International Journal for Housing Science and Its Applications
Publish August 6, 2025. Volume 46, Issue 3 Pages 6597-6608

https://doi.org/10.70517/ijhsa463567

A Computer Visual Interpretation of Time Consciousness and Symbolic Space in Digital Image Art in the Digital Image Era

Yiming Li[1], Chenxi Ye[2], Peiyi Wang[3], Lingjian Yang[4] and Shengdong Zhou[5,6,*]
[1] School of Liberal Arts, Zhuhai College of Science and Technology, Zhuhai, Guangdong, 519000, China
[2] Faculty of Science, Hongkong Baptist University, Hongkong, 999077, China
[3] School of Culture and Creativity, Hongkong Baptist University, Hongkong, 999077, China
[4] Film and Television Media College, Nanchong Film Industry Vocational College, Nanchong, Sichuan, 637000, China
[5] Fujian Digital Media Economy Research Center, Fujian Social Science Research Base, Minjiang University, Fuzhou, Fujian, 350108, China
[6] Faculty of Decorative Arts, Silpakorn University, Bangkok, 10700, Thailand
Corresponding authors: (e-mail: jason19960127@163.com).

Abstract With the development of digital technology, digital image art as an emerging art form is reshaping people's perception of space and time. Artificial intelligence technology provides a key impetus for digital image art, revolutionizing the means of artistic creation. This study explores the expression of time consciousness and symbolic space in digital image art in the digital image era, and proposes an interpretation method based on computer vision technology. The purpose of the study is to enhance the spatio-temporal expression of digital image art through computer vision technology and realize the effective integration of time consciousness and symbolic space. Methodologically, the article proposes a dual-stream spatio-temporal fusion algorithm for digital images based on Swin Transformer, which is divided into three modules, namely, temporal feature extraction network, spatial feature extraction network, and fusion network, with information optimization through the CBAM module, and feature processing through the RB module, and ultimately realizes the deep-learning-driven creation of digital image art. The results show that the proposed algorithm performs well in the visual interaction of digital image art, and its root mean square error, peak signal-to-noise ratio and structural similarity reach 1.217, 46.841 and 0.943, respectively, which are far superior to the image segmentation algorithm, the cyclic differential filtering algorithm, and the focusing shape restoration algorithm with robust focusing volume regularization. In addition, the running time of this algorithm is only 2.07 seconds, which is more than 50% shorter than other algorithms. The Swin Transformer-based dual-stream spatio-temporal fusion algorithm for digital images provides technical support for the expression of time consciousness and symbolic space in digital image art, which can effectively meet the requirements of digital image art design, promote the deep integration of digital image art and computer vision technology, and provide users with a more wonderful visual experience.

Keywords digital video art, time consciousness, symbolic space, computer vision, Swin Transformer, dual-stream spatio-temporal fusion algorithm

I. Introduction

Every change in video art is inseparable from the development and progress of technology, the emergence of digital imaging has enriched the artistic expression of the film, transformed the concept of image creation, and formed a new way of generating images, which has become an important part of the category of contemporary art [1]. Broadly speaking, digital image art is an art that includes animation, photography, film and television, virtual reality and other digital means, and narrowly speaking, digital image art is an image recorded through digitization, including moving images and static images, pictures, video single frames, movies, videos, and new media images [2]-[5]. Digital images depend on the medium of communication to exist. The Internet, DV, computers, mobile terminals, etc., are all different media, and interactivity, functionality and privacy, and portability have brought new vitality to traditional art forms.

In today's trend of information technology development, the rapid development of digital media technology, so that people's aesthetic needs and requirements for visual effects are increasingly high, the traditional two-dimensional visual viewing mode can no longer meet people's needs, the need for a more visually impactful art form to stimulate people's nerves [6]-[8]. In terms of space and time, the addition of digital technology has enriched people's sensory experience, revolutionized the feedback form of video art, broadened the dimension of symbolic space, and reshaped the spatial characteristics of digital video art, as well as the subjective experience of the passage of time [9]-[11]. Nowadays, there are fewer studies on spatio-temporal unlocking of digital image art, which provides theoretical support for the development of digital image art by exploring the spatio-temporal characteristics

of digital image art. And computer vision is a technology to study the field of artificial intelligence, through the input of images or video signals, to realize the automatic processing and analysis of image information, which can be applied to image analysis and spatio-temporal exploration [12], [13].

As a new art form in the digital era, digital image art has become an important medium for creators to express their cognition and interpret the world. From early digital image processing to today's complex 3D modeling and animation production, digital image art has experienced a remarkable technological evolution. Especially in recent years, the rise of artificial intelligence technology has injected new vitality into digital image art, which not only changes the means of art creation, but also prompts people to rethink the nature and process of art creation. In this context, computer vision technology, as the core technology in the field of artificial intelligence, provides innovative expression possibilities for digital image art. Time consciousness and symbolic space in digital video art are two key dimensions, the former focuses on how video art presents and manipulates the experience of time, while the latter explores the spatial expression of symbols and the construction of meaning in video art. In theater art, digital image technology extracts or simulates representations of the real world to construct a theater space that transcends the limits of real time, so that the audience can experience the ups and downs of emotions and plot changes in a short period of time. At the same time, the symbolic space of digital images, as a spatial symbol with social significance, carries cultural and historical significance, and its material form can be perceived while containing the cultural connotation of ephemeral coexistence. However, traditional digital image art creation faces problems such as lack of inspiration, uneven quality of materials and low efficiency, which prompts researchers to explore new technical methods to enhance the expressiveness and creative efficiency of digital image art. Computer vision technology, especially deep learning algorithms, offers the possibility to solve these problems.

Based on the above background, this study proposes a Swin Transformer-based dual-stream spatio-temporal fusion algorithm for digital images, aiming to enhance the expression of time-consciousness and symbolic space in digital image art through deep learning techniques. The study takes computer vision technology as the core and constructs three modules, namely, temporal feature extraction network, spatial feature extraction network and fusion network, to realize the innovative expression of digital image art by extracting and fusing the temporal and spatial features of digital images. The study firstly analyzes the multi-level expression of time consciousness in digital image art, including the temporal and spatial externalization of the subjective world, the temporal interaction between the image and the audience, and the paradox of the technological growth rate and the thinking stagnation; secondly, it explores the theoretical basis and expressive characteristics of the symbolic space in digital image art; then it elaborates the principle and framework of the dual-stream temporal and spatial fusion algorithm based on the Swin Transformer for digital images; Finally, we verify the application effect of the algorithm in the creation of digital image art through experiments. This study not only helps to promote the deep integration of digital image art and computer vision technology, but also provides technical support for the innovative development of digital image art.

II. Time Consciousness and Symbolic Space in Video Art

II. A. Digital Video Art

Digital image art in the digital era is a new art form, which carries the creator's knowledge and interpretation of the world. Digital image art has gone through the process from the original simple digital image processing to the current complex 3D modeling and animation production. In recent years, artificial intelligence technology has stood out and become a key driving force in digital image art. While revolutionizing the means of art creation, it has also prompted a re-examination of the nature and process of art creation. Artificial intelligence can independently study artistic styles, produce special images and pictures, and even assist creators in creative conceptualization, breaking through the traditional creation of inherent human thought patterns. This interdisciplinary combination brings computer science and art closely together, bringing unprecedented vitality to the art of digital imaging and opening a new page of exploration into the creative expression of art.

II. B. Time Consciousness in Video Art

II. B. 1) Spatio-temporal externalization of subjective worlds

Image art has a special potential and dynamism in theater art, which can present still or moving images and change flexibly according to the needs of the plot. Digital imaging technology creates image content in the theatrical space by extracting or simulating representations of the real world, thereby constructing a theatrical space that transcends the limits of real time. Image creators shape a collection of images that the audience can understand, thus constructing an aesthetic of time that can be perceived and accepted by the audience. This acceleration of images demonstrates the power of the moment, allowing the audience to experience dramatic emotional ups and downs and rapid plot changes within a short period of time. At the same time, it also allows the audience to deeply feel the passage of time and the fickleness of things, and to be empathetic to the inner changes of the characters.

II. B. 2) Temporal interaction between image and viewer

In the process of digital video art interpretation, wall projections and stage performances are synchronized, in which slow, low and dark long shot images complement the stage performances, vividly presenting the dreary atmosphere of grass-roots villages and towns, the conservative concepts and the slow process of innovation. The slowing down of the images gives a new meaning to each frame, while guiding the audience to incorporate their own thoughts and feelings into it. The sense of time created by the play is unique in that it takes the past, present, and future as a whole, constituting a “plausible present”. Further, in the interaction with the audience, the spatio-temporal mechanism of digital images prompts the audience and the creator to construct the energy field of time together, thus bringing about a direct experience of time.

II. B. 3) The paradox of technological growth and stagnant thinking

Time and space form the basic framework for human understanding of the world. As experiencers of this framework, people perceive their existence through time and space. With the advent of the accelerated era, new media video art is reshaping people's traditional perception of time and space, and providing a new way of experience for the public. However, it is worth cautioning that “the more speed grows, the more control tends to replace the environment, the time of interactive activities gradually replaces the space of physical activities, and the value of a meaningful space is judged by how much information it can provide and how fresh it is. Excessive use of digital video technology tends to make the audience expect more dazzling special effects, thus ignoring the deeper meaning and connotation inherent in the nature of art. Therefore, the purpose of digital video production should not be to “show off”, but to return to the original purpose of technology applied to stage performances. It will guide the audience into a more subjective and profound inner time dimension, build a bridge connecting virtual and real time and space, thus making the concept of time and space more clear and reachable.

II. C. Symbolic space in video art

If the space with social meaning is regarded as a symbol, and the semiotics is applied to understand the meaning of this social space, it will have the triadic representation, object, and interpretation items of symbolic meaning, and it will be found that the triadic combination of space and the triadic combination of symbolic meaning fit together. Spatial practice produces a material form of social space that can be perceived, especially since it is not only a productive process or result of society. It can be said that this space has a cultural significance that is ephemeral and coextensive in addition to its concrete materiality. Therefore, based on the theory of symbolic space, digital image art is explored. The empty symbols formed by landscape elements with the help of digital imaging technology gather and connect the thinking space of multiple perspectives in the presentation of meaning, then based on the powerful openness, inclusiveness and reproducibility of the meta-universe, its existence will also provide more space for the development of the empty symbols and people's multi-dimensional ideology. Based on digital photography technology, there is a kind of “neutral observation” perspective, and this kind of confirmation is often cold and emotionless, from the perspective of image presentation or spatial aesthetics, this kind of seemingly calm and suppressed observation angle is actually a kind of white space for the process of expression and decoding, and this kind of objective and calm thinking and visual presentation with emotional feedback can not only provide a space for empty symbols, but also for the development of multi-dimensional expression of empty symbols and people. This kind of objective and calm thinking and visual presentation with emotional feedback can not only provide a new emotional mode for the expression of empty symbols, but also provide an extension space for the aesthetics of photography in a higher dimension.

II. D. Digital Image Art Based on Computer Vision

II. D. 1) Computer vision technology

Computer vision technology is a crucial technology in the field of artificial intelligence, a discipline that analyzes and processes images or videos so as to give computers the ability to recognize the real world as such [14], [15]. However, computer vision technology has rarely been integrated into art displays. The use of computer vision technology in the presentation of art works will make the combination of technology and art closer. There is no special requirement for the level of technology used in digital image design, and mature computer vision libraries, such as OpenCV library and BoofCV library, are used to meet the application requirements of the technology to participate in the creation of art. Kinect body sensing device, with its integrated depth acquisition, human body recognition, motion capture and other features, is very popular among developers. It mainly uses related computer vision technology, which can provide technical support for extracting object or human body information in real-time images, and can provide ideal interactive information for producing interactive image art works. From the perspective of computer vision technology, the Swin Transformer-based dual-stream spatio-temporal fusion

algorithm for digital images is formulated, which is divided into a temporal feature extraction network, a spatial feature extraction network, and a fusion network, and the detailed principles and frameworks will be given in the following.

II. D. 2) Temporal feature extraction network

(1) Global branch of time network

After the activation of ReLU function, the intermediate feature $F_0^{T-NL} \in \mathbb{R}^{64\times H\times W}$ of the global branch is obtained, and its calculation process can be referred to Eq. (1):

$$F_0^{T-NL} = H_{ReLU}(H_{CONV_1}^{T-NL}(F_{in}^T)) \tag{1}$$

Subsequently, the feature F_0^{T-NL} is fed into a chained structure consisting of N_1 RSTB modules connected in series to further extract global information. Each RSTB module sequentially extracts the intermediate features $F_1^{T-NL}, F_2^{T-NL}\cdots F_{N_1}^{T-NL}$ one by one, which is computed as shown in equation (2):

$$F_i^{T-NL} = H_{RSTB_i}^{T-NL}(F_{i-1}^{T-NL}), i = 1,2,\cdots,N_1 \tag{2}$$

where $H_{RSTB_i}^{T-NL}(\cdot)$ denotes the ith RSTB module of the global branch $(T-NL)$ of the temporal feature extraction network, and considering the semantic tendency of the time-varying information, here, N_1 is set to 4, and a deeper network structure is used to extract the information.

After activation by ReLU function, $F_{N_1+1}^{T-NL}$ is obtained as shown in equation (3):

$$F_{N_1+1}^{T-NL} = H_{ReLU}(H_{CONV_2}^{T-NL}(F_{N_1}^{T-NL})) \tag{3}$$

(2) Local branching for temporal networks

The local branch uses a convolutional layer $H_{CONV_1}^{T-L}(\cdot)$ on the input feature F_{in}^T, the kernel size, number, convolutional step size and padding size of this convolutional layer are the same as the global branch's $H_{CONV_1}^{T-NL}(\cdot)$, after activation using the ReLU function, it yields The intermediate feature $F_0^{T-L} \in \mathbb{R}^{64\times H\times W}$ of the local branch, which is computed as shown in equation (4):

$$F_0^{T-L} = H_{ReLU}(H_{CONV_1}^{T-L}(F_{in}^T)) \tag{4}$$

This is similar to the method of extracting global features using series-connected RSTB modules in global branching, and the computational process of extracting local feature information is shown in Equation (5):

$$F_i^{T-L} = H_{RB_i}^{T-L}(F_{i-1}^{T-L}), i = 1,2,\cdots,N_1 \tag{5}$$

where $H_{RB_i}^{T-L}(\cdot)$ represents the ith RB module in the local branch $(T-L)$ of the temporal feature extraction network, and the parameter of the local branch, N_1, is set to 4, considering the characteristics of the time-varying features, and a deeper network is used for feature extraction.

After obtaining the output $F_{N_1}^{T-L}$ of the last RB module, the subsequent network structure of the local branch is the same as that of the global branch, and firstly, a second convolutional layer $H_{CONV_2}^{T-L}(\cdot)$ (with convolutional kernel size 3×3, number of kernels of convolution of 64, and step size 1) is used, which is activated by the ReLU function After that, the intermediate feature $F_{N_1+1}^{T-L}$ is obtained, and this process is shown in Equation (6):

$$F_{N_1+1}^{T-L} = H_{ReLU}(H_{CONV_2}^{T-L}(F_{N_1}^{T-L})) \tag{6}$$

(3) Temporal feature fusion

After obtaining the output of the global branch $F_{out}^{T-NL} \in \mathbb{R}^{64\times H\times W}$ and the output of the local branch $F_{out}^{T-L} \in \mathbb{R}^{64\times H\times W}$, the time feature extraction network sums them up element-by-element to get the complete time-varying features $F_{out}^T \in \mathbb{R}^{64\times H\times W}$, a process shown in equation (7):

$$F_{out}^T = F_{out}^{T-NL} + F_{out}^{T-L} \tag{7}$$

The F_{out}^{T} contains the time-varying information in the original image, which will be used as input to the fusion network and processed in the fusion network $H_{REC}(\cdot)$.

II. D. 3) Spatial feature extraction networks

(1) Global branching of spatial networks

In the global branch of the spatial network, the input feature F_{in}^{S} first passes through a convolutional layer $H_{CONV_1}^{S-NL}(\cdot)$, where the size of the convolutional kernel is 9×9, the number of convolutional kernels is 64, the step size is 1, and the Padding is 4, and the intermediate feature of the global branch is obtained after activation by the ReLU function $F_0^{S-NL}\in\mathbb{R}^{64\times H\times W}$, and the whole process can be represented by Equation (8):

$$F_0^{S-NL}=H_{ReLU}\left(H_{CONV_1}^{S-NL}\left(F_{in}^{S}\right)\right) \tag{8}$$

Subsequently, the intermediate feature F_0^{S-NL} is fed into a chained structure consisting of N_2 RSTB modules connected in series to further extract global information [16], [17]. Each RSTB module sequentially extracts the intermediate features $F_1^{S-NL},F_2^{S-NL}\cdots F_{N_1}^{S-NL}$, and the computation process is shown in equation (9):

$$F_i^{S-NL}=H_{RSTB_i}^{S-NL}\left(F_{i-1}^{S-NL}\right),i=1,2,\cdots,N_2 \tag{9}$$

where $H_{RSTB_i}^{S-NL}(\cdot)$ denotes the ith RSTB module of the global branch $(S-NL)$ of the spatial feature extraction network, which is further processed using the second convolutional layer $H_{CONV_2}^{S-NL}(\cdot)$ (with convolutional kernel size of 3×3, convolutional kernel number of 64, step size is 1, Padding is 0), and after activation by ReLU function, $F_{N_2+1}^{S-NL}$ is obtained as shown in equation (10):

$$F_{N_2+1}^{S-NL}=H_{ReLU}(H_{CONV_2}^{S-NL}(F_{N_2}^{S-NL})) \tag{10}$$

(2) Local branching for spatial networks

The local branch firstly uses the convolutional layer $H_{CONV_1}^{S-L}(\cdot)$ (convolutional kernel size is 9×9, the number of convolutional kernels is 64, the step size is 1, and the Padding is 4) to process the input feature F_{in}^{S}, and after activation by the ReLU function, the intermediate feature of the local branch $F_0^{S-L}\in\mathbb{R}^{64\times H\times W}$, as shown in equation (11):

$$F_0^{S-L}=H_{ReLU}\left(H_{CONV_1}^{S-L}\left(F_{in}^{S}\right)\right) \tag{11}$$

Subsequently, the intermediate feature F_0^{S-L} is fed into N_2 series-connected RB modules to further extract the local feature information of the image, which is computed as shown in Equation (12):

$$F_i^{S-L}=H_{RB_i}^{S-L}(F_{i-1}^{S-L}),i=1,2,\cdots,N_2 \tag{12}$$

where $H_{RB_i}^{S-L}(\cdot)$ refers to the ith RB module of the local branch $(S-L)$ of the spatial feature extraction network, and since the spatial texture features belong to the outputs of the shallower level network, N_2 is set to 2 to use a relatively shallow network for the processing, and the parameter settings here are consistent with the global branch.

(3) Spatial information fusion

After obtaining the output of the global branch $F_{out}^{S-NL}\in\mathbb{R}^{64\times H\times W}$ and the output of the local branch $F_{out}^{S-L}\in\mathbb{R}^{64\times H\times W}$, the spatial feature extraction The two features of the network are summed by elements to get the complete spatial feature $F_{out}^{S}\in\mathbb{R}^{64\times H\times W}$, which is computed as shown in Eq. (13):

$$F_{out}^{S}=F_{out}^{S-NL}+F_{out}^{S-L} \tag{13}$$

The F_{out}^{S} contains the spatial texture information in the original image, which will be used as input to the fusion network and processed in the fusion network $H_{REC}(\cdot)$.

II. D. 4) Converged networks

The entire processing of the fusion network can be summarized by Equation (14):

$$L_{2-Pre} = H_{REC}\left(F_{in}^{R}\right) \tag{14}$$

The input feature F_{in}^{R} is fed into the first convolutional layer $H_{CONV_1}^{R}(\cdot)$ for feature integration (convolution kernel size of 3×3, number of convolution kernels of 64, and step size of 1) after information optimization by the CBAM module, which is activated by the ReLU function to get the Intermediate features $F_0^R \in \mathbb{R}^{64\times H\times W}$, and this part of the computational process is shown in equation (15):

$$F_0^R = H_{ReLU}\left(H_{CONV_1}^{R}\left(H_{CBAM}\left(F_{in}^{R}\right)\right)\right) \tag{15}$$

Subsequently, the fusion network uses N_3 RB modules connected in series to further process the intermediate feature F_0^R, which avoids problems such as gradient vanishing while maintaining the depth of the network, thus improving the performance of the network, which is computed as shown in Eq. (16):

$$F_i^R = H_{RB_i}^{R}(F_{i-1}^{R}), i = 1, 2, \cdots, N_3 \tag{16}$$

where $H_{RB_i}^{R}(\cdot)$ denotes the ith RB module in the fusion network, where N_3 is set to 2 and a shallower network structure is used for processing. After obtaining the output $F_{N_3}^{R}$ of the last RB module, the fusion network is further processed by the second convolutional layer $H_{CONV_2}^{R}(\cdot)$ (convolutional kernel size of 3×3, number of convolutional kernels of 64, and step size of 1). Subsequently, it is activated by the ReLU function to obtain the intermediate feature $F_{N_3+1}^{R}$, as shown in equation (17):

$$F_{N_3+1}^{R} = H_{ReLU}\left(H_{CONV_2}^{R}\left(F_{N_3}^{R}\right)\right) \tag{17}$$

Similar to the feature extraction network, the fusion network adds the initial feature F_0^R to the intermediate feature $F_{N_3+1}^{R}$ through residual concatenation, and the summed features are fed into the final convolutional layer $H_{CONV_3}^{R}$ for the image fusion reconstruction, which ultimately produces the image $L_{2-Pre} \in \mathbb{R}^{C\times H\times W}$, in the sample of this paper, the number of channels of the input image $C = 6$, so the size of the convolution kernel of $H_{CONV_3}^{R}$ is 3×3, and the number of convolution kernels is 6, with a step size of 1 to ensure that the number of channels of the output image is consistent with that of the input image, which is shown in Eq. (18):

$$L_{2-Pre} = H_{ReLU}\left(H_{CONV_3}^{R}\left(F_{N_3+1}^{R} + F_0^R\right)\right) \tag{18}$$

III. Computer Visual Interpretation of Digital Video Art

III. A. Algorithm Validation Analysis

III. A. 1) Experimental environment

In this paper, Pycharm is used as the program development IDE tool, Python is used as the base development language, Python version is 3.7, PyTorch 1.7.1 is used as the deep learning development framework, and all the codes are compiled and debugged on the deep learning server. The server CPU is 16 vCPU Intel(R) Xeon(R) Platinum 8350CCPU@2.60GHz, and the GPU is Geforce RTX 3090.The main toolkits used in the program running are Opencv, Numpy and so on.

III. A. 2) Training design

A total of digital image datasets are used in this chapter to train the network and compare the efficiency of the algorithms. The follow-up experimental comparison process introduces datasets to test the advantages and disadvantages of each algorithm. And a total of 100 iterations of the above four image art datasets are used to train the network model, using the Adam optimizer during the training process, with the initial learning rate set to 10-4. Data enhancements are randomly performed during the training process, (e.g., image rotations, sequence inversions, etc.), with the batch size set to 2.

III. A. 3) Evaluation indicators

In order to verify the visual interaction capability of Swin Transformer-based dual-stream spatio-temporal fusion algorithm for digital images, root mean square error, peak signal-to-noise ratio, two-dimensional correlation

coefficient, logarithmic root mean square error, bumpiness, and structural similarity are proposed as the main evaluation metrics of this research, which are described as shown below:

(1) Root Mean Square Error:

$$RMSE(S,D)=\sqrt{\frac{1}{M}\sum_{p\in M}(S_p-D_p)^2} \tag{19}$$

(2) Peak signal-to-noise ratio:

$$PSNR(S,D)=10\ln\frac{L^2}{MSE(S,D)} \tag{20}$$

(3) Two-dimensional correlation coefficients:

$$Corr(S,D)=\frac{\sum_m\sum_n(S_{mn}-\bar{S})(D_{mn}-\bar{D})}{\sqrt{\left(\sum_m\sum_n\left(S_{mn}-\bar{S}^2\right)\right)\left(\sum_m\sum_n\left(D_{mn}-\bar{D}\right)^2\right)}} \tag{21}$$

(4) Log root mean square error:

$$logMSE(S,D)=\frac{1}{M}\sum_{p\in M}(\log(S_p)-\log(D_p))^2 \tag{22}$$

(5) Structural similarity:

$$\begin{gathered}l(X,Y)=\frac{2\mu_x\mu_y+C_1}{\mu_x^2+\mu_x^2+C_1}\\ c(X,Y)=\frac{2\delta_x\delta_y+C_2}{\delta_x+\delta_y+C_2}\\ s(X,Y)=\frac{\delta_{xy}+C_3}{\delta_x\delta_y+C_3}\,SSIM(S,D)=l(S,D)\times c(S,D)\times s(S,D)\end{gathered} \tag{23}$$

(6) Bumpiness:

$$\begin{gathered}Bumpiness(S,D)=\frac{\sum_{p\in M}\min\left(0.05,\left\|H_f(p)\right\|_F\right)}{M}\times 100\\ f=S-D\end{gathered} \tag{24}$$

Among them, S is the depth map predicted by the 3D topography reconstruction model, $l()$ represents the brightness similarity function, $c()$ represents the contrast similarity function, $s()$ represents the structural similarity function, μ_x and μ_y represent the mean of X and Y respectively, δ_x and δ_y represent the variance of X and Y respectively, and δ_{xy} represents the covariance of X and Y, D is the standard depth map of the scene to be measured, $\bar{S}$ and $\bar{D}$ are the average values of the depth matrix, and H is the Hesse matrix.

III. A. 4) Comparative analysis

After determining the evaluation indexes, the image segmentation algorithm (GC), the ring difference filtering algorithm (RDF), and the Robust Focusing Volume Regularized Focused Formal Recovery (RFVR-SFF) are set as control algorithms, and the analysis of algorithmic comparison results is shown in Fig. 1, in which (a)~(f) are the indexes RMSE, PSNR, SSIM, Correlation, log MSE, Bumpiness. the results show that the GC algorithm utilizes image segmentation to introduce a priori information about the scene, and the RDF algorithm utilizes full focus map guided filtering to repair the outlier focusing points, and the model is susceptible to the influence of the scene texture. The RFVR-SFF algorithm performs relatively well in the focusing evaluation, but there are deep outlier points. The Swin Transformer's dual-stream spatio-temporal fusion algorithm for digital images proposed in this paper performs the best in the visual interaction of digital image art, and the mean values of its indexes are 1.217, 46.841, 0.943,

0.949, 0.0209, and 4.144, which fully verifies the practical efficacy of the Swin Transformer-based digital image dual-stream spatio-temporal fusion algorithm's practical efficacy.

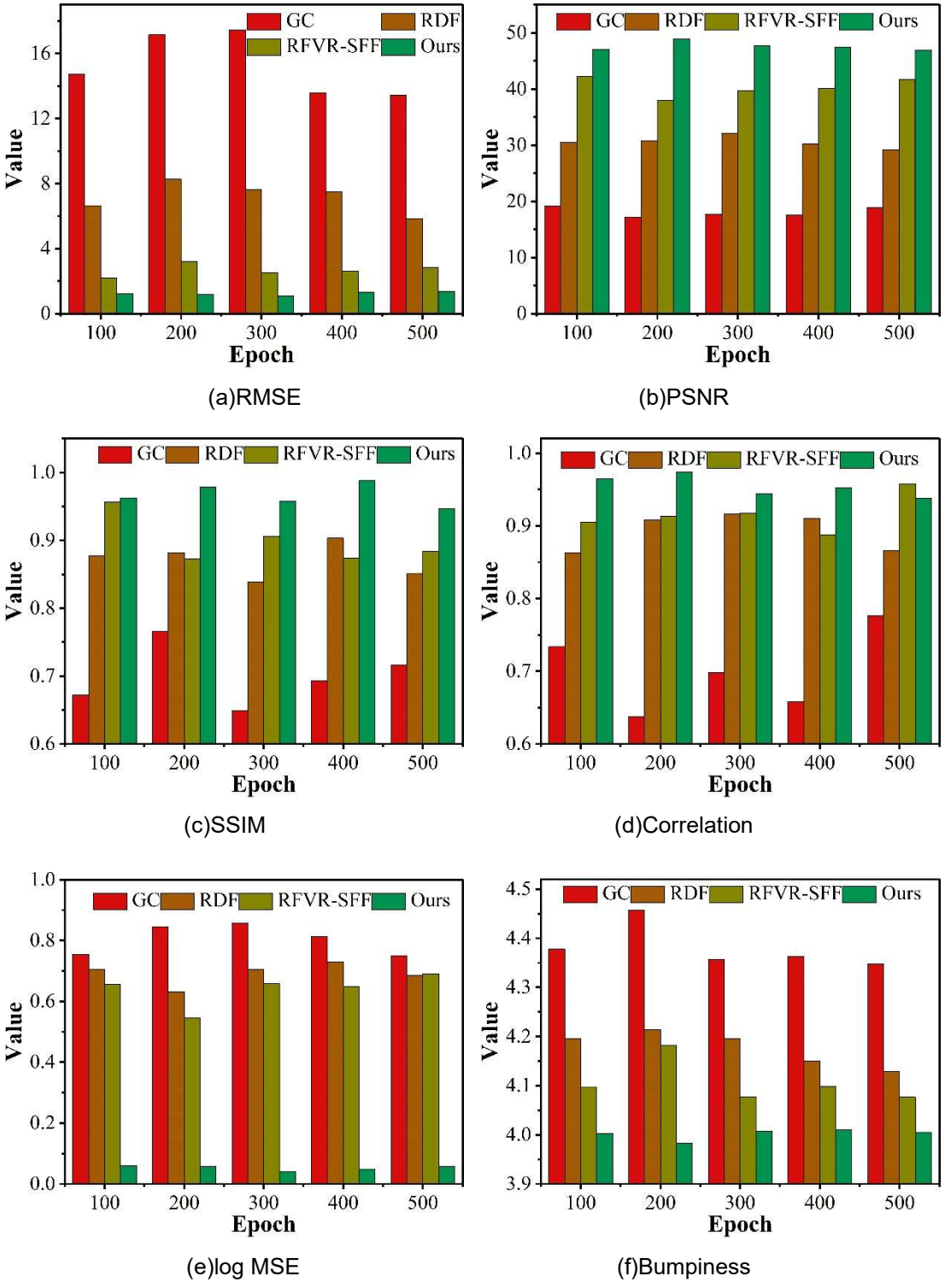

Figure 1: Analysis of Algorithm comparison results

III. A. 5) Comparative analysis of running time

The performance of the proposed algorithm in this paper was benchmarked on the research dataset, all the training and testing were conducted on the image processor with NVIDIA GeForce architecture, the average running time of the testing phase was controlled within 10s, the results of the running time comparison analysis are shown in Fig.

2, which shows that the GC algorithm, the RDF algorithm, the RFVR-SFF algorithm, the algorithm of this paper time 6.37s, 4.92s, 4.14s, 2.07s respectively, this paper's algorithm has a higher priority than the other three algorithms (GC algorithm, RDF algorithm, RFVR-SFF algorithm), and although this paper's algorithm's average time is not kept within 1s, it has been able to satisfy most of the real-time requirements of the digital image art, and it has been competitive in the latest research.

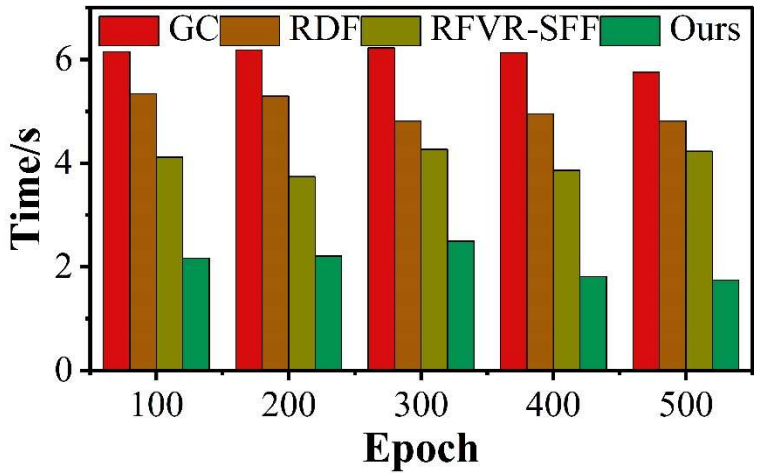

Figure 2: Comparison and analysis results of running time

III. B. Algorithm Application Analysis

III. B. 1) Specific applications

Digital imaging art creators in the design preparation session, in addition to good creative inspiration, should also collect a wealth of materials, otherwise the subsequent creation is difficult to achieve the desired effect". However, creators often face the problem of insufficient inspiration in the actual design process, and the time cost invested before creation is relatively insufficient, resulting in creative imagination is difficult to be fully stimulated. For example, when designing time consciousness and symbolic space in digital video art, the relevant creators should have a clear understanding of the requirements of the producer in order to effectively complete the design of time consciousness and symbolic space, and lay a good foundation for the subsequent scene layout and other links. In the preparation of scene concept design, designers should extract materials that meet the theme requirements from the library, but the quality of the materials is difficult to ensure, and the collection process is time-consuming, which will inevitably reduce the efficiency of the work. Under the trend of rapid development of artificial intelligence technology, it has become possible to generate the desired image content through Swin Transformer-based digital image dual-stream spatial-temporal fusion algorithm, which in turn improves the effect and efficiency of time-consciousness and symbolic-space design in digital image art. The integration of Swin Transformer spatio-temporal fusion algorithm into the design of digital video art can realize the enhancement of the user's interaction of time consciousness and symbolic space in digital video art, which can not only better guarantee the work efficiency, but also better design the film and television scenes to satisfy the people and present the users with a more wonderful visual feast.

III. B. 2) Algorithm application effect verification

This subsection will explore the application effect of Swin Transformer-based dual-stream spatio-temporal fusion algorithm for digital images from the time-awareness and symbolic-space fusion generation effect, and the validation analysis of the algorithm application effect is shown in Fig. 3, in which (a)~(f) are RMSE, PSNR, SSIM, Correlation, log MSE, and Bumpiness, respectively. The standard values of fusion generation indexes (RMSE<5, PSNR>40, SSIM>0.8, Correlation>0.8, log MSE<0.1, Bumpiness>4) of digital image art, this paper's method generates the effect of fusion generation of time-consciousness and symbolic space as a convergence of the standard, and the numerical results of the indexes are all in the reasonable aspect. Can meet the digital image art time consciousness and symbolic space design requirements. In summary, digital video art is an emerging art form that integrates science and technology and art, which can break through the limitations of traditional video art forms and create high-quality art works by computer vision, so that the audience can get a good audio-visual experience, and the aesthetic demand can be satisfied to a higher degree. Computer vision technology is the product of the integration of a number of advanced technical means, it has obvious intelligence, has been widely used in the field of art

creation, it is practical to introduce it into the design of digital video art, not only helps the innovation of the design content, but also makes the digital video art of time consciousness and symbolic space design in line with the needs of users, and promotes the rapid development of high quality digital video art.

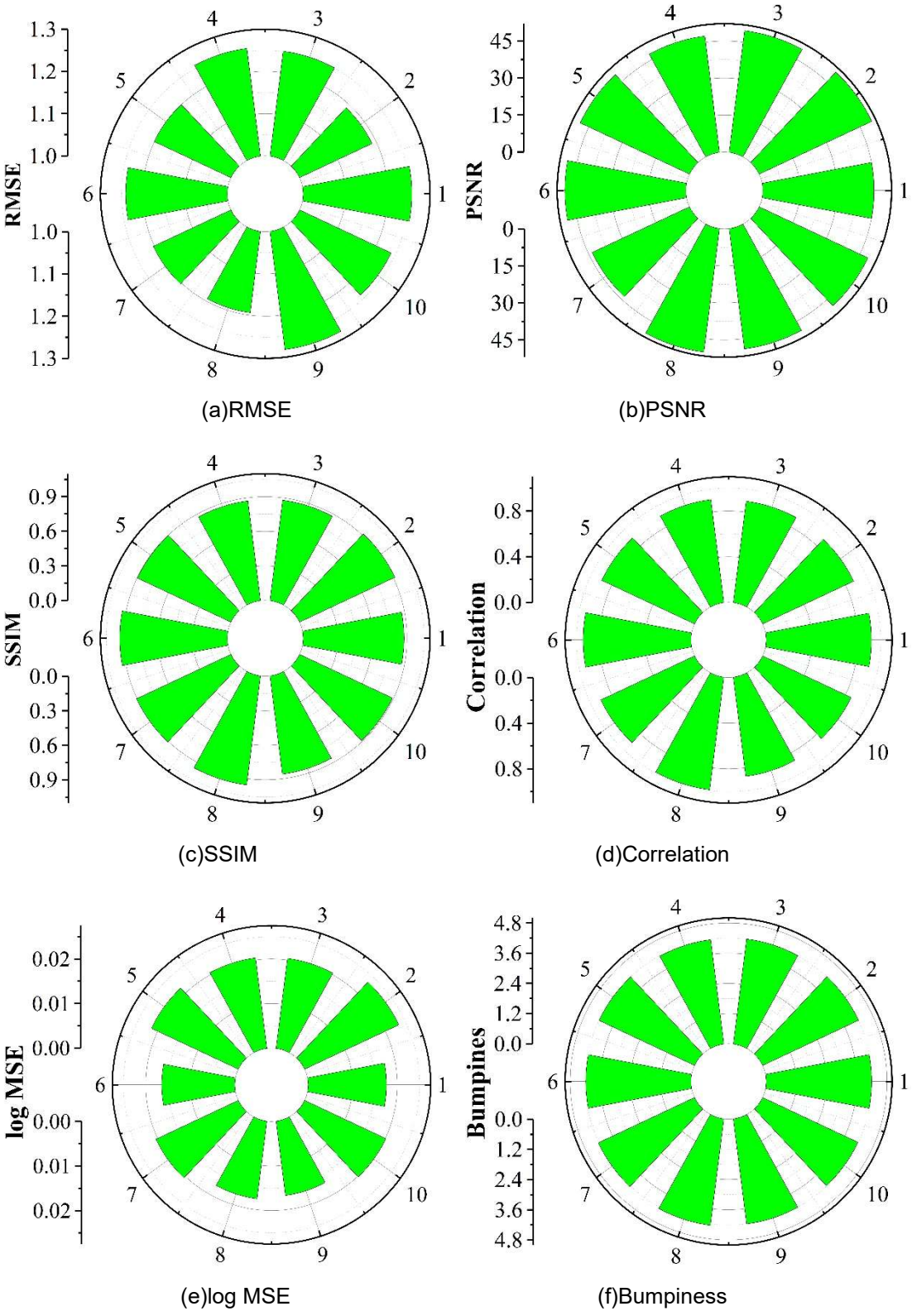

Figure 3: Verification of the application effect of the algorithm

IV. Conclusion

This study explores the computer visual interpretation of time consciousness and symbolic space in digital image art in the digital image era, proposes a dual-stream spatio-temporal fusion algorithm for digital images based on Swin Transformer, and experimentally verifies the effectiveness of the algorithm in the creation of digital image art. The results show that the proposed algorithm performs well in the visual interaction of digital image art, with the average values of root mean square error 1.217, peak signal-to-noise ratio 46.841, structural similarity 0.943, and two-dimensional correlation coefficient 0.949, which substantially exceed the performance of the image segmentation algorithm, the circular differential filtering algorithm, and the focusing shape recovery algorithm with robust focusing volume regularization. In addition, the average running time of this algorithm is only 2.07 seconds, which is significantly shorter than that of the comparison algorithms, which is from 4.14 to 6.37 seconds, and meets the demand of real-time processing of digital image art.

The Swin Transformer-based dual-stream spatio-temporal fusion algorithm for digital images realizes the effective expression of time-consciousness and symbolic space in digital image art through the collaborative work of three modules: temporal feature extraction network, spatial feature extraction network and fusion network. The algorithm not only solves the problems of insufficient inspiration and poor quality of materials in traditional digital image art creation, but also improves the creative efficiency and artistic expression, presenting users with a more wonderful visual experience.

Research shows that computer vision technology, as a product of the integration of a number of advanced technical means, has obvious intelligence and innovation, and its introduction into the field of digital image art design is not only feasible but also necessary. This technology can help digital video art break through the limitations of traditional forms, create more attractive and infectious works of art, and meet the growing aesthetic needs of the audience.

Future research should further explore the integration path between deep learning technology and art creation, develop more intelligent algorithms and tools applicable to digital imaging art, promote the deep integration of digital imaging art and computer vision technology, and provide stronger technical support for the innovative development of digital imaging art. At the same time, we should also pay attention to the balance between technology and art to ensure that the technology serves the essential needs of artistic expression rather than simple “showmanship”, so as to promote the high-quality development of digital imaging art.

Funding

1. This work was supported by Zhuhai College of Science and Technology Jinwan District Library Intangible Cultural Heritage Communication Practice Teaching Base (ZLGC20230303).

2. This work was supported by Fujian Provincial Science and Technology Plan Project for International Cooperation, Surface, China-South Africa Textile Apparel Digital Twin Collaborative Intelligent Manufacturing System Construction and Evaluation (No: 2023I0040).

About the Authors

Yiming Li was born in Runan County, Henan Province, China, in 1993. She obtained a Ph.D. from Wuhan University, China. She is currently a lecturer at the School of Liberal Arts, Zhuhai College of Science and Technology. Her main research interests include film and photography arts and traditional Chinese culture.

Chenxi Ye was born in Guangdong, China, in 2001. He obtained a bachelor's degree from Guangdong University of Technology in China and Vellore Institute of Technology in India. He is currently studying at the Faculty of Science, Hong Kong Baptist University. His main research direction is Interaction Design, Artificial Intelligence, and Digital Media.

Peiyi Wang was born in Changzhi, Shanxi,China,in 2001. She obtained a bachelor's degree from Guangxi Minzu University in China. She currently studying at the School of Culture and Creativity,Hongkong Baptist University, Her main research direction is Film,TV&New Media Concentration.

Yang Lingjian was born in Yibin, Sichuan. China, 2001. He obtained a bachelor's degree from Chengdu Neusoft. He is currently employed at the Film and Television College of Nanchong Film Industry Vocational College. His main research direction is the interdisciplinary integration of digital media art.

Shengdong Zhou is a PhD student in Art and Design at Silpakorn University, under the supervision of Professor Eakachat Joneurairatana. He is also an assistant researcher at the Fujian Digital Media Economy Research Center, Minjiang University. His research interests include graphic design, visual art, virtual reality, Artificial Intelligence and Design, and the intersection of computer and art market. Zhou has participated in the International Conference on Future Technology and Industry 5.0 and won awards. He has been involved in the design of Bangkok Design Week

2023-2024, the Design Workshop Project in cooperation with Tokyo University of the Arts, and the Design Workshop Project in cooperation with the Academy of Fine Arts in Florence, Italy.

References

[1] Alforova, Z., Marchenko, S., Kot, H., Medvedieva, A., & Moussienko, O. (2021). Impact of digital technologies on the development of modern film production and television. Rupkatha Journal on Interdisciplinary Studies in Humanities, 13(4), 1-11.

[2] Fang, J., & Gong, X. (2023). Application of visual communication in digital animation advertising design using convolutional neural networks and big data. Peerj Computer Science, 9, e1383.

[3] Sun, L. (2022). Research on Digital Media Art Film and Television Special Effects Technology Based on Virtual and Reality Algorithm. Scientific programming, 2022(1), 4424772.

[4] Mandal, P. C., Mukherjee, I., Paul, G., & Chatterji, B. N. (2022). Digital image steganography: A literature survey. Information sciences, 609, 1451-1488.

[5] Gerling, W. (2018). Photography in the digital: screenshot and in-game photography. photographies, 11(2-3), 149-167.

[6] Elitaş, T. (2020). Presentation of Visual Culture Elements in Digital Environments With Special Effect Technologies. In New Media and Visual Communication in Social Networks (pp. 1-16). IGI Global.

[7] Zhou, Y., Hu, X., & Shabaz, M. (2021). Application and innovation of digital media technology in visual design. International Journal of System Assurance Engineering and Management, 1-11.

[8] Alzubi, A. M. (2022). Impact of new digital media on conventional media and visual communication in Jordan. Journal of Engineering, Technology, and Applied Science (JETAS), 4(3), 105-113.

[9] Li, Y., Ye, C., Wang, P., Yang, L., & Zhou, S. (2025). An Interpretation of Computational Methods of Time Consciousness and Symbolic Space in Photographic Art in the Age of Digital Imaging. J. COMBIN. MATH. COMBIN. COMPUT, 127, 1477-1493.

[10] Du, J. (2024). The Application of Computer-Aided Film and Television Art Design Elements in Exhibition Space Design. International Journal of High Speed Electronics and Systems, 2540113.

[11] Spallone, R. (2017, November). In the space and in the time. Representing architectural ideas by digital animation. In Proceedings (Vol. 1, No. 9, p. 962). MDPI.

[12] Chen, Z., Li, H., Bao, Y., Li, N., & Jin, Y. (2016). Identification of spatio‐temporal distribution of vehicle loads on long‐span bridges using computer vision technology. Structural Control and Health Monitoring, 23(3), 517-534.

[13] Ma, D., Dang, B., Li, S., Zang, H., & Dong, X. (2023). Implementation of computer vision technology based on artificial intelligence for medical image analysis. International Journal of Computer Science and Information Technology, 1(1), 69-76.

[14] Sarah Nawoya,Quentin Geissmann,Henrik Karstoft,Kim Bjerge,Roseline Akol,Andrew Katumba... & Grum Gebreyesus. (2025). Prediction of black soldier fly larval sex and morphological traits using computer vision and deep learning. Smart Agricultural Technology,11,100953-100953.

[15] Yu Ling & Chung Wonjun. (2023). Analysis of material and craft aesthetics characteristics of arts and crafts works based on computer vision. Journal of Experimental Nanoscience,18(1),

[16] Jaewan Choi,Doochun Seo,Jinha Jung,Youkyung Han,Jaehong Oh & Changno Lee. (2024). Cloud Detection Using a UNet3+ Model with a Hybrid Swin Transformer and EfficientNet (UNet3+STE) for Very-High-Resolution Satellite Imagery. Remote Sensing,16(20),3880-3880.

[17] Akhaury U.,Jablonka P.,Starck J. L. & Courbin F. (2024). Ground-based image deconvolution with Swin Transformer UNet. Astronomy & Astrophysics,688,