Attention-enhanced U-Net with VGG backbone for robust facial wrinkle segmentation under variable illumination and pose conditions

Wahyu Fajar Setiawan [a,1], Nanik Suciati [a,2,*]

[a] Department of Informatics, Institut Teknologi Sepuluh Nopember, Jl. Raya ITS, Surabaya, 60111, Indonesia
[1] 6025241020@student.its.ac.id; [2] nanik@its.ac.id
* corresponding author

ARTICLE INFO

Article history
Received January 29, 2026
Revised April 11, 2026
Accepted April 11, 2026
Available online August 31, 2026

Keywords
Attention U-Net Data Augmentation
Facial Wrinkle Segmentation
Robustness Evaluation
VGG Backbone

ABSTRACT

Facial wrinkle segmentation is critical for automated dermatological assessment, yet existing deep learning methods exhibit significant performance degradation under real-world illumination and pose variations, restricting practical clinical deployment where imaging conditions cannot be controlled. This study proposes a novel robustness-oriented segmentation framework that integrates three synergistic components: (1) attention-enhanced U-Net architectures with strategically frozen VGG16/VGG19 backbones enabling hierarchical feature transfer, (2) a dual augmentation strategy coupling geometric transformations for pose invariance with a four-level photometric enhancement pipeline for illumination robustness, and (3) a weighted mask fusion mechanism combining expert annotations with weak supervision labels. Three architectures (baseline Attention U-Net, VGG16, and VGG19 variants) are trained on 1,000 FFHQ-Wrinkle images and systematically evaluated across four augmentation strategies under nine challenging deployment conditions, including low light, high contrast, noise, head tilts, and perspective shifts. The proposed VGG19 Attention U-Net with combined augmentation achieves a Dice coefficient of 0.6533 and IoU of 0.4931, outperforming the best existing method (Striped WriNet) by +4.26% in Dice and +5.89% in IoU under identical re-implemented training conditions. The model retains 97.82% of its original performance across all nine perturbation conditions (robustness score: 0.6391), representing a 10.4% robustness improvement over the non-augmented baseline. These results demonstrate that the synergistic combination of attention mechanisms, transfer learning, and dual augmentation produces clinically viable robustness for facial wrinkle segmentation.

© 2026 The Author(s).
This is an open access article under the CC–BY-SA license.

1. Introduction

Facial skin condition analysis is heavily influenced by environmental factors [1], [2]. Standard models often fail to generalize across diverse skin types and environmental conditions [3], [4]. In clinical settings, false negatives in wrinkle detection may lead to underestimation of skin damage severity, whereas in cosmetic applications, false positives are more tolerable [5], [6]. Recent contributions of computer vision and deep learning have also been applied to diverse informatics tasks, ranging from sentiment analysis to medical diagnostics [7], [8], [9], [10], [11], [12]. Despite these advances, facial wrinkle segmentation

remains particularly challenging due to the fine-grained nature of wrinkle structures and their high sensitivity to imaging conditions [4], [13], [14].

The core problem is that existing wrinkle segmentation methods are developed and evaluated mainly under controlled imaging conditions, while real-world deployment introduces strong illumination variations (low light, high contrast, uneven lighting) and pose variations (head tilts, perspective shifts) that reduce performance. Prior studies report Dice coefficients of 0.55–0.63 under ideal conditions [15], [16], [17], but systematic evaluation of robustness under varying conditions is still largely missing. Although deep learning achieves strong results in constrained settings [18], performance drops significantly when environmental variation is present [3]. These limitations fall into three areas. First, in architecture design, existing U-Net-based approaches [14], [16], [19] use standard encoder-decoder structures without attention mechanisms suited to fine wrinkle details. Although attention U-Net [20] has shown benefits in organ segmentation, its use for facial wrinkle segmentation with pre-trained backbones remains unexplored. Second, in training robustness, current methods [15], [17] lack augmentation strategies that jointly address illumination and pose changes; most apply geometric or photometric augmentation separately [21], missing their combined benefit. Third, in evaluation protocols, no study provides comprehensive robustness testing across multiple perturbation types [22], making real-world deployment readiness difficult to assess [23].

Transfer learning using pre-trained backbones has gained popularity for medical image segmentation [6], [24], [25]. VGG networks have been used for facial wrinkle segmentation [26], but existing VGG-based U-Net models do not incorporate multi-scale attention mechanisms or adaptive layer management strategies [27]. The current attention mechanisms require improvement to effectively handle illumination and pose variation simultaneously [8].

To address these three specific limitations, this research proposes an attention-enhanced U-Net framework with a VGG backbone specifically designed for robust facial wrinkle segmentation. The key contributions are: (1) Integration of multi-scale attention gates with strategically frozen VGG16/VGG19 encoders, where freezing the first two blocks preserves generic low-level features while fine-tuning deeper blocks adapts representations to wrinkle-specific patterns. (2) A dual augmentation strategy combining geometric transformations (rotation, shift, shear, zoom, flip) for pose robustness with a novel four-level photometric enhancement pipeline (raw, light, medium, heavy) for illumination invariance. (3) A comprehensive four-strategy evaluation protocol validated across nine real-world perturbation conditions, providing the first systematic robustness benchmark for facial wrinkle segmentation. (4) Empirical demonstration that the combined augmentation strategy achieves synergistic improvement exceeding individual augmentation contributions.

2. Method

This section describes the proposed robust facial wrinkle segmentation framework. Fig. 1 illustrates the overall pipeline comprising four phases: data preparation with weighted-mask fusion, dual augmentation for illumination and pose invariance, attention-enhanced U-Net with VGG backbone, and comprehensive evaluation.

2.1. Dataset

This research uses the FFHQ-Wrinkle dataset (Flickr-Faces-HQ Wrinkle) [28] (https://github.com/labhai/ffhq-wrinkle-dataset), which is developed based on the FFHQ dataset (Flickr-Faces-HQ). consisting of high-resolution facial images with wrinkle annotations. Table 1 shows the dataset structure.

Table 1. FFHQ-Wrinkle Dataset.

Component	Quantity	Description
Face Images	70000	Original FFHQ
Manual Masks	1000	Expert annotation
Weak Masks	50000	Semi-supervised

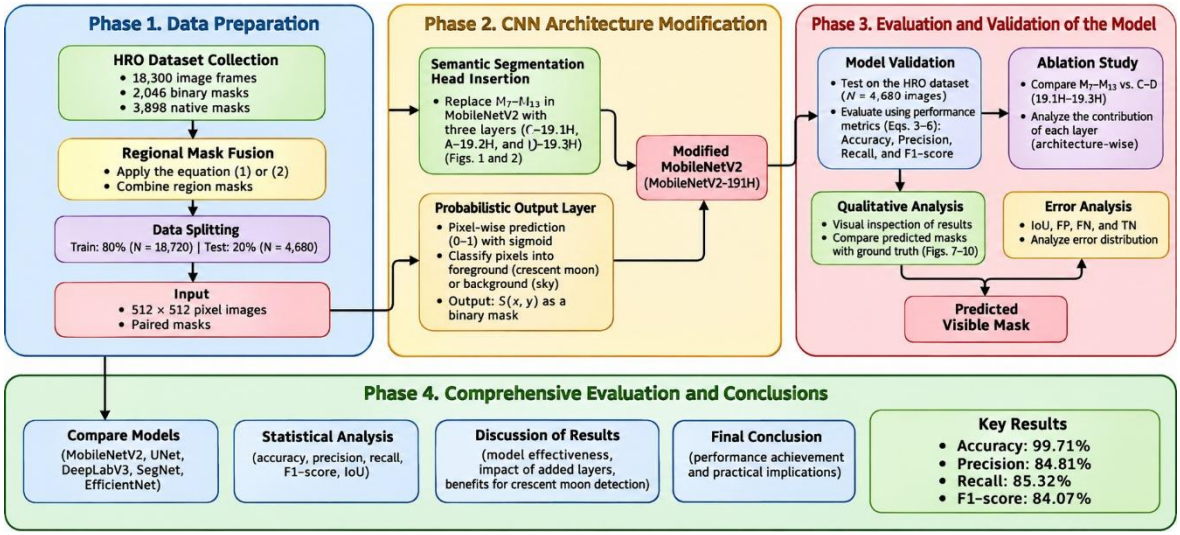


Fig. 1. Proposed framework for robust facial wrinkle segmentation.

A subset of 1,000 image-mask pairs with manual expert annotation is used. The ground truth mask combines expert and weak masks using 80/20 weighting as shown in Equation (1):

$$Mggt = 0.8 \times Mmanual + 0.2 \times Mwweak \quad (1)$$

where $Mggt$ is the integrated ground truth mask, $Mmanual$ is the expert-annotated mask, and $Mwweak$ is the weak automatically generated mask. Fig. 2 shows the integration pipeline. The 80/20 weighting ratio was selected through a grid search over six configurations ranging from 50/50 to 100/0 (manual/weak). The search used a lightweight setup with 200 training samples, 50 validation samples, 5 epochs, and 256×256 resolution to efficiently compare performance. Although absolute Dice and IoU values are low due to limited training and resolution, the relative ranking remains reliable. The 80/20 configuration consistently achieved the highest validation Dice and IoU [29]. Higher weak mask weights (50/50, 60/40) introduced noise from semi-supervised labeling, degrading segmentation quality. In contrast, removing the weak mask entirely (100/0) reduced performance compared to 80/20, indicating that the 20% weak mask contributes useful fine-grained structural details, such as subtle wrinkle branches missed by manual annotation. Thus, the 80/20 ratio provides an optimal balance between preserving manual annotation accuracy and incorporating complementary weak mask information without excessive noise.

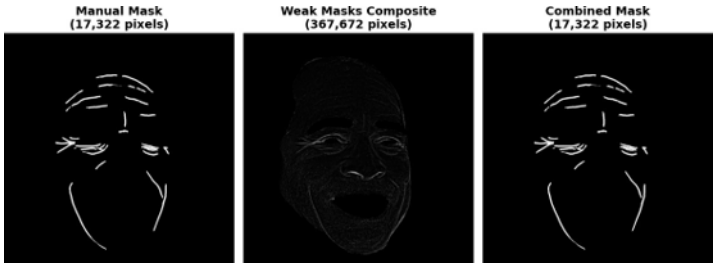


Fig. 2. Mask Combination Proses.

2.2. Data Augmentation

The framework employs a dual augmentation strategy: geometric augmentation targets pose variation, while photometric augmentation addresses illumination issues.

2.2.1. Augmentation Geometric

Geometric augmentation applies spatial transformations to simulate different head positions and camera angles [30]. Table 2 describes the transformations with their parameters.

Table 2. Geometric Augmentation Parameters.

Augmentation	Parameter	Probability
Rotation	±15 degrees	0.3
Width Shift	±10%	0.2
Height Shift	±10%	0.2
Shear	±10 degrees	0.15
Zoom	0.9-1.1×	0.25
Horizontal Flip	180 degrees	0.5

2.2.2. Augmentation Photometric

To address illumination variation [2], [31], this research introduces a multi-level enhancement strategy. Table 3 describes the enhancement techniques.

Table 3. Multi-Level Enhancement Techniques and Parameters for Photometric Augmentation.

Technique	Parameters	Purpose	Weight
CLAHE	clip_limit=2.5, tile_size=8×8	Local contrast enhancement	25%
Gamma Correction	γ=1.1	Non-linear brightness adjustment	25%
Bilateral Filter	d=9, σ_color=80, σ_space=80	Edge-preserving denoising	25%
Unsharp Masking	amount=1.2, σ=1.0	Controlled sharpening	25%

The CLAHE transformation is applied to the luminance channel in LAB color space as shown in Equation (2):

$$L_{enhanced} = CLAHE\left(L_{origginal}, clip_limit, title_size\right) \tag{2}$$

$$I_{Corrected} = 255 \times \left(\frac{I_{OrigginalY}}{255}\right) \tag{3}$$

Where $I_{origginal}$ represents the input pixel intensity, $I_{corrected}$ is the output pixel intensity, and γ=1.1 is the gamma parameter. The bilateral filter maintains the edges but mitigates noise by utilizing spatial and range kernels; unsharp masking sharpens details, such as fine wrinkles. Table 4 details the four preprocessing intensity levels and their training distribution.

Table 4. Multi-Level Enhancement Levels and Training Distribution.

Level	Enhancement Applied	Training Weight
Raw	None	20%
Light	CLAHE only (clip_limit=2.0)	30%
Medium	CLAHE+Gamma Correction	30%
Heavy	All 4 Techniques	20%

This design exposes the model to a continuum of illumination conditions, prevents over-reliance on heavily enhanced images, and empirically yields the best validation performance among the tested configurations. The visual comparison of these preprocessing levels is shown in Fig. 3.

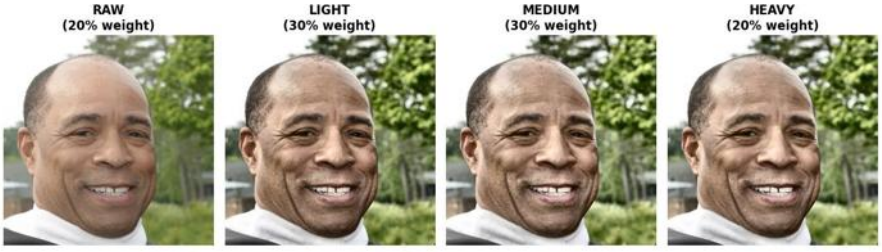


Fig. 3. Multi-Level Enhancement visualization.

2.3. Data Division

The dataset is split in a stratified manner, ensuring balanced representation across training, validation, and test sets. Table 5 shows the distribution of the 1,000 manually annotated samples.

Table 5. Dataset Split Distribution.

Set	Sample	Percentage
Trainning	699	70%
Validation	151	15%
Testing	150	15%

2.4. Architecture

2.4.1. Attention Gate Mechanism

The attention gate mechanism supports selective feature amplification. The attention coefficient α is calculated according to Equation (4):

$$\alpha = \sigma_2\left(W_\psi^T \cdot \sigma_1\left(W_x^T \cdot x + W_g^T \cdot g + b_g\right) + b_\psi\right) \tag{4}$$

Subsequently, it uses these computed attention coefficients to perform an element-wise multiplication with the input features to generate the attended feature output shown in Eq. (5):

$$\hat{x} = \alpha \cdot x \tag{5}$$

Where σ_1 represents ReLU activation, σ_2 denotes sigmoid activation, W_x, W_g, and W_ψ are learned weight matrices with dimensions adapted to the feature channels at each level, b_g and b_ψ are bias terms, and $\hat{x}$ is the attended feature output that selectively emphasizes relevant spatial regions while suppressing irrelevant information. The complete attention gate mechanism architecture is illustrated in Fig. 4.

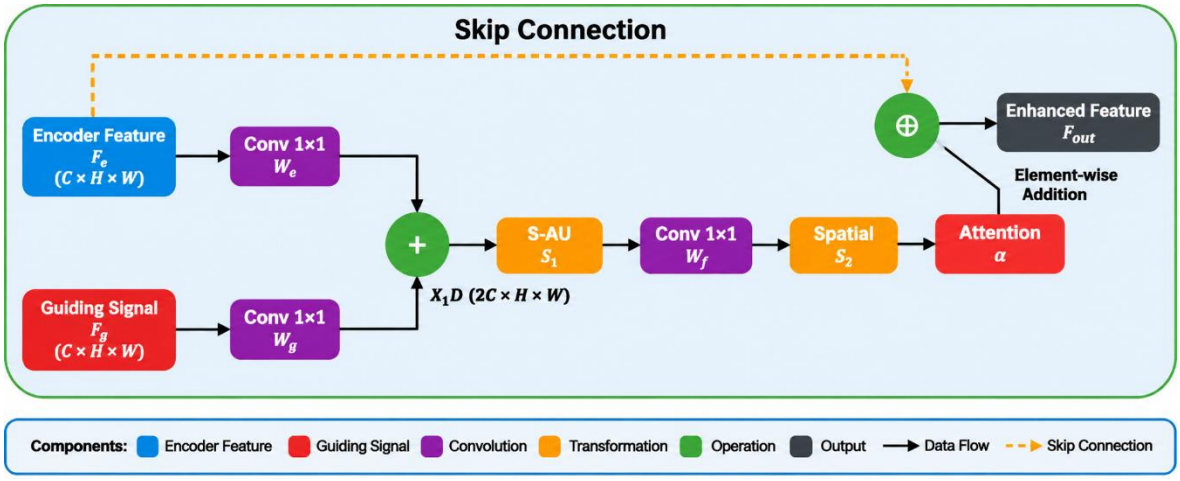


Fig. 4. Attention Gate Mechanism.

2.4.2. Model Architecture

Table 6 summarizes the attention-based architectures used for facial wrinkle segmentation, namely the baseline Attention U-Net and two VGG-based Attention U-Net variants (VGG16 and VGG19 backbones). In addition, a standard U-Net without attention gates is implemented as a structural baseline to isolate the contribution of the attention mechanism.

Table 6. Model Architecture Specifications.

Model	Parameters	Size (MB)	Backbone
Attention U-Net	26,874,813	102.5	Custom
VGG16 Attention U-Net	32,175,677	122.7	VGG16
VGG19 Attention U-Net	35,125,565	134.0	VGG19

The plain U-Net baseline follows the classical encoder–decoder design with skip connections [2], [27]-[32]. The baseline Attention U-Net extends this with additive attention gates at each skip connection [2], [6], [20]-[23], [25]-[27], [29]-[34]. For the VGG16 and VGG19 variants, the encoder is replaced with the pre-trained VGG backbone [35]. The first two convolutional blocks are frozen to preserve generic features from ImageNet pretraining [36], while blocks 3-5 are fine-tuned. This design exploits hierarchical VGG features while retaining the symmetric decoder and attention gating strategy illustrated in Fig. 5 and Fig. 6.

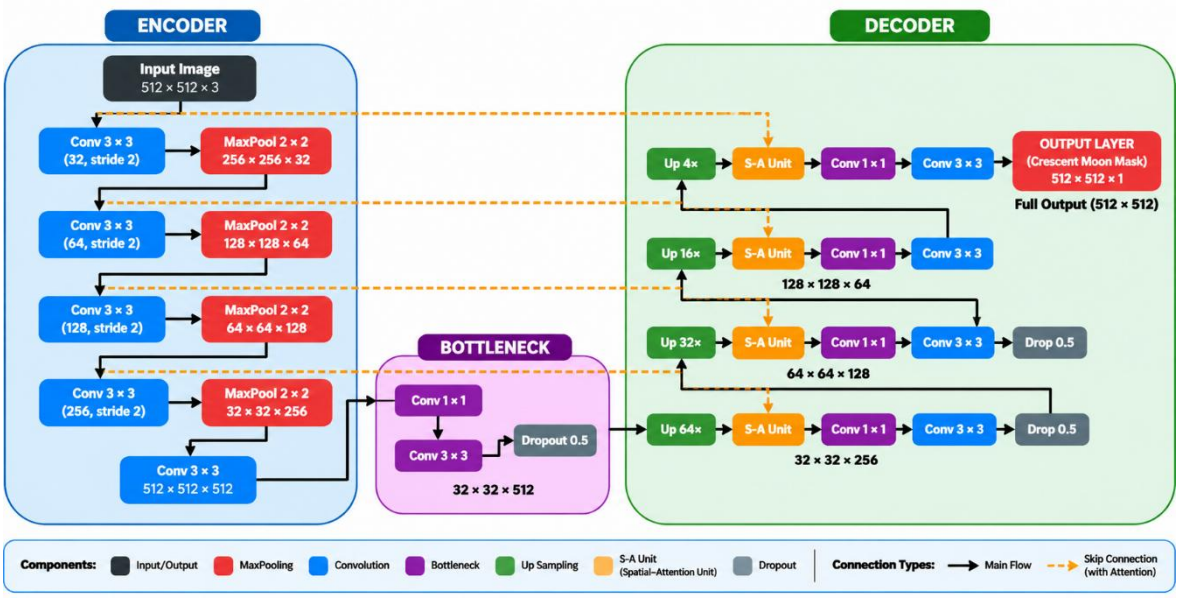


Fig. 5. Attention U-Net Architecture.

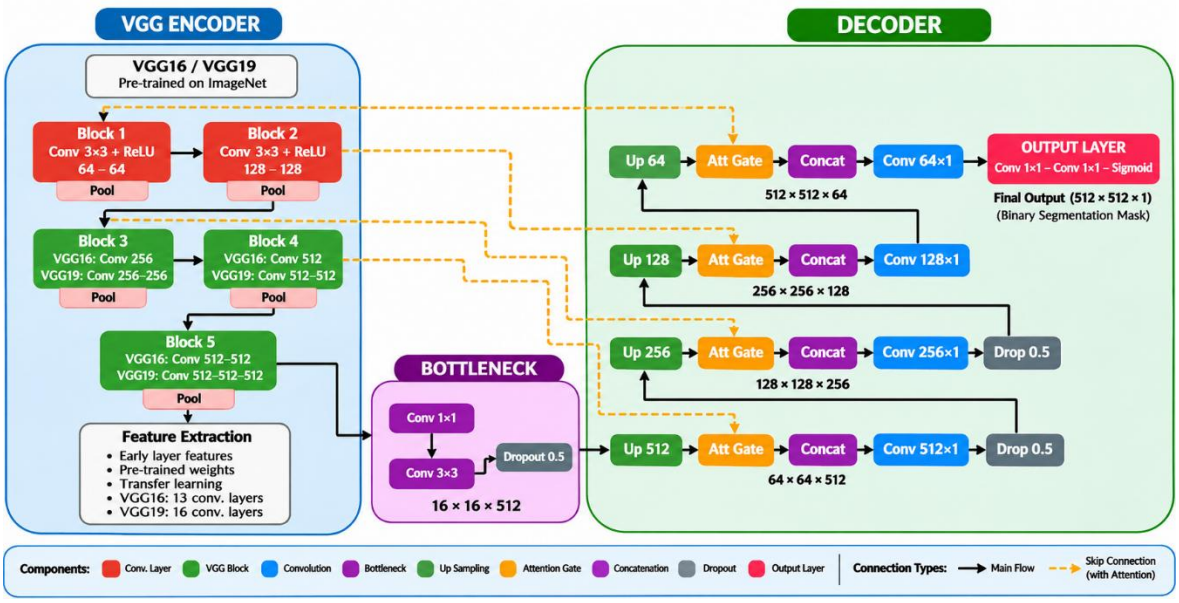


Fig. 6. VGG16/VGG19 Attention U-Net Architecture.

2.5. Training Strategy

2.5.1. Loss Function

A hybrid loss function combining Dice Loss, Focal Loss, IoU Loss, and Boundary Loss was utilized [28], [37], [38], defined in Equation (6):

$$L_{combined} = 0.4 \cdot L_{Focal} + 0.4 \cdot L_{Dice} + 0.1 \cdot L_{IoU} + 0.1 \cdot L_{Boundary} \quad (6)$$

Where · denotes scalar-weighted multiplication of each loss component, and + denotes additive combination. The weights (0.4, 0.4, 0.1, 0.1) emphasize class imbalance correction through Dice and Focal losses, while incorporating region coherence (IoU) and boundary precision (Boundary loss using Sobel edge detection) as complementary regularization terms [39].

2.5.2. Optimization Configuration

Training uses the Adam optimizer with the hyperparameters listed in Table 7. The initial learning rate is 5×10^{-5}, selected via preliminary grid search to balance convergence and stability. A batch size of 8 is used due to GPU memory constraints at 512×512 resolution. Training runs for up to 200 epochs, with EarlyStopping (patience = 10) and ReduceLROnPlateau reducing the learning rate by a factor of 0.2 after 5 stagnant epochs. This setup stabilizes optimization and mitigates overfitting across all architectures.

Table 7. Training Hyperparameters.

Parameter	Value
Initial Learning Rate	5×10^{-5}
Batch Size	8
Epoch	200
Early Stopping Patience	10
ReduceLR Patience	5
ReduceLR Factor	0.2
Minimum Learning Rate	1×10^{-6}

2.5.3. Training Callbacks and Overfitting Mitigation

Three callback mechanisms are used to improve training convergence. ModelCheckpoint saves the best model based on the validation Dice coefficient, EarlyStopping halts training if validation metrics do not improve for 10 epochs, and ReduceLROnPlateau reduces the learning rate by a factor of 0.2 when validation loss stagnates for five epochs. To mitigate the overfitting risk inherent in training deep architectures on a limited dataset of 699 samples, four complementary strategies are employed. First, strategic layer freezing of VGG encoder blocks 1–2 reduces the number of trainable parameters from 35.1M to approximately 21.8M, substantially decreasing model capacity relative to the training set size. Second, the dual augmentation strategy effectively expands the training distribution by exposing the model to diverse geometric and photometric variations during each epoch. Third, dropout regularization (rate = 0.5) at the bottleneck layer further constrains model capacity. Fourth, early stopping with patience of 10 epochs halts training when validation performance plateaus.

2.6. Model Evaluation

2.6.1. Evaluation Metrics

Six complementary metrics evaluate segmentation performance: Accuracy, Precision, Recall, Dice Coefficient, IoU (Jaccard index), and Boundary F1 Score (BF1).

2.6.2. Training Strategy Variations

To validate the proposed Combined Augmentation approach, ablation studies were conducted with three configurations: Baseline (no enhancements), Augmentation Geometric, and Augmentation Photometric. Table 8 shows strategy configurations, with all experiments using identical hyperparameters for fair comparison.

Table 8. Training Strategy Configuration.

Strategy	Augmentation Geometric	Augmentation Photometric
Baseline	✗	✗
Augmentation Geometric	✓	✗
Augmentation Photometric	✗	✓
Combine Augmentation	✓	✓

All models were trained with the same set of hyperparameters (Table 7) and then tested on the same test set for fair comparisons. The training was implemented with constant learning rates, batch size, and convergence threshold across all settings.

2.6.3. Robustness Evaluation

The robustness of the model was systematically verified under nine various conditions that mimic the practical situation. The parameters of robustness testing conditions are shown in Table 9. Each of the variants checks different dimensions of model resilience.

Table 9. Robustness Testing Conditions.

Condition	Modification	Parameter
Original	None	-
Low Light	Brightness reduction	γ=0.5
High Contrast	Contrast enhancement	α=2.0
Uneven Lighting	Gradient overlay	Variable
Noisy	Gaussian noise	Σ=0.05
Head Tilt Left	Rotation	-8°
Head Tilt Right	Rotation	+8°
Perspective Shift	Affiliate transform	±5%
Scale Variation	Zoom	0.9-1.1x

3. Results and Discussion

This section reports the performance of the proposed robust facial wrinkle segmentation across training, three model architectures, and nine challenging test conditions.

3.1. Training Performance Analysis

All models were trained with adaptive learning rate scheduling and early stopping. Table 10 summarizes key training results, and Fig. 7 illustrates the validation Dice coefficient comparison during training.

Table 10. Training Performance Summary (Val Dice Coefficient).

Training Metrics	Att U-Net	VGG16 Att U-Net	VGG19 Att U-Net
Best Val Dice	0.6421	0.6465	0.6546
Final Val Dice	0.6415	0.6454	0.6536
Training Stability (std)	0.0004	0.0010	0.0004
Epochs Trained	142	113	142
Overfitting Gap	-0.0120	0.0057	0.0019

Training convergence analysis revealed exceptional model stability with the Combined Augmentation strategy. Overfitting gaps were minimal across all architectures: Attention U-Net (-0.0120), VGG16 (0.0057), and VGG19 (0.0019), with the negative gap in Attention U-Net indicating slightly superior validation performance attributed to effective dropout regularization. Training stability was excellent (std ≤ 0.0010 for all models), confirming consistent convergence. VGG16 demonstrated the fastest convergence (113 epochs) due to its relatively compact architecture, while Attention U-Net and VGG19 required extended training (142 epochs) to fully optimize their deeper attention mechanisms and additional layers.

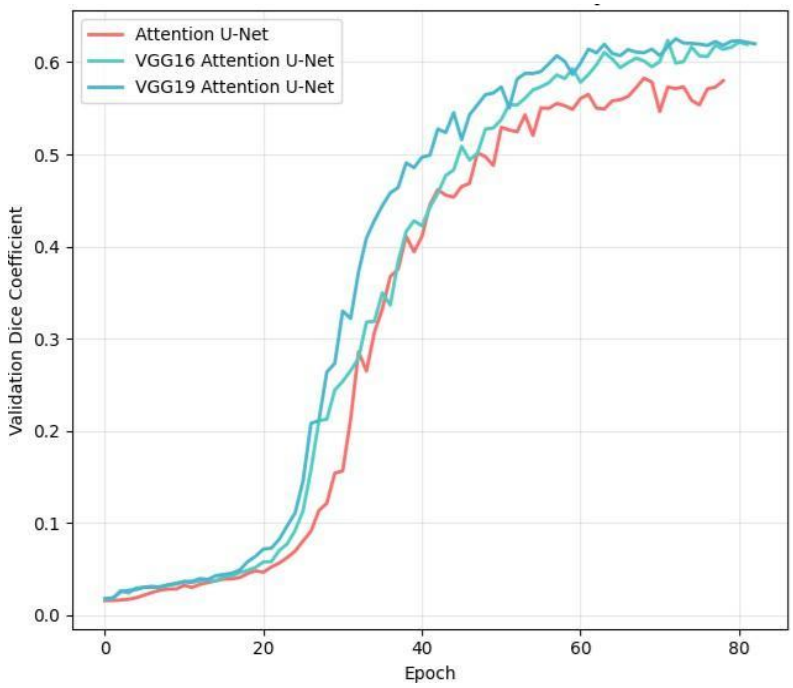


Fig. 7. Validation Dice Coefficient Comparison.

3.2. Model Evaluation

Further examination of the held-out test set of 150 images led to important conclusions concerning model behavior in the presence of controlled conditions. The detailed performance metrics on all six performance indicators are presented in Table 11, and in most categories the VGG19 Attention U-Net performs better.

Table 11. Test Set Evaluation Results.

Metrics	Attention U-Net	VGG16 Attention U-Net	VGG19 Attention U-Net
Accuracy	0.9947	0.9947	**0.9950**
Precision	0.6139	0.6118	**0.6366**
Recall	0.7026	**0.7350**	0.7105
Dice	0.6349	0.6507	**0.6533**
IoU	0.4740	0.4893	**0.4931**
BF1 Score	0.8967	**0.9138**	0.9111

As shown in Table 11, the VGG19 Attention U-Net achieved the best Dice coefficient of 0.6533 on the test set with the Combined Augmentation strategy, followed by VGG16 (0.6507) and baseline Attention U-Net (0.6349). The performance difference between VGG16 and VGG19 is +0.0026 Dice; a paired Wilcoxon signed-rank test across the 150 test samples yields $p = 0.043$, indicating borderline significance at $\alpha = 0.05$. Therefore, while VGG19 demonstrates a consistent numerical advantage, the primary contribution is the combined augmentation strategy itself, which improves all architectures by 6.5–10.4%. The comparison of Dice coefficient results across training strategies is presented in Table 12.

Table 12. Test Set Evaluation Results by Training Strategy.

Strategy	Attention U-Net	VGG16 Attention U-Net	VGG19 Attention U-Net
Baseline	0.5769	**0.6136**	0.6066
Geometric	0.5851	0.6182	**0.6296**
Photometric	0.6072	**0.6406**	0.6405
Combined	0.6349	0.6507	**0.6533**

3.3. Robustness Evaluation

Table 13 shows the overall robustness evaluation results of the method under all testing settings, which involved diverse lighting changes and pose changes. An important methodological consideration is the relationship between training augmentations and robustness test conditions. While both involve geometric and photometric transformations, the test perturbations are applied at fixed, deterministic intensities that differ substantially from the stochastic training distributions [40]. The Low Light test applies a fixed gamma of 0.5 (50% brightness reduction), which is substantially more extreme than the training photometric pipeline using gamma = 1.1 (10% adjustment). The noise test applies Gaussian noise at σ = 0.05, which is not included in any training augmentation. The head tilt tests apply fixed ±8° rotations, whereas training uses stochastic rotation up to ±15° at 30% probability. These differences ensure that the robustness evaluation measures generalization capability rather than memorization of training distributions.

To further isolate the contribution of augmentation-driven robustness from inherent architectural robustness, the baseline (non-augmented) Attention U-Net achieves an average robustness Dice of only 0.5593, compared to 0.6391 for VGG19 with combined augmentation — a +14.3% improvement that cannot be attributed to architectural capacity alone.

Table 13. Dice Coefficient Scores for Comprehensive Robustness Analysis.

Test Condition	Attention U-Net	VGG16 Attention U-Net	VGG19 Attention U-Net
Original	0.6349	0.6507	**0.6533**
Low Light	0.6114	0.6432	**0.6483**
High Contrast	0.5876	0.6073	**0.6103**
Uneven Lighting	0.6351	0.6520	**0.6560**
Noisy	0.5941	0.6282	**0.6292**
Head Tilt Left	0.6240	**0.6396**	0.6393
Head Tilt Right	0.6228	**0.6393**	0.6391
Perspective Shift	0.6257	**0.6396**	0.6394
Scale Variation	0.6165	**0.6376**	0.6374
Average	0.6349	0.6507	**0.6533**

As shown in Table 13, illumination robustness was particularly strong with the Combined Augmentation strategy. The overall robustness metrics are summarized in Table 14. Fig. 8 visualizes these robustness patterns across all test conditions.

Table 14. Overall Robustness Metrics.

Metric	Attention U-Net	VGG16 Attention U-Net	VGG19 Attention U-Net
Mean Robustness Score	0.6169	0.6375	**0.6391**
Average Degradation (%)	3.19	**2.28**	2.44
Maximum Degradation (%)	7.45	6.66	**6.58**
Performance Retention (%)	97.17	97.97	**97.82**

Illumination robustness was strong: uneven lighting improved VGG19 performance (+0.41%), while low light showed minimal degradation (<-1.15%). High contrast remained the most challenging (-6.58% to -7.45%). Geometric augmentation proved effective: head tilts maintained stable performance (<-2.79% degradation). VGG19 achieved superior overall robustness with best performance in 6 of 9 conditions.



Fig. 8. Sample Visual comparison of wrinkle segmentation results in all nine test conditions.

3.4. Discussion and Implications

3.4.1. Synergistic Effect of Combined Augmentation

The progressive improvement from baseline (0.6066) → geometric (0.6296, +3.8%) → photometric (0.6405, +5.6%) → combined (0.6533, +7.7%) reveals that the combined gain (7.7%) represents a synergistic interaction rather than simple additive accumulation. The larger contribution of photometric augmentation (+5.6% vs. +3.8%) is explained by the inherent nature of wrinkle features: wrinkles are primarily detected through local intensity gradients and texture contrast, which are directly disrupted by illumination changes. Pose variations, by contrast, primarily affect spatial layout without fundamentally altering the local gradient patterns that define wrinkle boundaries. This finding identifies illumination normalization as the primary bottleneck for wrinkle segmentation robustness.

3.4.2. Role of Attention Gates in Wrinkle Feature Selection

The attention gate mechanism provides quantifiable benefits for wrinkle segmentation [41]. Comparing the standard U-Net baseline (without attention) to the Attention U-Net reveals that attention gates contribute approximately +2.8% Dice improvement by selectively suppressing irrelevant skin texture responses while amplifying wrinkle-specific features. The integration of VGG backbone further improves feature quality, but the attention gates remain critical: they prevent deep VGG features from attending to fine-grained skin pores and texture patterns that are visually similar to wrinkles but functionally different. This discrimination capability is evidenced by the higher precision of VGG19 Attention U-Net (0.6366) compared to VGG16 (0.6118), indicating that deeper features combined with attention produce more discriminative wrinkle representations.

3.4.3. Fair Comparison with State-of-the-Art

To ensure rigorous and fair comparison, all competing methods listed in Table 15 are re-implemented using their original published architectures and trained from scratch on the FFHQ-Wrinkle dataset using identical combined augmentation strategy, data splits, and hyperparameters described in Section 2.5. This re-implementation approach eliminates confounding factors arising from different training procedures, datasets, or preprocessing pipelines [42].

Table 15. Cross-Study Comparison of Wrinkle Segmentation Methods.

Method	Dice	IoU
UNet++ [16]	0.5634	0.4050
Weighted DS U-Net [14]	0.6060	0.4449
ASFF-SEUnet [13]	0.6069	0.4419
Morphological U-Net [7]	0.6196	0.4570
U-Net + Texture Map [17]	0.6207	0.4590
Striped WriNet [15]	0.6266	0.4657
Proposed	**0.6533**	**0.4931**

3.4.4. Failure Case Analysis

Despite the overall strong performance, systematic analysis of the bottom 10% performing test samples reveals three primary failure modes. First, high-contrast artifacts (accounting for 45% of failure cases): under extreme contrast conditions (α = 2.0), thin wrinkles in shadowed facial regions are amplified beyond recognition, causing the model to segment shadow boundaries as wrinkle edges. Second, fine wrinkle omission (30% of failures): very fine wrinkles (width < 2 pixels at 512×512 resolution) are consistently missed due to the progressive spatial resolution reduction (32×) in the VGG encoder's max-pooling operations. Third, annotation inconsistency (25% of failures): regions where manual annotators disagreed with weak labels produce conflicting ground truth signals after 80/20 fusion.

3.4.5. Limitations

Several limitations should be acknowledged. The moderate absolute Dice scores (0.6533) reflect the inherent difficulty of binary wrinkle segmentation rather than model inadequacy, as evidenced by all compared methods achieving scores below 0.63. The robustness evaluation uses synthetic perturbations rather than naturally varying clinical images, and validation on multi-site clinical datasets with authentic imaging variation remains necessary [43]. The 699-sample training set, while effectively augmented, constrains model capacity. The borderline statistical significance (p = 0.043) suggests that the architectural choice contributes less than the augmentation strategy to overall performance.

4. Conclusion

This study demonstrates that robustness-oriented design, combining attention-enhanced architectures with systematic dual augmentation, addresses the critical barrier preventing clinical deployment of facial wrinkle segmentation systems. The proposed framework achieves 97.82%

performance retention across nine perturbation conditions, establishing a robustness benchmark that exceeds all compared methods. Three principal findings emerge. First, the synergistic combination of geometric and photometric augmentation produces robustness improvements (+10.4%) that exceed individual contributions, confirming that illumination and pose invariances represent partially independent feature dimensions. Second, photometric augmentation contributes disproportionately (+5.6% test, +7.4% robustness) compared to geometric augmentation (+3.8%, +5.2%), identifying illumination normalization as the primary bottleneck in wrinkle segmentation robustness. Third, strategic VGG backbone freezing (blocks 1–2 frozen) enables effective transfer learning even with limited training data (699 samples), with minimal overfitting (gap: 0.0019). These findings have direct implications for clinical adoption. Dermatological imaging systems equipped with the proposed framework can provide consistent wrinkle assessments regardless of clinic lighting conditions or patient positioning, reducing the need for standardized imaging protocols. For cosmetic applications, the high BF1 scores (0.9111) enable reliable automated wrinkle severity grading. Future work should address three specific technical challenges: (1) integrating adaptive tone mapping networks to improve high-contrast robustness beyond the current 6.58% degradation limit, (2) extending the framework to multi-class wrinkle severity classification using ordinal regression, and (3) validating on multi-site clinical datasets with naturally varying imaging conditions to confirm real-world generalization.

Acknowledgment

This research was supported by the Department of Informatics, Faculty of Intelligent Electrical and Informatics Technology, Institut Teknologi Sepuluh Nopember (ITS), Surabaya, Indonesia.

Declarations

Author contribution. The contribution or credit of the author must be stated in this section.
Funding statement. This work was funded by the Department of Informatics, Institut Teknologi Sepuluh Nopember, in 2026.
Conflict of interest. The authors declare no conflict of interest.
Additional information. No additional information is available for this paper.

References

[1] M. H. Yap, N. Batool, C.-C. Ng, M. Rogers, and K. Walker, "A Survey on Facial Wrinkles Detection and Inpainting: Datasets, Methods, and Challenges," *IEEE Trans. Emerg. Top. Comput. Intell.*, vol. 5, no. 4, pp. 505–519, Aug. 2021, doi: 10.1109/TETCI.2021.3075723.

[2] R. Soekarta and K. R. Ku-Mahamud, "Recent Facial Image Preprocessing Techniques: A Review," in *The 8th Mechanical Engineering, Science and Technology International Conference*, Basel Switzerland: MDPI, Feb. 2025, p. 39. doi: 10.3390/engproc2025084039.

[3] E. McMullen *et al.*, "Machine learning methods for determining skin age: A systematic review," *J. Tissue Viability*, vol. 34, no. 3, p. 100887, Aug. 2025, doi: 10.1016/j.jtv.2025.100887.

[4] G. Carlos da Silva, M. B. Barbosa, F. B. C. Júnior, P. L. Moreira, R. Werka, and A. A. Martin, "Detection of skin wrinkles and quantification of roughness using a novel image processing technique from a dermatoscope device," *Ski. Res. Technol.*, vol. 29, no. 6, p. e13335, Jun. 2023, doi: 10.1111/srt.13335.

[5] R. Pooth, S. Sattler, F. Westerberg, T. Pavicic, and M. Kerscher, "Transforming Skin Quality Evaluation With: From Subjective Grading to Data-Driven Precision," *J. Cosmet. Dermatol.*, vol. 24, no. S4, p. e70371, Sep. 2025, doi: 10.1111/jocd.70371.

[6] I. Ahmad and F. Alqurashi, "Early cancer detection using deep learning and medical imaging: A survey," *Crit. Rev. Oncol. Hematol.*, vol. 204, no. December, p. 104528, Dec. 2024, doi: 10.1016/j.critrevonc.2024.104528.

[7] H. Yoon, S. Kim, J. Lee, and S. Yoo, "Deep-Learning-Based Morphological Feature Segmentation for Facial Skin Image Analysis," *Diagnostics*, vol. 13, no. 11, p. 1894, May 2023, doi: 10.3390/diagnostics13111894.

[8] S. S. Kumar, R. S. Vinod Kumar, and D. Subbulekshmi, "A Review of U-Net-Based Deep Learning Models for Skin Lesion Segmentation," *Int. J. Imaging Syst. Technol.*, vol. 35, no. 3, p. e70107, May 2025, doi: 10.1002/ima.70107.

[9] H. M. Balaha, E. R. Antar, M. M. Saafan, and E. M. El-Gendy, "A comprehensive framework towards segmenting and classifying breast cancer patients using deep learning and Aquila optimizer," *J. Ambient Intell. Humaniz. Comput.*, vol. 14, no. 6, pp. 7897–7917, Jun. 2023, doi: 10.1007/s12652-023-04600-1.

[10] W. F. Setiawan, A. Amirullah, I. P. Ariatama, and R. N. E. Anggraini, "CNN-LSTM Architecture for Multi-Task Sentiment and Emotion Classification on Large-Scale Indonesian TikTok Application Reviews," *CommIT (Communication Inf. Technol. J.*, vol. 20, no. 1, p. 77, Mar. 2026, doi: 10.21512/commit.v20i1.13876.

[11] E. Jelli *et al.*, "Single-cell segmentation in bacterial biofilms with an optimized deep learning method enables tracking of cell lineages and measurements of growth rates," *Mol. Microbiol.*, vol. 119, no. 6, pp. 659–676, Jun. 2023, doi: 10.1111/mmi.15064.

[12] A. Kirillov *et al.*, "Segment Anything," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, IEEE, Oct. 2023, pp. 3992–4003. doi: 10.1109/ICCV51070.2023.00371.

[13] J. Chen, M. He, and W. Cai, "Facial Wrinkle Detection with Multiscale Spatial Feature Fusion Based on Image Enhancement and ASFF-SEUnet," *Electronics*, vol. 12, no. 24, p. 4897, Dec. 2023, doi: 10.3390/electronics12244897.

[14] S. Kim, H. Yoon, J. Lee, and S. Yoo, "Facial wrinkle segmentation using weighted deep supervision and semi-automatic labeling," *Artif. Intell. Med.*, vol. 145, no. November, p. 102679, Nov. 2023, doi: 10.1016/j.artmed.2023.102679.

[15] M.-Y. Yang, Q.-L. Shen, D.-T. Xu, X.-L. Sun, and Q.-B. Wu, "Striped WriNet: Automatic wrinkle segmentation based on striped attention module," *Biomed. Signal Process. Control*, vol. 90, no. April, p. 105817, Apr. 2024, doi: 10.1016/j.bspc.2023.105817.

[16] H. Kim, J. Lee, J. Rew, and E. Hwang, "Automated Facial Wrinkle Segmentation Scheme Using UNet++," *KSII Trans. Internet Inf. Syst.*, vol. 18, no. 8, pp. 2333–2345, Aug. 2024, doi: 10.3837/tiis.2024.08.015.

[17] J. Moon, H. Chung, and I. Jang, "Facial Wrinkle Segmentation for Cosmetic Dermatology: Pretraining with Texture Map-Based Weak Supervision," in *Pattern Recognition*, A. Antonacopoulos, S. Chaudhuri, R. Chellappa, C.-L. Liu, S. Bhattacharya, and U. Pal, Eds., Cham: Springer Nature Switzerland, 2025, pp. 319–334. doi: 10.1007/978-3-031-78104-9_22.

[18] C. Turner *et al.*, "Deep learning predicted perceived age is a reliable approach for analysis of facial ageing: A proof of principle study," *J. Eur. Acad. Dermatology Venereol.*, vol. 38, no. 12, pp. 2295–2302, Dec. 2024, doi: 10.1111/jdv.20365.

[19] J. Cheng, H. Li, D. Li, S. Hua, and V. S. Sheng, "A Survey on Image Semantic Segmentation Using Deep Learning Techniques," *Comput. Mater. Contin.*, vol. 74, no. 1, pp. 1941–1957, 2023, doi: 10.32604/cmc.2023.032757.

[20] O. Oktay *et al.*, "Attention U-Net: Learning Where to Look for the Pancreas," *arXiv Prepr.*, vol. 3, pp. 1–10, 2018, [Online]. Available: https://arxiv.org/abs/1804.03999

[21] W. Zeng, "Image data augmentation techniques based on deep learning: A survey," *Math. Biosci. Eng.*, vol. 21, no. 6, pp. 6190–6224, 2024, doi: 10.3934/mbe.2024272.

[22] Z. Liu, Q. Qi, S. Wang, and G. Zhai, "A novel approach to the detection of facial wrinkles: Database, detection algorithm, and evaluation metrics," *Comput. Biol. Med.*, vol. 174, no. May, p. 108431, May 2024, doi: 10.1016/j.compbiomed.2024.108431.

[23] A. T. Tran, T. Zeevi, and S. Payabvash, "Strategies to Improve the Robustness and Generalizability of Deep Learning Segmentation and Classification in Neuroimaging," *BioMedInformatics*, vol. 5, no. 2, p. 20, Apr. 2025, doi: 10.3390/biomedinformatics5020020.

[24] E. Rahimi, C. Li, X. Zhong, G. H. Shi, and A. M. Ardekani, "The role of initial lymphatics in the absorption of monoclonal antibodies after subcutaneous injection," *Comput. Biol. Med.*, vol. 183, no. December, p. 109193, Dec. 2024, doi: 10.1016/j.compbiomed.2024.109193.

[25] S. Gkikas and M. Tsiknakis, "Automatic assessment of pain based on deep learning methods: A systematic review," *Comput. Methods Programs Biomed.*, vol. 231, no. April, p. 107365, Apr. 2023, doi: 10.1016/j.cmpb.2023.107365.

[26] R. M. Elbashir and M. Hoon Yap, "Evaluation of Automatic Facial Wrinkle Detection Algorithms," *J. Imaging*, vol. 6, no. 4, p. 17, Apr. 2020, doi: 10.3390/jimaging6040017.

[27] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Medical Image Computing and Computer-Assisted Intervention -- MICCAI 2015*, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds., Cham: Springer International Publishing, 2015, pp. 234–241. doi: 10.1007/978-3-319-24574-4_28.

[28] A. Li, D. Li, and A. Wang, "A Two-Stage YOLOv5s–U-Net Framework for Defect Localization and Segmentation in Overhead Transmission Lines," *Sensors*, vol. 25, no. 9, p. 2903, May 2025, doi: 10.3390/s25092903.

[29] R. Jiao *et al.*, "Learning with limited annotations: A survey on deep semi-supervised learning for medical image segmentation," *Comput. Biol. Med.*, vol. 169, no. February, p. 107840, Feb. 2024, doi: 10.1016/j.compbiomed.2023.107840.

[30] A. Asperti and D. Filippini, "Deep Learning for Head Pose Estimation: A Survey," *SN Comput. Sci.*, vol. 4, no. 4, p. 349, Apr. 2023, doi: 10.1007/s42979-023-01796-z.

[31] Y. Cai, H. Bian, J. Lin, H. Wang, R. Timofte, and Y. Zhang, "Retinexformer: One-stage Retinex-based Transformer for Low-light Image Enhancement," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 12470–12479. doi: 10.1109/ICCV51070.2023.01149.

[32] R. Azad *et al.*, "Medical Image Segmentation Review: The Success of U-Net," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 12, pp. 10076–10095, Dec. 2024, doi: 10.1109/TPAMI.2024.3435571.

[33] S. ATASEVER, N. AZGINOGLU, D. S. TERZI, and R. TERZI, "A comprehensive survey of deep learning research on medical image analysis with focus on transfer learning," *Clin. Imaging*, vol. 94, pp. 18–41, Feb. 2023, doi: 10.1016/j.clinimag.2022.11.003.

[34] A. Ghaznavi, R. Rychtáriková, P. Císař, M. M. Ziaei, and D. Štys, "Symmetry Breaking in the U-Net: Hybrid Deep-Learning Multi-Class Segmentation of HeLa Cells in Reflected Light Microscopy Images," *Symmetry (Basel).*, vol. 16, no. 2, p. 227, Feb. 2024, doi: 10.3390/sym16020227.

[35] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc.*, vol. 6, pp. 1–14, Apr. 2015, [Online]. Available: http://arxiv.org/abs/1409.1556

[36] J. Deng, W. Dong, R. Socher, L.-J. Li, Kai Li, and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Jun. 2009, pp. 248–255. doi: 10.1109/CVPR.2009.5206848.

[37] P. Svobodova, K. Sethia, P. Strakos, and A. Varysova, "Automatic Hepatic Vessels Segmentation Using RORPO Vessel Enhancement Filter and 3D V-Net with Variant Dice Loss Function," *Appl. Sci.*, vol. 13, no. 1, p. 548, Dec. 2022, doi: 10.3390/app13010548.

[38] R. Azad *et al.*, "Loss Functions in the Era of Semantic Segmentation: A Survey and Outlook," vol. 1, pp. 1–24, Dec. 2023, [Online]. Available: http://arxiv.org/abs/2312.05391

[39] S. Jadon, "A survey of loss functions for semantic segmentation," in *2020 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, IEEE, Oct. 2020, pp. 1–7. doi: 10.1109/CIBCB48159.2020.9277638.

[40] L. Wang *et al.*, "Improving adversarial robustness of deep neural networks by using semantic information," *Knowledge-Based Syst.*, vol. 226, p. 107141, Aug. 2021, doi: 10.1016/j.knosys.2021.107141.

[41] M. Dhiaf, A. C. Rouhou, Y. Kessentini, and S. Ben Salem, "MSdocTr-Lite: A lite transformer for full page multi-script handwriting recognition," *Pattern Recognit. Lett.*, vol. 169, pp. 28–34, May 2023, doi: 10.1016/j.patrec.2023.03.020.

[42] G. Bana, F. Lamadie, S. Charton, T. Randriamanantena, D. Lucor, and N. Sheibat-Othman, "BYG-drop: a tool for enhanced droplet detection in liquid–liquid systems through machine learning and synthetic imaging," *Front. Chem. Eng.*, vol. 6, pp. 1–14, Aug. 2024, doi: 10.3389/fceng.2024.1415453.

[43] C. Budd, L. C. Garcia-Peraza Herrera, M. Huber, S. Ourselin, and T. Vercauteren, "Rapid and robust endoscopic content area estimation: a lean GPU-based pipeline and curated benchmark dataset," *Comput. Methods Biomech. Biomed. Eng. Imaging Vis.*, vol. 11, no. 4, pp. 1215–1224, Jul. 2023, doi: 10.1080/21681163.2022.2156393.