
Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2025 SCPE. Volume 26, Issues 5, pp. 1974–1984, DOI 10.12694/scpe.v26i5.4847

DEEP ANALYSIS ON THE COLOR LANGUAGE IN FILM AND TELEVISION ANIMATION WORKS VIA SEMANTIC SEGMENTATION TECHNIQUE

YANXIANG ZHANG*

Abstract. Color functions as a distinct type of ideographic symbol in animation for film and television, playing a crucial role in enhancing visual narratives and conveying emotions. In different types of animation, such as fantasy, horror, or children's genres, color language influences audience perception and can convey meanings beyond the capabilities of image language alone. For instance, bright colors may symbolize innocence in children's animation, while darker shades may evoke tension or fear in horror. However, current approaches to representing color in animation often fail to capture its full semantic richness and ideographic potential. Existing methods primarily focus on image-based analysis, overlooking the deeper layers of meaning encoded in color language. In this paper, we address these gaps by utilizing semantic segmentation techniques to combine the three modalities of color, content, and text to establish a consistent representation of color language in animation. We propose a method for semantic segmentation of color-depth (RGB-D) images using two-stream weighted Gabor convolutional network fusion. A weighted Gabor orientation filter builds a deep convolutional network (DCN) capable of extracting feature information adaptive to changes in orientation and scale, allowing for orientation- and scale-invariant features. Dual-stream picture features—color and depth—are extracted using a broad residual-weighted Gabor convolutional network and then combined into a lightweight feature extraction network. To evaluate the ideographic functions of color language quantitatively, we conducted extensive experiments using open databases. Our proposed method outperforms existing RGB-D picture semantic segmentation algorithms, demonstrating its effectiveness in representing color language in animation.

Key words: Semantic Segmentation, Video Animation, Color Language

1. Introduction. Text symbols and colour have a definite relationship in animation for film and television, and they can also naturally blend together with visuals. It is a kind of code in film and television works and has the function of expressing meaning. Ideology is the main feature of color [1]. The ability to combine the language of television and films, as well as its meaning, are qualities bestowed by life itself. People and colour have already had a significant impact on people's daily lives [2]. There is a certain experience in the expression of different colors, and different intentions are formed between different colors [3]. Not only can colour expressions convey meaning in the language of cinema and television animation, but colour also functions as a type of code with symbolic meaning [4]. It has distinct qualities. Colour can communicate diverse information when working in tandem with voice and vision. For example, as shown in Fig.1.1, due to the different shades of colors, the emotions of the characters are also expressed. Through the distinction of color and brightness, the contrast of the emotions of the protagonists in the film can be formed.

In animation production, color is not only an aesthetic element, but also a symbolic language that can convey emotions and suggest the direction of the plot. For example, in the same scene, by adjusting the brightness or hue of a color, the audience's perception of the emotion will change significantly. However, traditional image processing methods often fail to capture the potential meaning of color language in depth, especially when there is a difference between the semantic expression of color and image content. Simple color analysis methods cannot adequately express the semantic information in complex animations, which makes it challenging to study the color language.

Semantic segmentation techniques provide a way to solve this problem. By segmenting an image into regions with independent semantics, semantic segmentation enables a better understanding of objects and backgrounds in an image and associates colors with their symbolic meaning in a particular scene. Especially in animation, semantic segmentation can not only extract the boundaries and shapes of objects, but also identify the colors of different regions and their corresponding semantic information, so as to dig deeper into the expressive function

*HeNan Polytechnic Public Art Teaching Department, Zhengzhou 450046, China.(727842564@qq.com).

Fig. 1.1: Snow White Fragment.

of color language.

Therefore, this paper proposes to study the color language through semantic segmentation techniques and use the color-depth (RGB-D) image segmentation method to better understand the symbolic meaning of color in animation. This approach can effectively deal with the complex relationship between color language and image semantics, and provide a new perspective for understanding and analyzing the application of color in animation works.

If the color segments of different regions can be segmented, similar regions will present the same semantics, and information transmission can even be accomplished without subtitles. Consequently, the ideographic function of colour language in animation for film and television can be efficiently interpreted by semantic segmentation[5]. Traditional semantic segmentation methods generally use classifiers to perform pixel-level classification of artificial features and use conditional random fields (CRF) for refinement. However, the design of the classifier is generally aimed at a single category, and the classifier has a large training difficulty and high computational complexity when it is used for multi-category segmentation tasks [6]. In addition, the manually designed features have great limitations, resulting in poor model generalization ability and low segmentation accuracy. Deep learning-based semantic segmentation has demonstrated significant benefits thus far. It is capable of end-to-end training in addition to multi-category segmentation. Its varieties can be broadly categorised as color-depth (RGB-D), codec-based, and mixed with CRF. Three types of image fusion segmentation frameworks[7]. We propose a semantic segmentation method based on color-depth (RGB-D) images for better representation of color language in animations. When dealing with this kind of task, it is a major challenge to capture both color and depth information in an image and establish the semantic relationship between them. Especially in scenarios where color language and image language express different meanings, it is difficult for traditional methods to handle these complex semantic information effectively.

To solve this problem, we design a two-stream weighted Gabor convolutional network fusion method. First, a deep convolutional network is constructed by a Gabor directional filter to adapt to the orientation and scale variations of the image, so as to extract orientation- and scale-invariant features. Then, we extract image features from the color and depth channels separately and combine them into a lightweight feature extraction network.

The contributions of this paper are three points: (1) We present semantic segmentation as a means of understanding the ideographic role of colour language in animation for film and television. (2) Our suggestion is to use a two-stream weighted Gabor convolutional network for color-depth (RGB-D) picture fusion. Method of semantic segmentation (3) our approach is better than the RGBD picture semantic segmentation methods currently in use.

2. Related works. Colour matching in animated films and television shows must be determined by the story's narrative. The meanings that are created when various colours are combined also differ. It doesn't matter if it's colour matching within an image or colour matching across multiple photos. able to communicate the story's meaning.

2.1. In a single-frame composition, colour matching. When diverse colours are mixed in animation for film and television, the resultant colours frequently lose the original colour of the object. Together, these hues can convey a variety of connotations that together convey the main theme of the movie, particularly Certain spatial properties of colour elements can give a single image a dynamic beauty through various colour compositions, combining the storyline of the film and more effectively expressing its message. These spatial features are found in both cinema and television animation works [8]. When matching the colors of a single composition, the matching of colors can play a role in foiling, and several colors with relatively large contrast can be matched together, to form opposing colors, and people may feel opposed to one another when they see distinct colours. powerful visual impression that can support the plot's development [9]. When producing animation for film and television, the audience can be made to feel the emotions in the works by manipulating the colours. This can help to communicate feelings and emotions while also giving the viewers a powerful visual impact while they watch the animation. There will be enjoyment while working [10].For instance, the animation "Prince of Egypt" used a stark contrast between cool and warm tones to help the viewer distinguish between the good and the evil. When Jesus was crucified in Egypt, the film appeared black; during their fight, Moses appeared orange; and Pharaoh Ramses II is blue [11]. Characters and visual aesthetics in animated films and television shows can be shaped by colour choices. A single-frame composition can convey the stress of animation by combining contrasting colours. It usually has no visual effect in pure colours. When colours are combined, it can give the spectator a sense of movement [12]. For example, mixing cool and warm tones can work well and create great tension.

2.2. Colour coordination between images or lenses. Combining various colours in a single composition can produce a more logical impact and intensify the tension in the image. Especially in cinema and television animation works, continuity is created between the images by matching different colours to make the colour differences between the images obvious and enable more natural connections between the images, and compared with a single color, it can be more effective. It reflects the transmission effect of complex events [13]. After combining different colors between the pictures, it can show the sense of space of the picture, and allow the viewer to experience the change in time between various images. In animation, colour serves as a "time chain" in both film and television. It has the ability to produce many settings [14]. Colour is a type of expressive symbol that may be used to represent a movie's theme directly, set the mood, change and match colours, and then work in tandem with character language and actions to create the colour and theme of the movie. By coming to a consensus, it might make the movie's theme easier to explain [15]. For instance, the sun emerges from behind clouds in the opening scene of the motion picture and television cartoon "Prince of Egypt". The colours are used extremely well in this picture. The shot uses a mixture of white and blue colors. This kind of Cool tones give a sense of unease and make the viewer feel like something is going to happen right away.

As an important part of the model in this paper, the main role of the "pyramid pooling module" is to capture image information at different scales through multi-scale feature pooling. Simply put, pyramid pooling pools the input image features at different scales, enabling the model to focus on both global and local information in the image. This is important for complex scenes, especially in the presence of multi-scale objects, and the module can effectively alleviate problems caused by differences in object size and location.

Specifically for the combination of RGB and depth image features, the pyramid pooling module is able to process these two types of features separately to extract rich semantic information at different scales. For RGB images, it can help capture color and texture features; for depth images, it can better extract spatial contours and edge information. By pooling RGB and depth features at different scales and fusing them in subsequent steps, the model is able to more accurately deal with the problem of scale differences of objects in the scene, which in turn improves the accuracy and robustness of semantic segmentation.

This multi-scale processing makes the pyramid pooling module have obvious advantages in complex scenes, especially in dealing with objects of different sizes and shapes, and it can provide more accurate feature representation capability.

2.3. Semantic segmentation. The encoding and decoding-based segmentation method is divided into two parts: encoding and decoding. The encoding part is used to extract features, while the decoding part is used to gradually recover the lost spatial information, including U-Net, fully convolutional neural network (FCN),

Signet and DeepLabv3+ et al. The paper [16] regards the semantic segmentation problem as an instance segmentation problem and proposes a deep deconvolution network Deconned. The method firstly performs pixel-by-pixel category label recognition, predicts the segmentation mask, and then sends it into the network to obtain a combination of segmentation results through training, so this method can handle objects of different scales and enhance the processing of image details.

However, the extraction of target candidate boxes requires a lot of time and storage space, and the process is complicated, making it difficult to achieve fast and accurate segmentation. The paper [17] added the global context information to the fully convolutional network, and proposed a Parse Net network, which uses the average feature of any layer in the network to represent the feature of each position, It is employed to enhance the model's segmentation performance by capturing the image's global semantic information. The paper [18]proposed an improved symmetric encoding-decoding network SIPRNet, It builds an end-to-end semantic segmentation network by fusing semantic information and image data using the pooling index and convolution. Performance segmentation. In order to solve the problem that the up-sampling operation in the fully convolutional network cannot compensate for the loss of information, aurous convolution is used to achieve multi-level context aggregation, and the resolution of the feature image is not reduced. This method improves the performance based on FCN. Suggested an FCN model for the semantic segmentation of high-resolution remote sensing pictures that takes class imbalance into account. The end-to-end accuracy of small class prediction was improved by using the adaptive threshold method and the weighted cross-entropy loss function. Researchers have suggested techniques like aurous convolution and deconvolution to try and recover as much of the feature information that was lost during the downsampling process. However, aurous convolution requires a lot of processing power and storage space, while deconvolution is unable to restore low-level features. Both are not helpful for quick and precise semantic segmentation. suggested a network of many extraction stages. Refine Net to combine feature maps with varying resolutions and combine data that was lost during downsampling by using a high number of residual connections. Furthermore, residual connection and identity mapping are used to achieve end-to-end training. This can improve the segmentation impact by giving each layer's features a distinct pertinence. suggested an approach for semantic segmentation of polarization synthetic aperture radar images that combines the depth characteristics of each. Designed with a dense up sampling convolution (DUC) and hybrid dilated convolution (HDC) structure, the former can be utilized to expand the network and provide predictions at the pixel level, while the latter can be used to capture and decode missing information in bilinear up sampling. In order to prevent grid issues brought on by aurous convolution procedures, the receptive field is employed to aggregate global information. In general, lighting effects can be overcome and more edge and spatial information can be provided for indoor scene semantic segmentation by integrating RGB and depth images and taking advantage of depth channels. Unfortunately, the majority of existing semantic segmentation techniques either do not take into account the utilization of depth channels for context inference or use relatively single-source depth information that is solely utilized to create regional-level features. Additionally, the complexity of network training is increased by the filters' inability to leverage previous knowledge, such as the orientation and scale of indoor objects.

3. Methods. Fig.3.1 illustrates the general layout of the model used in this paper, which is centered on the convolution operation based on weighted Gabor directional filters, and combines the feature fusion and encoding-decoding frameworks in order to improve the model's ability to express the features abstractly. Specifically, the model in this paper takes RGB and depth images as inputs, extracts the depth features of the images through a novel wide residual Gabor convolutional network, and then pools the RGB and depth image features using a pyramid pooling module, respectively. The dual-stream features at different scales are acquired by this module to alleviate the object difference problem. In addition, RGB and depth image features of different scales are cascaded and fused. The fused multi-scale features are up-sampled and cascaded features in the form of decoding to obtain the fused features containing information about different scales, which are finally fed into the SoftMax classifier for classification.

Here, the role of the SoftMax classifier is particularly crucial.The SoftMax function converts the output of the model into a probability distribution, especially in multi-category classification tasks, where it is able to differentiate between features of different categories and assign a probability value to each category. In this way, the model is able to identify which category it belongs to from the multiscale fusion features extracted

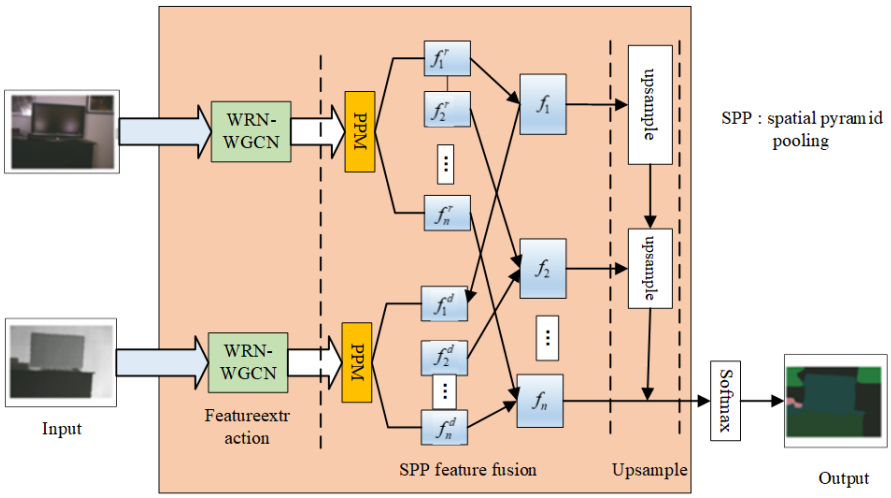

Fig. 3.1: Semantic segmentation of RGBD images fused by dual-stream weighted Gabor convolutional network.

from the RGB and depth features of the input image, thus accomplishing the multiclassification task.

3.1. Weighted Gabor directional filter. Deep convolutional neural networks have a long training period, high space complexity, difficulty adapting their extracted features to changes in direction and size, and an inability to dynamically update their own parameters based on feature differences. There are certain benefits to using traditional filters for feature extraction. Features that are invariant to spatial change are extracted through focused image processing, and feature redundancy is frequently lower than that of deep convolutional neural networks. The Abor filter exhibits good qualities in obtaining the target's local spatial-frequency information because it is a filter that closely resembles the basic cellular visual stimulus response. The Gabor filter resembles the convolutional neural network's shallow filter, according to the visualization results. A weighted Gabor direction filter is suggested, based on the developed Gabor direction filter. The convolution filter is modulated by the Gabor filter. The adaptability of the feature to the direction and scale is improved while lowering network parameters by modifying the convolution filter's feature extraction procedure. In order to highlight the differences between features in different directions, the method described in this paper involves the following specific processing steps: first, generate Gabor filters in different scales and directions; second, learn a weight coefficient for the filters; and finally, modulate the convolution filter. which, in order to make the output features flexible to the orientation and scale of the picture, generates a weighted Gabor orientation filter (Whoof) in each direction. Among them, Whoof is used as a filter with adjustable parameters, and the convolution filter is adjusted by the Gabor filter to enhance the expression of the feature map. The calculation process of the weighted Gabor direction filter is as follows: First, generate Gabor filters with U directions and V scales, weight each direction by learning a weight vector W, and then apply a learnable filter of size $N \times M \times M$. $M \times M$ represents the size of the two-dimensional filter and the modulation process is:

$$C_{i,u}^{v} = C_{i,o}^{o}[W.G(u,v)] \tag{3.1}$$

where u and v are the direction and scale indices, respectively; $G(u,v)$ is the corresponding Gabor filter; $C_{i,o}^{o}$ is the learnable filter; u is the dot product operation; $C_{i,u}^{v}$ is the modulation filter. Whoof can be expressed as:

$$C_{i}^{v} = \left(C_{i,1}^{v}, C_{i,2}^{v}, \cdots, C_{i,U}^{v}\right) \tag{3.2}$$

Since the Gabor filter has multiple directions, Whoof can be regarded as a three-dimensional convolutional filter, where the filter scale has different performances at different layers.

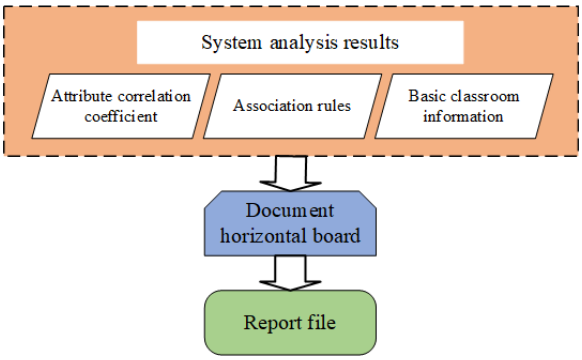

Fig. 3.2: Broad residual module. (a) initial residual module; (b) Extensive residual module L; (c) Extended residual module 2.

3.2. Wide Residual-Weighted Gabor Convolutional Network Module. The deep neural network model's layers are continuously deepened, which improves learning and allows for the extraction of richer characteristics. However, due to the influence of the gradient disappearing, the test accuracy of the model becomes harder to improve when the number of network layers reaches a certain point. Additionally, even as the training process goes on, the test accuracy gradually declines, and ultimately the loss function is unable to converge to the minimum value. To alleviate this problem, the residual module came into being.

The main idea is to use shortcut connections to skip multiple convolutional layers, and by continuously stacking modules, the network can continue to deepen without being affected by the disappearance of gradients. However, the deepening of the network makes the problem of decay feature reuse gradually prominent, resulting in only some parameters in the residual module participating in the update. A module-stacked network model is shallow and wide, so that a shallower model can be used to represent a deeper network. The background of the indoor scene is relatively complex, and there is interference from various illuminations, which makes feature extraction difficult. To construct features with high resolution, it is often necessary to use a deeper network model, and it is necessary to continuously fuse feature images between layers to alleviate the problem of excessive differences in multi-scale objects. To extract better features while building a lightweight network, this paper uses WRB to build a feature extraction network to extract RGB and depth image features respectively. The primary distinction between WRB and the standard residual module is the increase in the coefficient k and the quantity of convolution kernels. This guarantees the number of parameters while reducing the number of network layers, thereby fulfilling the goal of expediting the model training process.

Fig.3.2 displays a comparison of the structures of the various residual modules, with Fig.3.2a displaying the original residual module with two convolutional layers, batch normalisation layers, and REL layers. Figure 3(b) and Figure 3(c) represent two different WRBs, among them, X_1 and X_{1+1} are the input features of the lth layer and the two layer, respectively; FM is the feature learning with different number and width of convolution kernels. Compared with the original wide residual, which only adds different structure coefficients, WRB not only adds a different number of convolutional layers, but also adds different features.

The number of graphs, in turn, builds a wide and shallow network model. In this paper, the convolution filters used in WRB to build the network are all Woof's. On the one hand, the model is shallow by using WRB, and on the other hand, Woof's is used to extract the orientation and scale invariant features in the image, to better focus on the RGB image. It can improve the model's ability to represent information by extracting edge contour information from the depth image. To build a lightweight model and use a shallow neural network to achieve the same performance of a deep neural network, the wide residual-weighted Gabor convolutional network module in this paper has three broad residual groups and a network layer number of 13. The width k of each residual group is set to 4. In each residual group, the depth coefficient L determines the structure of the residual group. The first residual group GCConv2 and the second residual group GCConv3 adopt the structure of Fig.4.1.The specific structural parameters are shown in Table3.1. Fig.4.2 displays the schematic

Table 3.1: WRN-WGCN structural parameter configuration.

Group name	Output feature size	Block type
GCConv1	$N \times N$	[3×39]
GCConv2	$N \times N$	$\begin{bmatrix} 3\times 3 & 16\times k \\ 3\times 3 & 16\times k \end{bmatrix} \times L$
GCConv3	$N \times N$	$\begin{bmatrix} 3\times 3 & 16\times k \\ 3\times 3 & 16\times k \end{bmatrix} \times L$
GCConv4	$(N/2) \times (N/2)$	$\begin{bmatrix} 3\times 3 & 32\times k \\ 3\times 3 & 32\times k \end{bmatrix} \times L$

architecture of the WRN-WGCN module structure. First, the Gono is used to convolve the input image, and then the features are extracted through two wide residual groups, and finally the corresponding feature image is output.

4. Experiments.

4.1. Data set and experimental platform. NYUDv2 dataset the experiments in this paper are carried out on the mainstream semantic segmentation dataset NYUDv2, which contains 1449 densely annotated RGB and depth image pairs (image resolution is 640pixel×480pixel) and contains 44 scenes and 35064 images. target, with 894 target categories. Figure 8 shows an RGB image, a depth image, and a semantic label in the dataset. In the experiment, 40 types of semantic labels are used, and the dataset is divided according to the standard division strategy. There are 654 photos used for testing and 795 images utilised for training. During the training phase, the photos undergo flipping, translating, cropping, and colour jittering in order to address the issue of inadequate data.

4.2. Experimental results and discussion. This work presents a method comprising many functional modules, including a weighted Gabor direction filter, pyramid pooling feature fusion module, and wide residual convolution module. To validate the effectiveness of every module, the proposed method is extended to generate four alternative approaches. Variant model 1 should be set to Baseline in order to compare each module's effectiveness: Using an RGB image and a depth image as input, a conventional convolutional neural network is used to extract image features. Feature cascade is then used for fusing, and the network's convolution operation uses a conventional convolution filter. Variant approach 2 is configured as WRN-CNN in order to confirm the efficacy of the suggested wide residual feature extraction network: RGB and depth images make up the input, the wide residual network is used for feature extraction, and the two features are directly graded. Union and fusion, where the wide residual network consists of regular convolutional filters. In order to confirm the efficacy of the suggested weighted Gabor direction filter, variation method 3 is configured as WGCN: the weighted Gabor direction filter is utilised as the convolution filter with RGB and depth images as input, and the two resulting features are directly Fusion and cascading. Compared to variant model 1, this method replaces the weighted Gabor directional filter with a regular convolution filter. To verify the effectiveness of the feature fusion method proposed in this paper, the variant method 4 is set as PP-Fusion: taking RGB image and depth image as input, extracting features through conventional convolutional neural network, and utilising the fusion module's pyramid pooling capability for fusion.

In this paper, a training set is constructed based on the NYUDv2 data set, and the semantic segmentation model in the indoor scene is obtained by training, and then quantitative analysis is carried out on the test set. This work additionally tests the model and related techniques on the SUN-RGBD dataset, evaluating and visualising the results to confirm the model's generalisation performance. To verify the effectiveness of the design of each module, this paper designs ablation experiments, constructs different network models according to four variant methods, trains and tests the models, and obtains evaluation indicators. Additionally, employing FCN and Signee semantic segmentation, the approach in this work is contrasted with the current classical methods. Table4.1 displays the quantification results on the NYUDv2 dataset.

Based on the quantification results of the NYUDv2 data set, the network model based on the weighted Gabor

Table 4.1: Comparing the outcomes of various segmentation methods using the nyudv2 dataset.

Method	WRN-CNN	WGCN	PP-Fusion	Block typeA_{cc}%	mA_{cc}%	$mlon_{cc}$%	$Wmlon_{cc}$
Ours	.	.	.	60.4	50.9	40.2	53.2
Variant1				58.4	41.7	30.2	45.9
Variant2	.			58.7	42.5	31.8	45.4
Variant3	.			60.9	45.3	35.9	50.5
Variant4			.	63.3	45.9	36.5	46.7
FCN				65.5	45.2	34.4	48.7
Seg Net				56.3	47.7	35.2	50.2

Table 4.2: Comparison of various segmentation algorithms' output on the sun-robed dataset.

Method	WRN-CNN	WGCN	PP-Fusion	Block typeA_{cc}%	mA_{cc}%	$mlon_{cc}$%	$Wmlon_{cc}$
Ours	.	.	.	58.3	38.6	28.3	42.1
Variant1				45.3	33.8	21.9	38.5
Variant2	.			44.9	34.6	23.2	38.7
Variant3	.			54.7	35.2	27.4	37.8
Variant4			.	56.2	35.7	26.2	36.4
FCN				49.6	36.6	23.7	35.9
Seg Net				48.9	34.7	26.3	38.4

direction filter has greater advantages over the standard convolution filter network, and the four evaluation indicators are enhanced by 2.5% when compared to the baseline model 6.6%, 5.7%, and 4.6% demonstrate how the learnable convolution filter may be modulated with Gabor filters of various orientations and scales to efficiently extract features from the image, reducing the likelihood that orientation and scale variation interference will affect the segmentation results. To a certain extent, the segmentation performance of the network model can also be enhanced by the multi-scale fusion of RGB image data and depth image features. The accuracy of the proposed method has significantly improved when compared to the classic semantic segmentation models FCN and Signet. The Four index has also improved by 4.5% and 3.0%, respectively. This indicates that the method is effective when combined with dual-stream images and that multi-scale feature fusion and invariant feature extraction can enrich the information expression of images. The segmented images are labelled with different colours in order to intuitively depict the experimental results, and this yields the semantic results that are displayed in Fig.4.1. The figure shows that the traditional semantic segmentation techniques, FCN and Signet, are generally capable of segmenting a variety of objects. This suggests that the encoder-decoder structure may be able to recover certain features, but its fine-grained performance makes it unsuitable for small-scale objects. Even more so. Thanks to the benefits of the proposed pyramid pooling feature fusion module in multi-scale processing, the segmentation accuracy of some objects with smaller scales is rather good, and the segmentation of some objects' edges is more refined. In detail, the weighted Gabor direction filter-based model described in this study performs better in terms of both direction and scale invariance. This suggests that features linked to direction and scale invariance are extracted, and that this has an impact on the segmentation effect. WRB's dual-stream network model of colour and depth images fully leverages a range of various information representations in the image, offering definite advantages in object integrity and fine edge segmentation.Cross-dataset experimental results are a valuable tool in the research of picture semantic segmentation because they allow different approaches' generalisation performance to be tested. This paper trains the model on the NYUDv2 dataset and tests it on the SUN-RGBD dataset to validate the cross-dataset segmentation outcomes of the proposed technique. The SUN-RGBD dataset has 37 semantic categories and 10335 registered colour and depth image pairs. Only 37 common labels are quantified since the label set in the SUN-RGBD dataset is a subset of the label set in the NYUDv2 dataset. Table 4.2 displays the quantification outcomes of the various techniques.

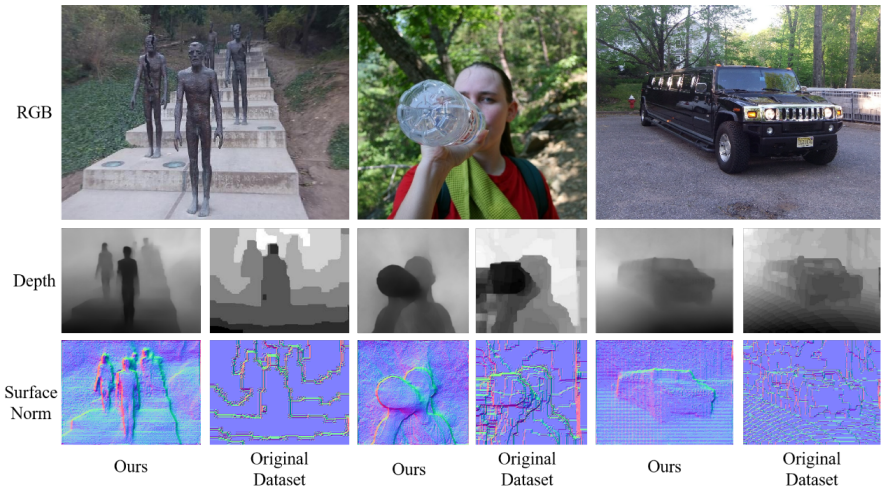

Fig. 4.1: Results of semantic segmentation using several techniques on the nyudv2 dataset. (a)RGB; (b) depth; (c)GT.

Although both NYUDv2 and SUN-RGBD datasets are derived from indoor scenes, they still have some differences in scene distribution, object categories, and labeling accuracy. Therefore, experiments on different datasets can comprehensively verify the generalization ability and robustness of the models.

Compared with the classical encoding-decoding semantic segmentation frameworks (e.g., FCN and Signet), the semantic segmentation method proposed in this study demonstrates stronger competitiveness and achieves certain performance improvement. By conducting ablation experiments on different modules, it can be found that each module contributes to the overall performance, especially the weighted Gabor direction filter and pyramid pooling feature fusion module, which have significant advantages in improving segmentation accuracy and model adaptation.

Fig. 4.2 shows the visualization results of different methods in the semantic segmentation task, and the colors represent different semantic categories. The experimental results show that the method in this study has better performance in detail portrayal, can accurately segment targets with large scale differences, and at the same time, effectively reduces the impact of environmental lighting changes and demonstrates strong scene adaptation ability. This indicates that the method has high practical value in semantic segmentation tasks in complex indoor environments.

4.3. Model complexity assessment. In order to create a relatively light network model, the approach suggested in this research takes into account both the model's complexity and the accuracy of semantic segmentation. Table4.3 compares the space complexity and minus value of different algorithm models. Considering the complexity of model space, Signet adopts pooling index for nonlinear up-sampling, and does not need to perform parameter learning in the up-sampling process, so the number of parameters is much less than that of the FCN method, but the segmentation performance of the two methods is comparable. This paper's model has a low complexity since it uses a large residual network for feature extraction, which drastically lowers the number of parameters. There are also less parameters when incorporating pyramid pooling. The Gabor directional filter also helps the network stay lightweight, which enables a more basic network to pick up complicated feature representations. When taking into account the model's temporal complexity, the network built using the conventional convolution filter has a longer inference time, whereas the network built using the broad residual network is shallower and offers some benefits during the inference process. When used in conjunction with the Gabor direction filter, this can efficiently shorten the model inference time by extracting the direction and scale features.

Taken together, this research method effectively reduces the model complexity and shortens the inference

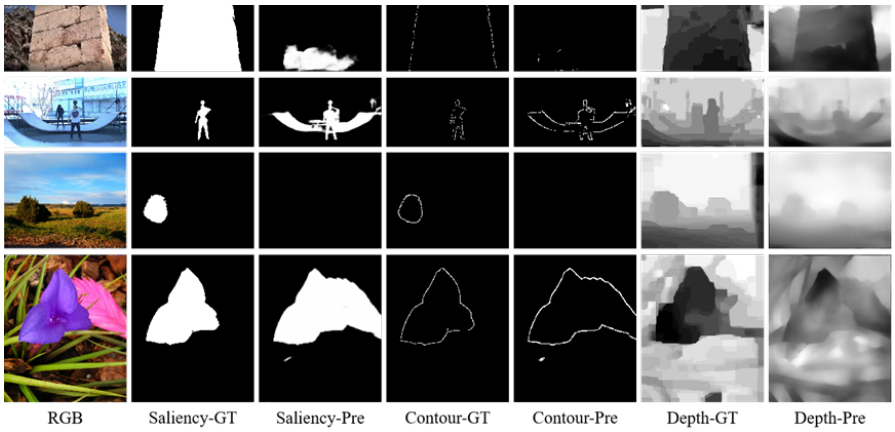

Fig. 4.2: Results of different techniques for semantic segmentation on the SUN-RGBD dataset.

Table 4.3: Comparison of the space complexity and reasoning time of several algorithms.

Method	WRN-CNN	WGCN	PP-Fusion	Model size/MB	Reasoning time/MS
Ours	.	.	.	118	43
Variant1				382	77
Variant2	.			116	36
Variant3	.			189	49
Variant4			.	246	52
FCN				548	44
Seg Net				127	59

time while ensuring high-quality semantic segmentation performance by adopting strategies such as wide residual network, large-scale residual feature extraction, Gabor direction filter and pyramid pooling, showing better lightweight characteristics and computational efficiency.

5. Conclusion. Our proposed approach addresses the challenge of color language expression in animation for film and television by introducing a semantic segmentation method for RGB-D images, leveraging a two-stream weighted Gabor convolutional network. The network first extracts image features, which are subsequently processed across multiple scales through pyramid pooling to capture both RGB and depth image features. The extracted dual-stream features undergo deep fusion at multiple scales via the pyramid pooling feature fusion module. To generate multi-scale fused features, the fused outputs are upsampled and cascaded in a decoding structure before being fed into a SoftMax classifier for final segmentation. Experimental results demonstrate that the proposed method achieves high-quality segmentation of object characteristics and edge contours, effectively adapts to variations in scale and orientation, and performs well in the semantic segmentation of indoor scenes.

Data Availability. The experimental data used to support the findings of this study are available from the corresponding author upon request.

REFERENCES

[1] Mirzakarimova, Z. D. *The acquisition of a subjective color as a result of the conversion of the meaning of the word.* ACADEMICIA An International Multidisciplinary Research Journal, 11(2),(2021) 1404-1413.

[2] Yanmin Xu, Yitao Tao, Chunjiong Zhang, Mingxing Xie, Wengang Li, Jianjiang Tai, *"Review of Digital Economy Research in China: A Framework Analysis Based on Bibliometrics"*, Computational Intelligence and Neuroscience, vol. 2022, Article ID 2427034, 11 pages, 2022.
[3] TU, Xing; WANG, DongRong; YANG, Qian. *Emotional Analysis in Animated Films Using Big Data and IoT: An In-Depth Study of'Krek'. In: Proceedings of the 2024 8th International Conference on Big Data and Internet of Things.* 2024. p. 175-182.
[4] Zeng, R. *Research on the application of computer digital animation technology in film and television.* Journal of Physics: Conference Series, 1915(3),(2021) 032047 (6pp).
[5] Xu, L. *Fast modelling algorithm for realistic three-dimensional human face for film and television animation.* Complexity, 2021(2), 1-10.
[6] Wang, S., Xu, Q., & Liu, Y. *Research on the creation of film and tv works based on virtual reality technology.* Journal of Physics Conference Series, 1744(3),(2021) 032015.
[7] Pan, Y. *Application of computer visual art in digital media art.* Journal of Physics: Conference Series, 1961(1)(2021), 012059 (6pp).
[8] Zhang, B., & Teng, Y. *A practical exploration of "ideological and political course" in film and television art education—take the "project training of 2d animation creation" as an example.* Open Journal of Social Sciences, 08(9)(2020), 229-236.
[9] Liu, X., & Pan, H. *The path of film and television animation creation using virtual reality technology under the artificial intelligence.* Scientific Programming, 2022, 1-8.
[10] Martin, A. E., & Baggio, G. *Modelling meaning composition from formalism to mechanism.* Philosophical Transactions of The Royal Society B Biological Sciences, 375(1791),(2020) 20190298.
[11] Avval, A. M., Hosseininejad, S. R., & Zahraei, S. H. *The most productive suffix - [i] of the persian language in the meaning of grammatical person and the ways of its expression in russian.* Bulletin of Udmurt University Series History and Philology, 30(3),(2020) 428-433.
[12] Stol, M. *The meaning of color in ancient mesopotamia. by shiyanthi thavapalan.* culture and history of the ancient near east 104. leiden: brill, 2020. pp. xiii + 509 + 30 plates. $163.00 (cloth). Journal of Near Eastern Studies, 80(1),(2021) 195-199.
[13] Kim, B.*A study on the content of eonmunjamo and the meaning of korean language education.* Korean Historical Linguistics, 32,(2021)33-80.
[14] Xu, L. *Fast modelling algorithm for realistic three-dimensional human face for film and television animation.* Complexity, 2021(2), 1-10.
[15] Pan, Y. *Application of computer visual art in digital media art.* Journal of Physics: Conference Series, 1961(1),(2021) 012059 (6pp).
[16] Arslan, G., & Gkearslan, A. *The use of clay animation in television advertisements.* e-Journal of New World Sciences Academy, 15(1),(2020)52-70.
[17] Sene, R. *The socio-historical factor behind change in meaning: the case of old french.* Taikomoji Kalbotyra, 15,(2021) 26-36.
[18] Cao, J. *Research on the application of color language in computer graphic design.* Journal of Physics Conference Series, 1915(4), (2021)042033.

Edited by: Ashish Bagwari
Special issue on: Adaptive AI-ML Technique for 6G/ Emerging Wireless Networks
Received: Sep 6, 2024
Accepted: Feb 25, 2025