CartoonNet: Cartoon Parsing with Semantic Consistency and Structure Correlation

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

Yu-Pei Song
Southwest Jiaotong University
Chengdu, China
yupei-song@my.swjtu.edu.cn

Abstract

Cartoon characters, with their complex appearances, abstract drawing styles, and irregular structures, pose great challenges for cartoon parsing. In this paper, a novel approach, named CartoonNet, is proposed for cartoon parsing that aims to recognize and segment various parts of cartoon characters. Semantic consistency and structure correlation, acting as two key factors, are integrated to address the visual diversity and structure complexity for cartoon images. A memory-based semantic consistency module is designed to encode and learn the diverse appearances exhibited by cartoon characters. It conducts consistent learning among the body parts that belong to the same class but from different characters, by storing, selecting and correlating different cartoon images. To recognize the intricate and irregular structures present in cartoon images, a structure correlation module is proposed. Leveraging graph attention networks and a main body-aware mechanism, the proposed approach enables structural learning and correlating, allowing it to recognize and parse cartoon images with significant complexity. Experiments conducted on cartoon parsing datasets demonstrate the effectiveness of the proposed method. Moreover, the proposed method also achieves competitive performance on human parsing dataset, proving its superiority.

CCS Concepts

• **Computing methodologies → Image segmentation**.

Keywords

Cartoon Parsing, Memory Structure, Graph Attention Network

ACM Reference Format:
Jian-Jun Qiao, Meng-Yu Duan, Xiao Wu, and Yu-Pei Song. 2024. CartoonNet: Cartoon Parsing with Semantic Consistency and Structure Correlation. 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, 9 pages. https://doi.org/10.1145/XXXXXX.XXXXXX

*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/XXXXXX.XXXXXX

Figure 1: Cartoon parsing remains a challenging task because the cartoon characters have various, complex and abstract appearances and structures. From human-like cartoon characters to animal-like cartoon images, the diversity and complexity of cartoon images pose great difficulties to cartoon parsing.

1 Introduction

Cartoon characters, with their diverse and abstract appearances and structures, are an important part of culture and media. From human-like cartoon characters to animal-like cartoon images, cartoons exhibit a wide range of forms and styles. As a result, recognizing the complex and varied cartoon appearances and structures poses significant challenges in computer vision. The task of cartoon parsing, which aims to segment different parts of cartoon characters, holds substantial importance for various applications. These applications span cartoon animation, content creation, virtual worlds, and more. However, the inherent irregularities, variations, and lack of uniformity in cartoon structures and appearances make this task particularly challenging.

Cartoon images usually have various, complex and abstract visual elements, including irregular lines, color variations, and shape changes. Some examples are illustrated in Fig. 1. Unlike traditional photographs, the style and grammar rules of cartoon images may vary depending on the work, author, or cultural background, which adds to the complexity of cartoon parsing. Therefore, cartoon parsing remains a challenging task for cartoon characters. To achieve effective cartoon parsing, it is essential to design image analysis algorithms specifically tailored for the domain of cartoons, accounting for the complexity, diversity, irregularity, and abstraction present in cartoon images.

For image parsing, existing methods have primarily focused on human parsing and achieved significant progress. Previous methods

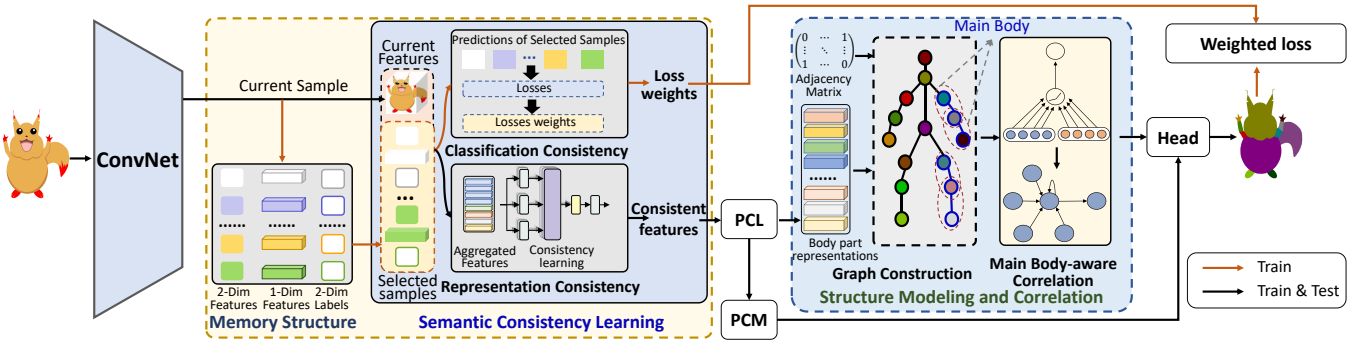

Figure 2: The framework of the proposed CartoonNet. It mainly consists of memory-based semantic consistency and main body-aware structure correlation. The memory-based semantic consistency utilizes a memory structure to achieve consistency learning across different samples. The main body-aware structure correlation focuses on correlating body structures, perceiving the structure of cartoons. PCL and PCM refers to pixel correlation learning and part correlation modeling [23], which are employed to solve the irregularities of cartoon images.

for human parsing have explored various approaches, such as hierarchical structure learning [12, 34, 35, 42], pose estimation-guided learning [19, 41], and the integration of self-correction strategy [14]. However, due to the significant gap and differences between real humans and cartoons, these human parsing methods exhibit suboptimal performance in cartoon parsing. In recent research, deep learning methods [23, 32] have been applied to cartoon parsing. They address the challenges in cartoon parsing by employing multi-scale learning structures, and leveraging pixel and part correlation to infer irregular cartoon structures. However, dealing with the diversity and complexity of cartoon images remains a challenge.

In this paper, CartoonNet is proposed to address the diversity and complexity of cartoon images in cartoon parsing. It augments the segmentation of cartoon images by designing memory-based semantic consistency structure and graph attention network-based structure correlation module. As illustrated in Fig. 2, the semantic consistency module encodes and learns the diverse appearances displayed by cartoon characters. It designs a memory-based learning structure to conduct consistent learning of diverse body parts from different cartoon images but belong to the same class. This enhances the method's ability to parse cartoon characters with substantial individual appearance differences, accommodating the rich diversity of visual styles encountered in cartoons. The structure correlation approach focuses on modeling and learning the intricate and irregular structures exhibited by cartoon characters. It leverages graph attention networks (GATs) [2, 11, 31] and a main body-aware mechanism to facilitate structural learning and correlation. This enables the method to recognize and parse cartoon characters characterized by substantial structural complexity. By correlating the distinctive and unique structural features of cartoon entities across global graphs and main bodies, the method excels in handling structural complexity. A series of experiments and evaluations are conducted to demonstrate the effectiveness of the proposed approach in recognizing and segmenting various parts of cartoon images. The contributions are summarized as follows:

- CartoonNet is proposed to address the challenges of visual diversity and structure complexity in cartoon images, which designs a memory-based semantic consistency approach and a main body-aware structure correlation strategy for cartoon parsing.
- A memory-based learning structure is proposed to facilitate appearance encoding, storage, selecting and consistency learning. By capturing the inconsistent appearance features of different cartoon characters and accommodating variations in color, texture, and details, the approach aims to address the diversity of cartoon images.
- A main body-aware structure correlation method is proposed, which enables the recognition of cartoon characters with substantial structural complexity. By modeling and correlating the main bodies with the graph attention networks, it addresses the complexity of cartoon structures in complex anthropomorphic and animal-like cartoons.
- The proposed method achieves state-of-the-art performance on the cartoon parsing datasets, which demonstrates the effectiveness of the proposed approach of semantic consistency and structure correlation.

2 Related Work

2.1 Human Parsing

Human parsing [5, 16, 29], the task of segmenting human body parts in images, has achieved significant progress. Previous human parsing methods develop approaches including hierarchical structure learning [12, 34, 35, 42], pose estimation-guided learning [19, 41], self-correction strategy [14], and so on, achieving satisfactory results on recognition of body parts of humans. However, applying human parsing methods to cartoon parsing presents challenges due to fundamental differences. Cartoons are known for their complexity, diverse styles, irregular lines, unique color schemes, and abstract representations, varying greatly based on creators and cultural contexts. Human parsing models [15, 18] are designed on real-world human images, making them unsuitable for cartoon parsing, where characters often have unconventional body proportions and articulations. Cartoon parsing necessitates dedicated research and specialized algorithms to address its unique characteristics, emphasizing accuracy and effectiveness in dealing with the diversity and complexity of cartoon images.

2.2 Cartoon Parsing

The task of cartoon parsing has gained increasing attention in recent years due to its relevance in various multimedia applications. Researchers have proposed several approaches to address the challenges posed by diverse and complex cartoon appearances

and structures. Early works in this domain primarily focused on rule-based methods and handcrafted feature extraction techniques. These approaches, while capable of capturing some aspects of cartoon images, often struggled with the diversity and complexity of cartoon imagery. The emergence of deep learning techniques, particularly the convolutional neural networks (CNNs) [1, 17, 25, 27], has significantly advanced the field of cartoon parsing. DFPNet [32] applies the feature pyramid network to cartoon dog parsing, which alleviates the varied scales in cartoon parsing. However, it ignores the correlations of visual cues and structural features in cartoon characters. CPNet [23] recognizes the irregular cartoon structures and body parts with different semantics but have visually akin appearances, by introducing a pixel and part correlation learning structure. However, it ignores the semantic consistency of diverse cartoon images and structure correlation of complex cartoon characters. Although CNN-based methods demonstrate improved performance in recognizing cartoon characters and their attributes, addressing the diversity and complexity of cartoon characters remains a challenge.

2.3 Memory-Augmented Networks

Memory-augmented neural networks [36, 37] have shown promise in capturing and utilizing long-term dependencies in data. These networks [13, 24] employ an external memory structure that can be read from and written to, enabling the storage and retrieval of information over extended sequences. In the context of cartoon parsing, memory-augmented networks can offer a promising avenue for encoding and learning the multifaceted appearances exhibited by cartoon characters. Despite the growing interest in memory-augmented networks, their potential for cartoon parsing remains unexplored. The ability to store appearance information and perform content-based consistent learning opens possibilities for addressing individual appearance differences. Therefore, this paper proposes a novel approach applicable to cartoon parsing, with the aim of addressing the issue of significant sample diversity in cartoon parsing through a memory-based structure.

3 Cartoon Parsing

3.1 Framework

The proposed CartoonNet for cartoon parsing is designed to address the intricacies of diverse cartoon characters. A two-fold strategy that combines appearance learning and structural modeling is introduced to achieve accurate and robust cartoon parsing. As shown in Fig. 2, a ConvNet [10] is employed to encode and extract features from cartoon images. The features related to diverse appearances of different cartoon images are stored in the memory structure and the model traverses the memory structure using the features of the current batch of cartoon images. Through the computation of feature similarity, the model selects the samples interrelate with the current sample, aiming to recognize the characteristics of the current sample based on these resemblances from different samples, enhancing the generalization ability of the network. With the current batch and the selected samples, a semantic consistency learning strategy is designed to recognize diverse and complex cartoon characters, focusing on the consistency of body parts from different cartoon characters. To learn the correlation of body parts, a main body-aware structure correlation strategy is designed. It incorporates graph attention networks and main body-aware mechanism to facilitate part correlation and learning, further improving the parsing capabilities of the network in modeling complex structures of cartoon characters.

3.2 Memory Structure Construction

Since previous cartoon parsing methods only consider multi-scale learning, as well as pixel and part correlation among a single cartoon character, they have difficulties in recognizing diverse cartoon images. Therefore, the first significant component of the approach is to encode and learn the varied appearances displayed by diverse cartoon characters. A memory-based learning structure is proposed to tackle the challenge of appearance diversity, aiming to achieve semantic consistency among different cartoon characters. The memory-based consistency learning includes two parts, representation consistency and classification consistency. The representation consistency is designed to conduct consistency learning on semantic parts from different cartoon characters at high-dimensional semantic feature level. The classification consistency is proposed to achieve consistency learning of semantic parts from various cartoon characters at the semantic recognition level, which leverages the constraint information.

Given a cartoon image X as an input to the ConvNet, $F \in \mathbb{R}^{C\times H\times W}$ is the encoded image features, where C denotes the number of channels, H and W refer to height and width of the feature maps. A memory structure with three columns is defined, where the first column stores the two-dimensional feature maps, the second column saves one-dimensional feature representations, and the third column retains labels. Specifically, F is fed to a decoder with three convolution layers [4] to obtain the two-dimensional feature maps $F_{mem} \in \mathbb{R}^{C_n\times H\times W}$, where C_n denotes the number of classes. F_{mem} is then stored in the first column of the memory structure. Additionally, the operation of information aggregation with global adaptive max pooling is performed on each feature map of F, which yields a feature vector $F_v \in \mathbb{R}^{C\times 1\times 1}$. F_v is the representation of X and it is stored in the second column of the memory structure. The label $L \in \mathbb{R}^{H\times W}$ corresponding to X is stored in the third column of the memory structure, which aims to provide supervision for the semantic consistency learning. In the initial stage, all three columns of the memory structure contain zeros. During the training process, image features are stored in the memory structure in a stacked manner. When the memory is full, the earliest stored features are popped out to ensure that the memory structure is continuously updated throughout the training process.

During each iteration, F not only gets stored in the memory structure but also undergoes further forward operations. A memory selection operation is performed to choose samples from the memory structure. These selected samples have some similarities to the current sample and they are used to achieve semantic consistency, which aims to enhances the semantic reasoning capability of network for diverse targets sharing similar characteristics. Specifically, when inputting a new sample, the network selects similar samples from the memory structure. By leveraging the knowledge acquired from these learned samples, the network accurately infers and identifies the characteristics of the current sample. To

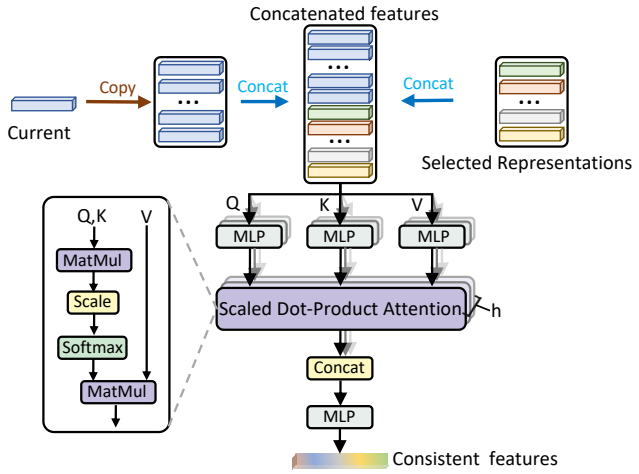

Figure 3: Memory-based representation consistency.

achieve memory selection, during each training iteration, a similarity calculation is performed between the current F^c_{mem} and the stored F^m_{mem} within the memory structure. This is done using the Euclidean distance formula, which retrieves n samples from the memory structure.

3.3 Representation Consistency

The problem of inconsistent appearances among different cartoon images poses great challenges to cartoon parsing. To achieve consistency learning for the body parts that belong to the same class but from different samples, the network initially performs consistency learning at the representation level. This involves the integration and consistency of the feature representations of the current sample and the feature representations of relevant samples retrieved from the memory structure, as can be seen in Fig. 3. The multi-head self-attention structure [30] is employed to achieve consistent learning among the semantic parts from different samples. It ensures that the current sample and the samples stored in the memory structure are associated and consistent at the feature representation level.

Specifically, the current features F is mapped into 1-dimensional feature representation with a global adaptive max pooling operation, which is denoted as F^c_v. Through the memory selection, n feature representations $F^{m,n}_v$ that represents the retrieved samples are selected for the current sample. F^c_v is then duplicated n times and concatenated with $F^{m,n}_v$ to aggregate information. This is formulated as follows:

$$F^{cat}_v = Cat(Copy(Gap(F)), Selection(M)) \tag{1}$$

where F^{cat}_v is the concatenated features. $Gap(\cdot)$ denotes global adaptive max pooling operation. $Copy(\cdot)$ is the duplication operation. $Cat(\cdot)$ refers to concatenation operation. $Selection(\cdot)$ is the memory selection operation which retrieves the memory M for related samples. Subsequently, multi-perceptual (MLP) layers are used to create queries for the multi-head self-attention structure. It is formulated as,

$$Q = F^{cat}_v W^Q, K = F^{cat}_v W^K, V = F^{cat}_v W^V \tag{2}$$

where W^Q, W^K, and W^V are learnable weight matrices that project the input vectors F^{cat}_v into query Q, key K, and value V, respectively.

The multi-head attention mechanism calculates the output vectors F^{out}_v as a weighted sum of the input vectors F^{cat}_v, where the weights are determined by the attention scores. The attention scores are calculated as,

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{3}$$

where dot products of the query Q with all keys K are computed and scaled by the scaling factor $\frac{1}{\sqrt{d_k}}$. d_k is the dimension of query. The softmax operation generates the attention scores. The dot products of the attention score and value V associate the representations of different samples.

The multi-head attention mechanism processes the projected vectors with h heads and performs the attention calculation in parallel for each head. The outputs of the heads are then concatenated and projected back to the original dimension of F^{cat}_v:

$$\text{head}_i = \text{Attention}(QW^Q_i, KW^K_i, VW^V_i) \tag{4}$$

$$F^{out}_v = \text{MultiHead}(Q, K, V) = Cat(\text{head}_1, \ldots, \text{head}_h)W^O \tag{5}$$

where W^Q_i, W^K_i, W^V_i, and W^O are learnable weight matrices.

This allows the model to attend to different representations of the same input sequence when calculating the output vectors F^{out}_v which denotes the consistent features. F^{out}_v is fused to the original features F as,

$$F_{rc} = Cat(F, F \odot Sig(F^{out}_v)) \tag{6}$$

where $Sig(\cdot)$ denotes sigmoid operation that aims to map the values of vectors into calibration weights and $\odot$ refers to element-wise multiplication.

3.4 Classification Consistency

In addition to representation consistency, the proposed method also performs classification consistency between the current sample and the samples retrieved from the memory structure. This enables the network to consistently recognize semantic parts that share the same semantics but may exhibit variations in appearances.

With the memory selection, the network obtains encoded predictions $F^{m,n}_{mem}$ of the retrieved n samples, which are related to the encoded feature F of the current sample X. F is mapped to the category dimension for semantic prediction. The mapping process uses the decoder of DeepLabV3+ [4], which is widely used in parsing methods [14, 26]. It is formulated as,

$$F^c_{mem} = Decoder(F, F') \tag{7}$$

where F' denote the shallow features from ConvNet that provide spatial details for the deep features F.

To ensure that the network consistently recognizes the body parts that belong to the same class in F^c_{mem} and $F^{m,n}_{mem}$, the classification consistency is performed by calibrating the recognition results of F^c_{mem} using the recognition results of $F^{m,n}_{mem}$. As can be seen in Fig. 4, for each sample in $F^{m,n}_{mem}$, losses of each sample are calculated for each category such as face and arm. This yields n losses values for the same category and they are summed for the

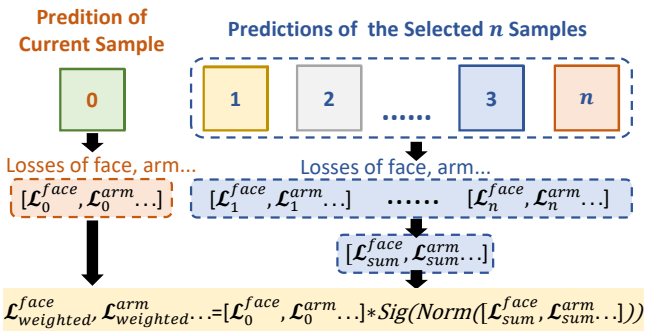

Figure 4: Memory-based classification consistency.

corresponding category, aggregating the losses from different samples. Subsequently, all summed losses are normalized and mapped to weights. The weights are multiplied to the loss of each category in F_{mem}^{c}. Consequently, the recognition of each category in F_{mem}^{c} is calibrated by the corresponding category in $F_{mem}^{m,n}$, which achieves classification consistency learning. It is formulated as follows:

$$\mathcal{L}_{cls} = CE_{cls}(F_{mem}^{m,n}, L^n) \tag{8}$$

$$\mathcal{L}_{wts} = Norm(Sum_{cls}(\mathcal{L}_{cls})) \tag{9}$$

$$\mathcal{L}_{weighted} = CE_{cls}(F_{mem}^{c}, L) \odot \mathcal{L}_{wts} \tag{10}$$

where $CE_{cls}(\cdot)$ means modified cross entropy loss function which is used for loss computation of each classes. $Sum_{cls}(\cdot)$ denotes gather the losses of the same classes from the selected samples. $Norm(\cdot)$ refers to normalization on the summed losses of different categories. F_{mem}^{c} and $F_{mem}^{m,n}$ are predictions of the samples. L and L^n are the corresponding labels.

3.5 Structure Correlation

In addition to appearance learning, structural learning is also crucial for cartoon parsing. Previous methods consider the correlation of different parts, but they overlook the learning of strongly related local limb structures. As can be seen in Fig. 6, some local limbs exhibit high correlations, such as the face, hair, and hat, which are strongly related. Similarly, the thighs, shins, and shoes, or the palms, upper arms, and forearms, are also strongly related. Therefore, this paper introduces the main body-aware recognition strategy, which focuses on structural modeling of strongly correlated limbs to enhance the capability of cartoon parsing.

Given F_{rc}, the output of representation consistency, pixel correlation learning (PCL) is performed to capture pixel correlation of the cartoon images [23]. After that, feature aggregation is performed using global average pooling along the channel dimension to obtain a one-dimensional feature representation. This one-dimensional feature representation is divided into n_c feature representations, which are treated as nodes of graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, with n_c representing the number of categories. $\mathcal{V}$ denotes the set of nodes that correspond to body parts of cartoon character and $\mathcal{E}$ refers to the set of edges that indicates the connection relations of adjacent body parts. For the construction of edge $\mathcal{E}$, the adjacency matrix that indicates the connections of semantic parts is computed to associate body parts, as can be seen in Fig. 5.

GAT layers are constructed to process the graph $\mathcal{G}$. Specifically, the network first computes the attention score a_{ij} between nodes i

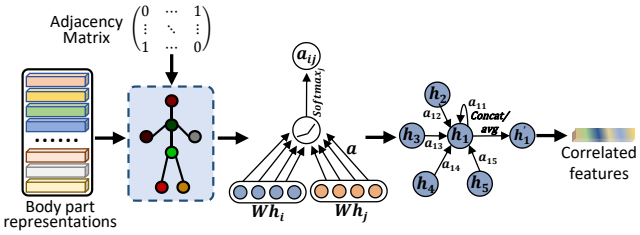

Figure 5: GAT-based structure correlation process.

and j. For a node i in graph $\mathcal{G}$, the attention over its neighbor j is defined as:

$$a_{ij} = \text{LeakyReLU}(a^T Cat(Wh_i, Wh_j)) \tag{11}$$

where a is a learnable vector. h_i and h_j are the feature vectors of nodes i and j. LeakyReLU is the activation function used to ensure non-linearity, and the concatenation allows the model to consider the features of both node i and j. The features are transformed with weight matrix W, learning to weigh the importance of different node pairs.

After that, the softmax function is applied to the attention scores, normalizing them to form weights α_{ij}. The attention coefficient α_{ij} represents the importance of node j when updating the representation of node i. The coefficient from node j to node i is calculated as,

$$\alpha_{ij} = \frac{\exp(a_{ij})}{\sum_{k \in \mathcal{N}(i)} \exp(a_{ik})} \tag{12}$$

where $\mathcal{N}(i)$ is the set of neighbors of node i.

The new representation of node i, h_i', is calculated as a weighted sum of the transformed features of its neighbors, with weights given by the attention coefficients α_{ij}. This allows each node to selectively attend to its neighbors, capturing the importance of different connections in the graph. It is formulated as,

$$h_i' = \sigma(\sum_{j \in \mathcal{N}(i)} \alpha_{ij} W h_j) \tag{13}$$

where σ denotes to a nonlinearity operation [31].

The updated node representation h_i' is obtained by employing multi-head attention [30]. Specifically, S independent attention mechanisms perform the transformation described in Equation 13. Their resulting features are then concatenated to produce the following output feature representation as,

$$h_i' = Cat(\sigma(\sum_{j \in \mathcal{N}(i)} \alpha_{ij}^0 W h_j), ..., \sigma(\sum_{j \in \mathcal{N}(i)} \alpha_{ij}^S W h_j)) \tag{14}$$

where $\alpha_{ij}^s (l \in [0, 1, 2, ..., S])$ refers to the normalized attention coefficients computed by the k-th attention mechanism.

Unlike the typical GAT, the GATs built is hierarchical, which are designed to process different structural levels. Firstly, a GAT layer processes the pre-constructed graph structure, updates all the node data, and returns the results. Subsequently, the subgraph structures are built by merging strongly related nodes among the n_c nodes, and then apply GAT to process the strongly related nodes and return the results. As can be seen in Fig. 6, the graph of the left hand, left arm and torso is treated as a main body graph because they are strongly connected through the left arm. By feeding the paired nodes associated with the left arm into the GAT for learning, the information related to the left arm is emphasized. This attention to

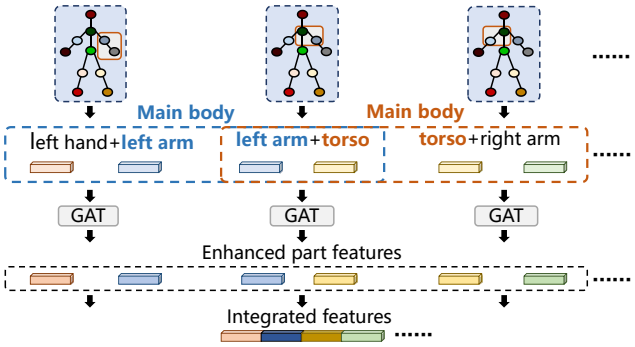

Figure 6: Main body-aware correlation.

the main body structure, including the left hand, left arm, and torso centered around it, facilitates the recognition of this specific region. Later, the learning results of various main bodies are amalgamated, and the integrated results are combined with the initially processed results from the global graph, producing the ultimate output. The output of the structure correlation is fused with the input features, resulting in a stronger structural feature representation that is used for more accurate cartoon parsing, which is formulated as,

$$h_i' = SC_{GATs}(Divide(Gap(PCL(F_{rc})))) \tag{15}$$

$$F_{rc-main} = Head(F_{rc} \oplus h_i', PCM(PCL(F_{rc}))) \tag{16}$$

where $SC_{GATs}(\cdot)$ denotes the structure correlation function based on GATs. $Divide(\cdot)$ is the operation dividing the one-dimensional feature representation into n_c nodes. $Head(\cdot)$ means a stacked module of concatenation and a convolution layer. $PCM(\cdot)$ denotes part correlation modeling [23].

3.6 Loss Functions

The objective function of this paper is composed of the proposed classification consistency loss, pixel correlation loss, part correlation loss, spatial loss and edge loss, which is formulated as,

$$\begin{aligned}\mathcal{L}_{obj} = \mathcal{L}_{cc} + CE(F^c_{mem}, L) + CE(F_{pcl}, L) + \\ CE(F_{rc-main}, L) + \mathcal{L}_{spatial} + \mathcal{L}_{edge}\end{aligned} \tag{17}$$

where $CE(\cdot)$ is the cross entropy loss. $CE(F^c_{mem}, L)$ is adopted to predict segmentation results using features from the ConvNet. $CE(F_{pcl}, L)$ and $CE(F_{rc-main}, L)$ are used to learn pixel correlations [23] and structure correlations. F_{pcl} is obtained by using a convolution layer to map the output of PCL into predictions. $\mathcal{L}_{spatial}$ and $\mathcal{L}_{edge}$ are utilized to learn spatial information [20, 23, 28] and edge information [14, 26]. $\mathcal{L}_{cc}$ is the proposed classification consistency loss and it is defined as follows:

$$\begin{aligned}\mathcal{L}_{cc} = (CE_{cls}(F^c_{mem}, L) + CE_{cls}(F_{pcl}, L) + \\ CE_{cls}(F_{rc-main}, L)) \odot \mathcal{L}_{wts}\end{aligned} \tag{18}$$

4 Experiments

4.1 Datasets and Evaluation Metrics

CartoonSet. The CartoonSet dataset [23] comprises 2,229 cartoon images, which are divided into 1,530 images for training, 510 for testing, and 189 for validation. This diverse dataset includes samples of children's drawings, illustrations with brief strokes, animated characters, and more. The various body parts of the cartoon images are annotated and categorized into 24 classes, including categories like hair, hat, face, and neck.

Cartoon dog. The cartoon dog dataset [32] contains 965 images of cartoon dogs. It consists of 773 images for training and 192 images for testing. This dataset focuses on cartoon dog parsing, with eight distinct classes for labeling, including categories such as the head and body.

LIP. LIP [8] is a human parsing dataset with 50,462 single-person images. This dataset is divided into 30,462 training images, 10,000 testing images, and an additional 10,000 images for validation. It provides annotations for 19 classes, which include categories such as face and right arm.

Evaluation Metrics. Following previous approaches [15, 23, 32], the utilized evaluation metric is the Mean Intersection over Union (Mean IoU). Mean IoU computes the average intersection-to-union ratio between the predicted body parts and ground truth. Additionally, Mean Accuracy (Mean Acc.) is employed to calculate per-class accuracy and Pixel Accuracy (Pixel Acc.) is used to compute the accuracy of the correctly predicted pixels in each class.

4.2 Implementation Details

The Stochastic Gradient Descent (SGD) is employed as the optimizer with a momentum of 0.9 and a weight decay of 5e-4. Following previous methods [3, 7, 21, 22, 38, 39], the "poly" learning rate strategy is utilized, defined as $lr = lr_{base} \times (1 - \frac{C_n}{T_n})^{power}$, where lr is the current learning rate, lr_{base} is the base learning rate, and $power = 0.9$. C_n and T_n represent the current iteration number and the total iteration number, respectively. As for data augmentation, random left-right flipping with a 0.5 probability and random scaling within the range of [0.25, 1.0] are applied.

To ensure a fair comparison on the CartoonSet, CartoonNet is trained with a batch size of 8 and image size of 384×384. The number of training epochs is 150 and the learning rate for CartoonSet is set to 7e-3. For the cartoon dog dataset, CartoonNet uses a learning rate of 1e-3, and the training images are resized to 384×384. All the methods adopt single-scale evaluation on the CartoonSet and the cartoon dog dataset. As for the LIP dataset, CartoonNet is trained with a learning rate of 7e-3, using training images resized to 473×473. The number of training epochs is 150. For the LIP dataset, the multi-scale evaluation approach [14, 15] is employed.

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

Table 1 presents a comparative analysis of the proposed method with state-of-the-art approaches like SCHP [14] and CPNet [23]. Human parsing methods exhibit limitations when applied to cartoon characters, mainly due to the differens between the abstract cartoon images and the real-world humans. Cartoon characters possess more intricate, variable, and irregular shapes and structures compared to humans, which is challenging for existing human parsing techniques. DeepLabV3+ [4], a model not specialized for cartoon parsing, demonstrates subpar performance in this domain. DFPNet [32], though dedicated to cartoon parsing, primarily focuses on scale variations, resulting in limited performance improvements. Methods like HHP [35] and CNIF [34] are designed to model

Table 1: Comparison on the CartoonSet.

Method	Pixel Acc.	Mean Acc.	Mean IoU
DeepLabV3+ [4]	87.28	63.42	50.12
DFPNet [32]	88.21	65.33	51.71
HHP [35]	87.51	64.42	51.98
CE2P [26]	88.06	66.55	52.90
CNIF [34]	87.74	66.01	53.21
CDGNet [15]	88.11	67.98	53.99
SCHP [14]	88.63	69.36	55.44
CPNet [23]	89.42	69.90	57.02
CartoonNet (Ours)	**89.61**	**70.59**	**57.84**

Table 2: Comparison on the cartoon dog dataset.

Method	Pixel Acc.	Mean Acc.	Mean IoU
Mask R-CNN [9]	89.21	57.78	50.56
CE2P [26]	92.63	76.43	65.32
DFPNet [32]	93.50	79.40	68.39
SCHP [14]	94.05	81.15	71.22
CDGNet [15]	94.09	80.03	71.44
CPNet [23]	94.32	82.60	72.28
CartoonNet (Ours)	**94.64**	**84.01**	**74.05**

the hierarchical body structures of real humans and excel in human parsing tasks. However, their performances are limited when segmenting the irregular and complex body structures of cartoon characters [23]. CE2P [26] leverages global context information and edge details for parsing but fails to account for appearance diversity and structure complexity of cartoon images, resulting in limited effectiveness. CDGNet [15] uses class distribution as spatial constraints, outperforming most of the methods. But it ignores the structure modeling and learning. SCHP [14] achieves improved results using a self-correction strategy, which helps differentiate visually complex body parts, making it suitable for cartoon parsing. Nonetheless, it still faces challenges in addressing the diversity and complexity of cartoon characters. CPNet [23] significantly improves the performance of cartoon parsing by addressing the irregular structures encountered in cartoon parsing with its pixel and part correlation learning structure. However, CPNet [23] mainly focuses on single cartoon character, which makes it less effective in recognizing a diverse range of cartoon characters.

CartoonNet is designed to address the challenges posed by diverse and complex cartoon characters in the field of cartoon parsing. The performance of CartoonNet, as demonstrated in Table 1, surpasses the state-of-the-art human parsing methods such as CDGNet [15] and cartoon parsing methods including DFPNet [32] and CPNet [23]. This notable result is attributed to CartoonNet's comprehensive exploration of consistency across different cartoon images and the correlation of main bodies.

Furthermore, when tested on the cartoon dog dataset, CartoonNet stands out as the top-performing method, achieving the optimal performance, as presented in Table 2. CartoonNet's specific design effectively alleviates the challenges posed by the diverse visual appearances and complex structures of cartoon characters. The results showcased in Table 2 serve as compelling evidence of the efficacy of the proposed method.

To verify the generalization ability of CartoonNet, it is also tested on the human parsing dataset LIP. Table 3 illustrates the comparative results, where CartoonNet outperforms the cartoon parsing method CPNet [23] significantly in terms of the performance on

Table 3: Comparison on the LIP dataset.

Method	Pixel Acc.	Mean Acc.	Mean IoU
DeepLabV3+ [4]	n/a	n/a	52.09
CE2P [26]	87.37	63.20	53.10
CorrPM [43]	87.68	67.21	55.33
OCR (ResNet101) [40]	n/a	n/a	55.60
HRNetV2 [33]	n/a	n/a	55.90
OCR (HRNetV2-W48) [40]	n/a	n/a	56.65
CPNet [23]	88.29	68.41	57.21
CNIF [34]	88.03	68.80	57.74
HHP [35]	**89.05**	70.58	59.25
SCHP [14]	n/a	n/a	59.36
CDGNet [15]	88.86	**71.49**	**60.30**
CartoonNet (Ours)	88.59	69.57	58.27

the LIP dataset. This demonstrate the effectiveness of the proposed method in human parsing. Compared to CartoonNet, HHP [35] achieves better performance on LIP dataset with its tailored GNN components designed for regular human bodies. However, compared to HHP [35], CartoonNet achieves better performance in recognizing the intricate and irregular structures of cartoon characters. SCHP [14] surpasses CartoonNet on LIP, benefiting from its self-correction strategy and label refinement process designed for human parsing. But CartoonNet outperforms SCHP [14] on both the CartoonSet and the cartoon dog dataset. CDGNet [15] excels in human parsing by learning the class distribution of real humans, achieving the optimal results on LIP. However, its performance in cartoon parsing remains limited. Despite methods like OCR (HRNetV2-W48) and HRNetV2 have achieved outstanding performance on segmentation datasets such as Cityscapes [6], their performance on LIP is limited due to the lack of specific designs for human parsing. Compared to these methods, the performance of CartoonNet on the LIP dataset is unexpected, which surpasses OCR (HRNetV2-W48) [40] and HRNetV2 [33] by a large margin. The performance of CartoonNet on the LIP dataset demonstrates that the proposed approach with semantic consistency and structure correlation is effective for human parsing as well.

4.4 Ablation Studies

To validate the effectiveness of the proposed method, ablation studies are conducted, as listed in table 4. The incorporation of representation consistency and classification consistency leads to a marked improvement in the Mean IoU metric, as compared to the baseline approach. The performance is improved because the proposed memory-based consistency learning strategy facilitates the semantic consistency of the body parts that belong to the same class but from diverse cartoon images. It enhances the generalization ability of the model, allowing it to correctly parse diverse cartoon images. The incorporation of the main body-aware structure correlation also leads to performance improvement, which helps the network to learn the main bodies. The integration of the memory-based consistency module and the main body-aware structure correlation achieves the optimal results, confirming that the proposed method is effective in addressing challenges posed by significant individual variations and complex structure correlations in cartoon parsing.

To further validate the effectiveness of the proposed modules, ablation studies on the cartoon dog dataset are also conducted. As

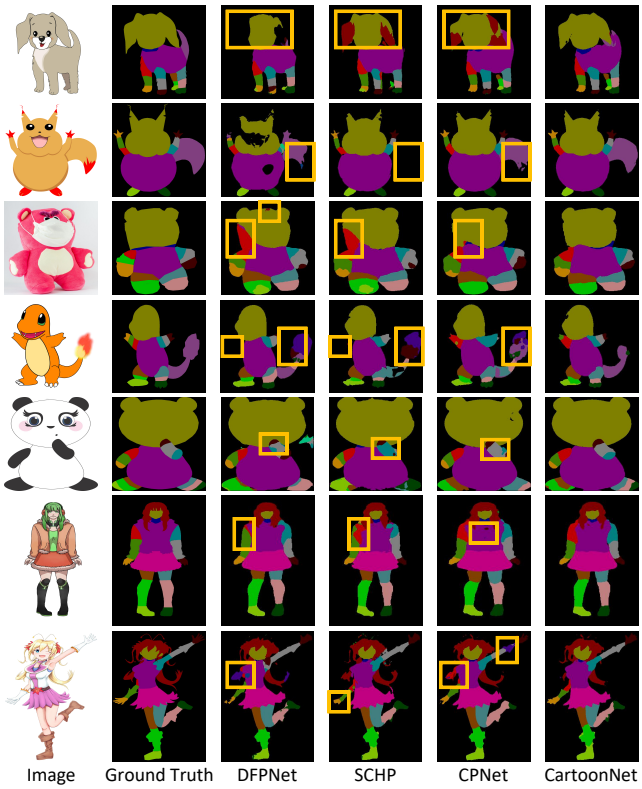

Figure 7: Visualization comparisons on the CartoonSet.

Table 4: Effect of the proposed modules on CartoonSet dataset.

Method	Pixel Acc.	Mean Acc.	Mean IoU
Baseline	89.19	68.91	56.12
Baseline + RC	89.33	70.24	57.11
Baseline + CC	89.51	69.71	57.08
Baseline + RC + CC	89.53	**71.00**	57.36
Baseline + Main	89.48	69.48	57.25
Baseline + RC + CC + Main	**89.61**	70.59	**57.84**

listed in table 5, compared to the baseline model, the proposed modules achieve significant performance improvements. CartoonNet with all the proposed modules significantly improves the Mean IoU metric based on the baseline, which proves the effectiveness of the proposed approach.

Ablation studies on the memory structure are conducted, involving ablations of the size of the memory structure and the number of the selected similar samples. As listed in the table 6, CartoonNet + M400 represents training CartoonNet with a memory storage size of 400 samples. 1530 is the number of images of training set in the CartoonSet dataset. It can be observed from the table that the storage size of the memory structure has little impact on the accuracy. CartoonNet + N3 represents selecting three similar samples in each iteration, and the data in the table indicates that the best result is obtained when the number of the selected samples is 5.

4.5 Qualitative Evaluation

To qualitatively evaluate the proposed method, Fig. 7 showcases the visual comparison between CartoonNet and the state-of-the-art approaches. For diverse cartoon characters, existing methods tend to make recognition errors when segmenting body parts that belong to the same class but exhibit significant appearance and structure differences. CPNet [23] have difficulties in recognizing diverse cartoon images as it primarily focuses on internal pixel and part correlation within a single cartoon character, which ignores the semantic consistency learning across different cartoon characters. Moreover, existing methods have suboptimal performance when the targets have intricate and irregular structures, as they lack specialized recognition structures for the irregularities in cartoon characters. CPNet [23] performs better among the previous methods by considering the irregularities of cartoon characters. However, it still falls short in recognizing complex structures as it doesn't emphasize the core and main bodies. The proposed method introduces the main body-aware strategy, addressing the recognition issues related to complex structures.

Table 5: Effect of the proposed modules on Cartoon Dog dataset.

Method	Pixel Acc.	Mean Acc.	Mean IoU
Baseline	94.09	82.22	72.13
Baseline + RC	94.47	83.05	73.24
Baseline + CC	94.37	81.91	73.09
Baseline + RC + CC	94.50	83.19	73.69
Baseline + Main	94.44	82.69	72.77
Baseline + RC + CC + Main	**94.64**	**84.01**	**74.05**

Table 6: Ablation studies of the memory size and the number of selected samples.

Method	Pixel Acc.	Mean Acc.	Mean IoU
CartoonNet + M400	**89.61**	70.59	**57.84**
CartoonNet + M800	89.51	70.48	57.45
CartoonNet + M1200	89.56	**70.94**	57.31
CartoonNet + M1530	89.41	70.45	57.67
CartoonNet + N3	89.49	70.04	57.15
CartoonNet + N5	**89.61**	70.59	**57.84**
CartoonNet + N7	89.67	**70.85**	57.47
CartoonNet + N9	89.51	69.64	57.40

5 Conclusion

In this paper, a cartoon parsing method named CartoonNet is proposed to address the challenges of diverse appearances and complex structures in cartoon characters. It primarily consists of a consistency learning strategy based on memory structure and a main body-aware structure correlation module. The memory-based consistency learning achieves semantic consistency among diverse cartoon characters, which explores deep relations of the body parts that have different appearances but belong to the same class. The main body-aware structure correlation enhances the recognition capability of complex structures by focusing on the core body parts. By combining the memory-based consistency learning with the main body-aware structure correlation strategy, the proposed method achieves state-of-the-art results in cartoon parsing.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (Grant No. 62372387), Key R&D Program of Guangxi Zhuang Autonomous Region, China (Grant No. AB22080038, AB22080039).

References

[1] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. 2017. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39, 12 (2017), 2481–2495.

[2] Shaked Brody, Uri Alon, and Eran Yahav. 2022. How Attentive are Graph Attention Networks?. In *International Conference on Learning Representations*.

[3] 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.

[4] 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.

[5] 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.

[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] 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.

[8] 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.

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

[10] 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.

[11] Adrián Javaloy, Pablo Sánchez-Martín, Amit Levi, and Isabel Valera. 2023. Learnable Graph Convolutional Attention Networks. In *International Conference on Learning Representations*.

[12] 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.

[13] Jin Kim, Jiyoung Lee, Jungin Park, Dongbo Min, and Kwanghoon Sohn. 2022. Pin the Memory: Learning to Generalize Semantic Segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*. 4340–4350.

[14] 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.

[15] 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.

[16] Yunan Liu, Liang Zhao, Shanshan Zhang, and Jian Yang. 2020. Hybrid Resolution Network Using Edge Guided Region Mutual Information Loss for Human Parsing. In *ACM International Conference on Multimedia*. 1670–1678.

[17] Jonathan Long, Evan Shelhamer, and Trevor Darrell. 2015. Fully Convolutional Networks for Semantic Segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*. 3431–3440.

[18] 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.

[19] 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.

[20] Tomas Pfister, James Charles, and Andrew Zisserman. 2015. Flowing ConvNets for Human Pose Estimation in Videos. In *IEEE International Conference on Computer Vision*. 1913–1921.

[21] 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.

[22] 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.

[23] 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.

[24] Tiago Ramalho and Marta Garnelo. 2019. Adaptive Posterior Learning: few-shot learning with a surprise-based memory module. In *International Conference on Learning Representations*.

[25] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention*. 234–241.

[26] 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.

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

[28] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. 2019. Deep High-Resolution Representation Learning for Human Pose Estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*. 5693–5703.

[29] 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.

[30] 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.

[31] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph Attention Networks. In *International Conference on Learning Representations*.

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

[33] 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.

[34] 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.

[35] 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.

[36] Zhonghua Wu, Xiangxi Shi, Guosheng Lin, and Jianfei Cai. 2021. Learning Meta-class Memory for Few-Shot Semantic Segmentation. In *IEEE International Conference on Computer Vision*. 497–506.

[37] Guo-Sen Xie, Huan Xiong, Jie Liu, Yazhou Yao, and Ling Shao. 2021. Few-Shot Semantic Segmentation with Cyclic Memory Network. In *IEEE International Conference on Computer Vision*. 7273–7282.

[38] 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.

[39] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. 2018. Learning a Discriminative Feature Network for Semantic Segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*. 1857–1866.

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

[41] 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.

[42] 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.

[43] 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.