# From Pixels to Objects: A Hierarchical Approach for Part and Object Segmentation Using Local and Global Aggregation

Yunfei Xie[1], Cihang Xie[2], Alan Yuille[3], and Jieru Mei[3]

[1] Huazhong University of Science and Technology
[2] UC Santa Cruz
[3] Johns Hopkins University

**Abstract.** In this paper, we introduce a hierarchical transformer-based model designed for sophisticated image segmentation tasks, effectively bridging the granularity of part segmentation with the comprehensive scope of object segmentation. At the heart of our approach is a multi-level representation strategy, which systematically advances from individual pixels to superpixels, and ultimately to cohesive group formations. This architecture is underpinned by two pivotal aggregation strategies: local aggregation and global aggregation. Local aggregation is employed to form superpixels, leveraging the inherent redundancy of the image data to produce segments closely aligned with specific parts of the object, guided by object-level supervision. In contrast, global aggregation interlinks these superpixels, organizing them into larger groups that correlate with entire objects and benefit from part-level supervision. This dual aggregation framework ensures a versatile adaptation to varying supervision inputs while maintaining computational efficiency.

Our methodology notably improves the balance between adaptability across different supervision modalities and computational manageability, culminating in significant enhancement in segmentation performance. When tested on the PartImageNet dataset, our model achieves a substantial increase, outperforming the previous state-of-the-art by 2.8% and 0.8% in mIoU scores for part and object segmentation, respectively. Similarly, on the Pascal Part dataset, it records performance enhancements of 1.5% and 2.0% for part and object segmentation, respectively.

**Keywords:** Semantic segmentation · Superpixels

## 1 Introduction

Joint part and object segmentation presents a formidable challenge that involves simultaneously performing holistic object segmentation and detailed part segmentation. Although current methods [12,31,34,44] have shown effectiveness in segmenting object parts, there often remains a gap in achieving simultaneous object-level segmentation, as well as constraints related to computational efficiency. In particular, many approaches [27–29,31,46] are tailored primarily for
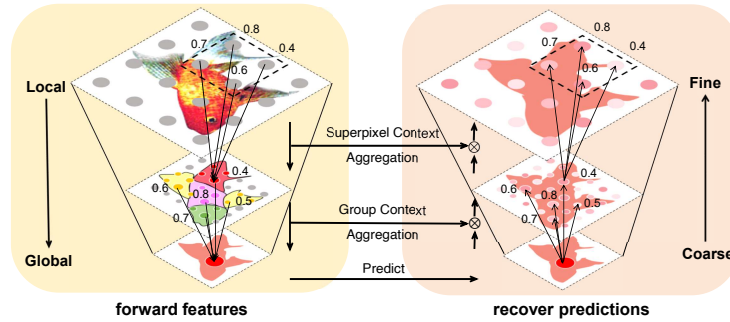
**Fig. 1: Conceptual illustrations of LGFormer.** On the left, LGFormer follows a hierarchical aggregation pathway, elevating features from pixels to parts to objects. The right figure shows the model's ability to progressively restore segmentation predictions from the object level to the original image resolution.

part segmentation and do not effectively address the dual task of object segmentation. Additionally, prevalent approaches [34] may employ separate specialized computational architectures for each segmentation task, thereby substantially increasing the computational overhead.

The underlying challenge stems from the inherently conflicting goals of part and object segmentation. Object segmentation necessitates the integration of broad features across an object to ensure a cohesive representation, in contrast to part segmentation, which requires distinguishing between the smaller, detailed features within the object. Moreover, these segmentation tasks differ fundamentally in their spatial emphasis: object segmentation is enhanced by a broader, global perspective which aids in object recognition, whereas part segmentation focuses more narrowly on local details essential for accurate delineation of component boundaries [34]. This inherent tension demands a strategy that can reconcile global coherence with detailed local recognition, highlighting the urgency for innovative solutions capable of efficiently bridging the conflicting demands of holistic object and intricate part segmentation within a cohesive framework.

To address these complexities, we introduce **Local Global Transformer** (LGFormer), a model inspired by the hierarchical structure of human visual perception, which starts with recognizing smaller components and their spatial relationships before synthesizing these elements into a holistic understanding of the object [19]. LGFormer is designed to innovatively manage the simultaneous segmentation of detailed parts and entire objects, thereby addressing the challenge from a foundational level.

**Hierarchical Representation.** At the core of our approach is the hierarchical organization of visual data, which simulates the natural progression from discrete pixels to complex object representations. This system organizes pixels into superpixels, and these superpixels into groups through advanced cross-attention mechanisms, enabling LGFormer to capture multi-scale information adeptly. This structured arrangement facilitates concurrent detailed part seg-

mentation and comprehensive object segmentation, maintaining precision and fidelity across varying scales.

**Association-Aware Upsampling.** Building on the structured hierarchy of pixels, superpixels, and groups, LGFormer incorporates an innovative upsampling technique, termed association-aware upsampling, engineered to preserve and accurately convey detailed spatial information through the hierarchical layers back to the original image scale. By utilizing attention scores from cross-scale interactions, this method preserves the integrity of fine details more effectively than traditional upsampling approaches, thereby enhancing both precision and detail retention. This association-aware upsampling serves as a critical component, ensuring the preservation of multi-scale segmentation insights in the final high-resolution output.

**Unified Multi-Task Framework.** The foundation of LGFormer is a unified architectural approach that obviates the need for separate frameworks for part and object segmentation. This integrated structure not only simplifies the segmentation workflow but also boosts the model's efficiency and interpretability. By employing a single, cohesive framework, LGFormer adeptly shifts between the local and global segmentation demands, marking a substantial advancement in replicating human-like perception in visual segmentation tasks.

Thus, LGFormer stands as a pioneering solution in the domain of joint part and object segmentation, offering a scalable and efficient approach that adeptly balances meticulous detail segmentation with global context comprehension, tackling both local and global redundancies with unmatched precision.

We empirically validate the superior performance of LGFormer on the benchmark datasets PartImageNet [13] and Pascal-Part [9]. The model demonstrates its capability in generating high-quality semantic parts, substantially enhancing object segmentation. Our evaluations show that LGFormer achieves mIoU scores of 67.4% and 79.8% for part and object segmentation, respectively, surpassing the previous state-of-the-art model, Compositer [12], by 2.9% and 0.9%. Remarkably, on the Pascal-Part dataset, LGFormer exceeded Compositer by 1.5% and 2.0% in part and object mIoU, respectively.

Further examination underscores the capability of LGFormer's hierarchical architecture to engender semantic understanding autonomously, illuminating a process that is both visually interpretable and inherently explainable. Notably, this architecture enables the derivation of groups from superpixels without explicit object-level supervision, and similarly allows superpixels to form solely from object supervision. These dynamics are comprehensively detailed in Fig. 5, showcasing the intrinsic adaptability and intelligence of our hierarchical representation in capturing complex semantic relationships.

1. We introduce a novel hierarchical representation that emulates the functioning of the human visual system, effectively reducing the computational complexity of Vision Transformers. This approach distinctively addresses both local and global redundancies, enhancing processing efficiency.
2. Leveraging this hierarchical structure, we develop an innovative upsampling technique termed association-aware upsampling. This method success-

fully overcomes the blurring commonly associated with existing upsampling strategies, thereby preserving the fidelity of fine details across various segmentation tasks.

3. We utilize our hierarchical framework to manage both part and object segmentation within a unified model, LGFormer, demonstrating our approach's versatility and efficiency.

## 2    Related Works

**Bridging Part and Object Segmentation**   Joint learning of objects and parts representation concurrently is an attractive topic and has been actively investigated. Wang *et al.* [34] introduced a novel approach with a dual-channel, fully convolutional network that predicts semantic compositional parts and object potentials at the pixel level, complemented by a fully connected conditional random field for refined predictions. Subsequently, Singh *et al.*'s FloatSeg [31] framework innovatively involves multiple decoders for objects and part attributes respectively. However, these methodologies share a common drawback: their reliance on multiple encoders or decoders for handling part and object representations, which increases computational demand and complexity in mapping part-to-object relationships. The most related work to ours is Compositor [12], which proposes a bottom-up strategy to compose embeddings from parts to objects. However, in Compositor, interactions among pixels, parts, and objects operate completely at a global scale, which is suboptimal due to the high computational cost and the absence of local attention as inherent inductive biases. In contrast, our model leverages the superpixels to utilize the local redundancy within the image, yielding segments that align with image parts even under solely object-level supervision, making it a natural fit for part segmentation. Thanks to the boundary preservation from our hierarchy representation, our association-aware upsampling produces sharper predictions.

**Overcoming Local and Global Redundancies**   High-resolution image processing often grapples with redundancies that challenge both resolution management and detail preservation. Traditional techniques employ max pooling to downscale images [14, 15, 18, 20, 26, 30], albeit at the cost of losing finer details. To counteract this loss, extensive decoders have been introduced to recover detailed information [1,3,22]. In contrast, the incorporation of superpixels into deep learning frameworks [16, 17, 21, 24, 25, 36, 38, 39, 42, 45, 47] presents a more nuanced approach, effectively addressing local image redundancy while conserving boundary integrity.

Yet, while superpixels adeptly handle local redundancy, global redundancy remains a concern. Recent models like GCViT [11], GroupViT [36], and GPViT [37] have attempted to bridge this gap by incorporating group tokens that facilitate global information exchange. However, due to computational demands, these models typically rely on either coarse patches or a limited number of groups, which restricts their capacity.

LGFormer innovates on this front by sequentially advancing from pixels to superpixels, and finally to groups. This progression addresses both local and

global redundancies through a hierarchical representation that aligns seamlessly with the joint demands of part and object segmentation. Unlike its predecessors, LGFormer harnesses varying levels of feature semantics, enabling a more comprehensive and detail-preserving segmentation approach.
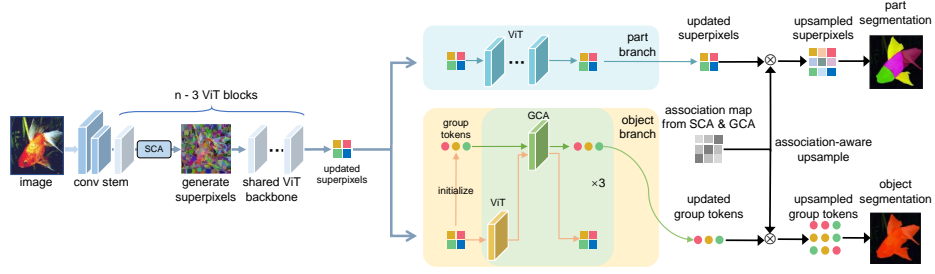
# 3 Method



**Fig. 2: Overview of LGFormer.** Pixel-level features are extracted by a light convolution stem. In the initial ViT stages, these features are refined into part-level superpixels via Superpixel Context Aggregation (SCA). In deeper ViT layers, superpixels are aggregated into object-level groups using Group Context Aggregation (GCA).

## 3.1 From Pixels to Objects

**Pixel Representation.** Given an input image $\mathbf{M} \in \mathbb{R}^{m_h \times m_w \times 3}$, the large spatial dimensions inherently introduce computational complexity. To address this, we use a lightweight convolution stem that downsamples the image and extracts pixel feature maps $\mathbf{I} \in \mathbb{R}^{i_h \times i_w \times i_c}$, thereby reducing initial redundancy and computational load while preserving crucial visual details.

**Superpixel Representation.** Noting that regions within an object often contain clusters of pixels with redundant information—due to similarity among adjacent pixels—we shift to a superpixel representation $\mathbf{S} \in \mathbb{R}^{s_h \times s_w \times s_c}$. Following [25], we apply a Superpixel Context Aggregation (SCA) technique, which aggregates pixels into superpixels by integrating local contextual information. This transition effectively reduces local redundancy and increases the model's efficiency and explainability by focusing on coherent contextual segments.

**Group Representation.** Although superpixels efficiently compress local information, they generally fail to capture the global semantics critical for object-level representations. This limitation is due to their focus on localized areas, which misses the broader, holistic view required for recognizing entire objects. To rectify this, we employ a method that groups multiple superpixels into groups $\mathbf{G} \in \mathbb{R}^{g_n \times g_c}$ using a Group Context Aggregation (GCA). This method reduces
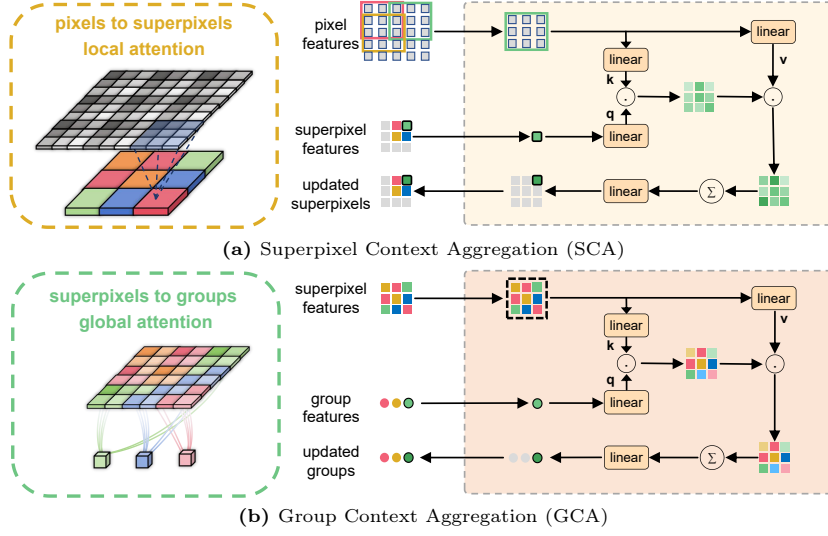
**(a)** Superpixel Context Aggregation (SCA)



**(b)** Group Context Aggregation (GCA)

**Fig. 3:** Illustration of interactions and spatial relationships among three hierarchical levels: pixels, superpixels, and groups. For clarity, the illustration presents a scenario involving only a single superpixel and a single group token. (a) Transitioning from pixels to superpixels involves iterative refinement of superpixels on a local scale through SCA. (b) Advancing from superpixels to groups, the refinement of groups on a global scale is facilitated by GCA.

global redundancy by abstracting similar or repeated part features across the image and simplifies computational demands. More importantly, it enhances global interpretability and overall model performance by forming higher-level abstractions that more accurately reflect object-level semantics.

**Local and Global Aggregation.** Hierarchical spatial downsampling, a prevalent technique in segmentation models [2, 20], typically fails to differentiate adequately between semantic elements, merging distinct features indiscriminately. To mitigate this, we introduce a novel semantic stratification process, by mapping part to superpixels and object to groups via SCA and GCA mechanisms, as shown in Fig. 3. This method ensures that semantic redundancies are efficiently managed, preserving essential features while reducing unnecessary information across scales.

In SCA, our approach aims to mitigate local redundancies by efficiently organizing pixels into superpixels. This crucial step not only reduces data complexity but also preserves vital details necessary for precise part segmentation. The process is articulated as:

$$\mathbf{S}_p^t = \mathbf{S}_p^{t-1} + \sum_{i \in \mathcal{N}_p} \mathrm{softmax}\left(\mathbf{q}_{\mathbf{S}_p^{t-1}} \cdot \mathbf{k}_{\mathbf{I}_i^{t-1}}\right) \mathbf{v}_{\mathbf{I}_i^{t-1}}, \tag{1}$$

where $\mathcal{N}_p$ identifies the pixels adjacent to superpixel $p$. The objective of this aggregation is to minimize local redundancy through optimized pixel-to-superpixel

assignments. The attention mechanism utilizes the softmax function to process the dot products of the query ($\mathbf{q}$) and key ($\mathbf{k}$) vectors, determining the significance of contributions from each pixel. These vectors, alongside the value vector ($\mathbf{v}$), are derived from linear transformations of the features of superpixels $\mathbf{S}_p^{t-1}$ and pixels $\mathbf{I}_i^{t-1}$ from the previous iteration. This permits the model to adaptively focus on and integrate the most relevant pixel data to enhance superpixel features. For comprehensive insights, we direct the reader to SPFormer [25].

Building upon local context aggregation, GCA employs global cross-attention to iteratively update groups and superpixels. This mechanism is made feasible by the reduced number of groups, which lowers computational costs. Additionally, the preceding mitigation of local redundancies by SCA aids in enhancing semantic abstraction at the object level, rendering the aggregation process more efficient. Each iteration $t$ consists of two critical phases: Superpixel-to-Group (S2G) and Group-to-Superpixel (G2S) cross-attention.

For S2G cross-attention within GCA, group tokens $\mathbf{G}_g^t$ are refined by aggregating information across all superpixels, with an FFN applied to elevate the aggregated features to a higher semantic level:

$$\mathbf{G}_g^t = \mathbf{G}_g^{t-1} + \text{FFN}\left(\sum_{p\in\mathcal{P}} \text{softmax}\left(\mathbf{q}_{\mathbf{G}_g^{t-1}} \cdot \mathbf{k}_{\mathbf{S}_p^{t-1}}\right) \mathbf{v}_{\mathbf{S}_p^{t-1}}\right), \qquad (2)$$

where $\mathcal{P}$ denotes the set of all superpixels, streamlining the enhancement of group token representations by encompassing comprehensive superpixel insights.

Conversely, the G2S cross-attention phase updates superpixel features $\mathbf{S}_p^t$ by assimilating global groups information, thus ensuring that each superpixel representation benefits from a broader context:

$$\mathbf{S}_p^t = \mathbf{S}_p^{t-1} + \text{FFN}\left(\sum_{g\in\mathcal{G}} \text{softmax}\left(\mathbf{q}_{\mathbf{S}_p^{t-1}} \cdot \mathbf{k}_{\mathbf{G}_g^{t-1}}\right) \mathbf{v}_{\mathbf{G}_g^{t-1}}\right), \qquad (3)$$

where $\mathcal{G}$ represents the collective set of group tokens, highlighting the interplay between superpixel and group token features for refined segmentation.

These phases facilitate a bidirectional information flow between superpixels and groups, promoting comprehensive semantic integration across both local and global scales. This method ensures that each group captures broader contextual insights while each superpixel receives enriched contextual feedback from the global perspective, thus optimizing the overall segmentation accuracy.

To further enhance global interactions among groups, a ViT block is integrated between the S2G and G2S stages, which further boosts the model's global semantic analysis capabilities:

$$\mathbf{G}^{t'} = \text{MHSA}\left(\text{LN}\left(\mathbf{G}^t\right)\right) + \mathbf{G}^t, \qquad (4)$$

$$\tilde{\mathbf{G}}^t = \text{MLP}\left(\text{LN}\left(\mathbf{G}^{t'}\right)\right) + \mathbf{G}^{t'}, \qquad (5)$$

(a) Input Image    (b) Final Feature Map    (c) Bilinear Upsampled Feature Map    (d) Association-Aware Upsampled Feature Map
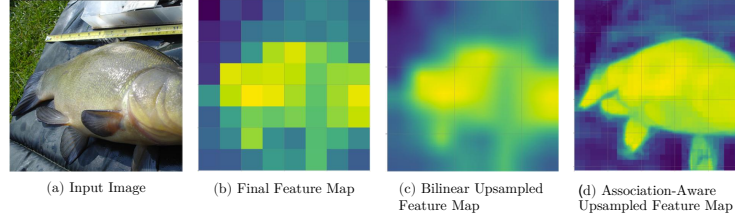
**Fig. 4: Enhanced Detail with Association-Aware Upsampling.** In contrast to the conventional bilinear upsampled feature map, our association-aware upsampled feature map achieves sharper boundary delineation and retains greater semantic detail, which is crucial for detailed segmentation tasks.

This ViT block integration, utilizing Multi-Head Self-Attention (MHSA) and Layer Norm (LN), enables the model to adeptly navigate complex global interactions at the groups level, further augmenting segmentation precision.

### 3.2   Association-Aware Upsampling

Unlike traditional hierarchical models with unidirectional information flow from fine to coarse levels [40, 41], our approach introduces bidirectional flow. This bilateral hierarchy enables data aggregation from pixels to superpixels, and then to groups, and facilitates detailed reconstruction from coarser to finer scales. This design enhances the model's ability to restore predictions to original resolution more accurately than traditional bilinear upsampling, which often lacks specificity to the data's inherent structure.

The core of this methodology lies in the attention scores detailed in Sec. 3.1, which define association matrices elucidating the intricate relationships among the pixels, superpixels, and groups. Leveraging these matrices enables the systematic upscaling of predictions from the group level ($\mathbf{O_G} \in \mathbb{R}^{g_n \times o_c}$), through the superpixel level ($\mathbf{O_S} \in \mathbb{R}^{s_h \times s_w \times o_c}$), and ultimately, back to the original pixel scale ($\mathbf{O_I} \in \mathbb{R}^{i_h \times i_w \times o_c}$):

$$\mathbf{O_S} = \mathbf{A}_{g \to p} \cdot \mathbf{O_G}, \tag{6}$$

$$\mathbf{O_I} = \mathbf{A}_{p \to i} \cdot \mathbf{O_S}. \tag{7}$$

Here, $\mathbf{A}_{g \to p}$ denotes the association matrix mapping from group token $g$ to superpixel $p$, and $\mathbf{A}_{p \to i}$ represents the matrix mapping from superpixel $p$ to pixel $i$. This procedural flow intricately enhances coarse object predictions, incrementally refining them to enrich part shapes at the superpixels level and meticulously refine boundaries at the pixel level. Through the bidirectional flow of information, our methodology not only preserves but also enhances the semantic integrity of the upscaled predictions, ensuring fidelity to the original data across all scales, as shown in Fig. 4.

### 3.3   LGFormer Architecture

The LGFormer architecture, depicted in Fig. 2, begins by extracting pixels features via a lightweight convolutional stem. Subsequently, SCA is employed to form superpixels, significantly reducing local redundancies. This enables the application of ViT blocks directly on superpixels, allowing the model to capture long-range dependencies and contextual information across superpixels.

To address the challenge of gradient conflicts, which arise from the simultaneous objectives of part and object segmentation [34], LGFormer incorporates two specialized branches designed to refine the segmentation process. In the part segmentation branch, superpixels are further processed with additional ViT blocks, classified, and then upsampled to generate part segmentation predictions through our association-aware upsampling. Concurrently, for object segmentation, superpixels are further abstracted by several ViT blocks, and undergo aggregation into groups via GCA. These groups are classified and upsampled to articulate the final object segmentation predictions, also utilizing the association-aware upsampling. This dual-branch architecture optimally balances the demands of both segmentation tasks, ensuring accurate delineation of both parts and objects within a cohesive framework.

## 4   Experiments

### 4.1   Datasets

To benchmark LGFormer, we evaluate on two benchmark datasets that include per-pixel part annotations: PartImageNet [13] and Pascal-Part [5]. PartImageNet augments 158 classes from the original ImageNet dataset with part annotations across 24,095 images. Pascal-Part is an enhancement of the VOC dataset [9] with 10,103 images across 20 classes. We focuses specifically on 16 classes with part-level annotations, following Compositor [12] protocols.

### 4.2   Implementation Details

**Hierarchical Feature Representation**   LGFormer delineates clear relationships among pixels, superpixels, and groups representations. The spatial dimensions of superpixel features are scaled down to a quarter of those of pixel features, while groups features further reduce to a 1/16 of superpixel number, which are initialized by a $4 \times 4$ average pooling from superpixels. This scaling strategy effectively balances detail retention and contextual abstraction. The model employs dual heads for superpixels and six heads for groups in SCA and GCA operations, optimizing local-global contextual interactions.

**Attention Mechanism Integration**   Cross-attention modules are integrated at strategic points within the ViT architecture, enhancing both shallow and deep layers. SCA blocks are positioned before the first and third self-attention layers, whereas GCA blocks are inserted ahead of the 9th, 10th, and

11th layers. We utilize the LayerScale technique [33] to promote stable training and faster convergence by ensuring uniform gradient distribution.

**Training Protocol**   Our training configuration mirrors the parameters set by Compositor to facilitate direct comparisons. We employ AdamW [23] with an initial learning rate of 0.0002, adjusting the ImageNet-pretrained backbone's learning rate to 10% of this value. Learning rates decrease tenfold at 90% and 95% of the training timeline. Models undergo training for 50k iterations on PartImageNet and 10k on Pascal-Part, with a batch size of 128. Data augmentation techniques include random cropping and large-scale jittering [8,10].

**Table 1:** Comparison of state-of-the-art methods on the PartImageNet and Pascal-Part validation splits.

| Method | Backbone | Params | Flops | PartImageNet | | | | Pascal-Part | | | |
| | | | | Part | | Object | | Part | | Object | |
| | | | | mIoU | mAcc | mIoU | mAcc | mIoU | mAcc | mIoU | mAcc |
|---|---|---|---|---|---|---|---|---|---|---|---|
| *Separate Training* | | | | | | | | | | | |
| DeeplabV3+ [4] | ResNet-50 [14] | 43M × 2 | 51G × 2 | 60.6 | 71.1 | 68.4 | 81.0 | - | - | - | - |
| Maskformer [6] | ResNet-50 [14] | 45M × 2 | 53G × 2 | 60.3 | 72.8 | 70.2 | 82.0 | 47.6 | 58.6 | 72.7 | 81.9 |
| SegFormer [35] | MiT-B2 [35] | 24M × 2 | 62G × 2 | 62.0 | 73.8 | 74.6 | 85.2 | - | - | - | - |
| Maskformer [6] | Swin-T [20] | 46M × 2 | 55G × 2 | 64.0 | 77.4 | 77.9 | 87.4 | 55.4 | 67.2 | 81.4 | 89.3 |
| **LGFormer** | ViT-S [7] | 34M × 2 | 50G × 2 | **69.4** | **80.0** | **80.0** | **89.3** | **57.5** | 67.2 | **85.1** | **92.0** |
| *Joint Training* | | | | | | | | | | | |
| Maskformer [6] | ResNet-50 [14] | 50M | 53G | 58.0 | 70.4 | 70.4 | 81.8 | 46.6 | 58.0 | 72.1 | 81.1 |
| Maskformer [6] | Swin-T [20] | 51M | 55G | 61.7 | 75.6 | 77.2 | 87.1 | 54.2 | 66.4 | 81.0 | 88.7 |
| Compositor [12] | ResNet-50 [14] | 50M | 54G | 61.4 | 73.4 | 71.8 | 83.0 | 48.0 | 58.8 | 74.4 | 83.8 |
| Compositor [12] | Swin-T [20] | 51M | 57G | 64.6 | 78.3 | 79.0 | 87.8 | 55.9 | 67.6 | 83.1 | 90.4 |
| **LGFormer** | ViT-S [7] | 38M | 50G | **67.4** | **79.6** | **79.8** | **88.4** | **57.4** | **67.9** | **85.1** | **91.8** |

### 4.3   Main Results

Following the experimental setup established by Compositor [12], we evaluated LGFormer in both specialized and dual-task scenarios. As shown in Tab. 1, LGFormer, when jointly trained on dual tasks, achieves a part mIoU of 67.4% and an object mIoU of 79.8% on PartImageNet. These findings represent improvements over Compositor, with increases of 2.8% and 0.8% in part and object mIoU, respectively. Further validation on Pascal-Part (Tab. 1) corroborates these advancements, with LGFormer achieving a part mIoU of 57.4% and an object mIoU of 85.1%, surpassing Compositor's performance by approximately 1.5% and 2.0%, respectively.

Considering Compositor's observation that dual-task frameworks might compromise individual task performance, we also investigated task-specific training for parts and objects, referred to as Separate Training in Tab. 1. This approach significantly improved part mIoU on PartImageNet, highlighting LGFormer's superior capability in precise part segmentation.
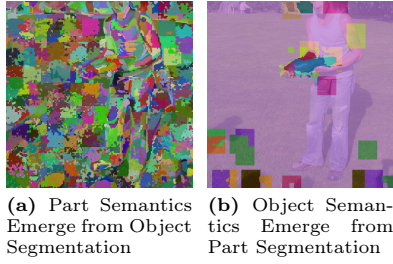
**(a)** Part Semantics Emerge from Object Segmentation

**(b)** Object Semantics Emerge from Part Segmentation

**Fig. 5: Visualization of Part and Object Semantic Emergence.** (a) In object segmentation, superpixels reveal emerging part semantics. (b) Conversely, during part segmentation, object semantics become apparent within groups.
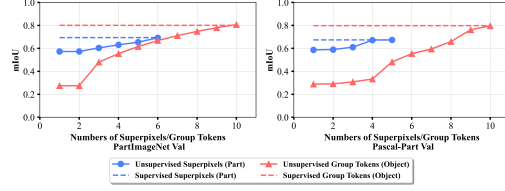


**Fig. 6. Quantitive Evaluation of Unsupervised Superpixels and Group Tokens.** Employing 6 superpixels for part segmentation and 10 groups for object segmentation in an unsupervised setting yielded mIoU scores comparable to those achieved with supervised methods.

A crucial factor in LGFormer's success is its optimized parameter use and computational efficiency. The model not only exceeds previous performance benchmarks but also does so with a lower total parameter count. This efficiency demonstrates the model's effectiveness in reducing redundancy across both local and global scales and enhancing accuracy through a sophisticated hierarchical semantic representation framework. The strategic balance between model complexity and performance underscores our approach's ability to refine segmentation outcomes without increasing computational demands.

### 4.4 Semantic Hierarchy Emergence in Part and Object Segmentation

Our model leverages a hierarchical design which mirrors natural segmentation layers of parts and objects. This structured approach prompts an examination of how LGFormer's segments, specifically superpixels and groups, handle semantic grouping under specific supervision scenarios. We investigate two key configurations: the capacity of superpixels to cluster semantically in line with objects when guided solely by part annotations and the ability of object-level supervision on groups to implicit guide the semantic understanding within superpixels. Such experiments concentrate on the object segmentation branch, as outlined in Fig. 2, where supervision is deliberately limited to part or object annotations to distinctly observe the impact of hierarchical representations on semantic discovery.

**Qualitative Emergence Evaluation.** The semantic arrangement of superpixels and groups is visually assessed by identifying the most relevant entities through the argmax across association matrices, as illustrated in Fig. 5. This approach reveals spontaneous semantic emergence: part-only supervision leads to object-level semantic recognition within groups. Inversely, object-only supervision enables the delineation of part semantics within superpixels, effectively

identifying parts as class-specific patterns [32]. Remarkably, both superpixels and groups show consistent alignment with the physical boundaries of parts or objects, even with a straightforward argmax selection. This alignment indicates that the hierarchical structuring effectively maintains detailed spatial information, which is critical in precise segmentation tasks.

**Quantitative Emergence Evaluation.** Complementing our qualitative analysis, a quantitative assessment further validates the emergence phenomena observed. We utilize an oracle setup where the model is trained with object-level annotations but evaluated superpixels against part-level ground truths, and vice versa for groups. Here, we employ the Mask-to-Attention conversion method from SegViT [43] to derive segmentation masks used for mIoU calculations. Remarkably, selecting a small subset of top-k superpixels or groups for evaluation yields mIoU scores that are on par with those obtained in fully supervised settings, as shown in Fig. 6. This experiment not only supports the model's ability to mimic aspects of human visual processing with minimal supervision but also showcases its efficiency in managing redundancies across various scales.

In summary, the evaluations both qualitative and quantitative, firmly establish that superpixels and groups are capable of semantic emergence without direct supervision. This success underscores the effectiveness of our hierarchical design in simulating the nuanced processes of the human visual system, as it categorizes and assimilates visual information into coherent entities.

### 4.5   Robustness to Occlusion

To further test the robustness of LGFormer, we evaluate its performance on the Occluded-PartImageNet-v1 dataset [12], where 20%-40% of the object region is obscured. Given the occlusions, LGFormer's performance drops 8.0% in part mIoU and 15.5% in object mIoU — positioning it favorably against benchmarks set by MaskFormer and Compositor. The qualitative results of this evaluation, illustrated in Fig. 7, further confirm LGFormer's capability to robustly handle segmentations even in scenarios involving significant occlusions, reflecting its practicality for real-world applications.
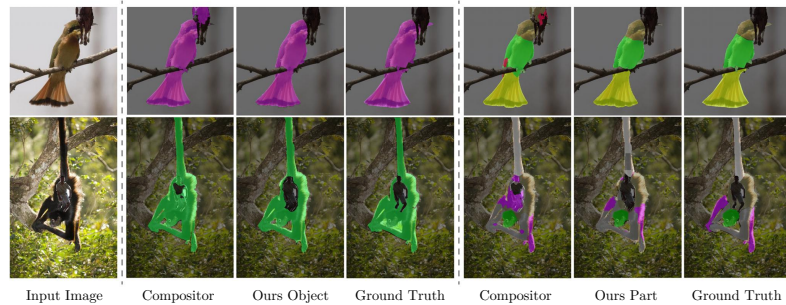


| Input Image | Compositor | Ours Object | Ground Truth | Compositor | Ours Part | Ground Truth |

**Fig. 7:** Qualitative evaluation of images with occlusions.

**Table 2:** Qualitative results on Occluded-PartImageNet-v1. LGFormer shows a smaller performance drop on occluded images compared to MaskFormer and Compositor.

| Method | Part mIoU | Object mIoU |
|---|---|---|
| MaskFormer | 50.2 (-13.7) | 56.7 (-21.2) |
| Compositor | **54.6** (-10.0) | **63.7** (-15.2) |
| **LGFormer(Ours)** | **59.3** (-8.1) | **64.2** (-15.6) |

## 4.6    Ablation Study

| Method | #Params | Part | | Object | |
|---|---|---|---|---|---|
| | | mIoU | mAcc | mIoU | mAcc |
| *Number of group tokens* | | | | | |
| 64 group | 39M | 67.4 | 79.6 | 79.8 | 88.4 |
| 256 group | 39M | 67.2 | 79.0 | 78.2 | 86.7 |
| 16 group | 39M | 67.3 | 79.1 | 78.6 | 87.2 |
| *Group token initialization method* | | | | | |
| avgpooling | 39M | 67.4 | 79.6 | 79.8 | 88.4 |
| learnable | 39M | 67.1 | 79.0 | 79.3 | 87.4 |
| conv | 41M | 66.7 | 78.2 | 79.1 | 88.0 |
| *Number of GCA stages* | | | | | |
| 3 stages | 39M | 67.4 | 79.6 | 79.8 | 88.4 |
| 4 stages | 40M | 67.0 | 78.9 | 78.5 | 87.3 |
| 2 stages | 37M | 67.1 | 79.0 | 77.7 | 86.4 |
| *Upsampling Method* | | | | | |
| Association-Aware Upsampling | 39M | 67.4 | 79.6 | 79.8 | 88.4 |
| Bilinear Upsampling | 39M | 65.3 | 76.8 | 73.2 | 82.8 |

**Table 3:** Ablation Study on PartImageNet val split.

In our ablation study, we systematically explore the design choices of LGFormer to validate the configuration's impact on segmentation performance. The results, detailed in Tab. 3, affirm the efficacy of our methodological choices by highlighting the role of group token quantity, branch block optimization, and initial group token methods.

**Optimal Group Token Count.** Adjusting the number of groups tokens demonstrates the critical balance required for precise semantic detail capture within hierarchical aggregation strategies. Reducing groups tokens to 64 or increasing to 256 affects object mIoU negatively by 1.6% and 1.2%, respectively. This phenomenon underscores that while fewer groups tokens (16) maintain

robust performance at sparse resolutions ($4 \times 4$), excessive quantities may dilute semantic richness and increase computational overhead. Our optimal count demonstrates how the model efficiently condenses semantic information without losing detail or explainability.

**Group Token Initialization.** Our comparative analysis highlights the superiority of average pooling over learnable tokens or convolution-based methods for initializing group tokens. This simplicity aligns with our methodological emphasis on efficiency, corroborating our claim that average pooling adequately prepares group tokens for subsequent hierarchical processing without necessitating complex initialization techniques.

**Branch Block Configuration.** The strategic arrangement of branch blocks within LGFormer is critical for dealing with gradient conflict and ensuring model capacity. Reducing branch blocks from three to two, or increasing them to four, adversely affects part and object segmentation mIoU by 0.3% and 2.1%, and 0.4% and 1.3%, respectively. These outcomes validate our choice of employing three branch blocks as the optimal configuration, effectively balancing effective gradient management with the preservation of hierarchical modeling capabilities.

**Upsampling Method.** The association-aware upsampling method surpasses traditional bilinear upsampling, enhancing part mIoU by 2.1% and object mIoU by 6.6%. The significant improvement in object mIoU can be attributed to the inherent limitations in bilinear upsampling, particularly its inability to recover substantial edge details lost when images are downsampled by 32. Our association-aware upsampling method progressively reconstructs details at the part and pixel levels, preserving essential information to refine prediction accuracy. This detailed recovery process significantly boosts segmentation performance, showcasing the method's capability to retain and reconstruct fine details for enhanced part and object segmentation outcomes.

## 5   Conclusion

In this paper, we introduce LGFormer, a hierarchical transformer-based model for advanced image segmentation, bridging the granularity of part segmentation with the comprehensive scope of object segmentation. Our multi-level representation strategy progresses from pixels to superpixels and finally to cohesive groups, supported by local and global aggregation strategies. Local aggregation forms superpixels aligned with object parts, while global aggregation organizes these superpixels into larger groups corresponding to entire objects. This dual framework ensures adaptability to various supervision inputs while maintaining computational efficiency and enhancing the segmentation performance.

## Acknowledgements

# References

1. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. TPAMI (2017)
2. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Transactions on Pattern Analysis and Machine Intelligence (2017)
3. Chen, L.C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. arXiv:1706.05587 (2017)
4. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation (2018)
5. Chen, X., Mottaghi, R., Liu, X., Fidler, S., Urtasun, R., Yuille, A.: Detect what you can: Detecting and representing objects using holistic models and body parts (2014)
6. Cheng, B., Schwing, A.G., Kirillov, A.: Per-pixel classification is not all you need for semantic segmentation (2021)
7. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale (2021)
8. Du, X., Zoph, B., Hung, W.C., Lin, T.Y.: Simple training strategies and model scaling for object detection (2021)
9. Everingham, M., Gool, L.V., Williams, C.K.I., Winn, J.M., Zisserman, A.: The pascal visual object classes (voc) challenge. International Journal of Computer Vision **88**, 303–338 (2010), `https://api.semanticscholar.org/CorpusID:4246903`
10. Ghiasi, G., Cui, Y., Srinivas, A., Qian, R., Lin, T.Y., Cubuk, E.D., Le, Q.V., Zoph, B.: Simple copy-paste is a strong data augmentation method for instance segmentation (2021)
11. Hatamizadeh, A., Yin, H., Heinrich, G., Kautz, J., Molchanov, P.: Global context vision transformers (2023)
12. He, J., Chen, J., Lin, M.X., Yu, Q., Yuille, A.: Compositor: Bottom-up clustering and compositing for robust part and object segmentation (2023)
13. He, J., Yang, S., Yang, S., Kortylewski, A., Yuan, X., Chen, J.N., Liu, S., Yang, C., Yu, Q., Yuille, A.: Partimagenet: A large, high-quality dataset of parts (2022)
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition (2015)
15. Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.: Mobilenets: Efficient convolutional neural networks for mobile vision applications. CoRR **abs/1704.04861** (2017)
16. Huang, H., Zhou, X., Cao, J., He, R., Tan, T.: Vision transformer with super token sampling. In: CVPR (2023)
17. Jampani, V., Sun, D., Liu, M., Yang, M., Kautz, J.: Superpixel sampling networks. In: ECCV (2018)
18. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems (2012)
19. Lake, B.M., Salakhutdinov, R., Tenenbaum, J.B.: Human-level concept learning through probabilistic program induction. Science **350**(6266), 1332–1338 (2015)

20. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows (2021)
21. Locatello, F., Weissenborn, D., Unterthiner, T., Mahendran, A., Heigold, G., Uszkoreit, J., Dosovitskiy, A., Kipf, T.: Object-centric learning with slot attention. Advances in Neural Information Processing Systems **33**, 11525–11538 (2020)
22. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR (2015)
23. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization (2019)
24. Ma, X., Zhou, Y., Wang, H., Qin, C., Sun, B., Liu, C., Fu, Y.: Image as set of points. In: ICLR (2023)
25. Mei, J., Chen, L., Yuille, A.L., Xie, C.: Spformer: Enhancing vision transformer with superpixel representation. CoRR **abs/2401.02931** (2024)
26. Mei, J., Li, Y., Lian, X., Jin, X., Yang, L., Yuille, A.L., Yang, J.: Atomnas: Fine-grained end-to-end neural architecture search. In: ICLR (2020)
27. Michieli, U., Borsato, E., Rossi, L., Zanuttigh, P.: Gmnet: Graph matching network for large scale part semantic segmentation in the wild (2020)
28. Peng, J., He, J., Kaushik, P., Xiao, Z., Mu, J., Yuille, A.: Learning part segmentation from synthetic animals. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 90–101 (2024)
29. Peng, J., Sun, Y., He, J., Chen, J., Kaushik, P., Ma, W., Zhang, Y., Wang, J., Wang, A., Yuan, X., Liu, Q., Kortylewski, A., Liu, Y., Yuille, A.: Dspart: A large-scale diffusion-generated synthetic dataset with annotations from 3d parts (2024)
30. Sandler, M., Howard, A.G., Zhu, M., Zhmoginov, A., Chen, L.: Mobilenetv2: Inverted residuals and linear bottlenecks. In: CVPR. pp. 4510–4520 (2018)
31. Singh, R., Gupta, P., Shenoy, P., Sarvadevabhatla, R.: Float: Factorized learning of object attributes for improved multi-object multi-part scene parsing (2022)
32. Tang, P., Zhang, J., Wang, X., Feng, B., Roli, F., Liu, W.: Learning extremely shared middle-level image representation for scene classification. Knowl. Inf. Syst. **52**(2), 509–530 (2017)
33. Touvron, H., Cord, M., Sablayrolles, A., Synnaeve, G., Jégou, H.: Going deeper with image transformers (2021)
34. Wang, P., Shen, X., Lin, Z., Cohen, S., Price, B., Yuille, A.: Joint object and part segmentation using deep learned potentials (2015)
35. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: Segformer: Simple and efficient design for semantic segmentation with transformers (2021)
36. Xu, J., Mello, S.D., Liu, S., Byeon, W., Breuel, T., Kautz, J., Wang, X.: Groupvit: Semantic segmentation emerges from text supervision (2022)
37. Yang, C., Xu, J., Mello, S.D., Crowley, E.J., Wang, X.: Gpvit: A high resolution non-hierarchical vision transformer with group propagation (2023)
38. Yang, F., Sun, Q., Jin, H., Zhou, Z.: Superpixel segmentation with fully convolutional networks. In: CVPR (2020)
39. Yu, Q., Wang, H., Kim, D., Qiao, S., Collins, M., Zhu, Y., Adam, H., Yuille, A., Chen, L.C.: Cmt-deeplab: Clustering mask transformers for panoptic segmentation. In: CVPR (2022)
40. Yu, Q., Wang, H., Kim, D., Qiao, S., Collins, M., Zhu, Y., Adam, H., Yuille, A., Chen, L.C.: Cmt-deeplab: Clustering mask transformers for panoptic segmentation (2022)
41. Yu, Q., Wang, H., Qiao, S., Collins, M., Zhu, Y., Adam, H., Yuille, A., Chen, L.C.: kmax-deeplab: k-means mask transformer (2023)
42. Yu, Q., Wang, H., Qiao, S., Collins, M.D., Zhu, Y., Adam, H., Yuille, A.L., Chen, L.: k-means Mask Transformer. In: ECCV (2022)

43. Zhang, B., Tian, Z., Tang, Q., Chu, X., Wei, X., Shen, C., Liu, Y.: Segvit: Semantic segmentation with plain vision transformers (2022)
44. Zhang, T., Yu, Q., Yuille, A., He, J.: Dictionary-based framework for interpretable and consistent object parsing (2024)
45. Zhang, Y., Pang, B., Lu, C.: Semantic segmentation by early region proxy. In: CVPR (2022)
46. Zhao, Y., Li, J., Zhang, Y., Tian, Y.: Multi-class part parsing with joint boundary-semantic awareness. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9177–9186 (2019)
47. Zhu, A.Z., Mei, J., Qiao, S., Yan, H., Zhu, Y., Chen, L.C., Kretzschmar, H.: Superpixel transformers for efficient semantic segmentation. IROS (2023)