

Coarse-to-Fine Amodal Segmentation with Shape Prior

Jianxiong Gao¹, Xuelin Qian^{1,†}, Yikai Wang¹, Tianjun Xiao^{2,†}, Tong He², Zheng Zhang², Yanwei Fu¹
¹Fudan University, ²Amazon Web Service

jxgao22@m.fudan.edu.cn, {xlqian, yikaiwang19, yanweifu}@fudan.edu.cn
{tianjux, htong, zhaz}@amazon.com

Abstract

Amodal object segmentation is a challenging task that involves segmenting both visible and occluded parts of an object. In this paper, we propose a novel approach, called Coarse-to-Fine Segmentation (C2F-Seg), that addresses this problem by progressively modeling the amodal segmentation. C2F-Seg initially reduces the learning space from the pixel-level image space to the vector-quantized latent space. This enables us to better handle long-range dependencies and learn a coarse-grained amodal segment from visual features and visible segments. However, this latent space lacks detailed information about the object, which makes it difficult to provide a precise segmentation directly. To address this issue, we propose a convolution refine module to inject fine-grained information and provide a more precise amodal object segmentation based on visual features and coarse-predicted segmentation. To help the studies of amodal object segmentation, we create a synthetic amodal dataset, named as MOViD-Amodal (MOViD-A), which can be used for both image and video amodal object segmentation. We extensively evaluate our model on two benchmark datasets: KINS and COCO-A. Our empirical results demonstrate the superiority of C2F-Seg. Moreover, we exhibit the potential of our approach for video amodal object segmentation tasks on FISHBOWL and our proposed MOViD-A. Project page at: <https://jianxgao.github.io/C2F-Seg>.

1. Introduction

Amodal instance segmentation [24] aims to extract complete shapes of objects in an image, including both visible and occluded parts. This task plays a vital role in various real-world applications such as autonomous driving [25, 10], robotics [4], and augmented reality [23, 20]. For instance, in autonomous driving, partial understanding of the scene may result in unsafe driving decisions.



Figure 1. Visualization of predicted amodal masks in KINS and COCOA by C2F-Seg. Images in the top row are from COCOA, while the others are from KINS.

Typically, existing approaches [21, 29, 33, 13, 9] build new modules on the detection framework, by additionally introducing an amodal branch that predicts the complete mask perception of the target object. The central idea lies in imbuing a holistic understanding of shape (*i.e.*, shape prior) through multi-task learning by harnessing the supervised signals of the visible and full regions. While these approaches have yielded promising outcomes in recent years, the task of amodal segmentation remains fraught with challenges. One of the main challenges of amodal segmentation is that it is an ill-posed problem, meaning that there are many non-unique and reasonable possibilities for perceiving occluded areas, particularly for elastic bodies like people and animals. On the other hand, there are intricate categories and shapes of objects in real-world scenarios, which would pose significant challenges to prior learning of shapes.

In this paper, we advocate that shape priors are essen-

†: Co-corresponding authors.

tial for amodal segmentation, since the shape of an object is usually determined by its function, physiology, and characteristics. For example, a carrot has a long shape, while an apple has a round shape. Thus, the potential distribution of this object can be learned via neural networks. Nevertheless, we argue that while shape prior can only provide a basic outline and may not capture individual differences or highly local information. Meanwhile, it is possible for the shape prior to being inconsistent with the observed visible area due to factors like pose and viewpoint. To this end, we in this paper propose to generate amodal segments progressively via a coarse-to-fine manner. Specifically, we divide the segmentation of amodal object into two phases: a coarse segmentation phase where we use the shape prior to generate a plausible amodal mask, and a refinement phase is adopted to refine the coarse amodal mask to get the precise segmentation.

In the coarse segmentation phase, as we only need to provide a coarse mask, we perform the segmentation in a low-dimension vector-quantized latent space to reduce the learning difficulty and accelerate the inference process. The segmentation in such latent space is resorted to the popular mask prediction task adopted in BERT [15] and MaskGIT [3]. Specifically, we adopt a transformer model which takes as inputs the ResNet visual feature, the vector-quantized visible segments, and the ground-truth amodal segments masked in a high ratio. Then the transformer is trained to reconstruct the masked tokens of the amodal segments. This mask-and-predict procedure [3] leads to natural sequential decoding in the inference time. Starting with an all-mask token sequence of amodal segments, our transformer gradually completes the amodal segments. Each step increasingly preserves the most confident prediction.

In the second refinement phase, our model learns to inject details to the coarse-prediction and provide a more precise amodal object segmentation. Our convolutional refinement module takes as inputs the coarse-predicted segments and the visual features. Imitating the human activity for visual stimulus, we construct a semantic-inspired attention module as an initial stimulus, and then gradually inject the visual features to the segments through convolution layers.

With this coarse-to-fine architecture design, our C2F-Seg complements the latent space that is easier-to-learn, the transformer that has superiority of long-range dependency, and the convolutional model that can supplement details, and results in a better amodal object segmentation. Our framework is flexible to generalize to video-based amodal object segmentation tasks. Guided by the shape prior and visual features of related frames, our model can generate precise amodal segments, and is even capable of generating amodal segments when the object is totally invisible, as shown in Figure 7.

In order to evaluate the performance of our C2F-Seg. We

conduct experiments both on image and video amodal segmentation benchmarks. For image amodal segmentation, our model reaches 36.5/36.6 on AP, 82.22/80.27 on full mIoU and 53.60/27.71 on occluded mIoU for KINS and COCOA respectively. For video amodal segmentation, our model reaches 91.68/71.30 on full mIoU and 81.21/36.04 on occluded mIoU for FISHBOWL and MOViD-A respectively. C2F-Seg outperforms all the baselines and achieves state-of-the-art performance.

Our contributions can be summarized as:

- We propose a novel coarse-to-fine framework, which consists of a mask-and-predict transformer module for coarse masks and a convolutional refinement module for refined masks. It imitates human activity and progressively generates amodal segmentation, mitigating the effect of detrimental and ill-posed shape priors.
- We build a synthetic dataset MOViD-A for amodal segmentation, which contains 838 videos and 12,299 objects. We hope it will advance research in this field. We release the dataset on our project page.
- Extensive experiments are conducted on two image-based benchmarks, showing the superiority of our methods over other competitors. Moreover, our framework can be easily extended to video-based amodal segmentation, achieving state-of-the-art performance on two benchmarks.

2. Related Works

Amodal Instance Segmentation [39] is a challenging task that involves predicting the shape of occluded objects in addition to the visible parts. To enhance the learning of such connection between the amodal segments and the category label, or prior, previous approaches design specific architectures. MLC [24] learns the visible masks and amodal masks separately via two network branches. AIS-Former [29] enhances the extraction of the long-range dependency via transformers, and utilizes multi-task training to learn a more comprehensive segmentation model. VRSP [33] for the first time explicitly designs the shape prior module to refine the amodal mask. There are also some approaches [39, 9, 38, 32, 36, 28, 14, 34, 27, 18] focus on modeling shape priors with shape statistics, making it challenging to extend their models to open-world applications where object category distributions are long-tail and hard to pre-define. SaVos [35] leverages spatiotemporal consistency and dense object motion to alleviate this problem. However, SaVos requires additional knowledge of optical flow known to cause object deformation in the presence of camera motion. In contrast, our method doesn't need the optical flow anymore. We propose a new framework to learn generic object prior in vector-quantized latent space with transformer to predict the coarse amodal masks of occluded objects. Then we use a CNN-based refine module to polish up the coarse mask in pixel-level to get the fine

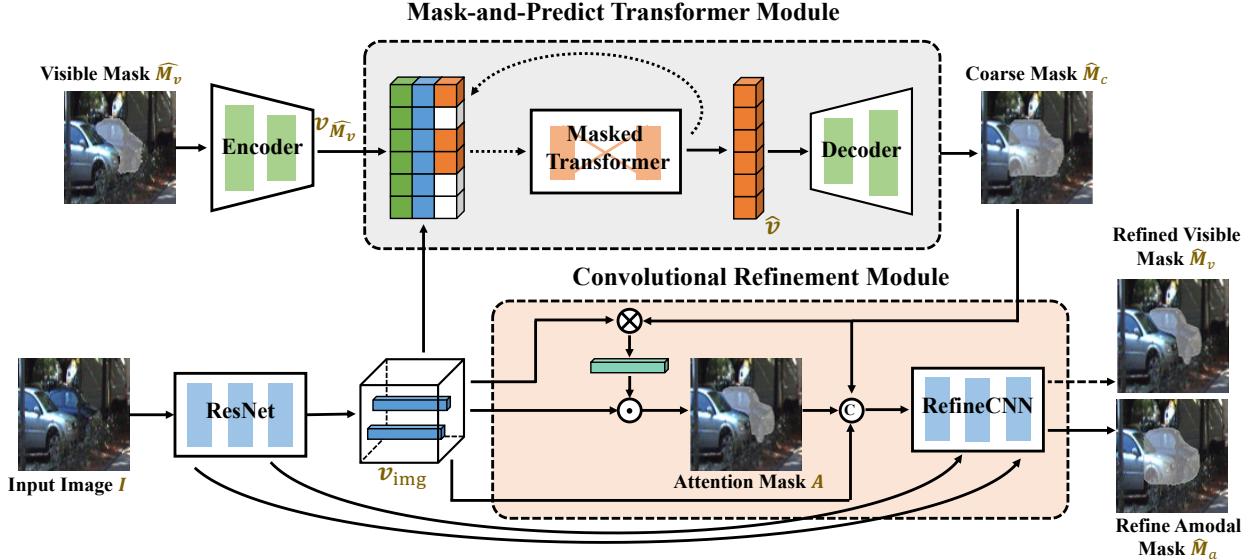


Figure 2. Illustration of our C2F-Seg framework. C2F-Seg first generates a coarse mask from the visible mask and visual features via the mask-and-predict procedure with transformers. Then this coarse amodal mask is refined with a convolutional module guided by human-imitated attention on visual features of the amodal object. The learning of visible mask is used as an auxiliary task in training, while in inference we only provide an estimation of amodal mask.

amodal mask.

Vision Transformer. The self-attention module [31] has enabled impressive performance in various natural language processing and vision tasks through transformer-based methods such as BERT [6] and ViT [7], specifically in vision tasks such as image classification [7], object detection [2], image/video synthesis [3, 11]. Nevertheless, applying transformers to autoregressively generate high-resolution images is computationally expensive and memory-intensive [5, 16]. Thus, new techniques like dVAE [26] and VQ-VAE [22] have been developed to represent images as discrete codes and shorten the sequence. VQ-GAN builds on VQ-VAE [8] by using GANs to improve efficiency, but both methods still use a single codebook to quantize the entire image. Amodal segmentation can also benefit from transformer adaptation. AISformer [29] employs transformer-based mask heads to predict amodal masks, following the approach of DETR [2]. Our framework adopts transformer to utilize the mask-and-predict formulation of amodal object segmentation, inspired by MaskGIT and MaskViT [3, 11]. Then, we refine the prediction using a CNN-based module for precise segmentation.

3. Coarse-to-Fine Segmentation

3.1. Problem Setup

Amodal object segmentation aims to segment not only the visible parts of an object but also its occluded parts.

Formally, amodal object segmentation takes as inputs an input image I , a bounding box of the Region-of-Interest (ROI). The amodal object is only partially visible in the image, decomposed into visible parts and occluded parts. The visible parts can be segmented via standard segmentation algorithms, but the invisible occluded parts needs to be estimated rather than segmented. Following [35], we denote the visible segment as M_v and the full/amodal segment as M_a such that M_a consisted of both the visible segment and the invisible segment. Thus our target is to estimate M_v and M_a simultaneously from the ROI of I .

We utilize current segmentation algorithms to provide an estimation of visible segment \hat{M}_v . Then based on I and \hat{M}_v , we construct our C2F-Seg framework by two stages. In the first stage, we estimate a coarse-grained segment M_c based on the vector-quantized latent space by transformer. Then we adopt a convolutional module to refine the estimation and provide a precise fine-grained prediction \hat{M}_a as the final estimation of M_a . In the following, we introduce each component of our C2F-Seg framework in details.

3.2. Vector-Quantized Latent Space

Our latent space is inspired by the well-known VQ-GAN [8]. Specifically, we adopt an encode-decode architecture with encoder E and decoder D with convolutional layers. For input mask $M \in \mathbb{R}^{H \times W}$, the encoder projects it to the continuous latent code $\hat{z} = E(M)$ from a learned, discrete codebook $\mathcal{Z} = \{z_k\}_{k=1}^K \subset \mathbb{R}^{n_z}$, where n_z is the dimension of codes. Then the closest codebook entry of

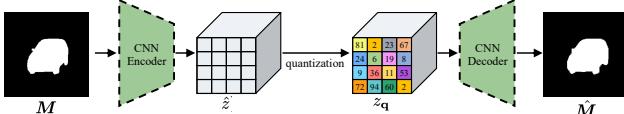


Figure 3. The architecture of Vector-Quantization model. The trained latent representation of masks are used in our transformer.

each spatial code z_{ij} is utilized to get a discrete representation from the codebook vocabulary

$$z_q = \mathbf{q}(\hat{z}) := \left(\arg \min_{z_k \in \mathcal{Z}} \| \hat{z}_{ij} - z_k \| \right) \in \mathbb{R}^{h \times w \times n_z}, \quad (1)$$

where z_k means the closest codebook entry of each spatial code z_{ij} . With this discrete representation, the decoder D reconstructs the input mask M as

$$\hat{M} := D(z_q) = D(\mathbf{q}(E(M))). \quad (2)$$

With properly trained encoder and decoder, we can get a latent representation $s \in \{0, \dots, |\mathcal{Z}| - 1\}^{h \times w}$ in terms of the codebook-indices for each mask, which consists the latent space on which our first learning stage performs. We initialize an embedding for the indices s as the input to the transformer model:

$$\mathbf{v}_M := \text{Embed}(s). \quad (3)$$

While it is common to utilize a VQ-GAN to encode the input image in the corresponding latent space, our preliminary experiments reveal a decrease in performance when employing this method. This could be due to the fact that current approaches utilizing the VQ-GAN for image encoding and latent space learning are typically reliant on a vast training dataset comprising millions, if not billions, of data points. Unfortunately, for amodal object segmentation tasks, we have access to only a limited training set, which may not be sufficient to train a potent embedding. As a result, in practical scenarios, we resort to utilizing a pretrained ResNet to extract and flatten the visual features of the input image as the transformer model input:

$$\mathbf{v}_{\text{img}} := \text{Flatten}(\text{ResNet}(\mathbf{I})). \quad (4)$$

By adopting this approach, we can alleviate the learning complexity and enhance the segmentation ability. Since the embedding of masks is initialized randomly, we choose to set the embedding dimension to match the size of the visual features for improved alignment.

3.3. Mask-and-Predict Transformer

Having established the latent space's architecture, as described in the preceding subsection, we now possess an image representation denoted as \mathbf{v}_{img} and a visible segment

representation referred to as $\mathbf{v}_{\hat{M}_v}$. Our aim is to predict the amodal object segmentation, denoted as \mathbf{v}_{M_a} .

To achieve this, we introduce a [MASK] token apart from the learned mask codebook. Then we initialize $\hat{\mathbf{v}}$ as all [MASK] tokens with the same dimension of $\mathbf{v}_{\hat{M}_v}$. Then we concatenate \mathbf{v}_{img} , $\mathbf{v}_{\hat{M}_v}$, and $\hat{\mathbf{v}}$ as the input of the transformer model.

The training objective of the transformer model is to minimize the negative log-likelihood as

$$\mathcal{L} := -\mathbb{E} \left[\sum_i \log p(\mathbf{v}_{M_a,i} | \hat{\mathbf{v}}, \mathbf{v}_{\text{img}}, \mathbf{v}_{\hat{M}_v}) \right]. \quad (5)$$

Nevertheless, learning to make one-step prediction are known to be challenging. Therefore, we draw inspiration from the general concept behind mask-and-predict approaches such as BERT [15], MaskGIT [3] and MaskVIT [11]. By doing so, we can simplify the objective by masking specific codes within the amodal segment representation, and then predicting the masked portions. Denote the masking operator as \mathcal{M} , our training objective now becomes

$$\mathcal{L} := -\mathbb{E} \left[\sum_i \log p(\mathbf{v}_{M_a,i} | \mathcal{M}(\mathbf{v}_{M_a}), \mathbf{v}_{\text{img}}, \mathbf{v}_{\hat{M}_v}) \right]. \quad (6)$$

Developing an appropriate masking policy is critical to the overall approach's success. If we only mask a negligible fraction of the amodal segments, the task becomes simplistic and fails to generalize to testing stages. Conversely, if we mask a substantial portion, it may prove too challenging to learn. To address this, we uniformly select the masking ratio from 50% to 100% in practical scenarios. This approach enables us to strike a balance between learning difficulty and training-testing consistency.

During inference, we take iterative inference method to complete the amodal masks in K steps. At each step, our model predicts all tokens simultaneously but only keeps the most confident ones. The remaining tokens are masked out and re-predicted in the next iteration. The mask ratio is made decreasing until all tokens are generated within T iterations.

After estimating $\hat{\mathbf{v}}$, we use the decoder D to reconstruct the coarse-predicted amodal mask

$$\hat{M}_c = D(\hat{\mathbf{v}}). \quad (7)$$

3.4. Convolutional Refinement

Although we train the VQ-GAN model to reconstruct the mask as precisely as possible, it inevitably losses some details of the mask and thus is only a coarse estimation. To recover these details, we adopt a convolutional refinement module.

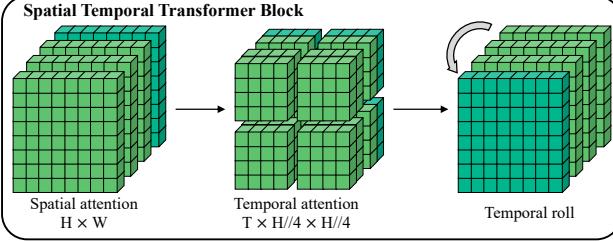


Figure 4. Architecture of Spatial Temporal Transformer Block(STTB). For video tasks, we roll the features in temporal dimension after each transformer block and recover the normal order at the end of our model.

Our convolutional refinement module takes as inputs the image features v_{img} and the estimated coarse amodal mask \hat{M}_c . We first downsample the coarse amodal mask to match the dimension of the ResNet features, yielding downsampled \hat{M}_{cd} . Note that in our mask-and-predict transformer, we adopt a vector-quantization module to align mask and visual features. However, this alignment requires extra training. Thus in our convolutional refinement module, we direct downsample the mask to avoid additional training for efficiency. As the mask can be regarded as a hard attention map, we directly encourage an attention on the amodal object via

$$\mathbf{A} := \text{softmax}\left(\frac{\hat{M}_{cd} v_{\text{img}}^\top}{\sqrt{d}}\right) \odot v_{\text{img}}, \quad (8)$$

where \odot is the element-wise multiplication.

Then the convolutional refinement module learns to predict the visible segment and amodal segment simultaneously

$$\hat{M}_a, \hat{M}_v = \text{Conv}(v_{\text{img}}, \mathbf{A}, \hat{M}_{cd}). \quad (9)$$

The convolutional refinement is trained to minimize binary cross-entropy loss for visible mask and amodal mask simultaneously.

$$\mathcal{L}_r := \text{BCE}(\hat{M}_a, M_a) + \text{BCE}(\hat{M}_v, M_v). \quad (10)$$

3.5. Extension to Video Amodal Segmentation

Our framework can generalize to video amodal object segmentation easily. Specifically, our model leverages temporal-spatial attention [11] to capture the temporal features throughout the entire video and model the amodal masks. Each transformer block comprises a spatial layer and a temporal layer, as shown in Figure 4. The spatial layer functions similar to classical self-attention layers, while the temporal layer splits the codes in the spatial dimension and stacks them into a (T, h, w) size in the temporal dimension. It then performs self-attention on these codes to capture the temporal relationships in the video.

Compared with object in a single image, object in video suffers from occlusion of different parts in different frames,



Figure 5. Overview of all the datasets used in our paper. The image at the top left panel is selected from COCOA. The top right panel shows an example of Fishbowl. Below are two images from MOViD-A. The bottom panel is an image from KINS.

and the object itself may also undergo rotation and deformation relative to the camera. Therefore, it is essential to enhance the spatial-temporal modeling ability of our model to accurately learn the complete shape of the target objects. To fully extract spatial-temporal information, our model rolls the features in the temporal dimension by $T/2$ frames after each transformer block. This operation significantly improves the performance, as discussed in the supplementary.

4. Experiments

Datasets. To evaluate the efficacy of our proposed model, we conduct comprehensive experiments on both image and video amodal segmentation benchmarks. **1) KINS** [24] is a large-scale amodal instance dataset, which is built upon KITTI [10]. It contains 7 categories that are common on the road, including car, truck, pedestrian, etc. There are 14,991 manually annotated images in total, 7,474 of which are used for training and the remaining for testing. **2) COCOA** [39] is derived from COCO dataset [17]. It consists of 2,476 images in the training set and 1,223 images in the testing set. There are 80 objects in this dataset. **3) FISHBOWL** [28] is a video benchmark, recorded from a publicly available WebGL demo of an aquarium[1]. Following [35], we select 10,000 videos for training and 1,000 for testing, each with 128 frames. **4) MOViD-A** is a video-based synthesized dataset. We create it from MOVi dataset ¹ for amodal segmentation. The virtual camera is set to go around the

¹<https://github.com/google-research/kubric/tree/main/challenges/movi>

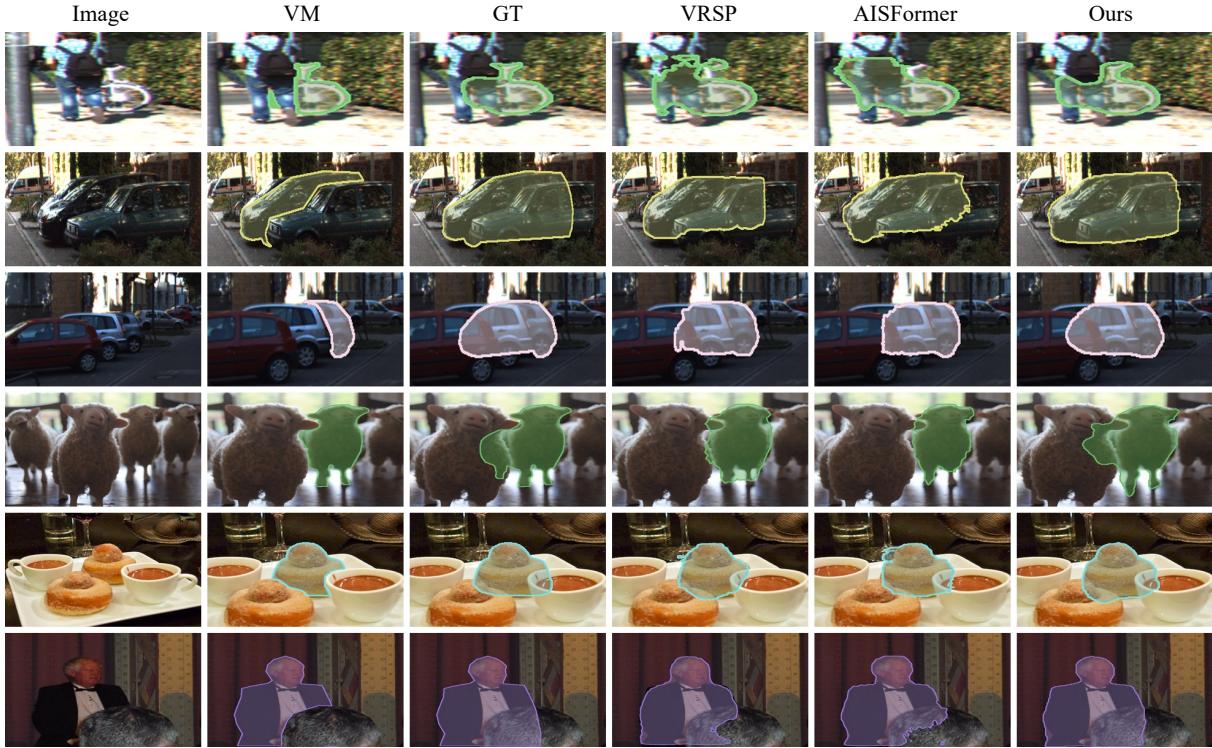


Figure 6. The qualitative results estimated by VRSP, AISFormer, and our method. VM and GT indicate ground-truth visible mask and amodal mask, respectively.

METHODS	KINS						COCOA					
	AP	AP ₅₀	AP ₇₅	AR	mIoU _{full}	mIoU _{occ}	AP	AP ₅₀	AP ₇₅	AR	mIoU _{full}	mIoU _{occ}
PCNet [36]	29.1	51.8	29.6	18.3	78.02	38.14	-	-	-	-	76.91	20.34
Mask R-CNN [12]	30.0	54.5	30.1	19.4	-	-	28.0	53.7	25.4	29.8	-	-
ORCNN [9]	30.6	54.2	31.3	19.7	-	-	28.0	53.7	25.4	29.8	-	-
VRSP [33]	32.1	55.4	33.3	20.9	80.70	47.33	35.4	56.0	38.7	37.1	78.98	22.92
AISformer [29] [†]	33.8	57.8	35.3	21.1	81.53	48.54	29.0	45.7	31.8	31.1	72.69	13.75
C2F-Seg (<i>ours</i>)	36.5	58.2	37.0	22.1	82.22	53.60	36.6	57.0	38.5	38.5	80.28	27.71

Table 1. Performance comparison on the KINS and COCOA. We fine-tune the AISformer (marked by [†]) on COCOA from the official model trained on KINS. Other results are reported in AISformer.

scene, capturing about 24 consecutive frames. We randomly place 10 ~ 20 static objects that heavily occlude each other in the scene. Finally, we collect 630 and 208 videos for training and testing. Examples are shown in Figure 5.

Metrics. For evaluation, we adopt standard metrics as in most amodal segmentation literature [13, 33, 29], namely mean average precision (AP) and mean average recall (AR). Furthermore, We use mean-IoU [24, 35] to measure the quality of predicted amodal masks. It is calculated against the ground-truth amodal mask (mIoU_{full}) or the occluded region (mIoU_{occ}). Occluded mIoU measures the complete quality of the occluded part of target objects directly. It is worth noting that occluded mIoU is a crucial indicator for amodal segmentation. Following [35], we specially only compute mIoU for objects on FISHBOWL with the occlu-

sion rate from 10 to 70%, and all detected objects on other datasets are involved for evaluation.

Implementation Details. Our framework is implemented on PyTorch platform. Considering that competitors for image-based amodal segmentation all include a detection branch, we use pre-detected visible bounding boxes and masks by AISFormer [29], for fair comparison. Particularly, since [29] does not provide weights on COCOA, we turn back to use the model trained by VRSP [33]. For video benchmarks, all baselines and our model take ground-truth visible masks as inputs. We use bounding boxes of visible regions, which enlarge 2 times, to crop images and masks as inputs. The inputs are all resized to 256 × 256. For data augmentation, morphology dilation, erosion and Gaussian blur are applied to mask inputs. AdamW optimizer [19]

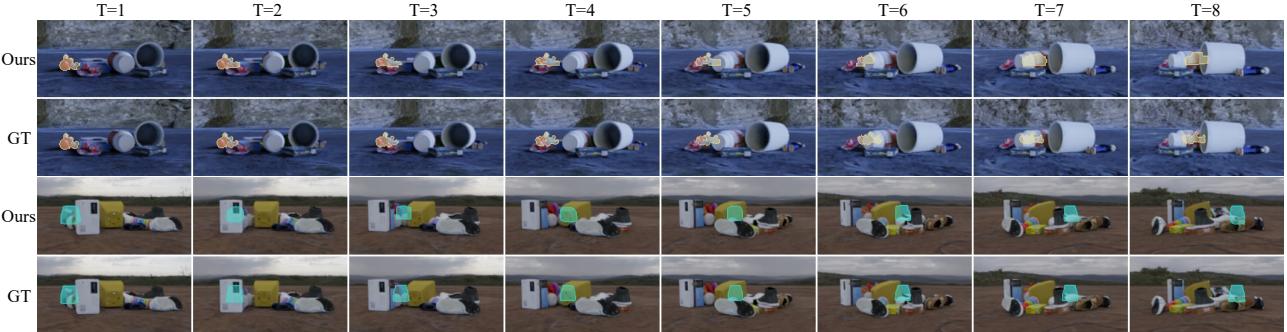


Figure 7. The visualization results of C2F-Seg on video dataset. The orange toy and the cyan box are invisible in a few frames, our model still produces approximate complete amodal masks. Best viewed in color and zoom in.

with a learning rate of 3e-4 is adopted for all experiments. We train the model using a batch size of 16 for all datasets except MOViD-A, which has a batch size of 24. The total number of iterations set for KINS, COCOA, FISHBOWL and MOViD-A datasets is 45K, 10K, 75K and 75K, respectively. For architecture design, we set the number of transformer layers is 12, and the feature dimension is 768 for image dataset. When training for FISHBOWL and MOViD-A, we adjust the number of transformer layers to 8, for we adjust the transformer block into spatial temporal one. For the codebook size of vector-quantized latent space, we set 256 for all the datasets. The vector-quantization model is trained for each dataset, respectively. During inference, we consistently set the iterative step K to 3 to demonstrate the generalizability of our method.

4.1. Results of Image Amodal Segmentation

We first compare our C2F-Seg with several image-based competitors on KINS and COCOA dataset. As shown in Table 1, we report results of both AP and mIoU. From the table, we can observe that (1) our model achieves state-of-the-art performance on both datasets across most of metrics. For KINS, we outperform the second-best method by a margin of at least 5 points on $mIoU_{occ}$. Despite COCOA being a more challenging dataset than KINS due to its diverse object categories and intricate shapes, our method still yields better results compared to other approaches. It clearly suggests the superiority of our proposed method. (2) Compared with AISFormer [29] that also utilizes transformers for amodal segmentation, we beat it by 2.7% and 1.0% on AP and AR metrics. Moreover, VRSP [33] utilizes learned shape prior to refine the predict amodal mask. Differently, our C2F-Seg leverages shape prior to obtain coarse amodal region and further complete it with visual features. As expected, we achieve 1.4% and 4.79% higher results on AR and $mIoU_{occ}$, which significantly shows the advantage of our design.

Qualitative results estimated by VRSP, AISFormer, and our method are further illustrated in Figure 6. As observed,

METHODS	FISHBOWL		MOViD-A	
	$mIoU_{full}$	$mIoU_{occ}$	$mIoU_{full}$	$mIoU_{occ}$
<i>visible masks</i>	68.53	-	56.92	-
Convex	77.61	46.38	60.18	16.48
PCNET [36]	87.04	65.02	64.35	27.31
SaVos [35]	88.63	71.55	60.61	22.64
AISformer [29]	-	-	67.72	33.65
C2F-Seg (<i>ours</i>)	91.68	81.21	71.67	36.13

Table 2. Quantitative results on FISHBOWL and MOViD-A. We report and compare the Mean-IoU metrics for FISHBOWL and MOViD-A of C2F-Seg with baselines.

our method can segment more occluded regions with accurate shapes, owing to the help of excellent shape prior and precise refine module. For visualizations from the 2nd to 4th row, the predictions of our method are not misled by occlusions which have the same category as target objects, especially when the occlusion rate is very large or relatively small. For objects that have intricate and delicate contours, such as the bicycle in 1st row, both VRSP and AISFormer fail to precisely segment the occluded area of rear-wheel and the visible region of saddle. By comparison, our method has shown to be successful in tackling this challenging case, and delivers good performance.

4.2. Results of Video Amodal Segmentation

We further investigate the efficacy of our model for video amodal segmentation task. Table 2 shows the mean-IoU metrics of C2F-Seg and baselines on FISHBOWL and MOViD-A datasets. Importantly, our method outperforms all the baselines, getting 81.21 and 36.04 results on occluded mIoU. Particularly, we achieve 4.64/16.19 and 3.05/9.66 higher performance than PCNET [36] and SaVos [35] on FISHBOWL, respectively. On MOViD-A, we also achieve 10.69/13.4, 3.58/2.39 higher performance than SaVos and AISformer [29], respectively. It is worth noting that MOViD-A presents much more challenges, including multiple objects, lens distortion, and change of view point. Nevertheless, our model remains effective in han-

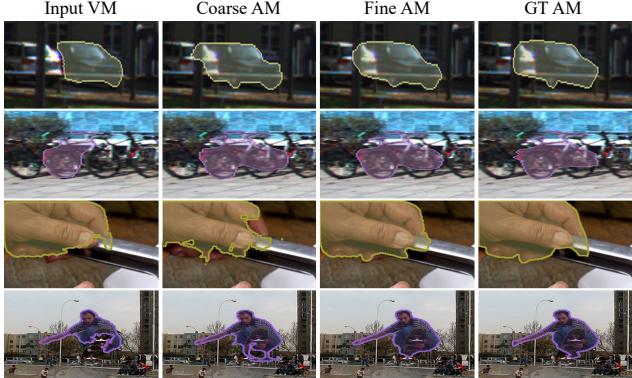


Figure 8. The coarse-to-fine progress in C2F-Seg. The estimated amodal mask is iteratively refined from the visible mask. VM indicates visible mask while AM indicates amodal mask.

METHODS	KINS		COCOA	
	mIoU _{full}	mIoU _{occ}	mIoU _{full}	mIoU _{occ}
w/o refine	81.81	52.57	79.52	24.25
single branch	81.95	52.78	80.03	26.12
full model	82.22	53.60	80.28	27.71

Table 3. **Ablation results for the refine module.** We report mean-IoU metrics on KINS and COCOA to evaluate our refine module.

dling these challenging scenarios. In addition, we provide the lower bound by directly evaluating with input visible masks. All of these observations strongly demonstrate the effectiveness of our proposed model, highlighting its generalization ability to video amodal segmentation task. Figure 7 shows the qualitative results of our model for MOViD-A. For the extreme case that the target is fully occluded by other objects, our model is capable of producing complete amodal masks that closely resemble the ground truth in terms of both position and shape. Thanks to our proposed shape prior generator and coarse-to-fine module.

4.3. Ablation Study

We further conduct ablation studies to evaluate the effectiveness of our model on both image and video datasets.

Effect of Convolutional Refinement. To investigate the effectiveness of our proposed refine module, we first verified the validity of the two-branches architecture which both predicts the visible masks and amodal masks. We train C2F-Seg with the refine module only predicting the amodal masks. Further more, we conduct another experiment without the refine module. These two experiments keep the same setting as claimed before. The results are shown in Table 3. We also visualize the process that C2F-Seg predicts the precise amodal mask based on the input visible mask. In Figure 8, the Coarse FMs reflect the shape prior our model has learned and the refine module improves them in detail by adding and removing some redundant regions. It indicates that both predict visible masks and amodal masks will

GT VM	KINS		COCOA	
	mIoU _{full}	mIoU _{occ}	mIoU _{full}	mIoU _{occ}
✗	82.22	53.60	80.28	27.71
✓	87.89	57.60	87.13	36.55

Table 4. The Upper bound mean-IoU metrics of C2F-Seg on KINS and COCOA. It indicates that high quality visible masks help our model reach better performance.

help the model to distinguish and figure out the difference between the two masks. We can draw the conclusion that the two-branches refine module helps our model to predict amodal masks better.

Effect of Parameter K . It’s noteworthy that K significantly influences MaskGIT [3]. To investigate the effect of K on our model, we assessed various K values on the COCOA and MOViD-A datasets. The results can be found in the supplementary (Table B.1). Interestingly, the table indicates that changes in the K value have only a marginal effect on our model. A potential explanation might be the difference in tasks: in our context, the module involving K is designed to generate coarse masks, whereas MaskGIT focuses on RGB image production.

Upper bound of C2F-Seg Since our model is driven by visible mask, we try to feed it with GT visible masks to explore the upper bound metric. We keep the same setting mentioned for Image Amodal Segmentation to evaluate the best performance of C2F-Seg. We only change the predicted visible masks to GT visible masks for KINS and COCOA. Table 4 shows the mIoU metrics for the two datasets. Our model achieves 5.67/4.0 and 6.86/9.44 on KINS and COCOA respectively. It shows that our model will reach much better results with high quality visible masks.

5. Conclusion

In this work, we introduce a novel framework, C2F-Seg, which harnesses transformers to learn shape priors in the latent space, enabling the generation of a coarse mask. Subsequently, we deploy a dual-branch refinement module to produce an attention mask. This mask is then combined with the coarse mask and features from ResNet-50 to predict both visible and amodal masks. For video datasets, we adapt our transformer block into a spatial-temporal version to effectively capture spatio-temporal features, leading to superior amodal mask predictions. Our model gets the precise amodal masks step by step and achieves new State-of-the-art performance both on image and video amodal segmentation.

Acknowledgements. This work is supported by China Postdoctoral Science Foundation (2022M710746). Yanwei Fu is with the School of Data Science, Shanghai Key Lab of Intelligent Information Processing, Fudan University, and Fudan ISTBI-ZJNU Algorithm Centre for Brain-inspired Intelligence, Zhejiang Normal University, Jinhua, China.

References

- [1] Webgl demo of an aquarium, 2022.
- [2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 213–229. Springer, 2020.
- [3] Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11315–11325, 2022.
- [4] Chilam Cheang, Haitao Lin, Yanwei Fu, and Xiangyang Xue. Learning 6-dof object poses to grasp category-level objects by language instructions. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 8476–8482. IEEE, 2022.
- [5] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pre-training from pixels. In *International Conference on Machine Learning*, pages 1691–1703. PMLR, 2020.
- [6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- [7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
- [8] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021.
- [9] Patrick Follmann, Rebecca König, Philipp Härtinger, Michael Klostermann, and Tobias Böttger. Learning to see the invisible: End-to-end trainable amodal instance segmentation. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1328–1336. IEEE, 2019.
- [10] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE, 2012.
- [11] Agrim Gupta, Stephen Tian, Yunzhi Zhang, Jiajun Wu, Roberto Martín-Martín, and Li Fei-Fei. Maskvit: Masked visual pre-training for video prediction. *arXiv preprint arXiv:2206.11894*, 2022.
- [12] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.
- [13] Lei Ke, Yu-Wing Tai, and Chi-Keung Tang. Deep occlusion-aware instance segmentation with overlapping bilayers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4019–4028, 2021.
- [14] Lei Ke, Yu-Wing Tai, and Chi-Keung Tang. Occlusion-aware video object inpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14468–14478, 2021.
- [15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186, 2019.
- [16] Manoj Kumar, Dirk Weissenborn, and Nal Kalchbrenner. Colorization transformer. *arXiv preprint arXiv:2102.04432*, 2021.
- [17] 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 *European conference on computer vision*, pages 740–755. Springer, 2014.
- [18] Huan Ling, David Acuna, Karsten Kreis, Seung Wook Kim, and Sanja Fidler. Variational amodal object completion. *Advances in Neural Information Processing Systems*, 33:16246–16257, 2020.
- [19] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.
- [20] Florian Mathis, John H Williamson, Kami Vaniea, and Mohamed Khamis. Fast and secure authentication in virtual reality using coordinated 3d manipulation and pointing. *ACM Transactions on Computer-Human Interaction (ToCHI)*, 28(1):1–44, 2021.
- [21] Rohit Mohan and Abhinav Valada. Amodal panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21023–21032, 2022.
- [22] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. *arXiv preprint arXiv:1711.00937*, 2017.
- [23] Youngmin Park, Vincent Lepetit, and Woontack Woo. Multiple 3d object tracking for augmented reality. In *2008 7th IEEE/ACM International Symposium on Mixed and Augmented Reality*, pages 117–120. IEEE, 2008.
- [24] Lu Qi, Li Jiang, Shu Liu, Xiaoyong Shen, and Jiaya Jia. Amodal instance segmentation with kins dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3014–3023, 2019.
- [25] Xuelin Qian, Li Wang, Yi Zhu, Li Zhang, Yanwei Fu, and Xiangyang Xue. Impdet: Exploring implicit fields for 3d object detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4260–4270, 2023.
- [26] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. *arXiv preprint arXiv:2102.12092*, 2021.
- [27] Yihong Sun and Adam Kortylewski. Amodal segmentation through out-of-task and out-of-distribution generalization with a bayesian model. cvpr 2022. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2022.
- [28] Matthias Tangemann, Steffen Schneider, Julius von Kügelgen, Francesco Locatello, Peter Gehler, Thomas Brox, Matthias Kümmeler, Matthias Bethge, and Bernhard Schölkopf. Unsupervised object learning via common fate. *arXiv preprint arXiv:2110.06562*, 2021.

- [29] Minh Tran, Khoa Vo, Kashu Yamazaki, Arthur Fernandes, Michael Kidd, and Ngan Le. Aisformer: Amodal instance segmentation with transformer. *arXiv preprint arXiv:2210.06323*, 2022.
- [30] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.
- [31] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- [32] Yuting Xiao, Yanyu Xu, Ziming Zhong, Weixin Luo, Jiawei Li, and Shenghua Gao. Amodal segmentation based on visible region segmentation and shape prior. *arXiv preprint arXiv:2012.05598*, 2020.
- [33] Yuting Xiao, Yanyu Xu, Ziming Zhong, Weixin Luo, Jiawei Li, and Shenghua Gao. Amodal segmentation based on visible region segmentation and shape prior. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2995–3003, 2021.
- [34] Jianwei Yang, Zhile Ren, Mingze Xu, Xinlei Chen, David J Crandall, Devi Parikh, and Dhruv Batra. Embodied amodal recognition: Learning to move to perceive objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2040–2050, 2019.
- [35] Jian Yao, Yuxin Hong, Chiyu Wang, Tianjun Xiao, Tong He, Francesco Locatello, David Wipf, Yanwei Fu, and Zheng Zhang. Self-supervised amodal video object segmentation. *arXiv preprint arXiv:2210.12733*, 2022.
- [36] Xiaohang Zhan, Xingang Pan, Bo Dai, Ziwei Liu, Dahua Lin, and Chen Change Loy. Self-supervised scene de-occlusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3784–3792, 2020.
- [37] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.
- [38] Ziheng Zhang, Anpei Chen, Ling Xie, Jingyi Yu, and Shenghua Gao. Learning semantics-aware distance map with semantics layering network for amodal instance segmentation. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 2124–2132, 2019.
- [39] Yan Zhu, Yuandong Tian, Dimitris Metaxas, and Piotr Dollár. Semantic amodal segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1464–1472, 2017.

A. Preliminary Knowledge

A.1. Detail of Vector-Quantization Module

This module draws inspiration from the well-known VQ-GAN [8]. Our aim is to reduce the learning complexity and expedite the inference process during the coarse segmentation phase. Therefore, we execute the segmentation within a low-dimensional vector-quantized latent space. Beyond what is mentioned in the main paper, the training objective is to identify the optimal compression model $\mathcal{Q}^* = \{E^*, G^*, \mathcal{Z}^*\}$, which can be expressed as:

$$\begin{aligned}\mathcal{Q}^* = \arg \min_{E, G, \mathcal{Z}} \max_D \mathbb{E}_{x \sim p(x)} [\mathcal{L}_{\text{VQ}}(E, G, \mathcal{Z}) \\ + \lambda \mathcal{L}_{\text{GAN}}(\{E, G, \mathcal{Z}\}, D)],\end{aligned}$$

where

$$\begin{aligned}\mathcal{L}_{\text{VQ}}(E, G, \mathcal{Z}) = \mathcal{L}_{\text{rec}} + \| \text{sg}[E(x)] - z_q \|_2^2 \\ + \beta \| \text{sg}[z_q] - E(x) \|_2^2\end{aligned}$$

and

$$\mathcal{L}_{\text{GAN}}(\{E, G, \mathcal{Z}\}, D) = [\log D(x) + \log(1 - D(\hat{x}))]$$

The adaptive weight λ is computed as:

$$\lambda = \frac{\nabla_{G_L} [\mathcal{L}_{\text{rec}}]}{\nabla_{G_L} [\mathcal{L}_{\text{GAN}}] + \delta}$$

In this context, \mathcal{L}_{rec} represents the perceptual reconstruction loss [37]. The symbol $\text{sg}[\cdot]$ indicates the stop-gradient operation, while $\| \text{sg}[z_q] - E(x) \|_2^2$ is referred to as the commitment loss and has a weighting factor of β [30]. The notation $\nabla_{G_L} [\cdot]$ signifies the gradient of its input with respect to the last layer L of the decoder. For numerical stability, we employ $\delta = 10^{-6}$.

In our experiments, we fixed the codebook size $|\mathcal{Z}|$ at 256 across all datasets. We also omitted the attention layer from the original model. The entire iteration process for the four datasets is configured at 100k.

A.2. Detail of Iterative Inference

Inspired by MaskGIT [3], the mask-and-predict procedure facilitates natural sequential decoding during inference. Beginning with a token sequence that masks all amodal segments, our transformer incrementally completes the amodal segments, preserving the most confident prediction with each step. In detail, to produce a coarse mask at inference time, we commence with a blank canvas where all tokens are masked, denoted as $Y_M^{(0)}$ (where Y_M represents the result after applying mask M to Y). For iteration t , our transformer operates as:

1. Parallel Prediction: Starting with the current set of masked tokens, $Y_M^{(t)}$, the transformer predicts the likelihoods for all masked positions at once, producing a probability matrix $p^{(t)} \in \mathbb{R}^{N \times K}$.

2. Token Sampling with Confidence Scoring: At every masked location, a token is sampled based on its associated probabilities. This token's prediction score is taken as a confidence measure, showing the model's trust in its prediction. Positions that are already unmasked are automatically given full confidence, scored at 1.0.

3. Dynamic Masking: The number of tokens that should remain masked in the next iteration is computed using the mask scheduling function γ . This accounts for the input length N and the progression of iterations t relative to the total T .

4. Update Masking Strategy: Tokens in $Y_M^{(t)}$ are then updated for the next iteration. Only tokens with lower confidence scores are re-masked, as determined by a threshold value derived from the sorted confidence scores. This ensures that the transformer focuses on refining less confident tokens in the subsequent iteration.

The Iterative Inference assembles a coarse amodal mask in K steps. During each iteration, the transformer anticipates all tokens concurrently, yet retains only the most confident selections. Subsequent tokens are masked again and re-predicted in the following iteration. The mask ratio diminishes until all tokens are formulated within K iterations.

B. Further Ablation Studies

B.1. Table of Ablation Study for K

We have carried out an ablation study to investigate the impact of K on our model. The performance of our model, across different values of K , on the COCOA and MOViD-A datasets, is detailed in Table B.1.

K	COCOA		MOViD-A	
	mIoU _{full}	mIoU _{occ}	mIoU _{full}	mIoU _{occ}
1	80.16	27.70	71.91	36.57
2	80.27	27.68	71.67	36.30
3	80.28	27.71	71.67	36.13
5	80.28	27.60	71.58	35.88
8	80.31	27.57	71.46	35.53
10	80.24	27.28	71.42	35.60
12	80.27	27.44	71.41	35.44

Table B.1. Ablation results for K on COCOA and MOViD-A.

In order to further evaluate the effectiveness of our model both on image and video datasets, we conduct the following two experiments.

B.2. Effect of Time Rolling in Transformer

We also investigate the effectiveness of Spatial Temporal(ST) module used in our video version of C2F-Seg. The ST module is proposed in [11] and we modify the module with an extra roll mechanism which will help C2F-Seg to model the whole video, and make full use of transformer to

METHODS	Fishbowl		MOViD-A	
	mIoU _{full}	mIoU _{occ}	mIoU _{full}	mIoU _{occ}
w/o ST module	89.64	78.93	67.19	26.48
w/o roll	90.91	80.01	69.92	32.35
full model	91.68	81.21	71.67	36.13

Table B.2. **Ablation results for our STTB module for Video task.** We report the mean-IoU metric for Fishbowl and MOViD-A to evaluate our design for spatio-temporal feature.

METHODS	KINS		COCOA	
	mIoU _{full}	mIoU _{occ}	mIoU _{full}	mIoU _{occ}
w/o attn	82.07	52.98	80.15	26.85
w. attn	82.22	53.60	80.28	27.71

Table B.3. **Ablation results for the attention mechanism.** Mean-IoU metrics on KINS and COCOA to evaluate this mechanism.

extract spatiotemporal information features over long distances. In this part, we evaluate the effect of each module. We train our model with full ST module, without ST module, and without roll mechanism respectively on the two video datasets. The results are shown in Table B.2. Results indicate the effectiveness of the ST module as well as our introduced roll mechanism.

B.3. The Effect of Attention Mechanism in Refinement Module

To investigate the effectiveness of the attention calculated in our proposed refine module, we train C2F-Seg with and without calculating attention separately on KINS and COCOA. Table B.3 shows the mIoU metrics for the two datasets. The results indicate our attention mechanism improves the quality of amodal masks.

C. Supports for the claim of shape prior

Our claim of shape prior is based on a common phenomenon, which is supported by Figure C.1 showcasing six randomly selected cases. In the figure, the arrangement from top to bottom includes the images, the visible masks, and the amodal masks. Specifically, (a) is from KINS, (b) is from COCOA, and the remaining cases are from MOViD-A. We can observe that:

(1) The visible masks of these cases exhibit significant differences compared to their corresponding amodal masks due to occlusion caused by different poses.

(2) Besides, viewpoint variations may lead to differences in the shape prior. This is exemplified by cases (b)-(d), where the shape prior differs from the original in regular view.

D. More Qualitative Results

In order to more intuitively illustrate the strengths of our algorithm, and to compare it with the baselines, we select

KINS and MOViD-A to show more qualitative results to demonstrate the effectiveness of our method.

D.1. Visualization on KINS Dataset

To show the performance of our method on real scenarios, we show more results from KINS in Figure D.1. In these images, for fair comparison, we select the intersection of the amodal masks predicted by VRSP [33] and AIS-Former [29]. Our algorithm completes the occluded cars better than all the baselines on KINS, which will help to improve the safety of autonomous driving significantly if applied to real scenarios.

D.2. Visualization on MOViD-A Dataset

We show the qualitative results estimated by the best baseline video and image-based amodal method on MOViD-A respectively in Figure D.2. Our method predicts the invisible masks excellently by extracting valid spatio-temporal features and outperforms all the baselines.

E. Limitations and Future Works

We propose a coarse-to-fine framework that leverages shape prior for amodal segmentation. Despite it has achieved significant advantages in both image and video-based benchmarks, our proposed C2F-Seg still faces several limitations. One is the additional input of the pre-detected visible mask. It is essential but not efficient, since we need to specify the target when multiple objects occur in the same scene. In future work, we will either replace it with a single point or incorporate our framework with an end-to-end detection branch, to effectively decrease the input requirement. Another limitation may lie in objects which are heavily or fully occluded. Though our introduced Spatial Temporal Transformer Block successfully mitigates this problem by aggregating multi-frame shape priors, amodal masks of some frames are not precise due to the ill-posed problem. We will explicitly design modules to utilize spatio-temporal prior and constraint the consistence of masks between adjacent frames.

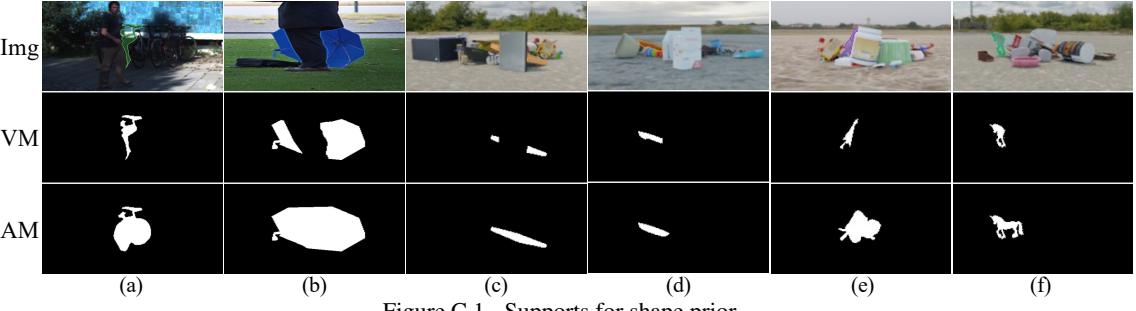


Figure C.1. Supports for shape prior.

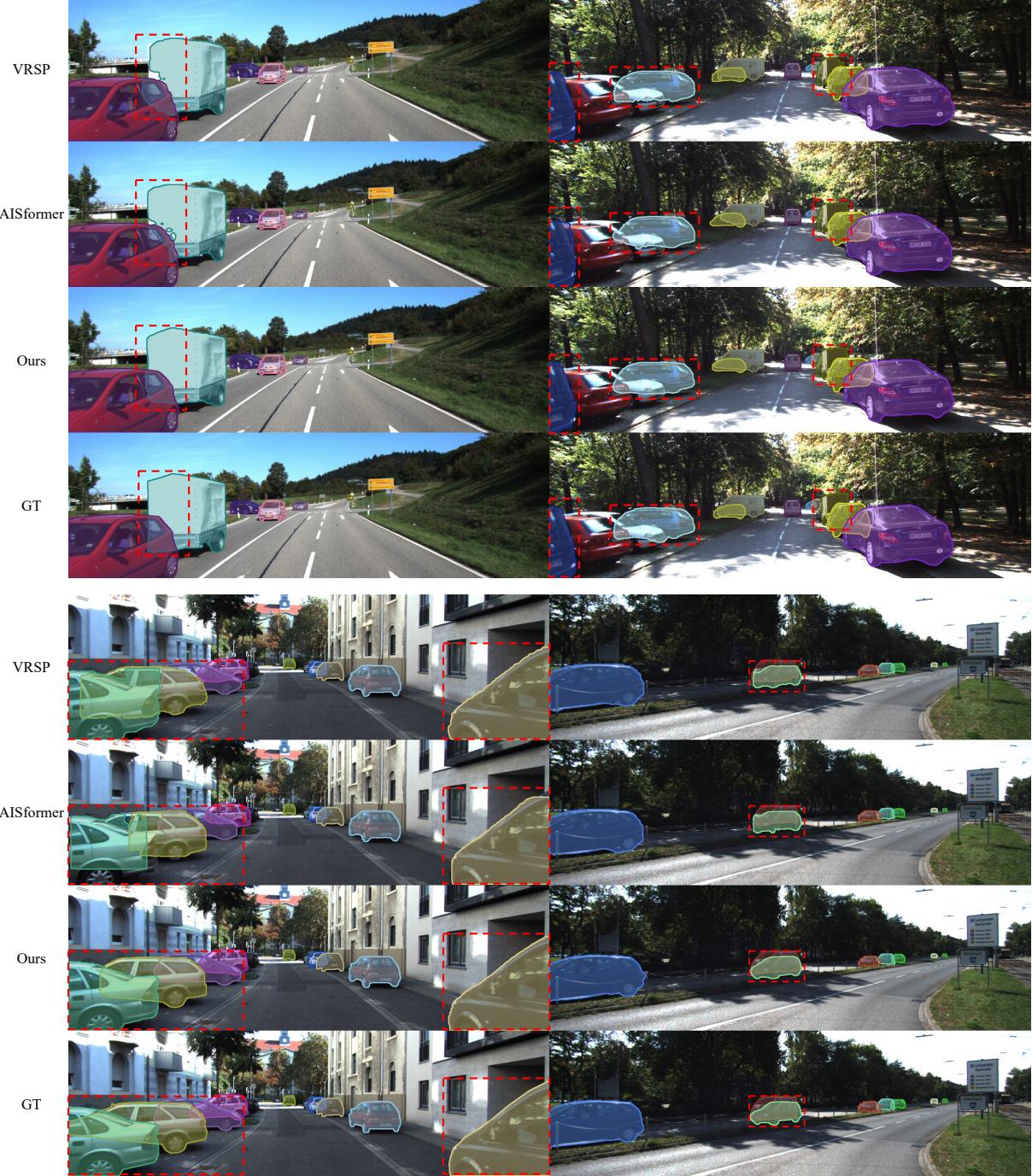


Figure D.1. The qualitative results estimated by VRSP, AISFormer, and our method. GT indicates ground-truth amodal mask.

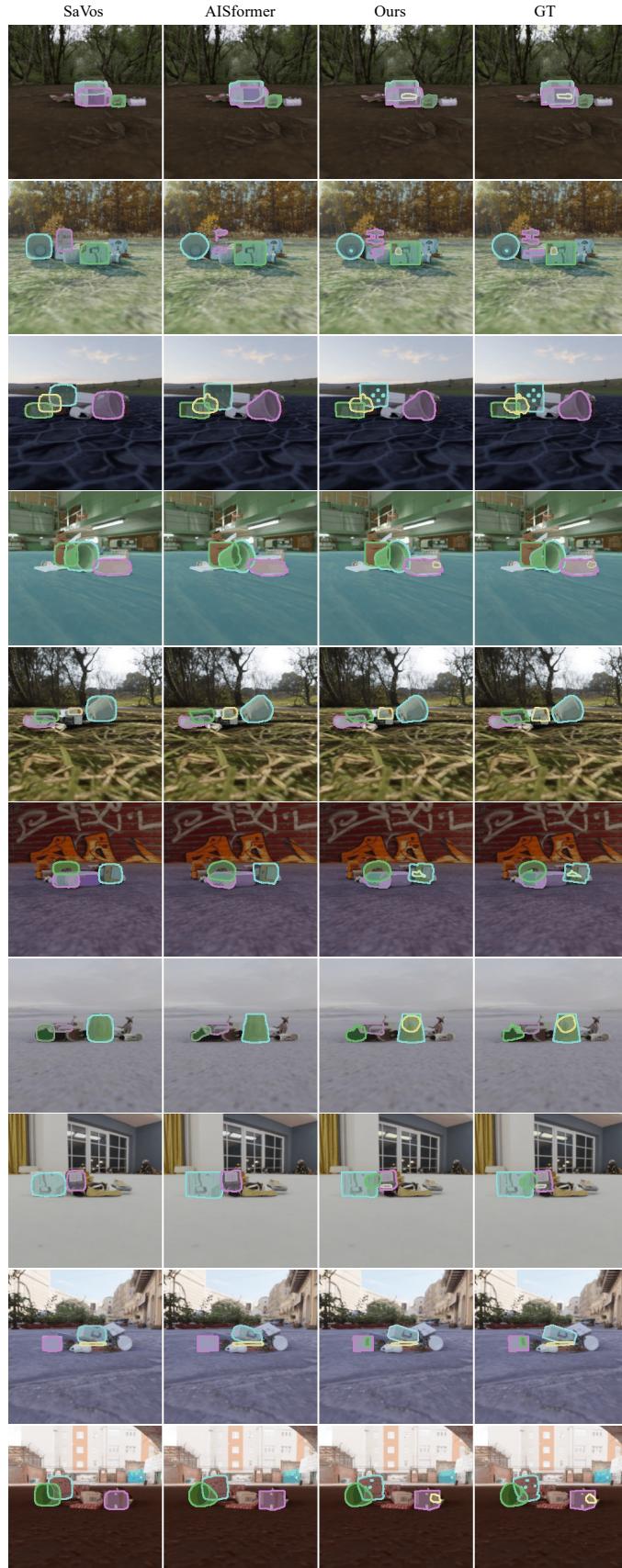


Figure D.2. The qualitative results estimated by SaVos, AISFormer, and our method. GT indicates ground-truth amodal mask.