Dictionary-based Framework for Interpretable and Consistent Object Parsing

Anonymous CVPRW submission

Paper ID *****

Abstract

*In this work, we present CoCal, an interpretable and consistent object parsing framework based on dictionary-based mask transformer. Designed around **Co**ntrastive **C**omponents and Logic**al** Constraints, CoCal rethinks existing cluster-based mask transformer architectures used in segmentation; Specifically, CoCal utilizes a set of dictionary components, with each component being explicitly linked to a specific semantic class. To advance this concept, CoCal introduces a hierarchical formulation of dictionary components that aligns with the semantic hierarchy. This is achieved through the integration of both within-level contrastive components and cross-level logical constraints. Concretely, CoCal employs a component-wise contrastive algorithm at each semantic level, enabling the contrasting of dictionary components within the same class against those from different classes. Furthermore, CoCal addresses logical concerns by ensuring that the dictionary component representing a particular part is closer to its corresponding object component than to those of other objects through a cross-level contrastive learning objective. To further enhance our logical relation modeling, we implement a post-processing function inspired by the principle that a pixel assigned to a part should also be assigned to its corresponding object. With these innovations, CoCal establishes a new state-of-the-art performance on both PartImageNet and Pascal-Part-108, outperforming previous methods by a significant margin of 2.08% and 0.70% in part mIoU, respectively. Moreover, CoCal exhibits notable enhancements in object-level metrics across these benchmarks, highlighting its capacity to not only refine parsing at a finer level but also elevate the overall quality of object segmentation.*

1. Introduction

Human perception involves the ability to decompose an object into its semantically meaningful components (*i.e.*, parts). For instance, when observing a dog, humans not only identify it as a dog but also simultaneously discover its head, torso, and other components, facilitating a more interpretable and resilient understanding of real-world scenarios. More specifically, humans can estimate the pose of a dog by considering the spatial arrangement of its parts, even in instances where some parts may be missing.

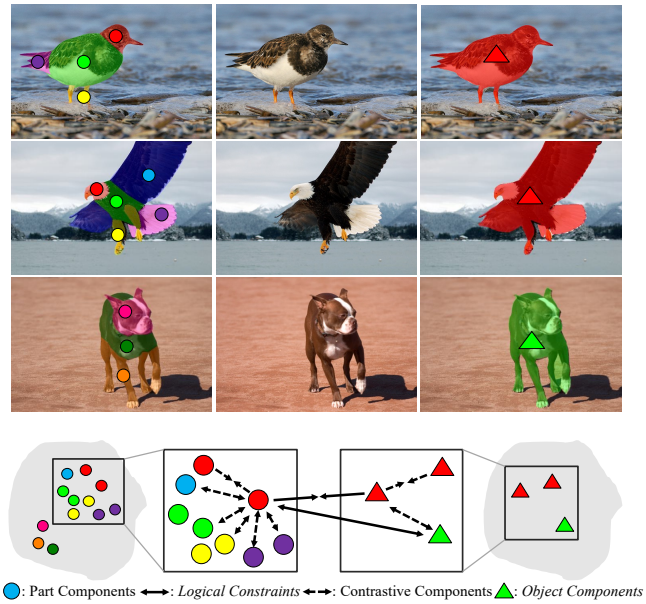


Figure 1. **Illustration of the proposed component-wise contrastive objectives**. CoCal establishes two discriminative dictionaries at the part and object levels. Within the same semantic level, part/object components of the same classes are pulled closer (→←), while those of different classes are pushed apart (←→) (*i.e.*, contrastive components). At the cross-semantic level, part components and their corresponding object components are pulled closer and vice versa (*i.e.* logical constraints).

By contrast, emulating this innate human visual capability presents a big challenge for modern computer vision models. The predominant focus within the field has been on addressing semantic segmentation at the object level, with minimal attention given to intermediate part representations. Notable works [15, 34, 48, 49] in object parsing primarily extend algorithms designed for general segmentation, overlooking the fact that parts, being at a lower seman-

tic level, can be captured more efficiently and interpretably through clustering. As a result, these works often adhere to frameworks tailored for object segmentation without incorporating specialized designs for handling parts. Moreover, even though certain studies [18, 22, 61] highlight the mutual benefit between object parsing and object segmentation, they typically treat these semantic levels separately, disregarding the logical relationship between them. Consequently, the optimization objectives for these two levels are disjoint, leading to sub-optimal predictions.

In this work, we propose CoCal, a dictionary-based framework built on top of an off-the-shelf cluster-based mask transformer, utilizing a set of dictionary components where each component is explicitly associated with a specific semantic class to facilitate the grouping of pixels belonging to that class. This enables CoCal to conduct inference in a straightforward parameter-free manner through nearest neighbor search on the pixel feature maps within the class dictionary. Taking this concept further, CoCal introduces a hierarchical formulation of dictionary components, aligning with the semantic hierarchy, which naturally forms the logical paths within the structure (*e.g.*, bird-head $\rightarrow$ bird). CoCal advances the learning of the above formulation through two simple yet effective targets: learning contrastive objectives for obtaining discriminative dictionary components and exploring logical relations for consistent predictions. Specifically, as depicted in Fig. 1, at each semantic level, CoCal employs a component-wise contrastive algorithm to pull closer the dictionary components withing the same class while pushing away those from different classes. Then to model the cross-level logical relations, CoCal further contrasts the positive pair between dictionary component representing a particular part and its corresponding object dictionary components against the negative pairs involving the part component and all other object components. In addition, CoCal applies a post-processing step enforcing that any pixel predicted as a certain part must also be labeled under its corresponding object class. Specifically, CoCal calculates a path probability by multiplying part- and object-level similarities, then assigns pixels to the path with the highest score. This mechanism captures cross-level semantics and resolves inconsistencies at inference.

2. Method

In this section, we begin by introducing the core principles behind mask transformer segmentation frameworks, focusing on cluster-based methods and showing how our proposed dictionary-based formulation evolves from these ideas. We then describe how CoCal leverages hierarchy, contrastive components, and logical constraints to achieve interpretable and consistent object parsing. Finally, we present the meta-architecture of our method, detailing all major components and their interactions.

2.1. Dictionary-based Mask Transformers

Problem Statement Semantic segmentation aims to partition an image $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$ into non-overlapping masks, where each mask is associated with a semantic label. Concretely, the ground-truth set of masks is:

$$\{y_i\}_{i=1}^{M} = \{(d_i, c_i)\}_{i=1}^{M}, \tag{1}$$

where $d_i \in \{0,1\}^{H\times W}$ is a binary mask denoting the pixels of the i-th region, c_i is its class label, and M represents the number of ground-truth masks. A typical segmentation model outputs a set of N predicted masks ($N \geq M$) and their corresponding classes:

$$\{\hat{y}_i\}_{i=1}^{N} = \{(\hat{m}_i, \hat{c}_i)\}_{i=1}^{N}. \tag{2}$$

Recap of Cluster-based Mask Transformers Cluster-based mask transformers [37, 71, 72] differ from standard query-based transformers in their use of a one-hot argmax assignment instead of a $\mathrm{softmax}$ for updating the query features. Specifically, let $\mathbf{O} \in \mathbb{R}^{N\times D}$ denote the N object queries and $\hat{\mathbf{O}}$ the updated queries. Similarly, let $\mathbf{Q}^o, \mathbf{K}^p, \mathbf{V}^p$ be the linearly projected features for queries, keys, and values. Then, instead of the $\mathrm{softmax}$ cross-attention,

$$\hat{\mathbf{O}} = \mathbf{O} + \underset{HW}{\mathrm{softmax}}\big(\mathbf{Q}^o \times (\mathbf{K}^p)^{\mathrm{T}}\big) \times \mathbf{V}^p, \tag{3}$$

the cluster-based mask transformer adopts a one-hot argmax:

$$\hat{\mathbf{O}} = \mathbf{O} + \underset{N}{\mathrm{argmax}}\big(\mathbf{Q}^o \times (\mathbf{K}^p)^{\mathrm{T}}\big) \times \mathbf{V}^p, \tag{4}$$

so each pixel is clustered to the single query with which it has the highest affinity. The prediction set $\{\hat{y}_i\}_{i=1}^{N}$ is then matched with $\{y_i\}_{i=1}^{M}$ through Hungarian Matching [30], which guides mask and classification loss computation.

Dictionary-based Formulation While cluster-based methods typically employ N object queries (often larger than M), our dictionary-based approach seeks a more direct mapping between queries and classes. We replace the learnable queries $\mathbf{O}$ with a dictionary $\mathbf{C} \in \mathbb{R}^{P\times D}$, where each of the P dictionary components serves as a *cluster center* for a specific class. Hence, there is a one-to-one correspondence between each component $\mathbf{C}_i$ and one of the P classes.

During training, the dictionary $\mathbf{C}$ updates in the same spirit as Eqs. 2–4, except that we now have a *fixed* alignment of dictionary elements to classes (so Hungarian Matching is unnecessary). At inference time, the dictionary-based mask transformer is fully parameter-free in its final assignment step, as each pixel's feature vector is classified by whichever dictionary component is closest in feature space. This streamlining removes the need for redundancies like 'void' labels and affords an inherently interpretable design: each dictionary component explicitly encodes a particular semantic class.

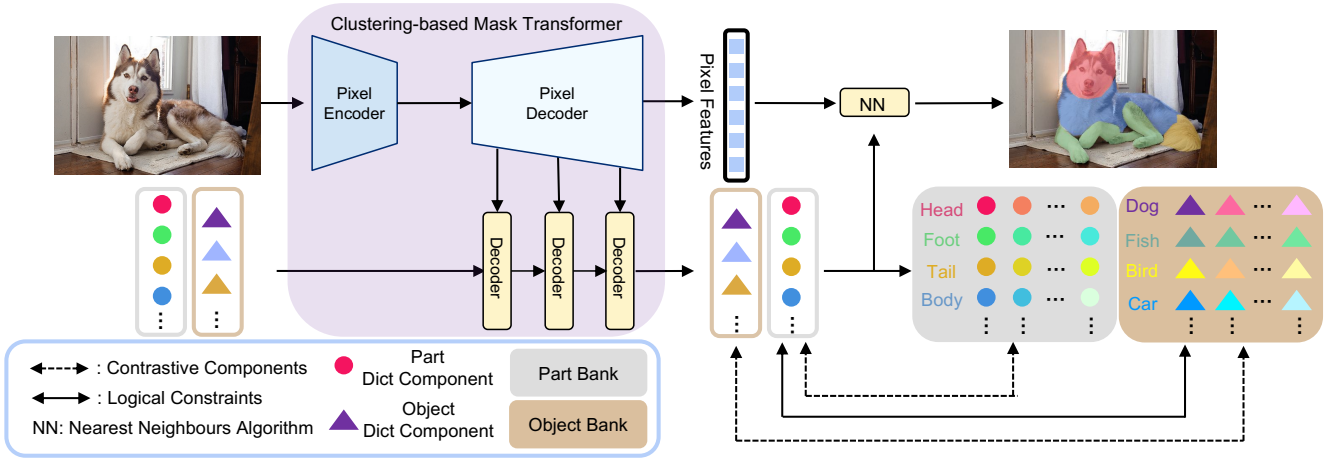


Figure 2. **Meta-architecture of the proposed CoCal.** CoCal builds on top of an off-the-shelf clustering-based mask transformer, incorporating dictionary components that function as the cluster centers for each semantic class. Throughout training, the dictionary components in CoCal are updated via both mask-wise objectives from the transformer and contrastive objectives from the dictionary. During testing, CoCal adopts a straightforward inference approach by executing nearest neighbor search of the pixel features on the dictionary components.

2.2. CoCal: Interpretable and Consistent Object Parsing

Although dictionary-based mask transformers already provide a concise per-class representation, object parsing often requires *hierarchical* reasoning (e.g., relating part classes to an object-level class) and advanced regularization to ensure consistent outputs. CoCal addresses these issues with three key ideas: (1) hierarchical dictionary components, (2) contrastive learning of dictionary elements, and (3) logical constraints at both training and inference time.

2.2.1. Hierarchical Structure of Dictionaries Across Multiple Levels

Part labels naturally imply rich logical relationships (e.g., *dog-head* is more semantically related to *dog-torso* than *fish-tail*). To exploit this hierarchy, CoCal extends the dictionary with an additional tier for object-level classes. Concretely, we have a part-level dictionary $\mathbf{C} \in \mathbb{R}^{P \times D}$ for part segmentation and an object-level dictionary $\widetilde{\mathbf{C}} \in \mathbb{R}^{\widetilde{P} \times D}$ for object classes, where $\widetilde{P}$ is the number of object classes. These two dictionaries are learned jointly to reflect the inherent relationship between objects and their parts.

2.2.2. Enhancing Dictionary Discrimination via Contrastive Objectives

To ensure discriminative power, CoCal applies contrastive learning on both part- and object-level dictionaries. Consider the part dictionary $\mathbf{C}$. We maintain a *part memory bank* $\mathbf{B} \in \mathbb{R}^{P \times S \times D}$, where S is the number of stored "samples" per class. After retrieving the relevant components $\mathbf{C}(y)$ for classes present in a training sample, we compute the contrastive loss as:

$$\mathcal{L}_{p\_con}(\mathbf{C}(y)) = \sum_{x \in M} \frac{-1}{|\mathbf{B}(x)|} \sum_{j \in \mathbf{B}(x)} \log\left(\frac{\exp(\mathbf{C}(y)_i \cdot B_j / \tau)}{\sum_{k \in \mathbf{B}} \exp(\mathbf{C}(y)_i \cdot B_k / \tau)} \right), \quad (5)$$

where $\mathbf{B}(x)$ are the stored features in $\mathbf{B}$ belonging to the same class x, B_j is one such feature, and τ is the temperature. Inspired by [41, 52], we employ hard negative mining to focus on the most challenging negatives. An analogous memory bank $\widetilde{\mathbf{B}}$ and contrastive loss $\mathcal{L}_{o\_con}$ are used for the object-level dictionary $\widetilde{\mathbf{C}}$:

$$\mathcal{L}_{o\_con}(\widetilde{\mathbf{C}}(y)) = \sum_{x \in M} \frac{-1}{|\widetilde{\mathbf{B}}(x)|} \sum_{j \in \widetilde{\mathbf{B}}(x)} \log\left(\frac{\exp(\widetilde{\mathbf{C}}(y)_i \cdot \widetilde{B}_j / \tau)}{\sum_{k \in \widetilde{\mathbf{B}}} \exp(\widetilde{\mathbf{C}}(y)_i \cdot \widetilde{B}_k / \tau)} \right), \quad (6)$$

2.2.3. Logical Constraints for Consistent Predictions

Even with hierarchical dictionaries, inconsistent predictions can arise—for example, labeling a pixel *snake-head* at the part level but *reptile* at the object level. CoCal incorporates two logical constraints to maintain cross-level consistency.

Cross-Level Contrastive Loss. First, parts belonging to the same object should be closer to *their* object-level component than to other objects. We encode this by an additional cross-level contrastive term:

$$\mathcal{L}_{logic}(\mathbf{C}(y)) = \sum_{x \in M} \frac{-1}{|\widetilde{\mathbf{B}}(x)|} \sum_{j \in \widetilde{\mathbf{B}}(x)} \log\left(\frac{\exp(\mathbf{C}(y)_i \cdot \widetilde{B}_j / \tau)}{\sum_{k \in \widetilde{\mathbf{B}}} \exp(\mathbf{C}(y)_i \cdot \widetilde{B}_k / \tau)} \right), \quad (7)$$

which brings part-level and object-level dictionary components closer when they belong to the same semantic object.

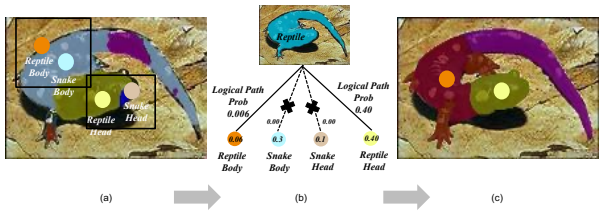


Figure 3. **Illustration of logical constraints at inference.** Here, a reptile-head and reptile-body are incorrectly predicted as snake-head and snake-body. CoCal corrects the wrong prediction by computing the logical path probability (multiplying part-level and object-level probabilities) and reassigning labels along the path to produce the correct part prediction.

Post-Processing at Inference. Second, we encode the fact that if a pixel belongs to a certain part, it must also belong to the corresponding object. CoCal thus multiplies part- and object-level probabilities and selects the top joint path (Fig. 3). This ensures if a pixel is labeled *dog-head*, it cannot simultaneously be labeled *fish* at the object level.

2.2.4. Meta-Architecture Overview

Figure 2 shows the overall pipeline of CoCal. We start with a cluster-based mask transformer backbone, which extracts pixel features. On top of these features, two dictionaries are learned: a part dictionary $\mathbf{C}$ and an object dictionary $\widetilde{\mathbf{C}}$. Two memory banks ($\mathbf{B}$ and $\widetilde{\mathbf{B}}$) store historical dictionary components, enabling within-level and cross-level contrastive training. Finally, inference proceeds via a parameter-free nearest neighbor search against both dictionaries, augmented by a logical consistency check that reassigns part labels according to the best object-part path. Therefore, CoCal delivers a highly interpretable and semantically coherent solution for object parsing.

3. Experiments

In this section, we first provide our main results on PartImageNet [21] and Pascal-Part-108 [49], followed by qualitative comparisons to highlight the effectiveness of CoCal. For extended ablation studies and additional implementation details, we refer readers to the supplementary materials(including detailed experimental settings as well as ablation studies.).

3.1. Main Results

We summarize our core findings on PartImageNet [21] and Pascal-Part-108 [49] in Table 1a and Table 1b, respectively. Notably, with a ResNet-50 [23] backbone on PartImageNet, CoCal surpasses kMaX-DeepLab [72] by 2.08% in part mIoU. When upgraded to the ConvNeXt-Tiny [42] backbone, CoCal continues to excel, reaching 70.31 part mIoU—an additional 1.79% gain over kMaX-DeepLab under the same backbone. Beyond improving part mIoU, CoCal also enhances super-category segmentation by over 1%, demonstrating robust performance at multiple semantic levels. Turning to Pascal-Part-108, CoCal achieves 49.8 part mIoU and 52.0 mAvg with a ResNet-101 backbone, thus establishing new state-of-the-art results. Compared to LOGISEG [32] and kMaX-DeepLab, CoCal offers improvements of 0.7% and 1.5% in part mIoU, respectively, alongside a notable 2.1% boost for object-level segmentation.

Table 1. PartImageNet *val* set and Pascal-Part-108 *test* set results. We report part-level and super-category/object-level mIoU (or mAvg), averaged over 3 runs.

(a) **PartImageNet *val* set**

method	backbone	mIoU	
		Part	Super-Category
DeepLabv3+ [6]	ResNet-50	60.57	-
MaskFormer [12]	ResNet-50	60.34	-
Compositor [22]	ResNet-50	61.44	-
kMaX-DeepLab [72]	ResNet-50	65.75	89.16
CoCal	ResNet-50	**67.83**	**90.41**
SegFormer [69]	MiT-B2	61.97	-
MaskFormer [12]	Swin-T	63.96	-
Compositor [22]	Swin-T	64.64	-
kMaX-DeepLab [72]	ConvNeXt-T	68.52	91.34
CoCal	ConvNeXt-T	**70.31**	**92.65**

(b) **Pascal-Part-108 *test* set**

method	Part mIoU	mAvg
SegNet [2]	18.6	20.8
FCN [44]	31.6	33.8
DeepLab [5]	31.6	40.8
DeepLabv3+ [6]	46.5	48.9
BSANet [74]	42.9	46.3
GMNet [49]	45.8	50.5
FLOAT [53]	48.0	**53.0**
HSSN [31]	48.3	-
DeepLabv3+ [6]+ LOGICSEG [32]	49.1	-
kMaX-DeepLab [72]	48.3	49.9
CoCal	**49.8**	52.0

3.2. Qualitative Results

Figure 4 shows three representative examples on PartImageNet. Compared to kMaX-DeepLab [72], CoCal produces more accurate boundaries (see the first row) and detects parts that kMaX-DeepLab misses (rows 2 & 3), demonstrating its superior ability to capture fine structures.

4. Conclusion

In conclusion, this paper introduces CoCal, an innovative model for object parsing that is rooted in a dictionary-based framework. A key aspect of CoCal is its emphasis on elucidating the intrinsic relationships between parts and objects, which significantly enhances the interpretability and consistency of parsing outcomes. This approach not only improves the accuracy of the parsing but also provides a deeper understanding of the complex interplay between part and object entities in images.

References

[1] Benjamin Alt, Minh Da Nguyen, Andreas Hermann, Darko Katic, Rainer Jaekel, Ruediger Dillmann, and Eric Sax. Efficientpps: Part-aware panoptic segmentation of transparent objects for robotic manipulation. In *ISR Europe 2023; 56th International Symposium on Robotics*, pages 131–138. VDE, 2023. 8

[2] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017. 4

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 8

[4] Liang-Chieh Chen, Yi Yang, Jiang Wang, Wei Xu, and Alan L Yuille. Attention to scale: Scale-aware semantic image segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3640–3649, 2016. 8

[5] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017. 4

[6] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. 4

[7] Mu Chen, Zhedong Zheng, Yi Yang, and Tat-Seng Chua. Pipa: Pixel-and patch-wise self-supervised learning for domain adaptative semantic segmentation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 1905–1914, 2023. 8

[8] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 8

[9] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big self-supervised models are strong semi-supervised learners. *Advances in neural information processing systems*, 33:22243–22255, 2020. 8

[10] Xianjie Chen, Roozbeh Mottaghi, Xiaobai Liu, Sanja Fidler, Raquel Urtasun, and Alan Yuille. Detect what you can: Detecting and representing objects using holistic models and body parts. In *CVPR*, pages 1971–1978, 2014. 8, 9

[11] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 8

[12] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *Advances in Neural Information Processing Systems*, 34:17864–17875, 2021. 4, 8, 10

[13] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299, 2022. 8, 10

[14] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation policies from data. *arXiv preprint arXiv:1805.09501*, 2018. 9

[15] Daan de Geus, Panagiotis Meletis, Chenyang Lu, Xiaoxiao Wen, and Gijs Dubbelman. Part-aware panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5485–5494, 2021. 1, 8

[16] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 8

[17] Ye Du, Zehua Fu, Qingjie Liu, and Yunhong Wang. Weakly supervised semantic segmentation by pixel-to-prototype contrast. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4320–4329, 2022. 8

[18] S Eslami and Christopher Williams. A generative model for parts-based object segmentation. *Advances in Neural Information Processing Systems*, 25, 2012. 2, 8

[19] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2010 (VOC2010) Results. http://www.pascal-network.org/challenges/VOC/voc2010/workshop/index.html. 8

[20] Bharath Hariharan, Pablo Arbeláez, Ross Girshick, and Jitendra Malik. Hypercolumns for object segmentation and fine-grained localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 447–456, 2015. 8

[21] Ju He, Shuo Yang, Shaokang Yang, Adam Kortylewski, Xiaoding Yuan, Jie-Neng Chen, Shuai Liu, Cheng Yang, Qihang Yu, and Alan Yuille. Partimagenet: A large, high-quality dataset of parts. In *ECCV*, pages 128–145. Springer, 2022. 4, 8, 9

[22] Ju He, Jieneng Chen, Ming-Xian Lin, Qihang Yu, and Alan L Yuille. Compositor: Bottom-up clustering and compositing for robust part and object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11259–11268, 2023. 2, 4, 8

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 4, 9

[24] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 8

[25] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 8

[26] Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q Weinberger. Deep networks with stochastic depth. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 646–661. Springer, 2016. 9

[27] Tsung-Wei Ke, Jyh-Jing Hwang, Yunhui Guo, Xudong Wang, and Stella X Yu. Unsupervised hierarchical semantic segmentation with multiview cosegmentation and clustering transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2571–2581, 2022. 8

[28] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020. 8

[29] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 9

[30] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955. 2

[31] Liulei Li, Tianfei Zhou, Wenguan Wang, Jianwu Li, and Yi Yang. Deep hierarchical semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1246–1257, 2022. 4, 8

[32] Liulei Li, Wenguan Wang, and Yi Yang. Logicseg: Parsing visual semantics with neural logic learning and reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4122–4133, 2023. 4, 8

[33] Tianyu Li, Subhankar Roy, Huayi Zhou, Hongtao Lu, and Stéphane Lathuilière. Contrast, stylize and adapt: Unsupervised contrastive learning framework for domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4868–4878, 2023. 8

[34] Xiangtai Li, Shilin Xu, Yibo Yang, Guangliang Cheng, Yunhai Tong, and Dacheng Tao. Panoptic-partformer: Learning a unified model for panoptic part segmentation. In *European Conference on Computer Vision*, pages 729–747. Springer, 2022. 1, 8

[35] Xiangtai Li, Shilin Xu, Yibo Yang, Haobo Yuan, Guangliang Cheng, Yunhai Tong, Zhouchen Lin, Ming-Hsuan Yang, and Dacheng Tao. Panopticpartformer++: A unified and decoupled view for panoptic part segmentation. *arXiv preprint arXiv:2301.00954*, 2023. 8

[36] Zhiheng Li, Wenxuan Bao, Jiayang Zheng, and Chenliang Xu. Deep grouping model for unified perceptual parsing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4053–4063, 2020. 8

[37] James Liang, Tianfei Zhou, Dongfang Liu, and Wenguan Wang. Clustseg: Clustering for universal segmentation. 2023. 2, 8

[38] Xiaodan Liang, Si Liu, Xiaohui Shen, Jianchao Yang, Luoqi Liu, Jian Dong, Liang Lin, and Shuicheng Yan. Deep human parsing with active template regression. *IEEE transactions on pattern analysis and machine intelligence*, 37(12):2402–2414, 2015. 8

[39] Xiaodan Liang, Ke Gong, Xiaohui Shen, and Liang Lin. Look into person: Joint body parsing & pose estimation network and a new benchmark. *IEEE transactions on pattern analysis and machine intelligence*, 41(4):871–885, 2018. 8

[40] Xiaodan Liang, Hongfei Zhou, and Eric Xing. Dynamic-structured semantic propagation network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 752–761, 2018. 8

[41] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 3

[42] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11976–11986, 2022. 4, 9

[43] Stuart Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137, 1982. 8

[44] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 4

[45] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018. 9

[46] Wenhao Lu, Xiaochen Lian, and Alan Yuille. Parsing semantic parts of cars using graphical models and segment appearance consistency. *arXiv preprint arXiv:1406.2375*, 2014. 8

[47] Lele Lv, Qing Liu, Shichao Kan, and Yixiong Liang. Confidence-aware contrastive learning for semantic segmentation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 5584–5593, 2023. 8

[48] Umberto Michieli and Pietro Zanuttigh. Edge-aware graph matching network for part-based semantic segmentation. *International Journal of Computer Vision*, 130(11):2797–2821, 2022. 1

[49] Umberto Michieli, Edoardo Borsato, Luca Rossi, and Pietro Zanuttigh. Gmnet: Graph matching network for large scale part semantic segmentation in the wild. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VIII 16*, pages 397–414. Springer, 2020. 1, 4, 8

[50] Shishir Muralidhara, Sravan Kumar Jagadeesh, René Schuster, and Didier Stricker. Jppf: Multi-task fusion for consistent panoptic-part segmentation. *SN Computer Science*, 5(1):187, 2024. 8

[51] Xuecheng Nie, Jiashi Feng, and Shuicheng Yan. Mutual learning to adapt for joint human parsing and pose estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 502–517, 2018. 8

[52] Joshua David Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples. In *International Conference on Learning Representations*, 2020. 3, 8

[53] Rishubh Singh, Pranav Gupta, Pradeep Shenoy, and Ravikiran Sarvadevabhatla. Float: Factorized learning of object attributes for improved multi-object multi-part scene parsing.

In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1445–1455, 2022. 4, 8, 9

[54] Yafei Song, Xiaowu Chen, Jia Li, and Qinping Zhao. Embedding 3d geometric features for rigid object part segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 580–588, 2017. 8

[55] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7262–7272, 2021. 8

[56] BLE Verboeket and Gijs Dubbelman. A hierarchical approach to part-aware panoptic segmentation. 2022. 8

[57] Changqi Wang, Haoyu Xie, Yuhui Yuan, Chong Fu, and Xiangyu Yue. Space engage: Collaborative space supervision for contrastive-based semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 931–942, 2023. 8

[58] Huiyu Wang, Yukun Zhu, Bradley Green, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Axial-deeplab: Stand-alone axial-attention for panoptic segmentation. In *European conference on computer vision*, pages 108–126. Springer, 2020. 8

[59] Huiyu Wang, Yukun Zhu, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Max-deeplab: End-to-end panoptic segmentation with mask transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5463–5474, 2021. 8

[60] Jianyu Wang and Alan L Yuille. Semantic part segmentation using compositional model combining shape and appearance. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1788–1797, 2015. 8

[61] Peng Wang, Xiaohui Shen, Zhe Lin, Scott Cohen, Brian Price, and Alan L Yuille. Joint object and part segmentation using deep learned potentials. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1573–1581, 2015. 2

[62] Wenguan Wang, Zhijie Zhang, Siyuan Qi, Jianbing Shen, Yanwei Pang, and Ling Shao. Learning compositional neural information fusion for human parsing. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5703–5713, 2019. 8

[63] Wenguan Wang, Hailong Zhu, Jifeng Dai, Yanwei Pang, Jianbing Shen, and Ling Shao. Hierarchical human parsing with typed part-relation reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8929–8939, 2020. 8

[64] Wenguan Wang, Tianfei Zhou, Fisher Yu, Jifeng Dai, Ender Konukoglu, and Luc Van Gool. Exploring cross-image pixel contrast for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7303–7313, 2021. 8

[65] Fangting Xia, Peng Wang, Liang-Chieh Chen, and Alan L Yuille. Zoom better to see clearer: Human and object parsing with hierarchical auto-zoom net. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, pages 648–663. Springer, 2016. 8

[66] Fangting Xia, Peng Wang, Xianjie Chen, and Alan L Yuille. Joint multi-person pose estimation and semantic part segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6769–6778, 2017. 8

[67] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European conference on computer vision (ECCV)*, pages 418–434, 2018. 8

[68] Binhui Xie, Shuang Li, Mingjia Li, Chi Harold Liu, Gao Huang, and Guoren Wang. Sepico: Semantic-guided pixel contrast for domain adaptive semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 8

[69] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021. 4

[70] Kota Yamaguchi, M Hadi Kiapour, Luis E Ortiz, and Tamara L Berg. Parsing clothing in fashion photographs. In *2012 IEEE Conference on Computer vision and pattern recognition*, pages 3570–3577. IEEE, 2012. 8

[71] Qihang Yu, Huiyu Wang, Dahun Kim, Siyuan Qiao, Maxwell Collins, Yukun Zhu, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Cmt-deeplab: Clustering mask transformers for panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2560–2570, 2022. 2, 8

[72] Qihang Yu, Huiyu Wang, Siyuan Qiao, Maxwell Collins, Yukun Zhu, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. k-means Mask Transformer. In *ECCV*, pages 288–307. Springer, 2022. 2, 4, 8, 9

[73] Wenwei Zhang, Jiangmiao Pang, Kai Chen, and Chen Change Loy. K-net: Towards unified image segmentation. *Advances in Neural Information Processing Systems*, 34, 2021. 8

[74] Yifan Zhao, Jia Li, Yu Zhang, and Yonghong Tian. Multi-class part parsing with joint boundary-semantic awareness. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9177–9186, 2019. 4, 8, 9

[75] Long Zhu, Yuanhao Chen, Chenxi Lin, and Alan Yuille. Max margin learning of hierarchical configural deformable templates (hcdts) for efficient object parsing and pose estimation. *International journal of computer vision*, 93:1–21, 2011. 8

5. Related Work

5.1. Object Parsing

The extensive literature on object parsing can be divided into single-object multi-part parsing [4, 20, 39, 51, 65, 66] and multi-object multi-part parsing [22, 49, 53, 74]. Single-object multi-part parsing has primarily focused on specific classes, such as humans [38, 70, 75], animals [60], and vehicles [18, 46, 54]. While the methodologies addressing multi-object multi-part parsing mainly focus on employing top-down or coarse-to-fine strategies. Specifically, Singh et al. [53] proposed FLOAT, a factorized top-down parsing framework by first detecting the object followed with zooming in for obtaining higher quality part masks. On the contrary, He et al. [22] introduced Compositor, a bottom-up architecture designed to iteratively learn objects by clustering pixels to derive parts. Recently, there are also explorations in the closely related area of panoptic part segmentation within the research community. Notable works such as [1, 15, 34, 35, 50, 56] have delved into the semantic parsing of objects while also distinguishing parts between different instances. However, a common trend in these works, whether focused on semantic object parsing or panoptic part segmentation, involves extending standard segmentation models, often overlooking the nuanced semantic levels of parts. In contrast, CoCal takes a novel approach by focusing specifically on semantic object parsing. It redefines the paradigm of cluster-based mask transformers and introduces a novel dictionary-based framework meticulously tailored for object parsing.

5.2. Cluster-based Mask Transformer

With the recent progress in transformers [3], a new paradigm named mask classification [12, 13, 55, 58, 59, 73] has been proposed, where segmentation predictions are represented by a set of binary masks with its class label, which is generated through the conversion of object queries to mask embedding vectors followed by multiplying with the image features. The predicted masks are trained by Hungarian matching with ground truth masks. Thus the essential component of mask transformers is the decoder which takes object queries as input and gradually transfers them into mask embedding vectors. Recently, cluster-based mask transformers are introduced in [37, 71, 72], which rethinks the design of the decoder by replacing the cross-attention with a k-means [43] attention. Building upon these explorations, CoCal introduces a global class dictionary and replaces the Hungarian matching with a fixed one-to-one matching, thereby establishing an interpretable dictionary-based framework for part segmentation.

5.3. Contrastive Learning in Segmentation

Contrastive learning [8, 9, 11, 24, 25, 28, 52] has emerged as a prominent technique in computer vision as an effective method for learning feature representation for self-supervised models. The core idea lies in contrasting similar (positive) data pairs against dissimilar (negative) pairs. Recently, Wang et al. [64] raise a pixel-to-pixel contrastive learning method for semantic segmentation, which enforces pixel embeddings belonging to the same semantic class to be more similar than embeddings from different classes. [7, 17, 33, 47, 57, 68] are built upon this concept, extending it to various segmentation domains. Motivated by these advancements, we propose a component-wise contrastive learning method tailored for modern cluster-based mask transformers, which effectively learns discriminative dictionary components within the clustering scheme.

5.4. Logical Constraints in Segmentation

Few segmentation models [27, 31, 32, 36, 40, 62, 63, 67] consider the implicit logic rules inherent in structured labels. While the majority of them are dedicated to human parsing, a few recent works [31, 32] tackle the general segmentation in a flexible function and avoid incorporating label taxonomies into the network topology. Concretely, Li et al. [31] enhance the logical consistency by modeling the segmentation as a pixel-wise multi-label classification. Li et al. [32] exploit neuro-symbolic computing for grounding logical formulae onto data. In contrast to these efforts, CoCal introduces an object level on top of the part and models logical rules as a contrastive objective during training.

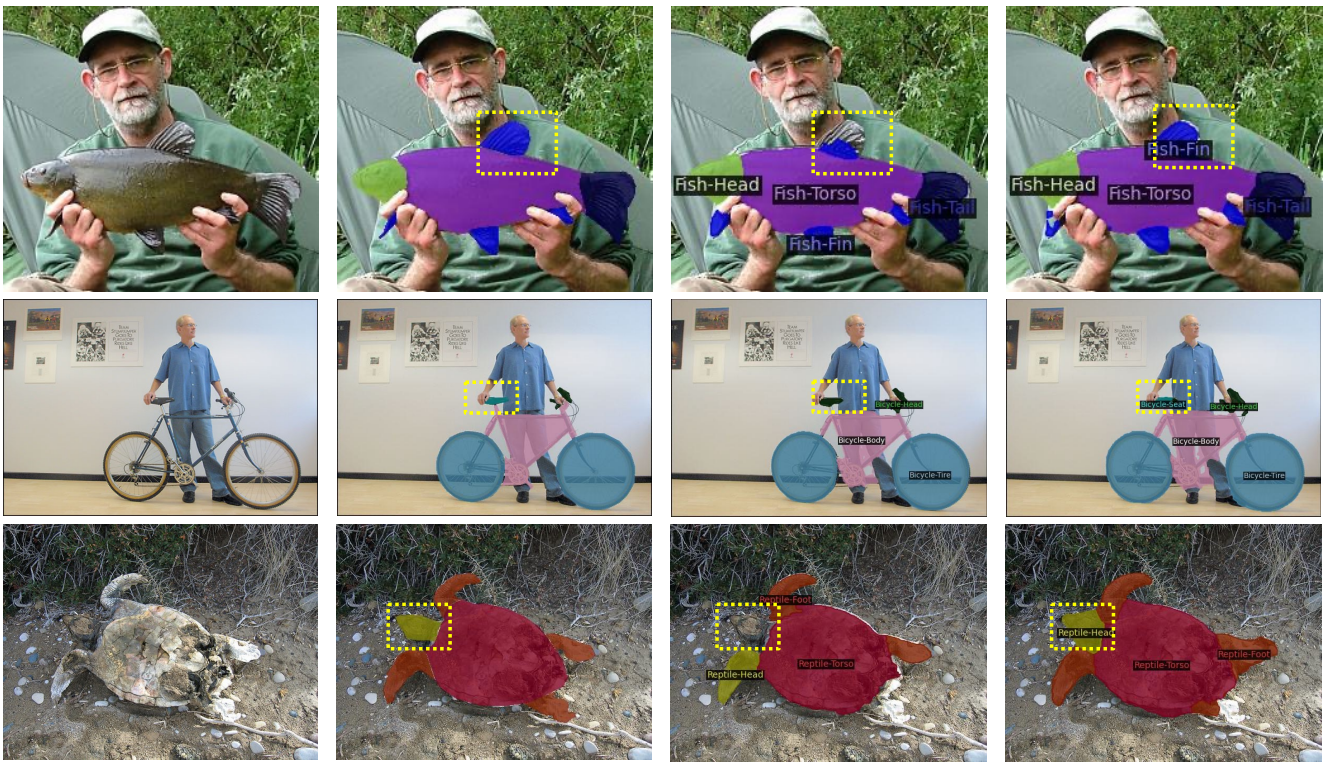
6. Extended Experiments and Analyses

This section provides details on our experiments, including ablation studies, dataset statistics, training parameters, and an in-depth exploration of CoCal's design components.

6.1. Experimental Setup and Datasets

Datasets We conduct experiments on two popular object parsing benchmarks: PartImageNet [21] and PASCAL-Part-108 [49]. We provide the detailed statistics of each dataset and the class definitions below:

- PartImageNet [21] contains 24095 elaborately annotated general images from ImageNet [16], which are split into 20481/1206/2408 for *train*/*val*/*test*. It is associated with 40 part classes, which are grouped into 11 object classes following the official class definition.
- Pascal-Part-108 [49] expands upon the part definition introduced in Pascal-Part-58 [10], providing a more intricate benchmark with finer part-level details. This extension maintains the original split of VOC [19] and encompasses a dataset of 10,103 images across 20 object classes and 108 part classes. Our experiments adhere to the orig-



(a) Image (b) Ground Truth (c) kMax-DeepLab **(d) CoCal**

Figure 4. **Qualitative comparison between CoCal and kMaX-DeepLab on PartImageNet.** Our CoCal yields more precise part boundaries (row 1) and captures missed parts (rows 2 & 3).

inal split, utilizing 4,998 images for training and 5,105 images for testing.

Evaluation Metrics We evaluate the performance of CoCal on the PartImageNet dataset [21] using the mean Intersection over Union (mIoU) on both part and super-category levels. It's important to note that for PartImageNet, we choose to report performance on the super-category level because the parts in PartImageNet are defined within the context of super-category. The hierarchy of super-category is inherited for training CoCal on this dataset. In the case of Pascal-Part-108, our evaluation includes reporting part mIoU, and additionally, we calculate the mAvg on the object level. The mAvg metric, as defined in the literature [74], provides the average mIoU score of all parts belonging to an object. We refer the reader to FLOAT [53] for a detailed explanation of these metrics.

Training details We implement CoCal based on the kMaX-DeepLab architecture [72], utilizing its official PyTorch re-implementation codebase. To ensure a fair comparison, we adopt the training settings from kMaX-DeepLab. The backbone, pretrained on ImageNet [23, 42], followed a learning rate multiplier of 0.1. For regularization and augmentations, we incorporate drop path [26] and random color jittering [14]. The optimizer used is AdamW [29, 45] with a weight decay of 0.05. Unless otherwise specified, we train all models with a batch size of 64 on a single A100 GPU, performing 40,000 iterations on PartImageNet [21] and 10,000 iterations on Pascal-Part-108 [10]. The first 2,000 and 500 steps serve as the warm-up stage, where the learning rate linearly increases from 0 to 5×10^{-4}. The training objective for CoCal includes the combination of kMaX-DeepLab's original losses and the proposed contrastive loss terms, as specified in Eq. 5, Eq. 6 and Eq. 7:

$$\mathcal{L} = \lambda_{\text{kMaX}}\mathcal{L}_{\text{kMaX}} + \lambda_{p\_con}\mathcal{L}_{p\_con} + \lambda_{o\_con}\mathcal{L}_{o\_con} + \lambda_{logic}\mathcal{L}_{logic}.$$

Here, $\mathcal{L}_{\text{kMaX}}$ represents the loss from kMaX-DeepLab [72], and λ_{kMaX} follows the official setting. The weights for the proposed loss terms are set to $\lambda_{p\_con} = 2$, $\lambda_{o\_con} = 2$, and $\lambda_{logic} = 1$. CoCal uses the exact same number of part and object queries corresponding to the part and object classes in the dataset. Specifically, we set P to 41 and 109, and $\widetilde{P}$ to 12 and 21 (with one additional learnable component for representing the background at both the part and object levels) in PartImageNet and Pascal-Part-108, respectively. This de-

sign enables a straightforward and highly interpretable inference process, using nearest neighbor search for parts and objects separately during inference. Afterward, we compute the top-scoring logical path and reassign the predicted classes based on that path.

6.2. Ablation Studies

Dictionary Components, Contrastive Objectives, and Logical Constraints Table 2 (reproduced from the main paper for convenience) shows an ablation on core design choices. Simply switching kMaX-DeepLab to a dictionary-based version slightly degrades performance (65.75 to 64.31), but adding contrastive objectives and logical constraints incrementally boosts part mIoU to 67.83, surpassing the original baseline.

Table 2. Ablation of CoCal components on PartImageNet *val* with ResNet-50.

method	Dict	$\mathcal{L}_{p\_con}$	$\mathcal{L}_{o\_con}$	$\mathcal{L}_{logic}$	Part mIoU
kMaX-DeepLab	✗	✗	✗	✗	65.75
Dictionary-based	✓	✗	✗	✗	64.31
CoCal (ours)	✓	✓	✗	✗	65.87
CoCal (ours)	✓	✓	✓	✗	66.53
CoCal (ours)	✓	✓	✓	✓	**67.83**

Memory Bank Size S In Table 3, we vary S and observe that excessively small or large values degrade performance due to insufficient or redundant samples.

Table 3. **Impact of memory bank size** on PartImageNet *val* (ResNet-50).

# memory bank S	Part mIoU
50	66.50
100	**67.83**
150	67.16
200	67.02

Number of Negative Samples k Table 4 shows that using too few negatives (e.g., $k = 50$) reduces performance to 66.28 part mIoU, while too many (e.g., $k = 200$ or "all") also hurts accuracy.

Table 4. **Influence of negative samples** on PartImageNet *val* (ResNet-50).

# negative sample k	Part mIoU
50	66.28
100	**67.83**
200	66.40
all	65.74

Generalizability of CoCal Finally, Table 5 demonstrates that CoCal also boosts other transformer-based frameworks such as MaskFormer [12] and Mask2Former [13], improving part mIoU by 3.18 and 2.77, respectively.

Table 5. **Generalizability to other baselines** on PartImageNet *val* (ResNet-50).

method	mIoU	
	Part	Super-Category
MaskFormer	60.34	-
CoCal (MaskFormer)	63.52	86.67
Mask2Former	63.62	87.20
CoCal (Mask2Former)	66.39	88.72