Applied Science and Innovative Research
ISSN 2474-4972 (Print) ISSN 2474-4980 (Online)
Vol. 7, No. 4, 2023
www.scholink.org/ojs/index.php/asir

Original Paper

Traditional Village Classification Model Based on Transformer Network

Qi Zhong[1*]

[1] College of Architecture & Environment, Sichuan University, Chengdu 610065, PR China

[*] Corresponding author: zhongqi77@stu.scu.edu.cn

Received: October 28, 2023 Accepted: November 17, 2023 Online Published: November 21, 2023

doi:10.22158/asir.v7n4p126 URL: http://doi.org/10.22158/asir.v7n4p126

Abstract

The study of traditional villages holds significant implications in cultural, historical, and societal contexts. Despite the considerable research focus on the architectural styles of Qiang, Tibetan, Han, and Hui ethnic villages due to their distinctiveness, rapidly and accurately identifying the types of traditional villages in practical surveys remains a challenge. To address this issue, this paper establishes an aerial image dataset for Qiang, Tibetan, Han, and Hui ethnic villages and introduces a specialized feature extraction network, Transformer-Village, designed for the classification and detection of traditional villages using deep learning algorithms. The overall structure of the network is lightweight, incorporating condconv dynamic convolution as the core layer structure; furthermore, a spatial self-attention-related feature extraction network is designed based on Transformer. In conclusion, through simulated experiments, Transformer-Village coupled with the YOLO detector achieves a 97.2% mAP on the test set, demonstrating superior detection accuracy compared to other baseline models. Overall, the experimental results suggest that this work is feasible and practical.

Keywords

Transformer, Classification model, Deep learning

1. Introduction

Chinese traditional villages (Gao & Wu, 2017) have evolved over thousands of years, witnessing complex and intense ethnic struggles and cultural integration among ethnic groups such as the Qiang, Tibetan, Hui, and the Han. Ultimately, they have formed unique cultural landscapes with significant research value. The study of traditional villages is based on a wealth of on-site research data. However, due to factors such as geographical conditions, lagging transportation, and underdeveloped information infrastructure, some traditional villages remain in a state of closed development. Conducting in-depth

research within these traditional villages presents significant challenges. Additionally, given the large number and widespread distribution of these villages, traditional manual mapping requires a considerable investment of time and effort. To address this difficulty, this paper employs aerial imagery and a deep learning model to establish an intelligent classification system for traditional villages using aerial images. This approach reduces dependence on experts, enables rapid identification of traditional village types, significantly enhances research efficiency, and to some extent, addresses the shortcomings of subjective judgment in traditional research methods.

Simultaneously, the current surge in research on deep learning has propelled the rapid development of computer vision technology. Deep convolutional neural networks (Aloysius & Geetha, 2017), capable of autonomously learning relevant features through iterative training, exhibit greater robustness and user-friendliness compared to traditional algorithms that rely on manually presetting feature detectors based on prior knowledge. Consequently, convolutional neural networks have gradually supplanted traditional algorithms, driving the practical application of numerous projects. Target detection algorithms based on deep learning can be categorized into single-stage and two-stage architectures depending on the application scenario. Single-stage networks integrate classifiers and locators into the same network framework after extracting feature information through convolutional networks, facilitating rapid detection of targets. However, this structure comes at the cost of a certain degree of loss in detection accuracy and is typically suitable for scenarios with low computational power and a demand for high-speed detection. Notable examples of such network algorithms include the YOLO series and SSD series.

In the context of this study, the distinctions between traditional villages of different ethnic groups are not immediately apparent, posing challenges in rapidly identifying the specific categories of villages during on-site investigations. To address this, the study utilizes unmanned aerial vehicles to capture comprehensive aerial images of traditional villages and establishes a dataset incorporating Qiang, Tibetan, Hui, and Han ethnic groups for training purposes. The designed classification model assists researchers in swiftly identifying the types of traditional villages during fieldwork, allowing them to determine research directions based on different types and thereby enhancing research efficiency. Consequently, the focal point of this study lies in the effective design of a traditional village feature extraction network.

2. Traditional Village Dataset

This study employed unmanned aerial vehicles to capture high-resolution aerial images at a resolution of 4056*3040 pixels. A total of 267 villages in the Sichuan province were comprehensively photographed from varying heights and angles. These villages represent Qiang, Tibetan, Hui, and Han ethnic groups. Additionally, satellite images from the internet were utilized to gather visuals of these four types of villages, collectively forming the dataset. The four distinct categories of traditional villages are as follows:

Figure 1. Qiang Village

Qiang ethnic villages primarily utilize rammed earth, stone masonry, and wood as their main building materials. The predominant color scheme consists of shades of yellow, complemented by accents of red and gray. Some villages are equipped with watchtowers. The roof styles vary, featuring both flat roofs and pitched roofs, with trapezoidal decorations at the corners of flat roofs. Courtyard spaces include both enclosed ground-level courtyards and roofed courtyards formed by setbacks. The architectural windows tend to be relatively small. Decorations on Qiang ethnic buildings are comparatively simple, with minimal embellishments using white stones and wood.

Figure 2. Tibetan Village

Tibetan ethnic villages often employ stone masonry as their primary building material. The color palette is primarily composed of shades of gray, with accents of yellow and red. Some villages include watchtowers. The roofs are flat, featuring a right-angled triangular decoration on the rooftop. The

architectural layout is square, forming a roofed courtyard through setbacks. Compared to Qiang ethnic architecture, the windows in Tibetan buildings are relatively larger. Tibetan architecture is known for its decorative elements, often adorned with painted patterns or colorful flags.

Figure 3. Hui Village

Hui ethnic villages primarily use stone and wood as their main building materials. The color scheme is predominantly white. The roofs are pitched, mainly covered with gray tiles, and the roof edges are adorned with white paint. The architectural layout consists of courtyard spaces enclosed on two or three sides.

Figure 4. Han Village

Traditional Han ethnic village architecture often features a brick-and-wood structure with pitched roofs covered in gray tiles, commonly displaying delicate eaves lines. The overall design places emphasis on

symmetrical beauty, and courtyard layouts typically follow a two-courtyard or three-courtyard configuration, creating enclosed and tranquil spaces. Windows, mostly made of wood, are adorned with carvings, lattice, and other decorations, showcasing exquisite traditional craftsmanship. Exterior walls are often painted in white or gray, presenting a clean and elegant appearance.

3. Related Technologies

Transformer (Parmar et al., 2018) was first proposed and applied in the NLP field. Its core idea is the self attention mechanism. Through encoding and decoding the phrases input into the network, it can improve the network's learning of the relevance between words. At present, more and more computer vision solutions use Transformer mechanism for reference to enhance the global awareness of the model by improving the global feature information of the target in the feature map. First, the working mechanism of Transformer is shown in the following figure:

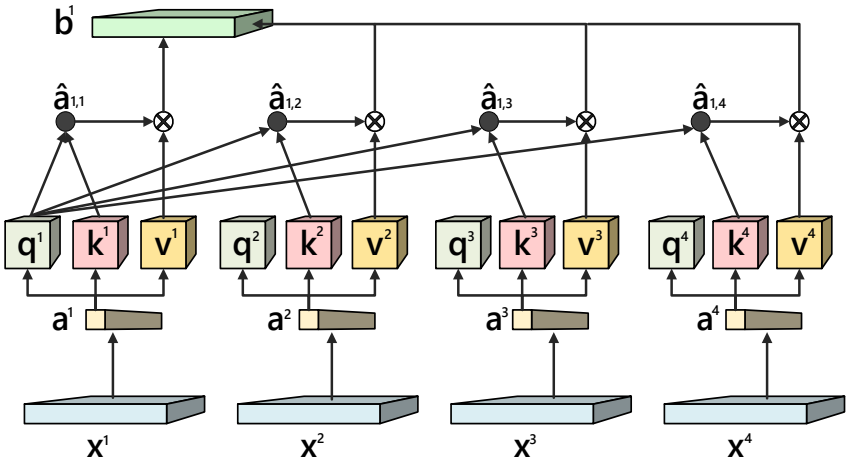

Figure 5. Transformer Workflow

As shown in the figure above x^1,x^2,x^3 and x^4 are input sequences (can be text or image), for each input sequence, there are certain correlation characteristics between them. To obtain the association information between input elements, each x^i is equipped with three vectors: q (query), k (key), and v (value). Take the x^1 as an example, its q vector is mainly used to query the relationship between its own element and other elements; The k-vector represents the relationship between itself and the q vector of other elements when querying; The v vector is the characteristic expression vector of the element. When extracting the relationship features of other input elements, x^1's q^1 will query and calculate with k^1、k^2、k^3、k^4 respectively, and obtain [$\hat{\alpha}_{1,1}, \hat{\alpha}_{1,2}, \hat{\alpha}_{1,3}, \hat{\alpha}_{1,4}$] such a group of weight vectors. After that, each weight vector is recombined with the input feature to obtain the relationship between x^1 and other elements to express the output b^1. In the field of computer vision, classification, detection and segmentation tasks can also obtain the association information between objects in the graph in the above way.

4. Network Structure Design

For the design of network structure, the structural scheme of "Conv+Transformer+Conv" is adopted. First, the convolutional network is used to convert the input image into a feature map. Then, the Transformer network acquires the global attention of the feature map. Finally, the convolutional network is used to transition the output feature map, and the detector detects the target in the map.

4.1 Preconditioning Convolutional Network

The function of the preprocessing convolutional network is to enhance the amount of information contained in the feature map, so that the subsequent Transformer network can build a global feature association based on more feature information. The preprocessing network is constructed based on the FPN (Feature Pyramid Networks) network structure (Deng et al., 2021). In addition, the DynamicConv dynamic convolution (Wu et al., 2019) and dense residual mapping structure (Huang et al., 2019) are used in the preprocessing network to enhance the feature extraction ability of the preprocessing network. Based on the above design points, the preprocessing network module designed in this paper is shown in the following figure:

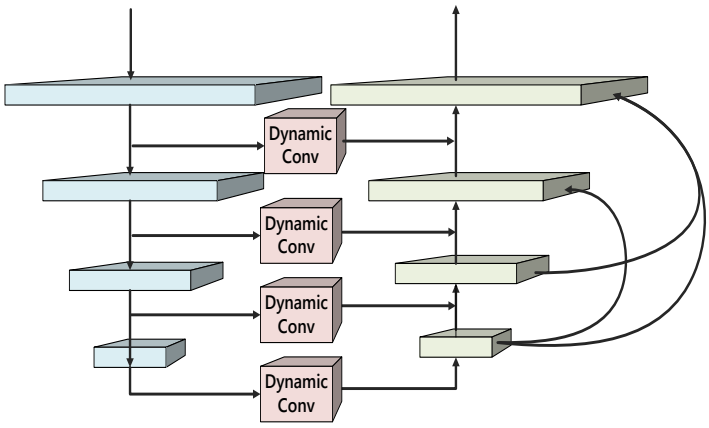

Figure 6. Pretreatment Network Module

As shown in the figure above, four layers of convolution layer are used for each "pyramid" branch in the construction of FPN network. When merging the corresponding feature maps in the "coding pyramid" and "decoding pyramid" network branches, we add dynamic convolution to the feature transmission channel to process the transmitted feature information twice. The dynamic convolution uses different convolution check feature maps for feature extraction, which enhances the flexibility of the convolution layer. At the same time, the dynamic convolution assigns the parameters of each convolution core based on the attention mechanism, which improves the feature extraction ability of the convolution core. The calculation process of dynamic convolution kernel in dynamic convolution can be expressed as:

$$y = g(\widetilde{W}^T x + b)$$
$$\widetilde{W} = \sum_{k=1}^{K} \pi_k(x) W_k$$

In the dynamic convolution kernel, $\widetilde{W}$ is the attention weighted weight parameter, b is the offset, g is the activation function, and the value range of the attention weight $\pi_k(x)$ is [0,1], which changes with the change of the input characteristic map. The implementation process of dynamic coiler is shown in the following figure:

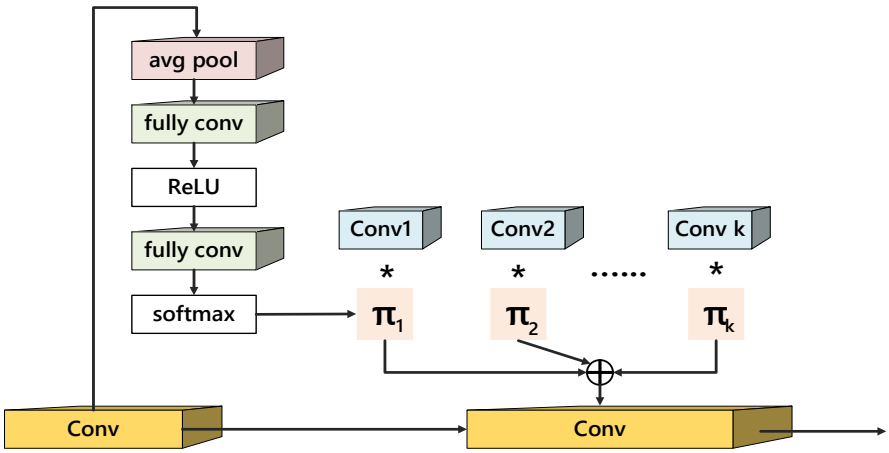

Figure 7. Structure of Dynamic Winding Layer

In order to obtain the attention weight $\pi_k(x)$ of each convolution kernel, first, compress the global spatial information of the input feature map through global average pooling, and then calculate the attention weight of each convolution kernel by using two full connections and ReLU and SoftMax activation functions respectively, and then multiply the attention weight with its corresponding convolution kernel. The assigned convolution kernels are added to form a dynamic convolution layer, and then the input characteristic graph is convolved. The dynamic convolution is used to extract the secondary features of the transferred feature map. Compared with the common FPN network, this method effectively increases the expression of the effective features in the feature map and the feature richness of the whole feature map. In addition, the output feature map of each layer of the coding pyramid is transmitted to the decoding pyramid after the weighting processing of the dynamic convolution layer, and is combined with the output of each layer of the decoding branch to be transmitted to the next layer. This combination operation can further enhance the proportion of effective features in the feature map.

4.2 Dual Transformer Branch Network

After the input data undergoes preprocessing by the convolutional network, a feature map containing low-level feature information can be obtained. Subsequently, a Transformer sub-network is designed to further extract the feature map output by the preprocessed network. In the detection task of this paper, one issue that needs attention is the significant size differences among detection categories. For

example, the aerial images of various villages have different scales, and the pixel sizes of individual buildings in the images vary. Therefore, addressing the feature extraction differences caused by different-sized targets is a core consideration in designing the Transformer sub-network. To tackle this problem, this paper adopts a multi-receptive field feature extraction method and designs a dual Transformer branch network (Yu et al., 2022). First, the structure of the single Transformer branch channel used in this paper is illustrated in the figure below:

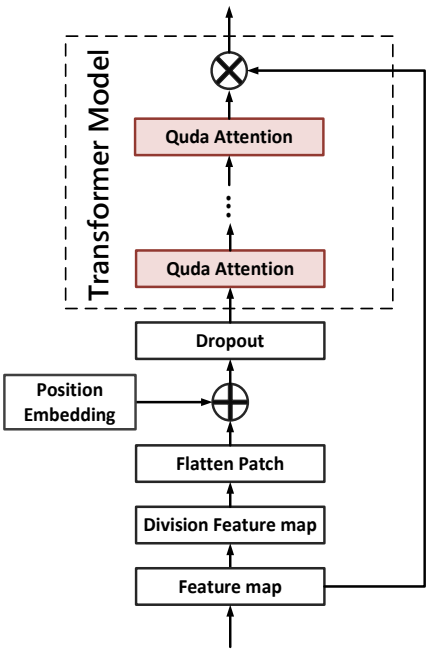

Figure 8. Transformer Branch Channel Structure

As shown in the above figure, the feature map input to the branch is first partitioned according to the preprocessing flow of the vision transformer, then processed by 1x1 convolution group, and then transferred to the transformer module group. Although Transformer mechanism can establish global feature attention, in the use environment of this paper, the background information will affect the results of Transformer global calculation for the sampled frame images due to the complexity of village information; At the same time, in order to improve the detection accuracy of small targets, the size of the input image is usually set as large as possible, but this also increases the amount of network computing. Therefore, for the design of Transformer module group, in this paper, the quadtree attention mechanism is used to build the Transformer pyramid structure, and the attention weight is calculated module by module from coarse to fine. The schematic diagram of this structure is as follows:

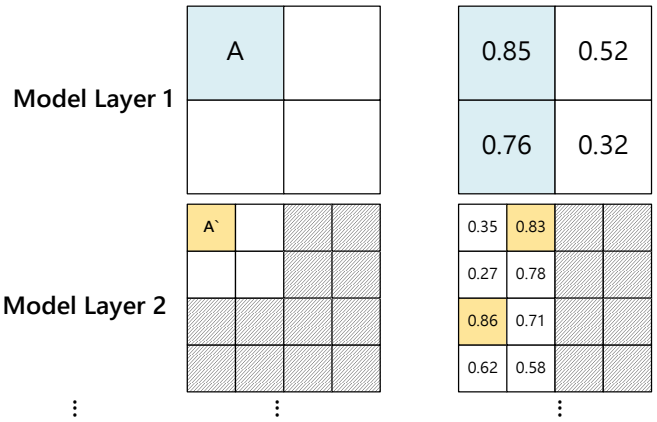

Figure 9. Transformer Pyramid Structure

As shown in the above figure, suppose that in the first module, the feature map is first divided into four patches (Wang et al., 2021). Take A patch as an example, calculate its attention weight with the four patches (including its own), and select the N patches with the highest weight (N is 2 in the above figure) to calculate the attention weight of the next module. In the second module, the A patch in the first module will be further partitioned. Similarly, the N patches with the highest degree of association with A patches in the first module are also divided at the same granularity; As shown in the above figure, the sub patch A' in the second module only performs attention calculation with the sub patches of two patches. Similarly, only the N sub patches with the highest weight are reserved for the calculation of the next module, and so on in subsequent module calculations. By ignoring irrelevant areas in this way, and increasing the segmentation granularity step by step, the attention weight of related targets can be improved. By ignoring irrelevant areas in this way, and increasing the segmentation granularity step by step, the attention weight of related targets can be improved. After the calculation of each module, the attention weight of each module needs to be aggregated. In this paper, use $m_i$to calculate the average weight of all modules. The formula is:

$$m_i = \sum_{1 \le l \le L} w_i^l m_i^l$$

In the above formula, w_i^l is a learnable offset parameter in the network. At the same time, Transformer modules at different levels are nested in calculation, so the calculation of m_i^l can be expressed as:

$$m_i^l = Attention(q_i^l, K_{\Gamma_i^l}^l, V_{\Gamma_i^l}^l)$$

In the above formula, $Attention$ is Transformer's softmax attention calculation formula, and $K_{\Gamma_i^l}^l$ and $V_{\Gamma_i^l}^l$ are the superposition of Key matrix and value matrix in each module. Finally, use m_i to weight the characteristic graph to obtain the characteristic graph calculated by the quad tree attention. Based on a single Transformer branch, this paper constructs a dual Transformer branch network. The structure of the network is shown in the following figure:

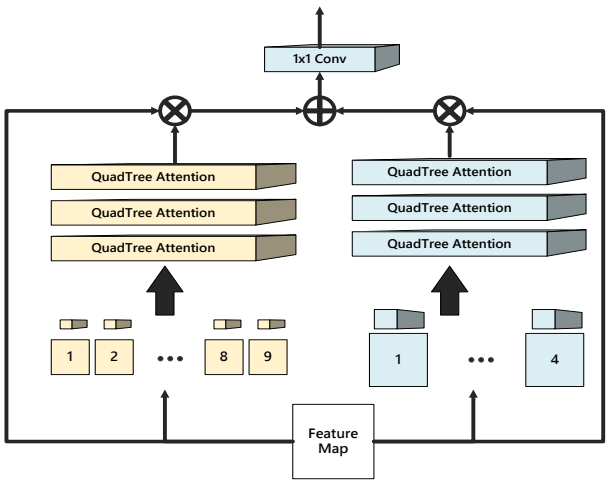

Figure 10. Dual Transformer Branch Network

As shown in the above figure, the feature map inputted to the network will be processed by two Transformer branch networks. The structures of the two branch networks are the same. The main difference is the fineness of the feature map segmentation. In the left branch, the patch granularity of feature graph segmentation is smaller, which is more conducive to weighting the feature information of small targets; On the contrary, in the right branch, the granularity of the split patch is larger, and more emphasis is placed on weighting the feature information of the large target. Two sets of patches with different granularity will be merged after being processed by different Transformer branch networks. In the merging operation, concat layer is used to stack the two feature maps on the channel dimension, and then 1x1 convolution layer is used to compress and integrate the channel of the feature map, remove redundant channels and reduce the parameter of the feature map at the same time. So far, all the work of feature extraction network has been completed.

Based on the usage scenario of the network and its main core methods, the feature extraction network is named Transformer-Village. In addition, in the detection module, in order to obtain faster detection speed, YOLO detector (Jiang et al., 2022) is selected to complete the target location and detection task based on linear regression. After obtaining the feature map output by the Transformer network, a standard convolution group modifies the size of the feature map and the number of channels, and then a single YOLO detector (Fang & Li, 2022) completes the detection.

5. Experimental Results and Analysis

To validate the detection performance and superiority of the network designed in this paper for traditional village classification, in the experimental phase, the constructed dataset was used for training and testing. Additionally, other types of target detection algorithms with different structures were trained and tested under the same experimental conditions. Subsequently, the obtained evaluation metrics were comprehensively analyzed.

	Precision	Recall	F1-Score	IoU	mAP
Model 1	89.8	91.7	91.2	81.2	91.3
Model 2	92.6	94.2	93.8	84.6	93.8
Model 3	93.5	95.8	94.9	85.4	94.7
Our	97.8	98.8	97.6	88.9	97.2

Model 1: YOLO v5 network

Model 2: The feature extraction network is a dual Transformer branch network (QuadTree Attention mechanism is not used)

Model 3: The feature extraction network is a dual Transformer branch network (using QuadTree Attention mechanism)

Our: Transformer-Village

As shown in the table above, compared to Model 1, Model 2 exhibits a certain degree of improvement in indicators, indicating that the use of the Transformer mechanism can enhance the detection performance of the network. In comparison to both Model 2 and Model 1, Model 3 utilizes QuadTree Attention to further enhance the correlation between feature block sequences, highlighting the representation of targets in the feature map. This results in improved precision and recall, surpassing the detection accuracy performance indicators of both Model 1 and Model 2. In contrast to Model 3, the shallow network structure of Model 4 incorporates the dynamic FPN preprocessing network designed in this paper, enhancing the reuse of features. Dynamic convolution is employed at the connection of the two "pyramids" network layers. By weighting the channels of the convolutional core, the weight of targets in the feature map is increased, providing the final preprocessing network output with richer information about the shapes of targets. Additionally, the IoU indicator also significantly improves. In summary, this paper utilizes the Transformer mechanism to establish the positional relationships between targets, enhancing the weights of targets in the feature map and exhibiting superior detection performance compared to pure convolutional neural networks. Simultaneously, the adoption of the QuadTree Attention mechanism reduces the computational load of the Transformer module, further improving the relevance of effective targets. A preprocessing module is designed to optimize the network process, leveraging convolutional networks for image processing to extract feature maps with rich morphological information. The dual Transformer branch network is then employed for block calculations.

Figure 11. Traditional Village Recognition Image (0: Tibetan, 1: Qiang, 2: Hui, 3: Han)

6. Conclusion

In this study, our primary focus lies in the design of a feature extraction network for traditional village classification. This feature extraction network is primarily tasked with extracting features from various traditional villages. The Transformer-Village feature extraction network proposed in this paper demonstrates accurate and rapid detection when paired with the YOLO detector. The design of the Transformer-Village feature extraction network incorporates a "Conv+Transformer" structure. The preprocessing module in the network is devised based on the FPN network structure with dynamic convolution, thereby enhancing the reuse rate of features. Within the Transformer module, a dual Transformer branch network is designed, accommodating features of targets with different sizes by configuring distinct block partition sizes. Furthermore, the Transformer incorporates the QuadTree Attention mechanism, reducing computational load while enhancing the correlation between targets. Experimental results illustrate that, compared to other networks, the Transformer-Village network designed in this paper significantly excels in extracting features from traditional villages.

References

Aloysius, Neena, & Geetha, M. (2017). A review on deep convolutional neural networks. International conference on communication and signal processing (ICCSP). *IEEE*, 2017.

Deng, C. F. et al. (2021). Extended feature pyramid network for small object detection. *IEEE Transactions on Multimedia*, *24*, 1968-1979.

Fang, J. X., & Li, X. W. (2022). Object Detection Related to Irregular Behaviors of Substation Personnel Based on Improved YOLOv4. *Applied Sciences*, *12*(9), 4301.

Gao, J., & Wu, B. H. (2017). Revitalizing traditional villages through rural tourism: A case study of Yuanjia Village, Shaanxi Province, China. *Tourism management*, *63*, 223-233.

Huang, L. Q. et al. (2019). Pyramid-structured depth map super-resolution based on deep dense-residual network. *IEEE Signal Processing Letters*, *26*(12), 1723-1727.

Jiang, P. Y. et al. (2022). A Review of Yolo algorithm developments. *Procedia Computer Science*, *199*, 1066-1073.

Parmar, Niki, et al. (2018). Image transformer. International conference on machine learning. *PMLR*, 2018.

Wang, W. H. et al. (2021). Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. *Proceedings of the IEEE/CVF international conference on computer vision*, 2021.

Wu, Felix, et al. (2019). Pay less attention with lightweight and dynamic convolutions. *arXiv preprint arXiv*, 1901.10430(2019).

Yu, G. C. et al. (2022). Dual-branch attention-in-attention transformer for single-channel speech enhancement. ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). *IEEE*, 2022.