

IRFusionFormer: Enhancing Pavement Crack Segmentation with RGB-T Fusion and Topological-Based Loss

Ruiqiang Xiao¹
rxiaoad@connect.ust.hk
Xiaohu Chen²
xiaohu@seu.edu.cn

¹ Robotics and Autonomous Systems
Hong Kong University of Science and
Technology (Guangzhou)
Guangzhou, China

² Intelligent Transportation System
Research Center
Southeast University
Nanjing, China

Abstract

Crack segmentation is a critical task in civil engineering applications, particularly for assessing pavement integrity and ensuring the durability of transportation infrastructure. While deep learning models have advanced RGB-based segmentation, their performance degrades under adverse conditions like low illumination and motion blur. Thermal imaging offers complementary information by capturing emitted radiation, enabling better differentiation of cracks in challenging environments. By integrating information from both RGB and thermal images, RGB-T pavement crack segmentation has demonstrated significant advantages in complex real-world environments such as adverse weather conditions. However, research in this area remains relatively limited, and current RGB-T crack segmentation methods do not fully and efficiently leverage the complementary relationships between different modalities during multi-level information interaction. To address this problem, we propose IRFusionFormer, a novel model for crack segmentation that effectively integrates RGB and thermal data. We introduce the Efficient RGB-T Cross Fusion Module (EGTCF) to capture extensive multi-scale relationships and long-range dependencies between modalities without incurring high computational costs. Additionally, we develop the Interaction-Hybrid-Branch-Supervision (IHBS) framework, which enhances modality interaction by distributing fused features across branches and enabling joint supervision. To preserve the topological structure of cracks, we propose a novel topology-based loss function that maintains connectivity and structural integrity during training. Our method achieves state-of-the-art results, surpassing existing approaches with a Dice score of 90.01% and an Intersection over Union (IoU) of 81.83%. These advancements address critical challenges in pavement crack segmentation by improving robustness and accuracy under varying environmental conditions. For access to the codes, data, and models pertinent to this study, please visit: [Code](#).

1 Introduction

Crack segmentation, which involves assigning binary labels—crack or background—to individual pixels in an image, has attracted growing attention in various civil engineering scenarios such as buildings[9], bridges[20], tunnels[22], and pavement[6] inspections. Among these, pavement crack segmentation is particularly crucial for assessing road quality and maintaining the longevity. Traditional image processing techniques like thresholding[10] and edge detection[13] have been used to segment pavement cracks in RGB images by exploiting grayscale value differences. However, these methods often suffer from low accuracy and lack robustness due to their reliance on handcrafted settings and sensitivity to varying imaging conditions. The advent of deep learning, particularly convolutional neural networks (CNNs), has revolutionized semantic segmentation tasks. Methods like FCN[14], U-Net[15], UNet++[24], and DeepLab V3+[1] have achieved impressive results on large-scale RGB image semantic segmentation task. Consequently, researchers have adapted these paradigms for pavement crack segmentation. For instance, DeepCrack[6] integrates multi-scale convolutional features from hierarchical stages to capture fine-grained line structures, leading to improved crack detection.

Despite these advancements, RGB-based pavement crack segmentation methods degrade rapidly under challenging conditions such as rainy or hazy weather and low illumination[5]. During pavement inspections, cameras mounted on fast-moving vehicles like inspection trucks or drones struggle to maintain stable footage, minimize motion blur, and ensure adequate illumination simultaneously. Consequently, underexposed and blurred RGB images yield unfavorable segmentation results compared to normal conditions, as shown in the **Figure 1 (a)**. Additionally, semantic interferences resembling cracks—such as scattered binding particles, tree shadows, water marks, and patch repairs—introduce additional complexities, leading to false detections and reduced reliability[25]. In contrast, thermal images rely on emitted radiation from objects and can capture stable images under complex conditions, albeit at lower resolutions. This capability allows for better differentiation between crack foregrounds and backgrounds in challenging environments[8, 9]. Therefore, RGB-Thermal (RGB-T) crack segmentation has gained increased attention[9]. By utilizing both RGB and thermal data and efficiently fusing their complementary information, the performance and stability of crack segmentation can be enhanced across various real-world scenarios. Leveraging the rich semantic information from both modalities is essential to overcome existing challenges[20]. Although some studies have employed both infrared and RGB images, a unified benchmark for pavement crack segmentation across these modalities is still lacking. To address this gap, we compiled existing crack segmentation approaches and hybrid methods integrating infrared and RGB data, constructing a novel and comprehensive benchmark for asphalt pavement crack segmentation. Our review identified several critical limitations in current methodologies.

One of the key challenges in RGB-T crack segmentation is the efficient fusion of cross-modal features. Traditional convolution-based attention mechanisms—such as channel, spatial, and hybrid attention—are limited by their local receptive fields, restricting their ability to model global contextual relationships[2]. While self-attention mechanisms can capture long-range dependencies, their high computational and memory costs make them impractical for large-scale inputs[16]. To address these issues, we introduce the Efficient RGB-T Cross Fusion Module (EGTCF), which effectively captures extensive multi-scale relationships and long-range feature interdependencies between RGB and thermal modalities without incurring prohibitive computational overhead. Another challenge lies in designing an

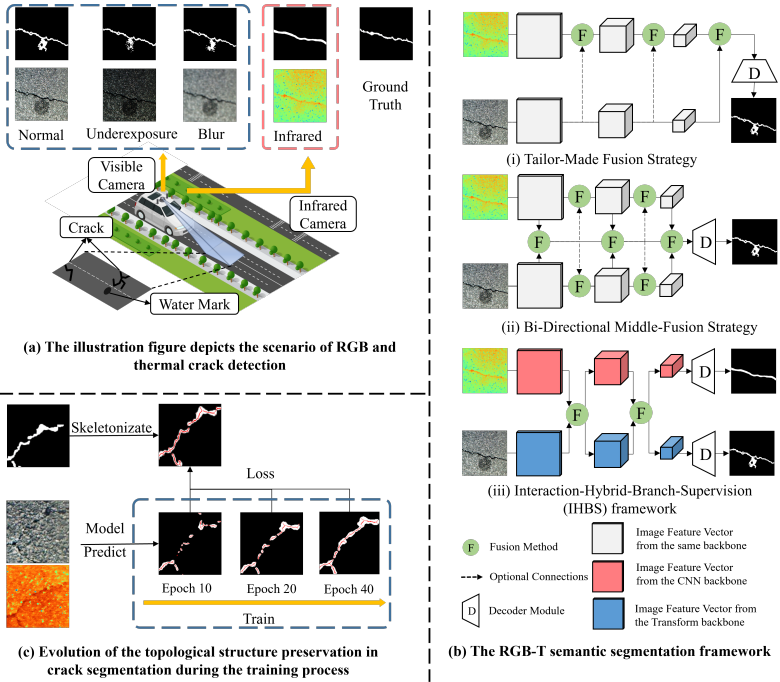


Figure 1: The illustration figure comprises (a) The illustration figure depicts the scenario of RGB and thermal crack detection, (b) The RGB-T semantic segmentation framework and (c) Evolution of the topological structure preservation in crack segmentation during the training process.

effective framework for multi-modal feature learning and supervision. Existing RGB-T segmentation frameworks can be divided into two main types (illustrated in **Figure 1(b)**): the tailor-made under-fused strategy and the bi-directional middle-fusion strategy [26]. The tailor-made approach integrates thermal features into RGB features within the encoder but lacks sufficient inter-modal interaction. The bi-directional middle-fusion strategy promotes interaction by allowing fused features to influence unimodal branches but struggles with supervising modality-specific learning effectively. To overcome these limitations, we propose the Interaction-Hybrid-Branch-Supervision (IHBS) framework. This framework enhances modality interaction by distributing fused feature information across branches and enabling joint supervision of RGB and thermal feature learning. Furthermore, an appropriate loss function is also critical for enhancing deep learning performance in crack segmentation models [23]. Commonly used loss functions in crack segmentation, such as Cross Entropy Loss and Dice Loss [7], focus on pixel-level prediction accuracy but often neglect the topological structure of cracks [24], leading to discontinuities in the segmented outputs. To remedy this, a topological-based loss function, illustrated in **Figure 1(c)**, is introduced to accurately capture and preserve the crack skeleton’s intrinsic topology.

In summary, the key contributions of our research are outlined as follows:

- We propose a novel method called IRFusionFormer for crack segmentation, achieving state-of-the-art results on benchmark datasets. Specifically, our framework outper-

forms existing approaches by attaining Dice and Intersection over Union (IoU) scores of 90.01% and 81.83%, respectively. This demonstrates its effectiveness in accurately identifying and delineating cracks, thereby enhancing both the efficiency and reliability of pavement maintenance practices.

- We propose the deployment of the Efficient RGB-T Cross Fusion Module (EGTCF) to capture multi-scale extensive relationships and long-range feature interdependencies between RGB and thermal modalities. Additionally, the Interaction-Hybrid-Branch-Supervision (IHBS) framework facilitates the sharing of fused feature information across multiple branches and supports simultaneous supervision of feature learning in different modalities.
- We introduce a topology-based loss function aimed at preserving the connectivity and topological structure of asphalt pavement cracks. This innovation significantly advances the accuracy and consistency of crack segmentation by systematically integrating topological considerations into the learning process.

These advancements significantly contribute to the field of pavement maintenance by improving the accuracy and efficiency of crack detection technologies, which are vital for prolonging pavement lifespan and ensuring road safety.

2 Methods

2.1 Overview

In modern image feature extraction, CNN and Transformer-based architectures serve as the primary methods. The CNN architecture is effective in extracting local features, but due to its limited receptive field and the pooling process, it may miss some global-scale correlations. In contrast, the Transformer architecture uses a self-attention mechanism to capture long-range dependencies and global context, enabling a comprehensive understanding of semantic entities. The proposed IHBS framework requires separate feature extraction for infrared and RGB images, followed by an interaction mechanism that enhances modality fusion and enables joint supervision. Infrared images, sensitive to thermal characteristics, highlight areas with temperature differences, while RGB images provide detailed visual information. To achieve this, as shown in **Figure 2**, ResNet is used for infrared image feature extraction, capturing local crack information, while the Segformer network, based on Transformer technology, is employed for RGB images, capturing both local and global features. After processing through the 1st, 2nd, and 4th ResNet or Segformer decoders, the infrared and RGB images of identical size are fused via the EGTCF. The fused features are then redistributed back into their respective modality-specific branches, enhancing inter-modal interaction as part of the IHBS framework, thereby promoting a more effective modality interaction. In the training phase, the IHBS framework employs joint supervision of RGB and thermal feature learning, while both the topological-based loss function and the infrared-based auxiliary loss function guide the segmentation training.

2.2 Efficient RGB-T Cross Fusion Module

Incorporating both thermal and RGB information has been shown to enhance segmentation performance[19]. Current techniques utilizing convolution or pooling layers have led to

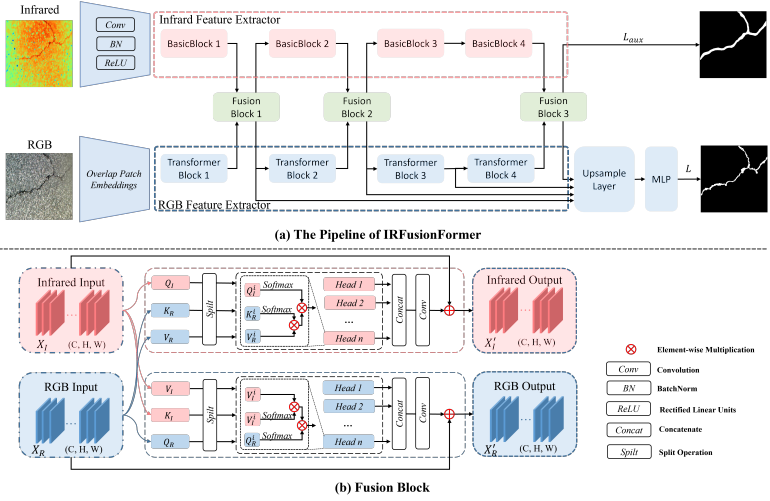


Figure 2: The IRFusionFormer framework comprises (a) the pipeline of the framework and (b) fusion block using efficient cross attention mechanism

restricted receptive fields, limiting the exploration of relationships between RGB images and thermal maps. Moreover, the self-attention mechanism is unsuitable for processing extensive inputs like shallow RGB and infrared feature maps. Instead of employing a basic non-local fusion strategy, we have devised the Enhanced Global Thermal and Color Feature (EGTCF) module for integrating multimodal features, as illustrated in **Figure 2**.

For infrared feature maps denoted as $X_I \in \mathbb{R}^{C \times H \times W}$ and RGB features denoted as $X_R \in \mathbb{R}^{C \times H \times W}$, we initially pass both features through separate convolutional layers to generate query tensor, key tensor, and value tensor. These are denoted as $Q_I \in \mathbb{R}^{C_k \times H \times W}$, $K_I \in \mathbb{R}^{C_k \times H \times W}$, $V_I \in \mathbb{R}^{C_v \times H \times W}$ for infrared features and $Q_R \in \mathbb{R}^{C_k \times H \times W}$, $K_R \in \mathbb{R}^{C_k \times H \times W}$, $V_R \in \mathbb{R}^{C_v \times H \times W}$ for visible features, as shown in Equation 1. Here, C_k and C_v correspond to the dimensions of the convolution matrix (W_k , W_q , W_v).

$$\begin{aligned} Q_I &= W_q^I \cdot X_I, & K_I &= W_k^I \cdot X_I, & V_I &= W_v^I \cdot X_I, \\ Q_R &= W_q^R \cdot X_R, & K_R &= W_k^R \cdot X_R, & V_R &= W_v^R \cdot X_R \end{aligned} \quad (1)$$

Drawing inspiration from the work of [16], a cross-efficient fusion module has been implemented to capture semantic relationships across multimodal features by leveraging an efficient attention mechanism that addresses long-range dependencies while minimizing memory and computational complexities. Initially, the query, key, and value tensors are partitioned into n segments to streamline computational operations. For instance, in the case of

the RGB query tensor, segments are defined as $Q_R^i = Q_R \left[\frac{C_k}{n} \cdot i, \frac{C_k}{n} \cdot (i+1) \right] \in \mathbb{R}^{\frac{C_k}{n} \times H \times W}$. This split approach is also applied to the query, key, and value tensors associated with the infrared and RGB features. Subsequently, the spatial dimensions of V^i and K^i are flattened to $\hat{V}^i \in \mathbb{R}^{\frac{C_v}{n} \times HW}$ and $\hat{K}^i \in \mathbb{R}^{\frac{C_k}{n} \times HW}$. The cross attention matrix A_R and A_I are computed by multiplying \hat{V}^i with the normalized \hat{K}^i from the complementary modality through the soft-

max function: $A_R^i = \text{softmax}(\hat{K}_I^i)^T \cdot \hat{V}_I^i$ and $A_I^i = \text{softmax}(\hat{K}_R^i)^T \cdot \hat{V}_R^i$. Next, we multiply the attention matrix A with the normalized query tensor Q_R within the same modality, concatenating all channels, applying a projection convolution layer, and adding the input feature residuals to generate new fusion features from the efficient RGB-T cross-fusion module:

$$\begin{aligned} X_I' &= \text{conv}(\text{concat}(\text{softmax}(Q_I^i) \cdot A_I^i)_{i \in [0, n-1]}) + X_I, \\ X_R' &= \text{conv}(\text{concat}(\text{softmax}(Q_R^i) \cdot A_R^i)_{i \in [0, n-1]}) + X_R \end{aligned} \quad (2)$$

This efficient RGB-T cross-fusion module facilitates feature interactions between parallel streams during each stage of the feature extraction process. It enables the learning of long-range dependencies from the other modality by correcting its own modal features through these interactions. Employing this methodology significantly reduces computational complexity. For instance, while a non-local module applied to a 256*256 image would require 17GB of memory, our efficient attention approach demands only 67 MB of memory.

2.3 Topological-based Loss Function

In the segmentation of tubular objects such as pavement cracks, considering their topological structure can significantly enhance the usability of the segmentation results[18]. Therefore, our research proposes the integration of a topological-based loss function into the segmentation process, aiming to preserve the structural integrity of cracks. In the context of asphalt pavement crack detection, the actual mask is referred to as V_L , while the model-generated predicted mask is symbolized as V_P . Maximum pooling is applied to refine images by smoothing object boundaries and removing minor noise. The difference between the images before and after this operation highlights the skeleton features, which are further refined through multiple iterations. The skeleton derived from the actual ground truth mask is S_L , while that from the model's prediction is termed as S_P . Topological precision is defined as $T_{\text{precision}} = \frac{|S_P \cap V_L|}{|S_P|}$, which is significantly impacted by False Positives (FP). And topological sensitivity is defined as $T_{\text{sensitivity}} = \frac{|S_L \cap V_P|}{|S_L|}$, which is markedly affected by False Negatives (FN). The loss function L_{Topology} is calculated using the harmonic mean of topological precision and sensitivity:

$$L_{\text{Topology}}(V_L, V_P) = 2 \times \frac{T_{\text{precision}}(S_P, V_L) \times T_{\text{sensitivity}}(S_L, V_P)}{T_{\text{precision}}(S_P, V_L) + T_{\text{sensitivity}}(S_L, V_P)}$$

However, L_{Topology} is primarily focused on the overall continuity and connectivity, which can result in challenges when attempting to accurately segment crack edges within images. Consequently, in the proposed loss function, the topological-base loss function L_{Topology} is combined with Cross Entropy Loss L_{CE} and Dice Loss L_{Dice} using the weights α , β and γ , thus constructing a more robust framework for crack segmentation. This composite loss function leverages the strengths of each component: L_{Topology} for maintaining topological integrity, L_{CE} for pixel-wise accuracy, and L_{Dice} for optimizing over imbalanced data conditions. Additionally, in the training phase of the proposed framework, features extracted from infrared images are fused with RGB image features to produce a fusion segmentation result, which is denoted as V_P^I . Unlike the final output, this result is based on the feature of the infrared images, incorporating their inherent physical constraint features. To further capitalize on these infrared-specific constraints and enhance crack segmentation, the proposed

loss function includes an auxiliary loss function, denoted as $L_{\text{aux}}(V_L, V_P^I)$. Ultimately, the proposed loss function consists of the aforementioned loss functions:

$$\begin{aligned} L(V_L, V_P) &= \alpha L_{\text{Topology}}(V_L, V_P) + \beta L_{\text{CE}}(V_L, V_P) + \gamma L_{\text{Dice}}(V_L, V_P) \\ L_{\text{aux}}(V_L, V_P^I) &= \alpha L_{\text{Topology}}(V_L, V_P^I) + \beta L_{\text{CE}}(V_L, V_P^I) + \gamma L_{\text{Dice}}(V_L, V_P^I), \\ L(V_L, V_P, V_P^I) &= L(V_L, V_P) + \delta L_{\text{aux}}(V_L, V_P^I) \end{aligned} \quad (3)$$

where α , β , γ and δ represent the weights for L_{Topology} , L_{CE} , L_{Dice} and L_{aux} .

3 Experiments

3.1 Dataset

The dataset employed in this study is an open-source dataset[[11](#)] dedicated to crack detection using Infrared Thermography (IRT), which is included in the RGB-T asphalt pavement crack segmentation benchmark. It comprises four image types: RGB images, infrared images, fused images (combined at a 50:50 ratio using IR-Fusion™ technology), and ground truth images manually annotated using Photoshop. Each category consists of 448 images, each with a resolution of 640x480 pixels. For training and evaluation purposes, the segmentation model divides the entire dataset into two subsets: 358 images for the training set and 90 images for the test set.

3.2 Training Details

To enhance the diversity of the training data and improve the model’s robustness, spatial, color, and numerical transformations were applied to the training set, such as random horizontal or vertical flips in spatial, randomly altering brightness or contrast in color. For images in the validation set, only resizing to 480x480 pixels and normalization processes were applied. The proposed IRFusionFormer was implemented using the PyTorch framework and optimized with AdamW, incorporating a weight decay of 1e-4. A batch size of 8 and 150 training epochs were designated for training. All experiments were performed on an NVIDIA GeForce RTX 4090 to expedite model training.

3.3 Evaluation Metrics

To accurately and objectively evaluate the performance of various models, we employed six widely used evaluation metrics in image segmentation: Dice, IoU (Intersection over Union), Accuracy, Precision, Specificity, and Recall. Higher values for these metrics indicate superior segmentation performance. Simultaneously, these metrics are also used as the evaluation criteria for the RGB-T asphalt pavement crack segmentation benchmark.

4 Results

4.1 Comparison with State-of-the-Art Methods

To validate the effectiveness of the proposed method, we compared it against eight main-stream models on the dataset. Among these, MCNet[[12](#)] uses only infrared images as input,

whereas U-net[15], UNet++[29], DeepLabV3[10], DeepCrack[30], and CrackFormer[16] use solely RGB images. CRM_RGBT_Seg[17] and CMNeXt[28] employ both infrared and RGB images as inputs. IRFusionFormer, along with eight comparative models, constitutes the benchmarking suite of the proposed benchmark. The quantitative results are summarized in **Table 1** and demonstrate that our proposed IRFusionFormer outperforms other SOTA methods on the dataset. **Figure 3** displays the visual comparison of 9 models on the test sets of the datasets, with some results highlighting the skeleton of the cracks.

Table 1: Quantitative results on various datasets. Best and second-best results are bold and underlined, respectively. For type, 'I' stands for Infrared, 'R' stands for RGB, and 'IR' stands for Infrared+RGB.

Type	Models(Venue)	Dice	IoU	Accuracy	Precision	Specificity	Recall
I	MCNet[12]	0.6844	0.5202	0.9467	0.7482	0.9786	0.6306
R	U-net[15]	0.7891	0.6517	0.9794	0.8161	0.9909	0.7639
R	UNet++[29]	0.8048	0.6733	0.9801	0.7937	0.9887	0.8574
R	DeepLabV3[10]	0.8338	0.7149	0.9828	0.8134	0.9896	0.8552
R	DeepCrack[30]	0.7406	0.5880	0.9787	0.6592	0.9837	0.8450
R	CrackFormer[16]	0.8489	0.7374	<u>0.9847</u>	0.8462	0.9918	0.8515
IR	CRM_RGBT_Seg[17]	0.8450	0.7370	0.9829	0.8651	<u>0.9921</u>	0.8293
IR	CMNeXt[28]	<u>0.8760</u>	<u>0.7794</u>	0.9835	<u>0.8885</u>	<u>0.9921</u>	<u>0.8639</u>
IR	<i>IRFusionFormer(ours)</i>	0.9001	0.8183	0.9899	0.9001	0.9947	0.9001

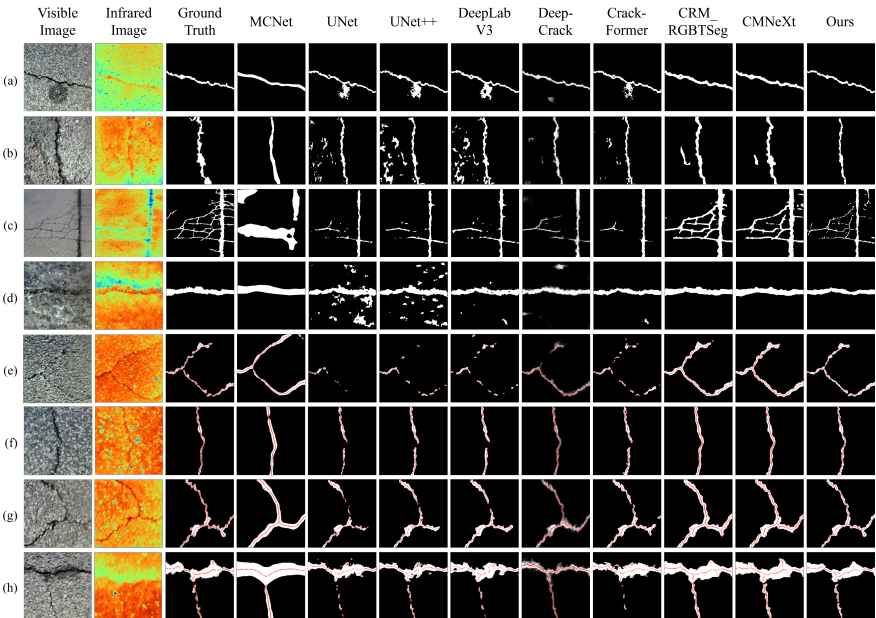


Figure 3: Comparison of crack segmentation results from nine models on test set of the dataset. Images (e)-(h) illustrate the skeleton of the cracks.

Results in **Table 1** indicate that the RGB-infrared integrated model outperforms the model that uses only RGB images, both of which are superior to the model that uses only

infrared images. As seen in **Figure 3**, MCNet’s segmentation results are smoother at the edges compared to other networks, indicating a lack of sufficient information. The proposed IRFusionFormer achieved the best results across all six evaluation metrics on the test dataset, outperforming the second-best model, CMNeXt, in Dice, IoU, Accuracy, Precision, Specificity, and Recall by 2.41%, 3.89%, 0.64%, 1.16%, 0.26%, and 3.62%, respectively. More importantly, infrared-integrated models were less affected by the presence of watermarks, shadows, or other disturbances on the pavement, compared to RGB-only models. These findings suggest that while infrared images alone are insufficient in isolation for accurate asphalt pavement crack segmentation, their integration with RGB images can markedly enhance prediction accuracy and reduce the impact of complex environmental factors. Additionally, as illustrated in **Figure 3**, the proposed method provides a more complete crack skeleton, attributed to the comprehensive crack information from infrared images and the specially designed loss function that accounts for the topology of cracks and integrates auxiliary loss from infrared feature fusion.

4.2 Ablation Study

Our IRFusionFormer mainly consists of two parts: the Efficient RGB-T Cross Fusion Module and the Topological-based Loss Function. Therefore, we conduct ablation studies to verify the effectiveness of each component, and then analyze the stages of fusion modules and loss function weights. Each of the three fusion modules in the model was evaluated both individually and in combination, with results detailed in **Table 2**. Additionally, the impact of varying weights of the topological-based loss function and the incorporation of auxiliary loss functions was examined, with findings presented in **Table 3** and **Table 4**.

Table 2: Quantitative results on various fusion stages. Best and second-best results are bold and underlined, respectively.

Fusion			Dice	IoU	Accuracy	Precision	Specificity	Recall
Stage 0	Stage 1	Stage 2						
✓			0.8706	0.7709	0.9869	0.8644	0.9927	0.8770
	✓		0.8717	0.7726	0.9870	0.8680	0.9929	0.8755
		✓	0.8737	0.7757	0.9872	0.8680	0.9929	0.8794
✓	✓		0.8819	0.7887	0.9881	0.8817	0.9937	0.8821
✓		✓	0.8838	0.7917	0.9883	0.8852	0.9939	<u>0.8823</u>
	✓	✓	<u>0.8841</u>	<u>0.7923</u>	<u>0.9884</u>	<u>0.8863</u>	<u>0.9940</u>	0.8820
✓	✓	✓	0.9001	0.8183	0.9899	0.9001	0.9947	0.9001

Analysis of **Table 2** reveals that experimental outcomes improve as the number of Fusion Blocks increases within the feature extraction network. Using Fusion Blocks across all three stages yielded optimal results. Notably, the application of a Fusion Block at the third stage, corresponding to the high-level feature stage, more effectively captures information from the alternate modality than at the low-level stages.

From **Table 3**, it is evident that the use of the auxiliary loss function significantly improves segmentation performance. However, when the weight of the auxiliary loss function increases, the results of crack segmentation decline. Specifically, when the weight is 0.1, the results are optimal. From **Table 4**, the topological-based loss function demonstrates an enhanced effect on crack segmentation. However, excessively high weights of the topological-based loss function lead to decreased segmentation performance. Consequently, an appropri-

Table 3: Quantitative results with varying auxiliary loss function weight. Best and second-best results are bold and underlined, respectively.

Aux Loss	Dice	IoU	Accuracy	Precision	Specificity	Recall
0	0.8759	0.7792	0.9875	0.8752	0.9934	0.8766
0.1	0.9001	0.8183	0.9899	<u>0.9001</u>	<u>0.9947</u>	0.9001
0.2	<u>0.8886</u>	<u>0.7996</u>	<u>0.9889</u>	<u>0.8949</u>	<u>0.9945</u>	0.8824
0.3	0.8875	0.7978	0.9888	0.9005	0.9949	0.8748
0.4	0.8857	0.7948	0.9885	0.8898	0.9942	0.8816
0.5	0.8834	0.7912	0.9883	0.8836	0.9938	<u>0.8833</u>

Table 4: Quantitative results with varying and topological-based loss function weight. Best and second-best results are bold and underlined, respectively.

L_{topology} Weight	Dice	IoU	Accuracy	Precision	Specificity	Recall
0	0.8792	0.7844	0.9881	<u>0.8989</u>	0.9949	0.8604
0.1	0.8842	0.7924	0.9885	<u>0.8959</u>	0.9946	0.8728
0.2	<u>0.8886</u>	<u>0.7996</u>	<u>0.9889</u>	0.8949	0.9945	0.8824
0.3	0.9001	0.8183	0.9899	0.9001	<u>0.9947</u>	0.9001
0.4	0.8861	0.7955	0.9884	0.8766	<u>0.9933</u>	<u>0.8958</u>
0.5	0.8815	0.7881	0.9880	0.8793	0.9936	<u>0.8836</u>
0.6	0.8729	0.7745	0.9871	0.8682	0.9929	0.8776

ately weighted topological-based loss function optimally enhances the model’s segmentation performance. The ablation study shows, a weight of 0.3 yields the best performance. Overall, these results demonstrate that the proposed multi-scale fusion module and appropriately weighted loss functions significantly contribute to the performance of the crack detection model.

5 Conclusions

To integrate RGB and infrared image information for segmenting asphalt pavement cracks in complex environments, we propose the RGB-T asphalt pavement crack segmentation benchmark. The benchmark includes a dataset of image pairs, a codebase comprising nine algorithms, six evaluation metrics, as well as all related results. This benchmark provided a new platform for various methods. IRFusionFormer, which was a new crack segmentation method proposed in this research, achieved state-of-the-art (SOTA) results in the established benchmark. In the proposed method, the Efficient RGB-T Cross Fusion Module was incorporated into the Interaction-Hybrid-Branch-Supervision (IHBS) framework, which was designed to efficiently fuse features from RGB and infrared images across three key stages. Additionally, a topological-based loss function was employed, specifically tailored to handle the topological structures of cracks, thereby improving the accuracy and robustness of the crack segmentation. Ablation study results demonstrated that these techniques significantly improve the performance of the IRFusionFormer network in crack segmentation tasks.

References

- [1] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018.
- [2] Qiangpu Chen, Wushao Wen, and Jinghui Qin. Globalsr: Global context network for single image super-resolution via deformable convolution attention and fast fourier convolution. *Neural Networks*, page 106686, 2024.
- [3] Dimitris Dais, Ihsan Engin Bal, Eleni Smyrou, and Vasilis Sarhosis. Automatic crack classification and segmentation on masonry surfaces using convolutional neural networks and transfer learning. *Automation in Construction*, 125:103606, 2021.
- [4] Yuchuan Du, Xiaoming Zhang, Feng Li, and Lijun Sun. Detection of crack growth in asphalt pavement through use of infrared imaging. *Transportation Research Record*, 2645(1):24–31, 2017.
- [5] Lili Fan, Shen Li, Ying Li, Bai Li, Dongpu Cao, and Fei-Yue Wang. Pavement cracks coupled with shadows: A new shadow-crack dataset and a shadow-removal-oriented crack detection approach. *IEEE/CAA Journal of Automatica Sinica*, 10(7):1593–1607, 2023.
- [6] Chengjia Han, Handuo Yang, Tao Ma, Shun Wang, Chaoyang Zhao, and Yaowen Yang. Crackdiffusion: A two-stage semantic segmentation framework for pavement crack combining unsupervised and supervised processes. *Automation in Construction*, 160: 105332, 2024.
- [7] Shruti Jadon. A survey of loss functions for semantic segmentation. In *2020 IEEE conference on computational intelligence in bioinformatics and computational biology (CIBCB)*, pages 1–7. IEEE, 2020.
- [8] Jiahao Jiang, Peng Li, Junjie Wang, Hong Chen, and Tiantian Zhang. Asphalt pavement crack detection based on infrared thermography and deep learning. *International Journal of Pavement Engineering*, 25(1):2295906, 2024.
- [9] Zülfiye Küçük and Görkem Algan. Semantic segmentation for thermal images: A comparative survey. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 286–295, June 2022.
- [10] Qingquan Li and Xianglong Liu. Novel approach to pavement image segmentation based on neighboring difference histogram method. In *2008 congress on image and signal processing*, volume 2, pages 792–796. IEEE, 2008.
- [11] Fangyu Liu, Jian Liu, and Linbing Wang. Asphalt pavement crack detection based on convolutional neural network and infrared thermography. *IEEE Transactions on Intelligent Transportation Systems*, 23(11):22145–22155, 2022.
- [12] Huajun Liu, Jing Yang, Xiangyu Miao, Christoph Mertz, and Hui Kong. Crackformer network for pavement crack segmentation. *IEEE Transactions on Intelligent Transportation Systems*, 24(9):9240–9252, 2023.

- [13] Huidrom Lokeshwor, Lalit K Das, and Savita Goel. Robust method for automated segmentation of frames with/without distress from road surface video clips. *Journal of Transportation Engineering*, 140(1):31–41, 2014.
- [14] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.
- [15] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention—MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.
- [16] Zhuoran Shen, Mingyuan Zhang, Haiyu Zhao, Shuai Yi, and Hongsheng Li. Efficient attention: Attention with linear complexities. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 3531–3539, 2021.
- [17] Ukcheol Shin, Kyunghyun Lee, In So Kweon, and Jean Oh. Complementary random masking for rgb-thermal semantic segmentation, 2024. URL <https://arxiv.org/abs/2303.17386>.
- [18] Suprosanna Shit, Johannes C Paetzold, Anjany Sekuboyina, Ivan Ezhov, Alexander Unger, Andrey Zhylyka, Josien PW Pluim, Ulrich Bauer, and Bjoern H Menze. cldice—a novel topology-preserving loss function for tubular structure segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16560–16569, 2021.
- [19] Yuxiang Sun, Weixun Zuo, Peng Yun, Hengli Wang, and Ming Liu. Fuseseg: Semantic segmentation of urban scenes based on rgb and thermal data fusion. *IEEE Transactions on Automation Science and Engineering*, 18(3):1000–1011, 2020.
- [20] Linfeng Tang, Jiteng Yuan, and Jiayi Ma. Image fusion in the loop of high-level vision tasks: A semantic-aware real-time infrared and visible image fusion network. *Information Fusion*, 82:28–42, 2022.
- [21] Thai Son Tran, Son Dong Nguyen, Hyun Jong Lee, and Van Phuc Tran. Advanced crack detection and segmentation on bridge decks using deep learning. *Construction and Building Materials*, 400:132839, 2023.
- [22] Hanxiang Wang, Yanfen Li, L Minh Dang, Sujin Lee, and Hyeonjoon Moon. Pixel-level tunnel crack segmentation using a weakly supervised annotation approach. *Computers in Industry*, 133:103545, 2021.
- [23] Qi Wang, Yue Ma, Kun Zhao, and Yingjie Tian. A comprehensive survey of loss functions in machine learning. *Annals of Data Science*, pages 1–26, 2020.
- [24] Wenjuan Wang, Allen Zhang, Kelvin CP Wang, Andrew F Braham, and Shi Qiu. Pavement crack width measurement based on laplace’s equation for continuity and unambiguity. *Computer-Aided Civil and Infrastructure Engineering*, 33(2):110–123, 2018.
- [25] Zuoxu Wang, Hancheng Zhang, Zhendong Qian, and Leilei Chen. A complex scene pavement crack semantic segmentation method based on dual-stream framework. *International Journal of Pavement Engineering*, 24(2):2286461, 2023.

- [26] Wei Wu, Tao Chu, and Qiong Liu. Complementarity-aware cross-modal feature fusion network for rgb-t semantic segmentation. *Pattern Recognition*, 131:108881, 2022.
- [27] Haitao Xiong, Wenjie Cai, and Qiong Liu. Mcnet: Multi-level correction network for thermal image semantic segmentation of nighttime driving scene. *Infrared Physics & Technology*, page 103628, 2021. ISSN 1350-4495. doi: <https://doi.org/10.1016/j.infrared.2020.103628>.
- [28] Jiaming Zhang, Huayao Liu, Kailun Yang, Xinxin Hu, Ruiping Liu, and Rainer Stiefelhagen. Cmx: Cross-modal fusion for rgb-x semantic segmentation with transformers. *IEEE Transactions on Intelligent Transportation Systems*, 2023.
- [29] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE Transactions on Medical Imaging*, 2019.
- [30] Qin Zou, Zheng Zhang, Qingquan Li, Xianbiao Qi, Qian Wang, and Song Wang. Deep-crack: Learning hierarchical convolutional features for crack detection. *IEEE Transactions on Image Processing*, 28(3):1498–1512, 2019.