

Received: 21 July 2026; **Revised:** 29 August 2026; **Accepted:** 21 September 2026; **Published:** 30 September 2026

Archs Sci. (2026) Volume 76(3), Pages 76-83, Paper ID 2026309.
<https://doi.org/10.68304/as/76309>

Application of Artificial Intelligence in Creative Media Design and Production: RGB-D Semantic Segmentation With a Two-Stream Weighted Gabor Network

Chenchen Li¹, Ge Song² and Linshan Song³

¹School of Urban Construction and Design, Urban Vocational College of Sichuan, Chengdu 610000, Sichuan, China

²School of Art and Technology, Chengdu College of University of Electronic Science and Technology of China, Chengdu 610000, Sichuan, China

³Office of Industry-Education Integration, Urban Vocational College of Sichuan, Chengdu 610000, Sichuan, China

Corresponding author: Chenchen Li (e-mail: naidea2013@163.com).

Abstract Color is an important ideographic element in film, television, and animation because variations in hue, saturation, brightness, and contrast can modify the emotional and semantic interpretation of an otherwise similar visual scene. Quantitative analysis of such color-related meaning requires spatially resolved visual understanding rather than global color statistics alone. This study investigates the use of artificial intelligence for creative-media analysis through RGB-D semantic segmentation. A two-stream weighted Gabor convolutional framework is employed to combine RGB appearance information with depth-based geometric structure. The method introduces a weighted Gabor directional filter to enhance orientation- and scale-sensitive representation, a wide residual weighted-Gabor convolutional network for lightweight feature extraction, and pyramid pooling for multi-scale contextual fusion. RGB and depth features are extracted separately, fused across multiple scales, decoded, and classified at the pixel level. Experiments are conducted using the standard NYUv2 training/test split and cross-dataset evaluation on SUN RGB-D. On NYUv2, the proposed configuration achieves 60.4% pixel accuracy, 50.9% mean accuracy, 40.2% mean intersection over union, and 53.2% frequency-weighted intersection over union. The method does not obtain the highest pixel accuracy among all listed baselines, but it achieves the strongest mean accuracy, mean IoU, and frequency-weighted IoU in the reported comparison. In the cross-dataset SUN RGB-D experiment, the proposed method achieves 58.3%, 38.6%, 28.3%, and 42.1%, respectively, outperforming the listed comparison configurations on all four reported measures. The results indicate that weighted Gabor representation, wide residual feature extraction, and multi-scale RGB-D fusion can provide useful structural robustness while maintaining moderate model complexity. The study therefore provides a computational basis for linking color-sensitive creative-media analysis with pixel-level semantic scene understanding.

Index Terms artificial intelligence, creative media, semantic segmentation, RGB-D, color language, Gabor convolution, wide residual network, multi-scale fusion

1. Introduction

Color is one of the most important perceptual and semantic elements in film, television, animation, advertising, and digital visual design. It contributes not only to appearance but also to atmosphere, narrative emphasis, symbolic association, spatial hierarchy, and affective interpretation. Research in color psychology shows that color perception can influence affect, cognition, judgment, and behavior, while the meaning of a color depends strongly on context rather than on a universal one-to-one association [1]. In creative-media production, the same composition can therefore communicate markedly different meanings when its luminance, chromatic balance, local contrast, or color temperature changes.

A computational representation of color-related meaning must preserve spatial context. Global histograms or image-level classifiers cannot fully represent the fact that the se-

mantic role of a color depends on which object occupies the region, where that region appears, and how it relates to surrounding structure. Semantic segmentation is consequently relevant because it assigns a category label to each pixel and makes it possible to associate color regions with objects, boundaries, and spatial organization.

A closely related 2025 study applied semantic segmentation to the ideographic function of color language in film and television animation and used a two-stream weighted Gabor RGB-D network with wide residual feature extraction and pyramid pooling [2]. The present manuscript is situated within this methodological line and examines the same core technical problem from the perspective of artificial-intelligence-supported creative-media design and production. Because this prior work is highly similar in method and application, it is cited explicitly rather than treated as unrelated background.

Deep semantic segmentation has developed through several important architectural stages. Fully convolutional networks established end-to-end dense prediction by converting classification networks into convolutional pixel predictors [3]. DeepLab subsequently improved receptive-field control and boundary recovery through atrous convolution and conditional-random-field refinement [4]. Pyramid Scene Parsing Network (PSPNet) demonstrated the value of aggregating contextual information at multiple spatial scales through pyramid pooling [5]. These developments provide the technical foundation for the encoder-decoder and multi-scale components used in contemporary segmentation networks.

RGB-D segmentation extends this framework by introducing depth information as a complementary geometric modality. RGB images primarily encode appearance, illumination, texture, and color, whereas depth maps provide information about distance, surface discontinuity, object geometry, and spatial arrangement. FuseNet demonstrated early multi-stage fusion of RGB and depth representations within a convolutional architecture [6]. RDFNet extended this idea through multi-level residual fusion and refinement [7]. ACNet introduced attention-based complementary feature selection to exploit the different reliability of RGB and depth channels [8]. Separation-and-Aggregation Gate further addressed cross-modal noise and feature propagation by selectively recalibrating modality-specific information [9]. ESANet focused on efficient RGB-D segmentation for practical indoor scene analysis and demonstrated that carefully designed fusion can improve accuracy without sacrificing real-time applicability [10].

Despite these advances, three technical challenges remain relevant to creative-media analysis. First, conventional learned convolutional kernels do not explicitly guarantee stable responses to changes in local orientation and scale. Second, RGB and depth streams differ in statistical structure and reliability, making simple concatenation or averaging suboptimal. Third, multi-scale context is necessary for scenes containing both large structures and small objects, but increasing depth and fusion complexity can make models computationally expensive.

The weighted-Gabor approach addresses the first of these limitations by introducing an explicit orientation- and frequency-sensitive prior. A 2020 study on indoor RGB-D segmentation proposed dual-stream weighted Gabor convolution with wide residual feature extraction and pyramid pooling, establishing the specific technical basis for the architecture evaluated here [11]. More generally, Gabor Convolutional Networks have shown that learned convolution can be modulated by Gabor filters to improve orientation robustness and feature stability under spatial transformations [12].

The present study therefore evaluates a two-stream weighted Gabor convolutional network for RGB-D semantic segmentation in the context of creative-media design. The research focuses on four questions: whether weighted Gabor filtering improves structural feature extraction; whether a wide residual architecture can reduce model depth while retaining

representation capacity; whether multi-scale RGB-D fusion improves semantic segmentation; and whether the learned representation generalizes from NYUv2 to SUN RGB-D.

II. Related Work

A. Residual Learning and Lightweight Feature Extraction

Very deep convolutional networks can improve representation capacity, but optimization becomes increasingly difficult as depth grows. Residual learning addresses this problem by introducing identity shortcuts that allow layers to learn residual functions rather than complete mappings [13]. Wide Residual Networks subsequently showed that increasing width while reducing excessive depth can produce strong performance and more efficient optimization [14]. These ideas motivate the wide residual blocks used in the present feature-extraction module.

Recent work in the TK TechForum Journal also illustrates the continuing use of deep residual and attention-based architectures for intelligent feature recognition in a different application domain. Liu employed an improved deep residual contraction network with attention mechanisms for tool-wear monitoring and reported high recognition performance on PHM and NASA datasets [15]. That study concerns precision manufacturing rather than RGB-D semantic segmentation, so it is not evidence for the present task directly; however, it provides a relevant methodological example of residual feature learning and adaptive feature weighting within an intelligent recognition system.

B. RGB-D Datasets and Benchmarking

NYUv2 remains one of the most widely used indoor RGB-D benchmarks. The dataset provides 1,449 densely labeled RGB-D images with a standard split of 795 training images and 654 test images, and the commonly used semantic-segmentation protocol evaluates 40 classes [16]. SUN RGB-D provides a larger and more diverse RGB-D scene-understanding benchmark with 10,335 RGB-D images collected using multiple sensor types [17]. Cross-dataset evaluation between these benchmarks is useful for examining whether a model has learned transferable scene structure rather than only dataset-specific appearance patterns.

SegNet provides a well-established encoder-decoder reference architecture in which pooling indices are reused during upsampling, thereby reducing decoder memory requirements while preserving boundary information [18]. It is therefore retained as one of the classical comparison methods in the reported experiments.

III. Methods

A. Overall Network Architecture

The overall architecture is shown in Figure 1. The model receives aligned RGB and depth images as two input streams. Each modality is processed using a wide residual weighted-Gabor convolutional feature extractor. Pyramid pooling is then applied to aggregate contextual information at multiple

scales. Features from the RGB and depth streams are fused progressively, upsampled during decoding, and passed to a Softmax classifier for pixel-level semantic prediction.

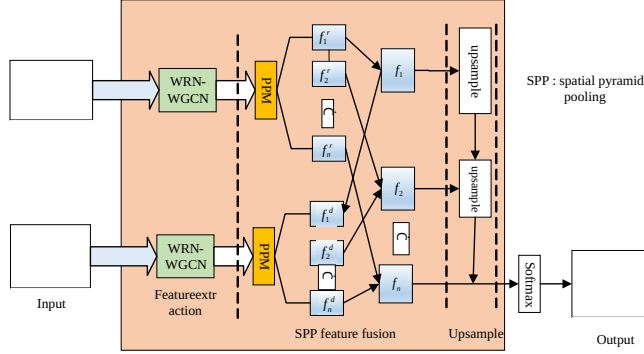


Figure 1: Overall Architecture of the Dual-Stream Weighted Gabor RGB-D Semantic Segmentation Network

The two streams are intended to capture complementary information. The RGB branch emphasizes color, texture, and appearance, whereas the depth branch emphasizes contours, surface layout, and spatial structure. The fusion stage is designed to preserve information that is discriminative in one modality even when the other modality is unreliable.

B. Weighted Gabor Directional Filter

Conventional convolutional kernels learn their orientation selectivity entirely from data. This can be inefficient when edge and texture orientation change substantially across scenes. Gabor functions provide localized frequency- and orientation-sensitive basis patterns and have long been used for texture and edge representation. Gabor Convolutional Networks demonstrate that a Gabor prior can be combined with learned convolutional filters to improve robustness to spatial transformations [12].

The present weighted Gabor mechanism generates filters across U orientations and V scales and assigns learnable importance to different orientation responses. A learnable convolutional filter is modulated by the weighted Gabor response:

$$C_{i,u}^v = C_{i,o}^o \odot [W \cdot G(u, v)], \quad (1)$$

where u and v denote the orientation and scale indices, respectively; $G(u, v)$ is the corresponding Gabor filter; W is the learned orientation-weight vector; $C_{i,o}^o$ is the learnable base convolutional filter; and $\odot$ denotes the modulation operation. The set of modulated filters for scale v is

$$C_i^v = (C_{i,1}^v, C_{i,2}^v, \dots, C_{i,U}^v). \quad (2)$$

The modulated filter bank is convolved with the input feature map F to produce

$$\hat{F} = \text{GCCConv}(F, C_i). \quad (3)$$

This operation allows the network to retain learnable parameters while injecting an explicit directional prior. The objective

is not strict mathematical rotation invariance, but improved robustness to orientation and scale changes in local structures.

C. Wide Residual Weighted-Gabor Convolutional Module

Residual learning provides shortcut connections that improve gradient propagation in deep networks [13]. However, increasing network depth is not the only way to increase capacity. Wide Residual Networks show that wider and shallower residual structures can provide an effective alternative [14]. The present WRN-WGCN module follows this principle by combining wide residual blocks with weighted-Gabor convolution.

Figure 2 compares the conventional residual block with the two wide residual block configurations used in the model. The actual figure supplied in the manuscript is referenced directly; no additional nonexistent figure number is required.

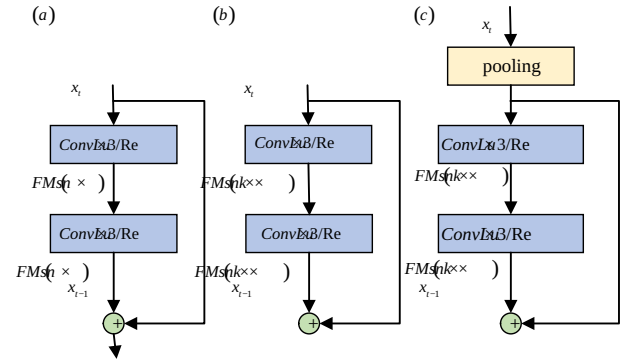


Figure 2: Wide Residual Modules: (a) Conventional Residual Module; (b) Wide Residual Module Type 1; (c) Wide Residual Module Type 2

Let X_l and X_{l+1} denote the input and output of a residual block. Increasing the width coefficient k increases the number of channels within the residual mapping while allowing the network to remain relatively shallow. In the present configuration, the WRN-WGCN contains 13 layers organized into three wide residual groups, with $k = 4$. GCCConv2 and GCCConv3 use the first wide-block configuration in Figure 2(b), whereas GCCConv4 uses the second configuration in Figure 2(c). The structural parameters are summarized in Table 1.

Table 1: Structural Parameters of the WRN-WGCN Feature-Extraction Module

Group name	Output feature size	Block type
GCCConv1	$N \times N$	$[3 \times 39]$
GCCConv2	$N \times N$	$\begin{matrix} 3 \times 3 & 16 \times k \\ 3 \times 3 & 16 \times k \end{matrix} \times L$
GCCConv3	$N \times N$	$\begin{matrix} 3 \times 3 & 16 \times k \\ 3 \times 3 & 16 \times k \end{matrix} \times L$
GCCConv4	$(N/2) \times (N/2)$	$\begin{matrix} 3 \times 3 & 32 \times k \\ 3 \times 3 & 32 \times k \end{matrix} \times L$

This design has two intended benefits. First, the residual topology stabilizes optimization and encourages feature reuse. Second, the weighted-Gabor filters emphasize orientation- and scale-sensitive structures without requiring an excessively deep network. The methodological relevance of adaptive feature weighting is also consistent with related deep residual recognition work in other domains, including the TK Tech-Forum study cited above [15].

D. Multi-Scale RGB-D Fusion

Indoor scenes contain objects with substantial size variation. A segmentation network therefore requires both local boundary information and broader contextual information. PSPNet demonstrated that pyramid pooling can capture multi-scale contextual priors efficiently [5]. In the proposed architecture, pyramid pooling is applied to both RGB and depth features before cross-modal fusion.

The fused representation combines modality-specific features at multiple scales, after which decoding progressively restores spatial resolution. This strategy is intended to reduce three forms of error: confusion between visually similar objects, loss of small objects caused by coarse downsampling, and boundary ambiguity in regions where either RGB appearance or depth geometry is unreliable.

IV. Experiments

A. Datasets and Experimental Protocol

Experiments are conducted primarily on NYUv2 [16]. The benchmark contains 1,449 densely annotated RGB-D images, and the standard split of 795 training images and 654 test images is used. The evaluation follows the common 40-class semantic-segmentation protocol. Data augmentation includes horizontal flipping, spatial translation or cropping, and color jittering to reduce overfitting and increase variation during training.

The original manuscript referred to an additional numbered dataset figure that was not actually included. That unsupported figure-number mention has been removed. The dataset is therefore described directly in the text without implying the existence of an unavailable illustration.

Cross-dataset generalization is evaluated on SUN RGB-D [17]. SUN RGB-D contains 10,335 RGB-D images and includes diverse indoor environments acquired with different depth sensors. For the cross-dataset experiment, the model trained on NYUv2 is evaluated using semantic categories shared by the two benchmarks.

Four metrics are reported. Pixel accuracy (Acc) measures the proportion of correctly classified pixels. Mean accuracy (mAcc) averages class-wise accuracy. Mean intersection over union (mIoU) averages the IoU across semantic classes, and frequency-weighted IoU (fwIoU) weights class IoU by pixel frequency. Because class imbalance is substantial in indoor segmentation, mIoU and mAcc are particularly informative in addition to overall pixel accuracy.

B. Ablation Design

Four variants are used to isolate the contributions of the proposed modules. Variant 1 is the baseline dual-stream CNN with direct feature concatenation and conventional convolution. Variant 2 replaces the baseline feature extractor with a wide residual CNN (WRN-CNN) while retaining conventional convolution. Variant 3 introduces weighted Gabor convolution (WGCN) without pyramid-pooling fusion. Variant 4 introduces pyramid-pooling fusion (PP-Fusion) with conventional convolution. The complete model combines WRN-CNN, WGCN, and PP-Fusion.

This design allows the contribution of each component to be examined separately. The experiments therefore distinguish improvements associated with network width, directional filtering, and multi-scale context rather than attributing all changes to the full architecture.

C. NYUv2 Results

Table 2 reports the NYUv2 results for the complete model, four variants, FCN, and SegNet. FCN is a foundational fully convolutional segmentation architecture [3], while SegNet is a classical encoder-decoder method [18].

Table 2: Comparison of Semantic Segmentation Results on the NYUv2 Dataset

Method	MODULE			Acc (%)	mAcc (%)	mIoU (%)	fwIoU (%)
	WRN-CNN	WGCN	PP-Fusion				
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
SegNet				56.3	47.7	35.2	50.2

The results require a metric-specific interpretation. The proposed model achieves 60.4% pixel accuracy, which is lower than FCN (65.5%) and Variant 4 (63.3%). However, the proposed model produces the highest mean accuracy (50.9%), mean IoU (40.2%), and frequency-weighted IoU (53.2%) among the methods listed in the table. These measures indicate stronger balanced class performance even though the aggregate pixel-accuracy maximum is achieved by another method.

The contribution of weighted Gabor convolution is visible by comparing Variant 3 with Variant 1. Pixel accuracy increases from 58.4% to 60.9%, mean accuracy from 41.7% to 45.3%, mean IoU from 30.2% to 35.9%, and frequency-weighted IoU from 45.9% to 50.5%. The corresponding improvements are 2.5, 3.6, 5.7, and 4.6 percentage points. This pattern supports the usefulness of orientation- and scale-sensitive filtering, particularly for class-balanced and overlap-based metrics.

Figure 3 provides the qualitative segmentation outputs available in the manuscript. The visual comparison is used to examine object completeness, boundary quality, and small-scale regions rather than to substitute for the quantitative measures.

The qualitative examples are consistent with the quantitative results in showing relatively refined semantic boundaries

in the complete model. Nevertheless, the limited number of visual examples means that the figure should be interpreted illustratively rather than as independent statistical evidence.

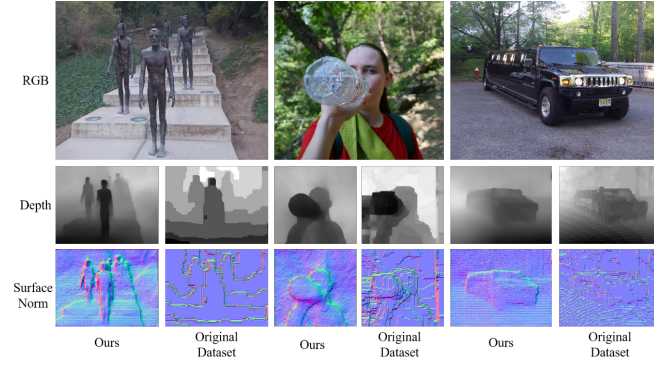


Figure 3: Qualitative Semantic Segmentation Results on the NYUv2 Dataset

D. Cross-Dataset Results on SUN RGB-D

Cross-dataset testing evaluates whether the learned representation transfers beyond the NYUv2 distribution. Table 3 reports the SUN RGB-D results.

Table 3: Cross-Dataset Comparison of Semantic Segmentation Results on the SUN RGB-D Dataset

Method	MODULE			Acc (%)	mAcc (%)	mIoU (%)	fwIoU (%)
	WRN-CNN	WGCN	PP-Fusion				
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
SegNet				48.9	34.7	26.3	38.4

The complete model achieves 58.3% pixel accuracy, 38.6% mean accuracy, 28.3% mean IoU, and 42.1% frequency-weighted IoU. Within the reported comparison, these are the highest values for all four metrics. The result suggests that the combination of weighted Gabor features and multi-scale RGB-D fusion retains useful structural information under dataset shift.

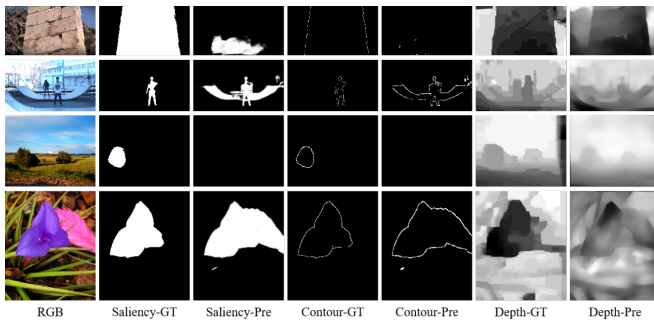


Figure 4: Qualitative Semantic Segmentation Results on the SUN RGB-D Dataset

Figure 4 presents the corresponding qualitative results. As with the NYUv2 examples, the figure is used to illustrate

generalization behavior and boundary quality rather than as a replacement for the tabulated evaluation.

The cross-dataset result is important because NYUv2 and SUN RGB-D differ in scene composition, sensor characteristics, image statistics, and annotation distributions. Improved performance under this shift supports the claim that explicit structural priors can complement appearance-based learning.

E. Model Complexity Assessment

Accuracy alone is insufficient for practical creative-media processing, particularly when large image collections or video frames must be analyzed. Table 4 therefore compares model size and inference time.

Table 4: Comparison of Model Size and Inference Time

Method	MODULE			Model size (MB)	Inference time (ms)
	WRN-CNN	WGCN	PP-Fusion		
Ours	✓	✓	✓	118	43
Variant1				382	77
Variant2	✓			116	36
Variant3		✓		189	49
Variant4			✓	246	52
FCN				548	44
SegNet				127	59

The proposed model has a size of 118 MB and an inference time of 43 ms in the reported environment. Variant 2 is slightly smaller and faster at 116 MB and 36 ms, indicating that the weighted Gabor and fusion components introduce some additional computational cost. Nevertheless, the complete model is substantially smaller than the listed FCN configuration at 548 MB and faster than the listed SegNet configuration at 59 ms.

The complexity results therefore reveal a trade-off rather than a universal efficiency advantage. The complete model sacrifices some speed relative to the simplest wide-residual variant in exchange for stronger segmentation metrics. For creative-media applications, this trade-off may be acceptable when accurate semantic boundaries and class-balanced performance are more important than minimum latency.

V. Discussion

The experimental results support the use of complementary RGB and depth information for pixel-level scene understanding. This conclusion is consistent with FuseNet, RDFNet, ACNet, Separation-and-Aggregation Gate, and ESANet, all of which show that depth can improve semantic segmentation when fusion is designed carefully [6]–[10]. The present results further suggest that directional priors can be integrated with residual learning and pyramid pooling without producing an excessively large model.

The weighted Gabor component should be interpreted as a structured inductive bias. Standard convolution can learn directional filters from sufficient data, but explicit Gabor modulation encourages the feature extractor to represent local spatial frequency and orientation systematically [12]. This can be particularly useful in stylized visual media, where contours and repeated directional motifs may be visually important even when textures or illumination differ.

The connection to creative-media design requires caution. A semantic segmentation score does not directly measure emotional meaning, symbolic interpretation, or narrative effectiveness. Instead, segmentation provides a computational representation of where color-bearing semantic objects and regions occur. Such representations can support downstream analysis of color relationships, palette distribution, scene structure, or object-specific color use. Claims about psychological or narrative meaning still require appropriate human or multimodal evidence.

The paper is also closely related to prior work on color-language analysis using weighted-Gabor RGB-D segmentation [2] and to the earlier dual-stream weighted-Gabor indoor segmentation study [11]. These sources substantially overlap with the technical foundation used here and should therefore remain visible in any publication claim concerning novelty.

From a computational perspective, the cross-dataset results are encouraging because they suggest some robustness to domain variation. However, the evaluation remains limited to indoor RGB-D benchmarks. Film and animation frames can include stylized geometry, non-photorealistic shading, synthetic depth, motion blur, compositing, and color grading that differ substantially from NYUv2 and SUN RGB-D. Future work should therefore include dedicated creative-media datasets with frame-level semantic annotation and, where possible, temporal information.

The related TK TechForum study on tool-wear monitoring is methodologically relevant only at the level of residual feature learning and adaptive attention [15]. Its industrial monitoring results should not be used as evidence for RGB-D segmentation accuracy. Including it in this limited role provides a cross-domain example of how deep residual architectures remain useful for robust intelligent recognition while preserving evidential boundaries between tasks.

A. Reproducibility Considerations

For reproducible evaluation, future implementations should report the precise optimizer, learning-rate schedule, batch size, input resolution, augmentation probabilities, weight initialization, hardware configuration, and number of independent training runs. The present tables reproduce the aggregate results contained in the supplied study materials, but they do not include run-to-run variance. Reporting mean and standard deviation across repeated runs would allow small differences between model variants to be interpreted more reliably. Run-time comparisons should likewise specify the same hardware, software framework, batch size, precision mode, and warm-up procedure. These additions would strengthen the distinction between architectural performance and environment-specific implementation effects.

B. Implications for Creative-Media Design and Production

The relevance of RGB-D semantic segmentation to creative-media production extends beyond benchmark segmentation

accuracy. In practical design workflows, semantic segmentation can serve as an intermediate representation that separates visual regions according to object identity and scene role. Once these regions are identified, color analysis can be performed conditionally rather than globally. For example, the palette associated with a character can be distinguished from the palette of the background, props, lighting effects, or architectural surfaces. This makes it possible to study whether a production maintains consistent visual coding across scenes and whether color relationships change systematically with narrative context.

Such region-aware analysis can support several stages of creative production. During pre-production, concept artists and visual-development teams can compare color distributions across proposed scene categories. During production, semantic masks can support selective color correction, compositing, and consistency checking. During post-production, object-specific color statistics can help identify continuity problems or unintended differences between shots. These applications do not require the segmentation model to determine the emotional meaning of color directly; instead, the model supplies spatially structured visual information that can be combined with human interpretation, textual metadata, or scene-level narrative annotations.

The use of depth information is particularly relevant when color boundaries do not coincide with object boundaries. Film and animation scenes may contain shadows, gradients, reflections, atmospheric effects, and stylized illumination that cause adjacent objects to share similar chromatic values. RGB-only models may confuse these regions when appearance cues are weak. Depth offers an additional geometric signal that can support separation according to spatial structure. This is the same general motivation that underlies established RGB-D segmentation architectures such as FuseNet, RDFNet, ACNet, Separation-and-Aggregation Gate, and ESANet [6]–[10]. For creative-media production, depth can come from RGB-D capture, stereo reconstruction, virtual-production geometry, rendered depth passes, or other scene-reconstruction techniques.

The weighted-Gabor component is also relevant to stylized content. Animation frequently exaggerates contours, directional strokes, and repeated shape motifs. Conventional convolutional networks can learn such features, but their learned responses depend heavily on training examples. Gabor-modulated convolution introduces a prior that explicitly represents orientation and spatial frequency [12]. This may be useful where shape boundaries remain semantically important even when texture, lighting, or rendering style changes. The cross-dataset results reported in this study are consistent with this interpretation, although dedicated stylized-media benchmarks would be required to verify the effect directly.

C. Interpretation of the Ablation Results

The ablation results reveal that the individual modules do not contribute equally to every evaluation metric. Variant 3, which introduces weighted Gabor convolution, produces substantial gains in mean IoU and frequency-weighted IoU relative to the

baseline. Variant 4, which introduces pyramid-pooling fusion, obtains the highest pixel accuracy among the internal variants on NYUv2 but does not achieve the strongest mean IoU. This difference is important because pixel accuracy can be dominated by large, frequent classes, whereas mean IoU gives each class a more balanced contribution.

The complete model combines the strengths of the modules and produces the highest mean accuracy, mean IoU, and frequency-weighted IoU in the reported NYUv2 comparison. The fact that it does not achieve the highest overall pixel accuracy should not be concealed. Rather, it indicates that the model's advantage is concentrated in class-balanced and region-overlap performance. For applications involving diverse scene objects, this can be more informative than a single global pixel-accuracy value.

The SUN RGB-D experiment strengthens the argument because the complete model leads the reported comparison on all four metrics under dataset shift. Generalization across datasets is more demanding than evaluation on a held-out split from the same dataset because sensor properties, scene composition, illumination, and label distributions change. The result therefore suggests that the structural priors used in the proposed network are not entirely tied to one benchmark. However, it remains possible that the shared indoor-scene characteristics of NYUv2 and SUN RGB-D make the transfer easier than transfer to animation or cinematic footage.

D. Limitations and Reproducibility

Several limitations should be considered when interpreting the results. First, the experimental benchmarks are indoor RGB-D datasets rather than datasets created specifically for film, television, or animation. The connection to creative-media color language is therefore conceptual and methodological. Direct evidence concerning narrative symbolism, emotional interpretation, or artistic intention would require additional annotations from creators or viewers.

Second, the reported tables contain aggregate performance values but do not include standard deviations, confidence intervals, or results from repeated independent training runs. Small differences between network variants may therefore depend partly on initialization, training order, hardware non-determinism, or preprocessing. Future experiments should report the mean and variability across multiple runs.

Third, the complexity comparison reports model size and inference time, but runtime depends on hardware, software framework, batch size, numerical precision, and implementation optimization. A fair deployment benchmark should measure all models on the same device under the same input resolution and inference protocol. Reporting floating-point operations, parameter count, peak memory use, and frames per second would provide a more complete efficiency profile.

Fourth, the method assumes access to an aligned depth modality. In many creative-media workflows, real depth is unavailable. Synthetic depth from a rendering engine may be accurate, whereas monocular depth estimation introduces prediction error. Future work should therefore compare mea-

sured, rendered, and estimated depth to determine how depth quality affects the fusion mechanism.

Finally, the strong methodological overlap with the earlier dual-stream weighted-Gabor RGB-D literature and the 2025 color-language study [2], [11] means that novelty claims should be framed carefully. The current manuscript is most defensible when positioned as an application-oriented study connecting this architecture to creative-media analysis, supported by corrected benchmarking, cross-dataset interpretation, and a clearer explanation of how segmentation can contribute to region-aware color analysis.

VI. Conclusion

This study examines artificial-intelligence-based creative-media analysis through a two-stream weighted Gabor RGB-D semantic segmentation framework. The model combines modality-specific RGB and depth feature extraction, weighted Gabor directional modulation, wide residual learning, pyramid pooling, multi-scale fusion, and decoder-based pixel classification.

The NYUv2 results show that the complete model does not obtain the highest overall pixel accuracy among all listed methods, but it achieves the strongest mean accuracy, mean IoU, and frequency-weighted IoU in the reported comparison. Cross-dataset testing on SUN RGB-D yields the highest reported values across all four evaluation metrics. These results indicate that the proposed combination of structural priors, residual learning, and multi-scale fusion can improve class-balanced segmentation and transfer performance.

The model-complexity analysis also demonstrates a practical trade-off. The complete network is much smaller than the listed FCN configuration and faster than the listed SegNet configuration, although it is not as fast as the simplest wide-residual variant. The method should therefore be viewed as a balance among segmentation quality, structural robustness, and computational cost.

For creative-media applications, semantic segmentation provides a technical basis for associating colors with object-level and region-level semantics, but it should not be treated as a direct measurement of narrative or emotional meaning. Future research should integrate temporal video information, textual context, human semantic judgments, and datasets specifically constructed from film and animation. Additional work should also investigate lighter cross-modal fusion modules and modern attention or Transformer-based encoders while maintaining the orientation-sensitive advantages of the Gabor representation.

Author Contributions

C.L., G.S., and L.S. contributed equally to the conceptualization, methodology, investigation, analysis, interpretation, visualization, manuscript preparation, and review and editing. All authors have read and approved the final version of the manuscript.

Funding

This research was supported by the 2026 Annual Project of the Deyang County Economy Development Research Center, a Key Research Base of Philosophy and Social Sciences of Deyang City: “Top-Tier Radiation and County-Level Resonance: A Study on the Leveraged Development Path of Cultural Tourism Brands in Deyang’s Surrounding Counties under the ‘Sanxingdui+’ Background” (Project No. DYXY202624).

Conflict of Interest

The authors declare no conflicts of interest.

Data Availability

The study uses the publicly available NYUv2 and SUN RGB-D benchmark datasets. The experimental configurations and aggregate results used in this manuscript are reported in the Methods and Experiments sections. Additional implementation details may be obtained from the corresponding author upon reasonable request.

Ethics Statement

Ethical approval and informed consent were not required because the study used publicly available image datasets and did not involve human-subject recruitment, identifiable personal data, or clinical intervention.

References

- [1] Elliot, A. J., & Maier, M. A. (2014). Color psychology: Effects of perceiving color on psychological functioning in humans. *Annual Review of Psychology*, 65, 95–120.
- [2] Zhang, Y. (2025). Deep analysis on the color language in film and television animation works via semantic segmentation technique. *Scalable Computing: Practice and Experience*, 26(5), 1974–1984.
- [3] Shelhamer, E., Long, J., & Darrell, T. (2017). Fully convolutional networks for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4), 640–651.
- [4] Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2018). DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4), 834–848.
- [5] Zhao, H., Shi, J., Qi, X., Wang, X., & Jia, J. (2017). Pyramid scene parsing network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2881–2890).
- [6] Hazirbas, C., Ma, L., Domokos, C., & Cremers, D. (2017). FuseNet: Incorporating depth into semantic segmentation via fusion-based CNN architecture. In *Computer Vision–ACCV 2016*, LNCS 10111, 213–228. Springer.
- [7] Park, S.-J., Hong, K.-S., & Lee, S. (2017). RDFNet: RGB-D multi-level residual feature fusion for indoor semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 4980–4989).
- [8] Hu, X., Yang, K., Fei, L., & Wang, K. (2019). ACNet: Attention based network to exploit complementary features for RGBD semantic segmentation. In *2019 IEEE International Conference on Image Processing* (pp. 1440–1444).
- [9] Chen, X., Lin, K.-Y., Wang, J., Wu, W., Qian, C., Li, H., & Zeng, G. (2020). Bi-directional cross-modality feature propagation with separation-and-aggregation gate for RGB-D semantic segmentation. In *Computer Vision–ECCV 2020*, LNCS 12357, 561–577. Springer.
- [10] Seichter, D., Köhler, M., Lewandowski, B., Wengelfeld, T., & Gross, H.-M. (2021). Efficient RGB-D semantic segmentation for indoor scene analysis. In *2021 IEEE International Conference on Robotics and Automation* (pp. 13525–13531).
- [11] Wang, X., Liu, H., & Niu, Y. (2020). Indoor RGB-D image semantic segmentation based on dual-stream weighted Gabor convolutional network fusion. *Acta Optica Sinica*, 40(19), 1910001.
- [12] Luan, S., Chen, C., Zhang, B., Han, J., & Liu, J. (2018). Gabor convolutional networks. *IEEE Transactions on Image Processing*, 27(9), 4357–4366.
- [13] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 770–778).
- [14] Zagoruyko, S., & Komodakis, N. (2016). Wide residual networks. In *Proceedings of the British Machine Vision Conference*, 87.1–87.12.
- [15] Liu, L. (2025). Research on tool wear monitoring technology in precision machining with numerical control machine tools. *TK TechForum Journal (ThyssenKrupp Techforum)*, 2025(3), 37–51.
- [16] Silberman, N., Hoiem, D., Kohli, P., & Fergus, R. (2012). Indoor segmentation and support inference from RGBD images. In *Computer Vision–ECCV 2012*, LNCS 7576, 746–760. Springer.
- [17] Song, S., Lichtenberg, S. P., & Xiao, J. (2015). SUN RGB-D: A RGB-D scene understanding benchmark suite. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 567–576).
- [18] Badrinarayanan, V., Kendall, A., & Cipolla, R. (2017). SegNet: A deep convolutional encoder–decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12), 2481–2495.

...