IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 41, No. 3, June, 2024, pp. 1485-1494
Journal homepage: http://iieta.org/journals/ts

Digital Reconstruction of Historical Cultural Landscapes Based on Image Recognition Technology

Chen Xiang[1], Yun Yang[2*], Tuo Zhou[3], Ting Wang[4]

[1] Department of Urban & Regional Planning, Faculty of Built Environment, Universiti Malaya, Kuala Lumpur 50603, Malaysia
[2] Director of Landscape Architecture Department, Faculty of Architecture, Chengdu College of Arts and Sciences, Chengdu 610401, China
[3] Department of Information Science, College of Information Engineering, Fuyang Normal University, Fuyang 236041, China
[4] Chengdu College of Arts and Sciences, Chengdu 610401, China

Corresponding Author Email: s2002999@siswa.um.edu.my

https://doi.org/10.18280/ts.410336

Received: 10 December 2023
Revised: 26 March 2024
Accepted: 3 May 2024
Available online: 26 June 2024

Keywords:
historical cultural landscapes, digital reconstruction, image recognition, Multi-scale Dilated Convolution (MSDC), YOLOv3, pyramid feature attention, Pixel2Mesh, 3D reconstruction

ABSTRACT

With the advancement of modern society and technology, the preservation and inheritance of historical cultural landscapes have become increasingly significant. These landscapes not only testify to the development of human civilization but are also an essential part of cultural heritage. However, the ravages of time, natural disasters, and human activities continually threaten these valuable cultural assets. To better preserve and pass on these landscapes, digital reconstruction using technological means has become a crucial method. The rapid development of image recognition technology offers new possibilities and solutions for the digital reconstruction of historical cultural landscapes. Although current digital reconstruction methods have improved in automation, they still require enhancements in recognition accuracy and three-dimensional reconstruction effects in complex scenes. Furthermore, the performance of existing methods in handling multi-scale and multi-perspective issues is not satisfactory. Therefore, this paper proposes a digital reconstruction method for historical cultural landscapes based on image recognition technology, comprising two main parts: historical cultural landscape target recognition based on Multi-Scale Dilated Convolution YOLOv3 (MSDC-YOLOv3) and three-dimensional reconstruction of historical cultural landscapes based on pyramid feature attention Pixel2Mesh. The MSDC-YOLOv3 technique enables more precise recognition of objects within historical cultural landscapes against complex backgrounds, while the pyramid feature attention Pixel2Mesh method achieves more efficient and accurate 3D reconstruction, providing detailed three-dimensional models. This research not only achieves technical breakthroughs, enhancing the precision and efficiency of recognition and reconstruction, but also holds significant value in the protection and inheritance of cultural heritage, offering new ideas and methods for future research in related fields.

1. INTRODUCTION

As modern society develops and technology advances, the preservation and inheritance of historical cultural landscapes have become increasingly important [1, 2]. Historical cultural landscapes not only testify to the development of human civilization but are also an integral part of cultural heritage. However, over time, natural disasters, human activities, and other factors continue to threaten these valuable cultural heritages [3-5]. To better protect and pass on historical cultural landscapes, using digital technology to reconstruct and preserve them has become an important means. The rapid development of image recognition technology offers new possibilities and solutions for the digital reconstruction of historical cultural landscapes [6-9].

Digital reconstruction of historical cultural landscapes not only aids in the protection of cultural heritage but also supports the development of academic research, educational dissemination, and the tourism industry [10-13]. By reconstructing historical cultural landscapes through digital means, historical scenes can be more intuitively and vividly recreated, allowing the public to experience and understand historical culture in a virtual environment [14, 15]. At the same time, this technology can also provide a wealth of data and analytical tools for research in related fields, promoting interdisciplinary collaboration and innovation, and advancing the in-depth development of historical and cultural studies.

Currently, although various methods have been applied to the digital reconstruction of historical cultural landscapes, there are still some shortcomings. Traditional methods often rely on manual modeling, which is inefficient and costly; some techniques based on image recognition, although they have improved the degree of automation, still need to improve accuracy in recognition and the effects of three-dimensional reconstruction in complex scenes [16, 17]. Additionally, existing methods perform poorly when dealing with multi-

scale and multi-perspective problems, failing to meet the needs of practical applications [18, 19]. Therefore, there is an urgent need for a more efficient and accurate technical means to solve these problems and promote the development of digital reconstruction technology for historical cultural landscapes.

This paper proposes a digital reconstruction method for historical cultural landscapes based on image recognition technology, divided into two main parts: first, historical cultural landscape target recognition based on MSDC-YOLOv3, and second, three-dimensional reconstruction of historical cultural landscapes based on pyramid feature attention Pixel2Mesh. Through MSDC-YOLOv3 technology, it is possible to more accurately recognize target objects in historical cultural landscapes against complex backgrounds; the pyramid feature attention Pixel2Mesh method, on the other hand, can achieve more efficient and accurate three-dimensional reconstruction, providing detailed three-dimensional models. This research not only achieves technical breakthroughs, enhancing the accuracy and efficiency of recognition and reconstruction, but also has significant application value in the protection and inheritance of cultural heritage, offering new ideas and methods for future research in related fields.

2. TARGET RECOGNITION OF HISTORICAL CULTURAL LANDSCAPES BASED ON MULTI-SCALE DILATED CONVOLUTION YOLOV3

2.1 Network structure

This paper applies the YOLOv3 network structure to the field of historical cultural landscape target recognition. The network structure of YOLOv3, through three steps of feature extraction, feature enhancement, and prediction module, fully extracts and utilizes features in images to achieve high recognition effectiveness. However, in complex historical cultural landscape scenes, the traditional YOLOv3 network structure may miss detecting some targets, primarily due to insufficient feature expression in the prediction module. Dilated convolution introduces dilation intervals to effectively expand the receptive field of the convolution kernel, capturing a broader range of image context information without increasing computational complexity. Additionally, mixed dilated convolution combines various scales of dilated convolutions, effectively avoiding the grid effect, thus enhancing the integrity and accuracy of feature expression.

This paper has made improvements to the YOLOv3 network to better adapt to the complex backgrounds and diverse targets of historical cultural landscapes. The YOLOv3 network uses an FPN structure, performing target recognition through three different scales of feature layers: 13×13, 26×26, and 52×52. The 13×13 layer is mainly used for recognizing large objects in historical cultural landscapes, the 26×26 layer for medium-sized objects, and the 52×52 layer for small objects. During the recognition process, specific improvements have been made to these prediction layers. For the 13×13 layer, the ordinary 3×3 convolution is changed to a dilated convolution with a dilation rate of 3. This modification integrates features over a larger field of view while ignoring some interfering features, thus improving the recognition of large objects in historical cultural landscapes. By adjusting the number of channels, the features on which the prediction results depend are further optimized, making the recognition of this layer more accurate. For the 52×52 layer, the ordinary 3×3 convolution is changed to a mixed dilated convolution with dilation rates of 1, 2, and 4. This improvement, while increasing the field of view, retains more local information, thereby enhancing the ability to recognize small objects in historical cultural landscapes. Adjustments in the number of channels optimize the features that the prediction results depend on, enhancing the recognition performance of this layer. See Figure 1 for the model structure diagram.

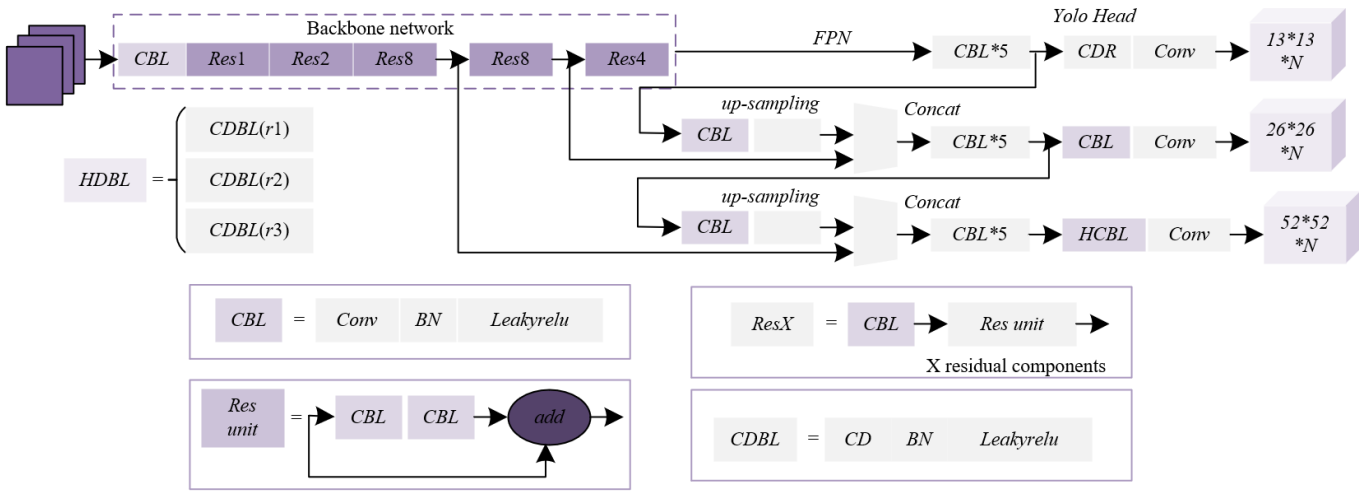

Figure 1. Network structure of the historical cultural landscape target recognition model

2.2 Loss function

In the study of *Target Recognition of Historical Cultural Landscapes Based on Multi-Scale Dilated Convolution YOLOv3*, to better enhance the model's recognition effectiveness and positioning accuracy, this paper adopts a new loss function to compute the localization loss. The loss function in YOLOv3 consists of localization loss, confidence loss, and classification loss, where the localization loss is calculated using Mean Squared Error (MSE) to determine the size and position errors between the actual and predicted boxes. However, traditional localization loss has certain issues in computational efficiency and accuracy, especially in scenes with complex backgrounds and diverse targets, where

inefficiency leads to slow convergence, thereby affecting the overall recognition effectiveness. In the domain of historical cultural landscape target recognition, due to the complexity of the landscape scenes and the diversity of target objects, there is a higher demand for localization accuracy. The presence of significant scale differences among targets and the complex and variable background enhances the impact of localization errors on the overall recognition effectiveness. Therefore, to better adapt to such complex environments, this paper implements a new loss function within the MSDC-YOLOv3 network model to compute localization loss. This new loss function not only improves localization accuracy but also accelerates model convergence, thereby enhancing recognition efficiency and effectiveness.

Specifically, the new localization loss function introduces Intersection over Union (IoU) loss or Generalized Intersection over Union (GIoU) loss in place of the traditional MSE loss. IoU loss more accurately reflects the degree of overlap between the predicted and actual boxes, avoiding the gradient vanishing problem caused by significant size or position differences between the boxes. Meanwhile, GIoU loss further improves upon IoU loss in handling non-overlapping areas. Suppose the width and length of the actual box are represented by q^- and g^-, respectively, the square of the distance between the centers of the two boxes by f^2, and the square of the diagonal length of the smallest enclosing rectangle around the boxes by z^2. The specific formulas are as follows:

$$CIoU = IoU - \left(\frac{f^2}{z^2} + \beta n \right) \tag{1}$$

$$n = \frac{4}{\pi^2} \left(\arctan \frac{\bar{q}}{g} - \arctan \frac{q}{g} \right)^2 \tag{2}$$

$$\beta = \frac{n}{(1 - IoU) + n} \tag{3}$$

Combining the above equations results in a localization loss composed of CIoU loss, given by the following equation:

$$LOSS_{LC} = 1 - CIoU \tag{4}$$

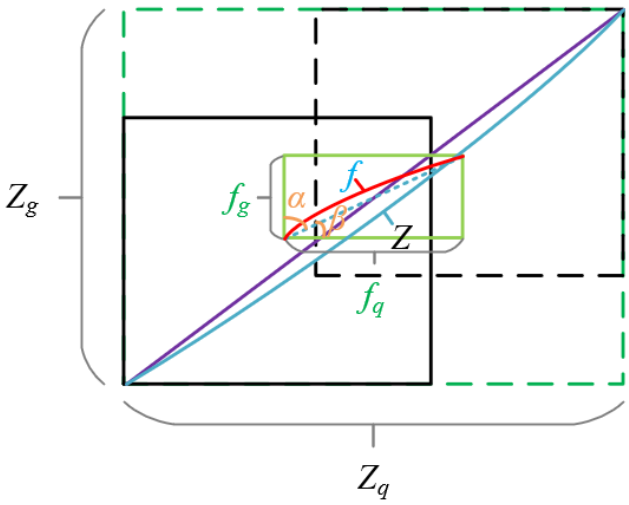

Figure 2. SIoU diagram

Although the CIoU loss function has shown significant improvements in localization accuracy compared to the original localization loss in YOLOv3, it lacks consideration from the perspective of whether the two boxes are well-matched. This could lead to poor model performance during the training process. The SIoI loss function, by considering the vector angle of the prediction box during regression, redefines the penalty metrics, significantly improving the convergence during network training, thereby enhancing the overall detection performance. Specifically, the SIoU loss function takes into account the vector angle of the prediction and actual boxes during regression, i.e., by calculating the vector differences between the centers of the two boxes to measure the degree of match. This method more accurately reflects the matching relationship between the prediction and actual boxes, reducing the deviation of the prediction box from the actual box. Furthermore, the SIoU loss function introduces a new penalty term on top of the traditional IoU, by comprehensively considering factors such as overlap area, distance between centers, aspect ratio, and vector angle, providing a more comprehensive loss evaluation. This makes the prediction box closer to the actual box during training, improving the model's localization accuracy. Figure 2 shows a diagram of SIoU. Assuming the distance between the centers of the two boxes is represented by f, and the length of the rectangle formed by the centers of the two boxes as diagonals is denoted by f_g, then the formulas are as follows:

$$\Lambda = 1 - 2 * \sin^2 \left(\arcsin(a) - \frac{\pi}{4} \right) \tag{5}$$

$$a = \frac{f_g}{f} = \sin(\beta) \tag{6}$$

Assuming the center coordinates of the actual box are represented by s_a^- and s_b^-, and the center coordinates of the prediction box are represented by s_a and s_b, and the width and length of the rectangle enclosing both boxes are represented by Z_q and Z_g, then the formulas are as follows:

$$\Delta = 2 - e^{-(2-\Lambda)\vartheta_a} - e^{-(2-\Lambda)\vartheta_b} \tag{7}$$

$$\vartheta_a = \left(\frac{\bar{s}_a - s_a}{Z_q} \right)^2 \tag{8}$$

$$\vartheta_b = \left(\frac{\bar{s}_b - s_b}{Z_g} \right)^2 \tag{9}$$

Assuming the width and length of the actual box are represented by q^- and g^-, and the width and length of the prediction box are represented by q and g, then the formulas are:

$$\Psi = \left(1 - e^{-\mu_q}\right)^{\varphi} + \left(1 - e^{-\mu_g}\right)^{\varphi} \tag{10}$$

$$\mu_q = \frac{|q - \bar{q}|}{MAX\left(q, \bar{q}\right)} \tag{11}$$

$$\mu_g = \frac{|g - \bar{g}|}{MAX\left(g, \bar{g}\right)} \tag{12}$$

The expression for the SIoU loss function is:

$$SIoU = IoU - \frac{\Delta + \Psi}{2} \tag{13}$$

The expression for the localization loss composed of SIoU is:

$$LOSS_{LC} = 1 - SIoU \tag{14}$$

3. PYRAMID FEATURE ATTENTION PIXEL2MESH FOR 3D RECONSTRUCTION OF HISTORICAL CULTURAL LANDSCAPES

3.1 Network structure

In the study of 3D reconstruction of historical cultural landscapes, this paper utilizes the Pixel2Mesh network structure, which is crucial for achieving high precision in reconstructing historical cultural landscapes. The Pixel2Mesh network initially extracts features from the input image of the historical cultural landscape, generating a series of multi-scale image features. Subsequently, the network introduces an initial mesh ellipsoid as the baseline model. Through multiple iterations, the initial mesh ellipsoid gradually deforms into the target 3D model. In each iteration, the network performs image up-pooling on the initial mesh to increase vertex count, enhancing the mesh model's detail and accuracy. This step-by-step refinement process ensures that the final 3D model accurately reflects the complex structure and details of the historical cultural landscape.

Although the Pixel2Mesh network has achieved notable results in the field of 3D reconstruction, it still exhibits some shortcomings during the 3D reconstruction of historical cultural landscapes, especially as the variety of models increases and their structures become more complex. Specifically, when reconstructing complex historical cultural landscapes, the resulting 3D model shapes may not be accurate, and the model details are poor. This is because the initial mesh ellipsoid needs to undergo multiple iterations to gradually deform into the target model, and during this process, if feature extraction is not precise or the deformation network is handled improperly, the final model's shape and details may not achieve the desired effects. After feature extraction, directly using features for subsequent network modules might be affected by irrelevant features, preventing relevant features from being effectively utilized in the cascade deformation network. These irrelevant features' interference can reduce the accuracy of model reconstruction, making the generated 3D model have errors and difficult to accurately reflect the details and structure of the historical cultural landscape.

To overcome these deficiencies, this study introduces a pyramid feature attention mechanism to enhance the performance of the Pixel2Mesh network. The pyramid feature attention mechanism can extract image features at different scales and selectively focus on important features through the attention mechanism, suppressing the impact of irrelevant features. This mechanism can improve the accuracy and effectiveness of feature extraction, allowing the model to more accurately capture key details and complex structures of the historical cultural landscape.

The image feature extraction module of the pyramid feature attention Pixel2Mesh contains five sub-modules, with outputs denoted as Conv1-2, Conv2-2, Conv3-3, Conv4-3, and Conv5-3. Outputs from Conv1-2 and Conv2-2 are considered low-level features, containing basic edge and texture information of the image. These features are inputted into the spatial attention module (SA) to suppress useless information and focus more on the key parts of the target object. Through this process, low-level features are effectively enhanced during feature extraction, removing interference and increasing the effectiveness of the features. Conv3-3, Conv4-3, and Conv5-3 outputs are considered high-level features, containing rich semantic information. These features are inputted into the Pyramid Context Feature Enhancement module (CPFE) and the Channel Attention module (CA) to obtain more semantic information about the target object. Through the CPFE and CA modules, high-level features are further optimized, enhancing the important information in the feature maps. Subsequently, through up-sampling, the feature maps are resized, allowing the low-level and high-level features to be fused at the same scale. After the fusion of the two different branches of features, the output of the pyramid feature attention network is formed. These fused features contain multi-layered, multi-scale information of the image, better describing the complex structure and details of the historical cultural landscape. To ensure that the output of the attention mechanism network modules can be better applied to the cascade deformation network, this study also adds a connection module between them to further process and optimize the features, making them more suitable for subsequent 3D reconstruction processes. The introduction of the connection module ensures the integrity and effectiveness of the features during transmission, improving the accuracy and stability of 3D reconstruction. Figure 3 provides the network structure of the historical cultural landscape 3D reconstruction model.

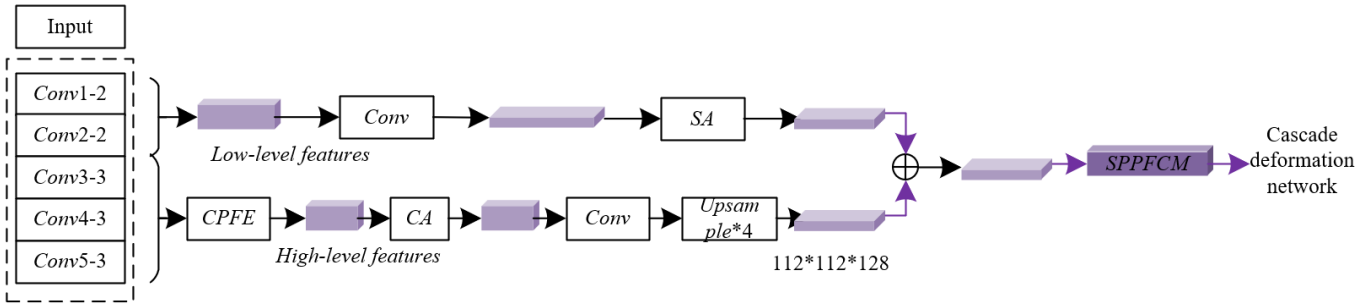

Figure 3. Network structure for 3D reconstruction of historical cultural landscapes

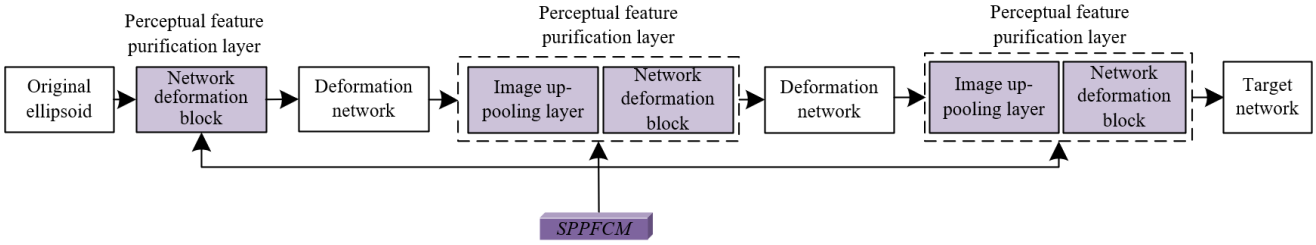

Figure 4. Structure of the utilized cascade deformation network

The SPPF structure processes input features through a series of concatenated Maxpool layers, with each Maxpool layer's output being progressively fused to form the final output of the module. This differs from the SPP structure, which processes inputs in parallel through multiple Maxpool layers before fusing the outputs of each layer. The design of the SPPF structure makes feature extraction more efficient, capturing more spatial information at different scales, thus enhancing feature expressiveness. In this research, the connection module is based on the SPPF structure, combined with the original Pixel2Mesh network design, to form the SPPF Connection Module (SPPFCM). The SPPFCM mainly includes pooling, fusion, and convolution operations. Specifically, through multiple pooling operations combined with the output features of the image feature extraction module, low-level and high-level features are fused. This multi-scale feature fusion better preserves and utilizes key information from different levels, enhancing feature expressiveness and adaptability. The SPPFCM uses 1×1 convolutions to adjust the number of channels, ensuring that the feature maps maintain a reasonable size and depth during transmission. This step not only optimizes the number of channels in the feature maps but also refines the features further, enhancing the model's reconstruction effectiveness. Through these processes, feature maps of different scales are obtained, providing richer and more accurate feature information for subsequent perceptual feature pooling layers and the cascade deformation network. Figure 4 shows the structure of the utilized cascade deformation network.

3.2 Loss function

In the study of 3D reconstruction of historical cultural landscapes based on Pyramid Feature Attention Pixel2Mesh, the selection and design of the loss function are crucial for the training of the model. This research continues the loss function design of the Pixel2Mesh network to ensure high precision and quality results during the 3D reconstruction process. Specifically, the model considers multiple aspects including vertex position, normal vectors, local geometric structures, and mesh edge lengths. These four parts of the loss function work together to ensure the model's comprehensive performance in terms of shape accuracy, surface smoothness, and geometric consistency. The model's loss function is primarily defined in terms of the model's vertices and normal vectors, including chamfer loss, normal loss, Laplacian regularization, and edge length regularization.

The chamfer loss is used to measure the distance between vertices of the reconstructed model and the real model. By calculating the minimum distances between predicted vertices and actual vertices and summing all these distances, the chamfer loss effectively assesses the overall shape accuracy of the reconstructed model. A smaller chamfer loss indicates that the vertices of the reconstructed model are closer to those of the real model, making the shape more accurate. Specifically, for each vertex of the reconstructed model, the nearest vertex in the real model is found, and the sum of these minimum distances squared is calculated. Conversely, for each vertex of the real model, the nearest vertex in the reconstructed model is found, and the sum of these minimum distances squared is calculated. The final chamfer loss is the average of these two sums. Assuming the sets of sampled points from the reconstructed mesh model and the real 3D model are represented by t_1, t_2, with points within these sets represented by a, b, the formula is as follows:

$$M_{ZF}(t_1,t_2)=\sum_{b\in t_2}\underset{b\in t_2}{MIN}\|a-b\|_2^2+\sum_{b\in t_2}\underset{a\in t_1}{MIN}\|a-b\|_2^2 \tag{15}$$

The normal loss function is used to measure the difference in normal vectors between the reconstructed model and the real model. Normal vectors reflect the directional information of the model's surface and have a significant impact on the detail and smoothness of the reconstructed model. By minimizing the differences between normal vectors, the normal loss function ensures that the surface of the reconstructed model is smoother and more realistic. Specifically, for each vertex of the reconstructed model, calculate its corresponding normal vector. Similarly, for each vertex of the real model, calculate its corresponding normal vector. Finally, compute the sum of squared differences of normal vectors between the reconstructed and the real model. Assuming a vertex in the reconstructed mesh model is denoted by constraint $j\in V$, o, and the nearest vertex in the real 3D model to o is denoted by w, with V representing the set of vertices adjacent to vertex o, and the normal vector derived from the real 3D model is denoted by v_w, the formula is as follows:

$$M_{NO}=\sum_{o}\sum_{w=AM_w\left(\|o-w\|_2^2\right)}\|<o-j,v_w>\|_2^2 \tag{16}$$

Laplacian regularization is used to constrain the local geometric structure of the model, preventing excessive distortion or deformation during reconstruction. By introducing the Laplacian operator, the regularization term smooths the model surface, maintaining its local consistency. This is particularly important for the 3D reconstruction of historical cultural landscapes, as these landscapes often have complex geometric shapes and rich detail information. Specifically, first calculate the Laplacian coordinates for each vertex in the reconstructed model, which represent the relative positions between the vertex and its neighboring vertices. Then, compare these Laplacian coordinates with the

corresponding Laplacian coordinates of the original model, minimizing the differences between them. Assuming the Laplacian coordinates of a vertex before and after deformation are denoted by σ'_o and σ_o, the formulas are as follows:

$$M_{LA} = \sum_{o} \| \sigma_o^{'} - \sigma_o \|_2^2 \tag{17}$$

$$\sigma_o = o - \sum_{j \in V(o)} \frac{1}{\| V(o) \|} j \tag{18}$$

Edge length regularization is used to control the lengths of the mesh edges in the model, preventing them from being overly long or short. By maintaining the consistency of edge lengths, the regularization term enhances the model's stability and visual quality. Edge length regularization plays a role in balancing the mesh structure and detail expression in 3D reconstruction. Specifically, first calculate the length of each edge in the reconstructed model, then compare these lengths with the corresponding lengths in the original model, minimizing the differences between them. The formula is as follows:

$$M_{ED} = \sum_{o} \sum_{j \in V(o)} \| o - j \|_2^2 \tag{19}$$

4. EXPERIMENTAL RESULTS AND ANALYSIS

Table 1. Ablation study results using CIoU, SIoU, and different convolutional layers

Experiment Number	*CIoU*	13*13 Layer	52*52 Layer	*SIoU*	*mAP* (%)
1	√				81.85
2	√	√			82.31
3	√		√		82.65
4				√	82.78
5	√	√	√		82.89
6		√	√	√	83.14

Based on the ablation study results in Table 1, the impact of different configurations on the target recognition of historical cultural landscapes is evident. Experiment 1, using only CIoU, achieved a mean Average Precision (mAP) of 81.85%. Experiment 2, which added a convolutional layer at the 13×13 level to CIoU, increased the mAP to 82.31%. Experiment 3, which added a 52×52 convolutional layer on top of CIoU, further increased the mAP to 82.65%. Experiment 4 introduced SIoU instead of CIoU, achieving a mAP of 82.78%. Experiment 5 combined CIoU with both 13×13 and 52×52 layers, reaching a mAP of 82.89%. Experiment 6 introduced SIoU along with the 13×13 and 52×52 layers, achieving the highest mAP of 83.14%. From these results, it can be concluded that the method proposed in this paper, based on MSDC-YOLOv3, significantly improves the accuracy of target recognition. Particularly when using mixed dilated convolution and SIoU, the model's mAP was highest, indicating that the combination of these technologies significantly enhances performance. These improvements allow for more precise recognition of target objects in historical cultural landscapes against complex backgrounds, providing a higher quality data basis for subsequent 3D reconstruction.

Table 2. Experimental results of target recognition in historical cultural landscapes using different algorithms

Method	*F1* (%)	*mAP* (%)	*FPS* (*f*/*s*)
SSD (ResNet)	74.23	74.52	92
Mask R-CNN	63.21	75.69	27
Original *Yolov*3	79.52	79.26	46
The proposed algorithm 1	81.24	82.31	46
The proposed algorithm 2	81.25	83.45	52

Table 2 shows the performance of different algorithms in the target recognition of historical cultural landscapes. SSD (ResNet) achieved a 74.23% F1 score, 74.52% mAP, and 92 FPS; Mask R-CNN had a 63.21% F1 score, 75.69% mAP, and 27 FPS; the original YOLOv3 achieved an F1 score of 79.52%, mAP of 79.26%, and 46 FPS. Algorithm 1 from this paper, which uses the GIoU loss function in a MSDC-YOLOv3, reached an F1 score of 81.24%, mAP of 82.31%, and 46 FPS; and Algorithm 2 from this paper, which uses the SIoU loss function in a MSDC-YOLOv3, further improved to an 81.25% F1 score, 83.45% mAP, and 52 FPS. From these results, it can be concluded that the two algorithms proposed in this paper significantly outperformed traditional methods in terms of accuracy and efficiency. The MSDC-YOLOv3 algorithm using the GIoU loss function (Algorithm 1) improves recognition accuracy while maintaining a high frame rate, showing a clear improvement over the original YOLOv3. The MSDC-YOLOv3 algorithm using the SIoU loss function (Algorithm 2) not only achieved the highest mAP at 83.45% but also increased the frame rate to 52 FPS, demonstrating superior real-time processing capabilities. Overall, these improvements make target recognition of historical cultural landscapes more precise and efficient against complex backgrounds, validating the effectiveness and superiority of the methods researched in this paper, and providing a solid data foundation for subsequent 3D reconstruction.

According to the Figure 5, different algorithms show varying AP values when recognizing targets in historical cultural landscapes. The original YOLOv3 performs consistently across most target categories, particularly excelling in categories such as ancient tower (97), stele (96), and ancient well (93), but it shows weaker performance in archway (64), ancient bridge (64), and city wall (57). Algorithm 1 of this paper, which uses the GIoU loss function with MSDC-YOLOv3, improves recognition accuracy in multiple categories, such as ancient buildings (81), sculptures (93), and murals (82). Algorithm 2, which utilizes the SIoU loss function with MSDC-YOLOv3, further increases the AP values in several categories, especially showing superior performance in ancient architecture (85), stele (98), and ancient tree (77), and it also shows significant improvement in archway (93) and traditional residence (92). From the experimental results, it can be concluded that the algorithms proposed in this paper have a significant advantage in recognizing targets in historical cultural landscapes. Algorithm 1, by adopting the GIoU loss function, has significantly improved recognition accuracy in several categories, especially in important target categories such as ancient architecture, sculpture, and mural. Algorithm 2, by further adopting the SIoU loss function, enhances recognition accuracy in more target categories, particularly excelling in categories like ancient architecture, stele, archway, and traditional residence. Figure 6 displays the experimental results of target recognition in historical cultural landscapes by the algorithms of this paper.

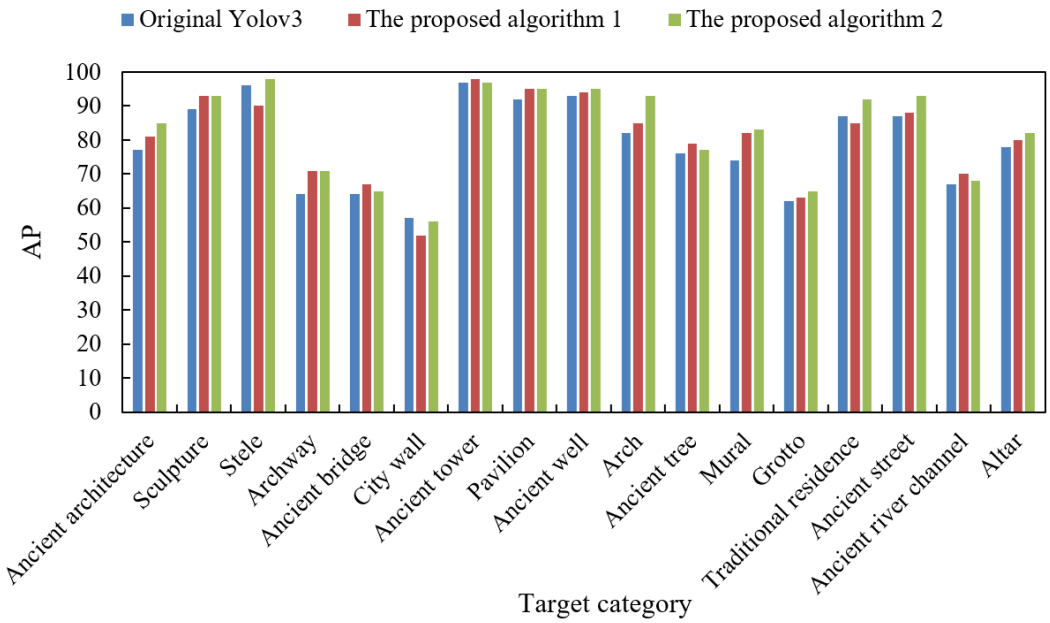

Figure 5. AP values for various categories of historical cultural landscape targets by different algorithms

Figure 6. Experimental results of target recognition in historical cultural landscapes by the algorithms of this paper

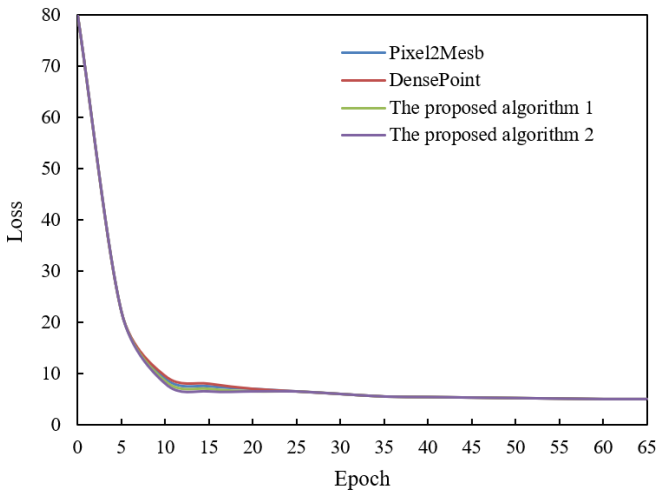

Figure 7. Loss function values during training of the historical cultural landscape 3D reconstruction model network

Figure 7 shows the changes in loss function values during the training process of different algorithms for 3D reconstruction models of historical cultural landscapes. At the initial epoch (0), the loss values for all algorithms start at 80. As the training epochs increase, the loss values for each algorithm gradually decrease. By the 10th epoch, the loss value for Pixel2Mesh drops to 9, DensePoint to 9.5, Algorithm 1 of this paper to 8.5, and Algorithm 2 to 8. By the 25th epoch, the loss values for all algorithms tend to stabilize, with Pixel2Mesh and DensePoint stabilizing at 6.5, and both Algorithm 1 and Algorithm 2 also stabilizing at 6.5. Finally, by the 65th epoch, the loss values for all algorithms reach their lowest points and stabilize at 5. From the experimental results, it is evident that the proposed algorithms show significant advantages in terms of the speed of loss reduction and final stable values in the 3D reconstruction of historical cultural landscapes. Especially, Algorithms 1 and 2 demonstrate a loss reduction speed in the first 10 epochs that is noticeably faster than Pixel2Mesh and DensePoint, showing higher training efficiency. Although the final stable values are similar across all algorithms, the quick convergence of the algorithms in the early training stages validates their effectiveness in efficient 3D reconstruction tasks. Additionally, by incorporating the pyramid feature attention mechanism, these algorithms improve the reconstruction accuracy in complex backgrounds, providing reliable technical support for the digital preservation and study of historical cultural landscapes.

Table 3 displays the F-score performance of different algorithms for various categories of historical cultural landscape targets at different thresholds (10^{-4} and 2×10^{-4}). For the threshold of 10^{-4}, Algorithm 1 (Pixel2Mesh network + pyramid feature attention mechanism) scores higher in most categories compared to both Pixel2Mesh and DensePoint, for example, temple (75.41), castle (61.25), and residential house (61.25). Algorithm 2 (DensePoint + pyramid feature attention mechanism) performs slightly lower in some categories compared to Algorithm 1, but still outperforms the original Pixel2Mesh and DensePoint. At the threshold of 2×10^{-4}, both Algorithm 1 and Algorithm 2 significantly improved their F-scores across all categories, particularly in temple (85.36 and 84.23), palace (71.23 and 68.23), and castle (73.54 and 71.24), where they exhibited especially strong performance. The experimental results show that the two algorithms proposed in this paper performed well at different thresholds, demonstrating their effectiveness in the 3D reconstruction of historical cultural landscapes. Algorithms 1 and 2, by incorporating the pyramid feature attention mechanism, were able to recognize historical cultural landscape targets more

accurately against complex backgrounds and achieved significantly higher F-scores than the traditional Pixel2Mesh and DensePoint methods. Particularly, when the threshold was raised to 2×10^{-4}, the performance of these algorithms was further enhanced, proving their stability and adaptability under higher precision requirements.

Table 3. F-scores (%) of various categories of historical cultural landscape targets at different thresholds

Historical Architecture Category	Threshold 10^{-4}				Threshold 2×10^{-4}			
	Pixel2Mesh	***DensePoint***	**Algorithm 1**	**Algorithm 2**	***P2M***	***Resnct_P2M***	**Algorithm 1**	**Algorithm 2**
Temple	74.12	72.31	75.41	72.31	86.54	83.24	85.36	84.23
Palace	54.23	52.69	54.23	52.34	68.93	66.23	71.23	68.23
Castle	57.36	55.41	61.25	57.59	71.23	69.58	73.54	71.24
Tower	40.12	38.62	41.58	39.65	56.23	52.31	56.39	54.23
Theater	44.25	42.31	45.36	43.25	61.54	61.45	62.31	61.23
Bridge	65.36	61.25	65.39	62.35	81.25	74.26	81.23	78.59
City gate	73.59	72.56	73.56	73.58	85.36	85.32	85.69	85.62
Residential house	58.69	56.39	61.25	56.39	71.26	71.26	73.45	73.21

Table 4. CD values of different algorithms on the test set of historical cultural landscape images

Historical Architecture Category	***Pixel2Mesh***	***DensePoint***	**Algorithm 1**	**Algorithm 2**
Temple	0.25	0.28	0.23	0.26
Palace	0.66	0.73	0.61	0.71
Castle	0.74	0.88	0.73	0.81
Tower	1.13	1.22	1.18	1.16
Theater	0.93	0.93	0.87	0.92
Bridge	0.43	0.51	0.46	0.46
City gate	0.38	0.37	0.36	0.36
Residential house	0.65	0.71	0.62	0.66

Table 5. CD values and f-scores at different thresholds for algorithms on the test set of historical cultural landscape images

		Temple	**Palace**	**Castle**	**Tower**	**Theater**	**Bridge**	**City Gate**	**Residential House**
10^{-4}	Algorithm 1	72.31	52.31	54.26	37.58	41.23	63.87	67.58	55.36
	Algorithm 2	75.32	54.36	61.58	41.23	46.58	66.25	73.26	61.24
Threshold 2×10^{-4}	Algorithm 1	84.56	67.89	71.69	54.69	58.69	79.58	81.46	72.31
	Algorithm 2	86.34	71.23	74.56	57.69	64.23	81.23	85.69	73.56
CD	Algorithm 1	0.24	0.66	0.75	1.12	0.88	0.41	0.37	0.66
	Algorithm 2	0.23	0.61	0.73	1.18	0.87	0.44	0.36	0.64

Table 4 presents the CD values of different algorithms on the test set of historical cultural landscape images, where a lower CD value indicates that the reconstructed model is closer to the real model. Algorithm 1 (Pixel2Mesh network + pyramid feature attention mechanism) shows lower CD values across most categories, such as temple (0.23), palace (0.61), and castle (0.73), performing better than both Pixel2Mesh and DensePoint. Algorithm 2 (DensePoint + pyramid feature attention mechanism) performs slightly better than Algorithm 1 in some categories like tower (1.16) and theater (0.92), though its overall CD values remain low. Overall, both algorithms proposed in this paper significantly outperform traditional methods, showing smaller CD values in multiple categories. The experimental results demonstrate that the two algorithms proposed in this paper significantly reduced the discrepancy between the reconstructed model and the real model in the 3D reconstruction of historical cultural landscapes, particularly noticeable in categories such as temple, palace, and castle, where Algorithm 1 exceled in reconstructing complex structures and details. Algorithm 2 also performed well in categories like tower and theater, further validating the effectiveness of the pyramid feature attention mechanism within the DensePoint network.

Table 5 shows the CD values and F-scores at different thresholds for the algorithms on the test set of historical cultural landscape images. CD represents the distance between corresponding points of the reconstructed and real models. At the threshold of 10^{-4}, Algorithm 1 shows uniform F-scores across various categories, with notable performances in temple (72.31), bridge (63.87), city gate (67.58), and residential house (55.36). Algorithm 2's F-scores are slightly higher than those of Algorithm 1, such as in temple (75.32) and residential house (61.24). At a threshold of 2×10^{-4}, both Algorithm 1 and Algorithm 2 show significant improvements in F-scores, particularly in temple (84.56 and 86.34), bridge (79.58 and 81.23), and city gate (81.46 and 85.69). In terms of CD values, both Algorithm 1 and Algorithm 2 show low values across categories, such as Algorithm 1's temple (0.24), city gate (0.37), and Algorithm 2's palace (0.61), castle (0.73), indicating high precision in model reconstruction. The experimental results demonstrate that the two algorithms based on the pyramid feature attention mechanism exhibited outstanding performance in the 3D reconstruction of historical cultural landscapes. Algorithms 1 and 2 show high efficiency and accuracy in target recognition and 3D reconstruction tasks, especially at the higher threshold of 2×10^{-4}, where the significant improvement in F-scores further validates the reliability of the algorithms. Additionally, the substantial reduction in CD values indicates minimal differences between the reconstructed and real models, particularly in complex structures such as temples, palaces, and castles, where the reconstruction effects are particularly refined.

In summary, the methods studied in this paper not only exhibited higher accuracy in target recognition but also

demonstrated excellent performance in 3D reconstruction, providing effective technical means for the digital preservation and study of historical cultural landscapes, fully proving the innovation and practicality of this research.

5. CONCLUSION

This paper presented a digital reconstruction method for historical cultural landscapes based on image recognition technology, primarily divided into two parts: historical cultural landscape target recognition using MSDC-YOLOv3, and 3D reconstruction of historical cultural landscapes using pyramid feature attention Pixel2Mesh. The MSDC-YOLOv3 technique enables more precise identification of target objects in historical cultural landscapes against complex backgrounds, while the pyramid feature attention Pixel2Mesh method facilitates more efficient and accurate 3D reconstruction, providing finely detailed 3D models. This paper conducted a series of experiments, including ablation studies with CIoU, SIoU, dilated and mixed dilated convolutions, performance evaluations of different algorithms in recognizing historical cultural landscape targets, assessments of AP values across various categories, analysis of loss function values in 3D reconstruction model network training, F-scores (%) across various thresholds, and CD values on the test set, comprehensively verifying the effectiveness of the proposed methods. Experimental results indicate that the algorithms presented excel in both the recognition and 3D reconstruction of historical cultural landscapes. Notably, Algorithm 1 (Pixel2Mesh network + pyramid feature attention mechanism) and Algorithm 2 (DensePoint + pyramid feature attention mechanism) demonstrated significant advantages in F-scores and CD values at different thresholds. Algorithm 1 showed high F-scores in categories such as temples (84.56) and bridges (79.58), and exhibited smaller differences in CD values across multiple categories (e.g., temples at 0.24), proving the high precision of model reconstruction. Algorithm 2 also performed excellently in certain categories (e.g., palaces at 0.61 in CD values) and F-scores, validating its effectiveness in the DensePoint network.

This study provides efficient and precise technical means for the digital preservation and reconstruction of historical cultural landscapes. With improved YOLOv3 and Pixel2Mesh techniques, this paper not only enhanced the accuracy of target recognition but also facilitated the construction of finely detailed 3D models, providing more reliable foundational data for digital preservation. Although the methods proposed performed well in numerous experiments, they still have limitations. First, the accuracy of recognition and reconstruction may decline when dealing with extremely complex backgrounds or severely damaged historical cultural landscapes. Secondly, the diversity and quality of training data significantly affect model performance, necessitating further enhancement with high-quality historical cultural landscape data in the future. Future research could improve in several areas: further optimizing the model structure to enhance recognition capabilities in complex backgrounds and damaged landscapes, expanding dataset diversity and scale to improve model generalizability, and integrating other cutting-edge technologies such as Generative Adversarial Networks (GANs) and image repair techniques to further enhance the precision and realism of 3D reconstruction. Through these improvements, more effective digital preservation and study of historical cultural landscapes can be achieved, advancing the field.

REFERENCES

[1] Yang, B.X. (2024). Research on the application of multidimensional collaborative landscape design course teaching in revolutionary historical and cultural landscape design-taking Zhuhai as an example. Applied Mathematics and Nonlinear Sciences, 9(1): 1-16. https://doi.org/10.2478/amns-2024-1218

[2] Zhou, W.H., Cenci, J., Zhang, J.Z. (2024). Systematic bibliometric analysis of the cultural landscape. Journal of Asian Architecture and Building Engineering, 23(3): 1142-1164. https://doi.org/10.1080/13467581.2023.2257276

[3] Lai, Y.L. (2024). Three-dimensional visualisation of cultural landscape under the perspective of culture and tourism integration. Applied Mathematics and Nonlinear Sciences, 9(1): 1-18. https://doi.org/10.2478/amns-2024-1466

[4] Xu, W.T. (2022). Ecological integrity evaluation of organically evolved cultural landscape. Mobile Information Systems, 2022(1): 9554359. https://doi.org/10.1155/2022/9554359

[5] Hu, Z., Wu, X. (2024). An object-oriented algorithm for constructing the conceptual lattice of cultural landscape genes of traditional settlements. Journal of Geo-Information Science, 26(3): 604-619. https://doi.org/10.12082/dqxxkx.2024.230518

[6] Valetti, L., Pellerey, F., Pellegrino, A. (2023). A novel approach for the assessment of the nocturnal image of the cultural landscape. LEUKOS-Journal of Illuminating Engineering Society of North America, 19(1): 71-93. https://doi.org/10.1080/15502724.2022.2057325

[7] Gao, W., Wang, S., Chen, S., Hu, S., Li, H. (2023). Identifying cultural ecosystem services and relevant landscape elements provided by urban green space throughout history from an information communication perspective. Forests, 14(5): 1045. https://doi.org/10.3390/f14051045

[8] Lin, Y.N., Yang, C., Ye, Y.H., Zhang, Z.R. (2021). Research on the transformation of historic patterns of cultural landscape using aerial photogrammetry and geo-database: A case study of Kuliang in Fuzhou, China. International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences-ISPRS Archives, XLVI-M-1-2021: 409-414. https://doi.org/10.5194/isprs-archives-XLVI-M-1-2021-409-2021

[9] Romanova, I.A., Poluboyarova, N.M. (2021). The virtual reconstruction of historical and cultural heritage monuments of the Vodyansky settlement. Scientific Visualization, 13(3): 9-21. https://doi.org/10.26583/sv.13.3.02

[10] Süvari, A., Okuyucu, Ş.E., Çoban, G., Eren Tarakci, E. (2023). Virtual reconstruction with the augmented reality technology of the cultural heritage components that have disappeared: The Ayazini Virgin Mary Church. Journal on Computing and Cultural Heritage, 16(1): 1-16. https://doi.org/10.1145/3579361

[11] Sun, J., Sun, Y., Chen, M. (2023). The digital reconstruction of rockery landscape based on NeRF. In

Proceedings of the 2023 7th International Conference on Big Data and Internet of Things, Beijing, China, pp. 22-27. https://doi.org/10.1145/3617695.3617699
[12] Wang, X. (2022). Artificial intelligence in the protection and inheritance of cultural landscape heritage in traditional village. Scientific Programming, 2022(1): 9117981. https://doi.org/10.1155/2022/9117981
[13] Razuvalova, E., Nizamutdinov, A. (2015). Virtual reconstruction of cultural and historical monuments of the middle volga. Procedia Computer Science, 75: 129-136. https://doi.org/10.1016/j.procs.2015.12.229
[14] Chen, G., Yang, R., Lu, P., Chen, P., Gu, W., Wang, X., Hu, Y., Zhang, J. (2022). How can we understand the past from now on? Three-dimensional modelling and landscape reconstruction of the shuanghuaishu site in the central plains of China. Remote Sensing, 14(5): 1233. https://doi.org/10.3390/rs14051233
[15] Nemtinov, V., Borisenko, A., Morozov, V., Nemtinov, K., Protasova, Y. (2023). Creation of a virtual environment for analysis of historical processes related to life of IV michurin in Russia. In International Conference on Intelligent Sustainable Systems, pp. 587-595. https://doi.org/10.1007/978-981-99-1726-6_45
[16] Matasov, V., Nizovtsev, V., Erman, N. (2019). Landscape-historical geoinformation system as a base for long-term land-use change retrospective modelling. International Multidisciplinary Scientific GeoConference: SGEM, 19(2.2): 895-901. https://doi.org/10.5593/sgem2019/2.2/S11.110
[17] Iakushkin, O., Selivanov, D., Tazieva, L., Fatkina, A., Grishkin, V., Uteshev, A. (2018). 3D reconstruction of landscape models and archaeological objects based on photo and video materials. In Computational Science and Its Applications–ICCSA 2018: 18th International Conference, Melbourne, VIC, Australia, pp. 160-169. https://doi.org/10.1007/978-3-319-95171-3_14
[18] Shiu, Y.S., Lei, T.C., Lee, R.Y., Lin, F.C., Chu, T.H. (2015). The reconstruction of urban cultural landscape and the change analysis of urban landscape and texture-A case study of the old downtown in Central District, Taichung City, Taiwan. In ACRS 2015-36th Asian Conference on Remote Sensing: Fostering Resilient Growth in Asia, Proceedings.
[19] De Kramer, M., Mersch, S., Morse, C. (2018). Reconstructing the historic landscape of Larochette, Luxembourg. In Euro-Mediterranean Conference, Nicosia, Cyprus, pp. 30-37. https://doi.org/10.1007/978-3-030-01765-1_4