Heritage Science

RESEARCH

Open Access

Automatic damage identification of Sanskrit palm leaf manuscripts with SegFormer

Yue Wang[1], Ming Wen[1], Xiao Zhou[2], Feng Gao[2], Shuai Tian[2], Dan Jue[3], Hongmei Lu[1,4*] and Zhimin Zhang[1*]

Abstract

Palm leaf manuscripts (PLMs) are of great importance in recording Buddhist Scriptures, medicine, history, philosophy, etc. Some damages occur during the use, spread, and preservation procedure. The comprehensive investigation of Sanskrit PLMs is a prerequisite for further conservation and restoration. However, current damage identification and investigation are carried out manually. They require strong professional skills and are extraordinarily time-consuming. In this study, PLM-SegFormer is developed to provide an automated damage segmentation for Sanskrit PLMs based on the SegFormer architecture. Firstly, a digital image dataset of Sanskrit PLMs (the PLM dataset) was obtained from the Potala Palace in Tibet. Then, the hyperparameters for pre-processing, model training, prediction, and post-processing phases were fully optimized to make the SegFormer model more suitable for the PLM damage segmentation task. The optimized segmentation model reaches 70.1% mHit and 51.2% mIoU. The proposed framework automates the damage segmentation of 10,064 folios of PLMs within 12 h. The PLM-SegFormer framework will facilitate the preservation state survey and record of the Palm-leaf manuscript and be of great value to the subsequent preservation and restoration. The source code is available at https://github.com/Ryan21wy/PLM_SegFormer.

Keywords Sanskrit palm leaf manuscript, Damage, Semantic segmentation, SegFormer

Introduction

Palm leaf manuscripts (PLMs) were an important writing medium in many Asian countries before the invention of papers [1, 2]. Sanskrit PLMs in Tibet are a kind of PLMs written in Sanskrit [3]. According to incomplete statistics, several museums and palaces in Tibet, such as Potala Palace and Norbulingka, preserve more than 60,000 folios of Sanskrit PLMs dating from about the third century AD to the thirteenth century AD. Many of them are the first-level cultural relics in China. After centuries of use, spread, and preservation, Sanskrit PLMs were inevitably aged and damaged [1, 4, 5].

There are many kinds of damage in Sanskrit PLMs. Among these damages, incompleteness, break, fiber delamination and warping, contamination, and improper restoration are five frequent damages. Incompleteness (Fig. 1b) refers to the lack of the main body of palm leaves. Break (Fig. 1c) refers to the transverse or longitudinal breaks formed along the texture of palm leaves by external force or excessive drying. Fiber delamination and warping (Fig. 1d) refers to the delamination of the fiber layers of palm leaves or the separation of the fiber layers from the leaf body. Contamination (Fig. 1e) refers to stains and traces formed on the surface of PLMs. Improper restoration (Fig. 1f) refers to manually restoring the damaged PLMs with inappropriate materials and methods.

*Correspondence:
Hongmei Lu
hongmeilu@csu.edu.cn
Zhimin Zhang
zmzhang@csu.edu.cn
[1] College of Chemistry and Chemical Engineering, Central South University, Changsha 410083, People's Republic of China
[2] Chinese Academy of Cultural Heritage, Beijing 100028, People's Republic of China
[3] Administrative office, the Potala Palace, Lhasa 850015, People's Republic of China
[4] Hunan Key Laboratory for Scientific Archaeology and Conservation Science, Changsha 410083, People's Republic of China

In 2019, the Chinese government launched the protection project of Sanskrit PLMs in Tibet. During the protection and restoration of Sanskrit PLMs, a comprehensive survey and record of the preservation state of Sanskrit PLMs is a prerequisite. The current survey of Sanskrit PLMs entirely relies on manual work, which requires strong professional skills and is extraordinarily time-consuming. And the manual identification of damages of Sanskrit PLMs is subjective and labor-intensive. Therefore, to develop a computer-aided damage identification is required for the efficient preservation state survey of Sanskrit PLMs.

With the view of image processing, the damage identification of Sanskrit PLMs can be seen as a semantic segmentation task, i.e., assigning the correct category label to each pixel in digital images of Sanskrit PLMs. Recently, deep learning methods have become prominent and potent in semantic segmentation [6–9]. They have been introduced to historical document analysis, such as binarization [10–12], text line segmentation [13–15], page segmentation [16, 17], Layout Analysis [18–20], and character recognition [21–23]. Based on the digital images of historical handwritten documents, Xu et al. [16] applied fully convolution networks (FCN) to classify the pixels of the documents into different categories: background, main text body, comments, and decorations. As for PLMs, several researchers applied deep learning methods to recognize Palm Leaf Characters [22–27]. Devi et al. [23]. manually built cursive training datasets and utilized a unique convolutional neural network (CNN) technique to identify the palm leaf characters. Sudarsan et al. [26] used a combination of Log-Gabor with uniform rotational invariant LBP for feature extraction. Then, a stacked ResNet-LSTM architecture was used for the classification of palm leaf characters.

In this research, a damage segmentation dataset named PLM dataset is established for the damage identification of Sanskrit PLMs in Tibet. It consists of five common damages, including incompleteness, break, fiber delamination and warping, contamination, and improper restoration. SegFormer [9] is chosen as the base segmentation network because it can balance segmentation efficiency and accuracy well. Based on SegFormer, the PLM-SegFormer framework is proposed to automatically identify

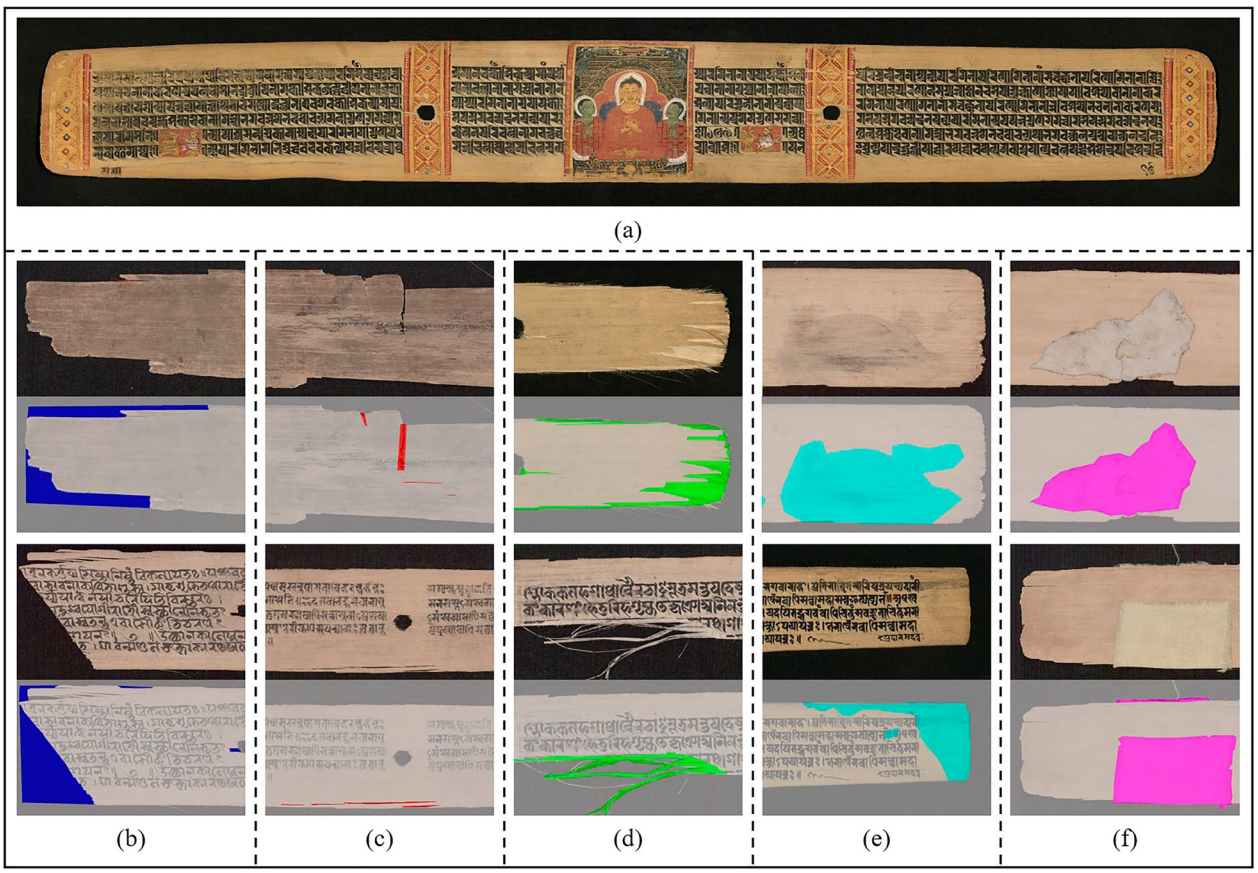

Fig. 1 The Sanskrit palm leaf manuscript in Tibet and five representative damages. **a** An example digital image of the Sanskrit PLM, **b** incompleteness, **c** break, **d** fiber delamination and warping, **e** contamination, **f** improper restoration. For each pair, the PLM image is shown at the top, and the corresponding manual annotation is at the bottom

damages of Sanskrit PLMs with their high-resolution digital images. The proposed PLM-SegFormer framework successfully adapts the SegFormer architecture to the PLM damage segmentation with significant improved performance. Furthermore, it can automatically and quickly complete the damage segmentation of a large number of PLMs and obtain the damage distribution information for subsequent protection and restoration of Sanskrit PLMs.

Methods

Overview of the damage segmentation framework

Figure 2 is the flowchart of the PLM-SegFormer framework. The framework includes two parts: training phase and inference phase. It consists of five steps: data collection and labeling, pre-processing, model training, prediction, and post-processing.

Image acquisition and damage labeling

The Nikon D5300 camera was selected as the image acquisition instrument of Sanskrit PLMs. A Camera Tripod was used to assist the acquisition with fixed space and angle. A black paperboard was placed underneath the Sanskrit PLMs during the image acquisition. The horizontal resolution and the vertical resolution of the images were both 300 dpi. The height and width of the images were in the range of [362, 2053], [2411, 4739] pixels, respectively. The aspect ratio (the ratio of width to height) of the images was in the range of [1.92, 10.82]. In total, 338 images were captured.

Five frequent damages, incompleteness, fiber delamination and warping, break, contamination, and improper restoration, were considered as the targets in this study. All the damages of Sanskrit PLMs were labeled manually by experts. The image polygonal annotation tool LabelMe (v4.5.6) [28] was used for damage labeling. The raw images with manual annotation were considered as the PLM dataset. Then, the PLM dataset was divided into the training set, validation set, and test set in the ratio of 6:2:2.

Pre-processing

The size of PLM images varies significantly from one another, and feeding large-size images directly into the model leads to out-of-memory (OOM) errors. Therefore, it is crucial to pre-process the original images to make them suitable for model training. Three pre-processing strategies were considered here: cropping, resizing, and resizing and cropping. The high-quality Lanczos filter from the Pillow package was used for image resizing.

Cropping. A common way to handle large-size images in semantic segmentation is to crop the original image into equal-sized patches. Then, the image patches are used to train the segmentation model [6]. All the PLM images were cropped into non-overlapped 512×512 patches, and the patches less than 512×512 were filled by adjacent pixels

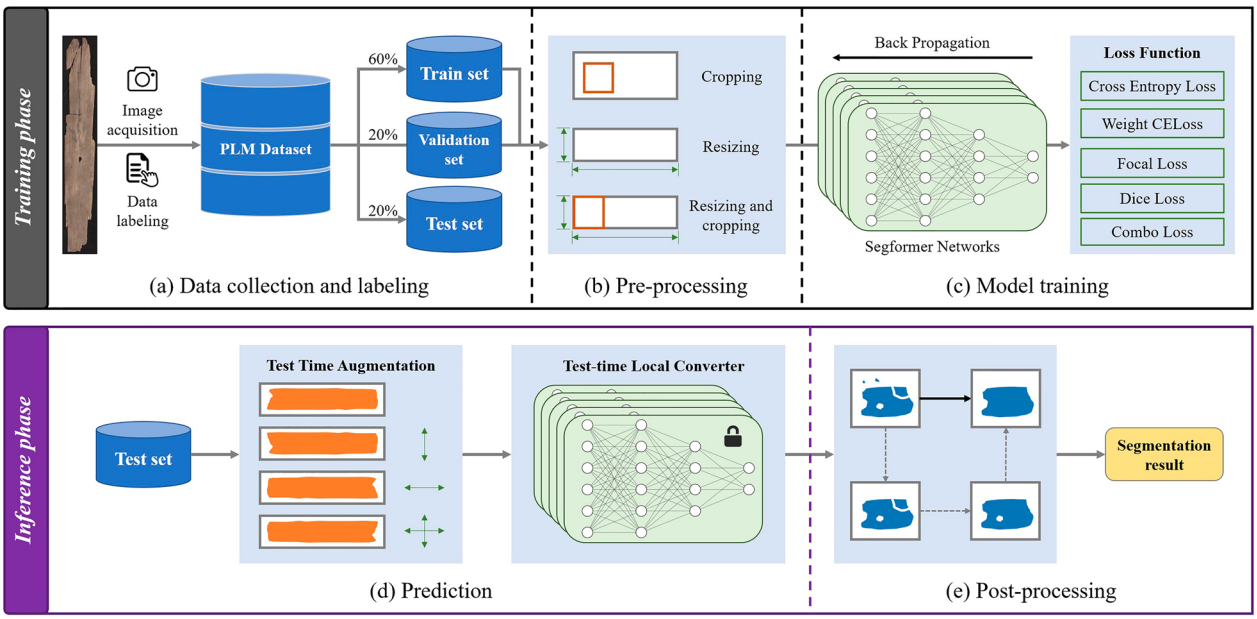

Fig. 2 The flowchart of the PLM-Segformer framework. **a** The PLM dataset is established by digital camera acquisition and manual annotation. It has been subsequently divided into the training set, validation set, and test set. Then, **b** various pre-processing methods and **c** loss functions are compared to find the best way to build the damage segmentation models. Finally, **d** test-time enhancement methods and **e** post-processing methods are used to optimize the prediction results in the inference phase

(e.g., a 512×768 image was cropped into two 512×512 patches with 512×256 overlapping area). For images with the size less than 512 pixels, their short sides were first resized into 512 while maintaining the aspect ratio. Then, the resized images were cropped into patches. In addition, a larger crop size (512×768) was also applied to investigate the effect of the crop size.

Resizing. Another way was to resize the images to a trainable size. Then the full images was directly used to train the segmentation model. There were two methods for image resizing: (I) one was to resize the short side of the image to 512, which maintained the original aspect ratio of the image; (II) another was to resize all the images to a fixed size, which was 512×976 or 512×3072 according to the minimum aspect ratio or average aspect ratio.

Resizing and Cropping. A resizing and cropping strategy that combined resizing and cropping was proposed. Firstly, the short sides of all the PLM images were first resized to 512 while maintaining the aspect ratio. Then, the resized images were cropped into 512×512 patches. Considering the significant reduction in image size after resizing, overlapping crop with an overlap area of half the patch size was used to increase the number of image patches. Similar to the cropping method, a larger crop size (512×768) was also applied.

Model training

In this study, SegFormer was used for the damage segmentation of PLMs. The details of the network architecture can be viewed in Additional file 1. A series of binary segmentation models were developed to predict each type of damage. As shown in Table 1, the damaged area is very small compared to the non-damage area of PLMs. The pixel percentages of the five damage area and the non-damage area in the dataset are 1.156%, 0.118%, 0.516%, 4.768%, 0.496%, and 92.946%, respectively. This extreme class imbalance causes the segmentation model to be biased toward non-damage area. This may severely affect the performance of the segmentation model. Here, different loss functions were used to find a suitable way to handle the class imbalance problem.

Cross-entropy loss. Cross-entropy (CE) loss is the most commonly used loss function in semantic segmentation tasks due to its simplicity and effectiveness. It examines each pixel individually by comparing the class prediction with the one-hot coded ground truth label. It is calculated by:

$$L_{CE} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{C} y_{n,c}\log(p_{n,c}), \tag{1}$$

Table 1 The image counts and pixel percentages of each damage and non-damage in the PLM dataset

Class index	Type	Image count	Pixel percentage (%)
1	Incompleteness	264	1.156
2	Break	217	0.118
3	Fiber delamination and warping	212	0.516
4	Contamination	322	4.768
5	Improper restoration	63	0.496
	Non-damage	338	92.946

where N is the number of pixels, C is the number of classes, $y_{n,c}$ is the one-hot coded ground truth label for the class c, and $p_{n,c}$ is the class prediction.

Weight cross-entropy loss. Since the cross-entropy loss evaluates the class predictions for each pixel individually and then averages over all pixels, the training procedure can be dominated by the majority class if there are imbalanced classes. A common solution is to turn standard cross-entropy loss into weighted cross-entropy (WCE) loss by adding a weighting factor to focus more on minority classes. The WCE loss is defined as:

$$\begin{aligned} L_{WCE} &= -\frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{C} w_c y_{n,c}\log(p_{n,c}), \\ w_c &= \frac{(1/f_c)^{\lambda}}{\sum_{i=1}^{C}(1/f_i)^{\lambda}}, \end{aligned} \tag{2}$$

where w_c is the class weight, f_c is the pixel percentage of class c, and λ controls the weighting factor. As λ increases, the weight value of minority classes increases. In this study, λ was selected in the range of [0, 1].

Focal loss. Focal loss [29] was initially used in image classification to solve the class imbalance problem. To reduce the influence of class imbalance, a modulating factor was added to the standard cross entropy loss function, aiming to put more focus on hard, misclassified pixels. The focal loss is defined as:

$$\begin{aligned} L_{focal} &= -\frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{C}(1-p_t)^{\gamma} y_{n,c}\log(p_{n,c}), \\ p_t &= y_{n,c}p_{n,c}, \end{aligned} \tag{3}$$

where γ controls the degree of down-weighting of easy-to-classify pixels. As γ increases, the degree of down-weighting increases. In this study, γ was set to 2 according to the reference [29].

Dice loss. Dice loss is based on the Dice coefficient, which measures the overlap between two data sets ranging from 0 to 1. A Dice coefficient of 1 denotes complete overlap. In practice, Dice coefficient was used as the loss function [30] to minimize the non-overlap between the prediction and the ground-truth label. Dice loss can be calculated as follows:

$$L_{Dice} = 1 - \frac{1}{C}\sum_{c=1}^{C}\frac{\sum_{n=1}^{N} y_{n,c} p_{n,c}}{\sum_{n=1}^{N} y_{n,c} + \sum_{n=1}^{N} p_{n,c}}. \quad (4)$$

Combo loss. The training procedure of Dice loss will be unstable when segmenting a small foreground from a large background. Thus, a combination of CE loss and Dice loss was used to stabilize the training procedure [31]. The combo loss is defined as:

$$L_{Combo} = (1-\lambda)L_{CE} + \lambda L_{Dice}, \quad (5)$$

where $\lambda \in [0, 1]$ controls the relative contribution of CE loss and Dice loss terms to the overall loss function.

Training setting. The MiT-B1 [9] network was used as the segmentation model. The last batch normalization layer was replaced with the group normalization (GN) [32] layer to improve the stability of the training procedure under a small batch size. Data augmentation was performed through random resized crop with a ratio of 0.8–1.5, random horizontal flipping, and random vertical flipping. The Adan [33] optimizer was used to train the models for 200 epochs with a learning rate of 0.0003 and a weight decay of 0.01. The batch size was set as 8. A cosine decay learning rate scheduler was used with a linear warm-up for 10 epochs.

Prediction

In this study, test time augmentation (TTA) and test-time local converter (TLC) were used in the inference phase for better performance. TTA duplicated and mirrored the input image along the horizontal, vertical, and diagonal axes. Given an image as input, four augmented images were obtained. Then, the predicted results of the augmented images were averaged as the final result. For patch-based training, it was common to use the full image directly as input during prediction. However, the information distribution inconsistency between patch-based training and full-image-based prediction led to performance degradation. TLC [34] was proposed to solve this problem by converting the spatial information aggregation operation from global to local. Before the global operation, TLC divided the feature map into patches in the spatial dimension. Each feature patch was operated separately, and then the feature patches were stitched back together according to their original spatial position. In SegFormer, the self-attention layer belonged to the spatially global operation, and TLC was applied to this operation during prediction.

Post-processing

Due to the complexity of the damages and the limited performance of the segmentation models, the damage segmentation results usually had some misclassified regions, such as small noise regions and discontinuous regions. Therefore, post-processing was used in the inference phase to alleviate these problems. First, regions with the area less than a given threshold were considered as noisy regions and were removed. Then, the morphological close operation was applied to connect the adjacent area. Finally, the small holes in the connected regions were filled.

Evaluation metrics

Two evaluation metrics were used in PLM damage segmentation for qualitative discovery and quantitative evaluation, respectively. In damage segmentation, finding the location of the damage region is as important as

Table 2 Comparison of different pre-processing methods on the PLM validation set

Method		Class IoU (%)					mIoU (%)
		INC	BRE	FIB	CON	IMP	
Cropping	512×512	49.5	33.3	41.2	29.3	69.2	44.5
	512×768	42.2	**36.1**	35.3	23.6	85.0	44.4
Resizing	512×w	47.2	35.0	31.6	27.6	**91.9**	46.7
	512×976	45.0	31.0	22.4	22.9	89.2	42.1
	512×3072	41.3	31.4	23.2	23.7	89.1	41.7
Resizing and cropping	512×512	46.3	31.6	38.9	26.1	87.9	46.2
	512×768	**54.8**	34.7	**41.4**	**29.8**	91.8	**50.5**

"512×w" indicates that the short side of the image is resized to 512, and the aspect ratio of the image is maintained. "INC", "BRE", "FIB", "CON", and "IMP" indicate incompleteness, break, fiber delamination and warping, contamination, and improper restoration, respectively. The CE loss was the default setting when comparing the different pre-processing methods. Best scores are highlighted in **bold**

Table 3 Comparison of different loss functions on the PLM validation set

Method	class IoU (%)					mIoU (%)
	INC	BRE	FIB	CON	IMP	
CE loss	54.8	34.7	41.4	29.8	91.8	50.5
WCE loss	56.5	41.6	42.6	**34.6**	**93.7**	53.8
Focal loss	54.4	39.6	33.9	28.5	93.0	49.9
Dice loss	26.1	8.2	11.1	24.7	16.6	17.3
Combo loss	**57.0**	**42.7**	**43.4**	33.8	93.5	**54.1**

For comparing loss functions, the resizing and cropping method with a crop size of 512 × 768 was used in all experiments. Best scores are highlighted in **bold**

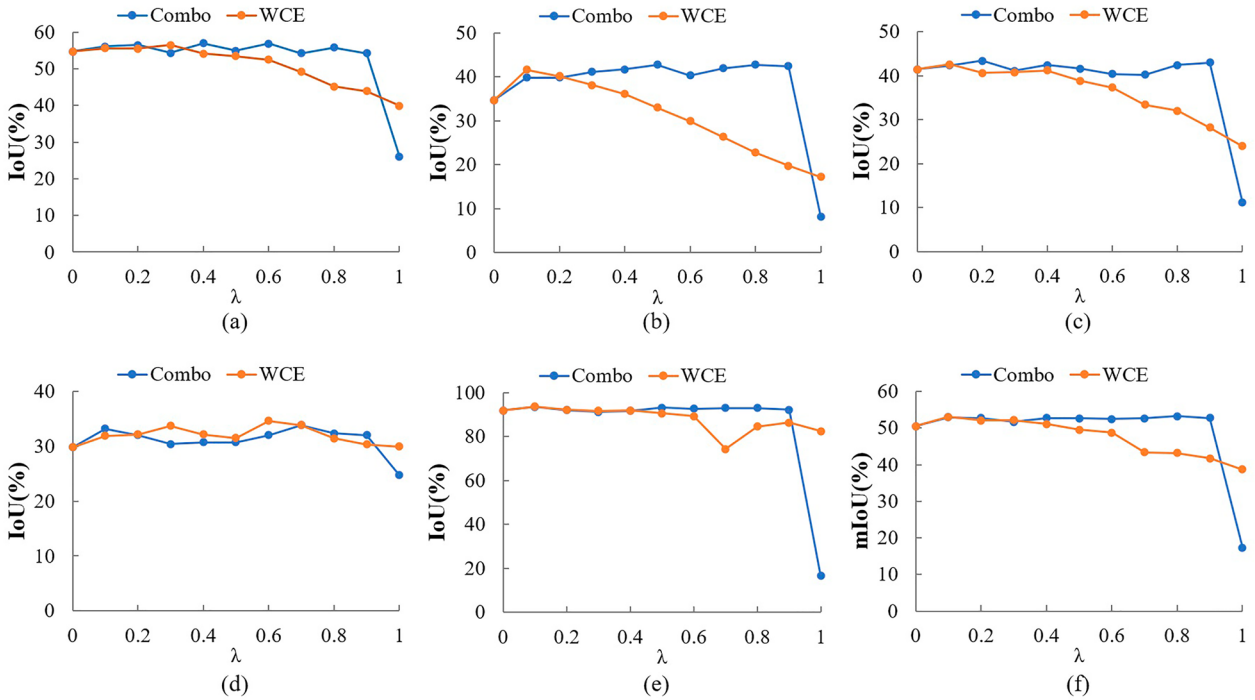

Fig. 3 Results of different λ in combo loss and WCE loss on the PLM validation set. **a** Incompleteness; **b** break; **c** fiber delamination and warping; **d** contamination; **e** improper restoration; **f** mean IoU of the five damages. When $\lambda = 0$, the combo loss and WCE loss simplify to the CE loss; when $\lambda = 1$, the combo loss is equal to the Dice loss. The results show that WCE loss is more sensitive to the choice of hyperparameter λ than combo loss

the precise segmentation of the damage regions. A qualitative discovery metric named hit area ratio (Hit) was proposed to evaluate the ability of damage region localization. If the recall value between the segmented region of the ground truth and the predicted result is greater than 0.5, this region is regarded as a Hit region. Hit is the ratio of the total area of hit regions to the total area of the ground truth and mHit is the mean value of Hit of all damages. Hit and mHit are defined as:

$$\begin{aligned} \text{Recall} &= \frac{\text{TP}}{\text{TP} + \text{FN}}, \\ S_{\text{r}} &= \begin{cases} S_{\text{region}}, & \text{if Recall}_{\text{r}} \geq 0.5 \\ 0, & \text{if Recall}_{\text{r}} < 0.5 \end{cases}, \\ \text{Hit} &= \sum_{r=1}^{R} \frac{S_r}{S}, \\ \text{mHit} &= \frac{1}{C} \sum_{c=1}^{C} \text{Hit}_{\text{c}}, \end{aligned} \tag{6}$$

where C is the number of damages, S_{region} is the area of the ground truth region, S is the area of the total regions

Table 4 Comparison of different inference phase strategies on the PLM validation set

Method	Hit (%)					class IoU (%)					mHit (%)	mIoU (%)
	INC	BRE	FIB	CON	IMP	INC	BRE	FIB	CON	IMP		
None	76.8	**71.2**	71.6	54.1	97.8	57.0	42.7	**43.4**	33.8	93.5	54.1	71.6
TTA	77.9	**71.2**	71.7	54.5	97.8	57.4	**43.3**	**43.4**	**35.3**	93.2	54.5	71.7
TLC	76.8	72.3	70.8	54.1	**99.6**	57.5	42.8	42.8	33.5	93.9	54.1	70.8
TTA + TLC	**79.6**	74.7	72.9	**54.9**	98.4	59.3	43.3	42.9	35.1	93.9	**54.9**	72.9
TTA + TLC + Post	**79.6**	74.8	**73.0**	54.8	98.4	**59.8**	42.3	42.5	35.2	**94.1**	54.8	**73.0**

"Post" indicates the post-processing method. The best scores are highlighted in **bold,** and the inference phase strategy used for each kind of damage is underlined

of the ground truth, and R is the total number of regions in the ground truth.

Interaction-over-Union (IoU) and mean IoU (mIoU) were used as the precise evaluation metrics. IoU is one of the mostly used metrics in semantic segmentation. It is the interaction set of the ground truth and the class prediction divided by the union of the ground truth and the class prediction of a specific damage. The mIoU is the mean value of IoU of all damages. IoU and mIoU are defined as:

$$\begin{aligned} \mathrm{IoU} &= \frac{\mathrm{TP}}{\mathrm{FN} + \mathrm{FP} + \mathrm{TP}}, \\ \mathrm{mIoU} &= \frac{1}{C} \sum_{c=1}^{C} \mathrm{IoU}_c, \end{aligned} \tag{7}$$

where C is the number of damages. TP, FP, TN, and FN represent the number of true positives, false positives, true negatives, and false negatives, respectively.

Results and discussion

Optimization of pre-processing method and loss function in the training phase

IoU and mIoU of the PLM validation set were used to compare different pre-processing methods and loss functions in the training procedure. When comparing the pre-processing methods, CE loss was used in all experiments.

The results of the different pre-processing methods are presented in Table 2. Regarding overall damage segmentation results, the resizing and cropping method with a crop size of 512×768 outperforms the other pre-processing methods by at least 3.8% mIoU. Compared to cropping methods, the image patches obtained by the resizing and cropping method gain more global information about the whole image at the expense of fine and local information. The importance of global information for damage segmentation can also be illustrated by the fact that the model performance can be improved by increasing the size of the image patches. Although the resizing methods preserve the most global information about the images, the small number of images in the training set (206) makes it difficult for models to learn damage features effectively. Furthermore, resizing the images to a fixed size yields poor mIoU because the images are distorted, and their original structures are lost.

In terms of single damage, break needs local and fine prediction because it often appears as elongated strips. Thus, the cropping method represents a 1.1% IoU improvement compared to the other pre-processing methods. However, global information is more important for the other four damage categories. From these results, the resizing and cropping method with a crop size of 512×768 is more suitable as the pre-processing method for the PLM damage segmentation and is used in subsequent experiments.

When the cropping size is expanded from 512×512 to 512×768, the mIoU of the resizing and cropping method obtains a substantial improvement of 4.3%. In contrast, the results of the resizing method are not improved and even worse. The reason for this phenomenon is unclear, which will be investigated in future works.

After determining the pre-processing method, different loss functions were compared to deal with the class imbalance problem. As shown in Table 3, combo loss achieves the highest mIoU of 54.1% and outperforms the CE loss in all damage classes. Compared to the CE loss, Focal loss achieves a 4.9% IoU improvement in the break but gets slightly worse results for the other damages. Notably, the performance of Focal loss underperforms CE loss by 7.5% IoU in fiber delamination and warping. Dice loss performs poorly and is worse than CE loss, which indicates that it is unsuitable for handling the severe class imbalance problem.

The effect of the values of the hyperparameters λ on both WCE loss and combo loss was investigated. For WCE loss and combo loss in the five damages, the optimal values of λ were [0.2, 0.1, 0.1, 0.5, 0.1] and [0.6, 0.5, 0.5, 0.9, 0.1], respectively. As shown in Fig. 3, WCE loss is more sensitive to the choice of hyperparameter λ than

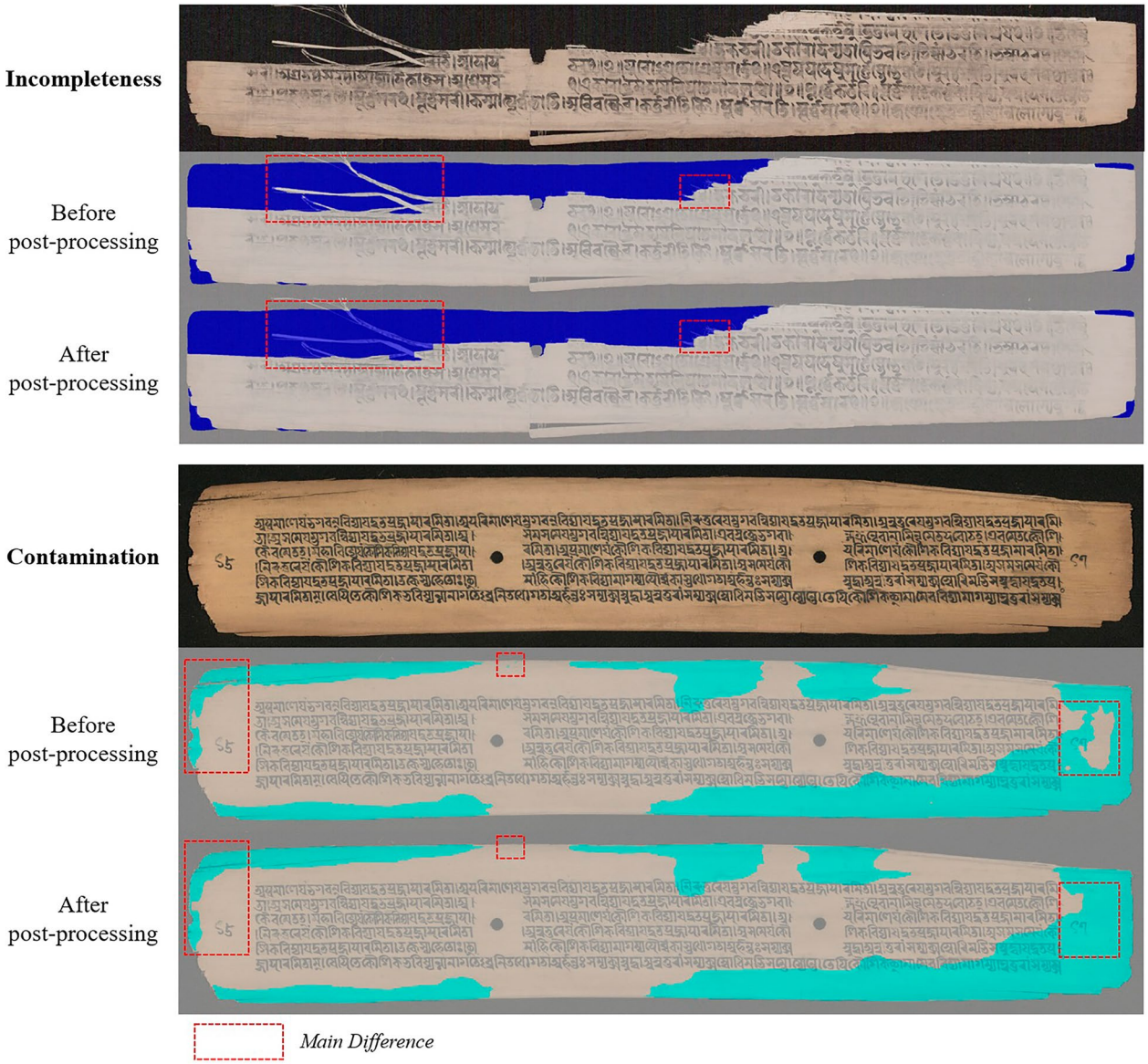

Fig. 4 Examples of differences in damage segmentation results before and after post-processing. The post-processing methods can deal with the small noise regions and discontinuous regions, but they also remove some details of the prediction results

combo loss. For WCE loss, as the λ increases, the IoU of each damage increases and then decreases. This decrease may be caused by giving too large weight to the damage class, resulting in many false positive regions. In particular, even given a small weight (0.1), CE loss can substantially improve the stability of Dice loss. Thus, combo loss was used as the default loss function.

Impact of inference phase strategy

Experiments were conducted to optimize the inference phase with several strategies, including TTA, TLC, and post-processing. IoU, mIoU, Hit, and mHit were used to evaluate the inference phase strategies. Since the inference phase takes significantly less time than the training phase, different combinations of inference phase strategies can be used for different damage types through multiple rounds of experiments.

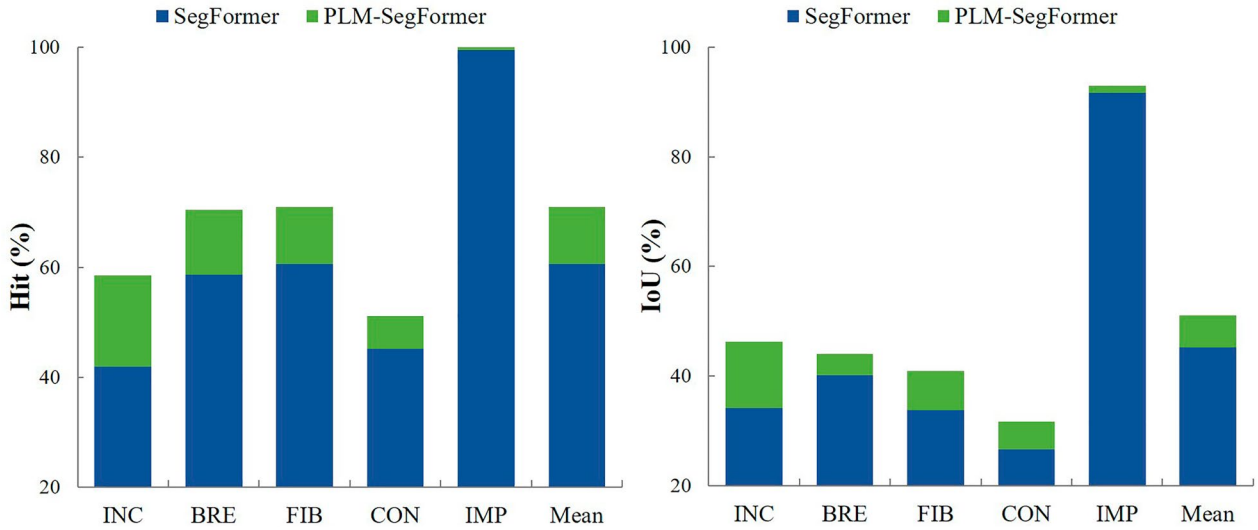

Fig. 5 Compare the performance of the PLM-SegFormer framework and the SegFormer baseline on the PLM test set. The SegFormer baseline models were trained by the cropping method with a size of 512 × 768 and the CE loss. The proposed PLM-SegFormer framework combines resizing and cropping methods for pre-processing, combo loss for training, and optimized post-processing methods with SegFormer models

As shown in Table 4, TTA slightly improves all damage segmentation results except improper restoration, resulting in 0.1% higher mHit and 0.4% higher mIoU than baseline (no extra strategies). TLC shows performance gains in three damages, but performance declines in other two damages, resulting in final performance slightly below the baseline. The combination of TTA and TCL brings a performance boost with a gain of 1.3% mHit and 0.8% mIoU. This combination significantly improves the performance in incompleteness by 2.8% Hit and 2.3% IoU, demonstrating an excellent synergistic effect between TTA and TLC. Integrating the post-processing methods can achieve performance gains in incompleteness, contamination, and improper restoration but lose the performance of IoU in break and fiber delamination and warping.

Two examples before and after post-processing are shown in Fig. 4. One can see that the post-processing method can deal with the small noise regions and some discontinuous regions, and improve the visual perception of the segmentation results. Meanwhile, it removes some details of prediction results.

Results of PLM damage identification

The best inference phase strategy chosen for each damage is underlined in Table 4, which mainly refers to IoU. To illustrate the effectiveness of the PLM-SegFormer framework, the SegFormer baseline was set using the cropping method with a crop size of 512×768 for data pre-processing and CE loss for model training. The results in Fig. 5 show that the PLM-SegFormer framework achieves consistency improvements over the Segformer baseline on five damages, especially for the incompleteness, which receives a substantial improvement in 16.6% Hit and 12.1% IoU (Additional file 1: Table S1). Furthermore, the PLM-SegFormer framework brings a performance boost with a gain of 10.4% mHit and 5.9% mIoU, which shows that the PLM-SegFormer framework can be well adapted to the damage segmentation of PLMs.

As a result, the PLM-SegFormer framework can reach 71.0 mHit and 51.2 mIoU on the PLM test set, indicating that the model can be used for damage identification of Sanskrit PLMs in Tibet.

The impact of each damage's characteristics on the performance

In this section, the impact of the characteristics of each type of damage and its segmentation results (Fig. 6) are discussed.

(a) *Incompleteness*: Incompleteness often appears in the edge area of PLMs. When incompleteness occurs in the length of the PLM, it tends to form a long-distance damage area. Thus, enough global information should be obtained for the segmentation of incompleteness. Moreover, since the boundary of the complete PLM is manually labeled, it is difficult for the segmentation model to accurately judge the boundary of the "imaginary" complete PLM. The challenge to determine where incompleteness occurs leads to a relatively lower Hit.

(b) *Break*: Break is shown as the horizontal or longitudinal fractures or cracks formed along the texture of

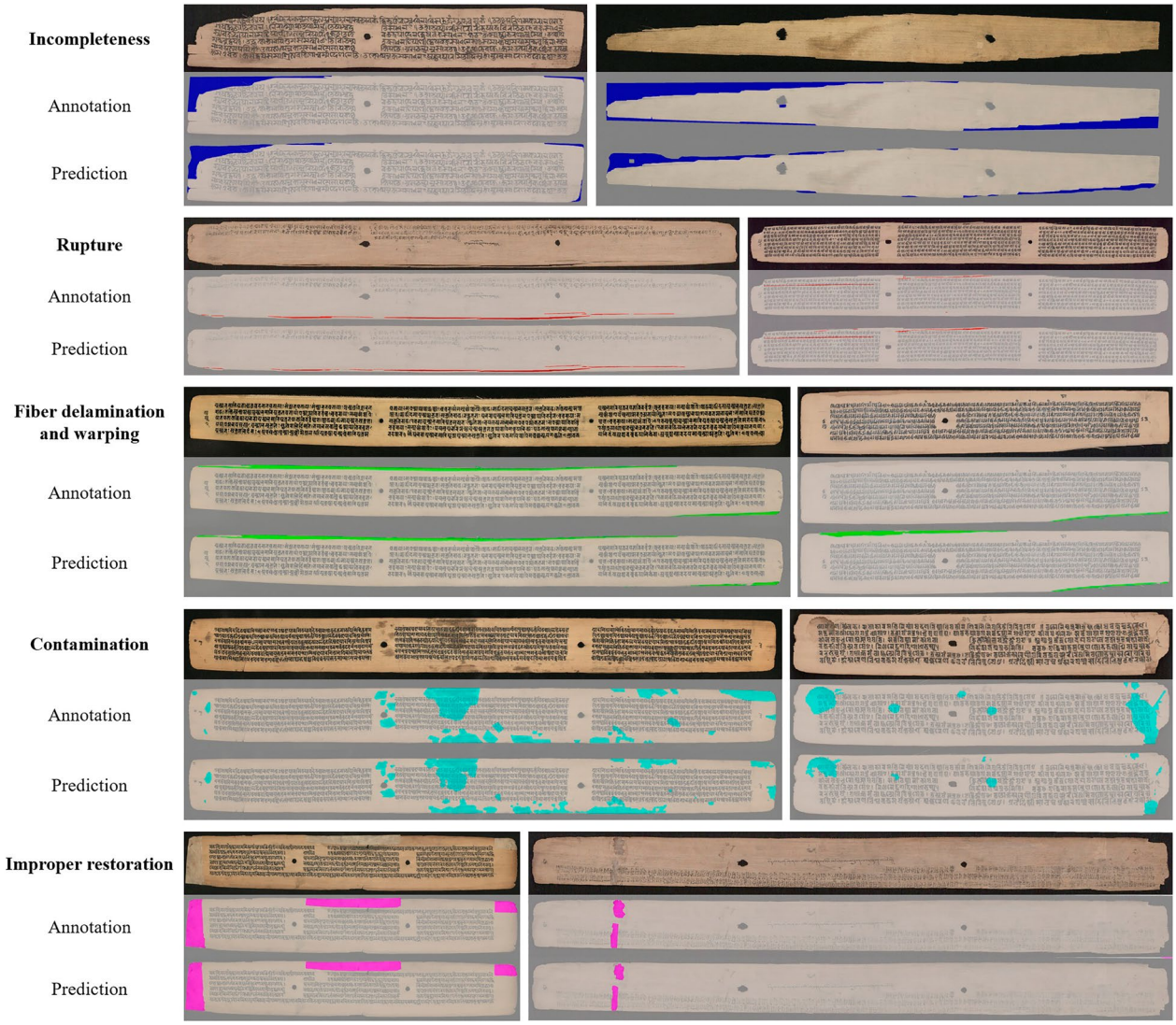

Fig. 6 Visual comparisons of annotation and segmentation results of the representative samples regarding five types of damages in the PLM test set

the palm leaf. Because the break regions are usually small, slight differences in the prediction region and the ground-truth label can significantly deteriorate evaluation metrics, especially for precise metrics like IoU. In addition, as shown in Table 1, the pixel percentage of the break regions is only 0.118%. This severe class imbalance degrades the performance of the segmentation model.

(c) *Fiber delamination and warping*: Fiber delamination and warping consists of the warped part and the remaining part. The warped part is the same as the body of PLMs but usually appears as slender and scattered strips. Therefore, fine-grained segmentation of the warped part is required to obtain satisfying results, which results in a high Hit but a low IoU, similar to the break. Due to the aging of the PLM, the surface of PLMs is usually darker in color compared to the remaining part. However, when contamination occurs together, the difference between the remaining part and the surface of PLMs is reduced, which increases the recognition difficulty of the segmentation model.

(d) *Contamination*: Contamination is the most challenging type of damage to identify. On the one hand, as a general term for a class of damages, contamination exists in various forms, such as water infiltration, tea infiltration, and stains. On the other hand, the gradual change of the contamination

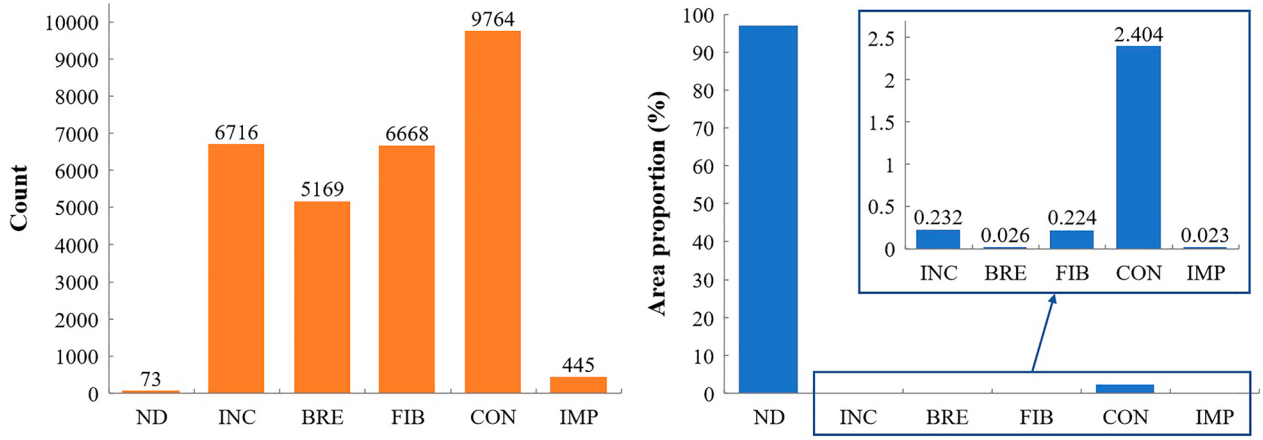

Fig. 7 Damage statistics on 10,064 digital images of the Sanskrit PLMs by the PLM-SegFormer framework. "ND", "INC", "BRE", "FIB", "CON", and "IMP" indicate non-damage, incompleteness, break, fiber delamination and warping, contamination, and improper restoration, respectively

color and its similarity to the color of aging PLMs lead to incomplete or even inaccurate annotation, therefore the performance of the segmentation model was seriously deteriorated.

(e) *Improper restoration*: Improper restoration is defined that the materials different from the body material of the PLM is used to repair. The repairing material is different from PLMs in color and often has a regular shape. Therefore, the segmentation model can reasonably identify the regions of improper repair, resulting in high values of all evaluation metrics.

Automatic damage segmentation on a large number of Sanskrit PLMs

From 2021 to 2022, 10,064 digital images of Sanskrit PLMs were collected to investigate the damage distribution information in Tibet. The resizing and cropping strategy was used as the pre-processing method. The short sides of all PLM images were resized to 512 while maintaining the aspect ratio. Thus, the aspect ratio had the greatest impact on the inference speed of the segmentation model. The minimum, maximum, and average aspect ratios of the PLM images were 1.89, 11.48, and 6.84, respectively.

The damage segmentation task was implemented on the PyTorch platform using a workstation with an i9-10900X CPU, 32GB RAM, and an NVIDIA RTX 3080 10GB GPU. The developed PLM-SegFormer framework can complete the damage segmentation of 10,064 PLM images within 12 h, significantly reducing the time required for investigating the damage information of the Sanskrit PLMs. In addition, the minimum, maximum, and average time costs for one image segmentation were 1.87 s, 6.44 s, and 4.08 s, respectively. The level of automation can be seen in Additional file 2: Video S1. During the entire damage segmentation process, the system is fully automated and does not require any human intervention.

After the segmentation, the distribution of each type of damages in the PLMs can be summarized. It can be seen from the results (Fig. 7) that, among all the damages, the image counts and pixel percentages of contamination are the highest, while that of improper restoration is the lowest. The number of non-damage PLMs is only 73 indicating that the existing Sanskrit PLMs in Potala Palace are seriously affected by various damages. However, the pixel percentage of all damages is low, only 2.9%, which indicates that the damage degree of PLMs is not high. Therefore, the preservation and restoration of Sanskrit PLMs should be carried out in time to prevent the deterioration of damages. The results of damage segmentation will facilitate the preservation state survey and record of the Palm-leaf manuscript, which is of great value to the following preservation and restoration.

Conclusion

In this study, a damage segmentation dataset for Sanskrit PLMs was created. The PLM-SegFormer framework was proposed for damage identification of the Sanskrit PLMs. The presented PLM dataset annotates five frequent damages, including incompleteness, break, fiber delamination and warping, contamination, and improper restoration. The PLM-SegFormer framework builds upon the SegFormer architecture and adapts it to damage segmentation of Sanskrit PLMs by optimizing the overall workflow, through pre-processing, model training, prediction, and post-processing.

The experimental results show that the resizing and cropping method, Combo loss for model training, are suitable for dealing with the inconsistent size problem and the class imbalance problem in the PLM dataset. The combination of TTA, TLC, and post-processing methods in the inference phase can further boost the performance of the damage segmentation models and reach 70.1% mHit and 51.2% mIoU. The developed PLM-SegFormer framework can complete 10,064 pages of PLM damage segmentation within 12 h, significantly reducing the time required for investigating the damage information of the Sanskrit PLMs. The proposed method will facilitate the preservation state survey and record of the Palm-leaf manuscript and be of great value to the following preservation and restoration.

Limits and outlook

The most significant barrier to the PLM damage semantic segmentation is the lack of accurate ground truth of labeled damages. The reasons come from two aspects. One is that it is hard to decide the boundary or the category of damages. The boundary of some damages is blurring, and some damages are overlapped, which leads to inaccurate damage annotation. The other one is that the annotations are labor-intensive and time-consuming, and the number of elaborately labeled PLM images is very small. The weakly supervised learning and self-supervised learning methods should be an option to handle the noise data problem and leverage a large amount of unlabeled data for future works.

Abbreviations

PLM	Palm leaf manuscript
FCN	Fully convolution networks
CNN	Convolution neural networks
OOM	Out-of-memory
CE	Cross-entropy
WCE	Weighted cross-entropy
GN	Group normalization
TTA	Test time augmentation
TLC	Test-time local converter
Hit	Hit area ratio
mHit	Mean hit area ratio
IoU	Interaction-over-union
mIoU	Mean interaction-over-union
TP	True positive
FP	False positive
TN	True negative
FN	False negative
ND	Non-damage
INC	Incompleteness
BRE	Break
FIB	Fiber delamination and warping
CON	Contamination
IMP	Improper restoration

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40494-023-01125-w.

Additional file 1: **Fig. S1**. The architecture of SegFormer. **a** The SegFormer framework consists of two main modules: A hierarchical Mix Transformer encoder to extract coarse and fine features. A lightweight decoder to up-sample and fuse these multi-level features and predict the semantic segmentation mask. **b** The detail of the transformer block in SegFormer. "Self-Attn" indicates the self-attention layer, and "FFN" indicates the feed-forward network. **Table S1.** Compare the performance of the PLM-SegFormer framework and the SegFormer baseline on the PLM test set. The SegFormer baseline models were trained by the cropping method with a size of 512 × 768 and the CE loss. The proposed PLM-SegFormer framework combines resizing and cropping method for pre-processing, combo loss for training, and optimized post-processing methods with the SegFormer models.

Additional file 2. Automatic damage segmentation on 10,064 Sanskrit PLMs with PLM-SegFormer framework.

Acknowledgements

We are grateful for resources from the High-Performance Computing Center of Central South University.

Author contributions

HL and ZZ conceived the idea. YW developed the PLM-SegFormer framework and was a major contributor in writing the manuscript. MW completed the damage labeling work and participated in writing the manuscript. DJ and XZ obtained the right to collect the image data of Palm leaf manuscripts. ST and FG collected the raw Palm leaf manuscripts data. HL supervised the project. All authors read and approved the final manuscript.

Funding

This work is financially supported by the National Natural Science Foundation of China (Grant Nos. 22273120, 21873116, and 22373117), the National Cultural Heritage Administration, and the Fundamental Research Funds for the Central Universities of Central South University (Grant No. 1053320212039).

Availability of data and materials

Raw data for the palm leaf manuscripts dataset was acquired from the Potala Palace. The authors do not have permission to share the data. We provide the Python code of the method. It is available at https://github.com/Ryan21wy/PLM_SegFormer.

Declarations

Competing interests

The authors declare that they have no competing interests.

Received: 19 September 2023 Accepted: 18 December 2023
Published online: 02 January 2024

References

1. Sah A. Palm Leaf manuscripts of the world: material, technology and conservation. Stud Conserv. 2002;47:15–24. https://doi.org/10.1179/sic.2002.47.Supplement-1.15.
2. Kumar DU, Sreekumar G, Athvankar U. Traditional writing system in southern India—palm leaf manuscripts. Design Thoughts. 2009;7:2–7.
3. Meinert C. Transfer of buddhism across central asian networks (7th to 13th Centuries). Leiden: Brill; 2016.
4. Crowley AS. Repair and conservation of palm-leaf manuscripts. Restaurator. 1970;1:105–14. https://doi.org/10.1515/rest.1970.1.2.105.
5. Wiland J, Brown R, Fuller L, Havelock L, Johnson J, Kenn D, Kralka P, Muzart M, Pollard J, Snowdon J. A literature review of palm leaf manuscript conservation—Part 1: a historic overview, leaf preparation, materials and

media, palm leaf manuscripts at the British Library and the common types of damage. J Inst Conserv. 2022;45:236–59. https://doi.org/10.1080/19455224.2022.2115093.
6. Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. In: Navab N, Hornegger J, Wells WM, Frangi AF, editors. Medical image computing and computer-assisted intervention–MICCAI 2015. Berlin: Springer; 2015. p. 234–41. https://doi.org/10.1007/978-3-319-24574-4_28.
7. Shelhamer E, Long J, Darrell T. Fully convolutional networks for semantic segmentation. IEEE Trans Pattern Anal Mach Intell. 2017;39:640–51. https://doi.org/10.1109/tpami.2016.2572683.
8. Wang W, Xie E, Li X, Fan D-P, Song K, Liang D, Lu T, Luo P, Shao L. Pyramid vision transformer: a versatile backbone for dense prediction without convolutions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021. p. 568–78.
9. Xie E, Wang W, Yu Z, Anandkumar A, Alvarez JM, Luo P. SegFormer: simple and efficient design for semantic segmentation with transformers. Adv Neural Inf Process Syst. 2021;34:12077–90.
10. Tensmeyer C., Martinez T. Document image binarization with fully convolutional neural networks. In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). 2017. p. 99–104. https://doi.org/10.1109/ICDAR.2017.25.
11. Tensmeyer C, Martinez T. Historical document image binarization: a review. SN Comput Sci. 2020;1:173. https://doi.org/10.1007/s42979-020-00176-1.
12. BJ BN, Nair AS. Ancient horoscopic palm leaf binarization using a deep binarization model-RESNET. In: 2021 5th International Conference on Computing Methodologies and Communication (ICCMC). 2021. p. 1524–9. https://doi.org/10.1109/ICCMC51019.2021.9418461.
13. Hu P, Chen Y, Hao Y, Wang Y, Wang W. Text line segmentation based on local baselines and connected component centroids for Tibetan historical documents. J Phys Conf Ser. 2020. https://doi.org/10.1088/1742-6596/1656/1/012034.
14. Renton G, Soullard Y, Chatelain C, Adam S, Kermorvant C, Paquet T. Fully convolutional network with dilated convolutions for handwritten text line segmentation. Int J Doc Anal Recognit. 2018;21:177–86. https://doi.org/10.1007/s10032-018-0304-3.
15. Chamchong R, Fung CC. Text line extraction using adaptive partial projection for palm leaf manuscripts from Thailand. In: 2012 International Conference on Frontiers in Handwriting Recognition. 2012. p. 588–93. https://doi.org/10.1109/ICFHR.2012.280.
16. Xu Y, He W, Yin F, Liu CL. Page segmentation for historical handwritten documents using fully convolutional networks. In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). 2017. p. 541–6. https://doi.org/10.1109/ICDAR.2017.94.
17. Can YS, Kabadayı ME. CNN-based page segmentation and object classification for counting population in ottoman archival documentation. J Imaging. 2020;6:32.
18. Xu Y, Yin F, Zhang Z, Liu C-L. Multi-task layout analysis for historical handwritten documents using fully convolutional networks. In: Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI). 2018. p. 1057–63. https://doi.org/10.24963/ijcai.2018/147.
19. Zhao P, Wang W, Cai Z, Zhang G, Lu Y. Accurate fine-grained layout analysis for the historical tibetan document based on the instance segmentation. IEEE Access. 2021;9:154435–47. https://doi.org/10.1109/ACCESS.2021.3128536.
20. Tarride S, Lemaitre A, Coüasnon B, Tardivel S. Combination of deep neural networks and logical rules for record segmentation in historical handwritten registers using few examples. Int J Doc Anal Recognit. 2021;24:77–96. https://doi.org/10.1007/s10032-021-00362-8.
21. Chamchong R, Fung CC. Character segmentation from ancient palm leaf manuscripts in Thailand. In: Proceedings of the 2011 Workshop on Historical Document Imaging and Processing. 2011. p. 140–5. https://doi.org/10.1145/2037342.2037366.
22. Sabeenian RS, Paramasivam ME, Anand R, Dinesh PM. Palm-leaf manuscript character recognition and classification using convolutional neural networks. In: Peng S-L, Dey N, Bundele M, editors. Computing and network sustainability. Berlin: Springer; 2019. p. 397–404. https://doi.org/10.1007/978-981-13-7150-9_42.
23. Devi SG, Vairavasundaram S, Teekaraman Y, Kuppusamy R, Radhakrishnan A. A deep learning approach for recognizing the cursive tamil characters in palm leaf manuscripts. Comput Intell Neurosci. 2022;2022:1–15. https://doi.org/10.1155/2022/3432330.
24. Kesiman M, Valy D, Burie J-C, Paulus E, Suryani M, Hadi S, Verleysen M, Chhun S, Ogier J-M. Benchmarking of document image analysis tasks for palm leaf manuscripts from Southeast Asia. J Imaging. 2018;4:43. https://doi.org/10.3390/jimaging4020043.
25. Haritha J, Balamurugan VT, Vairavel KS, Ikram N, Janani M, Indrajith K. CNN based character recognition and classification in tamil palm leaf manuscripts. In: 2022 International Conference on Communication, Computing and Internet of Things (IC3IoT). 2022. p. 1–6. https://doi.org/10.1109/IC3IOT53935.2022.9767866.
26. Sudarsan D, Sankar D. Development of an effective character segmentation and efficient feature extraction technique for malayalam character recognition from palm leaf manuscripts. Sādhanā. 2023;48:156. https://doi.org/10.1007/s12046-023-02181-5.
27. Bipin Nair BJ, Shobha Rani N, Khan M. Deteriorated image classification model for malayalam palm leaf manuscripts. J Intell Fuzzy Syst. 2023;45:4031–49. https://doi.org/10.3233/JIFS-223713.
28. Russell BC, Torralba A, Murphy KP, Freeman WT. LabelMe: a database and web-based tool for image annotation. Int J Comput Vision. 2008;77:157–73. https://doi.org/10.1007/s11263-007-0090-8.
29. Lin T-Y, Goyal P, Girshick R, He K, Dollar P. Focal loss for dense object detection. IEEE Trans Pattern Anal Mach Intell. 2020;42:318–27. https://doi.org/10.1109/tpami.2018.2858826.
30. Milletari F, Navab N, Ahmadi SA. V-Net: fully convolutional neural networks for volumetric medical image segmentation. In: 2016 Fourth International Conference on 3D Vision (3DV). 2016. p. 565–71. https://doi.org/10.1109/3DV.2016.79.
31. Taghanaki SA, Zheng Y, Kevin ZS, Georgescu B, Sharma P, Xu D, Comaniciu D, Hamarneh G. Combo loss: handling input and output imbalance in multi-organ segmentation. Comput Med Imaging Graph. 2019;75:24–33. https://doi.org/10.1016/j.compmedimag.2019.04.005.
32. Wu Y, He K. Group normalization. In: Proceedings of the European conference on computer vision (ECCV). 2018. p. 3–19.
33. Xie X, Zhou P, Li H, Lin Z,Yan S. Adan: adaptive nesterov momentum algorithm for faster optimizing deep models. arXiv e-prints. 2022; arXiv:2208.06677. https://doi.org/10.48550/arXiv.2208.06677.
34. Chu X, Chen L, Chen C, Lu X. Improving image restoration by revisiting global information aggregation. In: Avidan S, Brostow G, Cissé M, Farinella GM, Hassner T, editors. Computer vision–ECCV 2022. Berlin: Springer; 2022. p. 53–71. https://doi.org/10.1007/978-3-031-20071-7_4.

Publisher's Note