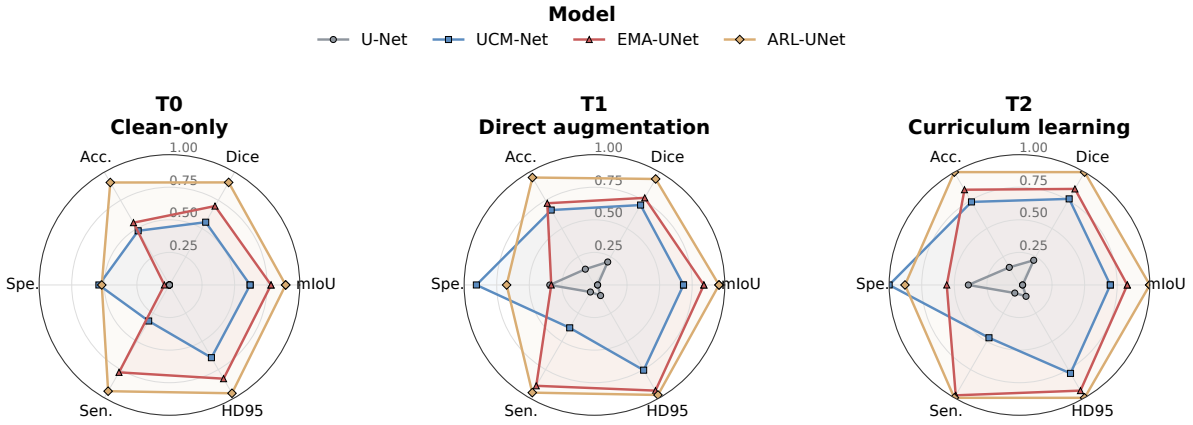


Figure 9. Zero-shot evaluation on PH2. Models are trained on ISIC 2018 and directly tested on PH2 without fine-tuning. (All axes are min-max normalized; HD95 is inverted so that larger is better.)

Table 8. Clean-domain comparison on ISIC 2018.† denotes that results for competing methods are taken directly from[46]. he best and second best results are highlighted in red and blue.

Model	mIoU	Dice	Accuracy	Specificity	Sensitivity	HD95
TransFuse†	80.10 ± 0.34	88.95 ± 0.21	94.62 ± 0.12	96.45 ± 0.43	88.94 ± 1.06	16.71 ± 0.32
Attention-UNet†	78.67 ± 0.23	88.06 ± 0.15	94.29 ± 0.12	96.77 ± 0.39	86.56 ± 0.76	19.09 ± 0.49
UNeXt-S†	79.18 ± 0.09	88.38 ± 0.06	94.37 ± 0.08	96.41 ± 0.60	88.02 ± 1.53	17.73 ± 0.33
MAL-UNet†	79.60 ± 0.31	88.64 ± 0.19	94.50 ± 0.13	96.57 ± 0.67	88.09 ± 1.73	17.02 ± 0.58
EGE-UNet†	80.28 ± 0.22	89.06 ± 0.14	94.66 ± 0.10	96.38 ± 0.56	89.32 ± 1.47	16.18 ± 0.22
UNet	78.20 ± 0.16	87.77 ± 0.11	94.16 ± 0.06	96.86 ± 0.36	85.74 ± 0.97	19.78 ± 0.71
UCM-Net	79.94 ± 0.20	88.88 ± 0.14	94.69 ± 0.08	97.17 ± 0.57	87.07 ± 1.58	16.48 ± 0.38
EMA-UNet	80.75 ± 0.41	89.37 ± 0.25	94.90 ± 0.12	96.98 ± 0.28	88.30 ± 0.81	15.81 ± 0.35
DCE-UNet	80.53 ± 0.34	89.47 ± 0.32	94.24 ± 0.14	96.74 ± 0.31	88.85 ± 1.56	16.22 ± 0.51
ARL-UNet (Ours)	81.72 ± 0.23	89.95 ± 0.12	95.17 ± 0.05	97.19 ± 0.19	88.89 ± 0.78	15.30 ± 0.30

Table 9. Module ablation of ARL-UNet on ISIC 2018 with segmentation performance and efficiency measurements. AvgCorr is computed over all artifact types and severity levels in the corrupted benchmark. Latency and memory are measured with 256×256 input and batch size 1.

Variant	mIoU		Dice		Efficiency			
	Clean	AvgCorr	Clean	AvgCorr	Params (M)	FLOPs (G)	Latency (ms)	Memory (MB)
w/o AAS	80.96	78.90	88.41	86.65	0.1470	0.1005	2.718	63.28
w/o REB	81.08	79.30	88.87	87.82	0.1209	0.0947	2.093	72.88
w/o MRSF	80.75	78.59	88.43	86.22	0.0998	0.0947	2.431	72.15
Full	**81.72**	**80.03**	**89.95**	**89.13**	0.1472	0.1135	2.939	83.53

5.8 PH2 Zero-shot Evaluation

We further evaluate cross-dataset generalization by directly testing ISIC-trained models on PH2 without fine-tuning. U-Net, UCM-Net, EMA-UNet, and ARL-UNet are compared under clean-only training $\mathcal{T}_0$, direct artifact augmentation $\mathcal{T}_1$, and curriculum learning $\mathcal{T}_2$.

Fig. 8 and Fig. 9 summarize the zero-shot results. Across both ISIC source datasets, ARL-UNet achieves the best mIoU, Dice, and HD95 in all three training regimes. This indicates that its gains are not limited to the corrupted ISIC benchmark, but also transfer to an external dermoscopic dataset. UCM-Net and EMA-UNet occasionally obtain slightly higher Specificity or Sensitivity, but these isolated gains do not lead to better overlap accuracy or boundary distance.

Curriculum learning $\mathcal{T}_2$ gives the best PH2 transfer performance for ARL-UNet on both source datasets. Compared with its clean-only baseline, $\mathcal{T}_2$ improves mIoU and reduces HD95 simultaneously, showing that gradual artifact exposure improves transferability without degrading cross-dataset segmentation accuracy.

5.9 Qualitative Visualization

We provide qualitative comparisons on representative corrupted dermoscopic images under the $\mathcal{T}_2$ training strategy. Each column shows a clean or corrupted input with the corresponding artifact type and severity, and each row compares the ground truth with predictions from ARL-UNet, UCM-Net, EMA-UNet, and U-Net.

As shown in Fig. 10, the baselines show different failure patterns. U-Net often over-expands the lesion mask, especially under severe highlight and hair corruptions. EMA-UNet and UCM-Net produce more compact masks, but they tend to miss boundary regions or deform the lesion contour under air-bubble and skin-line artifacts.

ARL-UNet gives predictions that are more aligned with the ground-truth contour. Under hair and skin-line corruptions, it reduces thin artifact-induced responses while preserving the main lesion boundary. Under highlight and air-bubble corruptions, it avoids the strong over-expansion of U-Net and retains more complete lesion regions than EMA-UNet. These visual results support the quantitative improvements in overlap accuracy and boundary-distance metrics under corrupted conditions.

5.10 Ablation Study on ARL-UNet Modules

To verify that the performance gains of ARL-UNet come from the proposed architectural design rather than incidental parameter changes, we ablate its three main components: AAS, REB, and MRSF. Removing AAS replaces the artifact-aware input stem

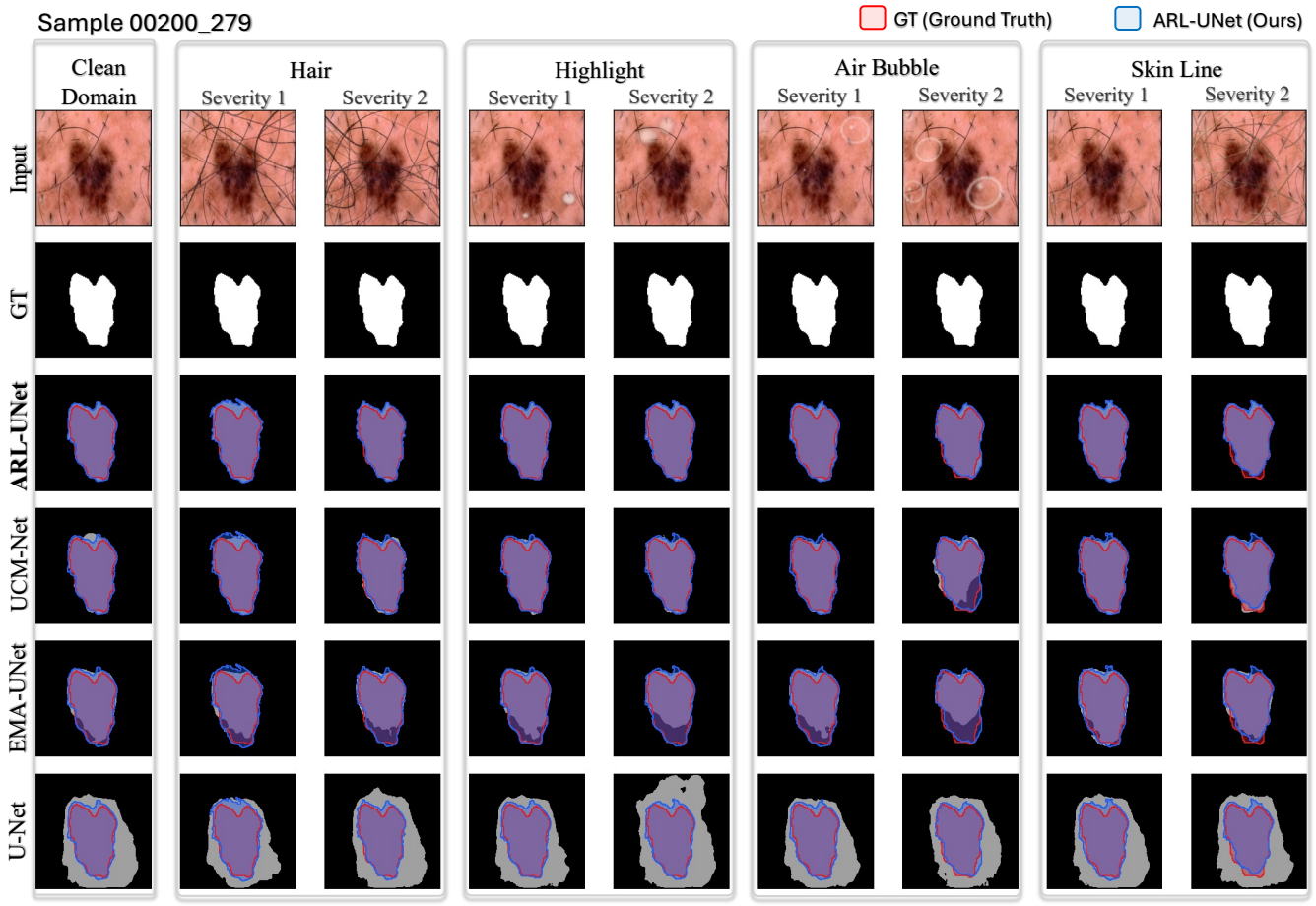


Figure 10. Qualitative comparison under different dermoscopic artifact corruptions. ARL-UNet better preserves lesion contours and shows fewer artifact-induced mask distortions than the competing baselines.

Table 10. Per-artifact module ablation on ISIC 2018. Each entry reports mIoU, with the mIoU drop relative to the full ARL-UNet shown in parentheses.

Sev.	Full	w/o AAS	w/o REB	w/o MRSF
Hair				
1	**80.48**	79.22 (↓1.26)	79.56 (↓0.92)	78.86 (↓1.62)
2	**80.20**	78.68 (↓1.52)	79.19 (↓1.01)	78.44 (↓1.76)
Highlight				
1	**80.37**	79.19 (↓1.18)	79.75 (↓0.62)	79.46 (↓0.91)
2	**78.94**	77.72 (↓1.22)	78.04 (↓0.90)	77.41 (↓1.53)
Air Bubble				
1	**80.04**	79.00 (↓1.04)	79.55 (↓0.49)	78.47 (↓1.57)
2	**78.87**	78.10 (↓0.77)	78.19 (↓0.68)	77.66 (↓1.21)
Skin Line				
1	**81.03**	80.15 (↓0.88)	80.42 (↓0.61)	79.54 (↓1.49)
2	**80.30**	79.12 (↓1.18)	79.64 (↓0.66)	78.86 (↓1.44)

with a plain convolutional stem, removing REB replaces the robust encoder block with a simpler depthwise separable residual block, and removing MRSF replaces mask-guided artifact-aware skip fusion with standard additive skip fusion. All variants are evaluated on the clean and corrupted ISIC 2018 benchmarks under the $\mathcal{T}_0$ clean-only training strategy.

Table 11. Ablation of training strategies for ARL-UNet on the corrupted ISIC 2018 benchmark.

Training strategy	Clean mIoU	AvgCorr mIoU	Severity-1 Avg	Severity-2 Avg
Clean-only	81.72	80.03	80.48	79.58
Direct augmentation	81.23	80.19	80.54	79.84
Curr. training	81.70	80.77	81.08	80.46
Curr. w/o clean	81.11	80.35	80.77	79.92
Curr. w/o Sever=2	81.56	80.24	81.11	79.37

As shown in Table 9, the full ARL-UNet achieves the best clean-domain and corrupted-domain performance. Removing any module degrades AvgCorr mIoU, with drops of 1.13, 0.73, and 1.44 percentage points for w/o AAS, w/o REB, and w/o MRSF, respectively. This indicates that artifact-aware skip fusion is particularly important for preventing corrupted shallow features from being directly propagated into the decoder.

The full model remains lightweight, with only 0.1472M parameters and 0.1135G FLOPs. Compared with the ablated variants, the additional latency is within a sub-millisecond range, while the full model consistently improves both clean and corrupted segmentation accuracy.

Table 10 further reports artifact-specific degradation relative to the full model. Removing AAS causes clear drops under boundary-like artifacts, with mIoU decreases of 1.52 and 1.18 points, respectively. Removing MRSF leads to broader degradation across all artifact categories; under Severity 2, the drops are 1.76, 1.53, 1.21, and 1.44 points for hair, highlight, air bubble, and skin line, respectively. In contrast, removing REB produces smaller but consistent decreases, suggesting that REB mainly improves general feature encoding, while AAS and MRSF more directly address artifact-induced degradation.

5.11 Ablation Study on Curriculum Training

We further examine the design of artifact curriculum training. Direct artifact augmentation may expose the model to corrupted samples too early, which can interfere with clean lesion representation. In contrast, curriculum training first learns from clean images and then gradually introduces artifact-corrupted samples. We compare clean-only training, direct artifact augmentation, the full curriculum strategy, and two curriculum variants. "Curr. w/o clean" removes the clean warm-up stage, while "Curr. w/o Sev=2" excludes Severity-2 corruptions during training. More training details are provided in Appendix D.

Table 11 shows that curriculum training achieves the best corrupted-domain robustness while almost preserving clean-domain accuracy. Compared with clean-only training, it improves AvgCorr mIoU by 0.74%, with gains of 0.60% and 0.88% under Severity=$\{1, 2\}$, respectively. Compared with direct augmentation, curriculum training further improves AvgCorr mIoU by 0.58% and recovers most of the clean-domain drop caused by direct exposure to corrupted samples.

The two curriculum variants further validate the training design. Removing the clean warm-up stage reduces clean mIoU by 0.59 points and AvgCorr mIoU by 0.42% compared with the full curriculum, indicating that early clean-domain learning remains useful for stable lesion representation. Excluding Severity=2 corruptions slightly improves Severity=1 performance, but reduces Severity=2 Avg by 1.09% and lowers AvgCorr mIoU by 0.53%. This suggests that mild artifact exposure alone is insufficient for severe corruptions.

6 Discussion

This study focuses on controlled robustness evaluation for lightweight dermoscopic lesion segmentation. The proposed corrupted benchmark provides a reproducible way to measure how different artifact types and severity levels affect segmentation performance. Because each corrupted image is paired with the original clean image and unchanged lesion mask, the benchmark enables artifact-specific analysis that is difficult to obtain from naturally collected images alone.

However, the benchmark is based on synthetic artifact generation and should not be regarded as a substitute for real-world clinical validation. Naturally acquired dermoscopic artifacts may have more complex appearances, mixed artifact patterns, device-dependent characteristics, and correlations with lesion type, skin tone, acquisition protocol, and image quality. In contrast, our synthetic corruptions are designed to isolate specific degradation factors such as hair-like structures, highlight saturation, air-bubble occlusion, and skin-line texture. Therefore, the benchmark is best interpreted as a controlled stress test for reproducible robustness analysis rather than direct evidence of clinical deployment readiness.

The proposed ARL-UNet improves robustness under this controlled setting by explicitly modeling artifact-prone regions and reducing the propagation of contaminated shallow features. Nevertheless, further validation on naturally acquired real-artifact images and multi-center clinical datasets is necessary to assess practical generalization. Future work will extend the benchmark with real artifact annotations, mixed-artifact cases, and device-specific acquisition variations, and will evaluate whether the proposed artifact-aware design remains effective under prospective clinical conditions.

7 Conclusion

In this paper, we presented an artifact-aware lightweight framework for robust dermoscopic lesion segmentation and a reproducible corrupted benchmark for systematic robustness evaluation. The proposed ARL-UNet uses artifact cue estimation, efficient robust encoding, and mask-guided skip fusion to reduce the influence of hair, specular highlights, air bubbles, and skin-line artifacts while preserving lesion boundaries. The benchmark is built from fixed ISIC validation splits with four clinically relevant artifact categories, two severity levels, unchanged ground-truth masks, and recorded artifact metadata. It supports reproducible evaluation of clean-domain accuracy, corrupted-domain robustness, and artifact-specific degradation.

Experiments show that ARL-UNet maintains competitive clean-domain performance, improves robustness under corrupted conditions, and transfers better to PH2 without fine-tuning. Ablation studies further confirm the contribution of the proposed modules and curriculum training. Overall, this work provides both a lightweight segmentation model and a reusable evaluation protocol for artifact-robust dermoscopic image analysis. Future work will extend the benchmark to more diverse real-world clinical artifacts and multi-center datasets.

Funding

This work was supported by the Guangxi Science and Technology Program under Grant Nos. GuiKeFN2600640336 and GuiKeFN2600640049, and the Natural Science Foundation of Guangxi under Grant No. 2026GXNSFAA00640439.

Author contributions statement

Haixiao Gong: Conceptualization, Methodology, Software, Validation, Formal analysis, Writing—original draft. Cuijiao Yin: Conceptualization, Supervision, Project administration, Writing—review and editing. Jie He: Software, Validation, Investigation. Hui Guo: Investigation, Data curation. Mei Tang: Data curation, Visualization. Linting Meng: Investigation. Chengchen Gong: Visualization. All authors reviewed the manuscript.

Additional information

The author declares no competing interests.

References

1. Esteva, A. *et al.* Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**, 115–118, DOI: 10.1038/nature21056 (2017).
2. Barata, C., Celebi, M. E. & Marques, J. S. A survey of feature extraction in dermoscopy image analysis of skin cancer. *IEEE J. Biomed. Heal. Informatics* **23**, 1096–1109, DOI: 10.1109/JBHI.2018.2845939 (2019).
3. Litjens, G. *et al.* A survey on deep learning in medical image analysis. *Med. Image Analysis* **42**, 60–88, DOI: 10.1016/j.media.2017.07.005 (2017).
4. Howard, A. G. *et al.* MobileNets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861* (2017). 1704.04861.
5. Ronneberger, O., Fischer, P. & Brox, T. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, vol. 9351 of *Lecture Notes in Computer Science*, 234–241, DOI: 10.1007/978-3-319-24574-4_28 (Cham, 2015).
6. Zhou, Z., Rahman Siddiquee, M. M., Tajbakhsh, N. & Liang, J. UNet++: A nested U-Net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support (DLMIA/ML-CDS 2018)*, vol. 11045 of *Lecture Notes in Computer Science*, 3–11, DOI: 10.1007/978-3-030-00889-5_1 (Cham, 2018).
7. Codella, N. C. F. *et al.* Skin lesion analysis toward melanoma detection: A challenge at the 2017 International Symposium on Biomedical Imaging (ISBI), hosted by the International Skin Imaging Collaboration (ISIC). *arXiv preprint arXiv:1710.05006* (2018). 1710.05006.
8. Codella, N. *et al.* Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the International Skin Imaging Collaboration (ISIC). *arXiv preprint arXiv:1902.03368* (2019). 1902.03368.
9. Yuan, C., Zhao, D. & Agaian, S. S. UCM-Net: A lightweight and efficient solution for skin lesion segmentation using MLP and CNN. *Biomed. Signal Process. Control.* **96**, 106573, DOI: 10.1016/j.bspc.2024.106573 (2024).

10. Yuan, C., Zhao, D. & Agaian, S. S. MUCM-Net: A Mamba powered UCM-Net for skin lesion segmentation. *arXiv preprint arXiv:2405.15925* (2024). 2405.15925.

11. Hu, J., Shen, L. & Sun, G. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 7132–7141, DOI: 10.1109/CVPR.2018.00745 (2018).

12. Woo, S., Park, J., Lee, J.-Y. & Kweon, I. S. CBAM: Convolutional block attention module. In *Computer Vision – ECCV 2018*, vol. 11211 of *Lecture Notes in Computer Science*, 3–19, DOI: 10.1007/978-3-030-01234-2_1 (Cham, 2018).

13. Oktay, O. *et al.* Attention U-Net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999* (2018). 1804.03999.

14. Nie, T., Zhao, Y. & Yao, S. ELA-Net: An efficient lightweight attention network for skin lesion segmentation. *Sensors* **24**, 4302, DOI: 10.3390/s24134302 (2024).

15. Farooq, H. *et al.* LSSF-Net: Lightweight segmentation with self-awareness, spatial attention, and focal modulation. *Artif. Intell. Medicine* **158**, 103012, DOI: 10.1016/j.artmed.2024.103012 (2024).

16. Ruan, J., Xie, M., Gao, J., Liu, T. & Fu, Y. MALUNet: A multi-attention and light-weight UNet for skin lesion segmentation. In *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 1150–1156, DOI: 10.1109/BIBM55620.2022.9995040 (2022).

17. Tschandl, P., Rosendahl, C. & Kittler, H. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Sci. Data* **5**, 180161, DOI: 10.1038/sdata.2018.161 (2018).

18. Lee, T., Ng, V., Gallagher, R., Coldman, A. & McLean, D. DullRazor: A software approach to hair removal from images. *Comput. Biol. Medicine* **27**, 533–543, DOI: 10.1016/S0010-4825(97)00020-6 (1997).

19. Toossi, M. T. B. *et al.* An effective hair removal algorithm for dermoscopy images. *Ski. Res. Technol.* **19**, 230–235, DOI: 10.1111/srt.12015 (2013).

20. Hammouda, N. G., Shalaby, M., Alfilh, R. H. & Singh, N. S. S. Hair and artifact removal in dermoscopic images using deep learning for enhanced skin cancer detection. *Sci. Reports* **16**, 14452, DOI: 10.1038/s41598-026-44545-x (2026).

21. Isensee, F., Jaeger, P. F., Kohl, S. A. A., Petersen, J. & Maier-Hein, K. H. nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. *Nat. Methods* **18**, 203–211, DOI: 10.1038/s41592-020-01008-z (2021).

22. Xu, Q., Ma, Z., Duan, W. *et al.* DCSAU-Net: A deeper and more compact split-attention U-Net for medical image segmentation. *Comput. Biol. Medicine* **154**, 106626, DOI: 10.1016/j.compbiomed.2023.106626 (2023).

23. Hendrycks, D. & Dietterich, T. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations (ICLR)* (2019).

24. Bengio, Y., Louradour, J., Collobert, R. & Weston, J. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning (ICML)*, vol. 382 of *ACM International Conference Proceeding Series*, 41–48, DOI: 10.1145/1553374.1553380 (2009).

25. Hasan, M. K., Dahal, L., Samarakoon, P. N., Tushar, F. I. & Martí, R. DSNet: Automatic dermoscopic skin lesion segmentation. *Comput. Biol. Medicine* **120**, 103738, DOI: 10.1016/j.compbiomed.2020.103738 (2020).

26. Iqbal, S. *et al.* LDMRes-Net: A lightweight neural network for efficient medical image segmentation on IoT and edge devices. *IEEE J. Biomed. Heal. Informatics* **28**, 3860–3871, DOI: 10.1109/JBHI.2023.3331278 (2024).

27. Tarn, W.-H., Chong, C. H., Wang, L., Kuo, C.-F. & Chen, J. A lightweight image segmentation network leveraging inception and squeeze-excitation modules for efficient skin lesion analysis. *Eng. Appl. Artif. Intell.* **159**, 111541, DOI: 10.1016/j.engappai.2025.111541 (2025).

28. Yang, B. *et al.* SLP-Net: An efficient lightweight network for segmentation of skin lesions. *Biomed. Signal Process. Control.* **101**, 107242, DOI: 10.1016/j.bspc.2024.107242 (2025).

29. Cao, H. *et al.* Swin-Unet: Unet-like pure transformer for medical image segmentation. In *Computer Vision – ECCV 2022 Workshops*, vol. 13803 of *Lecture Notes in Computer Science*, 205–218, DOI: 10.1007/978-3-031-25066-8_9 (Cham, 2023).

30. Chen, J. *et al.* TransUNet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306* (2021). 2102.04306.

31. Gu, A. & Dao, T. Mamba: Linear-time sequence modeling with selective state spaces. In *First Conference on Language Modeling* (2024).

32. Ma, J., Li, F. & Wang, B. U-Mamba: Enhancing long-range dependency for biomedical image segmentation. *arXiv preprint arXiv:2401.04722* (2024). 2401.04722.

33. Celebi, M. E., Mendonca, T. & Marques, J. S. (eds.) *Dermoscopy Image Analysis* (CRC Press, Boca Raton, FL, 2015).

34. Bardou, D., Bouaziz, H., Lv, L. & Zhang, T. Hair removal in dermoscopy images using variational autoencoders. *Ski. Res. Technol.* **28**, 445–454, DOI: 10.1111/srt.13145 (2022).

35. Li, W., Joseph Raj, A. N., Tjahjadi, T. & Zhuang, Z. Digital hair removal by deep learning for skin lesion segmentation. *Pattern Recognit.* **117**, 107994, DOI: 10.1016/j.patcog.2021.107994 (2021).

36. Li, Y., Hu, B., Wang, X. & Liu, K. Lightweight UNet with depthwise separable convolution for skin lesion segmentation. *Sci. Reports* **15**, 33211, DOI: 10.1038/s41598-025-16683-1 (2025).

37. Michaelis, C. *et al.* Benchmarking robustness in object detection: Autonomous driving when winter is coming. *arXiv preprint arXiv:1907.07484* (2019). 1907.07484.

38. Wang, S., Veldhuis, R., Brune, C. & Strisciuglio, N. A survey on the robustness of computer vision models against common corruptions. *arXiv preprint arXiv:2305.06024* (2023). 2305.06024.

39. Chen, J. & Hu, Z. Efficient sea-sky line detection via lightweight segmentation and dual multi-scale fusion. *The Vis. Comput.* **42**, 275, DOI: 10.1007/s00371-026-04491-1 (2026).

40. Wu, X., Zhan, G., Han, C. & Zhang, W. Multi-domain fusion network for enhanced polyp segmentation in colonoscopy images. *The Vis. Comput.* **42**, 209, DOI: 10.1007/s00371-026-04419-9 (2026).

41. Wang, A., Islam, M., Xu, M. & Ren, H. Curriculum-based augmented Fourier domain adaptation for robust medical image segmentation. *IEEE Transactions on Autom. Sci. Eng.* **21**, 4340–4353, DOI: 10.1109/TASE.2023.3295600 (2024).

42. Finlayson, S. G. *et al.* The clinician and dataset shift in artificial intelligence. *New Engl. J. Medicine* **385**, 283–286, DOI: 10.1056/NEJMc2104626 (2021).

43. Mirikharaji, Z. *et al.* A survey on deep learning for skin lesion segmentation. *Med. Image Analysis* **88**, 102863, DOI: 10.1016/j.media.2023.102863 (2023).

44. Valanarasu, J. M. J. & Patel, V. M. UNeXt: MLP-based rapid medical image segmentation network. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, vol. 13435 of *Lecture Notes in Computer Science*, 23–33, DOI: 10.1007/978-3-031-16443-9_3 (Cham, 2022).

45. Ruan, J., Xie, M., Gao, J., Liu, T. & Fu, Y. EGE-UNet: An efficient group enhanced UNet for skin lesion segmentation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, Lecture Notes in Computer Science, 481–490, DOI: 10.1007/978-3-031-43901-8_46 (Cham, 2023).

46. Ma, L., Wang, K., Wang, J., Shao, D. & Yi, S. EMA-U-Net: Efficient multi-attention U-Net for skin lesion segmentation. *The Vis. Comput.* **42**, 237, DOI: 10.1007/s00371-026-04432-y (2026).

47. Mendonça, T., Ferreira, P. M., Marques, J. S., Marcal, A. R. S. & Rozeira, J. PH2 – a dermoscopic image database for research and benchmarking. In *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 5437–5440, DOI: 10.1109/EMBC.2013.6610779 (2013).

48. Zhang, Y., Liu, H. & Hu, Q. TransFuse: Fusing transformers and CNNs for medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, vol. 12901 of *Lecture Notes in Computer Science*, 14–24, DOI: 10.1007/978-3-030-87193-2_2 (Cham, 2021).

A Detailed Results of Corrupted Dermoscopic Benchmarks

This section provides the full numerical results corresponding to the robustness analyses summarized in the main text of 5.7. While the main paper reports the key trends using compact visual summaries, the tables in this section list the complete evaluation metrics for each dataset, training regime, corruption severity, and artifact type.

Tables 12 and 13 report the main robustness comparison on the corrupted ISIC 2017 and ISIC 2018 benchmarks, respectively. For each training strategy $\mathscr{T}_0$–$\mathscr{T}_2$, we provide results under both Severity 1 and Severity 2 corruptions. These tables include mIoU, Dice, Accuracy, Specificity, Sensitivity, and HD95, allowing a more complete comparison of region overlap, pixel-level classification, and boundary localization performance.

Tables 14 and 15 further present the per-artifact mIoU breakdown for Terminal Hair, Specular Highlight, Air Bubble, and Skin Line corruptions. These results complement the main analysis by showing how each model behaves under different visual disturbance patterns. Overall, the detailed tables confirm the trends discussed in Section 5.7: ARL-UNet consistently achieves stronger corrupted-domain robustness, particularly under severe corruptions and artifact types that interfere with lesion boundaries.

Table 12. Main robustness comparison on the corrupted ISIC 2017 benchmark.

Model	Severity=1						Severity=2					
	mIoU	Dice	Acc.	Spe.	Sen.	HD95	mIoU	Dice	Acc.	Spe.	Sen.	HD95
$\mathcal{T}_0$: *Clean-only*												
U-Net	74.99	84.75	94.12	95.43	84.54	19.12	72.63	83.45	93.21	93.14	84.34	20.41
UCM-Net	75.40	85.56	94.88	96.78	83.11	18.21	72.99	85.16	94.10	96.60	82.63	19.50
EMA-UNet	77.75	87.47	95.77	97.77	87.31	15.73	76.00	86.36	95.33	96.79	88.13	17.37
ARL-UNet	79.39	88.50	96.18	97.84	87.91	15.12	77.19	87.12	95.68	97.38	87.25	17.16
$\mathcal{T}_1$: *Direct augmentation*												
U-Net	75.82	85.51	95.43	97.02	85.96	18.46	74.43	84.58	94.65	95.59	85.76	19.76
UCM-Net	76.78	86.68	95.32	97.15	83.40	16.88	74.84	85.67	94.97	97.16	83.66	18.41
EMA-UNet	79.11	88.24	96.19	98.15	88.33	15.26	77.15	86.94	96.04	97.86	88.65	17.27
ARL-UNet	80.35	89.32	96.98	98.22	88.92	15.01	78.33	87.80	96.68	98.34	87.92	17.04
$\mathcal{T}_2$: *Curriculum learning*												
U-Net	76.66	86.77	95.58	97.41	86.49	17.30	74.96	85.72	95.19	96.99	86.17	19.03
UCM-Net	77.55	87.35	95.92	98.27	84.18	15.59	76.00	86.35	95.55	97.86	84.03	17.57
EMA-UNet	79.52	88.47	96.77	98.77	88.45	15.12	78.23	87.33	96.33	97.98	88.90	17.37
ARL-UNet	80.46	89.59	97.54	98.92	89.03	14.80	79.52	88.42	96.95	98.53	88.29	16.87

Table 13. Main robustness comparison on the corrupted ISIC 2018 benchmark. AvgCorr denotes the average mIoU over all artifact types and severity levels. Drop denotes the decrease from clean mIoU to AvgCorr.

Model	Severity=1						Severity=2					
	mIoU	Dice	Acc.	Spe.	Sen.	HD95	mIoU	Dice	Acc.	Spe.	Sen.	HD95
$\mathcal{T}_0$: *Clean-only*												
U-Net	77.52	87.33	93.91	96.33	86.38	18.52	76.54	86.58	93.48	95.77	86.35	19.94
UCM-Net	78.51	87.96	94.32	97.24	85.25	17.62	77.45	87.29	93.99	96.96	84.75	18.85
EMA-UNet	80.26	89.04	94.71	96.79	88.25	16.30	79.03	88.29	94.30	96.29	88.12	17.85
ARL-UNet	80.48	89.52	94.98	97.19	88.10	16.13	79.58	88.74	94.58	96.81	87.65	17.65
$\mathcal{T}_1$: *Direct augmentation*												
U-Net	77.78	87.63	94.17	96.52	86.30	18.13	76.83	86.92	93.78	96.15	86.14	19.74
UCM-Net	79.04	88.44	94.87	97.19	85.88	16.84	78.13	87.64	94.23	96.90	85.47	18.21
EMA-UNet	80.02	89.38	94.79	96.76	88.14	16.14	79.45	88.40	94.42	96.40	88.16	17.53
ARL-UNet	80.54	89.75	95.34	97.36	88.10	15.98	79.84	89.22	94.87	96.93	87.53	17.58
$\mathcal{T}_2$: *Curriculum learning*												
U-Net	78.16	87.89	94.22	96.83	86.12	17.68	77.30	87.19	93.85	96.39	85.95	19.56
UCM-Net	79.44	88.54	94.57	97.29	86.13	16.48	78.49	87.95	94.26	96.95	85.93	17.63
EMA-UNet	80.57	89.56	94.92	97.11	88.11	16.12	79.84	88.59	94.61	96.46	88.23	17.32
ARL-UNet	81.08	89.89	95.34	97.48	88.05	15.84	80.46	89.45	94.91	97.20	87.66	16.46

Table 14. Per-artifact robustness mIoU on the corrupted ISIC 2017 benchmark. “Sever" denotes Severity.

Model	Hair		Highlight		Air bubble		Skin line	
	Sever=1	Sever=2	Sever=1	Sever=2	Sever=1	Sever=2	Sever=1	Sever=2
$\mathcal{T}_0$: *Clean-only*								
U-Net	70.03	69.57	76.32	74.43	75.81	71.80	76.80	74.73
UCM-Net	72.95	72.73	76.24	74.33	75.73	70.72	76.68	74.18
EMA-UNet	75.00	74.90	78.72	76.10	77.93	75.43	79.36	77.58
ARL-UNet	77.98	77.57	79.83	76.11	79.04	76.16	80.69	78.92
$\mathcal{T}_1$: *Direct augmentation*								
U-Net	73.04	72.48	76.71	76.07	75.68	73.54	77.84	75.63
UCM-Net	76.94	74.54	76.45	74.12	75.78	73.87	77.94	76.83
EMA-UNet	77.45	75.62	79.09	76.76	79.85	77.24	80.05	78.97
ARL-UNet	79.71	78.04	80.12	77.20	80.66	78.17	80.89	79.89
$\mathcal{T}_2$: *Curriculum learning*								
U-Net	73.99	72.68	77.46	76.61	76.84	73.94	78.35	76.60
UCM-Net	77.08	75.98	77.67	76.30	76.95	74.18	78.50	77.52
EMA-UNet	77.93	76.44	79.89	78.78	79.99	77.94	80.26	79.76
ARL-UNet	80.33	79.50	80.29	79.61	80.43	78.64	80.80	80.33

Table 15. Per-artifact robustness mIoU on the corrupted ISIC 2018 benchmark. “Sever" denotes Severity.

Model	Hair		Highlight		Air bubble		Skin line	
	Sever=1	Sever=2	Sever=1	Sever=2	Sever=1	Sever=2	Sever=1	Sever=2
$\mathcal{T}_0$: *Clean-only*								
U-Net	75.86	75.49	78.13	77.41	77.76	75.90	78.32	77.37
UCM-Net	77.84	77.72	78.53	77.00	78.57	76.92	79.10	78.15
EMA-UNet	80.15	79.82	80.32	78.08	79.84	78.37	80.71	79.86
ARL-UNet	80.48	80.20	80.37	78.94	80.04	78.87	81.03	80.30
$\mathcal{T}_1$: *Direct augmentation*								
U-Net	75.97	75.65	78.33	77.75	77.94	76.27	78.88	77.66
UCM-Net	78.34	78.53	79.12	77.39	79.07	77.81	79.61	78.78
EMA-UNet	79.54	79.84	80.04	78.49	79.79	79.54	80.71	79.94
ARL-UNet	80.17	80.11	80.37	79.41	80.25	79.59	81.37	80.23
$\mathcal{T}_2$: *Curriculum learning*								
U-Net	76.76	76.94	77.83	77.85	78.56	76.52	79.47	77.89
UCM-Net	79.03	78.80	79.57	77.99	79.50	78.30	79.66	78.88
EMA-UNet	80.11	80.02	80.41	79.41	80.53	79.63	81.22	80.28
ARL-UNet	80.51	80.25	80.98	80.48	81.04	80.16	81.79	80.94

B Detailed Results of PH2 Zero-shot Evaluation

This section provides the full numerical results for the PH2 zero-shot evaluation discussed in the main text. The main paper presents the results as compact visual summaries, while the tables in this section report the complete metric values for each source dataset and training regime.

Tables 16 and 17 show the zero-shot performance on PH2 when models are trained on ISIC 2017 and ISIC 2018, respectively. For each training strategy $\mathscr{T}_0$–$\mathscr{T}_2$, we report mIoU, Dice, Accuracy, Specificity, Sensitivity, and HD95. These metrics provide a detailed comparison of region overlap, pixel-level classification, and boundary localisation under cross-dataset transfer.

The detailed results support the observations in Section 5.8. ARL-UNet consistently achieves stronger performance in the principal segmentation metrics, especially mIoU, Dice, and HD95, across both ISIC source domains. The tables also show that curriculum learning $\mathscr{T}_2$ provides the best overall transfer performance for ARL-UNet, confirming that artifact-aware training improves robustness without degrading zero-shot generalisation to PH2.

Table 16. Zero-shot evaluation on PH2. Models are trained on ISIC 2017 and directly tested on PH2 without fine-tuning.

Model	mIoU	Dice	Accuracy	Specificity	Sensitivity	HD95
$\mathscr{T}_0$: *Clean-only*						
U-Net	80.76±0.58	89.30±0.38	93.29±0.33	95.00±0.33	88.81±0.56	18.86±0.87
UCM-Net	83.99±0.76	91.34±0.41	94.51±0.30	96.30±0.61	90.66±1.80	15.04±0.60
EMA-UNet	84.69±1.13	91.62±0.66	94.55±0.48	95.29±0.52	92.94±1.28	14.10±0.66
ARL-UNet	85.37±0.99	92.61±0.44	95.38±0.38	96.14±0.59	93.88±1.11	13.34±0.57
$\mathscr{T}_1$: *Direct augmentation*						
U-Net	80.86±0.55	90.20±0.34	93.33±0.29	95.67±0.31	88.90±0.44	18.44±0.72
UCM-Net	84.21±0.88	91.88±0.49	95.09±0.36	96.78±0.69	90.71±1.71	14.20±0.60
EMA-UNet	84.89±1.10	91.79±0.62	94.88±0.51	95.63±0.42	93.31±1.14	13.59±0.64
ARL-UNet	85.55±0.92	92.46±0.48	95.49±0.38	96.24±0.52	93.98±1.12	13.13±0.51
$\mathscr{T}_2$: *Curriculum learning*						
U-Net	80.86±0.48	90.29±0.34	93.77±0.40	96.18±0.33	89.27±0.35	18.19±0.68
UCM-Net	84.16±0.69	91.76±0.57	95.06±0.41	97.18±0.59	91.24±1.45	14.22±0.49
EMA-UNet	85.02±1.09	92.27±0.59	95.42±0.41	96.48±0.37	94.38±1.06	13.11±0.39
ARL-UNet	85.70±0.94	92.87±0.57	95.56±0.39	96.80±0.47	93.96±0.94	12.95±0.56

Table 17. Zero-shot evaluation on PH2. Models are trained on ISIC 2018 and directly tested on PH2 without fine-tuning.

Model	mIoU	Dice	Accuracy	Specificity	Sensitivity	HD95
$\mathscr{T}_0$: *Clean-only*						
U-Net	80.87±0.55	89.45±0.33	93.23±0.29	95.30±0.36	88.79±0.48	18.82±0.81
UCM-Net	83.96±0.83	91.28±0.49	94.43±0.24	96.28±0.67	90.53±1.93	14.98±0.65
EMA-UNet	84.76±1.16	91.75±0.68	94.61±0.43	95.37±0.49	93.01±1.25	13.87±0.66
ARL-UNet	85.32±0.91	92.44±0.43	95.50±0.32	96.24±0.51	93.92±1.10	13.10±0.53
$\mathscr{T}_1$: *Direct augmentation*						
U-Net	80.99±0.52	90.12±0.36	93.58±0.34	95.92±0.31	89.13±0.46	18.26±0.77
UCM-Net	84.28±0.81	91.78±0.44	94.89±0.28	96.94±0.66	90.86±1.64	14.32±0.49
EMA-UNet	85.06±1.00	91.99±0.64	95.04±0.48	95.90±0.51	93.66±1.18	13.24±0.53
ARL-UNet	85.64±0.88	92.54±0.45	95.61±0.32	96.52±0.54	93.99±1.08	13.01±0.49
$\mathscr{T}_2$: *Curriculum learning*						
U-Net	80.99±0.50	90.17±0.38	93.62±0.37	96.01±0.31	89.18±0.42	18.22±0.74
UCM-Net	84.35±0.75	91.96±0.45	95.07±0.29	97.11±0.51	91.34±1.60	14.15±0.40
EMA-UNet	85.00±1.00	92.25±0.66	95.34±0.48	96.31±0.51	94.12±1.18	13.24±0.47
ARL-UNet	85.86±0.84	92.74±0.47	95.73±0.36	96.89±0.50	94.24±0.97	12.86±0.47

C Visualization Results

We provide additional qualitative results to illustrate model behavior under different dermoscopic artifacts. Each figure compares predictions on the clean image and its corrupted variants, including hair, highlight, air bubble, and skin-line artifacts at two severity levels.

As shown in Fig. 11, artifact corruptions introduce different failure patterns. UCM-Net often produces smaller masks and misses peripheral lesion regions, while EMA-UNet tends to generate over-smoothed predictions. U-Net preserves larger regions in some cases, but its masks are less stable under severe highlights, bubbles, and skin-line artifacts.

In contrast, ARL-UNet better maintains the overall lesion shape across corruption types. It preserves the elongated lesion structure more reliably, especially when hair or skin-line artifacts overlap the lesion. These visual results are consistent with the quantitative findings, showing that ARL-UNet reduces artifact-induced mask distortion while retaining lesion boundaries.

D Curriculum Training Settings for Ablation

This section provides the detailed curriculum schedules used in the curriculum-training ablation study. In the main curriculum setting, the model is first trained on clean images and then exposed to artifact-corrupted samples with gradually increased corruption probability and severity. To analyze the contribution of different curriculum components, we consider two variants: curriculum training without clean warm-up and curriculum training without Severity-2 corruptions.

Table 18 shows the schedule for the curriculum variant without clean warm-up. In this setting, artifact-corrupted samples are introduced from the beginning of training. The first stage uses only mild corruptions with a sampling probability of 0.50, while the second stage increases the probability to 0.75 and includes both Severity 1 and Severity 2 corruptions. This variant is used to test whether early exposure to artifacts affects the learning of stable clean-domain lesion representations. Here, $\mathscr{A}$={Hair, highlight, air bubble, skin line}.

Table 19 presents the curriculum variant without Severity 2 corruptions. This setting keeps the clean warm-up stage but introduces only mild artifact corruptions afterward. It is designed to examine whether exposure to severe corruptions is necessary for improving robustness under stronger artifact perturbations.

Table 18. Curriculum schedule for the ablation variant without clean warm-up.

Stage	Epochs	Prob.	Artifacts	Sever
Stage 1	1–160	0.50	$\mathscr{A}$	1
Stage 2	161–320	0.75	$\mathscr{A}$	1, 2

Table 19. Curriculum schedule for the ablation variant without Severity-2 corruptions.

Stage	Epochs	Prob.	Artifacts	Sever
Stage 1	1–161	0	None	–
Stage 2	161–320	0.50	$\mathscr{A}$	1

Together, these schedules isolate two key factors in curriculum design: the clean warm-up phase and the inclusion of severe corruptions. The corresponding ablation results in the main text show that both components are important for achieving a favorable balance between clean-domain accuracy and corrupted-domain robustness.

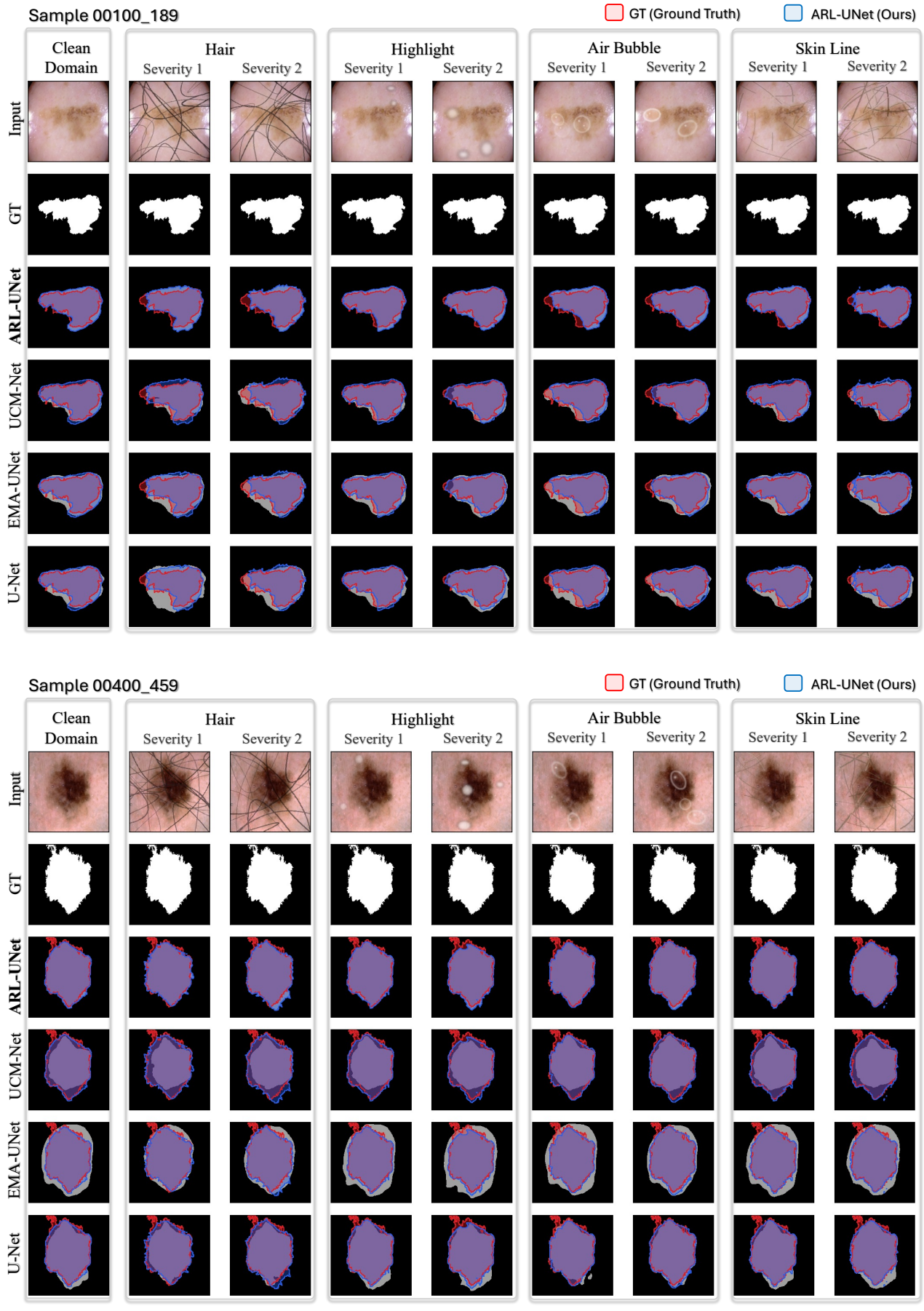


Figure 11. Qualitative comparison under different artifact corruptions.