symmetry

MDPI

Article

Intuitionistic Fuzzy Set Guided Fast Fusion Transformer for Multi-Polarized Petrographic Image of Rock Thin Sections

Bowei Chen [1], Bo Yan [2], Wenqiang Wang [2], Wenmin He [2], Yongwei Wang [2], Lei Peng [2], Andong Wang [2] and Li Chen [1,*]

[1] School of Information Science and Technology, Northwest University, Xi'an 710127, China; 202010284@stumail.nwu.edu.cn
[2] Shaanxi HPC Engineering Laboratory, Shaanxi Railway Institute, Weinan 714000, China; 202110855@sxri.edu.cn (B.Y.); 202110891@sxri.edu.cn (W.W.); 200100133@sxri.edu.cn (W.H.); chandlerrr@foxmail.com (Y.W.); ruthieee@foxmail.com (L.P.); fei258fei@163.com (A.W.)
* Correspondence: chenli@nwu.edu.cn

Abstract: The fusion of multi-polarized petrographic images of rock thin sections involves the fusion of feature information from microscopic images of rock thin sections illuminated under both plane-polarized and orthogonal-polarized light. During the fusion process of rock thin section images, the inherent high resolution and abundant feature information of the images pose substantial challenges in terms of computational complexity when dealing with massive datasets. In engineering applications, to ensure the quality of image fusion while meeting the practical requirements for high-speed processing, this paper proposes a novel fast fusion Transformer. The model leverages a soft matching algorithm based on intuitionistic fuzzy sets to merge redundant tokens, effectively mitigating the negative effects of asymmetric dependencies between tokens. The newly generated artificial tokens serve as brokers for the Query (Q), forming a novel lightweight fusion strategy. Both subjective visual observations and quantitative analyses demonstrate that the Transformer proposed in this paper is comparable to existing fusion methods in terms of performance while achieving a notable enhancement in its inference efficiency. This is made possible by the attention paradigm, which is equivalent to a generalized form of linear attention, and the newly designed loss function. The model has been experimented on with multiple datasets of different rock types and has exhibited robust generalization capabilities. It provides potential for future research in diverse geological conditions and broader application scenarios.

Keywords: image fusion; intuitionistic fuzzy set; asymmetric dependency; rock thin section; token merging

Citation: Chen, B.; Yan, B.; Wang, W.; He, W.; Wang, Y.; Peng, L.; Wang, A.; Chen, L. Intuitionistic Fuzzy Set Guided Fast Fusion Transformer for Multi-Polarized Petrographic Image of Rock Thin Sections. *Symmetry* **2024**, *16*, 1705. https://doi.org/10.3390/sym16121705

Academic Editor: Marek T. Malinowski

Received: 18 November 2024
Revised: 19 December 2024
Accepted: 21 December 2024
Published: 23 December 2024

1. Introduction

The analysis of rock thin section imagery represents an indispensable geological exploration [1] tool for understanding and recognizing the composition of the Earth. Furthermore, it serves as a significant evaluation method in oil and gas exploration and development [2]. This technique can be employed to identify petrological properties of reservoir rocks [3], ascertain genetic types [4], and differentiate the characteristics of reservoir space and pore structure [5]. Traditional analysis of rock thin section imagery primarily relies on manual methods, which are not only time-consuming and labor-intensive but also susceptible to subjective influences, making it difficult to guarantee the accuracy and consistency of analysis results. With the rapid development of artificial intelligence and computer vision technologies, research into feature fusion for rock thin section imagery has become an urgent necessity.

Image fusion [6] based on multiple light sources can integrate feature information from rock thin sections under different illumination conditions (such as plane-polarized light and orthogonal polarization [7]). Figure 1 illustrates the inclusion of multi-polarized

image data for three distinct rock types: sedimentary, metamorphic, and igneous [8]. The incorporation of such multidimensional information serves to enhance the precision of image analysis and mitigate errors that may arise from reliance on a single information source. Moreover, image fusion techniques provide a more abundant and accurate data foundation for automated rock classification and identification [9], thereby offering novel insights and directions for the interdisciplinary integration of geology with computer vision and artificial intelligence.

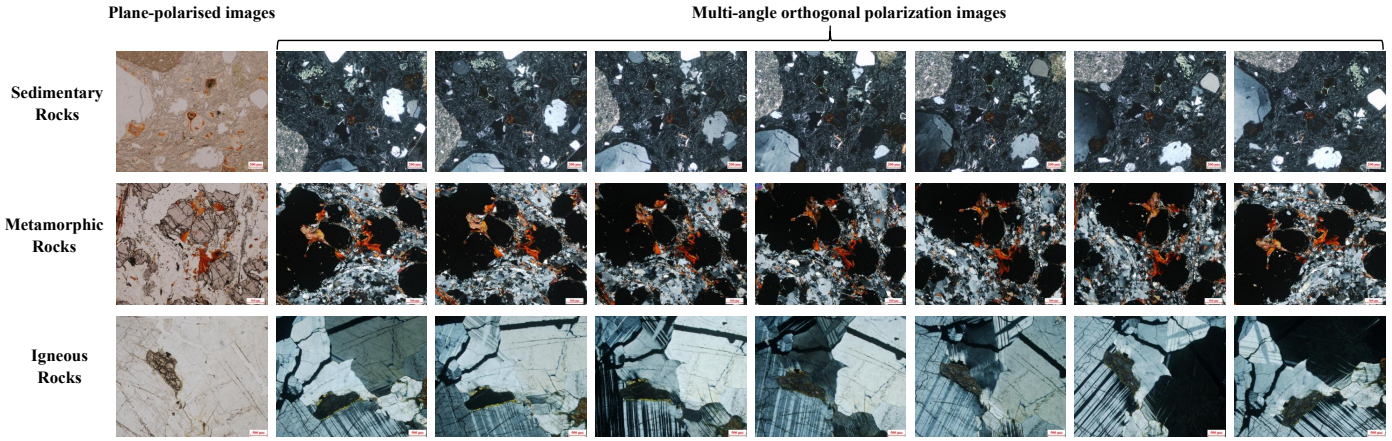

Figure 1. Thin section images of rocks of different species and polarization modes with a scaling dimension of 500 micrometer.

This task is also confronted with many challenges. Rock thin section images exhibit both local micro-features and global composite features. It is essential to strike a balance between these two aspects during the fusion process, ensuring that the fused images can reflect both local details and global structures. In this regard, Transformer [10] has demonstrated excellent global feature capture capability in the image fusion, due to its advantage of global attention. Unlike traditional sequential models such as RNN [11] or LSTM [12], the Transformer can process information within images in parallel, which significantly enhances computational efficiency. However, when dealing with the vast array of high-resolution and highly feature-complex rock thin section images, a formidable challenge lies in improving computational efficiency and reducing resource consumption while ensuring fusion effectiveness. For the processing of rock thin section image fusion, the model compression of models and the guarantee of computational speed are urgent issues to be addressed.

In light of the above, this paper introduces a compressed and rapid image feature fusion Transformer. The experimental results, based on diverse categories of rock data, indicate that this fusion model exhibits comparable accuracy to other fusion methods. The key contributions of the proposed method are as follows:

1. This paper presents a Transformer for fast fusion of rock thin section imagery. By amalgamating multi-scale features from both orthogonal polarization and plane-polarized images, it effectively preserves and enhances critical characteristic information in rock thin section imagery, including color, texture, shape, and mineral composition.
2. Based on the theory of intuitionistic fuzzy sets, a new merging strategy for the asymmetric dependence characteristics among tokens is proposed. The attention paradigm constructed by generating artificial tokens (broker) ensures the robustness of feature fusion and improves the processing speed of the model.
3. A new loss function has been designed for the fusion Transformer proposed in this paper. By optimizing the loss related to attention weights, it ensures that the Transformer focuses on crucial feature areas, such as mineral grains and texture.

2. Related Work

Recently, various techniques have been introduced to combine consistent features extracted from different polarized images. Shen et al. [13] suggested a fusion method for visible-light polarization images aimed at detecting mines during nighttime. They implemented a hybrid attention mechanism to improve the network's ability to extract important information from the feature tensor, ensuring the fused image retained significant details from the prominent pixel regions in the original images, ultimately producing an end-to-end output. Li et al. [14] proposed the Polarized Prior Guided Fusion Network for infrared polarization images. This approach employs a learned low-rank decomposition model to extract a low-rank representation that captures background details in infrared intensity, along with sparse features of key targets within the Degree of Linear Polarization (DoLP) images. Their fusion model effectively maintains prominent polarized targets while minimizing background interference with fewer parameters. Xu et al. [15] introduced an innovative unsupervised fusion network, PAPIF, which merges polarization and intensity images through pixel-based guidance and attention mechanisms. In this model, fusion targets include high-polarization elements from polarization images and detailed textures from intensity images. Both channel and spatial attention mechanisms are utilized to combine essential features while filtering out irrelevant information. Considering long-range dependencies between the fused and source images, Li et al. [16] developed the DFENet model. This model includes a Global Semantic Information Aggregation Module that efficiently gathers multi-scale features. Additionally, it integrates a fusion strategy that combines both local and gradient information to enhance performance. Liu et al. [17], in an effort to optimize modality discrepancies in multimodal images, introduced the Multimodal Feature Self-Adaptive Transformer model. This model integrates multimodal features via a self-adaptive fusion strategy during training. Yi et al. [18] proposed the TCPMFNet, which is based on an autoencoder network architecture. A hybrid fusion strategy, parallelly combining CNN and Transformer components, is also incorporated in their design. Li et al. [19] presented the CGTF model, which introduces skip connections within a hybrid CNN–Transformer framework. This model is engineered to extract both local and global features from images simultaneously. Tang et al. [20] put forward the MATR model for multimodal medical image fusion. This model incorporates an adaptive convolution mechanism that adjusts the convolution kernel based on the global background context. The adaptive Transformer component enhances the extraction of global semantic features and captures information across multiple scales. Wang et al. [21] introduced Res2Fusion, which utilizes a non-local fusion module.

3. Methodology

3.1. IFS Token Merging

Existing token merging methods are often constrained by the issue of asymmetric dependencies, which stem from the sequential and spatial nature of data. In such scenarios, a token may exert substantial influence on other tokens, while the reverse influence is often weaker or negligible. This imbalance can lead to suboptimal merging decisions, as strategies that rely solely on proximity principles risk introducing varying degrees of feature loss. For example, in rock imagery, the structural features of crystals serve as the core information transmitters, influencing surrounding textural features. However, the feedback from these peripheral features to the crystal core is typically minimal, further exacerbating the effects of asymmetric dependencies. Consequently, such approaches may fail to preserve critical features in tasks requiring fine-grained semantic understanding.

To address these challenges, this study proposes a token merging strategy based on intuitionistic fuzzy sets, which are particularly well suited for this task due to their ability to model complex dependencies and manage uncertainty. Unlike conventional methods, IFS introduces a richer representational framework by simultaneously quantifying membership, non-membership, and hesitation degrees. In the context of token merging, the membership degree measures the semantic similarity between a target token and a can-

didate token for merging, the non-membership degree assesses whether the candidate token might be better suited for an alternative match, and the hesitation degree reflects uncertainty in the merging decision, especially when multiple tokens exhibit comparable similarity. By leveraging this tripartite representation, the IFS-based method mitigates the adverse effects of asymmetric dependencies through a bidirectional evaluation process. Traditional merging algorithms, such as greedy soft matching, primarily focus on the perspective of the target token, identifying the most similar token to merge while neglecting the preferences or optimality from the candidate token's perspective. The inclusion of the non-membership degree explicitly addresses this limitation by ensuring that merging decisions account for both the target and candidate tokens, thereby balancing their mutual compatibility. Moreover, the IFS framework offers a robust mechanism for handling semantic conflicts and redundancies, which frequently arise in token merging tasks. For instance, when multiple tokens exhibit similar degrees of proximity to a target token, the hesitation degree allows the algorithm to dynamically identify ambiguous regions and adjust the merging strategy accordingly. This feature is particularly critical for applications like rock imagery analysis, where the preservation of fine-grained structural details is essential. By jointly considering membership, non-membership, and hesitancy, the proposed method systematically evaluates and resolves conflicting information, thereby minimizing semantic loss during the merging process.

Assuming we have a set of tokens T, it is evenly divided into sets A and B, $A \cup B = T$, and $A \cap B = \emptyset$. For each token a_i in set A, the token most similar to it (or tokens $B_{j_1}, B_{j_2}, \ldots, B_{j_n}$) is found in Set B. Membership $\mu(a_i, b_j)$ can be calculated as follows:

$$\mu(a_i, b_j) = \frac{\sum_{k \in \{j_1, j_2, \ldots, j_n\}} cosine\_sim(a_i, b_k) \cdot \delta(a_i, b_k)}{\sum_{k \in \{j_1, j_2, \ldots, j_n\}} cosine\_sim(a_i, b_k)} \tag{1}$$

where *cosine_sim* represents the cosine similarity [22]. $\delta(a_i, b_j)$ is an indicator function that takes the value 1 when b_j is the most similar match of a_i and 0 otherwise. However, since we only take the most similar one, the equation can be simplified to

$$\mu(a_i, b_{j^*}) = \frac{cosine\_sim(a_i, b_{j^*})}{\sum_{k=1}^{n^*} cosine\_sim(a_i, b_k)} \tag{2}$$

$$cosine\_sim(a_i, b_j) = \frac{a_i \cdot b_j}{\|a_i\| \|b_j\|} \tag{3}$$

where b_{j^*} is the most similar match of a_i in B and n^* is the number of most similar matches actually found. The hesitancy [23] indicates the difference in similarity for given matching token pairs. It can be obtained by calculating the standard deviation of all similarity matching memberships:

$$H(a_i, B^*) = \sqrt{\frac{1}{|B^*|} \sum_{b_j \in B^*} \left(\mu(a_i, b_j) - \bar{\mu}(a_i, B^*)\right)^2} \tag{4}$$

where B^* represents the set of all tokens for which there is at least one directionally most similar case. It is important to note that, as only the most similar b_j is taken, the computation of all possible memberships is not necessary, especially when B is large. $\bar{\mu}(a_i, B^*)$ is the average of the membership of all elements in a_i and set B^*:

$$\bar{\mu}(a_i, B^*) = \frac{1}{|B^*|} \sum_{b_j \in B^*} \mu(a_i, b_j) \tag{5}$$

Non-membership can be defined as 1 minus membership and an adjustment for hesitancy (which may not fall within the range [0, 1]):

$$\nu(a_i, B) = 1 - (\mu(a_i, b_{j^*}) + \alpha \cdot H(a_i, B^*)) \tag{6}$$

where α is an adjustment factor to control the effect of hesitancy on non-membership. The degree of similarity between a_i and b_j is proportional to the membership. When multiple tokens are similar, it can be inferred that the optimal match is significantly superior to the other options. A hasty merging would be disadvantageous for the remaining tokens, which would be compelled to merge with a second-best option. This may result in the loss of specific features. Low non-membership is more beneficial for matching and merging of tokens. A simplified Algorithm 1 is listed below.

Algorithm 1 Token soft merging algorithm with intuitionistic fuzzy set.

Require: T: Input tokens set
Ensure: Matched pairs and associated membership, non-membership and hesitancy
$A, B \leftarrow \text{Split}(T)$ ▷ Split T into A and B such that $A \cup B = T$ and $A \cap B = \emptyset$
for all $a_i \in A$ **do**
$B_{\text{similar}} \leftarrow \text{FindMostSimilar}(a_i, B)$ ▷ Find most similar token(s) in B for a_i
$\mu(a_i, B) \leftarrow 0$ ▷ Initialize membership
for all $b_{j^*} \in B_{\text{similar}}$ **do**
$sim \leftarrow \text{CosineSimilarity}(a_i, b_{j^*})$
$\mu(a_i, B) \leftarrow \mu(a_i, B) + sim$ ▷ Assuming single or aggregated similarity
end for
if $|B_{\text{similar}}| = 1$ **then**
$\mu(a_i, b_{j^*}) \leftarrow \frac{sim}{\sum_{k=1}^{|B_{\text{similar}}|} sim}$ ▷ Simplified to $\mu(a_i, b_{j^*}) = sim$ if only one
else
$\mu(a_i, b_{j^*}) \leftarrow \frac{\sum_{k=1}^{|B_{\text{similar}}|} sim}{\sum_{k=1}^{|B_{\text{similar}}|} sim}$ ▷ Aggregate if multiple, but simplified denominator may apply
end if
$H(a_i, B) \leftarrow \text{StdDev}(\mu(a_i, b_{j_1}), \mu(a_i, b_{j_2}), \ldots)$ ▷ Hesitancy, but may be 0 if single match
$\nu(a_i, B) \leftarrow 1 - \mu(a_i, b_{j^*})$ ▷ Non-membership, simplified
end for

The proposed IFS-based token merging strategy directly addresses the challenges posed by asymmetric dependencies and the inherent uncertainties of token merging tasks. Its ability to model bidirectional compatibility, resolve conflicts, and adaptively manage uncertainty makes it particularly well suited for applications requiring fine-grained semantic preservation, such as in rock imagery or other tasks involving complex data structures.

3.2. Broker Transformer

This paper presents an improved linear attention [24] broker, which is designed and applied to a trained Swin Transformer [25] module to construct a novel fusion network for multi-polarized rock thin section images. In order to more accurately represent the global context and spatial relationships, an encoding vector related to the relative positions of neighboring elements is initially generated for each image element. These encoding vectors can be combined with the embedding vectors of the image elements to form the input to the Transformer. It is necessary to transform the 2D embedding vector arrays using a reshape operation in order to satisfy the requirements of the Transformer input. The entire structure of the model is depicted in Figure 2. The feature extractor module underwent a process of pre-training, whereby each block of the three Swin Transformer modules contained six Swin Transformer layers. Each layer is composed of two consecutive components: Window Multihead Self-Attention (WMSA) and Moving Window Multihead Self-Attention (SW-MSA) [26]. The convolutional layers are employed to extract the shallow features of the image and map them to a high-dimensional space, thereby ensuring the extraction and fusion of features. The convolutional network encoder consists of a convolutional layer with a convolutional kernel of 3×3, as well as three convolutional modules with a step size of 1. The multiple layers of Swin Transformer are employed to extract deep features that

encompass global information. Within each layer, they execute self-attention computations within the windows and shifting operations between windows. The dimensions of the input images are uniformly set to 1280 × 1024, and subsequently these images are split into fixed patch blocks of size 32 × 32. Each image produces 1280 patch blocks, and the length of the input sequence becomes 1280. Consequently, the total number of tokens and the dimension of each token are determined to be 1280.

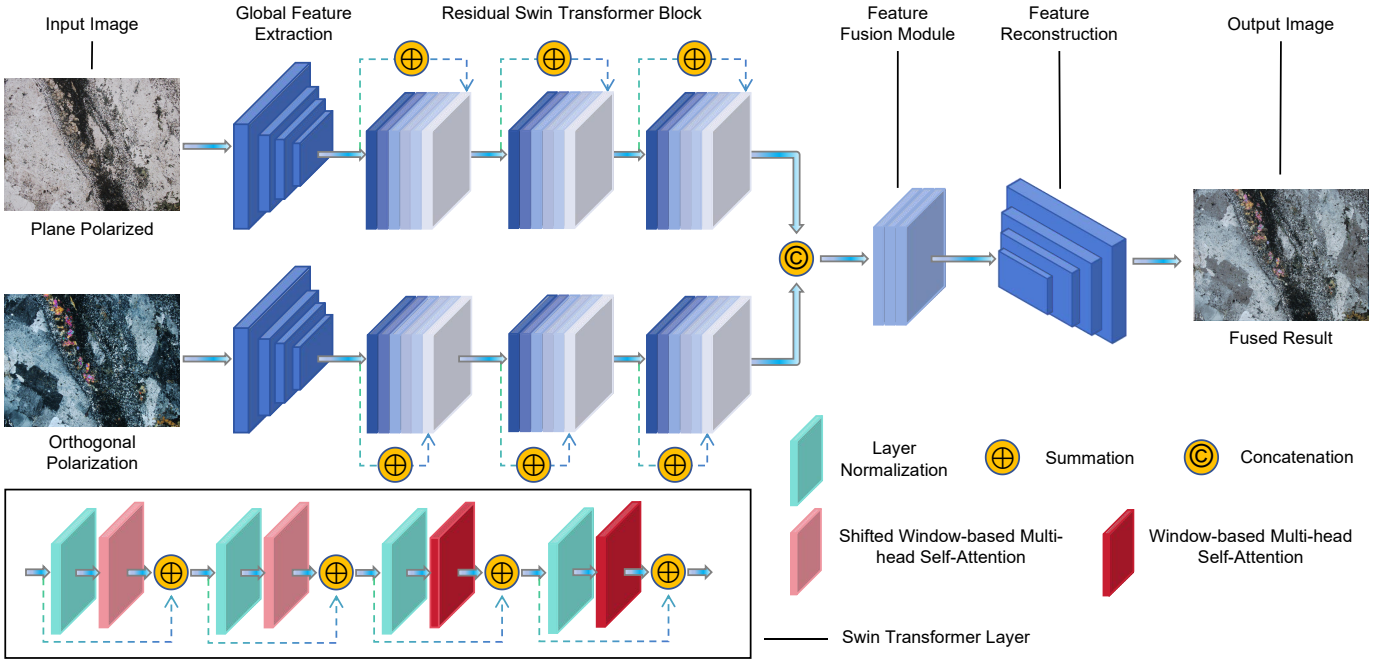

Figure 2. Structure of the proposed fast rock thin sections image fusion broker Transformer.

In the feature fusion module, the plane-polarized feature (P) is the primary feature, whereas the orthogonal polarization feature (C) serves as the auxiliary feature. The input features have been disassembled and mapped, with the main feature P mapped as Query and Value and the auxiliary feature C mapped as Key. The calculation of fusion attention is expressed in the following equation:

$$Q_p = X_p \cdot W^Q \tag{7}$$

$$K_c = X_c \cdot W^K \tag{8}$$

$$V_p = X_p \cdot W^V \tag{9}$$

$$Attention_1(Q_p, K_c, V_p) = Softmax\left(\frac{Q_s K_c^T}{\sqrt{d_K}} + B\right) \cdot V_p \tag{10}$$

where X_p and X_c represent plane-polarized and orthogonal polarization features, and W represents the corresponding feature mapping operation. The aforementioned attention calculation mechanism enables the auxiliary features to collaborate with the main features, thereby allowing the network to concentrate on the salient region of the auxiliary features and enhance the fusion effect. The process of fusing polarization information is essentially one of providing potential adaptations for another modality. Similarly, it is feasible to calculate the fusion attention ($Attention_2$) after the interchanging of modalities between the primary and auxiliary features. Following the acquisition of the two polarization features ($Attention_1$ and $Attention_2$), the fused features are output following channel-level splicing and serve as input to the decoder.

$$Attention_{Fused} = concat(Attention_1, Attention_2) \tag{11}$$

The decoder network comprises four cascaded convolutional layers with integrated 3 × 3 convolutional kernels, and BatchNorm is applied for normalization. The nonlinear

activation function is selected as LeakyReLU [27]. The decoder network receives the fused features as input and generates a fused image with identical spatial dimensions to the source image.

Both the broker attention designed in this paper and the linear attention that has been proposed attempt to reduce the computational complexity while maintaining the performance of the model. The construction of the two attention modules is shown in Figure 3. The linear attention employs a kernel-based self-attention mechanism and the associative property inherent in matrix products to reduce the complexity from quadratic to linear. Broker attention aims to measure the similarity between tokens using cosine distance and employs an intuitionistic fuzzy set-based soft matching algorithm to merge redundant tokens. In order to reduce the quadratic to linear complexity of Softmax attention [28], self-defined tokens are introduced as brokers between Query and Key. In contrast to the traditional token pruning technique, each amalgamated token encapsulates the information contained within the original tokens. Consequently, the model is able to reduce computational load effectively while experiencing minimal information loss.

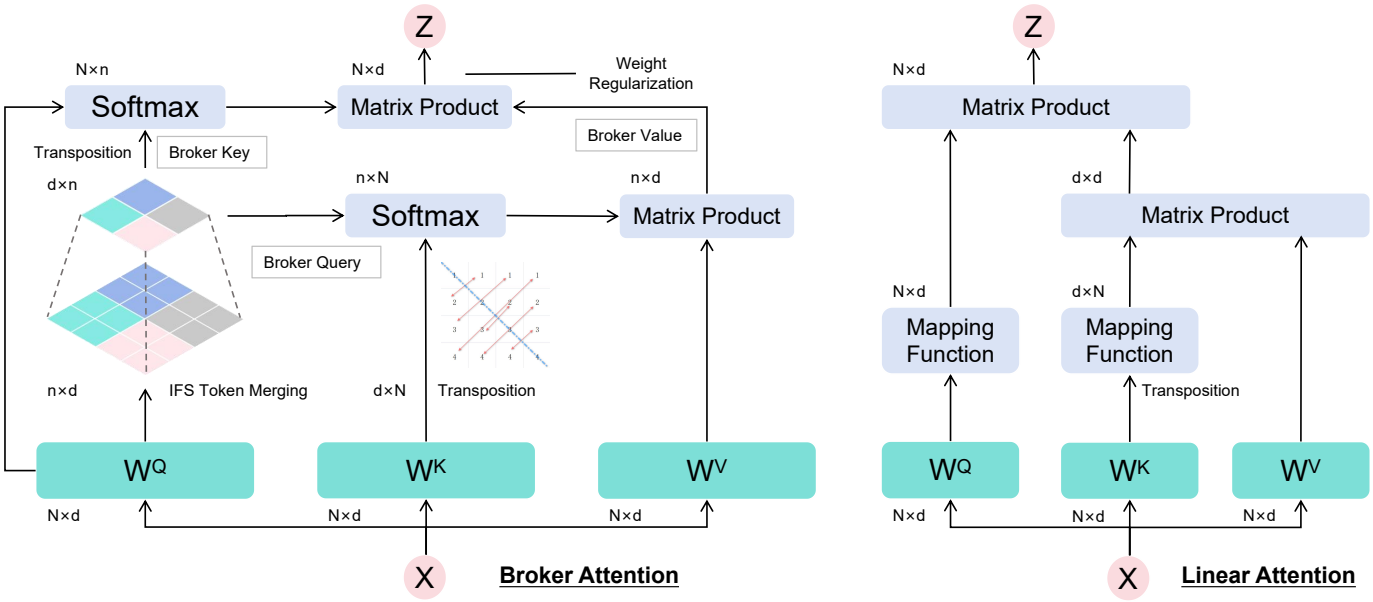

Figure 3. The diagrams on the **left** and **right** are respectively the schematic representations of the broker attention module and the linear attention module.

3.3. Design of Loss Function

The image fusion model is designed with the aim of maintaining the complementary information present in both the plane-polarized and orthogonal polarization images. At the pixel level, this paper employs the mean square error to quantify the pixel intensity discrepancy between the fusion outcome and the input. With regard to texture, the structural similarity loss serves to constrain the detailed information present in the fused image. The consolidation of tokens within the proposed novel attention paradigm may result in the expansion of some attention weights, as the merged token aggregates information from multiple original tokens. Conversely, a proportion of the weight may be reduced as a consequence of the original token information being either discarded or weakened during the merging process. In addition, in order to maintain the smoothness of tokens after merging, neighboring or similar tokens should have comparable attention weights after merging. Therefore, this paper encourages the attention weights to become sparser by Lasso sparsity regularization term [29]. The total variation smoothness regularization term encourages related tokens to maintain a smooth distribution of attention weights by calculating the differences between neighboring weight values and penalizing these differences.

The mean square error (MSE) [30] is employed as a metric for quantifying the discrepancy between the fused image and the reference at the pixel level. Where I_{fused} is the fused image, I_{ref} is the reference image, N is the total number of pixels, and i is the pixel index.

$$L_{pixel} = \frac{1}{N}\sum_{i=1}^{N}\left\|I_{fused,i} - I_{ref,i}\right\|^2 \tag{12}$$

A negative value of structural similarity (SSIM) is employed as a loss term to promote the fused image to exhibit a structural resemblance to the reference image.

$$L_{ssim} = 1 - SSIM(I_{fused}, I_{ref}) \tag{13}$$

The total loss of attention weights combines sparsity regularization and smoothness regularization in order to constrain and optimize the distribution of attention weights. Where L_{sparse} and L_{smooth} represent the sparsity and smoothness regularization terms, respectively. A is the attention weight matrix, h denotes the index of the attention head, i and j denote the index of the sequence position, and λ_1 and λ_2 are the weight coefficients of the sparsity and smoothness regularization, respectively.

$$L_{sparse} = \lambda_1 \sum_{h,i,j} |A_{hij}| \tag{14}$$

$$L_{smooth} = \lambda_2 \left(\sum_{h,i,j} (A_{h,i+1,j} - A_{hij})^2 + \sum_{h,i,j} (A_{h,i,j+1} - A_{hij})^2 \right) \tag{15}$$

$$L_{attention} = L_{sparse} + L_{smooth} \tag{16}$$

The total loss function is obtained by combining the aforementioned three loss terms, where α, β, and γ represent the weighting coefficients.

$$L_{total} = \alpha L_{pixel} + \beta L_{ssim} + \gamma L_{attention} \tag{17}$$

4. Experiments

4.1. Experiment Settings

In the multi-polarized image fusion task addressed in this paper, the dataset employed for model training and testing is derived from the 'Nanjing University Rock Teaching Thin Section Micrographic Image Dataset', provided by the Rock and Mining Department of the School of Earth Science and Engineering, Nanjing University. The rock thin section data have undergone a process of long-term loss, depletion, supplementation, and updating, resulting in the formation of a comprehensive and accurate thin section system. Concurrently, the dataset has been assembled through the application of electronic information processing techniques.

The dataset provides comprehensive coverage of the three principal categories of sedimentary, metamorphic, and igneous rocks. Each category contains detailed thin section images and micrograph datasets. The detailed information is shown in Table 1. The sedimentary rock dataset comprises 84 thin sections of 28 rock types, with a total of 699 images, including those of volcaniclastic rock, sandstone, mudstone, siltstone, greywacke, dolomite, siliceous rock, and evaporite. The igneous rocks comprise 120 thin sections of 40 rocks, with 963 images, and encompass a diverse range of rock types, including plutonic and extrusive rocks. The Metamorphic Rocks section also contains 120 thin sections of 40 rocks, comprising 972 images illustrating 17 fundamental categories of metamorphic rocks and their structures.

Table 1. Overview of the rock thin section dataset.

Rock Name	Categories	Sample Size	Sum of Micrographs
Sedimentary Rock	28	84	699
Igneous Rock	40	120	963
Metamorphic Rock	40	120	972

All rock thin section samples were prepared in accordance with the international standard thickness of 0.03 mm. The interference colors of the quartz grains observed in the same batch of rock flakes during micrographing and flake identification were all the primary interference color type. In order to ensure consistency between the colors observed visually and those captured by the system, the micrographs have been taken with automatic exposure and automatic white balance. The resolution of the micrographs is 1280 × 1024 pixels, and the images are uniformly saved in JPG format to ensure optimal quality and clarity.

The research was executed within a Windows 10 operating system environment, utilizing a desktop computer with 32 GB RAM, a Core i7-10700K CPU, and an NVIDIA RTX 3090 GPU. The experimental setup selected TensorFlow version 2.12.0, CUDA version 12.1, Python version 3.11, and cuDNN version 11.2.

4.2. Evaluation Metrics

Fusion assessment metrics are objective criteria responsible for evaluating the quality of the fused images. Multi-polarized image fusion requires multiple objective metrics for a comprehensive assessment of image fusion quality because of the lack of a reference image. The image quality evaluation metrics addressed in this paper can be broadly classified into two categories: one is the evaluation metrics based on the fused image for processing, and the other is the evaluation metrics based on the specific relationship between the fused image and the source image for processing.

The Cross Entropy (CE) [31] measure quantifies the information divergence between the fused image and the source. A lower CE value indicates greater consistency between the fused image and the source.

$$CE(I_a, I_b, I_f) = \frac{CE}{(I_a, I_f) + CE(I_b, I_f)^2} \tag{18}$$

$$CE(I_a, Ir_f) = \sum_{i=0}^{n} h_{I_a}(i) \log_2 \frac{h_{I_a}(i)}{h_{I_f}(i)} \tag{19}$$

$$CE(I_b, I_f) = \sum_{i=0}^{n} h_{I_b}(i) \log_2 \frac{h_{I_b}(i)}{h_{I_f}(i)} \tag{20}$$

The mutual information (MI) [32] is a measure of the similarity between the pixel distribution of the fused image and the source image. A larger value indicates a greater similarity between the two images.

$$I(X, Y) = \sum_{x,y} p_{XY}(x, y) \log_2 \frac{p_{XY}(x, y)}{P_X(x) P_Y(y)} \tag{21}$$

$$MI(I_a, I_b, I_f) = I(I_a, I_f) + I(I_b, I_f) \tag{22}$$

Peak signal-to-noise ratio (PSNR) [33], which measures the difference between the fused image and the source and the degree of detail preservation. A higher PSNR value indicates a superior quality of the fusion result.

$$PSNR(I,K) = 10\log_{10}(\frac{L^2}{MSE(I,K)}) \tag{23}$$

$$MSE(I,K) = \frac{1}{mn}\sum_{i=0}^{m}\sum_{i=0}^{n}\|I(i,j) - K(i,j)\|^2 \tag{24}$$

$$PSNR(I_a, I_b, I_f) = \frac{PSNR}{(I_a, I_f) + PSNR(I_b, I_f)2} \tag{25}$$

The root mean squared error (RMSE) [34] is a statistical measure that quantifies the similarity of the detailed information between the fused image and the source. A lower RMSE value indicates a smaller discrepancy between the two images, as well as a higher level of detail retention in the fused image.

$$RMSE(I_a, I_b, I_f) = \frac{RMSE}{(I_a, I_f) + RMSE(I_b, I_f)2} \tag{26}$$

$$RMSE(I_a, I_f) = \sqrt{\frac{1}{H*W}\sum_{i=1}^{H}\sum_{j=1}^{W}(I_a(i,j) - I_f(i,j))^2} \tag{27}$$

$$RMSE(I_b, I_f) = \sqrt{\frac{1}{H*W}\sum_{i=1}^{H}\sum_{j=1}^{W}(I_b(i,j) - I_f(i,j))^2} \tag{28}$$

The spatial frequency (SF) [35] is employed for the assessment of the fused image sharpness. The value of SF is positively correlated with the image quality, and the reference to the numerical results of the source map is unnecessary.

$$SF = \sqrt{RF^2 + CF^2} \tag{29}$$

$$RF = \sqrt{\frac{1}{HW}\sum_{i=1}^{H}\sum_{j=2}^{W}(I(i,j) - I(i,j-1))^2} \tag{30}$$

$$CF = \sqrt{\frac{1}{HW}\sum_{i=2}^{H}\sum_{j=1}^{W}(I(i,j) - I(i-1,j))^2} \tag{31}$$

The commonly appearing H and W in the above equation represent the width and height of the image, respectively. $I(i,j)$ denotes the pixel of the image at (i,j). I_a and I_b denote the source images of two modalities, and I_f implies the fused image.

The gradient-based fusion performance ($Q^{AB/F}$) [36] employs the Sobel operator to delineate local regions and assess the preservation of salient information and detail within these regions. A higher $Q^{AB/F}$ value indicates a superior fused image. Similarly, the Structural Similarity Index Measure (SSIM) [37], a metric based on structural similarity, is also employed as an evaluation metric. The derivation process of some of the functions is too cumbersome to be described in detail here.

In the evaluation of model lightweighting, the parameters encompass weights of convolutional kernels, scaling factors and shift parameters of batch normalization, biases, and so forth. The parameters serve as a pivotal metric for assessing model complexity, as they have a significant impact on the computational resources required and further influence the generalization and robustness of the model. Floating Point Operations Per Second (FLOPs) [38] is a metric used to quantify the number of floating-point operations required to be executed during a single forward pass of a model. A higher FLOPs value indicates that a greater number of floating-point operations need to be performed during a single forward pass, which consequently may result in lower computational efficiency. Memory footprint is one of the most important indicators of the effectiveness of model

lightweighting. Inference time refers to the time required for a model to complete a single forward pass given a specific input. A faster inference time implies that the model can accomplish the inference task in a shorter duration.

4.3. Visual Analysis

The training sets prepared in this paper have undergone precise registration and have been subjected to fusion testing on a variety of rock thin section images. Figure 4 presents the fusion results for mylonite, scapolite skarn, glaucophane schist, gneissic migmatite, and wollastonite skarn [39]. The two middle columns display the feature detection results derived from these two polarized light images. Each red circle represents a detected feature. For plane-polarized images, the features are uniformly distributed across the image. In contrast, due to the inherent optical characteristics of orthogonal polarization, the feature distribution exhibits a significant clustering pattern. The fifth column presents the checkerboard [40] diagrams of both, which aids in observing any discrepancies or shifts between the two types of images prior to their fusion and in assessing the continuity of structural lines by alternately displaying image patches within a gridded format. Precise fusion can only be achieved when the two images are perfectly aligned. The final column presents the ultimate outcome of the fused imagery. Regarding the fusion effectiveness, the model proposed in this paper successfully retains the respective feature information of both polarized light images within various types of rock data.

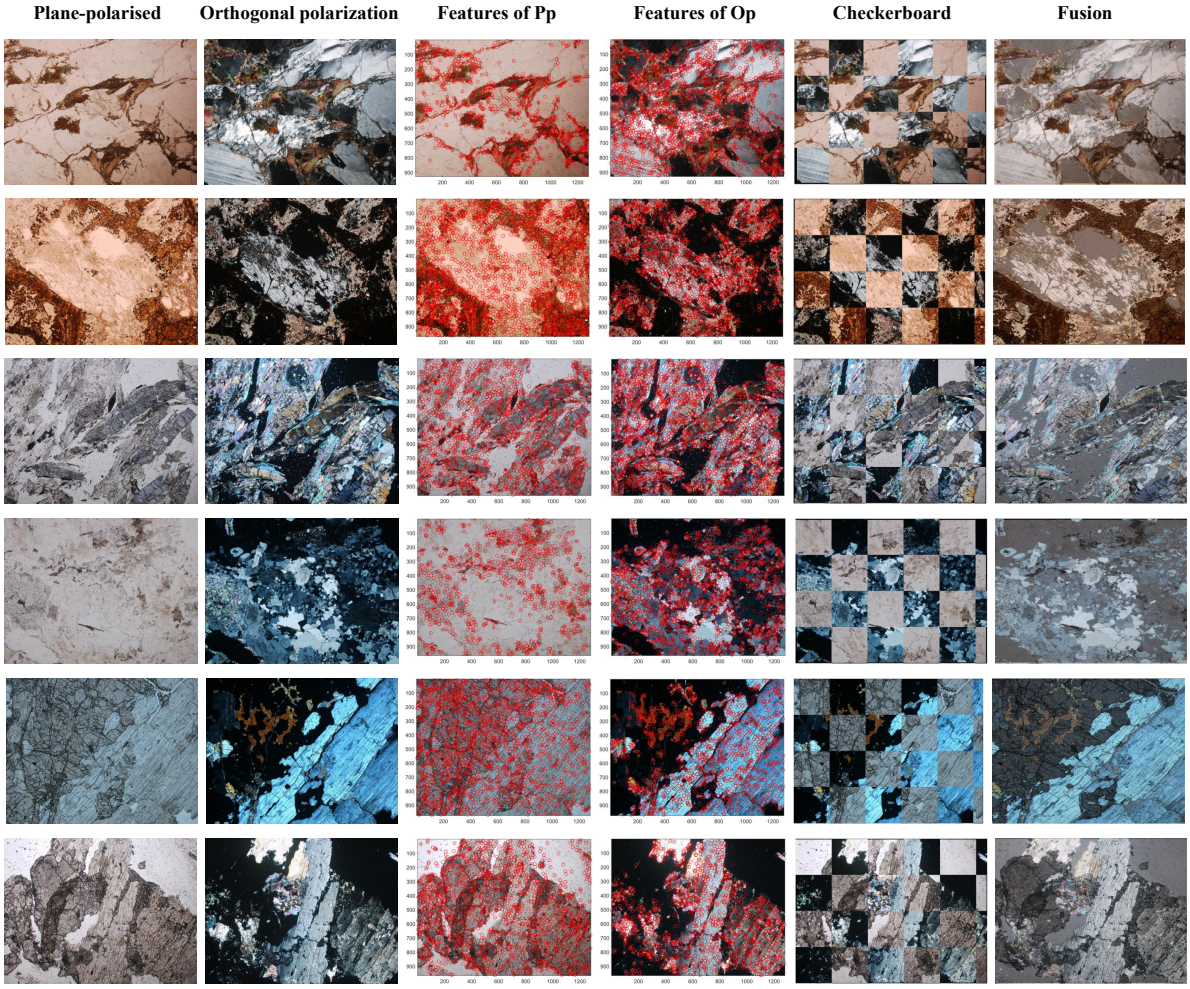

Figure 4. The demonstration of the fusion process of various types of rock thin section images. Each row represents a set of rock data, while each column corresponds to an image category. "Pp" and "Op" are abbreviations for "plane-polarized" and "orthogonal polarization", respectively. The small red circles represent feature markers that have been detected.

Following an analysis of the current research landscape on multi-polarization image fusion, this paper selects seven advanced algorithms (Nestfuse [41], SEDRFuse [42], DDcGAN [43], DenseFuse [44], DIDFuse [45], U2Fusion [46], and STDFusion [47]) published in various international journals and conferences for comparison with the proposed model. Figure 5 showcases a subjective comparison of the results obtained by the proposed model against those of the other comparative methods on the same test set. Furthermore, significant objects are annotated with red bounding boxes to facilitate a more intuitive analysis of the subjective outcomes. The contents of the larger box are a magnified view of the smaller box.

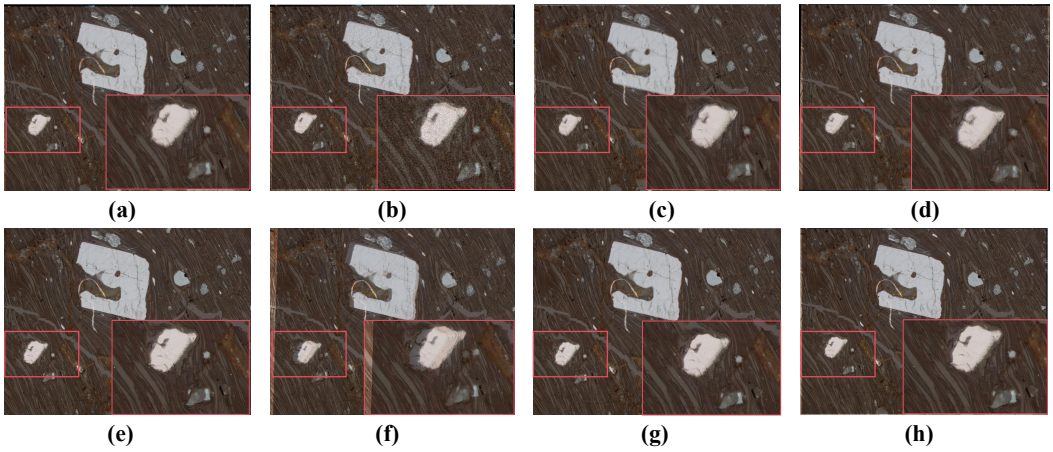

Figure 5. The fusion results of images of dacitic crystal–lithic–vitric welded tuff by different models. (**a**) Nestfuse. (**b**) SEDRFuse. (**c**) DDcGAN. (**d**) DenseFuse. (**e**) DIDFuse. (**f**) U2Fusion. (**g**) STDFusion. (**h**) Our proposed model. The small red boxes are areas of significant difference that have been selected. The larger box is a zoomed-in display of the area, for a clearer comparison of the fusion effect.

Apart from determining the accuracy of the edge structural information in the images, it can also be observed in Figure 6 whether the fine textures in the rock images can be completely retained after fusion. Based on the fusion results obtained from multi-polarized rock thin section images, the fusion method proposed in this paper is found to profoundly retain textural and detailed features while simultaneously achieving a more natural and clearer fusion effect. The fusion accuracy of this method is comparable to that of other superior fusion models trained using Transformer architectures. In the outcomes of multi-polarizing light rock thin section image fusion, it was found that the fusion method proposed in this paper not only preserves the textural details with emphasis but also achieves a more natural and clearer fusion effect. The fusion accuracy is comparable to that of other outstanding fusion models.

To evaluate the performance of the proposed fast fusion Transformer model on datasets with different resolutions, experiments were conducted on high-resolution (1280 × 1024) and low-resolution (480 × 384) petrographic thin section images. The low-resolution images were generated by artificially downscaling the high-resolution images, and multi-polarized image fusion was performed on both resolutions. As shown in Figure 7, the proposed model achieves effective feature fusion for both high-resolution and low-resolution data. The fused images retain critical information from petrographic thin sections under plane-polarized and orthogonal polarization light, with reconstructed texture details and characteristic features. However, a comparison of the fusion results indicates the presence of noise in the outputs for low-resolution images.

During the compression process, low-resolution images inevitably lose some microstructural textures and edge details, which hinders the model's ability to accurately capture the complete information of petrographic thin sections during feature extraction and matching. Furthermore, as the proposed model employs a soft matching mechanism

based on intuitionistic fuzzy sets, the reduced information redundancy in low-resolution images increases the likelihood of asymmetric dependencies between generated artificial tokens and original tokens. This asymmetry decreases the stability of the feature fusion process between high-resolution and low-resolution images, introducing noise into the results.

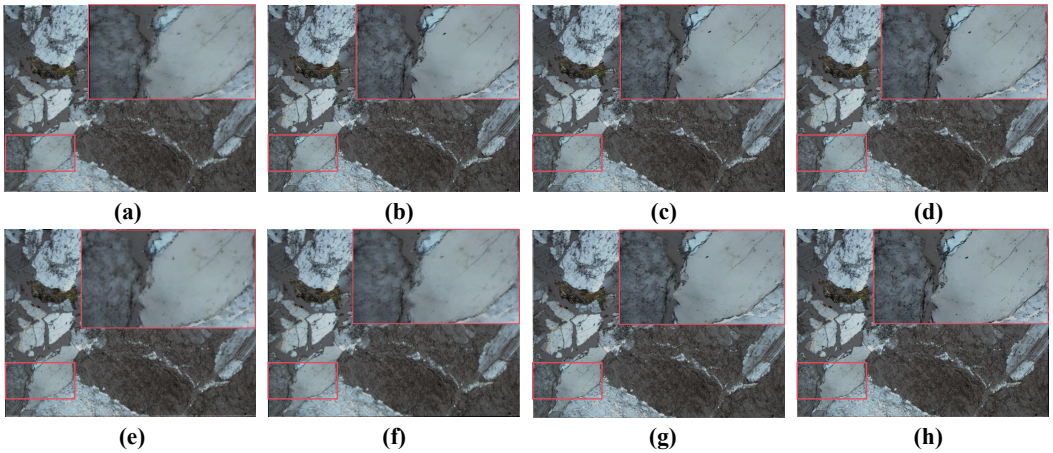

Figure 6. The fusion results of granite images by different models. (**a**) Nestfuse. (**b**) SEDRFuse. (**c**) DDcGAN. (**d**) DenseFuse. (**e**) DIDFuse. (**f**) U2Fusion. (**g**) STDFusion. (**h**) Our proposed model. The small red boxes are areas of significant difference that have been selected. The larger box is a zoomed-in display of the area, for a clearer comparison of the fusion effect.

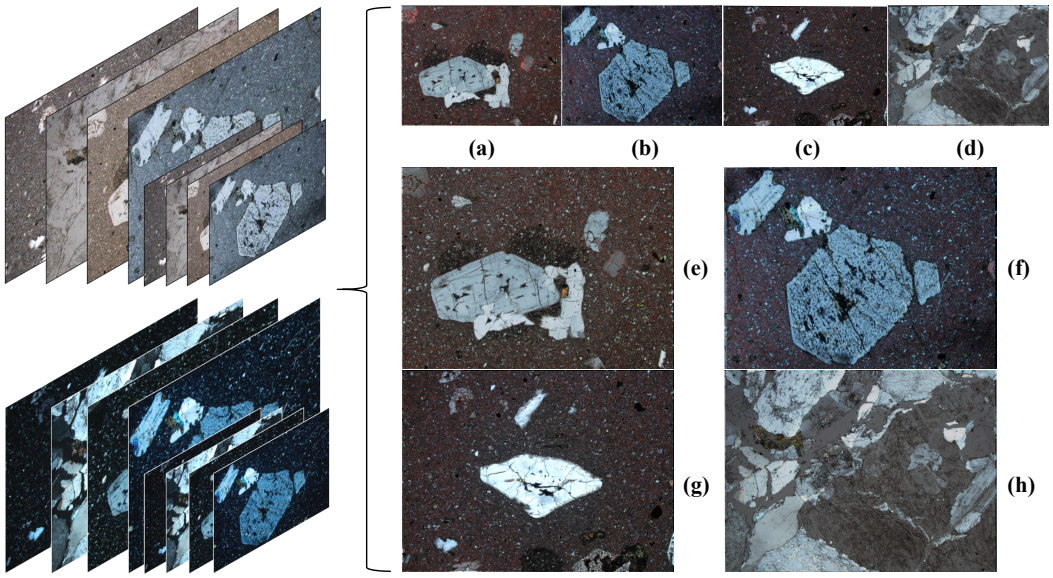

Figure 7. Fusion results for high- and low-resolution images: (**a**–**d**) show fused images with a resolution of 480 × 384, while (**e**–**h**) show fused images with a resolution of 1280 × 1024.

Feature matching experiments were conducted on high- and low-resolution datasets, with correctly matched features indicated by red lines in the images. As shown in Figure 8, the model achieves a higher number of correct feature matches in the high-resolution scenario. This improvement stems from the model's design. The soft matching algorithm based on intuitionistic fuzzy sets effectively preserves critical feature relationships, which is more pronounced in high-resolution images where fine-grained details are abundant. Additionally, the lightweight fusion strategy utilizing artificial tokens enhances feature alignment, particularly in high-resolution data, where richer texture and edge information are available. These design elements enable the model to achieve better feature matching accuracy in high-resolution settings.

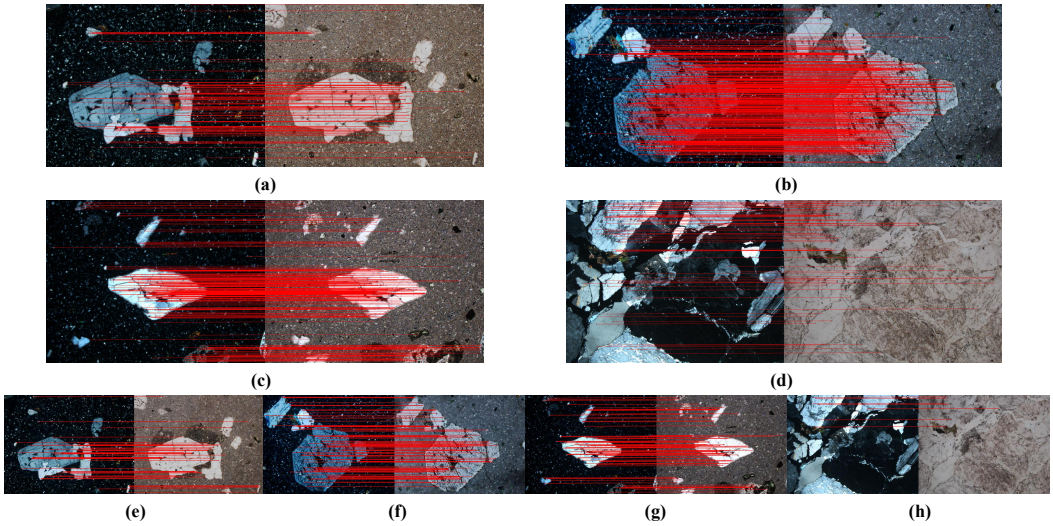

Figure 8. Feature matching results: (**a**–**d**) show images with a resolution of 1280 × 1024, while (**e**–**h**) represent images with a resolution of 480 × 384. Red lines indicate correctly matched feature pairs.

4.4. Quantitative Assessments

In the assessment of model compression, both size and time complexity are pivotal considerations. However, in terms of memory and inference time, additional attention must be given to the impact of GPU acceleration. The design of GPUs is inclined to provide optimal memory bandwidth, which is particularly crucial when dealing with Transformer models, as these types of models typically occupy large contiguous blocks of memory space. GPUs are capable of efficiently utilizing memory resources, thereby reducing memory footprint to a certain extent. Furthermore, through GPU acceleration, inference times can be significantly enhanced, with reductions from seconds to milliseconds. We selected 29 groups of images from the dataset as the test set and made comparisons on the four aforementioned model compression evaluation metrics, respectively. The average results of each group of metrics on the 29 test images are presented in Table 2.

In comparison with other state-of-the-art models, the model proposed in this paper attains optimality in FLOPs and inference time. This implies that, under the same hardware conditions, the model can accomplish the forward-propagation process more quickly, thus reducing the consumption of computational resources. A faster inference speed is conducive to shortening the overall processing time.

Table 2. The parameters, FLOPs, model sizes, and inference times of eight fusion methods are presented. Among them, the best results are indicated in red font, while the second-best results are denoted in blue.

Model	Parameters	FLOPs (G)	Memory (M)		Inference Time (S)	
			GPU-Acc	Indep	GPU-Acc	Indep
Nestfuse	2,732,761	706.027	8.981	10.931	1.838	2.272
SEDRFuse	40,153	22.028	0.127	0.159	0.141	0.159
DDcGAN	1,096,257	820.201	4.278	5.284	0.861	1.122
DenseFuse	74,193	45.475	0.226	0.297	0.135	0.247
DIDFuse	44,547	26.085	0.152	0.179	0.228	0.257
U2Fusion	659,217	404.722	1.846	2.637	0.875	1.163
STDFusion	886,901	538.102	2.648	3.311	0.773	1.079
Our	6,771,824	4.525	13.448	18.051	0.018	0.026

In Table 3, we have compared the accuracies of different models. During the model training process, due to the multi-angular nature of the orthogonal polarization image data, the training results of 0-degree deflection and multiple deflection angles are tested separately. Among the seven evaluation metrics, MI and SF achieve the best results in the single-angle case. This indicates that there is a high degree of pixel distribution similarity between the fused image and the source images. RMSE and $Q^{AB/F}$ achieve the second-best results, meaning that the fused image well preserves the detailed features and structural information of the source images and minimizes information distortion as much as possible. Concurrently, although the results for CE, PSNR, and SSIM have reached commendable levels, they still do not match the exceptional performance demonstrated by traditional fusion networks. This is attributed to the fact that Transformer models tend to focus more on global information and relatively significant regions within the source images while giving insufficient attention to all local areas of the entire image. Consequently, when the training set is relatively small, Transformers may exhibit deficiencies in capturing local or regional features. However, it can be seen that these deficiencies have been significantly improved when the data are expanded with multi-angle images.

Table 3. Comparison of the mean values of the algorithm's objective metrics in single-angle and multi-angle orthogonal polarization training sets. The numbers in red font represent the best fusion performance, while blue represents the next best.

Model (Single-angle)	**CE**	**MI**	**PSNR**	**RMSE**	**SF**	**SSIM**	**$Q^{AB/F}$**
Nestfuse	1.648	2.280	58.287	0.099	9.695	1.479	0.404
SEDRFuse	1.215	1.519	58.135	0.088	10.248	1.425	0.432
DDcGAN	0.963	1.456	54.742	0.219	13.109	1.118	0.343
DenseFuse	1.652	2.392	57.972	0.092	6.763	1.357	0.497
DIDFuse	1.538	2.245	58.517	0.098	15.821	1.482	0.328
U2Fusion	1.494	1.678	55.368	0.177	11.394	1.519	0.465
STDFusion	2.227	1.432	59.859	0.081	11.612	1.474	0.579
Our	1.571	2.665	58.438	0.085	17.278	1.431	0.506
Model (Multi-angle)	CE	MI	PSNR	RMSE	SF	SSIM	$Q^{AB/F}$
Nestfuse	1.481	2.813	62.357	0.089	16.693	1.536	0.485
SEDRFuse	1.196	1.976	61.682	0.083	18.271	1.447	0.449
DDcGAN	0.924	1.542	57.194	0.175	13.835	1.285	0.482
DenseFuse	1.385	2.981	59.528	0.081	14.516	1.401	0.557
DIDFuse	1.259	2.405	65.741	0.088	25.729	1.592	0.368
U2Fusion	1.263	1.864	58.903	0.117	22.347	1.554	0.509
STDFusion	1.827	1.412	59.265	0.076	19.652	1.538	0.612
Our	1.108	3.158	63.479	0.081	22.174	1.569	0.673

To verify the impact of data types on the model during the training process, three groups of datasets, namely Sedimentary, Metamorphic, and Igneous, are selected in this paper to train the same model. The objective average execution time results are presented in Table 4. Due to the advantage in the order of magnitude, the models trained under Metamorphic and Igneous have an edge in inference speed. However, whether it is a classified or a mixed training set, the model proposed in this paper can achieve the shortest inference time.

To evaluate the proposed model, Table 5 shows the differences in fusion quality between high- and low-resolution scenarios, calculated as the subtraction of their performance metrics, highlighting the model's superiority in high-resolution data handling. Values in red are the minimum difference, and in blue are suboptimal. In the task of image fusion across high- and low-resolution scenarios, metrics that rely on fine details and texture information (e.g., MI, PSNR, SF, and $Q^{AB/F}$) tend to exhibit significant differences between the two resolutions. In contrast, metrics that assess global consistency and structural sim-

ilarity (e.g., CE and SSIM) show relatively smaller variations. High-resolution scenarios place greater demands on detail preservation and local feature extraction, requiring models to possess stronger capabilities for feature capture and reconstruction. Accordingly, the proposed model demonstrates robust performance when applied to petrographic thin section images at varying resolutions.

Table 4. The average inference timetable (in seconds) of the other seven fusion methods and the model proposed in this paper on the dataset composed of sedimentary rocks, metamorphic rocks, and igneous rocks are presented. The numbers in red font represent the shortest reasoning time, and blue represents the next best.

Model	Sedimentary	Metamorphic	Igneous	Mixed
Nestfuse	2.953	2.045	1.791	2.272
SEDRFuse	0.288	0.175	0.157	0.159
DDcGAN	1.683	1.010	0.898	1.122
DenseFuse	0.371	0.272	0.148	0.247
DIDFuse	0.285	0.283	0.256	0.257
U2Fusion	1.745	1.047	0.930	1.163
STDFusion	1.618	0.971	0.863	1.079
Our	0.039	0.029	0.021	0.026

Table 5. High–low resolution fusion quality differences (high-resolution–low-resolution). Numbers in red font represent the smallest difference in fusion performance of the model between the two resolutions of data, and blue represents sub-optimal.

Method	CE	MI	PSNR	RMSE	SF	SSIM	$Q^{AB/F}$
NestFuse	0.262	0.425	−6.065	**0.004**	**5.054**	−0.132	−0.249
SEDRFuse	0.217	0.588	−5.024	0.014	7.483	−0.112	−0.052
DDcGAN	0.147	**0.295**	−3.507	0.008	9.283	−0.224	−0.044
DenseFuse	**0.003**	1.120	−1.274	0.006	15.956	**−0.025**	**−0.025**
DIDFuse	**0.004**	1.110	−1.167	0.007	14.795	−0.028	−0.048
U2Fusion	0.195	0.757	−2.304	0.045	7.476	−0.035	−0.086
STDFusion	0.005	1.125	**−1.159**	0.016	9.874	−0.030	−0.037
Our	0.006	**0.325**	**−0.875**	**0.002**	**2.264**	**−0.024**	**−0.029**

4.5. Ablation Study

In this section, a series of ablation experiments were conducted on several key modules of the fusion model, with the objective of verifying the significance of the employed modules and their associated settings. In the design of the attention module, we compared our broker attention with the Softmax and linear attention methods using the Swin Transformer. In model compression, linear attention has been widely recognized for its effectiveness in application as it improves the processing speed by reducing the computational complexity. The traditional Softmax attention can capture the dependency relationships between elements more precisely. To verify the potential of broker attention, this paper explores the correlation between the model output and the Dice coefficient. Based on the joint area of the fused image and the target, the average value of the model output is calculated. Subsequently, the average absolute value is subtracted from 1 to obtain the normalized value. Figure 9 provides evidence of whether there is a positive correlation between each mechanism and Dice. Softmax attention shows a significant positive correlation due to its powerful global performance. The broker attention proposed in this paper achieves a performance improvement over linear attention by enhancing the model's ability to capture global context information.

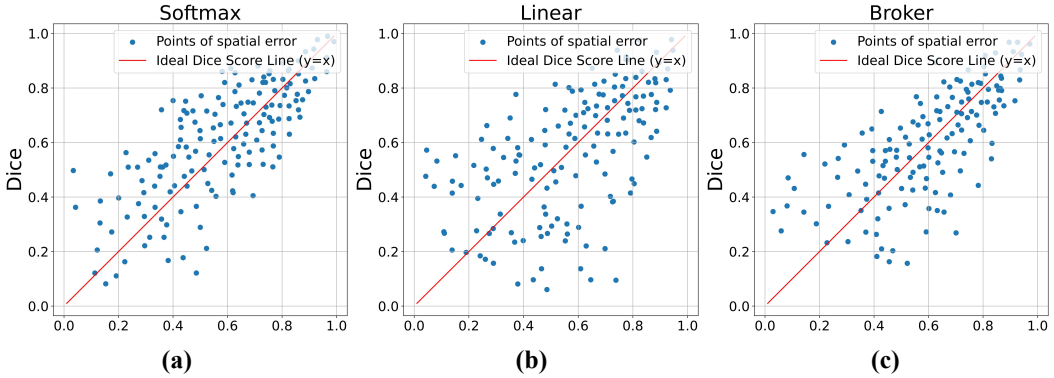

Figure 9. Correlation between the estimated spatial error and the Dice coefficient in three attention mechanisms: (**a**) Softmax, (**b**) Linear, and (**c**) Broker.

Capturing global information as much as possible is a common challenge that most lightweight models have to face. In terms of the selection of the Transformer backbone, ablation experiments on different backbone networks have also been conducted in this paper. In Table 6, the Swin Transformer adopted in this paper exhibits the highest performance in the metrics of CE, MI, RMSE, SSIM, and $Q^{AB/F}$. It also achieves the second-best results in PSNR and SF. This is attributed to the fact that the Swin Transformer further optimizes the attention by introducing the shift mechanism (introducing cyclic shift-rearrangement across spatial positions in the Swin Block) and masking operations, enabling it to capture the relationships between long-distance dependencies more effectively. The local window attention maintains computational efficiency and can effectively capture both local and global information in the image.

Table 6. The ablation experiments on applying broker Attention to different backbones. The numbers in red font represent the best fusion performance, while blue represents the next best.

Backbone	CE	MI	PSNR	RMSE	SF	SSIM	$Q^{AB/F}$
DeiT	2.072	1.346	59.198	0.192	20.413	1.562	0.535
PiT	2.253	1.707	51.815	0.154	10.258	1.489	0.648
ViL	1.519	2.182	55.276	0.227	32.739	1.445	0.479
CrossFormer	1.595	2.031	64.963	0.098	16.384	1.214	0.421
DW-ViT	1.628	2.474	61.351	0.085	21.506	1.429	0.603
Swin-T	1.108	3.158	63.479	0.081	22.174	1.569	0.673

Ablation experiments were also conducted on the impact of replacing Softmax attention with broker attention at different stages since the computational complexity of the model can be adjusted by changing the number of broker tokens. The evaluation metrics involved in Table 7 include NFM (number of feature matches), NBT (number of broker tokens), as well as parameters and FLOPs. In order, the four stages are Emb (embedding), DownS (down-sampling), Enh (enhancement), and Int (integration). In Stage 1, the input image is partitioned into patches, and a linear embedding operation is performed on these patches. In Stage 2, adjacent patch features are merged through the Patch Merging layer, thereby reducing the resolution and increasing the number of channels. Stage 3 continues to use the combination of Patch Merging and the Swin Transformer Block, but at this time, the size of the feature map being processed has decreased and the number of channels has increased. Stage 4 conducts feature integration and output generation operations.

Table 7. The ablation experiments on applying the broker attention module to the Swin Transformer at different stages.

Broker Attention				NFM	NBT	Parameters	FLOPs
Emb-1	DownS-2	Enh-3	Int-4				
●	○	○	○	426.978	44.492	6,771,824	4.525
●	●	○	○	309.216	188.954	6,771,824	4.525
●	●	●	○	251.597	297.109	6,771,824	4.525
●	●	●	●	230.638	318.765	6,771,824	4.525

It can be observed that an increased number of substitutions leads to a diminution in the number of detected features, accompanied by a concurrent reduction in the quantity of broker tokens. However, the subsequent impact of these changes is less significant than those observed in the initial stages. Furthermore, the decrease in the number of broker tokens within the model does not adversely affect its compression performance. Conversely, a higher number of broker tokens corresponds to a lesser quantity of feature matches, indicating an inverse proportional relationship between the two. Consequently, to ensure the quality of image fusion, the present study confines the application of token merging to the first three stages only.

There are multiple approaches for obtaining broker tokens. The commonly used ones include D-Points (deformed points) [48], pooling, and token merging. We have selected BSM (Bipartite Soft Matching) [49], K-means [50], Greedy-M [51] (Greedy matching based on attention weights), and P-Sampling [52] (Progressive Sampling) to participate in the ablation experiments for comparison with the IFS token merging proposed in this paper. The results of the ablation experiments are shown in Table 8. The IFS token merging approach demonstrates superior performance in terms of both FLOPs and inference time while also exhibiting a comparable algorithmic footprint relative to other methodologies. This superiority is primarily attributed to the necessity of computing global cosine distances during the token matching process within IFS. However, this computational burden is mitigated by the multidimensional decision-making advantages of the IFS, thereby not imparting undue adverse effects on the overall performance.

Table 8. The ablation experiments on the impact of broker token acquisition methods on the compression of Transformer. Numbers in red font represent the optimal modeling efficiency, and blue represents the sub-optimal.

Access	Parameters	FLOPs	Memory	Inference Time
BSM	6,515,781	7.941	12.714	0.174
K-means	6,798,011	15.275	22.803	0.859
Greedy-M	6,530,109	12.639	12.598	0.268
Pooling	6,525,065	37.157	8.923	0.123
D-Points	6,886,841	347.482	52.287	1.972
P-Sampling	6,502,259	15.386	11.135	0.517
IFS	6,771,824	4.525	13.448	0.018

To assess the influence of loss functions on image fusion performance, an ablation study was conducted on L_{pixel}, L_{ssim}, $L_{pixel} + L_{ssim}$, and the L_{total} loss employed in this paper. The outcomes of this analysis are depicted in Figure 10. On the vertical axis, y represents the specific numerical value corresponding to the cumulative probability on the horizontal axis. When the cumulative probability on the horizontal axis reaches x%, the y-value on the vertical axis indicates the maximum value that satisfies this cumulative condition. The x% on the horizontal axis denotes the proportion of observed values that are less than or equal to the metric value among all observed values. The legends within the figure explain the loss functions represented by the lines of different colors and markers.

From the ablation results, the fusion model achieved the highest average values in all metrics with the loss function proposed in this paper. In particular, significant improvements were shown in PSNR, CE, and RMSE. This indicates that our fusion results contain richer information, exhibiting high resolution and abundant feature details.

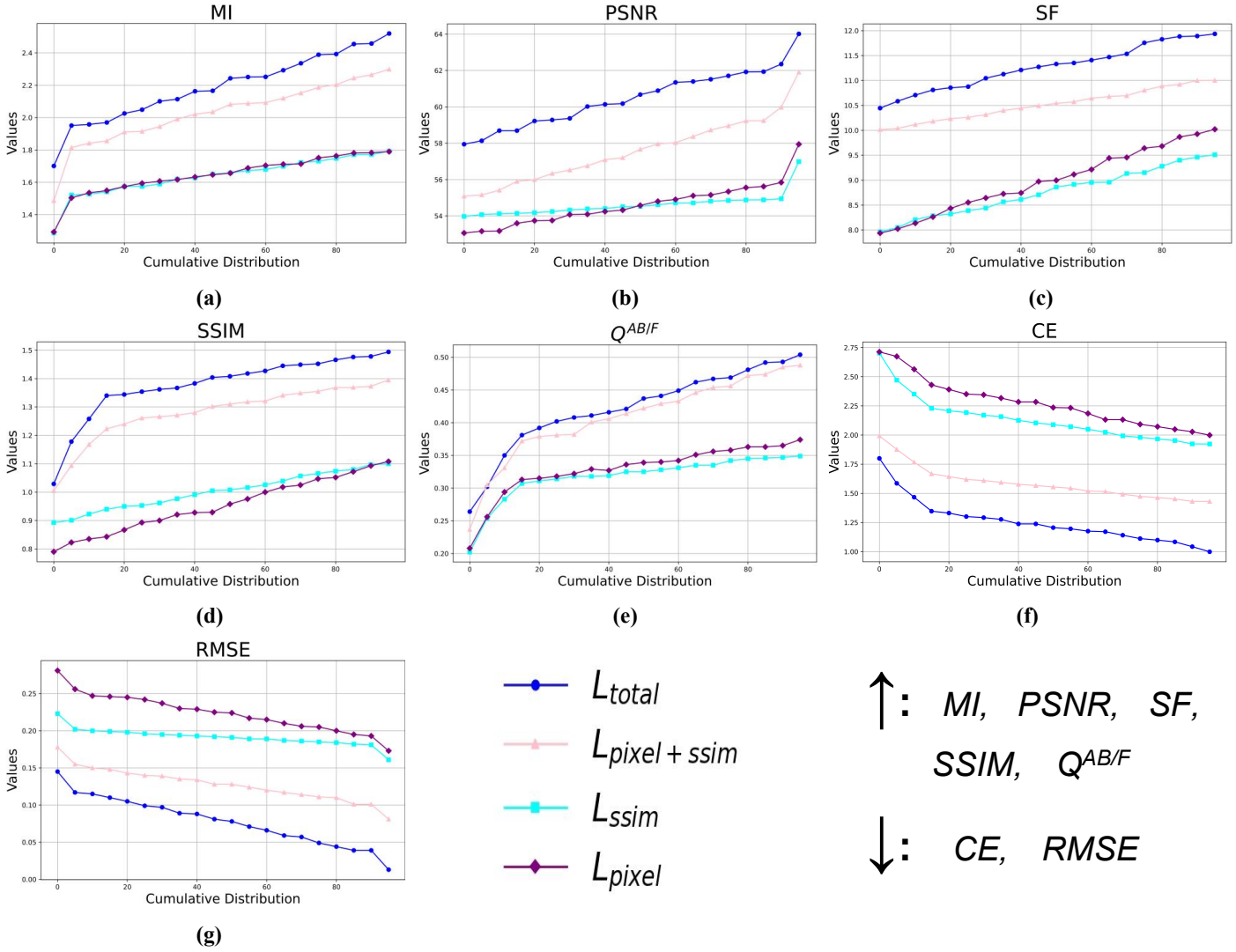

Figure 10. Comparison of cumulative probability distributions for different loss functions on image fusion performance. The metrics represented by each graph are: (**a**) MI, (**b**) PSNR, (**c**) SF, (**d**) SSIM, (**e**) $Q^{AB/F}$, (**f**) CE, and (**g**) RMSE.

5. Discussion

In addition to its application in the fusion of multi-polarized petrographic images, we have also explored the use of the proposed fast fusion Transformer in the field of medical imaging, specifically for mouse chemical exchange saturation transfer (CEST) MRI. CEST MRI is a non-invasive imaging technique that provides valuable information about tissue composition and metabolism, which is crucial for biomedical research and clinical applications. In CEST MRI, a primary challenge lies in the fusion of images acquired at different time points. As the water saturation within lesions changes over time, CEST MRI images obtained at different time intervals often exhibit variations in contrast and water saturation characteristics. This temporal variability complicates image fusion, as traditional fusion methods typically struggle to handle lesion features that evolve over time. To address this challenge, we have attempted to apply the model proposed in this study to fuse CEST MRI sequences acquired at different time points, with the aim of better

capturing the dynamic changes in lesions over time, thereby enhancing the extraction of relevant tissue features.

Our experimental results demonstrate that the proposed fusion algorithm significantly improves the representation of tissue features in mouse CEST MRI scans. Figure 11 presents the fusion results of the CEST MRI series. As identifying the optimal imaging results heavily relies on expert knowledge, the fusion of all image features helps to mitigate this limitation. The algorithm effectively combines high-contrast regions of interest from different time points, preserving important features that change over time while reducing unnecessary noise and background interference. By leveraging intuitionistic fuzzy set-guided fusion, we are able to more accurately capture the dynamic changes in water saturation within lesions, which is crucial for studying metabolic activity and tissue composition. Therefore, the work presented in this study not only shows promise for rock thin section image fusion but also has broad potential applications in biomedical imaging.

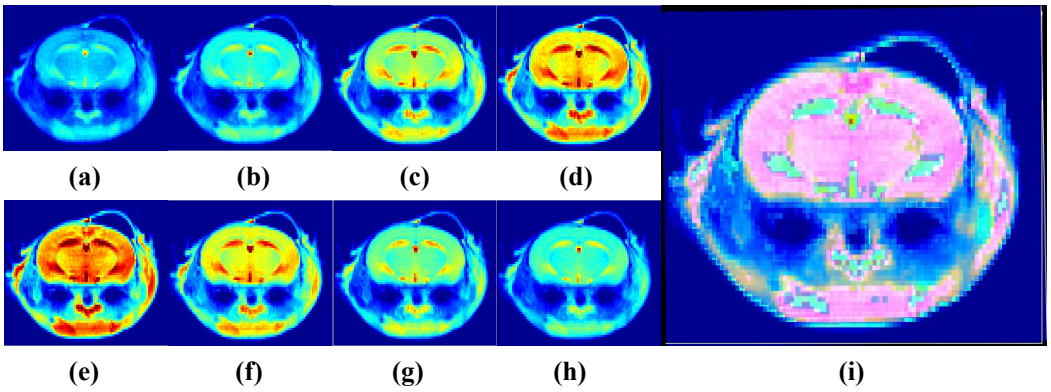

Figure 11. Panels (**a**–**h**) represent the CEST MRI images acquired at saturation durations of 17, 25, 33, 52, 60, 68, 76, and 84 min, respectively. Panel (**i**) shows the output result obtained by fusing this series of images.

The technology of image fusion plays a significant role in various application domains, including mineral resource exploration, geological structural analysis, medical imaging, and computer vision research, which undoubtedly constitute integral components of the development of intelligent manufacturing. However, this work still faces several issues and challenges that necessitate further research. In the fast process of feature fusion, a critical consideration is how to minimize the information loss incurred during the token merging. When dealing with numerous tokens, controlling the complexity of the matching algorithm remains a challenge. How should we address the variability in multi-polarized images and devise corresponding cross-polarization fusion strategies? Additionally, proposing more efficient approaches for model training and enhancing the creation of petrographic thin section datasets are imperative. These are the focal points that must be considered and resolved in future tasks related to multi-polarized petrographic thin section image fusion.

6. Conclusions

This paper presents a fast fusion Transformer guided by intuitionistic fuzzy set for multi-polarized petrographic images of rock thin sections. In this approach, we employ intuitionistic fuzzy set-based soft matching to merge tokens and generate novel intermediary tokens, addressing the limitations of asymmetric dependencies between tokens in petrographic features. The integration of plane-polarized and orthogonal polarization rock thin section images is achieved by replacing the traditional softmax function with a broker-attention in the Transformer. Based on this framework, a new loss function is designed to effectively integrate the deep features of both image types. The rationality of the proposed method is verified through ablation experiments, while its effectiveness is demonstrated through comparative tests. The image fusion results exhibit notable advantages in preserving the morphological, color, and edge features of plane-polarized

images, as well as the deep-level features such as interference colors, relief, and cleavage in orthogonal polarization images. Extensive fusion experiments conducted on datasets of various rock types have confirmed that the proposed image fusion method is comparable to current state-of-the-art methods in terms of both subjective visual evaluation and objective metrics. Moreover, it demonstrates a significant advantage in fusion speed. The lightweight model design of the proposed method provides the potential for its deployment on mobile devices.

Author Contributions: Conceptualization, B.C.; methodology, B.C.; software, W.H. and A.W.; validation, B.C.; formal analysis, B.Y.; investigation, B.Y.; resources, W.W. and W.H.; data curation, W.W. and L.P.; writing—original draft preparation, B.C.; writing—review and editing, B.C. and L.C.; visualization, B.Y. and Y.W.; supervision, L.C.; project administration, L.C.; funding acquisition, L.C. and B.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Scientific Research Program Funded by Education Department of Shaanxi Provincial Government (Program No. 23JP027) and the Key Research and Development Program of Shaanxi (Program No. 2024GX-YBXM-555).

Data Availability Statement: The details and code related to the content discussed in this study can be obtained by contacting the first author.

Acknowledgments: We would like to express our gratitude to the authors of Nestfuse, SEDRFuse, DDcGAN, DenseFuse, DIDFuse, U2Fusion, and STDFusion for sharing their code. This greatly assisted us in comparing our experiments. We also thank the anonymous reviewers for their insightful and valuable suggestions, which undoubtedly improved the quality of our paper.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Du, S.; Huang, C.; Ma, X.; Fan, H. A Review of Data-Driven Intelligent Monitoring for Geological Drilling Processes. *Processes* **2024**, *12*, 2478. [CrossRef]
2. Hou, D.; Han, G.; Chen, S.; Zhang, S.; Liang, X. A Study on a Novel Production Forecasting Method of Unconventional Oil and Gas Wells Based on Adaptive Fusion. *Processes* **2024**, *12*, 2515. [CrossRef]
3. Zhu, S.; Du, Q.; Dong, C.; Yan, X.; Wang, Y.; Wang, Y.; Wang, Z.; Lin, X. Reservoir Characteristics and Controlling Factors of Large-Scale Mono-Block Gas Field Developed in Delta-Front Sandstone—A Case Study from Zhongqiu 1 Gas Field in the Tarim Basin. *Minerals* **2023**, *13*, 1326. [CrossRef]
4. Meng, N.; Xiao, Q.; Li, W. Elemental Geochemistry and Pb Isotopic Compositions of the Thick No. 7 Coal Seam in the Datun Mining Area, China. *Minerals* **2024**, *14*, 848. [CrossRef]
5. Zhang, D.; Liu, Y.; Dong, G.; Liu, B.; Li, C.; Zeng, X. Study on the Pore Structure Characterization of the Limestone Reservoir of the Taiyuan Formation in the Ordos Basin. *Energies* **2024**, *17*, 3275. [CrossRef]
6. Tang, L.; Chen, Z.; Huang, J.; Ma, J. CAMF: An Interpretable Infrared and Visible Image Fusion Network Based on Class Activation Mapping. *IEEE Trans. Multimedias* **2024**, *26*, 4776–4791. [CrossRef]
7. Yuan, Y.; Zhang, K.; Wu, Q.; Burokur, S.N.; Genevet, P. Reaching the efficiency limit of arbitrary polarization transformation with non-orthogonal metasurfaces. *Nat. Commun.* **2024**, *15*, 6682. [CrossRef]
8. Rehman, H.U. Zircon Internal Deformation and Its Effect on U-Pb Geochronology: A Case Study from the Himalayan High-Pressure Eclogites. *Minerals* **2024**, *14*, 742. [CrossRef]
9. Feng, S.; Hu, Y.; Hu, D.; Li, Y.; Tan, Z.; Hu, R. Intelligent Classification of Rocks in Mountain Highway Tunnels Using ISSA-ELM Model. *Geotech. Geol. Eng.* **2024**, *42*, 7385–7405. [CrossRef]
10. Raymon, A. Transesterification Approaches to Natural Esters for Transformer Insulating Fluids: A Review. *IEEE Trans. Dielectr. Electr. Insul.* **2024**, *31*, 607–614. [CrossRef]
11. Ma, Z.; Zhang, H.; Liu, J. DB-RNN: An RNN for Precipitation Nowcasting Deblurring. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 5026–5041. [CrossRef]
12. Han, L.; Wang, X.; Yu, Y.; Wang, D. Power Load Forecast Based on CS-LSTM Neural Network. *Mathematics* **2024**, *12*, 1402. [CrossRef]
13. Shen, K.; Cui, B.; Lyu, Q.Visible-polarized image fusion for nighttime dispersal of mines. *Opt. Precis. Eng.* **2024**, *32*, 2439–2453. [CrossRef]
14. Li, K.; Qi, M.; Zhuang, S.; Liu, Y. Polarized Prior Guided Fusion Network for Infrared Polarization Images. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–17. [CrossRef]
15. Xu, H.; Sun, Y.; Mei, X.; Tian, X.; Ma, J. Attention-Guided Polarization Image Fusion Using Salient Information Distribution. *IEEE Trans. Comput. Imaging.* **2022**, *8*, 1117–1130. [CrossRef]

16. Li, W.; Zhang, Y.; Wang, G.; Huang, Y.; Li, R. DFENet: A dual-branch feature enhanced network integrating transformers and convolutional feature learning for multimodal medical image fusion. *Biomed. Signal Process. Control* **2023**, *80*, 104402. [CrossRef]
17. Liu, X.; Gao, H.; Miao, Q.; Xi, Y.; Ai, Y.; Gao, D. MFST: Multi-modal feature self-adaptive Transformer for infrared and visible image fusion. *Remote Sens.* **2022**, *14*, 3233. [CrossRef]
18. Yi, S.; Jiang, G.; Liu, X.; Li, J.; Chen, L. TCPMFNet: An infrared and visible image fusion network with composite auto encoder and transformer–convolutional parallel mixed fusion strategy. *Infrared Phys. Technol.* **2022**, *127*, 104405. [CrossRef]
19. Li, J.; Zhu, J.; Li, C.; Chen, X.; Yang, B. Cgtf: Convolution-guided transformer for infrared and visible image fusion. *IEEE Trans. Instrum. Meas.* **2022**, *71*, 1–14. [CrossRef]
20. Tang, W.; He, F.; Liu, Y.; Duan, Y. MATR: Multimodal medical image fusion via multiscale adaptive transformer. *IEEE Trans. Image Process.* **2022**, *31*, 5134–5149. [CrossRef]
21. Wang, Z.; Wu, Y.; Wang, J.; Xu, J.; Shao, W. Res2Fusion: Infrared and visible image fusion based on dense Res2net and double non-local attention models. *IEee Trans. Instrum. Meas.* **2022**, *71*, 5005012. [CrossRef]
22. Zhu, Y.; Xu, H.; Du, A.; Wang, B. Image–Text Matching Model Based on CLIP Bimodal Encoding. *Appl. Sci.* **2024**, *14*, 10384. [CrossRef]
23. Chen, B.; Chen, L.; Khalid, U.; Zhang, S. IFSrNet: Multi-Scale IFS Feature-Guided Registration Network Using Multispectral Image-to-Image Translation. *Electronics* **2024**, *13*, 2240. [CrossRef]
24. Xiao, Y.; Huang, J.; Liu, X.; Zhu, A. Sla-former: Conformer using shifted linear attention for audio-visual speech recognition. *Complex Intell. Syst.* **2024**, *10*, 5721–5741. [CrossRef]
25. Yu, L.; Wu, S.; Gabbouj, M. Multi-Swin Transformer Based Spatio-Temporal Information Exploration for Compressed Video Quality Enhancement. *IEEE Signal Process. Lett.* **2024**, *31*, 1880–1884. [CrossRef]
26. Yoo, D.; Kim, J.; Yoo, J. FSwin Transformer: Feature-Space Window Attention Vision Transformer for Image Classification. *IEEE Access* **2024**, *12*, *72598–72606*. [CrossRef]
27. Hu, X.; Gao, G.; Li, B.; Wang, W.; Ghannouchi, F.M. A Novel Lightweight Grouped Gated Recurrent Unit for Automatic Modulation Classification. *IEEE Wirel. Commun. Lett.* **2024**, *13*, 2135–2139. [CrossRef]
28. Hajra, S.; Alam, M.; Saha, S.; Picek, S.; Mukhopadhyay, D. On the Instability of Softmax Attention-Based Deep Learning Models in Side-Channel Analysis. *IEEE Trans. Inf. Forensics Secur.* **2024**, *19*, 514–528. [CrossRef]
29. Gao, Z.; He, X.; Ma, Z.; Wei, S.; Xiong, J.; Aoki, Y. Distributed Scatterer Interferometry for Fast Decorrelation Scenarios Based on Sparsity Regularization. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–14. [CrossRef]
30. Ren, Y.; Ye, J.; Wang, X.; Xiao, F.; Liu, R. SAM-Net: Spatio-Temporal Sequence Typhoon Cloud Image Prediction Net with Self-Attention Memory. *Remote Sens.* **2024**, *16*, 4213. [CrossRef]
31. Lim, B.; Yun, W.J.; Kim, J.; Ko, Y.-C. Joint User Clustering, Beamforming, and Power Allocation for mmWave-NOMA With Imperfect SIC. *IEEE Trans. Wirel. Commun.* **2024**, *23*, 2025–2038. [CrossRef]
32. Li, J.; Lam, L.C.W.; Lu, H. Decoding MRI-informed brain age using mutual information. *Insights Imaging* **2024**, *15*, 216. [CrossRef] [PubMed]
33. Lei, Y.; Xu, L.; Wang, X.; Fan, X.; Zheng, B. IFGAN: Pre- to Post-Contrast Medical Image Synthesis Based on Interactive Frequency GAN. *Electronics* **2024**, *13*, 4351. [CrossRef]
34. Tiwari, L.B.; Burman, A.; Samui, P. A Comparative Study of Soft Computing Paradigms for Modelling Soil Compaction Parameters. *Transp. Infrastruct.* **2024**, *11*, 4142–4160. [CrossRef]
35. Yang, Y.; Jiao, L.; Liu, F.; Liu, X.; Li, L.; Chen, P.; Yang, S. An Explainable Spatial–Frequency Multiscale Transformer for Remote Sensing Scene Classification. *IEEE Trans. Geosci. Remote Sens.* **2024**, *61*, 1–15. [CrossRef]
36. Shi, L.; Zhao, Y. Edge Detection of High-Resolution Remote Sensing Image Based on Multi-Directional Improved Sobel Operator. *IEEE Access* **2023**, *11*, 135979–135993. [CrossRef]
37. Baker, A.H.; Pinard, A.; Hammerling, D.M. On a Structural Similarity Index Approach for Floating-Point Data. *IEEE Trans. Vis. Comput. Graph.* **2024**, *30*, 6261–6274. [CrossRef]
38. Chen, J.; Dai, N.; Hu, X.; Yuan, Y. A Lightweight Barcode Detection Algorithm Based on Deep Learning. *Appl. Sci.* **2024**, *14*, 10417. [CrossRef]
39. He, L.; Liang, T.; Wang, D.; Zhang, J.; Liu, B. Skarn Formation and Zn–Cu Mineralization in the Dachang Sn Polymetallic Ore Field, Guangxi: Insights from Skarn Rock Assemblage and Geochemistry. *Minerals* **2024**, *14*, 193. [CrossRef]
40. Kaiser, M.; Brusa, T.; Bertsch, M.; Wyss, M.; Ćuković, S.; Meixner, G.; Koch, V.M. Extrinsic Calibration for a Modular 3D Scanning Quality Validation Platform with a 3D Checkerboard. *Sensors* **2024**, *24*, 1575. [CrossRef]
41. Li, H.; Wu, X.; Durrani, T. NestFuse: An Infrared and Visible Image Fusion Architecture based on Nest Connection and Spatial/Channel Attention Models. *IEEE Trans. Instrum. Meas.* **2020**, *99*, 1. [CrossRef]
42. Jian, L.; Yang, X.; Liu, Z.; Jeon, G.; Gao, M.; Chisholm, D. SEDRFuse: A Symmetric Encoder-Decoder with Residual Block Network for Infrared and Visible Image Fusion. *IEEE Trans. Instrum. Meas.* **2020**, *99*, 1. [CrossRef]
43. Khorasani, A.; Dadashi Serej, N.; Jalilian, M.; Shayganfar, A.; Tavakoli, M.B. Performance comparison of different medical image fusion algorithms for clinical glioma grade classification with advanced magnetic resonance imaging (MRI). *Sci Rep.* **2023**, *13*, 17646. [CrossRef] [PubMed]
44. Wu, Y. *DenseFuseNet: Improve 3D Semantic Segmentation in the Context of Autonomous Driving with Dense Correspondence*; IEEE: Piscataway, NJ, USA, 2021.

45. Zhao, Z.; Xu, S.; Zhang, C.; Liu, J.; Li, P.; Zhang, J. DIDFuse: Deep image decomposition for infrared and visible image fusion. In *International Joint Conference on Artificial Intelligence*; IJCAI: Yokohama, Japan, 2020; pp. 970–976.
46. Xu, H.; Ma, J.; Jiang, J.; Guo, X.; Ling, H. U2Fusion: A Unified Unsupervised Image Fusion Network. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *44*, 502–518. [CrossRef]
47. Ma, J.; Tang, L.; Xu, M.; Zhang, H.; Xiao, G. STDFusionNet: An infrared and visible image fusion network based on salient target detection. *IEEE Trans. Instrum. Meas.* **2021**, *70*, 1–13. [CrossRef]
48. Liu, L.; Xu, N.; Zhou, W.; Qin, Y.; Luan, S. Improvement of Coal Mining-Induced Subsidence-Affected (MISA) Zone Irregular Boundary Delineation by MT-InSAR Techniques, UAV Photogrammetry, and Field Investigation. *Remote Sens.* **2024**, *16*, 4221. [CrossRef]
49. Qiang, Z.; Shi, J.; Shi, F. Phenotype Tracking of Leafy Greens Based on Weakly Supervised Instance Segmentation and Data Association. *Agronomy* **2022**, *12*, 1567. [CrossRef]
50. Vardakas, G.; Likas, A. Global k-means++: An effective relaxation of the global k-means clustering algorithm. *Appl. Intell.* **2024**, *54*, 8876–8888. [CrossRef]
51. Ana, B.; Arnaud, C.; Josu, D.; FJean-Michel, O. Performance paradox of dynamic matching models under greedy policies. *Queueing Syst.* **2024**, *107*, 257–293. [CrossRef]
52. Seliger, N.; Faltlhauser, G. Progressive Expansion Sampling of Quasi-Static Magnetic Fields in Unconfined Regions. *IEEE Trans. Components Packag. Manuf. Technol.* **2023**, *13*, 1576–1583. [CrossRef]