SDGMamba: Stage-Aware Mamba with Deformable Detail and Gated Fusion for Clothing Parsing

Haitao Fu, Shaoyu Wang(✉), Zhongyuan Teng, Jiaxin He, Hong Wan, and Xiujin Shi

School of Information and Intelligent Sciences, Donghua University, Shanghai, China
sywang@dhu.edu.cn

Abstract. Clothing parsing is a pivotal subtask of semantic segmentation with significant research and practical implications. However, it faces persistent challenges, including substantial scale variations among clothing components, posture-induced occlusions, and high inter-class visual similarity, which necessitate models capable of synergizing global semantics with local details. While CNN-based models are restricted by narrow receptive fields, Transformer-based architectures—though adept at capturing long-range dependencies—often struggle with uniform input processing that weakens fine-grained details. To mitigate these issues, we introduce SDGMamba, an innovative clothing parsing framework which effectively exploits scale-dependent features for global modeling while preserving intricate characteristics. First, we design the Stage-Aware Attention VSS (SASS) Block, which dynamically allocates attention based on network depth to facilitate hierarchical structure-aware adjustments. Second, we introduce the Deformable Detail Retention (DDR) module, which employs deformable convolutions to dynamically align and aggregate multi-scale depth cues, thus strengthening texture representations. Finally, the Gated Cross-Scale Fusion (GCSF) module employs a gating mechanism to refine shallow features guided by high-level semantics, strengthening semantic coherence among components. Experimental results on the CFPD and ATR datasets demonstrate that SDGMamba attains 94.48% PA / 58.57% mIoU and 95.60% PA / 62.73% mIoU, respectively, outperforming CNN-, Transformer-, and Mamba-based counterparts.

Keywords: Clothing Parsing, Visual State Space Model, Stage-Aware Attention, Deformable Detail, Gated Fusion.

1 Introduction

Clothing parsing, a fine-grained subtask of semantic segmentation, aims to decompose clothing images into distinct semantic regions via pixel-level classification, accurately identifying components ranging from primary garments to intricate accessories. This technology not only serves as a cornerstone for emerging applications such as virtual try-on [1] and fashion clothes matching [2], but also functions as a vital instrument for designers to analyze consumer preferences, possessing substantial research and commercial value. However, as illustrated in Fig. 1, the task confronts several domain-specific challenges, including significant scale variations among clothing items, occlusions

and deformations caused by complex human poses, and high inter-class visual similarity. These factors severely exacerbate the difficulty of discriminating fine-grained features, rendering clothing parsing a formidable challenge in computer vision.

Fig. 1. Some images in the CFPD dataset, illustrating challenges like scale variations and complex poses.

With the emergence of Fully Convolutional Networks (FCN) [3], clothing parsing has witnessed substantial progress. However, standard convolutional layers suffer from local operations, leading to inadequate contextual modeling. Although DeepLabv3+ [4] incorporates context-aware mechanisms, it fails to fully exploit intra-class correlations and inter-class differences. Furthermore, Vision Transformers (ViTs) [5] enable global modeling but typically treat image patches uniformly. This process risks diluting discriminative local details with background interference, hindering fine-grained identification. Hence, it becomes essential to devise an architecture which seamlessly unifies global semantic context with local structural details.

In this paper, we propose SDGMamba (Stage-Aware Mamba with Deformable Detail and Gated Fusion), a novel framework designed to bolster the joint representation of global semantics and local details in clothing images. By explicitly addressing the multi-scale characteristics of clothing, our approach significantly improves robustness to deformations, preserves fine-grained details, and deepens semantic understanding. Experimental results demonstrate the competitive performance of SDGMamba on clothing parsing benchmarks.

2 Related Work

2.1 Clothing Parsing

Clothing parsing, a pivotal component of fashion analysis, aims to perform pixel-level semantic segmentation of garment images. Early methodologies predominantly utilized probabilistic graphical models, specifically MRF and CRF, incorporating hand-crafted

features and human pose priors for categorization. For instance, Hasan et al. [6] introduced an MRF-based method for segmenting instance components, while Yamaguchi et al. [7] established the Fashionista dataset to facilitate pose-aided parsing. However, these conventional approaches relied heavily on labor-intensive feature engineering. With the rise of deep learning, FCN [3] and its variants became dominant. Yoo et al. [8] designed a Context-aware Encoder for improving accuracy, while He et al. [9] incorporated a color attention component in TANet to leverage color consistency. Nevertheless, the inherent locality of convolutional operations constrains global context modeling. Recently, Transformer architectures have shown considerable promise in visual tasks by modeling long-range interactions through self-attention mechanisms. Although Tang et al. [10] integrated this mechanism with feature projection to achieve promising results, the feature dilution arising from the uniform processing of image patches remains a critical bottleneck for fine-grained clothing parsing.

2.2 Mamba Based Method

State Space Models (SSMs) are distinguished by their linear computational complexity with respect to sequence length. Early advancements, such as S4 [11], mitigated computational and memory bottlenecks by imposing diagonal structures on state matrices. Building on this, S5 [12] enhanced performance by integrating Multi-Input Multi-Output (MIMO) SSMs with parallel scanning mechanisms. Mamba [13] further advanced the field by proposing a hardware-conscious selective scanning strategy, which leverages input-dependent dynamics to filter information. This architecture has demonstrated the capability to outperform Transformers in natural language processing while significantly reducing parameter counts and computational overhead. In the visual domain, adaptations like VMamba [14] have bridged the gap by devising scanning strategies tailored for non-causal image data. Subsequent works, such as MSVMamba [15], have incorporated multi-scale feature fusion to optimize performance in tasks including classification and segmentation. However, existing research primarily targets general scenes, and the potential of Mamba architectures for fine-grained tasks, particularly clothing parsing, remains largely unexplored.

3 Our Method

3.1 Overall Architecture

As depicted in Fig. 2(a), SDGMamba employs an encoder-decoder design. Taking an image $I \in \mathbb{R}^{3\times H\times W}$ as input, a patch embedding layer projects it into feature tokens $X \in \mathbb{R}^{C\times H/4\times W/4}$ prior to the encoder. The encoder is organized into four stages, each comprising downsampling layers and stacked Stage-Aware Attention VSS (SASS) Blocks, with the SASS block counts configured as [2, 2, 5, 2] respectively. Specifically, the SASS Block adaptively modulates attention at each stage, identifying fine-grained structures while maintaining consistency between stripe directions and global semantics. Concurrently, Deformable Detail Retention (DDR) modules are integrated across stages 1-4. These modules dynamically adjust sampling positions to align with local

deformations, while simultaneously employing parallel multi-kernel depth-wise convolutions to preserve medium-to-small-scale details. Furthermore, Gated Cross-Scale Fusion (GCSF) modules in stages 1-3 facilitate precise, spatially-aware fusion across high-resolution details and low-resolution semantic features. Finally, the standard UPerNet [16] serves as the decoder to generate the final segmentation masks.

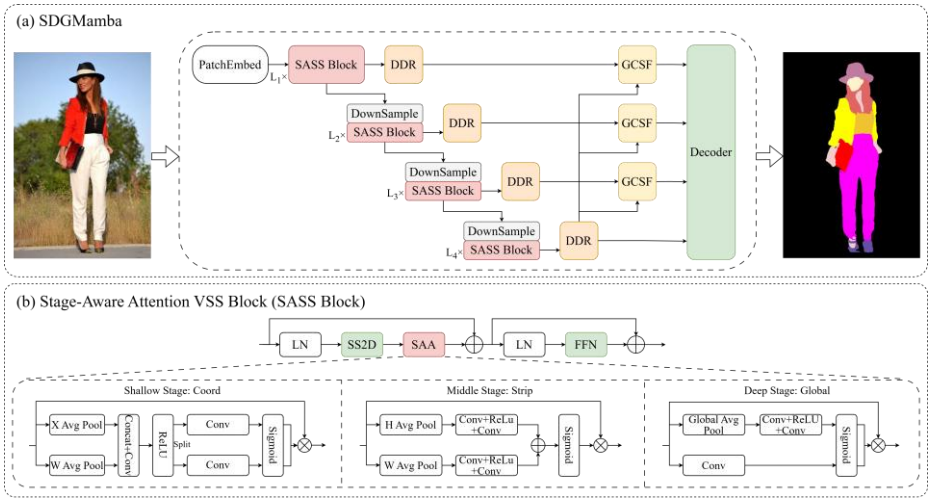

Fig. 2. (a) Overview of the SDGMamba architecture. (b) Detailed structure of the Stage-Aware Attention VSS (SASS) Block.

3.2 Stage-Aware Attention VSS Block

Serving as the core component within the VMamba framework, the VSS Block leverages a selective scan mechanism for efficient visual feature extraction. However, when applied to the fine-grained task of clothing parsing, its homogenous feature processing mechanism exposes critical limitations. Specifically, the SS2D module employs a channel-agnostic scanning strategy, causing subtle semantic cues—such as those defining intricate accessories—to be overshadowed by background noise. Moreover, its fixed spatial interaction mechanism struggles to accommodate the diverse morphological and structural variations inherent in clothing components, impeding the precise delineation of structured regions like collars and cuffs. To overcome these obstacles, we propose the Stage-Aware Attention VSS (SASS) Block, which employs differentiated attention mechanisms to perceive clothing structures, thereby sustaining efficient global modeling capabilities while simultaneously empowering the model to dynamically capture and selectively enhance clothing structural features.

The structure of the SASS Block is shown in Fig. 2(b). Input features $x \in \mathbb{R}^{C\times H\times W}$ first undergo Layer Normalization (LN) and the SS2D transformation to extract long-range dependencies, yielding the intermediate representation u. The formulation is defined as follows:

$$u = \mathrm{SS2D}(\mathrm{LN}(x)) \tag{1}$$

Subsequently, the features are enhanced by applying Stage-Aware Attention (SAA) for dynamic refinement and gated fusion, as follows:

$$u' = u + \gamma \cdot \sigma(g) \odot \mathrm{A}(u) \tag{2}$$

The refined features are subsequently combined with the initial input x through a skip connection to produce the merged features v. formulated as:

$$v = x + \mathrm{DropPath}(u') \tag{3}$$

Finally, v undergoes normalization once more and is transformed through a Feed-Forward Network (FFN), culminating in the final output y through a residual connection. This formulation is expressed as follows:

$$y = v + \mathrm{DropPath}(\mathrm{FFN}(\mathrm{LN}(v))) \tag{4}$$

Here, $\mathrm{A}(\cdot)$ denotes the Stage-Aware Attention (SAA) operator, which instantiates as Coord, Strip, or Global attention variants depending on the stage. $\gamma \in \mathbb{R}$ represents the learnable residual scaling factor, $g \in \mathbb{R}^{C\times 1\times 1}$ denotes the learnable channel gate, σ denotes the Sigmoid function, and $\odot$ signifies element-wise multiplication.

The SAA mechanism dynamically allocates attention operators according to network depth. In the shallow stages (Stages 1-2), Coord Attention [17] is employed. This mechanism encodes spatial coordinate information via axial pooling to enhance sensitivity to the orientation of clothing edges. Given an input tensor $z \in \mathbb{R}^{C\times H\times W}$, direction-aware feature maps z_h and z_w are generated by aggregating features across the height together with width spatial axes, individually. The formulation is given as follows:

$$\begin{aligned} z_h &= \mathrm{AvgPool}_{(1,W)}(z) \\ z_w &= \mathrm{AvgPool}_{(H,1)}(z) \end{aligned} \tag{5}$$

Subsequently, the two feature representations are concatenated along the spatial axis, followed by a shared convolution and non-linear activation. The resulting tensor is then split along the spatial dimension to yield the coordinate-specific vectors z_h' and z_w'. This process is expressed as follows:

$$[z_h', z_w'] = \mathrm{Split}(\mathrm{ReLU}(\mathrm{Conv}_{1\times 1}([z_h; (z_w)^{\intercal}])), [H, W]) \tag{6}$$

Here, $[\cdot;\cdot]$ represents the concatenation operation, while $\mathrm{Split}(\cdot, [H, W])$ indicates splitting the tensor along the spatial dimension into heights and widths. Each vector undergoes independent convolutional transformation to yield attention maps. Finally, these maps are passed through a Sigmoid function and imposed on the original features via element-wise multiplication to perform re-weighting, as follows:

$$\mathrm{Coord}(z) = z \odot \sigma(\mathrm{Conv}_{1\times 1}^{h}(z_h')) \odot \sigma(\mathrm{Conv}_{1\times 1}^{w}((z_w')^{\intercal})) \tag{7}$$

In the intermediate stage (Stage 3), Strip Attention [18] is utilized to perform strip-wise aggregation along both the height and width axes, thereby capturing long-range structural dependencies. This mechanism is particularly effective for modeling elongated structures, such as pleats and sleeves. With the input tensor $z \in \mathbb{R}^{C\times H\times W}$, the

axial features z_h and z_w are computed as defined in Eq. (5). Subsequently, both feature vectors undergo independent sequential convolutional transformations and non-linear activations. They are then broadcast back to the original spatial dimensions to yield the horizontal and vertical attention weights a_h and a_w, as follows:

$$\begin{aligned} a_h &= \mathrm{Conv}_{1\times1}^{(h,2)}(\mathrm{ReLU}(\mathrm{Conv}_{1\times1}^{(h,1)}(z_h))) \\ a_w &= \mathrm{Conv}_{1\times1}^{(w,2)}(\mathrm{ReLU}(\mathrm{Conv}_{1\times1}^{(w,1)}(z_w))) \end{aligned} \tag{8}$$

The attention maps from both directions are aggregated via summation and processed by a Sigmoid function to produce a unified weight map capable of highlighting long-range dependencies. This map is subsequently used to re-weight the original features in an element-wise manner, as illustrated below:

$$\mathrm{Strip}(z) = z \odot \sigma(a_h + a_w) \tag{9}$$

In the deep stage (Stage 4), Global Attention is deployed to synergize global channel context with spatial gating mechanisms, enabling high-level semantic verification. For an input tensor $z \in \mathbb{R}^{C\times H\times W}$, global average pooling is first performed to compress spatial dimensions into a global channel descriptor, as follows:

$$f = \mathrm{AvgPool}_{(H,W)}(z) \tag{10}$$

Subsequently, the channel pathway computes a channel focus descriptor a_c via two layers of 1×1 convolutions and activations to model global dependencies. Concurrently, the spatial attention branch produces the spatial gating map a_s through a 1×1 convolutional projection, facilitating pixel-level semantic refinement, as follows:

$$\begin{aligned} a_c &= \sigma(\mathrm{Conv}_{1\times1}^{(2)}(\mathrm{ReLU}(\mathrm{Conv}_{1\times1}^{(1)}(f)))) \\ a_s &= \sigma(\mathrm{Conv}_{1\times1}^{s}(z)) \end{aligned} \tag{11}$$

Finally, the two attention maps are combined via element-wise multiplication and imposed on the input representations to perform selective modulation of higher-order semantics. The formulation is defined as follows:

$$\mathrm{Global}(z) = z \odot a_c \odot a_s \tag{12}$$

3.3 Deformable Detail Retention

The Deformable Detail Retention (DDR) module is explicitly designed to bolster robustness against non-rigid deformations while effectively preserving fine-grained garment details. It achieves this objective by synergizing adaptive deformable receptive fields with multi-scale depth-wise feature aggregation. The structure of the DDR component is shown in Fig. 3.

Given input features $z \in \mathbb{R}^{C\times H\times W}$, a dynamic transformation is initiated via offset-guided convolutions. Specifically, a learned offset map Δ guides adaptive deformable convolution sampling from the regular grid $\mathcal{R} = \{(-1,-1),(-1,0),\dots,(0,1),(1,1)\}$.

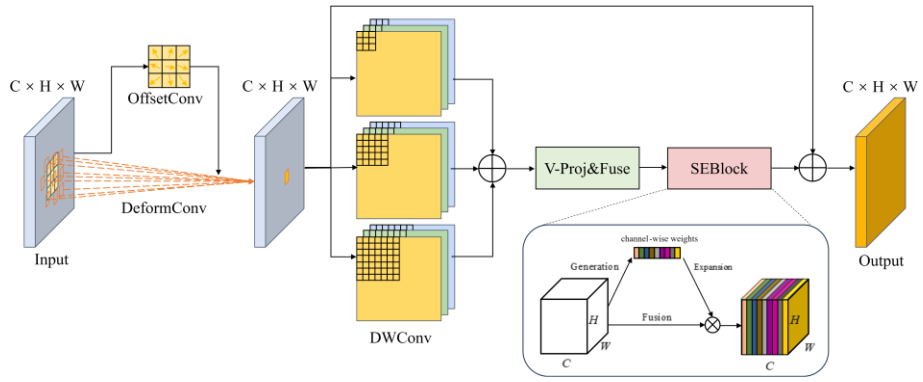

Fig. 3. Architecture of the Deformable Detail Retention (DDR) module.

Batch Normalization (BN) together with a non-linear activation is subsequently employed to stabilize training. This process is formulated as follows:

$$\begin{aligned} \Delta &= \mathrm{Conv}_{\mathrm{offset}}(z) \\ f &= \mathrm{ReLU}(\mathrm{BN}(\mathrm{DeformConv}(z, \Delta))) \end{aligned} \tag{13}$$

To capture small-to-medium scale details, the aligned feature f undergoes further multi-scale refinement. First, the channel dimension is expanded via a pointwise convolution. parallel multi-scale depth-wise operators with kernels of 3×3, 5×5, and 7×7 are leveraged to extract structural details over a range of receptive field sizes. The outputs from these branches are aggregated via summation. To synthesize these multi-scale features, the Value Projection and Fusion block, denoted as Φ, is applied to prepare for channel recalibration:

$$z_{\mathrm{val}} = \Phi\left(\sum_{k\in\{3,5,7\}} \mathrm{DWConv}_k\left(\mathrm{Conv}_{1\times1}(f)\right)\right) \tag{14}$$

Here, $\Phi(\cdot)$ denotes the Conv-BN-ReLU sequence. Following this, a Squeeze-and-Excitation (SE) block [19] employs global average pooling for extracting per-channel statistics and producing scaling weights via a pair of fully connected layers, which are subsequently used to recalibrate the feature representations. Finally, a channel projection is carried out through a 1×1 convolution, whose output is merged with the local representation f through a skip connection. The final output o is obtained as follows:

$$o = \mathrm{ReLU}(\mathrm{BN}(\mathrm{Conv}_{1\times1}(z_{\mathrm{val}} \odot \mathrm{SE}(z_{\mathrm{val}})) + f)) \tag{15}$$

3.4 Gated Cross-Scale Fusion

The Gated Cross-Scale Fusion (GCSF) module is designed to bridge the gap between high-resolution garment details and low-resolution semantic context. It dynamically filters and integrates cross-level information via a gating mechanism, the detailed architecture of which is illustrated in Fig. 4.

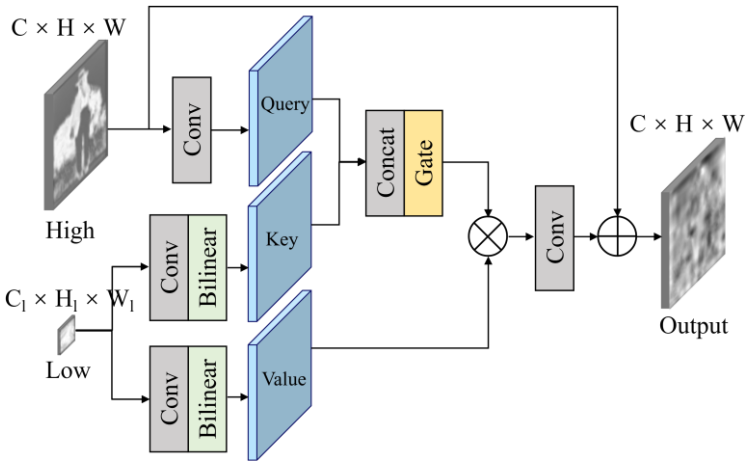

Fig. 4. Architecture of the Gated Cross-Scale Fusion (GCSF) module.

Given high-resolution features $z_{high} \in \mathbb{R}^{C_h \times H \times W}$ and coarse-grained features $z_{low} \in \mathbb{R}^{C_l \times H_l \times W_l}$, both are first projected into a unified latent space. The query branch (q) emerges from the high-resolution feature maps, whereas the key (k) and value (v) branches are built upon the coarse-grained representations, facilitating cross-scale semantic alignment. To resolve spatial discrepancies, the low-resolution feature maps are upsampled to the high-resolution space via bilinear interpolation, ensuring spatial consistency for subsequent operations:

$$\begin{aligned} q &= \mathrm{Conv}_{1\times1}^{high}(z_{high}) \\ k &= \mathrm{Bilinear}(\mathrm{Conv}_{1\times1}^{low}(z_{low}), (H, W)) \\ v &= \mathrm{Bilinear}(\mathrm{Conv}_{1\times1}^{value}(z_{low}), (H, W)) \end{aligned} \tag{16}$$

Here, (H, W) denotes the target spatial resolution. The query is merged with the upsampled key across the channel axis. Subsequently, a 3×3 convolution extracts local fusion patterns, then Batch Normalization, ReLU activation, plus a 1×1 compression layer. Finally, a Sigmoid function generates a spatial gating map to model adaptive cross-scale attention, as follows:

$$gate = \sigma(\mathrm{Conv}_{1\times1}(\mathrm{ReLU}(\mathrm{BN}(\mathrm{Conv}_{3\times3}([q; k]))))) \tag{17}$$

This gating signal applies element-wise weighting to the value features, enabling adaptive filtering that effectively suppresses irrelevant contextual noise. The modulated features are then projected back into the high-resolution channel space and fused with the original features via a residual connection. This approach preserves fine-grained details while injecting semantic context, thereby refining structural boundaries. This process is formulated as follows:

$$o = z_{high} + \mathrm{Conv}_{1\times1}(gate \odot v) \tag{18}$$

4 Experiments

4.1 Dataset

We assess our approach on two standard datasets. The primary evaluation is conducted on the Colorful Fashion Parsing Dataset (CFPD) [20], which comprises 2,682 images (600×400 pixels) with 23 pixel-level semantic categories covering main garments, accessories, and background regions. To further validate the generalization capability of SDGMamba on large-scale data, we additionally adopt the ATR human parsing dataset [21], which contains 17,706 images annotated with 18 semantic classes.

4.2 Implementation Details

All experiments were conducted using PyTorch with NVIDIA RTX 3090 GPU (24GB VRAM). The backbone was initialized with ImageNet-pretrained VMamba-tiny [14] weights, while the proposed SAA, DDR, and GCSF modules were randomly initialized. Both datasets employed the AdamW optimizer [22] and a batch size 4, with an initial learning rate of 2.5×10^{-4}, and gradient clipping with a maximum norm of 2.0. Evaluation metrics are Pixel Accuracy (PA) and mean Intersection over Union (mIoU).

On the CFPD dataset, training was performed over 30,000 iterations using a three-stage schedule: linear warm-up for the first 3,000 iterations (starting factor 0.03), polynomial decay (power 0.9) for the middle 26,000 iterations, and a constant learning rate for the final 1,000 iterations. On the ATR dataset, the model was trained for 100,000 iterations following a similar three-stage schedule, consisting of linear warm-up for the initial 8,000 iterations (starting factor 0.03), polynomial decay for the subsequent 90,000 iterations, and a constant learning rate for the last 2,000 iterations.

4.3 Comparison Experiments

For evaluating the effectiveness of SDGMamba, it is benchmarked against CNN-, Transformer-, and Mamba-based counterparts using the CFPD dataset (Table 1).

Table 1. Quantitative results against compared methods on the CFPD Dataset.

Method	PA (%)	mIoU (%)
Deeplabv3+ [4]	94.15	56.18
ConvNeXt-T [23]	94.26	56.37
SegNeXt-B [24]	94.41	57.57
Segformer-B2 [26]	94.16	55.17
Segmenter-S [27]	93.90	54.56
Swin-T [25]	94.18	56.44
VMamba-T [14]	94.24	56.49
MSVMamba-T [15]	94.34	57.55
SDGMamba (Ours)	**94.48**	**58.57**

DeepLabv3+ [4] and ConvNeXt [23] yield notable results via atrous spatial pyramid pooling and modern convolutional architectures, respectively. SegNeXt [24] advances accuracy by aggregating local details via multi-scale convolutions, while Swin [25] leverages shifted windows to capture global dependencies. VMamba [14] achieves global receptive fields, further enhanced by MSVMamba [15] via multi-scale improvements. Notably, SDGMamba achieves superior performance (94.48% PA, 58.57% mIoU). Synergizing global semantics with local details, it exceeds the runner-up by 0.07% PA and 1.00% mIoU.

To further examine per-category segmentation accuracy, we report IoU scores on representative clothing components from the CFPD dataset. As presented in Table 2, SDGMamba surpasses the competing approaches over all selected categories. The improvements are particularly pronounced on structurally complex items such as stocking and sweater, where our method yields gains of 8.08% and 6.64% IoU over the second-best result, respectively. For small accessories including bag, belt, and scarf, which pose challenges in precise boundary localization, SDGMamba also maintains clear advantages over competing approaches. The above findings validate the capability of our approach in addressing diverse clothing parsing challenges, spanning fine-scale details, elongated structures, and semantically ambiguous regions.

Table 2. Per-category IoU (%) comparison on the CFPD dataset.

Method	bag	belt	blazer	jeans	scarf	shorts	stocking	sweater
ConvNeXt-T [23]	73.46	45.83	40.00	50.38	36.98	65.25	66.33	23.07
Swin-T [25]	71.95	44.61	42.45	57.81	45.49	58.97	67.06	18.60
MSVMamba-T [15]	72.97	47.41	39.26	55.59	40.08	66.54	69.24	23.80
SDGMamba (Ours)	**74.24**	**48.75**	**46.48**	**59.46**	**46.27**	**69.70**	**77.32**	**30.44**

We further evaluate SDGMamba on the ATR dataset to assess its generalization capability. Table 3 summarizes the comparison regarding Pixel Accuracy (PA), mean Intersection over Union (mIoU), number of parameters, FLOPs, and inference speed (FPS). SDGMamba achieves the best PA (95.60%) and mIoU (62.73%), outperforming ConvNeXt-T, Swin-T, and MSVMamba-T, and it surpasses the second-best mIoU by 1.91%. SDGMamba contains 67.36M parameters and 256G FLOPs, slightly higher than ConvNeXt-T and Swin-T, yet it runs at 12.29 FPS, substantially faster than MSVMamba-T (4.47 FPS) and competitive with the convolutional architectures. These results highlight a favorable accuracy–efficiency balance.

Table 3. Quantitative results on the ATR dataset.

Method	**PA (%)**	**mIoU (%)**	**Params (M)**	**FLOPs (G)**	**FPS (img/s)**
ConvNeXt-T [23]	95.30	60.68	59.25	234	10.64
Swin-T [25]	95.19	59.51	58.95	236	17.42
MSVMamba-T [15]	95.32	60.82	63.70	234	4.47
SDGMamba (Ours)	**95.60**	**62.73**	**67.36**	**256**	**12.29**

Fig. 5 visualizes qualitative results of SDGMamba against the compared approaches for the CFPD dataset. As shown in the first row, our approach effectively rectifies pixel-level misclassifications, demonstrating superior global semantic consistency. The second and third rows highlight the discriminative capability of our model in distinguishing visually similar categories, such as T-shirts versus blouses and varying pant styles. Accurate separation of these categories is particularly challenging due to their overlapping appearances, and the results indicate that SDGMamba can reliably capture the subtle differences between them. The fourth row underscores performance on fine-scale details, where our model precisely segments small components like socks that require high spatial precision. Overall, these visual comparisons indicate that SDGMamba produces more refined boundary delineations across diverse and challenging cases.

Fig. 5. Visual results of SDGMamba and the compared methods on the CFPD dataset.

4.4 Ablation Study

To investigate the contribution of individual components within SDGMamba, we conducted ablation studies on the CFPD dataset, as summarized in Table 4. The baseline model yields 94.24% PA and 56.49% mIoU. Incorporating the SAA module yields gains of 0.20% in PA and 1.07% in mIoU—the largest single-component improvement—demonstrating that stage-aware attention effectively strengthens clothing structural perception. The subsequent integration of the DDR module further elevates PA and mIoU by 0.02% and 0.45%, respectively, underscoring its capability in refining fine-grained feature representation. The addition of GCSF brings the full configuration to peak performance (94.48% PA / 58.57% mIoU), demonstrating that the gating mechanism effectively suppresses irrelevant cross-scale information and strengthens semantic coherence. Overall, the complementary gains validate the robustness of the SDGMamba framework.

Table 4. Ablation study of our method on the CFPD Dataset.

Components			Metrics	
SAA	**DDR**	**GCSF**	**PA (%)**	**mIoU (%)**
			94.24	56.49
✓			94.44	57.56
✓	✓		94.46	58.01
✓	✓	✓	94.48	58.57

5 Conclusion

This work presents SDGMamba, an innovative clothing parsing architecture that strengthens the joint representation of holistic semantics as well as fine-grained details..Specifically, the Stage-Aware Attention VSS (SASS) Block performs hierarchical structural calibration, while the Deformable Detail Retention (DDR) module improves robustness against non-rigid deformations and preserves multi-scale details. The Gated Cross-Scale Fusion (GCSF) module further refines semantic coherence across components. Together, these modules address the challenges of intricate clothing structures and varied visual details. Evaluations on the CFPD as well as ATR benchmarks demonstrate that SDGMamba attains competitive results against CNN, Transformer, and Mamba counterparts, validating its effectiveness for clothing parsing.

References

1. Ge, Y., Song, Y., Zhang, R., Ge, C., Liu, W., Luo, P.: Parser-free virtual try-on via distilling appearance flows. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8485--8493 (2021)
2. Zhu, C., Du, P., Zhang, W., Yu, Y., Cao, Y.: Combo-fashion: Fashion clothes matching ctr prediction with item history. In: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. pp. 4621--4629 (2022)
3. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3431--3440 (2015)
4. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European conference on computer vision (ECCV). pp. 801--818 (2018)
5. Dosovitskiy, A.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
6. Hasan, B., Hogg, D.C.: Segmentation using deformable spatial priors with application to clothing. In: BMVC. pp. 1--11 (2010)
7. Yamaguchi, K., Kiapour, M.H., Ortiz, L.E., Berg, T.L.: Parsing clothing in fashion photographs. In: 2012 IEEE Conference on Computer vision and pattern recognition. pp. 3570--3577. IEEE (2012)
8. Yoo, C.H., Shin, Y.G., Kim, S.W., Ko, S.J.: Context-aware encoding for clothing parsing. Electronics Letters **55**(12), 692--693 (2019)

9. He, R., Cheng, M., Xiong, M., Qin, X., Liu, J., Hu, X.: Triple attention network for clothing parsing. In: Neural Information Processing. pp. 580--591. Springer (2020)
10. Tang, G., Yu, F., Li, H., Shi, Y., Liu, L., Peng, T., Hu, X., Jiang, M.: Clothseg: semantic segmentation network with feature projection for clothing parsing. Journal of Visual Communication and Image Representation **97**, 103980 (2023)
11. Gu, A., Goel, K., Ré, C.: Efficiently modeling long sequences with structured state spaces. arXiv preprint arXiv:2111.00396 (2021)
12. Smith, J.T., Warrington, A., Linderman, S.W.: Simplified state space layers for sequence modeling. arXiv preprint arXiv:2208.04933 (2022)
13. Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. In: First conference on language modeling (2024)
14. Liu, Y., Tian, Y., Zhao, Y., Yu, H., Xie, L., Wang, Y., Ye, Q., Jiao, J., Liu, Y.: Vmamba: Visual state space model. Neural Information Processing **37**, 103031--103063 (2024)
15. Shi, Y., Dong, M., Xu, C.: Multi-scale vmamba: Hierarchy in hierarchy visual state space model. Neural Information Processing **37**, 25687--25708 (2024)
16. Xiao, T., Liu, Y., Zhou, B., Jiang, Y., Sun, J.: Unified perceptual parsing for scene understanding. In: Proceedings of the European conference on computer vision (ECCV). pp. 418--434 (2018)
17. Hou, Q., Zhou, D., Feng, J.: Coordinate attention for efficient mobile network design. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 13713--13722 (2021)
18. Hou, Q., Zhang, L., Cheng, M.M., Feng, J.: Strip pooling: Rethinking spatial pooling for scene parsing. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4003--4012 (2020)
19. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7132--7141 (2018)
20. Liu, S., Feng, J., Domokos, C., Xu, H., Huang, J., Hu, Z., Yan, S.: Fashion parsing with weak color-category labels. IEEE Transactions on Multimedia **16**(1), 253--265 (2013)
21. Liang, X., Liu, S., Shen, X., Yang, J., Liu, L., Dong, J., Lin, L., Yan, S.: Deep human parsing with active template regression. IEEE transactions on pattern analysis and machine intelligence **37**(12), 2402--2414 (2015)
22. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
23. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11976--11986 (2022)
24. Guo, M.H., Lu, C.Z., Hou, Q., Liu, Z., Cheng, M.M., Hu, S.M.: Segnext: Rethinking convolutional attention design for semantic segmentation. Neural Information Processing **35**, 1140--1156 (2022)
25. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012--10022 (2021)
26. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: Segformer: Simple and efficient design for semantic segmentation with transformers. Neural Information Processing **34**, 12077--12090 (2021)
27. Strudel, R., Garcia, R., Laptev, I., Schmid, C.: Segmenter: Transformer for semantic segmentation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 7262--7272 (2021)