CAPNet: Cartoon Animal Parsing with Spatial Learning and Structural Modeling

Jian-Jun Qiao
Southwest Jiaotong University
Chengdu, China
qjjai56@gmail.com

Meng-Yu Duan
Southwest Jiaotong University
Chengdu, China
duanmengyu369@gmail.com

Xiao Wu*
Southwest Jiaotong University
Chengdu, China
wuxiaohk@gmail.com

Wei Li
Southwest Jiaotong University
Chengdu, China
liwei@swjtu.edu.cn

Abstract

Cartoon animal parsing aims to segment the body parts such as heads, arms, legs and tails from cartoon animals. Different from previous parsing tasks, cartoon animal parsing faces new challenges, including irregular body structures, abstract drawing styles and diverse animal categories. Existing methods have difficulties when addressing these challenges caused by the spatial and structural characteristics of cartoon animals. To address these challenges, a novel spatial learning and structural modeling network, named CAPNet, is proposed for cartoon animal parsing. It aims to address the critical problems of spatial perception, structure modeling and spatial-structural consistency learning. A spatial-aware learning module integrates deformable convolutions to learn spatial features of the irregular shapes of cartoon animals. The multi-task edge and center point predictions are incorporated to capture intricate spatial patterns. A structural modeling method is proposed to model the intricate structural representations of cartoon animals, which integrates a graph neural network with a shape-aware relation learning module. To mitigate the significant differences among animals, a spatial and structural consistency learning mechanism is proposed to capture and learn feature correlation across different animal categories. Extensive experiments conducted on benchmark datasets demonstrate the effectiveness of the proposed approach, which outperforms state-of-the-art methods.

CCS Concepts

• **Computing methodologies → Image segmentation**.

Keywords

Cartoon Parsing, Deformable Convolution, Graph Neural Network, Shape-aware Convolution

*Corresponding author: Xiao Wu

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
MM '24, October 28–November 1, 2024, Melbourne, VIC, Australia.
© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-0686-8/24/10
https://doi.org/10.1145/3664647.3680570

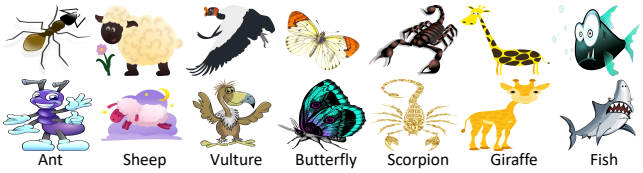

Figure 1: Cartoon animal parsing is challenging due to irregular body structures, complex visual appearances, abstract drawing styles and substantial variations among different categories. The significant differences of body parts across different species and individuals are caused by the spatial and structural characteristics of cartoon animals.

ACM Reference Format:
Jian-Jun Qiao, Meng-Yu Duan, Xiao Wu, and Wei Li. 2024. CAPNet: Cartoon Animal Parsing with Spatial Learning and Structural Modeling. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM '24), October 28–November 1, 2024, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3664647.3680570

1 Introduction

Cartoon characters are important components of various multimedia applications with their vivid, attractive and imaginative expressive capabilities. The cartoon-centric applications include the metaverse, animated films, virtual reality and artistic creations. A major category of cartoons is animals, which consist of a wide range of diverse species, including reptiles, birds, mammals, fish, and so on. They typically exhibit diverse, rich and complex body parts, such as ears, limbs, wings, claws, and so on.

Cartoon animal parsing, an emerging frontier of cartoon animals, seeks to semantically delineate various body parts. It enhances the comprehension of cartoon characters and makes progress in cartoon-centric applications. Unfortunately, previous cartoon parsing methods are proposed to parse anthropomorphic cartoon characters [26] or single animal category like cartoon dog [33], which exhibit more visual and structural consistency. On the contrary, cartoon animals have significant diversity in visual appearances and structures. As a result, techniques of cartoon parsing have unsatisfactory performance in the domain of cartoon animal parsing, which motivates this work.

Cartoon animal parsing is a challenging task due to the irregular structures, complex appearances, abstract styles and various animal categories, which is illustrated in Fig. 1. Compared to real-world humans or animals, the body parts in cartoon animals usually

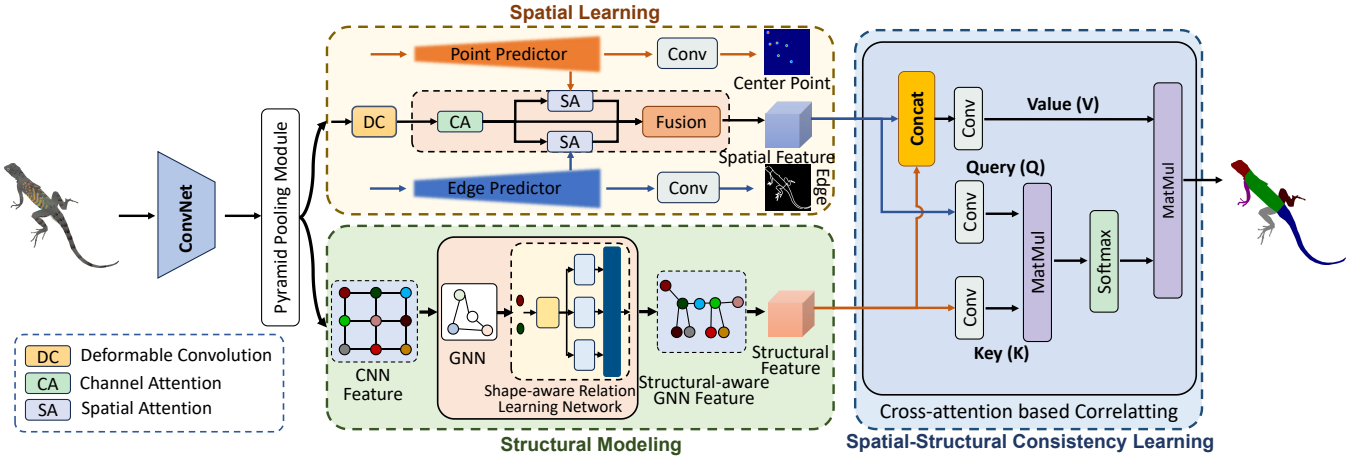

Figure 2: The framework of the proposed cartoon parsing network (CAPNet). It consists of spatial learning branch, structural modeling branch and spatial-structural consistency learning. The spatial learning branch captures the intricate spatial patterns and the structural modeling branch models the intricate structural features of cartoon animals. The spatial-structural consistency learning addresses the significant diversity and complexity by conducting consistent feature correlating and learning.

demonstrate inconsistent spatial and structural attributes. This includes complex and uncertain spatial distributions of the center points and edges of the body parts, and irregular structural shapes and sizes of the semantic parts. The spatial and structural differences among diverse cartoon animals are conspicuous even within the animals of the same category. More specifically, the body parts of real-world humans or animals possess consistent spatial and structural attributes. The spatial properties of real-world humans or animals, such as the center point positions and edge contexts of the body parts, or the structure properties like the shapes and sizes of their semantic parts, are usually consistent. However, when facing the cartoon animals, the spatial distributions of the center points and edges of the body parts, or the structure shapes and sizes of the semantic parts, are not consistent due to the abstract properties of cartoon images. Therefore, improving the spatial learning, structural modeling and achieving spatial-structural consistency learning in cartoon animal parsing is challenging.

Conventional human parsing methods have good performance in extracting visual features and predicting pixel-wise labels for human bodies. Although some approaches have incorporated attention mechanisms and graphical models to capture spatial information and structural intricacies, they are designed for the real-world human bodies, which have limited performance in irregular cartoon images. Cartoon parsing methods are proposed to address the problems caused by the scale variations [33] and irregular body structures [26] of cartoon images. Unfortunately, the spatial learning and structural modeling of cartoon animal parsing have not been fully explored.

To alleviate the challenging problems, a novel spatial learning and structural modeling network, named CAPNet, is proposed for cartoon animal parsing. The framework is illustrated in Fig. 2. The cartoon animal image is fed into ConvNet [12] followed by Pyramid Pooling Module (PPM) [47] to extract contextual features. The spatial-aware structure is designed to capture and learn spatial information, with deformable convolutions to capture spatial features of irregular shapes. In addition, it predicts the edges and center points of the body parts to obtain edge-aware and center-aware spatial features. The structural modeling approach is designed to capture the intricate structures of cartoon animals. A graph neural network (GNN) is employed to learn the animal structure. The shape-aware relation learning network is designed to correlate the node features of GNN, mitigating the complexity and diversity led by the intricate and diverse structures. To integrate the spatial and structural features and obtain consistency among diverse cartoon animals, a spatial-structural consistency learning mechanism based on the cross-attention mechanism is proposed. With the designed novel structures, CAPNet alleviates the challenges of spatial perception, structure modeling, and spatial-structural consistency learning in cartoon animal parsing. It is evaluated on a cartoon animal parsing dataset, which achieves the state-of-the-art performance. The contributions are listed as follows:

- A novel spatial learning and structural modeling network named CAPNet is proposed for cartoon animal parsing, which addresses the challenges of spatial learning, structure modeling, and spatial-structural consistency learning.
- To capture and learn the complex and inconsistent spatial attributes in diverse cartoon animals, a spatial learning branch with deformable convolutions and a multi-task prediction strategy is designed, which enhances the spatial awareness of the network.
- To model the structural information and capture the intricate structural relationships within cartoon animals, a structural modeling branch is proposed with a graph neural network and a shape-aware relation network.
- A spatial-structural consistency learning strategy is proposed with a cross-attention mechanism, which aims to achieve consistent learning among spatial and structural features and mitigates the problems caused by the complexity and diversity in cartoon animals.
- The proposed method achieves state-of-the-art performance on the cartoon parsing datasets, which demonstrates the effectiveness of the proposed spatial learning and structural modeling network.

2 Related Work

2.1 Human Parsing

The human parsing methods [4, 18, 31] have made significant progress in recent years, with numerous approaches proposed to parse body parts of humans.

Spatial Learning. Spatial information plays a crucial role in parsing tasks by providing crucial information about the relative positions. It facilitates the recognition of complex and ambiguous regions by considering the spatial distribution of the human body, achieving more accurate parsing results. Existing methods [22, 43] enhance the learning ability of spatial patterns by embedding a pose estimation branch into the human parsing framework. MMAN [21] improves the capability of capturing consistent spatial information through local and global learning. CDGNet [20] learns the spatial characteristics of the human body in horizontal and vertical directions to improve the performance of human parsing. Overall, the integration of spatial information enhances the performance of human parsing algorithms.

Structural Modeling. In addition to spatial features, structural information are also important for human parsing. To transcend the limitations of conventional CNN-based approaches, previous works have proposed structural modeling modules. These existing approaches have integrated the prowess of convolutional neural networks (CNNs) [12, 30] and graphical models [13, 19, 23, 28] to extract visual features and model structural information. Previous methods [35, 36] introduce graph neural networks (GNNs) to model the regular structure of human body, leveraging the powerful structural reasoning ability of GNNs to achieve promising results in human parsing. Some methods [17, 45] decompose and model the human body with attention mechanisms [5, 15, 16, 37, 38, 40] to learn and capture important structural information, improving the performance of human parsing. However, these techniques, tailored for human parsing, have limited performance when processing the cartoon animal images with irregular body structures, complex visual appearances, abstract drawing styles, and so on.

2.2 Cartoon Parsing

In recent years, there has been a growing interest in the field of cartoon parsing. The traditional cartoon parsing approach [39] segments cartoon images into different regions using conventional linear iterative clustering superpixels and adaptive region propagation merging techniques. However, it lacks semantic discrimination for each segmented region. Recently, pioneering works [26, 33] have adapted deep learning-based human parsing techniques to the realm of anthropomorphic cartoon characters and cartoon dogs, achieving commendable results. DFPNet [33] designs a dense multi-scale pyramid network to capture multi-scale information of cartoon dogs, addressing the scale issue in cartoon dog parsing. CPNet [26] proposes pixel and part correlation method to address the irregularities of cartoon characters. By correlating pixel-level and part-level information, it can recognize complex cartoon structures and improve the performance of cartoon parsing. However, these methods have limitations since they primarily focus on human-like forms or single-category cartoon characters such as dogs, ill-equipped to appearance diversity and structural complexity in different cartoon animals like alligators and butterflies.

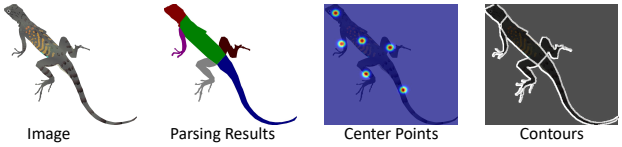

Figure 3: Center points and contours of cartoon image.

3 Cartoon Animal Parsing

3.1 Framework

To alleviate the challenges of cartoon animal parsing, a novel spatial learning and structural modeling network, named CAPNet, is proposed to integrate spatial perception, structure modeling, and spatial-structural consistency learning. The framework is illustrated in Fig. 2. ConvNet [12] and Pyramid Pooling Module (PPM) [47] are used to extract convolutional features. The spatial learning branch introduces deformable convolutions to learn spatial information related to the complex cartoon structures. This is augmented by multi-task edge and center point prediction to capture edge-aware and center-aware spatial features. The structural modeling branch employs a graph neural network to model the cartoon animal structures. It designs a body part relation learning network based on shape-aware convolutions, which enables the modeling of complex and variable structural representation. The captured spatial information and structural representation are seamlessly integrated with the cross-attention mechanism, which achieves consistency among the complex and diverse cartoon animals.

3.2 Spatial Learning

The spatial information is of great importance for cartoon animal parsing. The spatial distributions of the center points and edges of the body parts are critical to delineate the body parts of cartoon animal images, which could help recognize the complex and diverse cartoon animals. The center points and contours of cartoon image are illustrated in Fig. 3. To learn the intricate spatial patterns, CAPNet integrates deformable convolutions to transcend the rigid receptor fields of conventional convolution, capturing spatial features of irregular cartoon images. It incorporates multi-task edge and center point predictions to capture edge-aware and center-aware spatial features. This makes the network adapt to the distortions and abstraction in spatial dimension.

The framework of spatial learning is illustrated in Fig.4, in which the middle branch with deformable convolutions is adopted to extract spatial features. The edge prediction module aims to capture discriminative representations of boundaries, while the center point estimation part provides spatial cues that facilitate the localization and delineation of body parts.

Formally, given features $X \in \mathbb{R}^{C\times H\times W}$, which is generated by PPM module [48], the output features of the deformable convolutions is X_d. C refers to the number of channels, H and W denote height and width of the feature maps. The edge features of X_e are generated by an edge predictor consisting of 5 convolutional layers, with shallow and deep features of the backbone used as inputs [18, 27]. X is also processed by a heat map decoder which consists of 2 deconvolution layers to generate center-aware spatial features X_h. X_d is then fed into a channel attention module [15], aiming to build the connections of feature maps in X_d. The channel attention-based

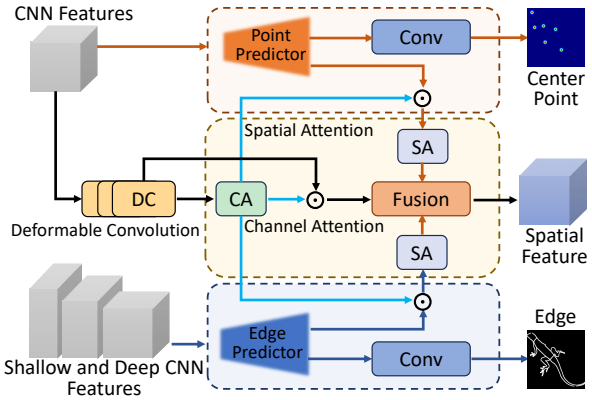

Figure 4: The framework of spatial learning.

refinement bridges the gap among the edge prediction task, the center point prediction task, and the parsing task. It is formulated as,

$$X_d^c = X_d \oplus Sig(Conv(Mp(X_d))) \odot X_d \quad (1)$$

$$X_e^c = X_e \oplus Sig(Conv(Mp(X_d))) \odot X_e \quad (2)$$

$$X_h^c = X_h \oplus Sig(Conv(Mp(X_d))) \odot X_h \quad (3)$$

where $Sig(\cdot)$ denotes sigmoid operation. $Conv(\cdot)$ means 1×1 convolution. $Mp(\cdot)$ refers to global max pooling. $\oplus$ and $\odot$ denote element-wise addition and multiplication, respectively.

The proposed approach leverages a spatial attention mechanism [38] to effectively integrate the complementary spatial information from the edge prediction and center point prediction branches, and subsequently transfer this enriched spatial information to the parsing branch. This enables the parsing branch to obtain discriminative spatial cues captured by the auxiliary tasks. The synergistic integration of multi-task spatial features, mediated by attention-based fusion, helps the network learn important spatial information in cartoon images. It is formulated as follows:

$$X_e^{sp} = Cat[X_d^c \odot Sig(Conv(Cat[Max(X_e^c), Avg(X_e^c)])), X_e^c] \quad (4)$$

$$X_h^{sp} = Cat[X_d^c \odot Sig(Conv(Cat[Max(X_h^c), Avg(X_h^c)])), X_h^c] \quad (5)$$

where $Cat(\cdot)$ denotes concatenation operation. $Max(\cdot)$ and $Avg(\cdot)$ are the functions aggregating salient spatial features by computing max and average values alongside the channel dimension, respectively.

Finally, the parsing features X_d^c operated by the deformable convolution, the edge-aware spatial features X_e^{sp}, and the canter-aware spatial features X_h^{sp} are integrated to obtain the spatial-aware parsing features. It is formulated as,

$$X_{sp} = Cat[X_d^c, X_e^{sp}, X_h^{sp}] \quad (6)$$

Through the integration of deformable convolutions and multi-task predictions, this branch helps the network learn spatial awareness. This is indispensable to address the parsing problem caused by the complex spatial distributions of cartoon animals.

3.3 Structural Modeling

The structural information, including shapes and structures, are also important for cartoon animal parsing. To model the complex structural features that characterize cartoon animals, the structural modeling module is designed. CAPNet designs a shape-aware graph

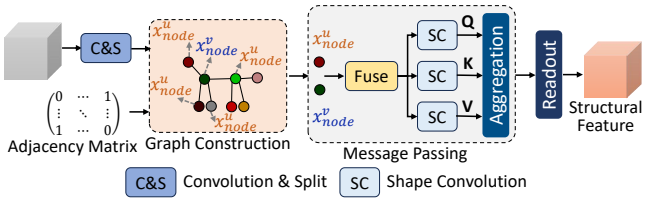

Figure 5: The framework of structural modeling.

neural network to model graphical cartoon animal representations. It employs the shape convolutions to capture shape information from graph nodes, which represent the body parts. A self-attention mechanism is adopted to learn the correlation of these graph nodes with shape information.

The framework of the structure modeling is illustrated in Fig. 5. CAPNet leverages the shape-aware GNN to construct and refine a graphical representation $\mathcal{G} = (\mathcal{N}, \mathcal{E})$, where $\mathcal{N}$ denotes the set of nodes corresponding to body parts, and $\mathcal{E}$ encodes the structural relationships of these components. The shape-aware GNN operates by updating the node representations through a message passing scheme.

In the structure modeling, X is processed to node features with a convolutional layer to adjust channel dimension and a split operation to split the features into node features $X_{node} \in \mathbb{R}^{|\mathcal{N}| \times c \times H \times W}$. $|\mathcal{N}|$ denotes the number of nodes and $|\mathcal{N}| \times c = C$. The adjacency matrix is calculated to establish connections between neighboring semantic parts. CAPNet leverages Message Passing Neural Network (MPNN) [9] to model the correlations among nodes. MPNN is structured into two stages: message passing and readout. CAPNet designs a shape-ware relation network as the message-passing function for cartoon animal features, aiming to connect different nodes with the guidance of shape information. For every node v, the message passing gathers messages m_v from its neighbors $\mathcal{N}_v$, which is formulated as,

$$m_v = \sum\nolimits_{u \in \mathcal{N}_v} M(X_{node}^u, X_{node}^v) \quad (7)$$

where X_{node}^v and X_{node}^u are elements of X_{node}. $M(\cdot)$ is the message function used to collect part relations. $M(\cdot)$ is formulated as follows:

$$M(X_{node}^u, X_{node}^v) = \phi(Cat[X_{node}^u, X_{node}^v]) \quad (8)$$

where ϕ denotes the shape-ware relation network.

The proposed shape-aware relation network leverages the information association capability of self-attention, which captures the long-range dependencies of the body parts. It distinguishes itself by incorporating shape-aware convolution to perceive geometric shapes and utilizing prior geometric information. By combining self-attention with shape-aware convolution and geometric priors, the method aims to capture both long-range dependencies and salient structural representations effectively.

The framework of the shape-aware relation network is illustrated in Fig. 5, which takes the graph nodes as inputs. The graph nodes are fed into three shape-aware convolutions with kernel size of 3, which aims to capture shape information. The vanilla convolution layer is defined as,

$$X_{vanilla} = Conv(W_K, X_{cat}^{u,v}) \quad (9)$$

where $X_{cat}^{u,v} = Cat[X_{node}^{u}, X_{node}^{v}]$. W_K denotes the learnable weights of kernels in a convolution layer.

CAPNet employs the shape-aware convolution [1] to replace the vanilla convolution in the shape-aware relation network, which aims to capture shape information. It is formulated as,

$$X_{sc}^{Q} = SC(W_K, W_B, W_S, X_{cat}^{u,v}) = Conv(W_K, W_B \diamond X_B + W_S * X_S) \tag{10}$$

where $SC(\cdot)$ denotes shape convolution. $\diamond$ and $*$ are base-product and shape-product operator [1], respectively. $X_{cat}^{u,v}$ is decomposed into two components: X_B and X_S. X_B is the mean of X and $X_S = X - X_B$ is the relative features of $X_{cat}^{u,v}$. W_B and W_S are two learnable weights to separately consume X_B and X_S. Similarly, X_{sc}^{K} and X_{sc}^{V} are computed as the key and value for the self-attention module, respectively.

The shape-aware structure learning process is defined as,

$$m_v = \text{SelfAtten}(X_{sc}^{Q}, X_{sc}^{K}, X_{sc}^{V}) \tag{11}$$

where the self-attention strategy with the geometric priors is defined as follows:

$$\text{SelfAtten}(X_{sc}^{Q}, X_{sc}^{K}, X_{sc}^{V}) = \text{Softmax}(A + G)X_{sc}^{V} \tag{12}$$

where G is initialized and learned as [14]. The difference is that the geometric priors of CAPNet are not applied to local regions but for the global feature maps. A is the attention maps calculated from X_{sc}^{Q} and X_{sc}^{K} [8, 32].

With the gathered message m_v, the node feature X_{node}^{v} is updated as,

$$X_{up}^{v} = ConvGRU(X_{node}^{v}, m_v) \tag{13}$$

where $ConvGRU$ [29] is adopted as update function. In the readout process, the node features are combined and projected into segmentation features as,

$$X_{st} = Conv(Cat[X_{up}^{0}; X_{up}^{1}; ...; X_{up}^{|\mathcal{N}|-1}]) \tag{14}$$

By integrating the GNN-based structure modeling and shape-aware relation learning, the network captures the intricate structural information of diverse cartoon animals. This is of great importance to delineate the complex and diverse body parts.

3.4 Spatial-Structural Consistency Learning

The body parts of cartoon animals exhibit substantial variations across different animal categories, which are usually caused by the complex and inconsistent spatial and structural characteristics. Therefore, a novel spatial-structural consistency learning (SSCL) mechanism to seamlessly integrate the spatial and structural features, achieving consistency among the spatial and structural dimensions. SSCL employs the cross-attention mechanism to correlate the spatial and structural features. The goal is to learn a consistent feature representation by fully utilizing the complementary spatial and structural information. Specifically, the spatial information is employed as the query, while the structural information serves as the key. The rationale behind this design is to exploit the query-key matching mechanism, enabling the model to learn the correspondence between spatial and structural features. For diverse cartoon animals, their spatial and structural characteristics may vary significantly. The cross-attention mechanism can adaptively adjust the fusion weights between these two modalities, effectively capturing the correlations between spatial and structural information. This results in a more consistent feature representation that better encapsulates the holistic characteristics of the cartoon animals.

In the consistency learning, the spatial features X_{sp} are treated as the query, the structural features X_{st} are used as the key, and the concatenated features $X_{cat} = cat[X_{sp}, X_{st}]$ as the value. The consistency learning is formulated as follows:

$$Q = X_{sp}W_q, \quad K = X_{st}W_k, \quad V = X_{cat}W_v \tag{15}$$

$$X_{fused} = \text{CrossAtten}(Q, K, V) \tag{16}$$

where W_q, W_k, W_v are learnable weights from convolution layers. Through this aggregation mechanism, a fused feature representation X_{fused} that incorporates consistent spatial and structural information is obtained.

3.5 Loss Functions

The objective function of this paper consists of parsing loss, center point loss and edge loss. The parsing loss is set to predict the segmentation results. The center point loss is used to predict the center points of the body parts. The edge loss is utilized to predict the edges of the body parts. The objective function is defined as,

$$\mathcal{L}_{all} = \mathcal{L}_{parsing} + \mathcal{L}_{spatial} + \mathcal{L}_{edge} \tag{17}$$

where $\mathcal{L}_{edge}$ [20, 27] denotes the weighted cross-entropy loss between the detected edge map and the binary edge label map.

The parsing loss is defined as,

$$\mathcal{L}_{parsing} = -\frac{1}{M}\sum_{i}^{M}\sum_{n}^{|\mathcal{N}|} g_{in} log p_{in} \tag{18}$$

where g_{ip} denotes the ground truth label of the n-th body part on the i-th pixel and p_{ip} is the related prediction result. M is the number of pixels in the image.

The center point loss [10, 26] is defined as,

$$\mathcal{L}_{spatial} = \frac{1}{|\mathcal{N}|}\sum_{n}^{|\mathcal{N}|}\sum_{x,y} \|H_n(x,y) - G_n(x,y)\|_2 \tag{19}$$

where $H_n(x,y)$ represents the predicted result for the n-th center point of the n-th body part at the pixel location (x,y). $G_n(x,y)$ is the ground truth.

The proposed method alleviates the problems caused by the complexity and diversity in the cartoon animal domain with the above objective function.

4 Experiments

4.1 Datasets and Evaluation Metrics

CASet. To evaluate the proposed methods, a cartoon animal parsing dataset called CASet is collected from the Internet and annotated. It contains 2,643 images of cartoon animals. The dataset is divided into 1,718 images for training, 262 images for validation and 663 images for testing. The diversity of the dataset comprises a wide range of cartoon animal categories, including 52 classes such as ant, horse, monkey and penguin. Each category contains dozens of cartoon images. The images are annotated according to different body parts, which include 9 categories: head, body, wing, tail, left arm, right arm, left leg, right leg, and a background class.

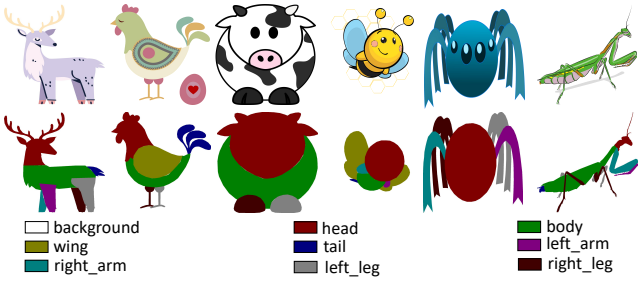

Figure 6: Visualizations of CASet.

CartoonSet. The CartoonSet dataset [26] consists of 2,229 cartoon images. Among these images, 1,530 are used for training, 510 for testing, and 189 for validation. This dataset encompasses a wide range of cartoon styles, including drawings of children, illustrations with brief strokes, and animated characters. Each image contains detailed annotations of 24 classes for the body parts.
Cartoon dog. The cartoon dog dataset [33] has 965 images of diverse cartoon dogs. It is divided into 773 images for training and 192 images for testing. The cartoon dog dataset provides annotations of eight classes for the body parts.
LIP. The LIP dataset [10] is a human parsing dataset containing 50,462 single-person images. It is divided into 30,462 images for training, 10,000 for testing, and an additional 10,000 for validation. 19 classes of the body parts of humans are annotated.
Evaluation Metrics. Following previous studies [20, 25, 26, 33], Mean Intersection over Union (Mean IoU) is adopted as the evaluation metric. It computes the average intersection-over-union ratio between predicted body parts and ground truth. Moreover, Mean Accuracy (Mean Acc.) is utilized for per-class accuracy calculation, and Pixel Accuracy (Pixel Acc.) measures the accuracy of correctly predicted pixels within each class.

4.2 Implementation Details

To implement the proposed CAPNet, a ResNet-101 backbone [12] with PPM [47] and DeepLabv3+ decoder [3] is used as the baseline model. Stochastic Gradient Descent (SGD) is employed as the optimizer with a momentum of 0.9 and a weight decay of 5e-4. Following previous works [2, 8, 24, 41, 44], CAPNet employs the "poly" learning rate strategy, defined as $lr = lr_{init} \times (1 - \frac{C_n}{T_n})^{power}$, where lr and lr_{init} are the current learning rate and the base learning rate, respectively. $power = 0.9$. C_n and T_n represent the current iteration number and the total iteration number, respectively. For data augmentation, CAPNet applies random left-right flipping with a 0.5 probability and random scaling within the range of [0.25, 1.0].

For a fair comparison on CASet, all methods are trained with a batch size of 8 and image size of 384×384. The number of training epochs is 150. For the compared methods, other train settings are kept the same as their original papers. For CartoonSet, CAPNet is trained with 150 epochs. The learning rate is 7e-3 and the training size of images is 384×384. For Cartoon dog dataset, CAPNet uses a learning rate of 3e-3, and the training images are resized to 384×384. The number of the training epochs is 300. All methods undergo single-scale evaluation on CASet, CartoonSet, and Cartoon dog datasets. For LIP dataset, CAPNet is trained with a learning rate of 7e-3, using training images resized to 473×473. The number of the training epochs is 150. For a fair comparison, CAPNet utilizes the multi-scale evaluation approach [18, 20, 36] for the LIP dataset.

Table 1: Comparison on the CASet.

Method	Pixel Acc.	Mean Acc.	Mean IoU
DeepLabV3+ [3]	93.71	80.05	69.74
DFPNet [33]	94.38	80.81	70.74
CE2P [27]	94.30	81.07	70.81
HHP [36]	94.08	80.23	70.84
SCHP [18]	94.37	82.24	72.32
CDGNet [20]	94.56	83.02	73.43
CPNet [26]	94.67	83.28	73.58
CAPNet (Ours)	**94.86**	**84.33**	**74.57**

4.3 Comparisons with the State-of-the-Art Approaches

To evaluate the performance of the proposed method, it is compared with the state-of-the-art cartoon parsing, dog parsing and human parsing methods. The comparison results are listed in Table 1. The general semantic segmentation method DeepLabV3+ [3] lacks dedicated modules for cartoon animal parsing. Accodingly, its performance is limited. Human parsing methods such as CDGNet [20], CE2P [27], and SCHP [18] focus on spatial information extraction and label noise learning, which improves the accuracy of cartoon animal parsing. But they ignore the modeling and learning of structural information, which limits the performance. HHP considers structural information modeling but overlooks spatial information learning. Moreover, when capturing spatial and structural information, they mainly consider the discriminative spatial patterns and regular structures of the human body, without taking into account the abstract and irregular characteristics of cartoon animals. DFPNet [33] is used for cartoon dog parsing, but the proposed dense pyramid structure is primarily used to address the multi-scale problem in cartoon parsing, without solving the complexity and significant diversity caused by the spatial and structural characteristics of cartoon animals. CPNet [26] improves the segmentation of the irregular structures in cartoon characters using pixel and part correlation strategy. However, it still ignores the huge differences within the body parts belong to the same classes, thus limiting its performance. Overall, these existing methods ignore the intrinsic differences between cartoon animals and real-world humans. The distinctions between cartoon animals and common cartoon characters are also not noticed. As a result, they have difficulties in cartoon animal parsing. For the challenges in cartoon animal parsing, CAPNet captures the important spatial information and complex structural context by introducing the spatial learning method and structural modeling approach. In addition, it deeply explores the essential characteristics of cartoon characters by associating the spatial and structural features, achieving consistency among diverse and complex cartoon animals. The proposed method achieves the highest results on CASet, outperforming the compared methods.

Table 2 gives the comparison results on the CartoonSet dataset. CartoonSet is a cartoon parsing dataset containing diverse and multi-style cartoon characters. Existing methods have a similar trend as in CASet. Overall, the approach using conventional CNN method have relatively worse performance, for example, DeepLabV3+

Table 2: Comparison on the CartoonSet.

Method	Pixel Acc.	Mean Acc.	Mean IoU
DeepLabV3+ [3]	87.28	63.42	50.12
DFPNet [33]	88.21	65.33	51.71
HHP [36]	87.51	64.42	51.98
CE2P [27]	88.06	66.55	52.90
CNIF [35]	87.74	66.01	53.21
CDGNet [20]	88.11	67.98	53.99
SCHP [18]	88.63	69.36	55.44
CPNet [26]	89.42	69.90	57.02
CAPNet (Ours)	**89.51**	**70.83**	**57.61**

Table 3: Comparison on the cartoon dog dataset.

Method	Pixel Acc.	Mean Acc.	Mean IoU
Mask R-CNN [11]	89.21	57.78	50.56
CE2P [27]	92.63	76.43	65.32
DFPNet [33]	93.50	79.40	68.39
SCHP [18]	94.05	81.15	71.22
CDGNet [20]	94.09	80.03	71.44
CPNet [26]	94.32	82.60	72.28
CAPNet (Ours)	**94.58**	**83.78**	**74.30**

[3]. When structural representations are captured with GNN, the performance is boosted by HHP [36] and CNIF [35]. CE2P [27] and CDGNet [20] incorporate the spatial information, including edges and class distributions, which further improves the performance of cartoon parsing. When focusing on the irregularity and complexity of cartoon characters, the performance is improved by CPNet [26]. CAPNet outperforms the state-of-the-art methods by learning both spatial and structural context, achieving consistency learning among diverse and complex cartoon animals.

The comparative results of different methods on the Cartoon dog dataset are listed in Table 3. The Cartoon dog dataset contains cartoon dogs of diverse styles and appearances, for which traditional segmentation method like Mask R-CNN [11] and human parsing approach like CE2P [27] are not well-suited. DFPNet [33] is proposed specifically for cartoon dog parsing and improves the results, but it primarily focuses on multi-scale feature learning. The performance is boosted by CDGNet [20] with its class distribution learning method. SCHP [18] further improves the results by proposing a self-correction strategy to correct misclassified regions. CPNet [26] learns the irregularity of cartoon characters, achieving better results. However, these methods still overlook the unique spatial and structural characteristics of diverse cartoon dogs, resulting in limited performance. Compared to these methods, the proposed CAPNet leverages the essential characteristics of cartoon dogs by focusing on the diversity and complexity of cartoon images, which achieves the state-of-the-art parsing results on the cartoon dog dataset.

To evaluate the generalization ability of the proposed method for human parsing, the comparison of different methods on the LIP dataset is listed in Table 4. LIP is a large-scale human parsing dataset that is widely used [18, 20]. DeepLabV3+ [3], OCR (ResNet101) [42], HRNetV2 [34] and OCR (HRNetV2-W48) [42] are common semantic segmentation methods. They have limited performance on LIP dataset due to the lack of specific components for human parsing. HHP [36], SCHP [18], CDGNet [20] and similar approaches have been designed with specific modules tailored for human parsing, achieving good performance in human parsing. Although CAPNet is proposed for cartoon animal parsing, it achieves competitive results on the LIP dataset, outperforming the cartoon parsing method CPNet [26] and most of the human parsing methods. The general semantic segmentation methods such as HRNetV2 [34] and OCR (HRNetV2-W48) [42], although have good performance on generic datasets like Cityscapes [6], exhibiting suboptimal results on the human parsing task. Compared to these common segmentation methods, CAPNet, which is mainly designed for cartoon animal parsing, achieves better results on LIP dataset. This suggests that the proposed method is effective for human parsing as well, which demonstrate its generalization ability from cartoon animal parsing to human parsing.

Table 4: Comparison on the LIP dataset.

Method	Pixel Acc.	Mean Acc.	Mean IoU
DeepLabV3+ [3]	n/a	n/a	52.09
CE2P [27]	87.37	63.20	53.10
CorrPM [46]	87.68	67.21	55.33
OCR (ResNet101) [42]	n/a	n/a	55.60
HRNetV2 [34]	n/a	n/a	55.90
OCR (HRNetV2-W48) [42]	n/a	n/a	56.65
CPNet [26]	88.29	68.41	57.21
CNIF [35]	88.03	68.80	57.74
HHP [36]	**89.05**	70.58	59.25
SCHP [18]	n/a	n/a	59.36
CDGNet [20]	88.86	**71.49**	**60.30**
CAPNet (Ours)	88.40	68.80	57.57

Table 5: Effect of the proposed modules on CASet dataset. SLB: Spatial Learning Branch. SMB: Structural modeling Branch. SSCL: Spatial-Structural Consistency Learning.

Method	Pixel Acc.	Mean Acc.	Mean IoU
Baseline	93.69	79.38	69.33
Baseline + SLB	94.56	82.54	72.24
Baseline + SMB	94.46	81.84	71.97
Baseline + SLB + SMB	94.82	83.52	74.04
Baseline + SLB + SMB + SSCL	**94.86**	**84.33**	**74.57**

4.4 Ablation Studies

To validate the effectiveness of each module in CAPNet, ablation studies are conducted. Table 5 gives the experimental results of the proposed three modules: SLB (Spatial Learning Branch), SMB (Structural Modeling Branch), and SSCL (Spatial-Structural Consistency Learning). The baseline model, which lacks specialized designs for cartoon animal parsing, has limited performance on the CASet dataset. When SLB module is added to learn spatial information, a significant performance boost is obtained, which proves the importance of spatial learning. With the SMB module utilized to the baseline, the performance is substantially improved, which demonstrates the crucial role of structural information modeling. The performance is further enhanced by integrating the spatial features and structural contexts, which indicates the complementary nature of these two types of information and their ability to collectively improve cartoon animal parsing. SSCL is proposed to learn the correlation between spatial cues and structural representations, which results in an additional performance gain. This suggests that the integration and correlation of the spatial and structural features is essential to address the challenges in cartoon animal parsing.

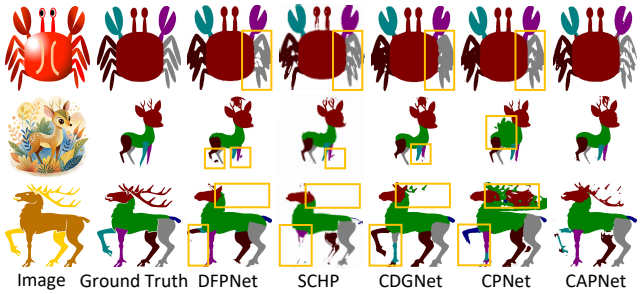

Figure 7: Visualizations of the state-of-the-art methods.

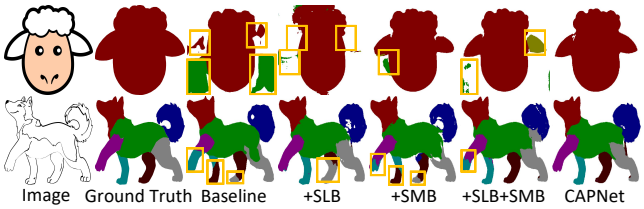

Figure 8: Visualizations of different components.

Table 6: Effect of the deformable convolutions (DC), shape-aware convolutions (SC), and geometry prior (GP). The symbol w/o means without.

Method	Pixel Acc.	Mean Acc.	Mean IoU
CAPNet	**94.86**	**84.33**	**74.57**
CAPNet (w/o SC)	94.78	83.48	73.85
CAPNet (w/o DC)	94.57	83.18	73.22
CAPNet (w/o GP)	94.75	83.53	73.72

To effectively capture the spatial patterns and structural representations of cartoon animals, CAPNet employs deformable convolution [7], shape-aware convolution [1] and geometry prior [14]. Ablation studies are further conducted to demonstrate the effectiveness of these components. The results of the ablation experiments are listed in Table 6. As can be seen from the table, CAPNet without the deformable convolution, shape-aware convolution or geometry prior leads to a decline in performance, which indicates that these structures are effective for the network to capture and learn spatial distributions and structural shapes. Deformable convolution, with its unique convolution design, enables the network to learn irregular shapes and structures. It enhances the ability of CAPNet to capture spatial patterns of cartoon images. Shape-aware convolution helps the network capture the shape information of the body parts, which boosts the structural representation ability of CAPNet. The geometry prior strengthens the perception and learning of the geometric structure of cartoon animals. The integration of these components contribute to the state-of-the-art performance of CAPNet in cartoon animal parsing by strengthening the learning and modeling of the important spatial cues and structural contexts.

4.5 Qualitative Results

Fig. 7 showcases the parsing results of different methods on the CASet dataset. Compared to other methods, the proposed method has better performance in segmenting various and intricate body parts of cartoon animals such as crab and deer. Such results also demonstrate that the proposed method has good generalization capabilities across different styles and species of cartoon animals. The ability of CAPNet in recognizing diverse and complex animals is attributed to its capabilities of learning the spatial patterns and modeling the structural information of cartoon animals.

Figure 9: Examples of failure cases.

The visualization results of the proposed modules are illustrated in Fig. 8. Due to the significant differences and complex structures of the cartoon animals, the baseline model struggles to segment the body parts correctly. By incorporating the spatial learning and structural modeling method, the parsing results of cartoon animals are noticeably improved. However, simply combining the spatial learning and structural modeling still leads to limited performance. By leveraging the spatial-structural consistency learning to learn the spatial and structural correlations, the performance of the network is further improved, leading to more accurate parsing results for cartoon animals.

Although CAPNet achieves promising performance for cartoon animal parsing, there is still some space to improve. Currently, it focuses on the spatial learning and structural modeling of cartoon animals. But it ignores the interference and influence of complex backgrounds. When the backgrounds have similar colors, visual appearances or visual textures to cartoon animals, CAPNet has difficulties in capturing and learning the relationship between cartoon animals and the complex backgrounds. As a result, it cannot distinguish cartoon animals from complex backgrounds. Some failure examples are illustrated in Fig. 9, which illustrate the limitations of CAPNet.

5 Conclusion

In this paper, a spatial learning and structural modeling network, named CAPNet, tailored for cartoon animal parsing is introduced. It addresses the challenges in spatial information learning, structural representation modeling, as well as spatial-structural consistency learning in cartoon animal parsing. The spatial-ware learning structure utilizes deformable convolutions to capture spatial features of irregular cartoon animals. In addition, it predicts edges and center points of semantic parts to learn edge-aware and center-aware spatial patterns. The structural modeling approach incorporates GNN and shape-aware relation learning network to model intricate structures. The spatial and structural consistency learning module enhances feature representation consistency among dissimilar cartoon animals, which demonstrates improved parsing accuracy. In the future, the proposed method will be improved to address the problems caused by complex backgrounds. The dataset will be expanded to include more animal categories with more diverse and intricate backgrounds.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (Grant No. 62372387, 62001400), Key R&D Program of Guangxi Zhuang Autonomous Region, China (Grant No. AB22080038, AB22080039), Sichuan Science and Technology Program (Grant No. 2024NSFSC0494), Fundamental Research Funds for

the Central Universities (2682024ZTPY044) and China Postdoctoral Science Foundation (Grant No. 2021M702713).

References

[1] Jinming Cao, Hanchao Leng, Dani Lischinski, Danny Cohen-Or, Changhe Tu, and Yangyan Li. 2021. ShapeConv: Shape-aware Convolutional Layer for Indoor RGB-D Semantic Segmentation. In *IEEE International Conference on Computer Vision*. 7068–7077.

[2] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. 2017. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40, 4 (2017), 834–848.

[3] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. 2018. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In *European Conference on Computer Vision*. 833–851.

[4] Weihua Chen, Xianzhe Xu, Jian Jia, Hao Luo, Yaohua Wang, Fan Wang, Rong Jin, and Xiuyu Sun. 2023. Beyond Appearance: A Semantic Controllable Self-Supervised Learning Framework for Human-Centric Visual Tasks. In *IEEE Conference on Computer Vision and Pattern Recognition*. 15050–15061.

[5] Zhi-Qi Cheng, Qi Dai, Siyao Li, Teruko Mitamura, and Alexander Hauptmann. 2022. GSRFormer: Grounded Situation Recognition Transformer with Alternate Semantic Attention Refinement. In *ACM International Conference on Multimedia*. 3272–3281.

[6] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. 2016. The Cityscapes Dataset for Semantic Urban Scene Understanding. In *IEEE Conference on Computer Vision and Pattern Recognition*. 3213–3223.

[7] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. 2017. Deformable Convolutional Networks. In *IEEE International Conference on Computer Vision*. 764–773.

[8] Jun Fu, Jing Liu, Haijie Tian, Zhiwei Fang, and Hanqing Lu. 2019. Dual Attention Network for Scene Segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*. 3146–3154.

[9] Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. 2017. Neural Message Passing for Quantum Chemistry. In *IEEE International Conference on Machine Learning*. 1263–1272.

[10] Ke Gong, Xiaodan Liang, Dongyu Zhang, Xiaohui Shen, and Liang Lin. 2017. Look into Person: Self-Supervised Structure-Sensitive Learning and a New Benchmark for Human Parsing. In *IEEE Conference on Computer Vision and Pattern Recognition*. 6757–6765.

[11] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. 2017. Mask R-CNN. In *IEEE International Conference on Computer Vision*. 2980–2988.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*. 770–778.

[13] Yixuan He, Quan Gan, David Wipf, Gesine D. Reinert, Junchi Yan, and Mihai Cucuringu. 2022. GNNRank: Learning Global Rankings from Pairwise Comparisons via Directed Graph Neural Networks. In *IEEE International Conference on Machine Learning*. 8581–8612.

[14] Han Hu, Zheng Zhang, Zhenda Xie, and Stephen Lin. 2019. Local Relation Networks for Image Recognition. In *IEEE International Conference on Computer Vision*. 3463–3472.

[15] Jie Hu, Li Shen, and Gang Sun. 2018. Squeeze-and-Excitation Networks. In *IEEE Conference on Computer Vision and Pattern Recognition*. 7132–7141.

[16] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. 2019. CCNet: Criss-Cross Attention for Semantic Segmentation. In *IEEE International Conference on Computer Vision*. 603–612.

[17] Ruyi Ji, Dawei Du, Libo Zhang, Longyin Wen, Yanjun Wu, Chen Zhao, Feiyue Huang, and Siwei Lyu. 2020. Learning Semantic Neural Tree for Human Parsing. In *European Conference on Computer Vision*. 205–221.

[18] Peike Li, Yunqiu Xu, Yunchao Wei, and Yi Yang. 2022. Self-Correction for Human Parsing. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 6 (2022), 3260–3271.

[19] Wanyu Lin, Hao Lan, Hao Wang, and Baochun Li. 2022. OrphicX: A Causality-Inspired Latent Variable Model for Interpreting Graph Neural Networks. In *IEEE Conference on Computer Vision and Pattern Recognition*. 13719–13728.

[20] Kunliang Liu, Ouk Choi, Jianming Wang, and Wonjun Hwang. 2022. CDGNet: Class Distribution Guided Network for Human Parsing. In *IEEE Conference on Computer Vision and Pattern Recognition*. 4463–4472.

[21] Yawei Luo, Zhedong Zheng, Liang Zheng, Tao Guan, Junqing Yu, and Yi Yang. 2018. Macro-Micro Adversarial Network for Human Parsing. In *European Conference on Computer Vision*. 424–440.

[22] Xuecheng Nie, Jiashi Feng, and Shuicheng Yan. 2018. Mutual Learning to Adapt for Joint Human Parsing and Pose Estimation. In *European Conference on Computer Vision*. 519–534.

[23] Yinhua Piao, Sangseon Lee, Dohoon Lee, and Sun Kim. 2022. Sparse Structure Learning via Graph Neural Networks for Inductive Document Classification. In *Association for the Advancement of Artificial Intelligence*. 11165–11173.

[24] Jian-Jun Qiao, Zhi-Qi Cheng, Xiao Wu, Wei Li, and Ji Zhang. 2022. Real-time Semantic Segmentation with Parallel Multiple Views Feature Augmentation. In *ACM International Conference on Multimedia*. 6300–6308.

[25] Jian-Jun Qiao, Xiao Wu, Jun-Yan He, Wei Li, and Qiang Peng. 2022. SWNet: A Deep Learning Based Approach for Splashed Water Detection on Road. *IEEE Transactions on Intelligent Transportation Systems* 23, 4 (2022), 3012–3025.

[26] Jian-Jun Qiao, Jie Zhang, Xiao Wu, Yu-Pei Song, and Wei Li. 2023. CPNet: Cartoon Parsing with Pixel and Part Correlation. In *ACM International Conference on Multimedia*. 6888–6897.

[27] Tao Ruan, Ting Liu, Zilong Huang, Yunchao Wei, Shikui Wei, and Yao Zhao. 2019. Devil in the Details: Towards Accurate Single and Multiple Human Parsing. In *Association for the Advancement of Artificial Intelligence*. 4814–4821.

[28] Simon Schaefer, Daniel Gehrig, and Davide Scaramuzza. 2022. AEGNN: Asynchronous Event-based Graph Neural Networks. In *IEEE Conference on Computer Vision and Pattern Recognition*. 12361–12371.

[29] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. 2015. Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting. In *Advances in Neural Information Processing Systems*. 802–810.

[30] Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *International Conference on Learning Representations*.

[31] Shixiang Tang, Cheng Chen, Qingsong Xie, Meilin Chen, Yizhou Wang, Yuanzheng Ci, Lei Bai, Feng Zhu, Haiyang Yang, Li Yi, Rui Zhao, and Wanli Ouyang. 2023. HumanBench: Towards General Human-Centric Perception with Projector Assisted Pretraining. In *IEEE Conference on Computer Vision and Pattern Recognition*. 21970–21982.

[32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*. 5998–6008.

[33] Jerome Wan, Guillaume Mougeot, and Xubo Yang. 2020. Dense Feature Pyramid Network for Cartoon Dog Parsing. *The Visual Computer* 36, 10 (2020), 2471–2483.

[34] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao. 2021. Deep High-Resolution Representation Learning for Visual Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43, 10 (2021), 3349–3364.

[35] Wenguan Wang, Zhijie Zhang, Siyuan Qi, Jianbing Shen, Yanwei Pang, and Ling Shao. 2019. Learning Compositional Neural Information Fusion for Human Parsing. In *IEEE International Conference on Computer Vision*. 5702–5712.

[36] Wenguan Wang, Hailong Zhu, Jifeng Dai, Yanwei Pang, Jianbing Shen, and Ling Shao. 2020. Hierarchical Human Parsing with Typed Part-Relation Reasoning. In *IEEE Conference on Computer Vision and Pattern Recognition*. 8926–8936.

[37] Xiaolong Wang, Ross B. Girshick, Abhinav Gupta, and Kaiming He. 2018. Non-Local Neural Networks. In *IEEE Conference on Computer Vision and Pattern Recognition*. 7794–7803.

[38] Sanghyun Woo, Jongchan Park, Lee Joon-Young, and In So Kweon. 2018. CBAM: Convolutional Block Attention Module. In *European Conference on Computer Vision*. 3–19.

[39] Huisi Wu, Yilin Wu, Shenglong Zhang, Ping Li, and Zhenkun Wen. 2017. Cartoon Image Segmentation based on Improved SLIC Superpixels and Adaptive Region Propagation Merging. In *IEEE International Conference on Signal and Image Processing*.

[40] Lingxiao Yang, Ru-Yuan Zhang, Lida Li, and Xiaohua Xie. 2021. SimAM: A Simple, Parameter-Free Attention Module for Convolutional Neural Networks. In *IEEE International Conference on Machine Learning*. 11863–11874.

[41] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. 2018. BiSeNet: Bilateral Segmentation Network for Real-Time Semantic Segmentation. In *European Conference on Computer Vision*. 334–349.

[42] Yuhui Yuan, Xilin Chen, and Jingdong Wang. 2020. Object-Contextual Representations for Semantic Segmentation. In *IEEE International Conference on Computer Vision*. 173–190.

[43] Dan Zeng, Yuhang Huang, Qian Bao, Junjie Zhang, Chi Su, and Wu Liu. 2021. Neural Architecture Search for Joint Human Parsing and Pose Estimation. In *IEEE International Conference on Computer Vision*. 11365–11374.

[44] Ji Zhang, Xiao Wu, Zhi-Qi Cheng, Qi He, and Wei Li. 2023. Improving Anomaly Segmentation with Multi-Granularity Cross-Domain Alignment. In *ACM International Conference on Multimedia*. 8515–8524.

[45] Xiaomei Zhang, Yingying Chen, Bingke Zhu, Jinqiao Wang, and Ming Tang. 2020. Blended Grammar Network for Human Parsing. In *European Conference on Computer Vision*. 189–205.

[46] Ziwei Zhang, Chi Su, Liang Zheng, and Xiaodong Xie. 2020. Correlating Edge, Pose with Parsing. In *IEEE Conference on Computer Vision and Pattern Recognition*. 8897–8906.

[47] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. 2017. Pyramid Scene Parsing Network. In *IEEE Conference on Computer Vision and Pattern Recognition*. 6230–6239.

[48] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. 2017. Pyramid Scene Parsing Network. In *IEEE Conference on Computer Vision and Pattern Recognition*. 6230–6239.