


Article

Multi-Modal Vision Transformer with Explainable Shapley Additive Explanations Value Embedding for *Cymbidium goeringii* Quality Grading

Zhen Wang [†], Xiangnan He [†], Yuting Wang and Xian Li ^{*†} 

Agricultural Information Institute, Chinese Academy of Agricultural Sciences, Beijing 100081, China; zhenskar@163.com (Z.W.); 82101225582@caas.cn (X.H.); wangyuting@caas.cn (Y.W.)

* Correspondence: lixian@caas.cn

[†] These authors contributed equally to this work.

Abstract: *Cymbidium goeringii* (*Rchb. f.*) is a traditional Chinese flower with highly valued biological, cultural, and artistic properties. However, the valuation of *Rchb. f.* mainly relies on subjective judgment, lacking a standardized digital evaluation and grading methods. Traditional grading methods solely rely on unimodal data and are based on fuzzy grading standards; the key features for values are especially inexplicable. Accurately evaluating *Rchb. f.* quality through multi-modal algorithms and clarifying the impact mechanism of key features on *Rchb. f.* value is essential for providing scientific references for online orchid trading. A multi-modal Transformer for *Rchb. f.* quality grading combined with the Shapley Additive Explanations (SHAP) algorithm was proposed, which mainly includes one embedding layer, one UNet, one Vision Transformer (ViT) and one Encoder layer. A multi-modal orchid dataset including images and text was obtained from Orchid Trading Website, and seven key features were extracted. Based on petals' RGB segmented from UNet and global fine-grained features extracted from ViT, text features and image features were organically fused into Transformer Encoders throughout concatenation operation, a 93.13% accuracy was achieved. Furthermore, SHAP algorithm was utilized to quantify and rank the importance of seven features, clarifying the impact mechanism of key features on *Rchb. f.* quality and value. This multi-modal Transformer with SHAP algorithm for *Rchb. f.* grading provided a novel idea to represent the explainable features accurately, exhibiting good potential for establishing a reliable digital evaluation method for agricultural products with high value.

Keywords: multi-modal feature fusion; explainable features representation; quality grading; transformer encoder; *Cymbidium goeringii*



Citation: Wang, Z.; He, X.; Wang, Y.; Li, X. Multi-Modal Vision Transformer with Explainable Shapley Additive Explanations Value Embedding for *Cymbidium goeringii* Quality Grading. *Appl. Sci.* **2024**, *14*, 10157. <https://doi.org/10.3390/app142210157>

Academic Editor: Pedro Couto

Received: 26 August 2024

Revised: 25 October 2024

Accepted: 30 October 2024

Published: 6 November 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Cymbidium goeringii (*Rchb. f.*) is a traditional Chinese flower that embodied biological, cultural, and artistic properties, and is also known as Chinese Orchid and flower gentleman [1]. Its features include petal type, flower color, leaf length, leaf width, etc. [2]. With an elegant appearance, beautiful flowers, and slender leaves, *Rchb. f.* has been loved by ancient and modern literati [3,4]. Some precious varieties have even been auctioned for millions [5,6]. In recent years, as living standards have improved and people's appreciation for beauty has deepened, orchids have become increasingly prevalent in countless homes [7]. However, in current orchid industry trading, the valuation of *Rchb. f.* relies on the subjective preferences of consumers, and even the opinions of orchid experts with long experience in planting and studying orchids are divided [8]. The evaluation standard and key features of *Rchb. f.* with a high value is still unclear. Especially for rising online trading, labeled data of *Rchb. f.* is lacking, and none of the valuing guidelines can be supplied to consumers. Therefore, accurately correlating the key features with values and establishing

digital quality evaluation indicators are crucial for promoting the development of the *Rchb. f.* industry.

This study proposed a multi-modal explainable *Rchb. f.* grading model based on Transformers. Inspired by the self-attention of Transformer Encoder, this model can provide more attention score to key features during training and handle variable length and incomplete data. Vision Transformer (ViT) transformed UNet-segmented petals into global fine-grained feature vectors, which were input into Transformer Encoders along with RGB values and textual features. To enhance features' explainability and model decision's reliability, Shapley Additive Explanations (SHAP) algorithm was adopted to sort features' importance, further clarifying the impact mechanism of key features on *Rchb. f.* quality and value. Finally, Transformer Encoder's performance was evaluated against the Vector Quantised-Variational AutoEncoder (VQ-VAE) and the Conditional Variational AutoEncoder (C-VAE), demonstrating its effectiveness in *Rchb. f.* quality grading.

2. Related Works

2.1. Orchid Variety Classification

Some studies have focused on classifying varieties or extracting the number of flowers and buds in orchid images. A hybrid model architecture, called the Homogeneous Ensemble Convolutional Neural Network (HE-CNN), was presented to classify the orchid species based on the global features of petal images, with an accuracy of 94.30% [9]. By training collected *Cymbidium* images with ten varieties, the Convolutional Neural Network (CNN) classified different varieties of orchids, achieving a 94.13% accuracy [8]. However, the above CNN algorithms just focused on the relationship between orchid images and varieties, without any key features. The PA-YOLO algorithm was applied to count the blooms and buds of *Phalaenopsis*, achieving 95.4% mean average precision (mAP) for buds and 91.9% for blooms by integrating a dual-scale detection branch and dynamic head framework [10]. However, the above studies did not extend to grading the quality of orchids.

2.2. Flower Quality Grading

Image processing and object detection algorithms have been employed for grading the quality of flowers. Bubble sequencing and cluster segmentation were used to measure the crown spathe width and spathe number of potted *Anthuriums*, respectively, and a minimum closed rectangle was used to measure crown width and spathes. Based on these features and the existing grading standard of potted *Anthurium*, a grading accuracy of 85.86% was achieved [11]. YOLO-V3 was utilized to detect objects' position on *phalaenopsis*, including red flowers, white flowers, stems, red buds, white buds, and inappropriate deciduous buds on stems, so as to determine the three grades of *phalaenopsis*, with an accuracy of 82% [12]. A CNN was employed to grade *Gloxinia* quality based on images, with its grading capability enhanced by determining whether samples containing buds belonged to medium grade, achieving an accuracy of 89.6% [13]. However, the above models merely output grades based on images, and the importance of each feature to the grading result was not clear.

2.3. Agricultural Product Quality Grading

Quality grading algorithms have been applied to edible agricultural products, based on different features. The Back Propagation (BP) Neural Network and Support Vector Machine (SVM) were utilized to grade pear appearance quality based on shape, surface color, and defects according to industry standards, achieving an accuracy of 80.5% [14]. Machine learning ensemble models (XGBoost, LightGBM, and Linear models) were employed to estimate apple ripeness by training spectral sensor data, facilitating apple quality grading based on a custom relationship between the ripeness and quality levels, achieving an accuracy of 80.5% [15]. The relationship between the RGB and HSV color maps of mushrooms' freshness was analyzed, and the grading of two quality classes of mushrooms (fresh and deteriorated) were carried out using an Artificial Neural Network (ANN) and

SVM, with accuracies of 94.7% and 93.4%, respectively [16]. However, the above adopted grading methods lacked textual data, and were based on existing grading standards, which are insufficient for classifying orchids that lack defined indicators and grading standards. Especially, the outputs and features of these models do not clearly represent the key features affecting quality, nor their importance, providing limited guidance for the grading results of agricultural production processes. In contrast, this study fully utilizes both image and textual features of orchids, effectively uncovering the impact mechanism of key features on the orchid's quality and value. This provides a scientific basis for *Rchb. f.* cultivation and online transactions.

3. Materials and Methods

3.1. Dataset Acquisition and Definition

Information on 4556 *Rchb. f.* sales was collected from Orchid Trading Website (<https://www.hmlan.com>, (accessed on 2 January 2024)). To ensure transparency and authenticity in transactions, the website requires sellers to provide detailed information about their orchids' attributes, which are also essentially of interest to orchid lovers and buyers. Based on these attributes, the recognized text features were classified into three categories (flower, leaf, and bud) and seven specific indicators (petal type, longest leaf, widest leaf, leaf height, leaf number, bud number, seedling number). Then, these key features were extracted from the text using regular expression matching, forming 4556 groups of joint datasets with both images and multiple key features (Table 1). Each sample in the dataset consists of a feature vector and a label. Each feature vector has eight features, including one image and seven key features. *Rchb. f.* samples with different price ranges are defined as different grades. In order to facilitate the tuning of model hyperparameters, prevent overfitting, and monitor training effectiveness, 10% of the samples (455 samples) from each level were selected as validation set, and 10% of the samples (449 samples) as the test set. Finally, the remaining 3652 samples were used as the training set (Table 2).

Table 1. The multi-modal *Rchb. f.* dataset and (taking 4 samples as examples; “-” denotes missing feature).





Image	Flower		Leaves			Bud		Label (Grade)
	Petal Type	Longest Leaf	Widest Leaf	Leaves Height	Leaves Number	Bud Number	Seedlings Number	
	Lotus	20	0.6	-	-	-	-	1
	Lotus	12	0.6	-	5	2	2	2
	Lotus	23	1.1	-	27	-	5	3
	Plum	39	1.2	-	-	-	3	4

Table 1 showed the varying degrees of missing multi-modal features due to the lack of an evaluation standard for values. To solve this problem, the model received an array of feature indices (for identifying feature names) and an array of corresponding feature values, respectively. Therefore, feature indices and feature values were separated. Each feature name was mapped to a unique numerical index, transforming it into a recognizable digital format. Each feature value corresponds to its respective feature name. Given the presence of missing values, padding ensures consistent input tensor dimensions for the Neural

Network. Even if some features were missing, their feature indices were still included in feature indices array, represented by the padding value of 9, and their feature values were included in feature values array, indicated by the padding value of 0. This allows the model to dynamically handle variable-length inputs and missing features without compromising the integrity of the information processed. Then, the preprocessed data were fed into the input embedding layer and Transformer Encoder of the developed model.

Table 2. Label, reference price, and sample numbers for training, validation, and test sets.

Label (Grade)	Reference Price Level (RMB)	Numbers of Samples	Size of Training Set	Size of Validation Set	Size of Test Set
1	0–100	722	587	73	62
2	100–300	696	556	70	70
3	300–600	673	538	66	69
4	600–1500	668	534	65	69
5	1500–3000	624	499	63	62
6	3000–5000	593	474	60	59
7	>5000	580	464	58	58
Total		4556	3652	455	449

3.2. Transformer Encoder

The Transformer Encoder structure is the key of the multi-modal *Rchb. f.* quality grading model, and its construction was shown in Figure 1a. Self-attention is the core of the Transformer Encoder layer, which enables the model to dynamically weigh the importance of different input features. The Transformer Encoder consists of a bunch of identical layers, each of which consists of a multi-head self-attention and a fully connected feedforward network as its main components. Each major component incorporates normalization and residual concatenation to mitigate the vanishing gradient problem. Multiple Transformer Encoder layers are connected back and forth to form a deep Transformer Encoder structure [17,18].

The self-attention allows the encoder to consider other parts of the input sequence when encoding a particular piece. It computed the attention scores based on the query (Q), key (K), and value (V) matrices derived from the input embedding layer. The attention function can be described as

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

where d_k is the dimensionality of the key vectors, which helps in stabilizing the gradients. Multi-head attention extends this by parallelizing multiple attention heads, each focusing on different parts of the input sequence, thereby capturing a richer representation of the input data. Each Transformer Encoder layer includes a feed-forward network applied to each position separately and identically. Since the order of seven features was irrelevant to grading results, the positional encoding in Transformer Encoder was removed in this study. This adjustment allows the model to focus on the key features itself, adapting to the disorder of features.

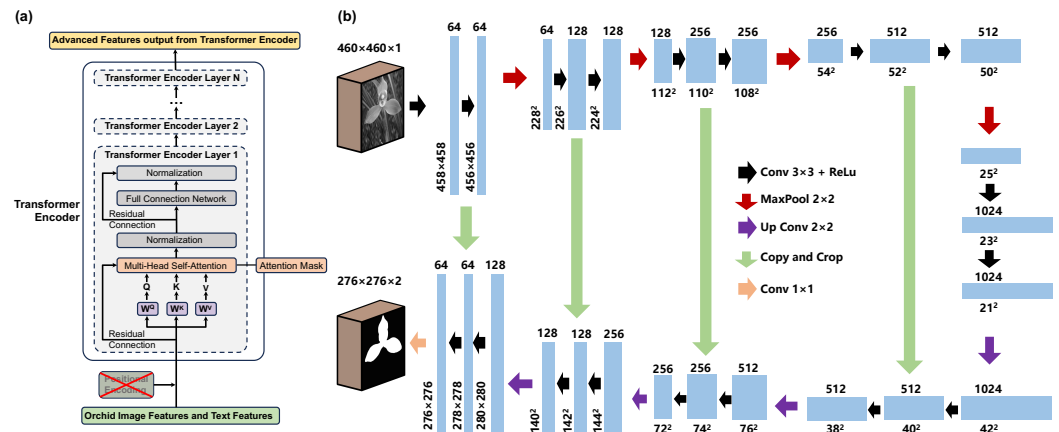


Figure 1. (a) Construction of Transformer Encoder and its self-attention. (b) UNet construction for petal segmentation.

3.3. UNet Architecture

The chaotic and diverse backgrounds in images from orchid trading website necessitated UNet to segment the petals. UNet has a symmetrical U-shaped architecture, including a contraction path (encoder) and an expansion path (decoder), as shown in Figure 1b. The encoder is a typical Convolutional Network architecture, consisting of repeated application of two 3×3 convolutions (each followed by a ReLu activation), and a 2×2 max pooling operation with stride of 2 for down-sampling. At each down-sampling step, the number of feature channels was doubled, enhancing the network’s capacity to capture high-level semantic information at multiple scales. The decoder inversely mirrors the encoder, employing up-sampling operations that increase the resolution of the output. Each step in the expansive path involved up-sampling of the feature map, followed by a 2×2 up-convolution that halves the number of feature channels, a concatenation with the correspondingly cropped feature map from the contraction path, and two 3×3 convolutions, each followed by a ReLu. This process recovered spatial information lost during down-sampling, enabling precise localization for segmentation tasks. UNet used jump connections to feed the feature map directly from the encoder to the decoder. These connections helped to recover spatial context lost during down-sampling, which is essential for accurate pixel classification [19].

3.4. Model Architecture Design

The multi-modal Transformers for *Rchb. f.* quality grading was developed, which includes 1 input embedding layer, 1 UNet, 1 ViT, N Transformer Encoder layers, 1 subsequent pooling layer, 1 linear layer, and 1 Softmax layer. The input embedding layer transformed each feature indices into a dense vector representation. This enabled the model to capture the semantics of features within a continuous numerical vector space, while also placing the features in a learnable embedding space, providing more accurate inputs for subsequent operations. Subsequently, the dense vectors containing feature indices information are element-wise multiplied with the feature values through padding and dot product operations. Since missing feature values have been replaced with padding value 0 in feature values array, the padding value canceled out the corresponding missing feature indices by dot product operation. Through this novel design, the model can precisely learn both the existence of input features and the feature values.

UNet was used to perform semantic segmentation on *Rchb. f.* images. Dot product operation combined the average RGB value of UNet-segmented petal images with the input embedded feature indices and the corresponding feature values, so that the model understood color information. ViT acted as the primary feature extractor for UNet-segmented petal images, producing a rich feature vector that encapsulates the global visual information of orchid petal. The vector concatenated with the output from the previous dot

product operation of the embedded feature indices and corresponding values. Then, the concatenated features were input into Transformer Encoder and processed through self-attention. The Transformer Encoder’s layer needs to be deepened to improve the network’s ability to learn from the training set, tentatively to N layers. A pooling operation averaged the output from Transformer Encoder, reducing the data’s dimensionality while retaining critical information. The pooled output was then fed into a linear layer and mapped to a dimension consistent with the number of orchid levels. A Softmax layer converts these linear outputs into probabilities, each reflecting the likelihood that an orchid belongs to a particular quality grade. Finally, the architecture of multi-modal Transformers for *Rchb. f.* quality grading was constructed, as shown in Figure 2.

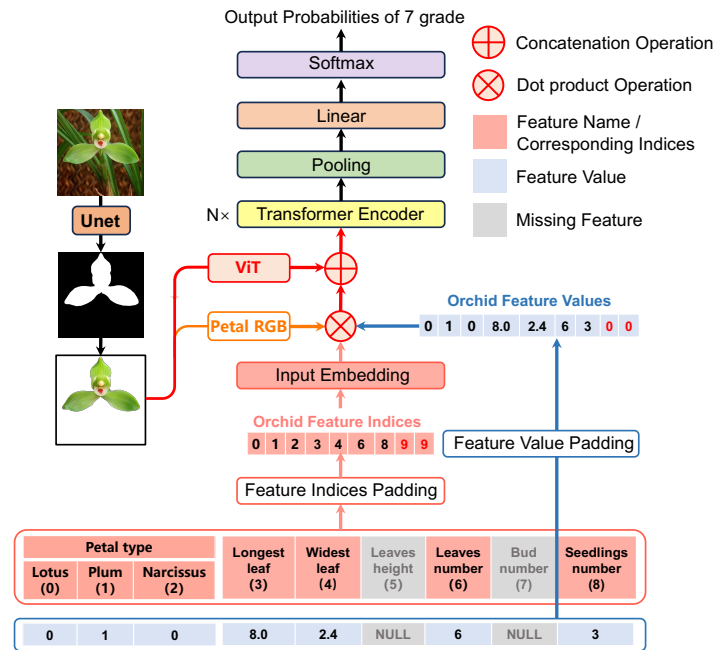


Figure 2. Architecture of multi-modal Transformers for *Rchb. f.* quality grading. Red numbers 9 and 0 are padding values of orchid feature indices and orchid feature values, respectively.

3.5. SHAP Algorithm for Feature Importance Calculation

The SHAP algorithm can output the quantifiable importance of distinct features in orchid quality grading, facilitating the establishment of a standardized, data-driven paradigm for orchid assessment. This enhanced the reliability, transparency, and explainability of deep learning models. SHAP is grounded in the concept of game theory, particularly Shapley values, a method for assigning payouts to players depending on their contribution to the total payout. In the realm of machine learning, SHAP values interpret the prediction of an instance by computing the contribution of each feature to the prediction. The SHAP value is the average marginal contribution of a feature value across all possible combinations of features [20]. The core calculation process that encapsulated the computation of SHAP values for the prediction model f based on a feature set x , as shown in Figure 3. The formula of SHAP algorithm was given by

$$v(S \cup \{j\}) = \int f(x_{S \cup \{j\}}^{(i)} \cup X_{C \setminus j}) d\mathbb{P}_{X_{C \setminus j}} - \mathbb{E}[f(X)], C = F \setminus S \tag{2}$$

$$v(S) = \int f(x_S^{(i)} \cup X_C) d\mathbb{P}_{X_C} - \mathbb{E}[f(X)] \tag{3}$$

$$\phi_j^{(i)} = \sum_{S \subseteq F \setminus \{j\}} \frac{|S|!(|F| - |S| - 1)!}{p!} [v(S \cup \{j\}) - v(S)], \phi_j = \frac{1}{N} \sum_{i=1}^N \phi_j^{(i)} \tag{4}$$

where $\phi_j^{(i)}$ is the SHAP value for feature j , F is the set of all features, S is a subset of features excluding j , $|S|$ is the number of features in subset S , $|F|$ is the total number of features, and $f(x_S^{(i)} \cup X_C)$ is the prediction model's output given the features in S . $f(x_{S \cup \{j\}}^{(i)} \cup X_{C \setminus j})$ is the output of model when feature i is added to S [20]. In this study, by computing and averaging SHAP values for each feature across all samples, 7 features were ranked based on their importance in determining *Rchb. f.* grades, and the output of model was explainable. This ranking mined the key features affecting model grading decisions, providing valuable information for growers and consumers.

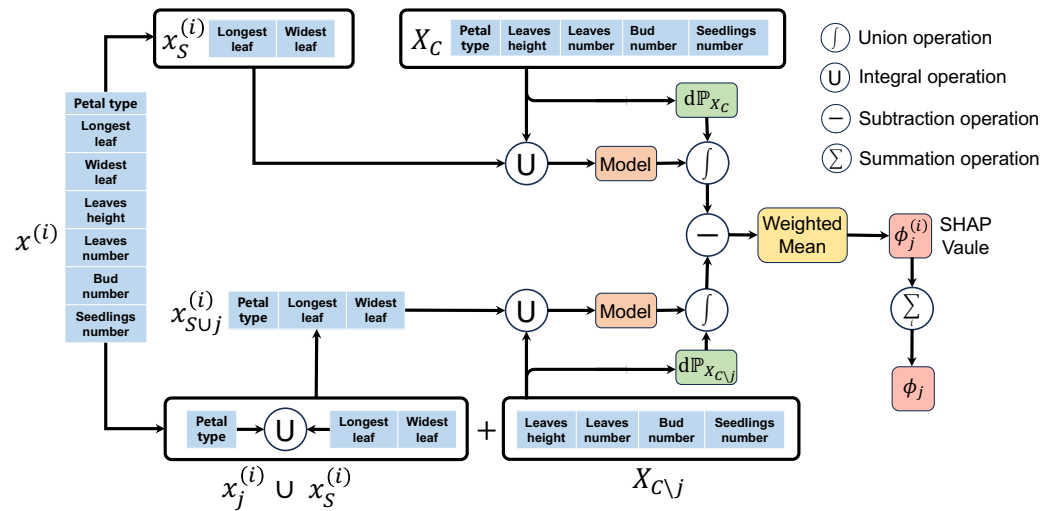


Figure 3. Computing process of SHAP algorithm for explainable features representation.

4. Results and Discussion

4.1. Combination of Image and Text Basically Achieved Quality Grading

Multiple regular expression matching patterns were designed for each feature based on all of the possible formats in which each feature could appear in the descriptive text, obtaining a value for each feature present in each sample. Numerical features extracted from the textual data were solely utilized to train the model (Experiment (Exp) 1, Figure 4a). The influence of the layer number of the Transformer Encoder was studied (insert of Figure 4a). As the layer number increased, the accuracy of test set generally increased first, and reached the maximum (59.36%) at three layers, then sharply fell. Therefore, the structure of Transformer Encoder was optimized with three layers. Due to the prevalent features missing in most samples within the orchid dataset, the feature matrix exhibited significant sparsity, with most elements being empty [21]. A few non-zero elements cannot sufficiently represent the grading pattern of orchids. The sparsity led to the model overfitting to limited available and frequently occurring features, resulting in poor generalization ability [22]. The subjectivity of sellers or different grading standards contributed to inconsistency in the dataset. *Rchb. f.* with similar grades often had vastly different numerical features. It is difficult for the model to stably learn quality grading patterns from a dataset [23]. After 224 epochs, peak accuracies of 63.02% and 59.36% on validation set and test set were achieved, respectively (Figure 4b).

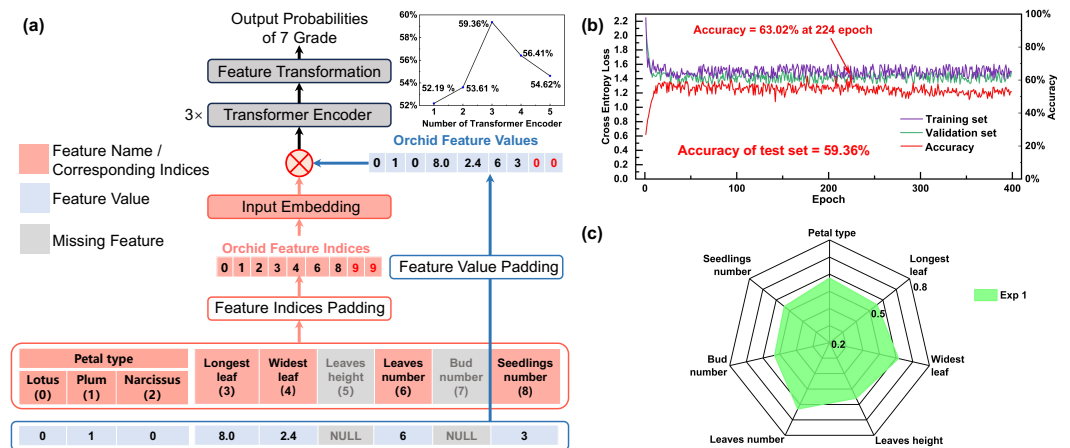


Figure 4. (a) Illustration of the model architecture with numerical features extracted from textual data and accuracy of the model in the test set under different numbers of Transformer Encoder layers (insert). (b) The value curves of Cross Entropy Cost and accuracy of the model. (c) Radar chart of Feature importance based on SHAP values.

SHAP analysis was integrated to quantify the importance of seven orchid numerical features based on their contribution to the model’s grading decisions. The analysis revealed that ‘Leaf number’ had the most significant impact on the model, whereas ‘Bud Number’ had the least. However, the SHAP values for all seven features were closely aligned, indicating a challenge for the model in distinguishing the differential impact of each feature on grading outcomes, as depicted in Figure 4c.

To address the limitations brought by numerical features extracted from textual data, visual datasets were incorporated into dataset. UNet was utilized to segment the petals from orchid images. After manually annotating all orchid sample images and conducting 42 epochs, the test set achieved a maximum mIOU of 89.16%. The average RGB values were computed from the *i*th segmented petal images by averaging the pixel values across the red, green, and blue channels, defined as

$$\text{Average_RGB}_i = \left(\frac{1}{N_i} \sum_{j=1}^{N_i} R_j, \frac{1}{N_i} \sum_{j=1}^{N_i} G_j, \frac{1}{N_i} \sum_{j=1}^{N_i} B_j \right) \tag{5}$$

where R_j, G_j, B_j are the red, green, and blue pixel values of the segmented petals, respectively, and N_i is the total number of pixels of the *i*th image in the segmented area.

Dot product operation between Average_RGB_i and numerical features extracted from textual data were integrated (Figure 5a). The model was then retrained, resulting in a significant performance improvement. The peak accuracy of validation set reached 84.92%, and the test set’s accuracy improved to 81.69% (Figure 5b). Color carries the health, maturity, and rarity of orchids [24,25]. The RGB values of color introduced a new dimension of visual features for orchids, enhancing the model’s capability to detect subtle quality variations. Moreover, the continuity and density of RGB information reduced data sparsity and increased signal stability [26], providing a more accurate foundation for the model’s decisions. By integrating RGB with textual features, the model initially gained the advantage of multi-modal learning, establishing deeper correlations between visual and textual levels, thereby comprehensively understanding the characteristics of orchids. Subsequent SHAP analysis revealed a new impact of features on the model’s predictions. ‘Petal type’ emerged as the most influential feature. The disparity in SHAP values among the seven features became pronounced, indicating that the integration of petal color and textual data made the model preliminarily distinguish the importance of different features (Exp 2, Figure 5c).

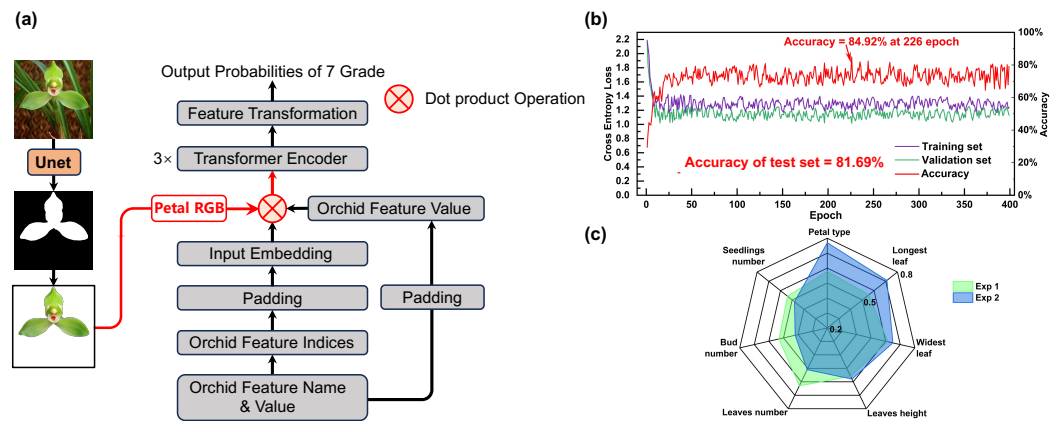


Figure 5. (a) Illustration of model architecture with UNet and numerical features extracted from textual data. (b) The value curves of Cross Entropy Cost and accuracy of the model. (c) Radar chart of feature importance based on SHAP values.

4.2. ViT and Global Fine-Grained Features Achieved Explainable Features Representation

Petal color and textual information alone are not enough to distinguish the value of different qualities of orchids. The incorporation of ViT was crucial for extracting global fine-grained features from orchid petals, such as shape and texture of petals. The feature vector was concatenated with the output from the previous dot product operation of the embedded feature indices and corresponding values, as shown in Exp 3 in (Figure 6a). The accuracy of validation set soared to 93.39% after 39 epochs, while the test set accuracy reached 91.86% (Figure 6b). The improvement underscores the efficacy of ViT in discerning subtle patterns in the petal images that were not captured by color information alone [27]. By focusing on all locations and taking their weighted average to compute the response at a given location in image sequence, self-attention within ViT allowed the model to selectively focus on different parts of the petal images and analyzed the spatial hierarchical interrelations among petal features [28]. SHAP analysis post-ViT integration revealed that ‘Seeding number’ had the least impact on the model’s predictions, indicating a shift in feature importance hierarchy (Exp 3 in Figure 6d). This shift emphasized the model’s capacity to prioritize the most informative visual and textual cues for grading, highlighting ViT’s nuanced understanding of *Rchb. f.* quality.

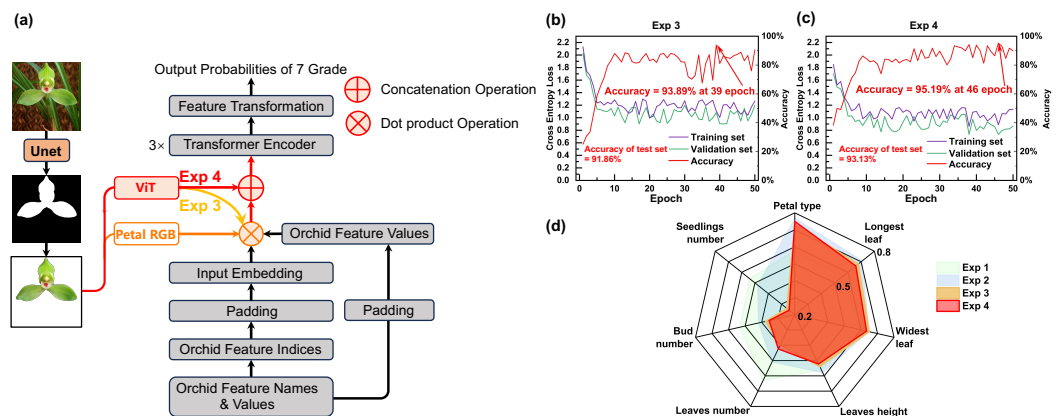


Figure 6. (a) Illustration of model architecture with UNet, ViT, and numerical features extracted from textual data (Exp 3) and concatenation operation for connecting ViT’s output to Transformer Encoder (Exp 4). The value curves of Cross Entropy Cost and accuracy of (b) Exp 3 and (c) Exp 4. (d) Radar chart of feature importance based on SHAP values.

4.3. Concatenation Replaced Dot Production Reflecting the Relationship Between Key Features and Value Adequately

The dot product in feature integration reduces the dimensionality of the output of ViT. Since the dot product is highly sensitive to the scale of the feature values, it can also merge multiple vectors into a scalar or a shorter vector, and the values with opposite signs will be canceled out, resulting in information loss [29,30]. The transformation of the integration strategy of feature vectors refined the model architecture. The dot product operation between ViT's output and previous dot product operation result was deprecated, and optimized to concatenation operation, as shown in Exp 4 in (Figure 6a). The accuracy of validation set and test set has been further improved to 95.19% and 93.13%, respectively (Figure 6c). The concatenation operation preserved original information contained in the global fine-grained petal features of ViT, compensating for the loss caused by the dot product operation [31]. The output of the concatenation operation provided an optimal combined input to Transformer Encoder adaptively, thus addressing the shortcomings of fixed integration of the dot product operation [32], realizing the organic fusion of the multi-modal features. Therefore, the performance of the grade model reached its optimal level. Exp 4 in (Figure 6d) displayed the final SHAP value sorting, determined by the orchid quality grading model. Compared with the dot product operation, the feature importance sorting of the concatenation operation was unchanged, but the difference of seven SHAP values were more significant. This transformation showed that concatenation operation not only improved model's accuracy, but also further optimized its ability to assess the importance of features.

To verify the performance of ViT in image encoding, both HE-CNN [9] and CNN [8] from related works replaced ViT as image encoders for global fine-grained orchid features, and the results are shown in Table 3. It can be observed that the performance of HE-CNN and CNN is inferior to ViT for both the validation and test sets.

Table 3. Accuracy's comparison of development set and test set with HE-CNN, CNN and ViT.

Encoder Input	Accuracy of Validation Set			Accuracy of Test Set		
	HE-CNN	CNN	ViT	HE-CNN	CNN	ViT
Numerical features ⊗ RGB ⊕ global feature from HE-CNN, CNN or ViT	81.22%	74.46%	94.19%	76.58%	70.63%	93.13%

CNN gradually extracts local features through convolutional layers. Although multiple layers can be stacked to capture a larger receptive field, the inherent local convolution operations may limit its ability to effectively capture global dependencies. HE-CNN combines a global prediction network (GPN) for global feature extraction and a local prediction network (LPN) for local feature extraction, aiming to capture both global and local information. However, as it is still based on the CNN architecture, it remains constrained by the locality of convolution operations. In contrast, ViT can directly capture complex global information of the entire image by self-attention without relying on local convolutions. Therefore, ViT is more suitable for extracting images such as *Rchb. f.* that rely on global fine-grained features. Combined with text features, the model can better understand the relationship between *Rchb. f.* features and value, thereby accurately establishing a digital indicator system for *Rchb. f.* quality evaluation.

The performance of Transformer Encoder was further verified against existing commonly used feature encoders, detailed in Table 4. Initially, the Transformer Encoder employed for processing the combined features of orchids was replaced with the VQ-VAE Encoder. For the #1 training, the performance indicated minor decline from Transformer-based model. However, the #2 and #3 training showed a noticeable performance decline compared to the Transformer Encoder. VQ-VAE Encoder created a static, discrete codebook representation of the input, which limited its flexibility in capturing the nuanced variations

in orchid features [33]. The VQ-VAE Encoder is used to generate compact, discrete representations, only for datasets with well-defined and consistent patterns [34,35]. However, the *Rchb. f.* dataset exhibited significant feature sparsity and inconsistency caused by missing values, and is not an ideal input for the VQ-VAE Encoder. The Transformer Encoder was then replaced with C-VAE Encoder. Three identical trainings were performed, resulting in a further performance decrease compared to both Transformer Encoder and VQ-VAE Encoder. The conditional generation of C-VAE is designed to generate outputs based on fixed conditions, which cannot capture the full range of variability and nuances of orchids [36,37]. Additionally, the variational method introduced variability while processing missing data based on the probability framework [38], leading to poor performance in the classification task. Conversely, the Transformer Encoder can dynamically focus on important features through self-attention spontaneously [39], rather than static codebook and fixed conditions generation, providing adaptability and good performance that is not achieved by rigid structures of VQ-VAE Encoder and C-VAE Encoder, and becoming an excellent choice for orchid quality grading model.

Table 4. Accuracy comparison of validation set and test set with VQ-VAE Encoder, C-VAE Encoder, and Transformer Encoder

Training ID	Encoder Input	Accuracy of Validation Set			Accuracy of Test Set		
		VQ-VAE Encoder	C-VAE Encoder	Transformer Encoder	VQ-VAE Encoder	C-VAE Encoder	Transformer Encoder
#1	Numerical features	57.21%	51.36%	63.02%	52.56%	48.71%	59.36%
#2	Numerical features ⊗ RGB	75.67%	71.32%	84.92%	72.81%	69.10%	81.69%
#3	Numerical features ⊗ RGB ⊕ global feature from ViT	85.36%	81.29%	94.19%	82.48%	77.88%	93.13%

In summary, through the integration of Transformer Encoder's self-attention and the combination of global fine-grained features, RGB and text features, this model achieved accurate grading of *Rchb. f.*, clarified the impact mechanism of key features on its quality and value, which provided a scientific reference for the online trading between orchid growers and hobbyists.

5. Conclusions

A multi-modal Transformer for *Rchb. f.* quality grading combined with explainable features representation was firstly developed. A dataset of 4556 *Rchb. f.* samples was obtained from the Orchid Trading Website. Seven key features were extracted through regular expression matching. Four strategies were proposed for excellent accuracy. Firstly, the Transformer Encoder adapted missing features, and its self-attention dynamically prioritized important features. The dot product input feature's existence and feature value into the Transformer Encoder simultaneously. Secondly, UNet was employed for petal segmentation, integrating the average RGB of petals into the Transformer Encoder, addressing data sparsity and overfitting in textual data, and enabling the model to grade the orchid quality by color. Thirdly, ViT was utilized to transform petal images into global fine-grained features, which were combined with color extracted from UNet and numerical features. Finally, the concatenation operation helped the model to understand the relationship between key features and values adequately, and an excellent accuracy of 93.13% was achieved. Furthermore, SHAP analysis algorithm ranked feature importance and gave the model explainability for grading results. Therefore, this multi-modal Transformer exhibited its advantage of explainable features representation affecting the *Rchb. f.* value. This study

suggested an effective way to establish a reliable digital index system for agricultural product quality evaluation without referred standards.

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

Funding: This research was funded by the Innovation Program of Chinese Academy of Agricultural Sciences (Funder No: CAAS-CAE-202302, CAAS-ASTIP-2024-AII).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors upon request.

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

References

1. Yang, F.; Zhu, G.; Wang, Z.; Liu, H.; Xu, Q.; Huang, D.; Zhao, C. Integrated mRNA and microRNA transcriptome variations in the multi-tepal mutant provide insights into the floral patterning of the orchid *Cymbidium goeringii*. *BMC Genom.* **2017**, *18*, 1–24. [[CrossRef](#)] [[PubMed](#)]
2. Chen, J.; Song, J.; Han, W.; Chen, B.; Zhang, X. Morphological diversity of wild *Cymbidium goeringii* and *Cymbidium faberi* in the Qinling Mountains. *J. Northwest A F Univ.-Nat. Sci. Ed.* **2017**, *45*, 143–150.
3. Balilashaki, K.; Martinez-Montero, M.E.; Vahedi, M.; Cardoso, J.C.; Silva Agurto, C.L.; Leiva-Mora, M.; Feizi, F.; Musharof Hossain, M. Medicinal Use, Flower Trade, Preservation and Mass Propagation Techniques of *Cymbidium* Orchids—An Overview. *Horticulturae* **2023**, *9*, 690. [[CrossRef](#)]
4. Yang, F.; Gao, J.; Li, J.; Wei, Y.; Xie, Q.; Jin, J.; Lu, C.; Zhu, W.; Wong, S.M.; Zhu, G. The China orchid industry: Past and future perspectives. *Ornam. Plant Res.* **2024**, *4*, e002. [[CrossRef](#)]
5. Seyler, B.C. *The Role of Botanical Gardens in the Conservation of Orchid Biocultural Diversity in Sichuan Province, China*; University of Hawai'i at Manoa: Honolulu, HI, USA, 2017.
6. Tiwari, P.; Sharma, A.; Bose, S.K.; Park, K.I. Advances in Orchid Biology: Biotechnological Achievements, Translational Success, and Commercial Outcomes. *Horticulturae* **2024**, *10*, 152. [[CrossRef](#)]
7. Shefferson, R.P.; Jacquemyn, H.; Kull, T.; Hutchings, M.J. The demography of terrestrial orchids: Life history, population dynamics and conservation. *Bot. J. Linn. Soc.* **2020**, *192*, 315–332. [[CrossRef](#)]
8. Fu, Q.; Zhang, X.; Zhao, F.; Ruan, R.; Qian, L.; Li, C. Deep feature extraction for *cymbidium* species classification using global-local CNN. *Horticulturae* **2022**, *8*, 470. [[CrossRef](#)]
9. Sarachai, W.; Bootkrajang, J.; Chaijaruwanich, J.; Somhom, S. Orchid classification using homogeneous ensemble of small deep convolutional neural network. *Mach. Vis. Appl.* **2022**, *33*, 17. [[CrossRef](#)]
10. Yang, Y.; Zhang, G.; Ma, S.; Wang, Z.; Liu, H.; Gu, S. Potted phalaenopsis grading: Precise bloom and bud counting with the PA-YOLO algorithm and multiviewpoint imaging. *Agronomy* **2024**, *14*, 115. [[CrossRef](#)]
11. Wei, H.; Tang, W.; Chu, X.; Mu, Y.; Ma, Z. Grading method of potted anthurium based on RGB-D features. *Math. Probl. Eng.* **2021**, *2021*, 1–8. [[CrossRef](#)]
12. Chang, Y.W.; Hsiao, Y.K.; Ko, C.C.; Shen, R.S.; Lin, W.Y.; Lin, K.P. A Grading System of Pot-Phalaenopsis Orchid Using YOLO-V3 Deep Learning Model. In Proceedings of the Advances in Networked-Based Information Systems: The 23rd International Conference on Network-Based Information Systems (NBIS-2020) 23, Victoria, Canada, 31 August–2 September 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 498–507. [[CrossRef](#)]
13. Sun, Y.; Zhu, L.; Wang, G.; Zhao, F. Multi-input convolutional neural network for flower grading. *J. Electr. Comput. Eng.* **2017**, *2017*, 9240407. [[CrossRef](#)]
14. Yang, Z.; Li, Z.; Hu, N.; Zhang, M.; Zhang, W.; Gao, L.; Ding, X.; Qi, Z.; Duan, S. Multi-Index Grading Method for Pear Appearance Quality Based on Machine Vision. *Agriculture* **2023**, *13*, 290. [[CrossRef](#)]
15. Chopra, H.; Singh, H.; Bamrah, M.S.; Mahbubani, F.; Verma, A.; Hooda, N.; Rana, P.S.; Singla, R.K.; Singh, A.K. Efficient fruit grading system using spectrophotometry and machine learning approaches. *IEEE Sens. J.* **2021**, *21*, 16162–16169. [[CrossRef](#)]
16. Mukherjee, A.; Sarkar, T.; Chatterjee, K.; Lahiri, D.; Nag, M.; Rebezov, M.; Shariati, M.A.; Miftakhutdinov, A.; Lorenzo, J.M. Development of artificial vision system for quality assessment of oyster mushrooms. *Food Anal. Methods* **2022**, *15*, 1663–1676. [[CrossRef](#)]
17. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*. [[CrossRef](#)]

18. Yan, H.; Deng, B.; Li, X.; Qiu, X. TENER: Adapting transformer encoder for named entity recognition. *arXiv* **2019**, arXiv:1911.04474. [[CrossRef](#)]
19. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, 5–9 October 2015; Springer: Berlin/Heidelberg, Germany, 2015; Volume 3, pp. 234–241. [[CrossRef](#)]
20. Gramegna, A.; Giudici, P. SHAP and LIME: An evaluation of discriminative power in credit risk. *Front. Artif. Intell.* **2021**, *4*, 752558. [[CrossRef](#)]
21. Luo, X.; Zhou, M.; Li, S.; Hu, L.; Shang, M. Non-negativity constrained missing data estimation for high-dimensional and sparse matrices from industrial applications. *IEEE Trans. Cybern.* **2019**, *50*, 1844–1855. [[CrossRef](#)]
22. Chen, H.; Li, H.; Li, Y.; Chen, C. Sparse spatial transformers for few-shot learning. *Sci. China Inf. Sci.* **2023**, *66*, 210102. [[CrossRef](#)]
23. Maharana, K.; Mondal, S.; Nemade, B. A review: Data pre-processing and data augmentation techniques. *Glob. Transit. Proc.* **2022**, *3*, 91–99. [[CrossRef](#)]
24. Jain, A.; Sarsaiya, S.; Chen, J.; Wu, Q.; Lu, Y.; Shi, J. Changes in global Orchidaceae disease geographical research trends: Recent incidences, distributions, treatment, and challenges. *Bioengineered* **2021**, *12*, 13–29. [[CrossRef](#)] [[PubMed](#)]
25. Zhao, X.; Li, Y.; Zhang, M.M.; He, X.; Ahmad, S.; Lan, S.; Liu, Z.J. Research advances on the gene regulation of floral development and color in orchids. *Gene* **2023**, *888*, 147751. [[CrossRef](#)] [[PubMed](#)]
26. Hadsell, R.; Rao, D.; Rusu, A.A.; Pascanu, R. Embracing change: Continual learning in deep neural networks. *Trends Cogn. Sci.* **2020**, *24*, 1028–1040. [[CrossRef](#)]
27. Arshed, M.A.; Rehman, H.A.; Ahmed, S.; Dewi, C.; Christanto, H.J. A 16 × 16 Patch-Based Deep Learning Model for the Early Prognosis of Monkeypox from Skin Color Images. *Computation* **2024**, *12*, 33. [[CrossRef](#)]
28. Mehta, S.; Rastegari, M. Separable self-attention for mobile vision transformers. *arXiv* **2022**, arXiv:2206.02680. [[CrossRef](#)]
29. Li, R.; Zheng, S.; Zhang, C.; Duan, C.; Su, J.; Wang, L.; Atkinson, P.M. Multiattention network for semantic segmentation of fine-resolution remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–13. [[CrossRef](#)]
30. Zhang, B.; Tian, Z.; Tang, Q.; Chu, X.; Wei, X.; Shen, C. Segvit: Semantic segmentation with plain vision transformers. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 4971–4982. [[CrossRef](#)]
31. Zheng, Q.; Zhu, J.; Li, Z.; Pang, S.; Wang, J.; Li, Y. Feature concatenation multi-view subspace clustering. *arXiv* **2019**, arXiv:1901.10657. [[CrossRef](#)]
32. Henderson, M.; Casanueva, I.; Mrkšić, N.; Su, P.H.; Wen, T.H.; Vulić, I. ConveRT: Efficient and accurate conversational representations from transformers. *arXiv* **2019**, arXiv:1911.03688. [[CrossRef](#)]
33. Pucci, R.; Micheloni, C.; Foresti, G.L.; Martinel, N. CVGAN: Image Generation with Capsule Vector-VAE. In Proceedings of the International Conference on Image Analysis and Processing, Lecce, Italy, 23–27 May 2022; Springer: Berlin/Heidelberg, Germany, 2023; pp. 536–547. [[CrossRef](#)]
34. Nash, C.; Menick, J.; Dieleman, S.; Battaglia, P.W. Generating images with sparse representations. *arXiv* **2021**, arXiv:2103.03841. [[CrossRef](#)]
35. Ye, N.; Tang, J.; Deng, H.; Zhou, X.Y.; Li, Q.; Li, Z.; Yang, G.Z.; Zhu, Z. Adversarial invariant learning. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA 20–25 June 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 12441–12449. [[CrossRef](#)]
36. Carraro, T.; Polato, M.; Aioli, F. A look inside the black-box: Towards the interpretability of conditioned variational autoencoder for collaborative filtering. In Proceedings of the Adjunct Publication of the 28th ACM Conference on User Modeling, Adaptation and Personalization, New York, NY, USA, 14–17 July 2020; pp. 233–236. [[CrossRef](#)]
37. Cheng, Y.C.; Lee, H.Y.; Sun, M.; Yang, M.H. Controllable image synthesis via segvae. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part VII 16; Springer: Berlin/Heidelberg, Germany, 2020; pp. 159–174. [[CrossRef](#)]
38. Yeung, S.; Kannan, A.; Dauphin, Y.; Feifei, L. Tackling over-pruning in variational autoencoders. *arXiv* **2017**, arXiv:1706.03643. [[CrossRef](#)]
39. Yang, B.; Li, J.; Wong, D.F.; Chao, L.S.; Wang, X.; Tu, Z. Context-aware self-attention networks. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 387–394. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.