SAKARYA UNIVERSITY JOURNAL OF COMPUTER AND INFORMATION SCIENCES
http://saucis.sakarya.edu.tr/
e-ISSN: 2636-8129
Publisher: Sakarya University
Vol. 8, No. 1, 47-57, 2025
DOI: 10.35377/saucis...1517723
SAUCIS
SAKARYA UNIVERSITY
RESEARCH ARTICLE

Joint Detection and Removal of Specular Highlights using Vision Transformer with Multi-scale Patch Attention

Levent Karacan[1]

[1] Gaziantep University, Department of Computer Engineering, Gaziantep, Türkiye, ror.org/020vvc407

Corresponding author:
Levent Karacan, Department of Computer Engineering, Gaziantep University, Gaziantep, Türkiye
lkaracan@gantep.edu.tr

Article History:
Received: 17.07.2024
Revised: 19.12.2024
Accepted: 22.02.2025
Published Online: 27.03.2025

ABSTRACT
Specular highlights play a pivotal role in comprehending scenes within developed visual environment. Nevertheless, their presence can adversely affect the efficacy of solutions in various computer vision tasks. Current methodologies typically use Convolutional Neural Network (CNN)-based Unet architectures for specular highlight detection. However, CNNs exhibit limitations in capturing global contextual information, despite excelling in local context analysis. To utilize global context information, it is proposed a novel network architecture leveraging Vision Transformers (ViTs) to jointly detect and remove specular highlights for a given image. Developed model incorporates a multi-scale patch-based self-attention mechanism to effectively capture global context, alongside a CNN-based feed-forward network for local contextual cues. Experimental results with both quantitative and qualitative evaluations demonstrate that the proposed approach achieves state-of-the-art performance.

Keywords: Specular highlight detection, Specular highlight removal, Vision transformers, Convolutional neural networks

1. Introduction

Specular highlights are visual phenomena that appear on smooth and shiny surfaces. They are essential in helping the human visual system to interpret the environment by conveying information about light sources and surface materials. However, these highlights pose significant challenges in computer vision tasks such as image segmentation, text detection, object recognition, and scene understanding. They appear as bright and intense regions in images, as shown in Figure 1, and can degrade the performance of computer vision algorithms. Researchers have proposed various methods to detect and remove specular highlights to mitigate their impact on these tasks. Additionally, detecting these regions can be useful for light source detection and intrinsic image decomposition [1].

Early methods for detecting specular highlights [2] - [8] relied on the assumption that specular highlight regions contain the brightest pixels. These methods defined a threshold process to identify specular highlights. However, the bright pixel assumption does not hold for complex cases. In the context of removing specular highlights, traditional methods [9], [10], [11] - [14] predominantly depend on color values derived from the Dichromatic Reflection Model [2]. While previous highlight removal techniques have demonstrated success, they are constrained in their ability to handle large-scale removal tasks due to their reliance on prior information regarding material, color, or lighting conditions.

Learning-based highlight detection and removal methods [15] - [20] leverage the convolutional neural networks (CNNs) to train a highlight detection or removal model on carefully curated datasets. Recent research [18], [21] has shown that joint highlight detection and removal models produce more effective results than single detection or removal models. The CNN-based methods achieved effective results by leveraging the convolution operation to model local hierarchical image context. However, they are not well suited to model global context, which is important to model scene illumination.

Wu et al. [21] defined a ViT-based network to remove specular highlights, incorporating global context information into joint detection and removal models. However, they use a CNN-based network to detect specular highlights in this joint model, similar to previous methods.

Cite as: L. Karacan, "Joint detection and removal of specular highlights using vision transformer with multi-scale patch attention," *Sakarya University Journal of Computer and Information Sciences,* vol. 8, no. 1, pp. 47-57, 2025, doi: 10.35377/saucis...1517723

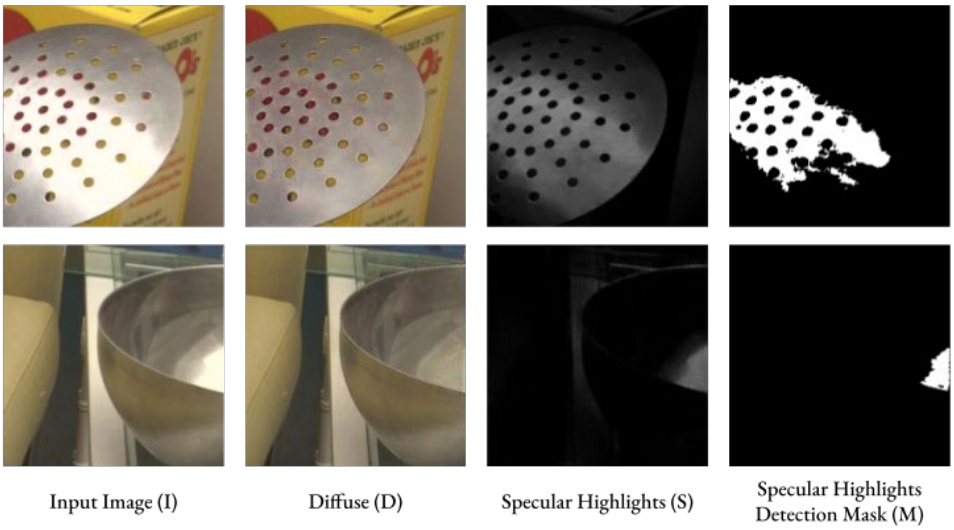

Figure 1. Specular Highlights Detection and Removal. The Proposed Model Detects and Remove Specular Highlights as Shown

The primary purpose of this study is to address the challenges posed by specular highlights in images, which can significantly degrade the performance of computer vision systems by affecting tasks like object detection, segmentation, and recognition. To tackle this issue, a novel network architecture is proposed that jointly detects and removes specular highlights in a given color image. At the core of the proposed network is a modified vision transformer, which employs a multi-scale patch-based self-attention mechanism to reduce scale dependency. The developed approach ensures a unified and efficient framework using a shared transformer-based backbone for detection and removal processes. Quantitative and qualitative results demonstrate that the proposed model achieves state-of-the-art performance on a standard dataset for both tasks. Additionally, an ablation study is conducted to analyze the multi-scale patch-based attention mechanism's effectiveness, further validating the developed method's robustness.

The manuscript is constructed as follows: Section 2 reviews the literature on specular highlight detection and removal. Section 3 provides a detailed description of the proposed method, encompassing its architectural framework, training protocol, and experimental configuration. Section 4 presents and critically examines the experimental results. Finally, Section 5 discusses the findings and potential future directions for the conclusion.

2. Related Works

Table 1 provides a detailed overview of significant studies on specular highlight detection and removal. It highlights the methods, key contributions, and tasks each work addresses. The initial approaches to highlight detection were largely based on color constancy models [2] - [4], which employed thresholding techniques to identify specular pixels. Incorporating the dark channel prior into their optimization scheme, Kim et al. [10] employed a different approach to that of Liu et al. [12], who estimated specular highlight reflection ratio and tuned saturation of the input image to remove highlights. Park et al. [5] proposed using two images in a least-squares regression scheme. These images are captured under distinct illumination conditions. One image was captured with specularities to be detected, while the other served as a reference, largely devoid of specularities, to generate a threshold map for image pixels. Building on the dichromatic reflection model [2], Meslouhi et al. [6] applied a specularity condition in the CIE-XYZ color space. Despite their successes, these methods were constrained by their reliance on specific assumptions and inability to cope with complex illumination scenarios or images with textured colors.

Learning-based approaches for specular highlight detection and removal offer more generalized solutions but necessitate diverse and extensive datasets. Fu et al. [17] introduced a real-world highlight dataset with annotated ground-truth masks, covering various material categories by providing different highlight shapes and appearances. They also trained a convolutional neural network that is used on this dataset for specular highlight detection. However, since this dataset lacks ground-truth diffuse images, it is unsuitable for training models aimed at removing specular highlights.

Table 1. Summary of Related Works on Specular Highlight Detection and Removal. The Table Lists the Reference, Authors, Title, Approach, Key Contributions, Publication, and Task Categories, Including Specular Highlight Detection, Removal, or Both Detection and Removal

Ref.	Authors	Title	Task	Approach	Key Contributions	Publication
[29]	Shen et al.	Chromaticity-based separation of reflection components in a single image	Specular Highlight Detection	Chromaticity-based method	Separates reflection components in single images using chromaticity information.	Pattern Recognit., 2008
[14]	Yamamoto et al.	Efficient improvement method for separation of reflection components based on an energy function	Specular Highlight Detection	Energy function-based optimization	A method to separate reflection components efficiently using an energy function.	IEEE ICIP, 2017
[8]	Zhang et al.	Improving shadow suppression for illumination robust face recognition	Specular Highlight Removing	Chromaticity-based method	Enhances shadow suppression for better face recognition under varying illumination conditions.	IEEE Trans. Pattern Anal. Mach. Intell., 2018
[7]	Li et al.	Specular reflection removal for endoscopic image sequences with adaptive-RPCA decomposition	Specular Highlight Removing	Adaptive-RPCA decomposition	Removal of specular reflections in endoscopic image sequences using adaptive Robust Principal Component Analysis (RPCA).	IEEE Trans. Med. Imaging, 2019
[30]	Lin et al.	Deep multi-class adversarial specularity removal	Specular Highlight Removing	Convolutional Neural Networks	A deep learning model for multi-class specularity removal with adversarial training.	SCIA 2019, Springer
[16]	Muhammad et al.	Spec-Net and Spec-CGAN: Deep learning models for specularity removal from faces	Specular Highlight Removing	Convolutional Neural Networks Generative Adversarial Networks	Proposes two deep learning models for removing specularity from facial images.	Image Vis. Comput., 2020
[17]	Fu et al.	Learning to Detect Specular Highlights from Real-world Images	Specular Highlight Detection	Convolutional Neural Networks	Highlights detection in real-world images using a deep learning approach.	ACM Multimedia, 2020
[18]	Fu et al.	A multi-task network for joint specular highlight detection and removal	Joint Specular Highlight Detection and Removing	Convolutional Neural Networks	Joint detection and removal of specular highlights using a multi-task network.	IEEE/CVF CVPR, 2021
[21]	Wu et al.	Joint specular highlight detection and removal in single images via Unet-Transformer	Joint Specular Highlight Detection and Removing	Vision Transformers	Proposes a joint detection and removal approach using a hybrid Unet-Transformer model.	Comput. Vis. Media, 2023

Shi et al. [23] proposed a CNN model comprised of encoder-decoder architectures to train on a large-scale object intrinsic database for the intrinsic image decomposition. Funke et al. [15] proposed a Generative Adversarial network (GAN)--based deep learning model capable of removing specular highlights from an endoscopic image. They trained this model on a small image patches dataset. The small patch images are extracted from the endoscopic video. Muhammad et al. [16]

proposed a facial specular highlight removal dataset and presented two alternative models, Spec-Net and Spec-CGAN. A real-world dataset with a pair of color images and ground-truth diffuse components was collected by Wu et al. [19] to train the specular highlights removal model. They proposed a GAN-based approach to remove specular highlights considering polarization theory. [20] introduced a specular highlight removal network comprising three stages in which the image is first decomposed into albedo, shading, and specular components. Subsequently, refinement and tone correction networks are employed to obtain decent specular-free images.

Recently, joint highlight detection and removal models [18] and [21] have been proposed using a multi-task network trained on a large dataset with corresponding diffuse and highlight components. Fu et al. [18] proposed the SHIQ dataset, a new large-scale specular highlight detection and removal dataset including ground-truth diffuse and specular highlights components of each image sample. They also introduced a multi-task CNN-based Unet architecture trained on the SHIQ dataset to detect and remove specular highlights jointly. Wu et al. [21] proposed a Swin Transformer-based [24] highlight removal model taking input image and specular highlight mask predicted by a CNN-based Unet detection model. They trained both removal and detection models jointly in an end-to-end manner. While Wu et al. demonstrate the efficacy of vision transformers for highlight removal tasks, it is asserted that a separate detection model is unnecessary. Hence, Fu et al. [18] used the same architecture backbone for detection and removal. Rather than relying solely on a CNN-based architecture like in Fu et al., a novel joint specular highlight detection and removal model that combines both CNN and ViT [25] architectures is proposed to leverage local and global context information. Moreover, while vanilla ViT employs the same patch scale across the different heads, the proposed MsPAT operates each scale in different heads.

3. Method

The Dichromatic Reflection Model (DFM) [2] defines a color image as the following composition of diffuse and specular highlight components:

$$I = D + S \tag{1}$$

The diffuse D, represents light uniformly scattered across an object's surface. In contrast, the specular highlight component S, accounts for the concentrated reflections of light from smooth surfaces. The model follows Fu et al. [18] to jointly detect and remove specular highlights for a given image. More clearly, it adopted the generalized version of DFM:

$$I = D + S \times M \tag{2}$$

where M denotes the pixel-wise binary mask indicating the location of the specular highlights. The point-wise multiplication of S and M provides to restrict the specular highlight removal in the masked regions.

Given a dataset of input RGB images I with ground-truth D, S, and M for each sample. Main goal is to train the proposed ViT-based model G to predict $\hat{D}$, $\hat{S}$, and $\hat{M}$ from a given single RGB image I:

$$[\hat{D}, \hat{S}, \hat{M}] = G(I) \tag{3}$$

where $\hat{D}$, $\hat{S}$, and $\hat{M}$ are predicted diffuse, specular highlights, and binary masks, respectively.

3.1. Network Architecture

The proposed model's overall structure is illustrated in Figure 2. The model includes an Encoder, multiple Multi-scale Patch Attention Transformers (MsPAT), and a Decoder. As shown in the Figure 2, an input RGB image $\boldsymbol{I} \in \boldsymbol{R^{H\times W\times 3}}$ with a height $\boldsymbol{H}$ and width $\boldsymbol{W}$is first passed through three convolutional blocks that reduce its spatial dimensions by half twice. The first convolutional block comprises a $\mathbf{7\times 7}$ convolution layer, batch normalization (BN), and ReLU activation layer, respectively. The next two convolutional blocks consist of $\mathbf{3\times 3}$ convolutions with stride $\mathbf{2}$, BN, and ReLU activation layers. Consequently, 256-dimensional feature representation $\boldsymbol{f} \in \boldsymbol{R^{H/4\times W/4\times 256}}$ is obtained, where the spatial dimensions are a quarter of the input image's dimensions. Subsequently, the feature representation is passed on to residually connected to multiple MsPAT modules. The transformed features $\boldsymbol{f}$ processed by the MsPAT modules are forwarded to the decoder network. The decoder network comprises two blocks, each containing two $\mathbf{4\times 4}$ transposed convolution layers with a stride of 2, followed by batch normalization (BN) and ReLU activation layers. These operations double the input feature maps' spatial resolution, restoring them to their original dimensions. Subsequently, the features are directed to three distinct prediction heads, each composed of $\mathbf{7\times 7}$convolutional layers. As shown in Figure 2, the predicted specular highlight mask highlight mask $\hat{\boldsymbol{M}}$ is concatenated with the feature maps before entering the specular highlight prediction head. Similarly, the predicted specular highlight mask $\hat{\boldsymbol{M}}$ and the specular highlight $\hat{\boldsymbol{S}}$ are concatenated with the feature maps before passing the residual prediction head.

3.2. Multi-scale Patch Attention Transformer (MsPAT)

The proposed Multi-scale Patch Attention Transformer (MsPAT) is illustrated in Figure 3. The MsPAT employs three basic steps: Embedding, Matching, and Attending.

3.2.1. Embedding

The feature maps $f \in R^{H/4\times W/4\times 256}$ gathered from the encoder or previous transformer are first embedded into query Q, key K, and value V features $Q,K,V \in R^{H/4\times W/4\times 256}$ using 1×1 convolution with a stride of 1. These features are then decomposed into multi-scale non-overlapping patches in parallel. As a result, it is obtained c-dimensional $(c = 256/s)$ patch embeddings at different scales s, $(q_i^s, k_i^s, v_i^s) \in R^{h^s\times w^s\times c}$, where i denotes the index of patches of height h^s and width w^s in each scale. In contrast to the vanilla Vision Transformer (ViT) [24], where patches are embedded into one-dimensional tokens, it is maintained the patch embeddings as two-dimensional. Additionally, while the vanilla ViT processes the same scale across different heads, the proposed MsPAT assigns each head to operate on different scales, enhancing its multi-scale processing capability.

3.2.2. Matching

In the matching step, the patch embeddings are initially flattened to the one-dimensional vector, and then patch similarities between query and key patches are calculated by dot product as follows:

$$S_{i,j}^s = \frac{q_i^s . k_j^s}{\sqrt{h^s \times w^s \times c}} \tag{4}$$

where i and j are indices within the N^s patches at each scale $(1 \le i,j \le N^s)$, h^s and w^s denote the height and width of the patch at scale s, and c denotes the dimension of the feature embedding. Based on this similarity, the weights in the attention map are calculated using the expression given below:

$$A_{i,j}^s = \frac{exp(S_{i,j}^s)}{\sum exp\,(S_{i,n}^s)} \tag{5}$$

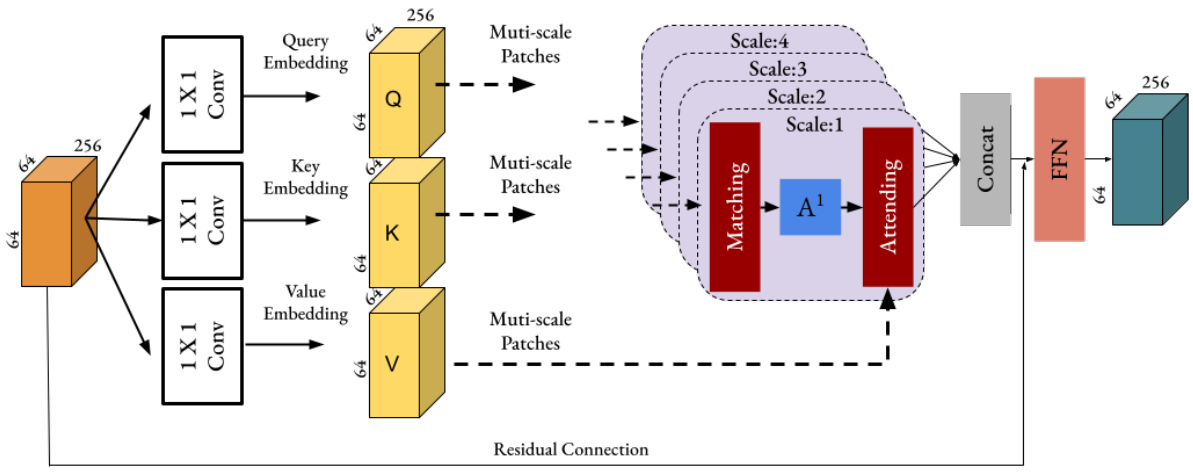

Figure 2. Proposed Multi-scale Patch Attention Transformer Module. This Module Applies Patch-Based Attention to Provide Global Connectivity. It Effectively Captures Global Dependencies by Incorporating Attention Mechanisms at Multiple Patch Scales

3.2.3. Attending

After the attention map $A_{i,j}^s$ is calculated for each scale s i.e. head, the output feature is obtained as the weighted summation of the value patches using the $A_{i,j}^s$:

$$f_i^s = \sum A_{i,j}^s \, v_j^s \tag{6}$$

The output feature patches f_i^S are reshaped to 2D and then recomposed to the input feature dimensions. Lastly, the output features $f^s \in R^{H/4\times W/4\times c}$ from each head are concatenated along the feature dimension:

$$f = [f^1, f^2, \ldots, f^s] \tag{7}$$

Lastly, the transformed feature $f \in R^{H/4\times W/4\times 256}$ is given to a feed-forward network and then passed to the subsequent transformer block or decoder. Note that there is a residual connection from the transformer block's input before passing to the feed-forward network.

As for the feed-forward network (FFN), it adopted the LocalViT [26], [27] that consists of convolutional layers of ReLU activation functions and squeeze-excitation (SE) module [28] to enrich the local context information in the transformer block.

3.3. Loss Function

To jointly detect and remove specular highlights, the proposed model is trained using a hybrid loss function comprising binary cross-entropy (BCE) loss, dice loss, and mean squared error (MSE) loss:

$$L = L_{BCE}(M, \hat{M}) + L_{Dice}(M, \hat{M}) + L_{MSE}^{S}(S, \hat{S}) + L_{MSE}^{D}(D, \hat{D}) \tag{8}$$

The BCE loss $\boldsymbol{L_{BCE}}$ and dice loss $\boldsymbol{L_{Dice}}$ is employed for the specular highlight detection as a binary mask segmentation. The BCE loss is commonly used for segmentation and mask prediction for specular highlight detection. The loss function between the predicted value $\hat{\boldsymbol{M}}_{\boldsymbol{p}}$ and the actual value $\boldsymbol{M_p}$ for each pixel $\boldsymbol{p}$ in the image is the sum of the binary cross-entropies for all pixels:

$$L_{BCE} = -\sum \left[M_p log(\hat{M}_p) + (1 - M_p) log(1 - \hat{M}_p) \right] \tag{9}$$

The regions of specular highlight are relatively small compared to the non-specular regions. To alleviate the effect of this imbalance, it is employed Dice loss [29], which measures the overlap between the predicted binary mask and the ground truth mask for each pixel $\boldsymbol{p}$:

$$L_{Dice} = 1 - \frac{1 + 2 \times \sum M_p \hat{M}_p}{1 + \sum (M_p + \hat{M}_p)} \tag{10}$$

To provide the model to remove specular highlight, it is defined the mean squared error (MSE) loss on the ground truth $\boldsymbol{S}$ and predicted $\hat{\boldsymbol{S}}$ specular highlights:

$$L_{MSE}^{S} = \frac{1}{N} \sum (S_p - \hat{S}_p)^2 \tag{11}$$

Similarly, it is employed mean squared error loss for the ground truth diffuse image $\boldsymbol{D}$ and the predicted diffuse image $\hat{\boldsymbol{D}}$ to predict the diffuse image as follows:

$$L_{MSE}^{D} = \frac{1}{N} \sum (D_p - \hat{D}_p)^2 \tag{12}$$

4. Experiments

4.1. Dataset and Implementation Details

It is trained the proposed model on the SHIQ dataset [18], which includes input images and their corresponding ground-truth binary masks M, specular highlight components S, and diffuse images D. The SHIQ dataset consists of 9825 training and 1000 test samples, with 200×200 image dimensions. It contains challenging examples of highly reflective objects like metal, plastic, and glass.

It is implemented with the proposed model network using PyTorch. The training was conducted over 60 epochs by using the Adam optimizer with a learning rate of 2×10^{-5}, and parameter settings of $\beta_1 = 0.5$ and $\beta_2 = 0.999$. During training, random horizontal flip data augmentation was applied.

4.2. Evaluation Metrics

It is used accuracy (Acc) and Balanced Error Rate (BER) metrics to evaluate the specular highlight detection results by following the previous works. Acc and BER can be defined as follows:

$$Acc = \frac{TP + TN}{TP + FP + TN + FN} \tag{13}$$

$$BER = \frac{1}{2}\left(\left(\frac{FP}{TN + FP}\right) + \left(\frac{FN}{TN + TP}\right)\right) \tag{14}$$

where TP, FP, TN, and FN refer to pixel-wise true positives, false positives, true negatives, and false negatives, respectively. Better detection results are indicated by a higher accuracy value and a lower BER value. Three commonly used metrics are utilized to evaluate the specular highlight removal performance: mean squared error (MSE), peak signal-to-noise ratio (PSNR), and structural similarity index (SSIM). Higher values of SSIM and PSNR, as well as lower MSE, imply improved performance.

4.2. Quantitative Results

It quantitatively compared specular highlight detection results with traditional methods NMF [7] and ATA [8], as well as deep learning-based methods SHDN [17], JSHDR [18], and Unet-Trans [21], using accuracy (Acc) and balanced error rate (BER) metrics commonly used for this task. Table 2 presents the average accuracy (Acc) and balanced error rate (BER) metric results obtained on the SHIQ test set. According to the results, the developed method produces the best results in terms of both accuracy and balanced error rate.

Table 2. Quantitative Specular Highlight Detection Results with Comparisons on the SHIQ Dataset. ↑ Denotes Better Performance with Higher Values, while ↓ Indicates Better Performance with Lower Values

Method	**Accuracy ↑**	**BER ↓**
NMF [7]	0,700	18,80
ATA [8]	0,710	24,40
SHDN [17]	0,910	6,180
JSHDR [18]	0,930	5,920
Unet-Trans [21]	0,970	5,920
Developed Model	**0,980**	**5,260**

Table 3. Quantitative Specular Highlight Removal Results with Comparisons on the SHIQ Dataset. ↑ Denotes Better Performance with Higher Values, while ↓ Indicates Better Performance with Lower Values

Method	**MSE ↓**	**SSIM ↑**	**PSNR ↑**
Shen et al. [29]	5,44	0,4596	19,2
Yamamoto et al. [14]	12,86	0,2945	9,15
Multi-class GAN [30]	0,3375	0,9103	24,49
Spec-CGAN [16]	0,6575	0,9154	23,53
JSHDR [18]	0,2525	0,9614	28,19
Trans-Unet [21]	0,1425	**0,9669**	29,85
Developed Model	**0,1353**	0,9523	**30,82**

For specular highlight removal, it is compared results with both traditional Shen et al. [29] and Yamamoto et al. [14] and state-of-the-art learning-based methods Multi-class GAN [30], Spec-CGAN [16], JSHDR [18], and Trans-Unet [21] as

shown in Table 2. The developed approach outperforms the others in mean squared error (MSE) and peak signal-to-noise ratio (PSNR) while delivering a highly competitive structural similarity index (SSIM) score in comparison to the latest methods.

4.3 Visual Results

Figure 4 presents the proposed method's visual specular highlight detection results and the corresponding ground truth. As shown, the proposed method provides highly accurate detection performance. Additionally, Figure 5 compares detection performance with the recent Unet-Trans method. It is worth noting that Unet-Trans employs a Unet architecture for specular highlight detection while utilizing a transformer architecture for specular highlight removal. In contrast, the method employs the proposed transformer-based approach for joint specular highlight detection and removal. As demonstrated in Figure 5, detection masks are more accurate than those produced by Unet-Trans.

Visual results of specular highlight removal are presented in Figure 6, alongside the corresponding input images and ground truth specular highlight-free images. The developed method demonstrates a high capability in accurately removing specular highlights and closely approximating the ground truth images. This highlights the robustness and effectiveness of the approach in preserving underlying image details while eliminating unwanted specular highlights. The method of comparison with the Unet-Trans approach is shown in Figure 7 to evaluate it further. The comparison reveals that it surpassed Unet-Trans in terms of visual quality or achieved competitive performance. The accuracy and consistency of the removal process are evident, showcasing cleaner and more natural-looking results. This can be attributed to innovative transformer-based architecture for detection and removal tasks, providing a cohesive and powerful framework for handling specular highlights.

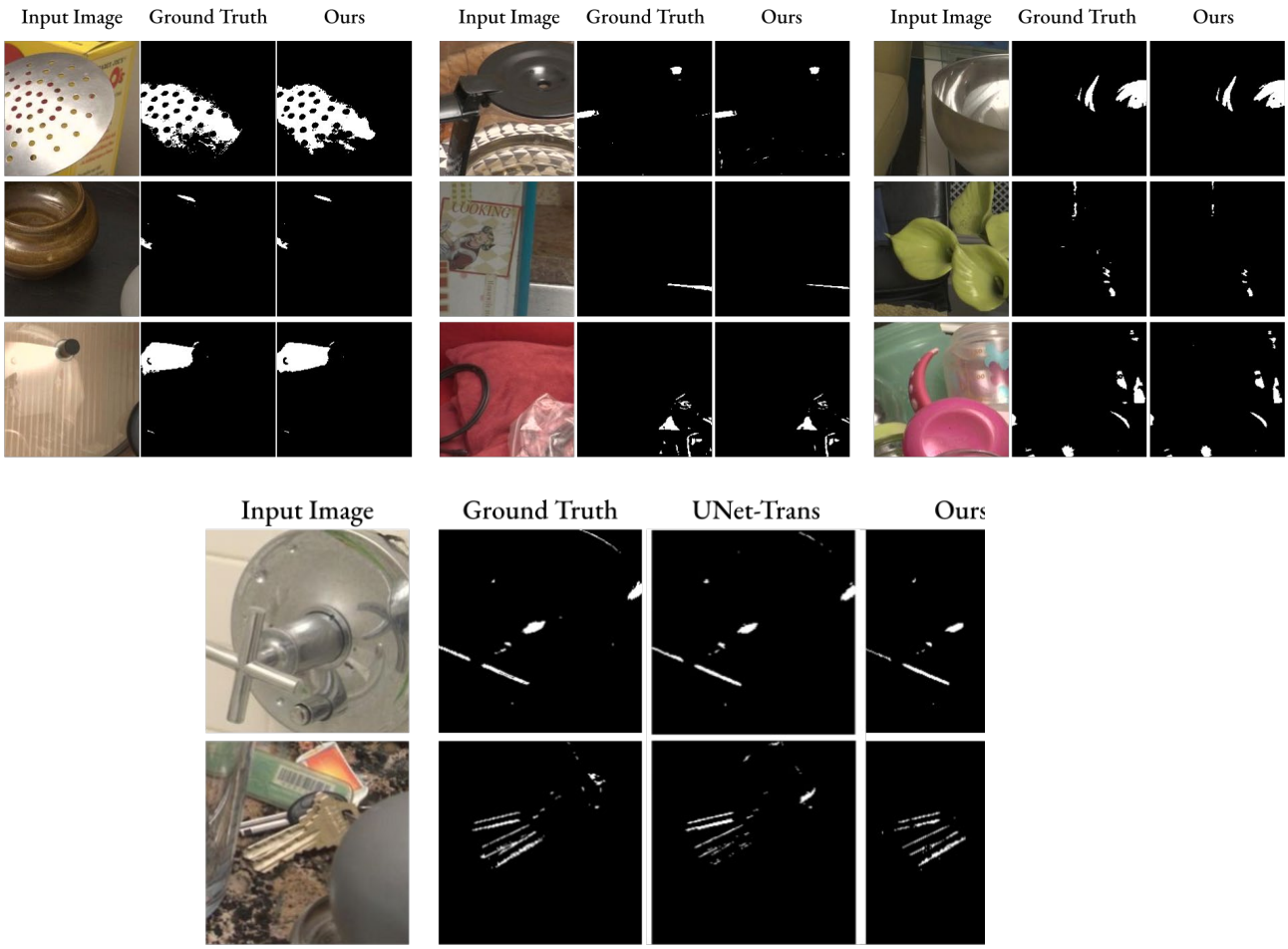

Figure 3. Visual Comparison of Specular Highlight Detection Results with A Recent Unet-Based Method Unet-Trans

4.4. Patch Scale Analysis on Attention

Table 4 analyzes the impact of the multi-scale approach on patch-based self-attention, where each head processes different scales. It compares multi-scale attention with various single-scale patch-based multi-head self-attention methods, where

each head processes the same scale. The observations indicate that different scales contribute to different metrics. Overall, multi-scale patch attention provides a balanced compromise across metrics. A trade-off between detection metrics, Accuracy (Acc), and Balanced Error Rate (BER) is also found. While the smallest scale provides the best BER score, it deteriorates the Accuracy. In this regard, the multi-scale approach offers a compromise, balancing these metrics. For specular highlight removal, the multi-scale approach achieves the best scores for MSE and PSNR, slightly trailing behind the largest single-scale attention for SSIM.

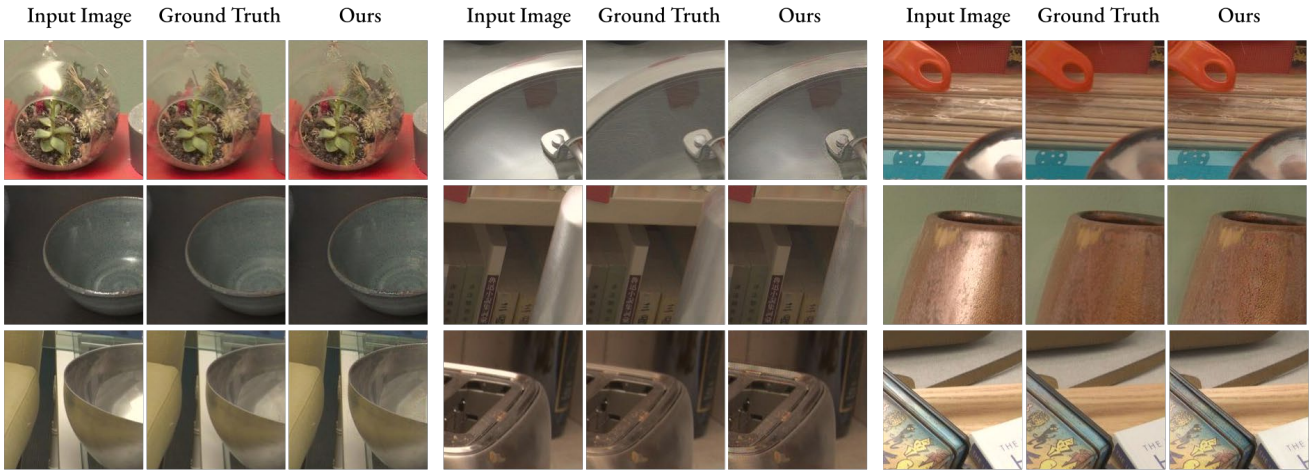

Figure 6. Visual Specular Highlight Removal Results

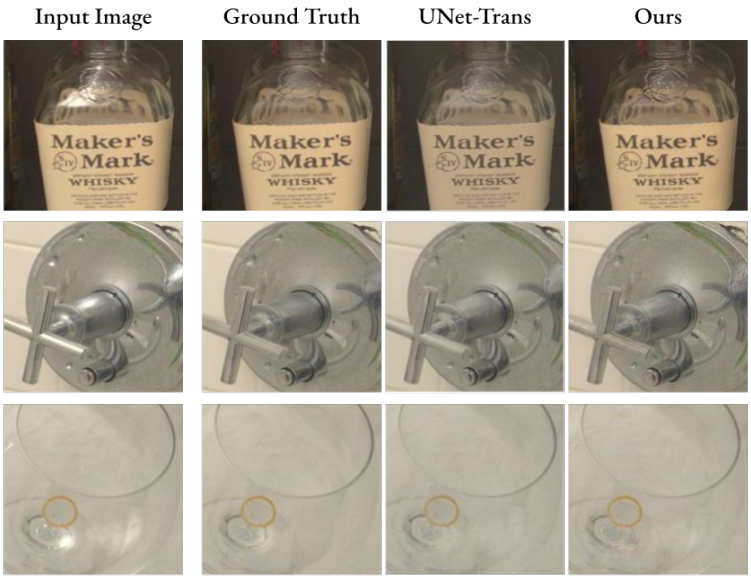

Figure 5. Visual Comparison of Specular Highlight Removal Results with A Recent Method Unet-Trans

Table 4. Patch Scale Analysis on Multi-Head Self-Attention. ↑ Denotes Better Performance with Higher Values, while ↓ Indicates Better Performance with Lower Values

Scale	**Acc ↑**	**BER ↓**	**MSE ↓**	**SSIM ↑**	**PSNR ↑**
Multi-scale	**0,9825**	5,26	**0,1353**	0,9523	**30,8200**
Patch Scale: (56 × 56)	0,9775	4,69	0,1624	**0,9541**	29,7693
Patch Scale: (28 × 28)	0,9813	5,41	0,1445	0,9513	30,5848
Patch Scale (14 × 14)	0,9794	5,32	0,159	0,9475	30,3146
Patch Scale: (7 × 7)	0,9725	**3,55**	0,2129	0,9434	28,9624

5. Conclusion

This study proposes a new ViT-based model architecture to jointly detect and remove specular highlights in a given image, defining a multi-scale patch self-attention in the transformer block. The proposed model demonstrated superior performance in specular highlight detection tasks, achieving the best accuracy and balanced error rate (BER) results. This approach enhances detection accuracy and removal quality. The experiments showed that multi-scale attention outperforms single-scale attention, particularly in MSE and PSNR metrics, while maintaining competitive SSIM scores. The multi-scale patch attention mechanism allows the model to process different scales within each attention head, leading to a comprehensive understanding of the image features.

Several enhancements can be explored for future work to improve the model's performance. Incorporating overlapping patches could provide better coverage and finer granularity in detection and removal processes. Exploring new transformer models with advanced architecture might yield additional performance gains. Utilizing pre-trained models on large-scale datasets could offer a strong initial foundation, reduce training time, and improve generalization. These directions offer promising opportunities to refine and enhance the effectiveness of specular highlight detection and removal.

References

[1] S. Jiddi, P. Robert, and E. Marchand, "Detecting specular reflections and cast shadows to estimate reflectance and illumination of dynamic indoor scenes," *IEEE Trans. Vis. Comput. Graph.*, vol. 28, no. 2, pp. 1249–1260, 2020.
[2] S. A. Shafer, "Using color to separate reflection components," *Color Res. Appl.*, vol. 10, no. 4, pp. 210–218, 1985.
[3] L. T. Maloney and B. A. Wandell, "Color constancy: a method for recovering surface spectral reflectance," in *Readings in Computer Vision*, Elsevier, 1987, pp. 293–297.
[4] Osadchy and Ramamoorthi, "Using specularities for recognition," in *IEEE ICCV*, IEEE, 2003, pp. 1512–1519.
[5] J. B. Park and A. C. Kak, "A truncated least squares approach to detecting specular highlights in color images," in *IEEE ICRA*, IEEE, 2003, pp. 1397–1403.
[6] O. El Meslouhi, M. Kardouchi, H. Allali, T. Gadi, and Y. A. Benkaddour, "Automatic detection and inpainting of specular reflections for colposcopic images," *Cent. Eur. J. Comput. Sci.*, vol. 1, pp. 341–354, 2011.
[7] R. Li, J. Pan, Y. Si, B. Yan, Y. Hu, and H. Qin, "Specular reflections removal for endoscopic image sequences with adaptive-RPCA decomposition," *IEEE Trans. Med. Imaging*, vol. 39, no. 2, pp. 328–340, 2019.
[8] W. Zhang, X. Zhao, J.-M. Morvan, and L. Chen, "Improving shadow suppression for illumination robust face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 3, pp. 611–624, 2018.
[9] Q. Yang, S. Wang, and N. Ahuja, "Real-time specular highlight removal using bilateral filtering," in *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part IV 11*, Springer, 2010, pp. 87–100.
[10] H. Kim, H. Jin, S. Hadap, and I. Kweon, "Specular reflection separation using dark channel prior," in *IEEE CVPR*, 2013, pp. 1460–1467.
[11] Q. Yang, J. Tang, and N. Ahuja, "Efficient and robust specular highlight removal," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 6, pp. 1304–1311, 2014.
[12] Y. Liu, Z. Yuan, N. Zheng, and Y. Wu, "Saturation-preserving specular reflection separation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3725–3733.
[13] J. Suo, D. An, X. Ji, H. Wang, and Q. Dai, "Fast and high quality highlight removal from a single image," *IEEE Trans. Image Process.*, vol. 25, no. 11, pp. 5441–5454, 2016.
[14] T. Yamamoto, T. Kitajima, and R. Kawauchi, "Efficient improvement method for separation of reflection components based on an energy function," in *2017 IEEE international conference on image processing (ICIP)*, IEEE, 2017, pp. 4222–4226.
[15] I. Funke, S. Bodenstedt, C. Riediger, J. Weitz, and S. Speidel, "Generative adversarial networks for specular highlight removal in endoscopic images," in *Medical Imaging 2018: Image-Guided Procedures, Robotic Interventions, and Modeling*, SPIE, 2018, pp. 8–16.
[16] S. Muhammad, M. N. Dailey, M. Farooq, M. F. Majeed, and M. Ekpanyapong, "Spec-Net and Spec-CGAN: Deep learning models for specularity removal from faces," *Image Vis. Comput.*, vol. 93, p. 103823, 2020.
[17] G. Fu, Q. Zhang, Q. Lin, L. Zhu, and C. Xiao, "Learning to Detect Specular Highlights from Real-world Images," in *ACM Multimedia*, 2020, pp. 1873–1881.
[18] G. Fu, Q. Zhang, L. Zhu, P. Li, and C. Xiao, "A multi-task network for joint specular highlight detection and removal," in *IEEE/CVF CVPR*, 2021, pp. 7752–7761.
[19] Z. Wu, C. Zhuang, J. Shi, J. Xiao, and J. Guo, "Deep specular highlight removal for single real-world image," in *SIGGRAPH Asia 2020 Posters*, 2020, pp. 1–2.
[20] G. Fu, Q. Zhang, L. Zhu, C. Xiao, and P. Li, "Towards High-Quality Specular Highlight Removal by Leveraging Large-Scale Synthetic Data," in *IEEE/CVF ICCV*, 2023, pp. 12857–12865.

[21] Z. Wu, J. Guo, C. Zhuang, J. Xiao, D.-M. Yan, and X. Zhang, "Joint specular highlight detection and removal in single images via Unet-Transformer," *Comput. Vis. Media*, vol. 9, no. 1, pp. 141–154, 2023.
[22] J. Shi, Y. Dong, H. Su, and S. X. Yu, "Learning non-lambertian object intrinsics across shapenet categories," in *IEEE CVPR*, 2017, pp. 1685–1694.
[23] Z. Liu *et al.*, "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, QC, Canada: IEEE, Oct. 2021, pp. 9992–10002. doi: 10.1109/ICCV48922.2021.00986.
[24] A. Dosovitskiy *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in *International Conference on Learning Representations*, 2020.
[25] Y. Li, K. Zhang, J. Cao, R. Timofte, and L. Van Gool, "Localvit: Bringing locality to vision transformers," *ArXiv Prepr. ArXiv210405707*, 2021.
[26] L. Karacan, "Multi-image transformer for multi-focus image fusion," *Signal Process. Image Commun.*, vol. 119, p. 117058, 2023.
[27] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *IEEE CVPR*, 2018, pp. 7132–7141.
[28] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. Jorge Cardoso, "Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 14, Proceedings 3*, Springer, 2017, pp. 240–248.
[29] H.-L. Shen, H.-G. Zhang, S.-J. Shao, and J. H. Xin, "Chromaticity-based separation of reflection components in a single image," *Pattern Recognit.*, vol. 41, no. 8, pp. 2461–2469, 2008.
[30] J. Lin, M. El Amine Seddik, M. Tamaazousti, Y. Tamaazousti, and A. Bartoli, "Deep multi-class adversarial specularity removal," in *Image Analysis: 21st Scandinavian Conference, SCIA 2019, Norrköping, Sweden, June 11–13, 2019, Proceedings 21*, Springer, 2019, pp. 3–15.

Author(s) Contributions
The article was authored and prepared solely by the author. The author was responsible for all aspects of the research, including conceptualization, methodology, implementation, analysis, and writing of the manuscript.

Conflict of Interest Notice
The author declares that there is no conflict of interest regarding the publication of this paper.

Ethical Approval and Informed Consent
It is declared that during the preparation process of this study, scientific and ethical principles were followed, and all the studies benefited from are stated in the bibliography.

Plagiarism Statement
This article has been scanned by iThenticate ™.