

Mask Transfomer for High-Quality Instance Segmentation

Lei Ke^{1,2} Martin Danelljan¹ Xia Li¹ Yu-Wing Tai³ Chi-Keung Tang² Fisher Yu¹
¹ETH Zürich ²HKUST ³Kuaishou Technology

Abstract

Two-stage and query-based instance segmentation methods have achieved remarkable results. However, their segmented masks are still very coarse. In this paper, we present Mask Transfomer for high-quality and efficient instance segmentation. Instead of operating on regular dense tensors, our Mask Transfomer decomposes and represents the image regions as a quadtree. Our transformer-based approach only processes detected error-prone tree nodes and self-corrects their errors in parallel. While these sparse pixels only constitute a small proportion of the total number, they are critical to the final mask quality. This allows Mask Transfomer to predict highly accurate instance masks, at a low computational cost. Extensive experiments demonstrate that Mask Transfomer outperforms current instance segmentation methods on three popular benchmarks, significantly improving both two-stage and query-based frameworks by a large margin of +3.0 mask AP on COCO and BDD100K, and +6.6 boundary AP on Cityscapes. Our code and trained models will be available at <http://vis.xyz/pub/transfomer>.

1. Introduction

Advancements in image instance segmentation has largely been driven by the developments of powerful object detection paradigms. Approaches based on Mask R-CNN [12, 19, 22, 25, 32] and more recently DETR [14, 15, 21] have achieved ever increasing performance on, for instance, the COCO challenge [31]. While these methods excel in detection and localization of objects, the problem of efficiently predicting highly accurate segmentation masks has so far remained elusive.

As shown in Figure 3, there is still a significant gap between the bounding box and segmentation performance of the recent state-of-the-art methods, especially for the recent query-based methods. This strongly indicates that improvements in mask quality has not kept pace with the advancements detection capability. In Figure 2, the predicted masks of previous methods are very coarse, most often over-smoothing object boundaries. In fact, efficient and accurate mask prediction is highly challenging, due to the need for

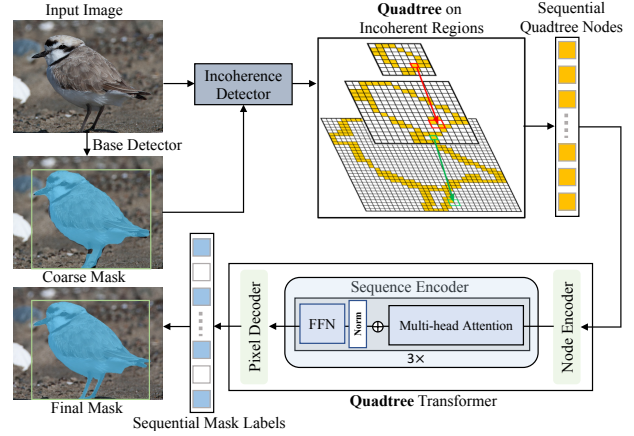


Figure 1. We propose Mask Transfomer for high-quality instance segmentation. It first builds a quadtree based on the sparse incoherent regions on the RoI pyramid and then jointly refines all tree nodes using the refinement transformer with quadtree attention.

high-resolution deep features, which demands large computational and memory costs [38].

To address these issues, we propose Mask Transfomer, an efficient transformer-based approach for high-quality instance segmentation. In Figure 1, our approach first identifies error-prone regions, which are mostly strewn along object boundaries or in high-frequency regions. To this end, our network learns to detect *incoherent regions*, defined by the loss of information when downsampling mask itself. These incoherent pixels are sparsely located, consisting only of a small portion of the total pixels. However, as they are shown to be critical to the final segmentation performance, it allows us to only process small parts of the high-resolution feature maps in the refinement process. Thus, we build a hierarchical quadtree [16] to represent and process the incoherent image pixels at multiple scales.

To refine the mask labels of the incoherent quadtree nodes, we design a refinement network based on the transformer instead of standard convolutional networks because they require operating on uniform grids. Our transformer has three modules: node encoder, sequence encoder and pixel decoder. The node encoder first enriches the feature embedding for each incoherent point. The sequence encoder then takes these encoded feature vectors across multiple quadtree levels as input queries. Finally, the pixel decoder predicts their corresponding mask labels. Comparing



Figure 2. Instance Segmentation on COCO [31] validation set by a) Mask R-CNN [19], b) BMask R-CNN [12], c) SOLQ [14], d) PointRend [25], g) Mask Transfimer (Ours) using R50-FPN as backbone, where Mask Transfimer produces significantly more detailed results at high-frequency image regions by replacing Mask R-CNN’s default mask head. Zoom in for better view.

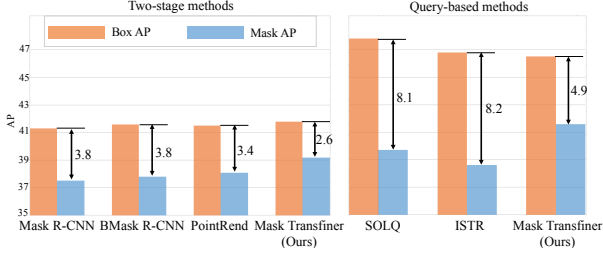


Figure 3. The performance gap between object detection and segmentation for instance segmentation models on COCO *test-dev* set using R50-FPN as backbone. Detailed comparisons are in Table 9.

to MLP [25], the sequential representation and multi-head attention enables Mask Transfimer to flexibly takes as input sparse feature points across levels in parallel, models their pixel-wise relations, and then propagates information among them even in a long distance range.

We extensively analyze our approach on COCO, Cityscapes and BDD100K benchmarks, where quantitative and qualitative results show that Mask Transfimer not only outperforms existing two-stage and query-based methods, but also is efficient in computation and memory cost compared to standard transformer usages. We establish a new state-of-the-art result on COCO *test-dev* of 41.6 AP^{Mask} using ResNet-50, outperforming most recent SOLQ [14] and QueryInst [15] by a significant margin.

2. Related Work

Instance Segmentation Two-stage instance segmentation methods [2, 6, 8, 19, 22, 23, 28, 29] first detects bounding boxes and then performing segmentation in each RoI region. Mask R-CNN [19] extends Faster R-CNN [34] with an FCN branch. The follow-up works [7, 12, 24, 32, 33] also contribute to the family of Mask R-CNN models. One-stage methods [5, 8, 26, 27] and kernel-based method [48], such as PolarMask [44], YOLOACT [1], and SOLO [40, 41] remove the proposal generation and feature re-pooling steps, achieving comparable results with higher efficiency.

Query-based instance segmentation methods [14, 15, 17, 21, 42], which are inspired by DETR [4], have emerged very recently by treating segmentation as a set prediction problem. These methods use queries to represent the interested objects and jointly perform classification, detection and mask regression on them. In [14, 21], the object masks

are compressed as encoding vectors using DCT or PCA algorithms, while QueryInst [15] adopts dynamic mask heads with mask information flow. However, the large gaps between the detection and segmentation performance in Figure 3 reveals that the mask quality produced by these query-based methods are still unsatisfactory. In contrast to the above methods, Mask Transfimer is targeted for high-quality instance segmentation. In our efficient transformer the input queries are incoherent pixels nodes, instead of representing the objects. Our method is applicable to and effective in both the two-stage and query-based frameworks.

Refinement for Instance Segmentation Most existing works on instance segmentation refinement rely on specially designed convolutional networks [36, 47] or MLPs [25]. PointRend [25] samples feature points with low-confidence scores and refines their labels with a shared MLP, where the selected points are determined by the coarse predictions of the Mask R-CNN. RefineMask [47] incorporates fine-grained features with an additional semantic head as the guidance. The post-processing method BPR [36] crops boundary patches of images and initial masks as input and use [38] for segmentation. Notably some methods [11, 35, 38, 46] focus on refining semantic segmentation details. However, it is much more challenging for instance segmentation due to the more complex segmentation setting, with varying number of objects per image and the requirement of delineating similar and overlapping objects.

Compared to these refinement methods, Mask Transfimer is an end-to-end instance segmentation method, using a transformer for correcting errors. The regions to be refined are predicted using a lightweight FCN, instead of non-deterministic sampling based on mask scores [25]. Different from the MLP in [25], the sequential and hierarchical input representation enables Mask Transfimer to efficiently take non-local sparse feature points as input queries, where the strong global processing of transformers is a natural fit for our quadtree structure.

3. Mask Transfimer

We propose an approach to efficiently tackle high-quality instance segmentation. The overall architecture of Mask Transfimer is depicted in Figure 5. From the base object detection network, *e.g.* Mask R-CNN [19], we employ a multi-scale deep feature pyramid. The object detection head then

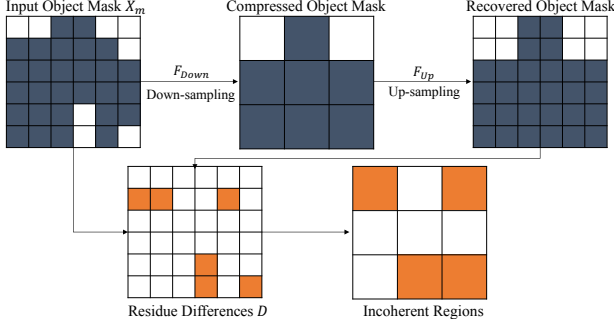


Figure 4. Illustration on incoherent regions definition by simulating mask information loss.

predicts bounding boxes as instance proposals. This component also generates a coarse initial mask prediction at low resolution. Given this input data, our aim is to predict highly accurate instance segmentation masks.

Since much of the segmentation errors are attributed to the loss of spatial resolution, we first define such *incoherent regions* and analyze their properties in Section 3.1. To identify and refine incoherent regions in multiple scales, we employ a quadtree, discussed in Section 3.2. The lightweight incoherent region detector takes as input the coarse initial mask alongside the multi-scale features, and predicts the incoherent regions for each scale in a cascaded manner. This allows ours Mask Transfiner to save huge computational and memory burdens, because only a small part of the high-resolution image features are processed by the refinement network itself. Our refinement transformer, detailed in Section 3.3, operates in the detected incoherent regions. Since it operates on feature points on the constructed quadtree, and not in a uniform grid, we design a transformer architecture which jointly processes all incoherent nodes in all levels of the quadtree. Finally, we present the training strategy of Mask Transfiner along with the implementation details.

3.1. Incoherent Regions

Much of the segmentation errors produced by existing instance segmentation methods [14, 19] are due to the loss of spatial resolution, such as the mask downsampling operations, small RoI pooling size, and coefficients compression [14, 21], where mask prediction itself is performed at a coarse feature scale. Despite its efficiency, low spatial resolution makes it challenging to predict accurate object boundaries, due to the loss of high-frequency details. In this section, we first define *incoherent regions*, where mask information is lost due to reduced spatial resolution. Then, by analyzing their properties, we observe that a large portion of the errors are indeed located in these regions.

Definition of Incoherent Regions To identify incoherent regions, we simulate the loss of information due to downsampling in the network by also downsampling the mask itself. Specifically, information is lost in regions where the

mask cannot be correctly reconstructed by a subsequent up-sampling step, as illustrated in Figure 4. Formally, let M_l be a binary ground-truth instance mask of an object at scale level l . The resolution at each scale level differs by a factor of 2, where $l = 0$ is the finest and $l = L$ is the coarsest scale. We denote $2\times$ nearest neighbor down and up-sampling by \mathcal{S}_\downarrow and \mathcal{S}_\uparrow respectively. The incoherent region at scale l is then the binary mask achieved as,

$$D_l = \mathcal{O}_\downarrow(M_{l-1} \oplus \mathcal{S}_\uparrow(\mathcal{S}_\downarrow(M_{l-1}))). \quad (1)$$

Here, \oplus denotes the logical ‘exclusive or’ operation and \mathcal{O}_\downarrow is $2\times$ downsampling by performing the logical ‘or’ operation in each 2×2 neighborhood. A pixel (x, y) is thus incoherent $D_l(x, y) = 1$ if the original mask value M_{l-1} differs from its reconstruction in at least one pixel in the finer scale level. Intuitively, incoherent regions are mostly strewn along object instance boundaries or high-frequency regions, consisting of points with missing or extra predicted wrong labels by coarse masks. We provide the visualizations of them in Figure 5 and Supp. file, which are sparsely and non-contiguously distributed on a typical image.

Table 1. Experimental analysis of the incoherent regions on COCO *val* set. Percent denotes the area ratio of incoherent regions in the object bounding boxes. Recall_{Err} is the ratio for all wrongly predicted pixels per object. Acc is the accuracy rate for coarse mask predictions inside incoherent regions. AP_{Coarse} is measured by using coarse mask predictions for whole object regions while AP_{GT} only fills the incoherent regions with the ground truth labels.

Percent	Recall _{Err}	Acc	AP _{GT}	AP _{Coarse}
14%	43%	56%	51.0	35.5

Properties of Incoherent Regions In Table 1, we provide an analysis of the incoherent regions defined above. It shows that a large portion of prediction errors are concentrated in these incoherent regions, occupying 43% of all wrongly predicted pixels, while only taking 14% to the corresponding bounding box areas. The accuracy of the coarse mask prediction in incoherent regions is 56%. By fixing the bounding boxes detector, we conduct an oracle study to fill all these incoherent regions for each object with ground truth labels, while leaving the remaining parts as initial mask predictions. Compared to using initial mask predictions in the incoherent regions, the performance surges from 35.5 AP to 51.0 AP, indeed justifying they are critical for improving final performance.

3.2. Quadtree for Mask Refinement

In this section, we describe our approach for detecting and refining incoherent regions in the image. Our approach is based on the idea of iteratively detecting and dividing the incoherent regions in each feature scale. By only splitting the identified incoherent pixels for further refinement, our approach efficiently processes high-resolution features

by only focusing on the important regions. To formalize our approach, we employ a quadtree structure to first identify incoherent regions across scales. We then predict the refined segmentation labels for all incoherent nodes in the quadtree, using our network detailed in Section 3.3. Finally, our quadtree is employed to fuse the new predictions from multiple scales by propagating the corrected mask probabilities from coarse to finer scales.

Detection of Incoherent Regions The right part of Figure 5 depicts the design of our lightweight module to efficiently detect incoherent regions on a multi-scale feature pyramid. Following a cascaded design, we first concatenate the smallest features and coarse object mask predictions as input, and use a simple fully convolutional network (four 3×3 Convs) followed by a binary classifier to predict the coarsest incoherence masks. Then, the detected lower-resolution masks are upsampled and fused with the larger-resolution feature in neighboring level to guide the finer incoherence predictions, where only single 1×1 convolution layer is employed. During training, we enforce the groundtruth incoherent points in lower-level generated by Eq. 1 within the coverage of their parent points in higher-level feature map.

Quadtree Definition and Construction We define a *point quadtree* for decomposing the detected incoherent regions. Our structure is illustrated in Figure 5, where one yellow point in higher-level of FPN feature (such as feature resolution 28×28) has four quadrant points in its neighboring lower-level FPN feature map (such as resolution 56×56). These are all feature points but with different granularities because they are on different pyramid levels. In contrast to the conventional quadtree ‘cells’ used in computer graphics, where a quadtree ‘cell’ can have multiple points, the subdivision unit for our point quadtree is always on a single point, with the division of points decided by the detected incoherent values and the threshold for the binary classifier.

Based on the detected incoherent points, we construct a multi-level hierarchical quadtree, beginning from using the detected points in the highest-level feature map as root nodes. These root nodes are selected for subdividing to their four quadrants on the lower-level feature map, with larger resolution and more local details. Note that at the fine level, only the quadrant points detected as incoherent could make a further break down and the expansion of incoherent tree nodes is restricted in regions corresponding to the incoherent predictions at the previous coarse level.

Quadtree Refinement We refine the mask predictions of the incoherent nodes of the quadtree using a transformer-based architecture. Our design is described in Sec. 3.3. It directly operates on the nodes of the quadtree, jointly providing refined mask probabilities at each incoherent node.

Quadtree Propagation Given the refined mask predictions, we design a hierarchical mask propagation scheme that exploits our quadtree structure. Given the initial coarse masks

predictions in low-resolution, Mask Transfmer first corrects the points labels belong to the root level of the quadtree, and then propagates these corrected point labels to their corresponding four quadrants in neighboring finer level by nearest neighbor interpolation. The process of labels correction is efficiently conducted on the incoherent nodes in a level-wise manner until reaching the finest quadtree level. Comparing to only correcting the labels of finest leaf nodes on the quadtree, it enlarges the refinement areas with negligible cost by propagating refinement labeled to leaf nodes of the intermediate tree levels.

3.3. Mask Transfmer Architecture

In this section, we describe the architecture of the refinement network, which takes as input the incoherent points on the built quadtree (Section 3.2) for final segmentation refinement. These points are sparsely distributed along the high-frequency regions across levels and not spatially contiguous. Thus, standard convolutional networks operating on uniform grids are not suitable. Instead, we design a refinement transformer, Mask Transfmer, that corrects the predictions of all incoherent quadtree nodes in parallel.

Accurately segmenting ambiguous points requires both fine-grained deep features and coarse semantic information. The network therefore needs strong modeling power to sufficiently relate points and their surrounding context, including both spatial and cross-level neighboring points. Thus, a transformer, which can take sequential input and perform powerful local and non-local reasoning through the multi-head attention layers, is a natural choice for our Mask Transfmer design. Compared to the MLP in [25], the strong global processing of transformers is a natural fit for our quadtree structure. It benefits the effective fusion of the multi-level feature points information with different granularities and the explicit modeling of pairwise point relations.

Figure 5 shows the overall architecture of our Mask Transfmer. Based on the hierarchical FPN [30], instance segmentation is tackled in a multi-level and coarse-to-fine manner. Instead of using single-level FPN feature for each object [19], Mask Transfmer takes as input sequence the sparsely detected feature points in incoherent image regions across the RoI feature pyramid levels, and outputs the corresponding segmentation labels.

RoI Feature Pyramid Given an input image, the CNN backbone network equipped with FPN first extracts hierarchical feature maps for downstream processing, where we utilize feature levels from P_2 to P_5 . The base object detector [14, 19] predicts bounding boxes as instance proposals. Then the RoI feature pyramid is built by extracting RoI features across three different levels $\{P_i, P_{i-1}, P_{i-2}\}$ of FPN with increasing square sizes $\{28, 56, 112\}$. The starting level i is computed as $i = \lfloor i_0 + \log_2(\sqrt{WH}/224) \rfloor$, where $i_0 = 4$, W and H are the RoI width and height. The coarsest

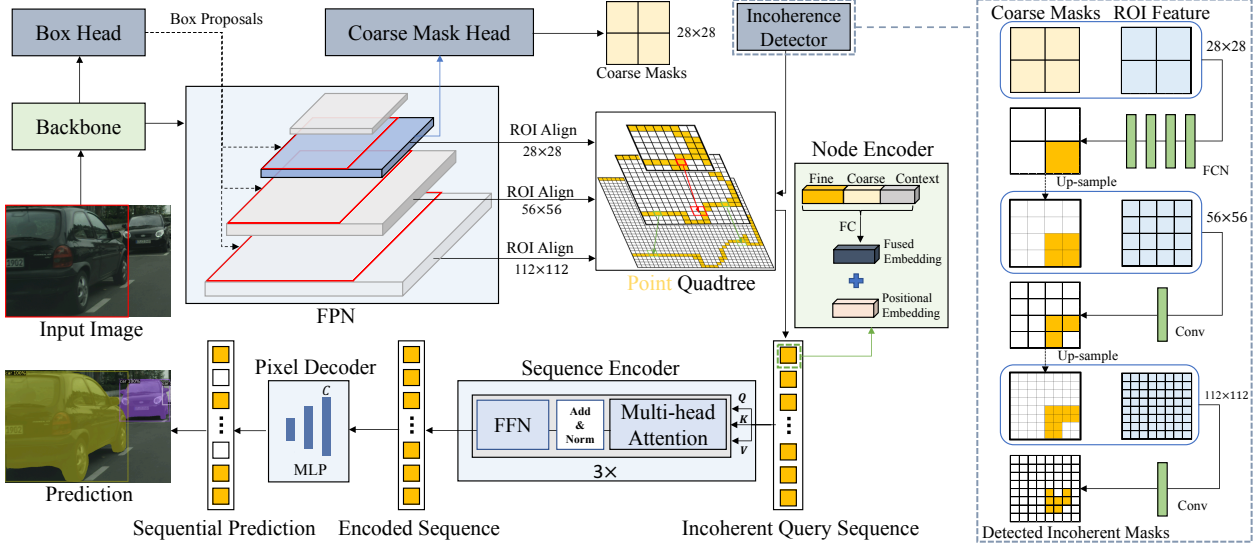


Figure 5. The framework of Mask Transfimer. On the point quadtree, yellow point grids denote detected incoherent regions requiring further subdivision to four quadrants. The incoherent query sequence is composed of points across three levels of the quadtree for joint refinement. The encoder of Transfimer consists of node encoder and sequence encoder, while the pixel decoder is on top of each self-attended query pixel and output their final labels. The incoherence detector is detailed in the right part of the figure with detections on multi-level incoherent regions (Yellow). The higher-resolution detection is under the guidance of the predicted incoherent mask up-sampled from lower level.

level features contain more contextual and semantic information, while the finer levels resolve more local details.

Input Node Sequence Given the quadtree discussed in Section 3.2 along with the associated FPN features for each node, we construct the input sequence for our transformer-based architecture. The sequence consists of all incoherent nodes from all three levels of the quadtree. The resulting sequence thus has a size of $C \times N$, where N is the total number of nodes and C is the feature channel dimension. Notably, $N \ll HW$ due to the high degree of sparsity. Moreover, the ordering of the sequence does not matter due to the permutation invariance of transformer. In contrast to standard transformer encoder, the encoder of Transfimer has two parts: the node encoder and the sequence encoder.

Node Encoder To enrich the incoherent points feature, the node encoder of Mask Transfimer encodes each quadtree node using the following four different information cues: **1)** The fine-grained features extracted from corresponding location and level of the FPN pyramid. **2)** The initial coarse mask prediction from the base detector provides region-specific and semantic information. **3)** The relative positional encoding in each RoI encapsulates spatial distances and relations between nodes, capturing important local dependence and correlations. **4)** The surrounding context for each node captures local details to enrich the information. For each node, we use features extracted from the 3×3 neighborhood, compressed by a fully connected layer. Intuitively, this helps in localizing edges and boundaries, as well as capturing the local shape of the object. As illustrated in Figure 5, the fine-grained features, coarse segmentation cues and context features are first concatenated and fused

by a FC layer to original feature dimension. The positional embedding is then added to the resulting feature vector.

Sequence Encoder and Pixel Decoder Then, the sequence transformer encoder of Transfimer jointly processes the encoded nodes from all levels in the quadtree. The transformer thus performs both global spatial and inter-scale reasoning. Each sequence encoder layer has a standard transformer structure, formed by a multi-head self-attention module and a fully connected feed forward network (FFN). To equip the incoherent points sequence with adequate positive and negative references, we also use all feature points from the coarsest FPN level with small size 14×14 . Different from the standard transformer decoder [4] with deep attention layers, the pixel decoder in Mask Transfimer is a small two-layer MLP, which decodes the output query for each node in the tree, in order to predict the final mask labels.

Training and inference Based on the constructed quadtree, we develop flexible and adaptive training and inference schemes for Mask Transfimer, where all detected incoherent nodes across quadtree levels are formed into a sequence for parallel prediction. During inference, to obtain final object masks, Mask Transfimer follows the quadtree propagation scheme (Section 3.2) after obtaining the refined labels for incoherent nodes. During training, the whole Mask Transfimer framework can be trained in an end-to-end manner. We employ a multi-task loss,

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{Detect}} + \lambda_2 \mathcal{L}_{\text{Coarse}} + \lambda_3 \mathcal{L}_{\text{Refine}} + \lambda_4 \mathcal{L}_{\text{Inc}}. \quad (2)$$

Here, $\mathcal{L}_{\text{Refine}}$ denotes the refinement with L1 loss between the predicted labels for incoherent nodes and their ground-truth labels. A Binary Cross Entropy loss \mathcal{L}_{Inc} is for detecting incoherent regions. The detection loss $\mathcal{L}_{\text{Detect}}$ in-

cludes the localization and classification losses from the base detector, *e.g.* Faster R-CNN [34] or DETR detector. Finally, $\mathcal{L}_{\text{Coarse}}$ represents the loss for the initial coarse segmentation prediction used by [19]. $\lambda_{\{1,2,3,4\}}$ are hyper-parameter weights $\{1.0, 1.0, 1.0, 0.5\}$.

Implementation Details Mask Transfinner is implemented on both the two-stage detector Faster R-CNN [34] and query-based detector [4]. We design a 3-level quadtree and use the hyper-parameters and training schedules of Mask R-CNN implemented in Detectron2 [43] for the backbone and coarse mask head. The Mask Transfinner encoder consists of three standard transformer layers. Each layer has four attention heads with feature dimension at 256. In our ablation study, R-50-FPN [20] and Faster R-CNN with $1\times$ learning schedule are adopted. For COCO leaderboard comparison, we adopt the scale-jitter with shorter image side randomly sampled from [640, 800], following training schedules in [8, 24, 27]. More details are in the Supp. file.

4. Experiments

4.1. Experimental Setup

COCO We perform experiments on COCO dataset [31], where we train our networks on 2017train and evaluate our results on both the 2017val and 2017test-dev. We employ the standard AP metrics and the recently proposed boundary IoU metrics [10]. Notably, AP^B for boundary IoU is a measure focusing on boundary quality. Following [25], we also report AP^* , which evaluates the val set of COCO with significantly higher-quality LVIS annotations [18] that can better reveal improvements in mask quality.

Cityscapes We report the results on Cityscapes [13], a high-quality instance segmentation dataset containing 2975, 500, 1525 images with resolution of 2048×1024 for training, validation and test respectively. Cityscapes focus on self-driving scenes with 8 categories (*e.g.*, car, person, bicycle).

BDD100K We further train and evaluate Mask Transfinner on the BDD100K [45] instance segmentation dataset, which has 8 categories with 120K high-quality instance mask annotations. We follow the standard practice, using 7k, 1k, 2k images for training, validation and testing respectively.

4.2. Ablation Experiments

We conduct detailed ablation studies on the COCO validation set, analyzing the impact of the proposed incoherent regions and individual components of Mask Transfinner.

Effect of the Incoherent Regions Table 1 presents an analysis on the properties of incoherent regions described in Section 3.1. It reveals they are critical to the final segmentation performance. Table 2 presents analyzes the effectiveness of the detected incoherent regions by replacing the refinement regions with full RoIs or detected object boundary regions. Due to memory limitation, the full RoIs only uses output size 28×28 . The comparison shows the advantage

Table 2. Effect of the incoherent regions on COCO val set. AP^B is evaluated Boundary IoU [10] while AP^* uses LVIS annotations.

Region Type	AP	AP^B	AP^*	AP^*_{50}
Full RoIs (28×28)	35.5	21.4	38.3	59.5
Boundary regions	36.6	23.8	40.1	60.2
Incoherent regions	37.3	24.2	40.5	60.7
Incoherent regions (w/o L_1)	36.5	23.5	39.8	59.7
Incoherent regions (w/o L_2)	36.8	23.8	40.2	60.1
Incoherent regions (w/o L_3)	36.7	23.6	40.0	59.9

Table 3. Effect of lower-level masks guidance in detecting incoherent regions on COCO val. AP and AP^B are final performance.

Lower-level Guidance	Acc	Recall	AP	AP^B
	79%	73%	36.6	23.7
✓	84%	86%	37.3	24.2

Table 4. Analysis of node encoding cues on COCO val set.

Fine	Coarse	Pos.	Context	AP	AP^B	AP^*	AP^*_{50}
✓				33.8	20.1	37.0	53.8
✓	✓			34.2	20.4	37.3	54.3
✓	✓	✓		36.8	23.9	40.1	60.1
✓	✓	✓	✓	37.3	24.2	40.5	60.7

of incoherent regions, with 1.8 AP and 0.7 AP gain over the use of full RoIs and detected boundary regions respectively.

To study the influence of incoherent regions on different pyramid levels, in Table 2, we also perform ablation experiments by removing the refinement regions of the Mask Transfinner in a level-wise order. We find that all three levels are beneficial to the final performance, while L_1 contributes most with 0.8 AP increase, where L_1 denotes the root level of Mask Transfinner with the smallest feature size.

Ablation on the Incoherent Regions Detector We evaluate the performance of the light-weight incoherent region detector by computing its recall and accuracy rates. In Table 3, with the guidance of the predicted incoherent mask up-sampled from lower level (Figure 5), the recall rate of detected incoherent regions has an obvious improvement from 74% to 86%, and the accuracy rate also increases from 79% to 84%. Note that recall rate is more important here to cover all the error-prone regions for further refinements.

Effect of Incoherent Points Encoding We analyze the effect of the four information cues in the incoherent points encoding. In Table 4, comparing to only using the fine-grained feature, the coarse segmentation features with semantic information brings a gain of 0.4 point AP. The positional encoding feature has a large influence on model performance by significantly improving 2.6 points on AP and 3.5 points on AP^B respectively. The positional encoding for incoherent points are crucial, because transformer architecture is permutation-invariant and the segmentation task is position-sensitive. The surrounding context feature further promotes the segmentation results from 36.8 AP to 37.3 AP by aggregating local neighboring details.

Influence of Quadtree Depths In Table 5, we study the influence on hierarchical refinement stages by constructing the quadtree in our Mask Transfinner with different depths.

Table 5. Analysis of the quadtree depth on the COCO *val* using R50-FPN as backbone.

Depth	Output size	AP	AP*	AP _L	AP _M	AP _S	FPS
0	28×28	35.2	37.6	50.3	37.7	17.2	12.3
1	28×28	35.5	38.4	50.9	38.1	17.2	10.6
2	56×56	36.2	39.1	51.9	38.7	17.3	8.9
3	112×112	37.3	40.5	52.9	39.5	17.5	7.1
4	224×224	37.1	40.7	53.1	39.3	17.4	5.2

Table 6. Mask Transfuser vs. MLP and CNN on COCO *val* set using ResNet-50-FPN.

Model	AP	AP ^B	AP*	AP ₅₀ *
CNN (full regions, 56 × 56)	35.7	21.8	38.7	58.8
MLP (full regions, 56 × 56)	36.1	23.4	39.2	59.2
MLP (PointRend [25], 112 × 112)	36.2	23.1	39.1	59.0
MLP (incoherent regions)	36.4	23.7	39.7	59.8
Mask Transfuser (D = 3, H = 4)	37.3	24.2	40.5	60.7
Mask Transfuser (D = 3, H = 8)	37.1	24.1	40.2	60.8
Mask Transfuser (D = 6, H = 4)	37.4	24.4	40.6	60.9

Table 7. Efficacy of Transfuser compared to standard attention models on COCO *val*. NLA denotes non-local attention [39].

Model	AP	FLOPs (G)	Memory (M)	FPS
NLA [39] (112×112)	36.3	24.6	8347	4.6
NLA [39] (224×224)	36.6	80.2	18091	2.4
Transformer [4] (28×28)	36.1	37.2	4368	6.9
Transformer [4] (56×56)	36.5	68.3	17359	2.1
Mask Transfuser (112×112)	37.3	16.8	2316	7.1
Mask Transfuser (224×224)	37.1	38.1	4871	5.2

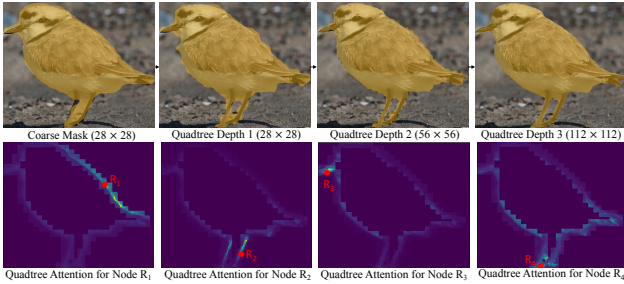


Figure 6. Qualitative results comparison between the coarse mask predictions by our baseline [20] and the refinement results with various depths of the quadtree built on detected incoherent regions. The bottom row visualizes the quadtree attention weights distribution in the sparse incoherent regions for four sampled red nodes.

Depth 0 denotes the baseline using coarse head mask prediction **w/o** refinement steps. The output size grows twice larger than its preceding stage. By varying the output sizes from 28×28 to 224×224, the mask AP* increases from 38.4 to 40.7 with increased tree depth. This reveals that models with more levels and larger output sizes for an object indeed brings more gain to segmentation performance. The large objects benefit most from the increasing sizes with an improvement of 2.8 point in AP_L. We further find that the performance saturates when the output size is larger than 112×112, while the 3-stage Transfuser also has a lower computational cost and runs at 7.1 fps. Figure 6 visualizes results with increasing quadtree depths, where masks become substantially finer detail around object boundaries.

Mask Transfuser vs. MLP and CNN We compare different popular choices of the refinement networks, including the MLP and CNN structures. MLP is implemented with three hidden layers of 256 channels [25], while CNN is a FCN with four convolution layers with 3×3 kernels [19]. Note that for full refinement regions, CNN and MLP are limited to the RoI size 56 × 56 due to memory limitations, and CNN is not suitable for incoherent regions because uniform grids are required. In Table 6, our Mask Transfuser outperforms the MLP by 0.9 AP, benefiting from the non-local pixel-wise relation modeling, where we use the same incoherent regions on all three quadtree levels for fair comparison. Moreover, we investigate the influence of layer depth *D* and width *W* of Mask Transfuser and find that deeper and wider attention layers only lead to minor performance change. In Figure 6, we visualize the sparse quadtree attention maps of the last sequence encoder layer of the Transfuser, focusing on a few incoherent points. The encoder already seems to distinguish between foreground instances

and background, where the neighboring attended regions of point *R*₁ are separated by the object boundary.

Efficacy of Quadtree Structure Table 7 compares Mask Transfuser with different attention mechanisms. Compared to pixels relation modeling using 3-layer non-local attention [39] or standard transformer [4, 37], Mask Transfuser not only obtains higher accuracy but also is very efficient in computation and memory consumption. For example, Mask Transfuser with multi-head attention uses 3 times less memory than the non-local attention given same output size, due to the small number of incoherent pixels. Compared to standard transformer operating on full RoI regions of much smaller size 56×56, the quadtree subdivision and inference allows Mask Transfuser to produce a high-resolution 224×224 prediction using only half of the FLOPs computation. Note that the standard transformer with output size 112×112 runs out of memory in our experiments.

Effect of Multi-level Joint Refinement Given incoherent nodes from the 3-level quadtree, Transfuser forms all of them into a sequence for joint refinement in single forward pass. In Table 8, we compare it with separately refining the quadtree nodes on each level with multiple sequences. The performance boost of 0.6 AP* shows the benefit of multi-scale feature fusion and richer context in global reasoning.

Effect of Quadtree Mask Propagation During inference, after Mask Transfuser has refined all incoherent points, we utilize a hierarchical coarse-to-fine mask propagation scheme along the quadtree levels to obtain the final predictions. Comparing to only correcting the labels of finest leaf nodes on the quadtree in Table 8, the propagation enlarges the refinement areas and improves the performance from 36.5 AP to 37.0 AP. The propagation brings negligible computation because the new labels for the quadrant leaf (coherent) nodes in intermediate tree levels are obtained via duplicating the refined label values of their parents.

Table 8. Effect of the multi-level joint refinement (MJR) and quadtree mask propagation (QMP) on COCO *val* set.

MJR	QMP	AP	AP ^B	AP*	AP ₅₀ *
		36.5	23.7	39.6	59.7
✓		36.9	23.9	40.2	60.2
	✓	37.0	24.0	40.1	60.2
✓	✓	37.3	24.2	40.5	60.7

4.3. Comparison with State-of-the-art

We compare our approach with the state-of-the-art methods on the benchmarks COCO, Cityscapes and BDD100K,

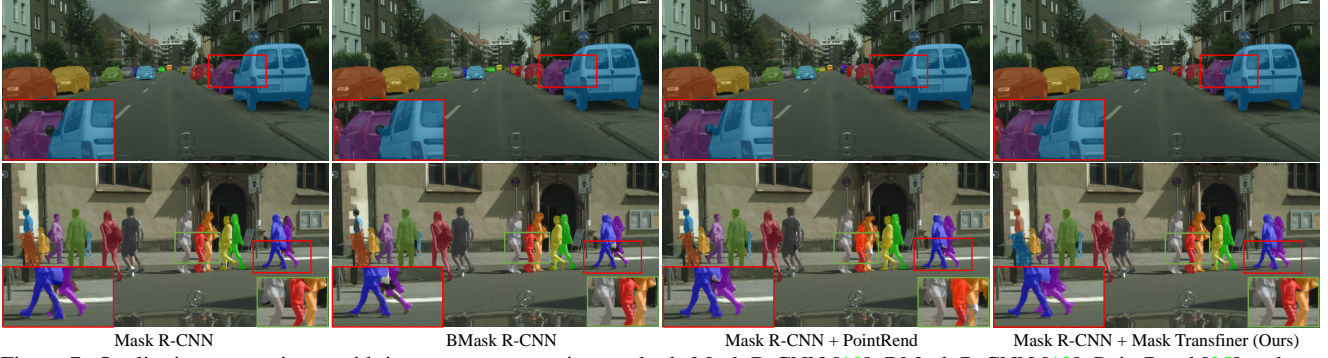


Figure 7. Qualitative comparisons with instance segmentation methods Mask R-CNN [19], BMask R-CNN [12], PointRend [25] and our Mask Transfomer on Cityscapes *val* set. Mask Transfomer produces more natural boundaries while revealing details for small parts, such as the rear mirrors of the car and the high-heeled shoes. Zoom in for better view. Refer to the supplemental file for more visual comparisons.

Table 9. Comparison with SOTA methods on COCO *test-dev* and *val* set. All methods are trained on COCO *train2017*. †: trained with DCN [49]. AP* denotes evaluation using LVIS [18] annotation and AP^B denotes using Boundary IoU [10].

Method	Backbone	Type	AP	AP* _{val}	AP ^B _{val}	AP ^{Box}	AP _S	AP _M	AP _L
Mask R-CNN [19]	R50-FPN	Two-stage	37.5	38.2	21.2	41.3	21.1	39.6	48.3
PointRend [25]	R50-FPN	Two-stage	38.1	39.7	23.5	41.5	18.8	40.2	49.4
B-MRCNN [12]	R50-FPN	Two-stage	37.8	39.8	23.5	41.6	19.7	40.3	49.6
BPR [36]	R50-FPN	Two-stage	38.4	40.2	24.3	41.3	20.2	40.5	49.7
Mask Transfimer	R50-FPN	Two-stage	39.4	42.3	26.0	41.8	22.3	41.2	50.2
Mask Transfimer†	R50-FPN	Two-stage	40.5	43.1	26.8	43.2	22.8	42.3	52.5
Mask R-CNN [19]	R101-FPN	Two-stage	38.8	39.3	23.1	43.1	21.8	41.4	50.5
PointRend [25]	R101-FPN	Two-stage	39.6	41.4	25.3	43.3	19.8	42.6	53.7
MS R-CNN† [22]	R101-FPN	Two-stage	39.6	41.1	25.0	44.1	18.9	42.7	55.1
HTC [6]	R101-FPN	Two-stage	39.7	42.5	25.4	45.9	21.0	42.2	53.5
RefineMask [47]	R101-FPN	Two-stage	39.4	42.3	26.8	43.8	21.6	42.0	53.1
BCNet [24]	R101-FPN	Two-stage	39.8	41.9	26.1	43.5	22.7	42.4	51.1
Mask Transfimer	R101-FPN	Two-stage	40.7	43.6	27.3	43.9	23.1	42.8	53.8
Mask Transfimer†	R101-FPN	Two-stage	42.2	45.0	28.6	45.8	24.1	44.8	55.4
ISTR [21]	R50-FPN	Query-based	38.6	39.5	23.0	46.8	22.1	40.4	50.6
QueryInst [15]	R50-FPN	Query-based	39.9	42.1	25.1	44.5	22.9	41.7	51.9
SOLQ [14]	R50-FPN	Query-based	39.7	39.8	23.3	47.8	21.5	42.5	53.1
Mask Transfimer	R50-FPN	Query-based	41.6	45.4	28.2	46.5	24.2	44.6	55.2

Table 10. Performance comparison between two-stage instance segmentation methods on Cityscapes *val* set using R50-FPN.

Method	AP ^B	AP ^B ₅₀	AP	AP ₅₀
Mask R-CNN (Baseline) [19]	11.4	37.4	33.8	61.5
PointRend [25]	16.7	47.2	35.9	61.8
BMask R-CNN [12]	15.7	46.2	36.2	62.6
Panoptic-DeepLab [9]	16.5	47.7	35.3	57.9
RefineMask [47]	17.4	49.2	37.6	63.3
Mask Transfimer (Ours)	18.0	49.8	37.9	64.1

Table 11. Performance comparison between instance segmentation methods on BDD100K *val* set.

Method	Backbone	AP _{mask}	AP _{box}
Mask R-CNN (Baseline) [19]	R101-FPN	20.5	26.1
Cascade Mask R-CNN [3]	R101-FPN	19.8	24.7
Mask R-CNN + DCNv2 [49]	R101-FPN	20.9	26.0
HRNet [38]	HRNet-w32	22.5	28.2
Mask Transfimer (Ours)	R101-FPN	23.6	26.2

where Mask Transfimer outperforms all existing methods without bells and whistles, demonstrating efficacy on

both two-stage and query-based segmentation frameworks. Codes and models will be released upon publication.

COCO Table 9 compares Mask Transfimer with state-of-the-art instance segmentation methods on COCO dataset. Transfimer achieves consistent improvement on different backbones and object detectors, demonstrating its effectiveness by outperforming RefineMask [47] and BCNet [24] by 1.3 AP and 0.9 AP using R101-FPN and Faster R-CNN, and exceeding QueryInst [15] by 1.7 AP using query-based detector [4]. Note QueryInst consists of six-stage refinement in parallel with far more parameters to optimize. Besides, we find that Transfimer using Faster R-CNN and R50-FPN with much lower object detection performance still achieves comparable segmentation results with query-based methods [14, 21] on mask AP, and over 2 points gain in boundary AP^B, further validating the higher AP achieved by Transfimer is indeed contributed by the fine-grained masks.

Cityscapes The results of Cityscapes benchmark is tabulated in Table 10, where Mask Transfimer achieves the best

mask AP 37.6 and boundary AP^B 18.0. Our approach significantly surpasses existing SOTA methods, including PointRend [25] and BMask R-CNN [12] by a margin of 1.3 AP^B and 2.3 AP^B using the same Faster R-CNN detector. Compared to our baseline Mask R-CNN [19], Transfimer greatly improves the boundary AP from 11.4 to 18.0, which shows the effectiveness of the quadtree refinement.

BDD100K Table 11 shows results on BDD100K dataset, where Mask Transfimer obtains the highest AP_{mask} of 23.5 and outperforms the baseline [20] by 3 points under the comparable AP_{Box} . The significant advancements reveals the high accuracy of the predicted masks by Transfimer.

Qualitative Results Figure 7 shows qualitative comparisons on Cityscapes, where our Mask Transfimer produces masks with substantially higher precision and quality than previous methods [12, 19, 25], especially for the hard regions, such as the small rear mirrors and high-heeled shoes. Refer to supplementary file for more visual comparisons.

5. Conclusion

We present Mask Transfimer, a new high-quality and efficient instance segmentation method. Transfimer first detects and decomposes the image regions to build a hierarchical quadtree. Then, all points on the quadtree are transformed into to a query sequence for our transformer to predict final labels. In contrast to previous segmentation methods using convolutions limited by uniform image grids, Mask Transfimer produces high-quality masks with low computation and memory cost. We validate the efficacy of Transfimer on both the two-stage and query-based segmentation frameworks, and show that Transfimer achieves large performance advantages on COCO, Cityscapes and BDD100K. A current limitation is the fully supervised training required by our Mask Transfimer as well as the competing methods. Future work will strive towards relaxing this assumption.

6. Appendix

We first provide more implementation and training/inference details of Mask Transfimer on three instance segmentation benchmarks (Section 6.1). Then we conduct more experimental analysis and discussion of comparison between Mask Transfimer and other methods (Section 6.2). We further present more qualitative results comparisons on COCO [31], BDD100K [45] and Cityscapes [13] datasets in various scenes (Section 6.3). Finally, we visualize quadtree attention weights, detected incoherent regions and segmentation results with various quadtree depths (Section 6.4), including failure cases analysis.

6.1. More Implementation Details

Implementation and Training/Inference Details We implement Mask Transfimer based on Detectron2 [43],

where SGD is used with 0.9 momentum and 1K constant warm-up iterations. The weight decay is set to 0.0001. On the two-stage and query-based frameworks, we employ Mask Transfimer using Faster R-CNN [34] and DETR [4] detectors respectively while leaving the RoI pyramid construction and refinement transformer unchanged.

To make the detection on incoherent regions more robust, we adopt jittering operations along the boundaries of the ground truth incoherent regions because in our case, the recall rate of the detection to cover all the incoherent regions play a more critical role in influencing final performance. We use 0.5 as the threshold for the binary incoherence classifier. For the experiment of Table 2 in the paper, the boundary regions are pixels within two-pixel Euclidean distance to the detected object mask contours on all three levels of the object feature pyramid, where the object boundary detector [24] is used. The coarse mask head is composed of a FCN network with four 3×3 Convs attached on the ROI feature of size 28×28 .

During training, we randomly permute the order of the incoherent points for each object and select 300 of them (100 per quadtree level), so as to maintain the same sequence length for each object for batch efficiency. We adopt the horizontal flipping and scale data augmentation during training following [25].

During inference, no test-time augmentation is used. We employ a hierarchical propagation scheme based on the quadtree structure from coarse to finer scales (detailed in Section 3.1 of the paper) and the refined incoherent nodes predictions. In Figure 8, we further illustrate the mask propagation process with a simplified 3-level quadtree. The incoherent nodes number N with their refined predictions value V_n are formatted in $N : V_n$, where $\{1 : v1, 2 : v2, 5 : v5, 7 : v7, 8 : v8, 10 : v10, 13 : v13\}$ are incoherent nodes numbers and prediction values pairs in our given example. We break down the mask correction and propagation into 3 steps corresponding to 3 levels of the quadtree with visualizations. Comparing to only correcting the labels of finest leaf nodes on the quadtree, it enlarges the refinement areas with negligible cost by propagating refinement labeled to leaf nodes $\{3, 4, 6, 9, 11, 12\}$. We validate the effect of quadtree mask propagation in Table 8 of the paper.

COCO: We set 16 images per mini-batch. Following [25], our training schedule is 60k / 20k / 10k with updating learning rates 0.02 / 0.002 / 0.0002 respectively. For ablation study, our method is trained on four GPUs using ResNet-50, where we use SGD for optimization and set initial learning rate to 0.01 with total batch size 8. We train Mask Transfimer for 12 epochs (taking about 8 hours with NVIDIA RTX 2080 Ti), and decrease the learning rate by 0.1 after 8 and 11 epochs.

Cityscapes: We adopt 8 images per mini-batch and the training schedule is 18k / 6k updates at learning rates of 0.01

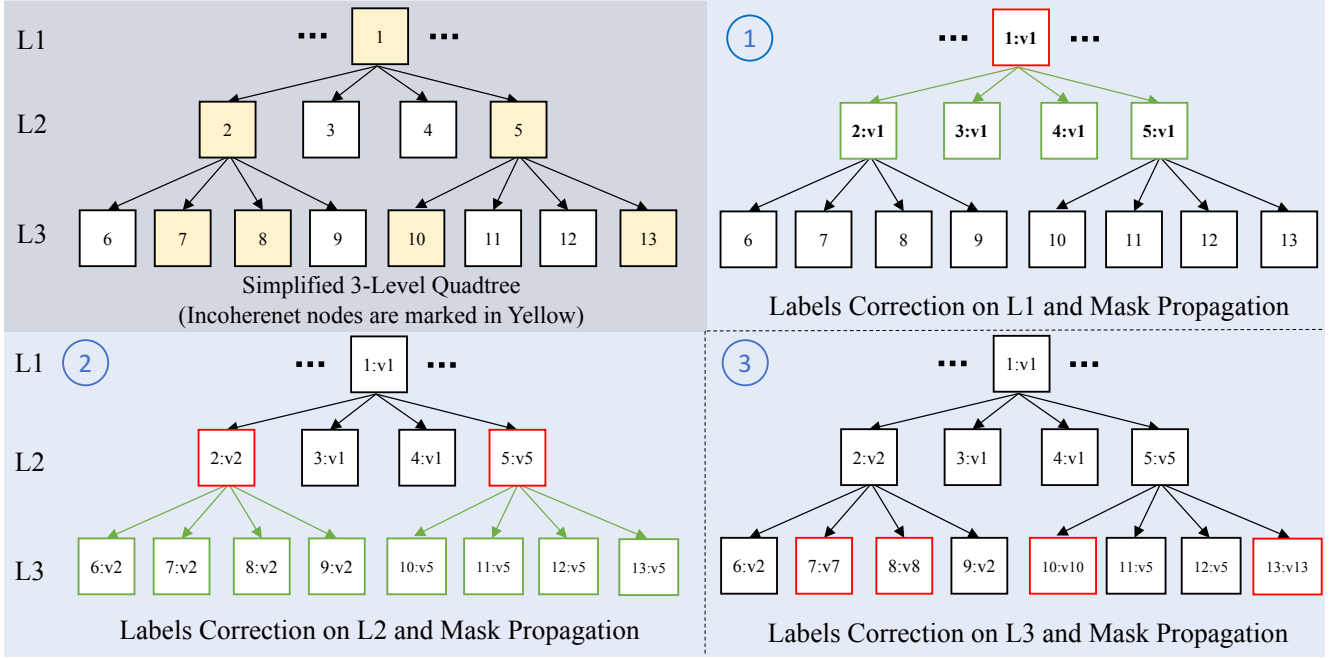


Figure 8. The simplified illustration of mask propagation on a 3-level quadtree during inference. Given the detected incoherent nodes and their refined predictions, in Step 1, Mask Transfomer corrects the nodes labels belonging to L1 level of the quadtree, and then propagates these corrected labels to their corresponding four quadrants in L2 level. In Step 2, the process of labels correction is efficiently conducted on the incoherent nodes in L2 and further propagating to L3. This process is recursive until reaching the finest quadtree level.

/ 0.001 respectively. During training, the images are resized randomly to a shorter edge from [800, 1024] pixels with a step of 32 pixels. The inference images are resized to a shorter edge size of 1024 pixels. For Cityscapes evaluation, we train the models on the fine annotations of the train set with 64 epochs following [12, 25].

BDD100K: We use 16 images per mini-batch and the training schedule is 22k / 4k / 4k updates at learning rates of 0.02 / 0.002 / 0.0002 respectively. During training, the images are resized randomly to a shorter edge from [600, 720] pixels with a step of 24 pixels. During inference, the images are resized to a shorter edge size of 720 pixels. Note all compared methods are trained with the same schedules and image size settings.

6.2. More Experimental Analysis

Accuracy Comparison In Table 9 of the main text, we compare the accuracy of Mask Transfomer with previous methods and find that Mask Transfomer achieves consistently large improvements on different backbones and object detectors. We further observe that the usage of DCN [49] with Mask Transfomer can bring a surge in performance. We compare Transfomer with Mask Scoring R-CNN [22] trained with DCN under the same setting and training schedules. Using ResNet-101 and Faster R-CNN [34] detector, the mask AP of Mask Transfomer on COCO *test-dev* is 42.2, while Mask Scoring R-CNN is 39.6 in Table 9 of the paper. For more comprehensive comparisons on two-stage instance segmentation methods, in Table 12, we also train

Mask R-CNN [19], PointRend [25], BCNet [24], Cascade Mask R-CNN [3] and HTC [6] with the multi-scale $3\times$ training schedule with DCN, and submit their predictions to the evaluation server for obtaining their accuracies on the test-dev split. The performance advantages of Mask Transfomer are consistently significant, improving the baseline Mask R-CNN[†] for 2.8 mask AP and outperforming PointRend by 0.9 AP.

Table 12. Performance comparison between two-stage instance segmentation methods on COCO *test-dev* set using R101-FPN. The dagger [†] denotes training with DCN [49] and $3\times$ training schedule in our implementation. HTC and Cascade Mask R-CNN use 3-stage cascade refinement with multiple object detectors and mask heads. The standard transformer with output size 112×112 runs out of memory in our experiments.

Method	Output Size	AP	AP _S	AP _M	AP _L	FPS
Mask R-CNN [†] [19] (Baseline) [19]	28×28	39.4	18.6	42.8	54.5	9.6
Mask Scoring R-CNN [†] [22]	28×28	39.6	18.9	42.7	55.1	9.2
BCNet [†] [24]	28×28	41.2	23.6	43.9	52.8	8.9
PointRend [†] [25]	224×224	41.3	20.6	44.0	55.3	7.2
Cascade Mask R-CNN [†] [3]	28×28	41.5	22.1	42.6	54.2	4.8
HTC [†] [6]	28×28	41.7	23.3	44.2	53.8	2.1
Standard Transformer [†]	56×56	41.3	23.4	43.5	53.2	1.4
Mask Transfomer [†] (Ours: Quadtree Transformer)	112×112	42.2	24.1	44.8	55.4	6.1

Inference Speed We adopt frames per second (FPS) to evaluate the inference speed of the models. In Table 12, we benchmark all the compared two-stage methods using a Titan RTX GPU. The reported FPS is the average obtained in five runs, where each run measures the FPS of a model through 200 iterations. Compared to the Cascade Mask R-CNN and HTC with three-stage cascade refinement and multiple object detectors/mask heads (output

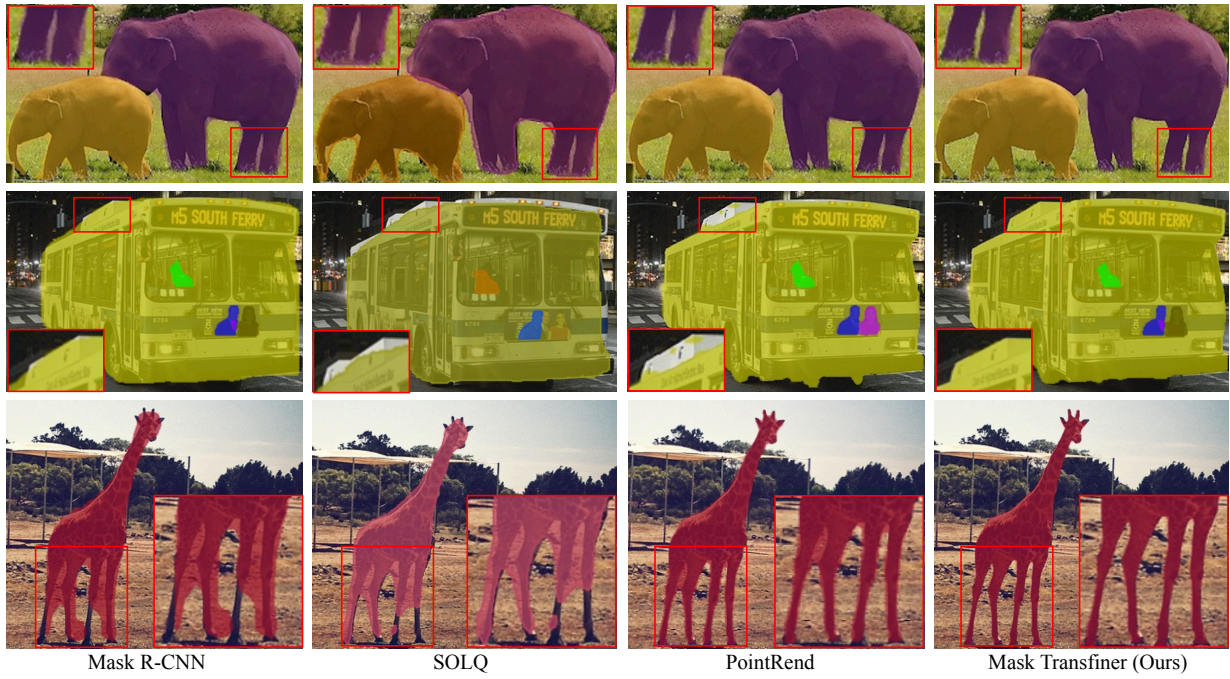


Figure 9. Instance Segmentation on COCO [31] validation set by a) Mask R-CNN [19], b) SOLQ [14], c) PointRend [25], d) Mask Transfmer (Ours) using R50-FPN as backbone, where Mask Transfmer produces significantly more detailed results at high-frequency image regions by replacing Mask R-CNN’s default mask head. Zoom in for better view.



Figure 10. Qualitative comparisons with baseline method Mask R-CNN [19] and our Mask Transfmer on BDD100K [45] val set. Mask Transfmer produces more correct and natural segmentation results by revealing details for high-frequency regions.

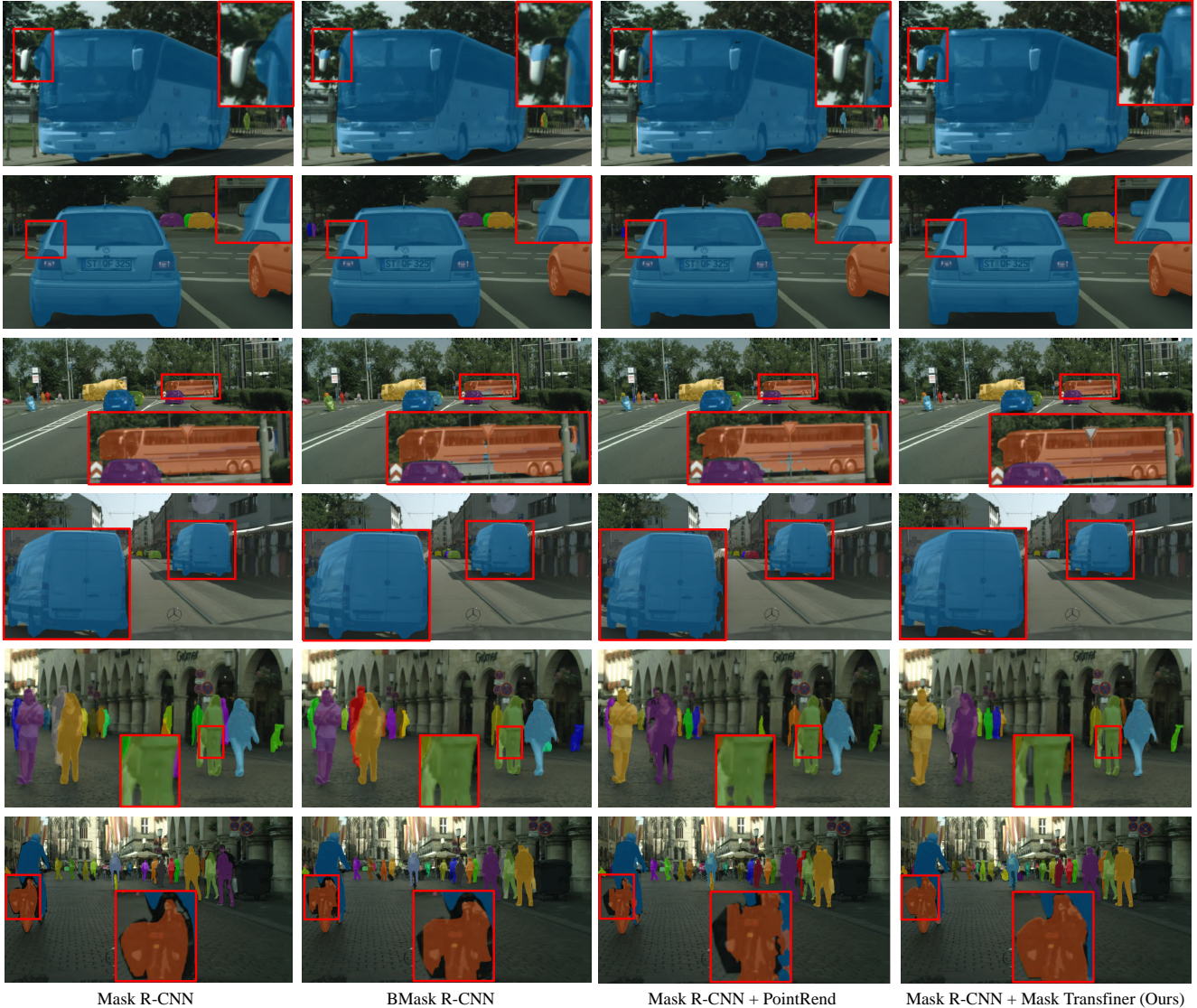


Figure 11. Qualitative comparisons with instance segmentation methods Mask R-CNN [19], BMask R-CNN [12], PointRend [25] and our Mask Transfiner on Cityscapes [13] *val* set. Mask Transfiner produces more precise and natural segmentation results, where even the small triangle-shaped traffic sign occluding the bus (3rd row) and the gap between the hand and leg (5th row) could be correctly separated.

size 28×28), our Transfiner using 3-level quadtree is much faster and more accurate with higher-resolution predictions (112×112). Comparing to the baseline Mask R-CNN, although there is a drop on inference speed for about 35% due to multi-head attention modeling between hierarchical incoherent regions, the significant performance boost of 2.8 mask AP and 4 times larger output height/width are good compensation trade-offs. Note that standard transformer (3 layers and 4 attention heads in each layer) operating on uniform grids with output size 56×56 only runs at 1.4 FPS, which is much slower than our method.

6.3. More Qualitative Comparisons

We provide more qualitative results comparisons on three evaluation benchmarks COCO (Figure 9), B100K

(Figure 10) and Cityscapes (Figure 11), where our Mask Transfiner consistently produces masks with substantially higher precision and quality than previous methods [12, 14, 19, 25]. Take the third case in Figure 9 as an example, SOLQ and the baseline Mask R-CNN only provides very coarse mask predictions in the high-frequency regions, such as the giraffe’s head and feet regions, due to their low-resolution output sizes 28×28 . Although PointRend employs large output size 224×224 , it still fails to delineate the thin gap between the left legs of giraffe. Note that the mask output size of Mask Transfiner only is 112×112 . These segmentation errors on ambiguous regions reveal the limitation of segmenting each pixel separately only by a share MLP [25] without global reasoning and hierarchical pixel-wise relations modeling.

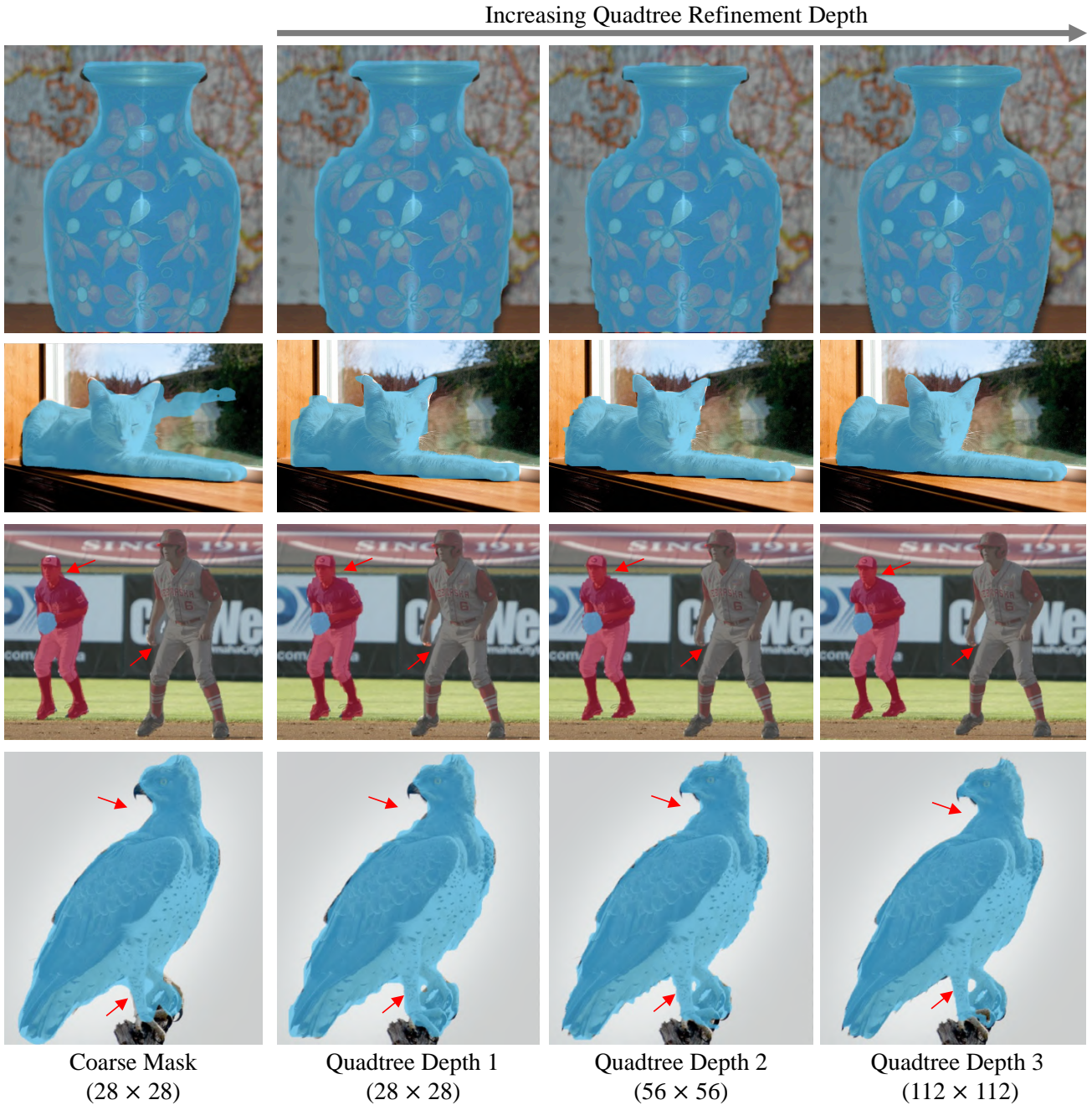


Figure 12. Qualitative results comparison between the coarse mask predictions by our baseline [20] and the refinement results of Mask Transfuser on COCO with various depths of the quadtree built on detected incoherent regions.

6.4. Visual Analysis

Visualization Multi-level Refinement In Figure 12, we analyze how the mask predictions evolve with increasing quadtree depths. The predicted masks become substantially finer in detail around object boundaries, which reveals that the quadtree nodes with more levels at larger output sizes for an object preserves more low-level details for fine-grained segmentation.

Failure Cases We also analyze the failure cases and find one typical failure mode shown in the last row of Figure 12,

where a small portion of the bird’s paw is wrongly predicted as background wood due to their highly similar appearance and texture.

Visualization on Quadtree Attention and Incoherent Regions In Figure 13 and Figure 14, we further provide more quadtree attention visualization examples and their detected incoherent regions on RoI pyramid, where the outline of objects can be observed and the sparsity of quadtree attention is clearly shown. The quadtree nodes with higher appearance or positional similarity has larger attention weights attending between them.

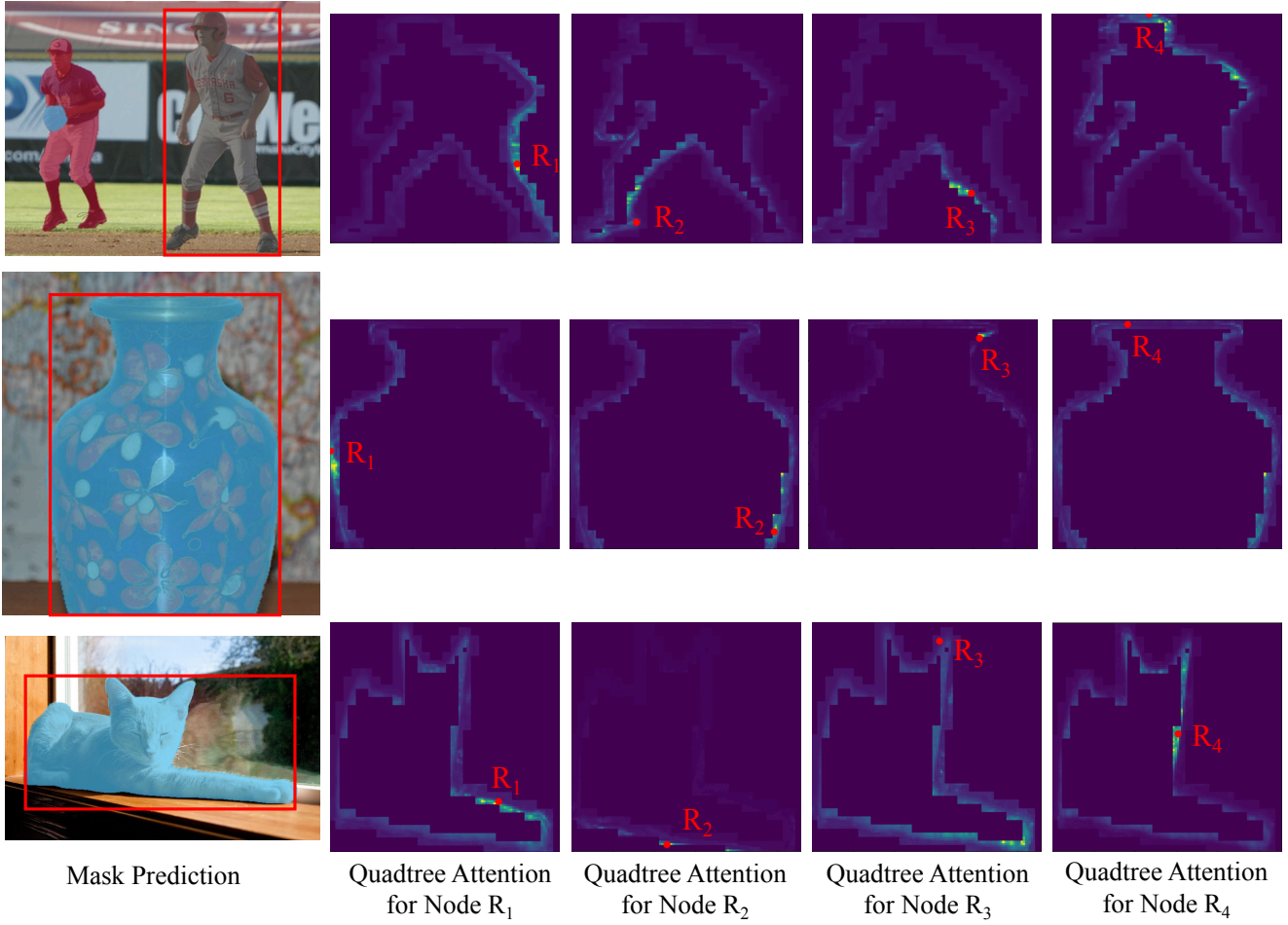


Figure 13. Visualization on the quadtree attention weights distribution in the sparse incoherent regions for four sampled red nodes.



Figure 14. Visualization of detected incoherent regions on different levels of the constructed quadtree based on the RoI pyramid, where the incoherent nodes regions in deeper quadtree levels with larger resolution size are distributed more sparsely.

References

- [1] Daniel Bolya, Chong Zhou, Fanyi Xiao, and Yong Jae Lee. Yolact: real-time instance segmentation. In *ICCV*, 2019. 2
- [2] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *CVPR*, 2018. 2
- [3] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: High quality object detection and instance segmentation. 2019. 8, 10
- [4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 2, 5, 6, 7, 8, 9
- [5] Hao Chen, Kunyang Sun, Zhi Tian, Chunhua Shen, Yongming Huang, and Youliang Yan. BlendMask: Top-down meets bottom-up for instance segmentation. In *CVPR*, 2020. 2
- [6] Kai Chen, Jiangmiao Pang, Jiaqi Wang, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jianping Shi, Wanli Ouyang, et al. Hybrid task cascade for instance segmentation. In *CVPR*, 2019. 2, 8, 10
- [7] Liang-Chieh Chen, Alexander Hermans, George Papandreou, Florian Schroff, Peng Wang, and Hartwig Adam. Masklab: Instance segmentation by refining object detection with semantic and direction features. In *CVPR*, 2018. 2
- [8] Xinlei Chen, Ross Girshick, Kaiming He, and Piotr Dollár. Tensormask: A foundation for dense object segmentation. In *ICCV*, 2019. 2, 6
- [9] Bowen Cheng, Maxwell D Collins, Yukun Zhu, Ting Liu, Thomas S Huang, Hartwig Adam, and Liang-Chieh Chen. Panoptic-deeplab: A simple, strong, and fast baseline for bottom-up panoptic segmentation. In *CVPR*, 2020. 8
- [10] Bowen Cheng, Ross Girshick, Piotr Dollár, Alexander C Berg, and Alexander Kirillov. Boundary iou: Improving object-centric image segmentation evaluation. In *CVPR*, 2021. 6, 8
- [11] Ho Kei Cheng, Jihoon Chung, Yu-Wing Tai, and Chi-Keung Tang. Cascadepsp: toward class-agnostic and very high-resolution segmentation via global and local refinement. In *CVPR*, 2020. 2
- [12] Tianheng Cheng, Xinggang Wang, Lichao Huang, and Wenyu Liu. Boundary-preserving mask r-cnn. In *ECCV*, 2020. 1, 2, 8, 9, 10, 12
- [13] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016. 6, 9, 12
- [14] Bin Dong, Fangao Zeng, Tiancai Wang, Xiangyu Zhang, and Yichen Wei. Solq: Segmenting objects by learning queries. In *NeurIPS*, 2021. 1, 2, 3, 4, 8, 11, 12
- [15] Yuxin Fang, Shusheng Yang, Xinggang Wang, Yu Li, Chen Fang, Ying Shan, Bin Feng, and Wenyu Liu. Instances as queries. In *ICCV*, 2021. 1, 2, 8
- [16] Raphael A Finkel and Jon Louis Bentley. Quad trees a data structure for retrieval on composite keys. *Acta informatica*, 4(1):1–9, 1974. 1
- [17] Ruohao Guo, Dantong Niu, Liao Qu, and Zhenbo Li. Sotr: Segmenting objects with transformers. In *ICCV*, 2021. 2
- [18] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *CVPR*, 2019. 6, 8
- [19] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017. 1, 2, 3, 4, 6, 7, 8, 9, 10, 11, 12
- [20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 6, 7, 9, 13
- [21] Jie Hu, Liujuan Cao, Yao Lu, ShengChuan Zhang, Ke Li, Feiyue Huang, Ling Shao, and Rongrong Ji. Istr: End-to-end instance segmentation via transformers. *arXiv preprint arXiv:2105.00637*, 2021. 1, 2, 3, 8
- [22] Zhaojin Huang, Lichao Huang, Yongchao Gong, Chang Huang, and Xinggang Wang. Mask scoring r-cnn. In *CVPR*, 2019. 1, 2, 8, 10
- [23] Lei Ke, Xia Li, Martin Danelljan, Yu-Wing Tai, Chi-Keung Tang, and Fisher Yu. Prototypical cross-attention networks for multiple object tracking and segmentation. In *NeurIPS*, 2021. 2
- [24] Lei Ke, Yu-Wing Tai, and Chi-Keung Tang. Deep occlusion-aware instance segmentation with overlapping bilayers. In *CVPR*, 2021. 2, 6, 8, 9, 10
- [25] Alexander Kirillov, Yuxin Wu, Kaiming He, and Ross Girshick. Pointrend: Image segmentation as rendering. In *CVPR*, 2020. 1, 2, 4, 6, 7, 8, 9, 10, 11, 12
- [26] Weicheng Kuo, Anelia Angelova, Jitendra Malik, and Tsung-Yi Lin. Shapemask: Learning to segment novel objects by refining shape priors. In *ICCV*, 2019. 2
- [27] Youngwan Lee and Jongyoul Park. Centermask: Real-time anchor-free instance segmentation. In *CVPR*, 2020. 2, 6
- [28] Yi Li, Haozhi Qi, Jifeng Dai, Xiangyang Ji, and Yichen Wei. Fully convolutional instance-aware semantic segmentation. In *CVPR*, 2017. 2
- [29] Justin Liang, Namdar Homayounfar, Wei-Chiu Ma, Yuwen Xiong, Rui Hu, and Raquel Urtasun. Polytransform: Deep polygon transformer for instance segmentation. In *CVPR*, 2020. 2
- [30] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017. 4
- [31] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 1, 2, 6, 9, 11
- [32] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *CVPR*, 2018. 1, 2
- [33] Kemal Oksuz, Baris Can Cam, Emre Akbas, and Sinan Kalkan. Rank sort loss for object detection and instance segmentation. In *ICCV*, 2021. 2
- [34] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2015. 2, 6, 9, 10

- [35] Towaki Takikawa, David Acuna, Varun Jampani, and Sanja Fidler. Gated-scnn: Gated shape cnns for semantic segmentation. In *ICCV*, 2019. 2
- [36] Chufeng Tang, Hang Chen, Xiao Li, Jianmin Li, Zhaoxiang Zhang, and Xiaolin Hu. Look closer to segment better: Boundary patch refinement for instance segmentation. 2021. 2, 8
- [37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 7
- [38] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *TPAMI*, 2020. 1, 2, 8
- [39] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, 2018. 7
- [40] Xinlong Wang, Tao Kong, Chunhua Shen, Yuning Jiang, and Lei Li. Solo: Segmenting objects by locations. *arXiv preprint arXiv:1912.04488*, 2019. 2
- [41] Xinlong Wang, Rufeng Zhang, Tao Kong, Lei Li, and Chunhua Shen. Solov2: Dynamic and fast instance segmentation. In *NeurIPS*, 2020. 2
- [42] Yuqing Wang, Zhaoliang Xu, Xinlong Wang, Chunhua Shen, Baoshan Cheng, Hao Shen, and Huaxia Xia. End-to-end video instance segmentation with transformers. In *CVPR*, 2021. 2
- [43] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. <https://github.com/facebookresearch/detectron2>, 2019. 6, 9
- [44] Enze Xie, Peize Sun, Xiaoge Song, Wenhai Wang, Xuebo Liu, Ding Liang, Chunhua Shen, and Ping Luo. Polarmask: Single shot instance segmentation with polar representation. In *CVPR*, 2020. 2
- [45] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *CVPR*, 2020. 6, 9, 11
- [46] Yuhui Yuan, Jingyi Xie, Xilin Chen, and Jingdong Wang. Segfix: Model-agnostic boundary refinement for segmentation. In *ECCV*, 2020. 2
- [47] Gang Zhang, Xin Lu, Jingru Tan, Jianmin Li, Zhaoxiang Zhang, Quanquan Li, and Xiaolin Hu. Refinemask: Towards high-quality instance segmentation with fine-grained features. In *CVPR*, 2021. 2, 8
- [48] Wenwei Zhang, Jiangmiao Pang, Kai Chen, and Chen Change Loy. K-net: Towards unified image segmentation. In *NeurIPS*, 2021. 2
- [49] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results. In *CVPR*, 2019. 8, 10