Knowledge-Guided Part Segmentation

Supplementary Material

In the following, we provide more details in Section A. We further provide additional experimental results and ablation study in Section B. Finally, we present qualitative results for the benchmarks in Section C.

A. More Details

A.1. Coarse-grained Perception Module

The Coarse-grained Perception Module, shown in Figure 7, is tailored to capture object-level contextual cues, specifically for segmenting larger objects in the scene. This module alternates between Swin Transformer blocks and Text Transformer blocks to iteratively refine the visual and text features. The Swin Transformer blocks apply local and shifted self-attention mechanisms, allowing the model to capture spatial relationships across object-level regions, enhancing the detail of more significant segments. The Text Transformer blocks refine these features by focusing on object-level text embeddings, supporting more robust coarse-grained differentiation between distinct objects. The outputs pass through a Conv Decoder with upsampling layers to complete the process, producing a high-resolution object-level segmentation output. This structure enables the module to effectively segment objects by leveraging visual structure and coarse-grained object representations.

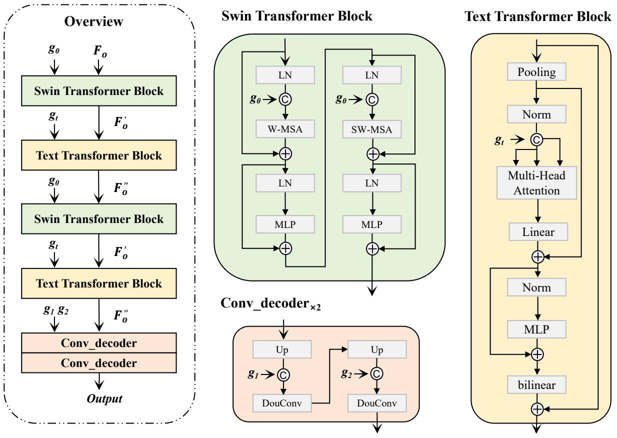

Figure 7. The Coarse-grained Perception Module focuses on segmenting larger objects. The terms g_1 and g_2 represent features extracted from the 3rd and 7th layers of the CLIP model (ViT-B/16).

A.2. Mask Encoder

This encoder is designed to capture high-level, abstract features by progressively reducing the spatial dimensions of the input through a series of convolutional operations. The mask encoder processes coarse-grained, object-level class probability maps, refining them to serve as targeted guidance for fine-grained segmentation. It consists of convolutional layers, each followed by a ReLU activation function, introducing non-linearity to enable the model to capture intricate spatial patterns. Following the convolutional layers, max pooling operations downsample the feature maps, reducing spatial dimensions while retaining essential features. Finally, a 1×1 convolution layer consolidates these refined representations, preparing them as guidance signals for the subsequent fine-grained segmentation pipeline.

This architecture effectively distills critical information from coarse-grained inputs, producing abstract, context-enriched features that enhance the model's ability to perform precise part-level segmentation.

A.3. Fine-grained Perception Module

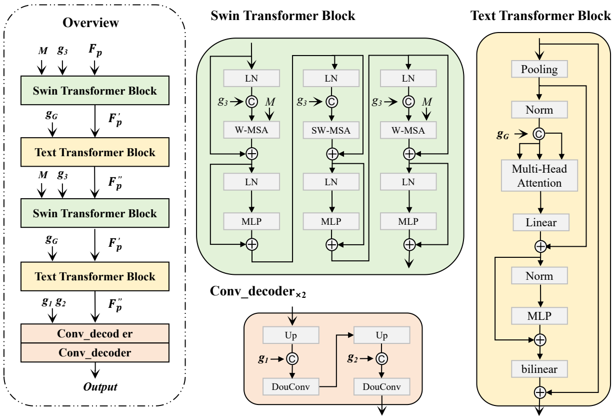

Figure 8. Fine-grained Perception Module, focuses on segmenting finer parts.The terms g_1 and g_2 represent intermediate features extracted from the 3rd and 7th layers of the CLIP model (ViT-B/16).

The structure of the Fine-grained Perception Module is shown in Figure 8. This module comprises three main components: the Swin Transformer for visual perception, the Text Transformer for text-based perception, and a decoder head for segmentation. Three consecutive blocks are employed in the visual perception pathway to capture fine-grained features. The first and third blocks apply self-attention within a local window guided by coarse-grained information, while the second block employs shifted window self-attention without guidance. This configuration allows the module to focus on foreground regions, reduce background noise, and enhance the understanding of relationships among parts. In the text perception pathway, part structure-enriched text features replace the original part-level textual features, embedding structural relationships directly into the text representation. Overall, this module leverages text features containing part structure relationships and coarse-grained object information to guide precise part-level segmentation, further enhancing the model's ability to focus on fine-grained details within parts.

B. Additional Ablation Study

B.1. Ablation on input text templates

We study the importance of the template choice for the entries in the text list in Table 7. We experiment with "A photo of a {CLS} in the scene" template for our text entries where {CLS} is the class name for the object mask. We also experiment with the identity template "{CLS}". Our template choice: "A photo of a {CLS} in the scene" gives a strong performance as a baseline. We believe more exploration in the text template space could help improve performance.

Text Templates	mIOU	mACC
{CLS}	61.57	71.28
"A photo of a {CLS}"	62.38	72.19
"A photo of a {CLS} in the scene"	**62.42**	**72.13**

Table 7. Ablation on input Text Templates.

The results in Table 7 highlight the importance of text template choice for segmentation performance. The template "A photo of a {CLS} in the scene" achieves the highest mIOU and mACC scores, outperforming both the simple "{CLS}" and the shorter template "A photo of a {CLS}". This suggests that more descriptive templates that provide contextual information can better capture the semantic richness needed for accurate segmentation. While our selected template serves as a strong baseline, further exploration of alternative text templates could potentially lead to additional performance improvements by better aligning textual and visual features.

B.2. Effectiveness of Graph Convolution Types

We assess several representative graph convolution types within our SKGM architecture, including GraphConv [43], ResGatedGraphConv [6], TransformerConv [51], MFConv [17], LEConv [48], and TAGConv [16]. As shown in Table 8, most graph convolution variants surpass the baseline model in terms of mIOU and mACC, highlighting the adaptability of the SKGM architecture to different convolutional designs. Among these, TAGConv achieves the highest accuracy and is therefore selected as the default configuration for all further experiments unless otherwise noted.

The experimental results highlight the varying performance of different graph convolution types within the SKGM architecture, reflecting their distinct characteristics and adaptability. TAGConv achieves the best results, leveraging multi-scale aggregation to capture both local and global relationships, enhancing feature representation and recognition. TransformerConv [51] and ResGatedGraphConv [6] also perform well, with TransformerConv [51] improving long-range dependency modeling via self-attention, while ResGatedGraphConv [6] utilizes gated residual connections to refine information flow and mitigate degradation in deeper networks. Their strong performance underscores the benefits of flexible information propagation in GCNs.

Conv Layers	mIOU(%)	mACC(%)
GraphConv [43]	61.38	70.85
ResGatedGraphConv [6]	62.23	71.76
TransformerConv [51]	62.21	72.10
MFConv [17]	60.64	70.31
LEConv [59]	60.44	69.83
TAGConv [16]	**62.42**	**72.13**

Table 8. GCN of different types of graph convolution.

In contrast, GraphConv [43] provides solid performance but lacks the advanced propagation and feature extraction capabilities of newer variants. MFConv [17] and LEConv [59] show lower scores, likely due to MFConv [17]'s limitations in modeling long-range dependencies and LEConv [59]'s design for low-energy structures, making them less suitable for this task.

B.3. Effectiveness of the loss

The parameters λ_o and λ_p control the segmentation loss for the coarse-grained and fine-grained modules, respectively. As illustrated in Table 9, increasing the weight of the fine-grained module's loss initially improves KPS performance, as it strengthens the focus on detailed segmentation. However, as λ_p continues to increase, the gains diminish, indicating that overemphasizing fine-grained detail can lead to suboptimal results. The best performance is achieved when λ_o= 0.2 and λ_p= 0.8, highlighting the importance of balancing these parameters to optimize both coarse- and fine-grained segmentation objectives.

λ_p λ_o	0.2 0.8	0.4 0.6	0.6 0.4	**0.8** **0.2**	0.9 0.1
mIOU(%)	53.74	60.53	60.79	62.42	59.42
mACC(%)	65.61	70.24	70.42	72.13	70.48

Table 9. The ablation studies for two coefficients λ_o and λ_p.

B.4. Effectiveness of the N_m in the Mask Encoder

We assess the effect of varying the number of convolutional layers N_m in the mask encoder, as detailed in Table 10, using configurations of N_m= [2, 3, 4, 5, 6, 8]. Results indicate that performance improves with an increase in N_m up to N_m= 4, which achieves the highest mIOU (62.42%) and mACC (72.13%). Beyond this point, further increases in N_m result in diminishing returns, and even slight declines in performance due to potential overfitting, emphasizing the importance of selecting N_m= 4 for optimal balance.

N_m	mIOU(%)	mACC(%)	N_m	mIOU(%)	mACC(%)
N_m=2	60.54	70.29	N_m=6	60.93	71.05
N_m=3	61.69	72.04	N_m=8	60.62	71.49
N_m=5	62.34	71.88	N_m=4	**62.42**	**72.13**

Table 10. Ablation Study on Conv Count in the Mask Encoder.

The results in Table 10 show that increasing N_m in the mask encoder improves segmentation performance up to N_m= 4, where the highest mIOU (62.42%) and mACC (72.13%) are achieved. This indicates that adding layers initially enhances feature extraction, supporting better segmentation outcomes. However, further increases beyond N_m= 4 lead to diminishing returns and slight declines in performance, likely due to overfitting as model complexity increases. Thus, N_m= 4 provides an optimal balance, effectively capturing rich features without excessive complexity.

B.5. Effectiveness of FPM Design

To evaluate the impact of specific design choices within the fine-grained perception module, we performed ablation studies focusing on two primary factors: the number of Swin Transformer blocks and the application of coarse-grained guided self-attention within a local window. We tested configurations with both two and three consecutive Swin Transformer blocks to observe how deeper feature extraction layers affect performance. Additionally, we compared configurations using the original, unconditioned self-attention within a local window against a variant with foreground-focused, coarse-grained guided self-attention.

EXP	Block Config	Guidance	mIOU(%)	mACC(%)
(I)	2 blocks	✗	62.22	71.70
(II)	2 blocks	✓	62.29	71.39
(III)	3 blocks	✗	61.72	71.51
(IV)	**3 blocks**	✓	**62.42**	**72.13**

Table 11. Impact of different modules in our KPS. We show the results of integrating different modules into the baseline.

Following Table 11, the results demonstrate that different configurations of Swin Transformer blocks and the application of coarse-grained guided self-attention impact the model's segmentation performance. Specifically, using three Swin Transformer blocks with coarse-grained guided self-attention (setting IV) yields the highest mIOU and mACC scores, achieving 62.42% and 72.13%, respectively. This configuration highlights the benefits of deeper feature extraction in combination with foreground-focused guidance, effectively enhancing fine-grained segmentation. Conversely, reducing the number of blocks or omitting the coarse-grained guidance leads to lower accuracy, suggesting that both additional layers and targeted attention are key contributors to model improvement.

C. More Qualitative Results

We provide more qualitative results on PartImageNet and Pascal-Part in the Figure 9.

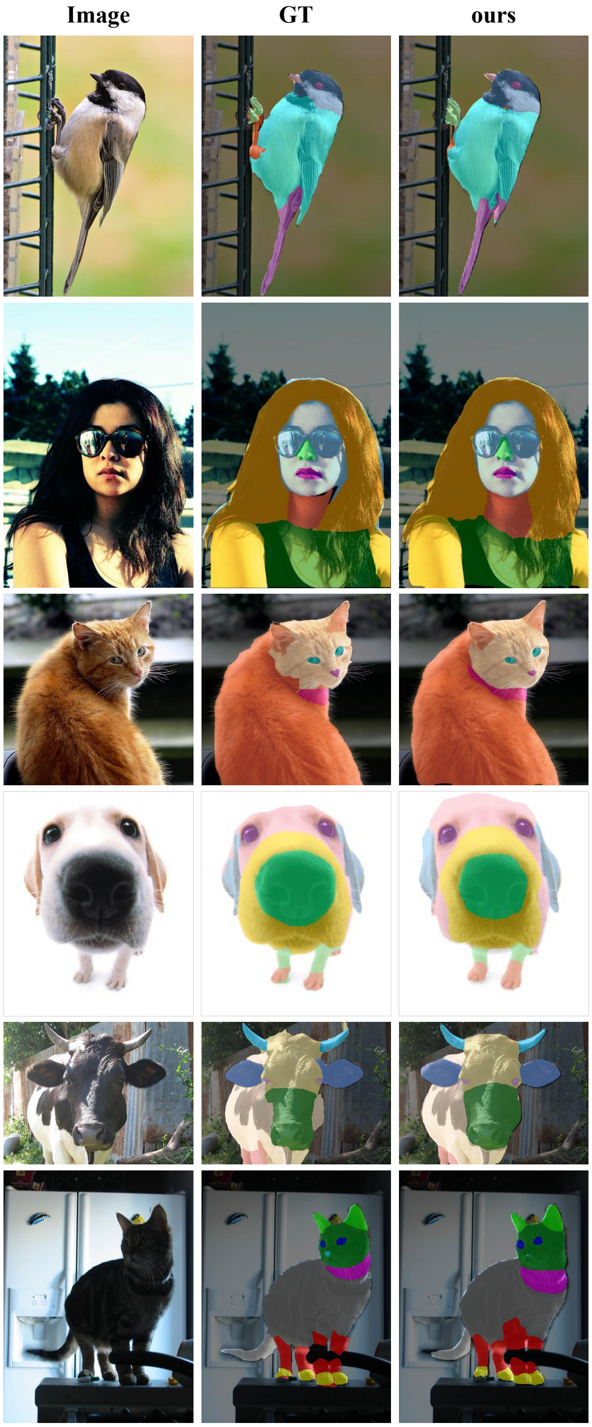

Figure 9. Qualitative Results on Pascal-Part: Additional visualization results.