

CAPNet: Cartoon Animal Parsing with Spatial Learning and Structural Modeling (Supplementary Materials)

Anonymous Author(s)

In this document, more information about the CASet and more experimental results not put in the original manuscript due to page limitation are provided.

1 MORE INFORMATION ABOUT THE CASET

In Fig.1, the number of image samples for different body parts such as heads, wings, and tails within the CASet are provided. It can be observed that the heads and bodies have the highest number of samples, as these categories are present in the majority of animals. In contrast, the number of samples for wings and tails is relatively lower, as wings are only present in certain animals such as birds, mosquitoes, and bees, and tails sometimes are occluded from view. Other categories, such as arms and legs, have a more balanced number of image samples.

Additionally, the supplementary materials provide the categories of the 52 animal species in the CASet, along with the sample counts for training, testing, and validation sets for each species, as detailed in Fig. 2, Fig. 3, Fig. 4, and Fig. 5. From these figures, it is evident that the CASet encompasses a diverse array of animal categories, including mammals such as elephants, birds such as chickens and hawks, reptiles such as alligators and lizards, and other cartoon animal categories. Each animal category comprises dozens of diverse images sourced from several search engines including Pixabay, Google, Baidu and Bing.

2 PER CLASS PARSING RESULTS ON THE CASET

In Table 1, the per class IoU and Mean IoU of different methods on the CASet are provided, which help us know more about the performance of different methods on cartoon animal parsing. CAPNet outperforms the state-of-the-art methods on most of the categories, including head, wing, left/right legs, etc. CAPNet outperforms other methods by a large margin in categories with significant complexity and diversity among different animals. Legs, for instance, vary greatly among animal categories and exhibit complex spatial structures in animals with multiple legs or irregular leg structures. This performance gap demonstrates the effectiveness of CAPNet in capturing and learning spatial and structural information, which is crucial for accurate cartoon animal parsing.

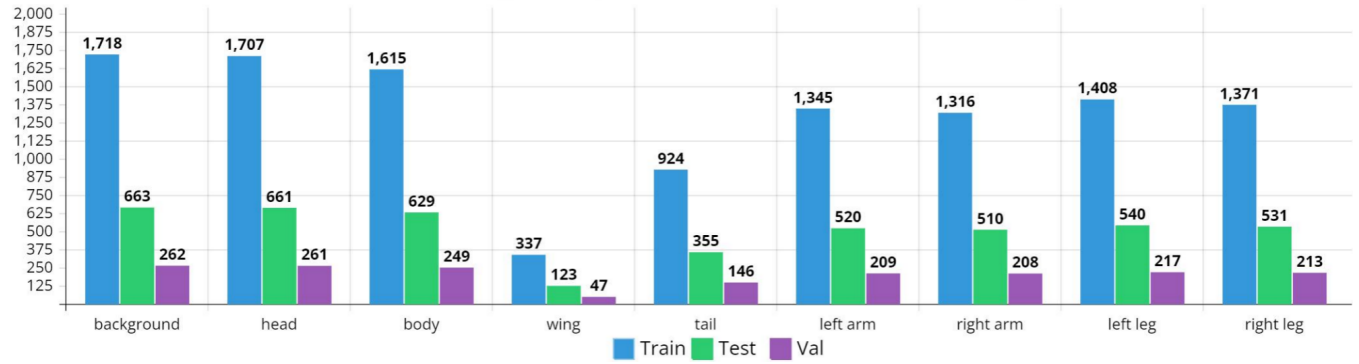
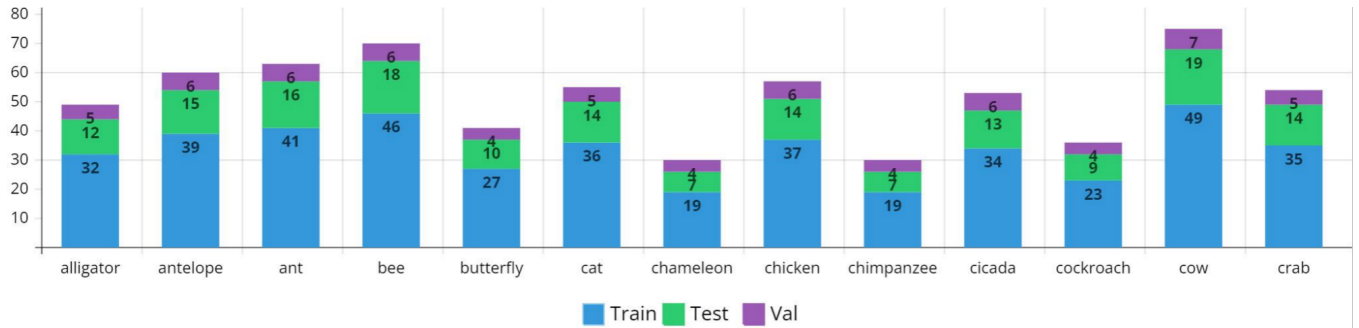
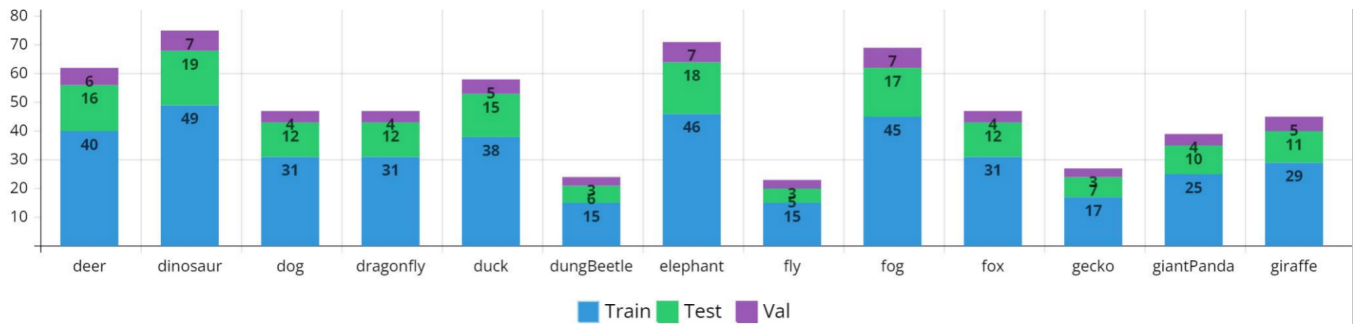
REFERENCES

- [1] 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.
- [2] 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.
- [3] 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.
- [4] 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.

- [5] 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.
- [6] Jerome Wan, Guillaume Mougeot, and Xubo Yang. 2020. Dense feature pyramid network for cartoon dog parsing. *The Visual Computer* 36, 10 (2020), 2471–2483.
- [7] 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.

Table 1: Comparison on the CASet. BG means the class of background.

Method	BG	Head	Body	Wing	Tail	Left Arm	Right Arm	Left Leg	Right Leg	Mean IoU
DeepLabV3+ [1]	95.26	87.45	70.16	79.56	63.96	56.60	60.00	58.27	56.43	69.74
DFPNet [6]	96.74	88.25	70.75	81.32	66.96	57.69	57.29	60.13	57.52	70.74
CE2P [5]	96.40	88.65	71.20	79.41	67.08	54.76	58.09	62.46	59.25	70.81
HHP [7]	96.24	85.06	72.06	76.78	63.83	62.59	61.27	60.69	59.03	70.84
SCHP [2]	96.37	88.29	69.48	79.40	68.57	59.78	63.14	64.59	61.26	72.32
CDGNet [3]	96.24	89.19	70.94	81.58	70.50	61.34	63.58	66.21	61.30	73.43
CPNet [4]	96.60	88.46	71.60	79.40	71.34	63.01	62.36	66.45	62.98	73.58
CAPNet (Ours)	96.58	89.81	71.72	82.60	72.04	62.67	62.00	69.14	64.58	74.57

**Figure 1: Number of Image Samples of Different Body Parts.****Figure 2: Number of Image Samples of Different Animal Categories (The first 13 categories).****Figure 3: Number of Image Samples of Different Animal Categories (The second 13 categories).**

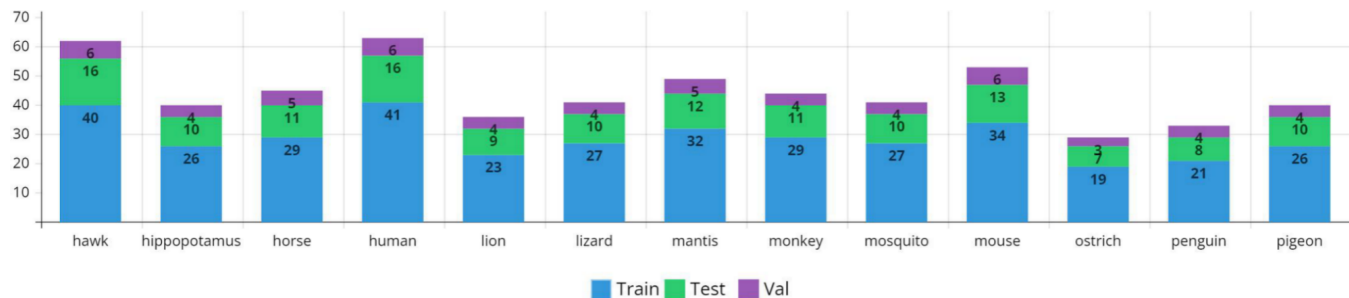


Figure 4: Number of Image Samples of Different Animal Categories (The third 13 categories).



Figure 5: Number of Image Samples of Different Animal Categories (The last 13 categories).