Coarse-to-Fine Amodal Segmentation with Shape Prior (Supplementary Material)

Jianxiong Gao[1], Xuelin Qian[1,†], Yikai Wang[1], Tianjun Xiao[2,†], Tong He[2], Zheng Zhang[2], Yanwei Fu[1]
[1]Fudan University, [2]Amazon Web Service
jxgao22@m.fudan.edu.cn, {xlqian,yikaiwang19,yanweifu}@fudan.edu.cn
{tianjux,htong,zhaz}@amazon.com

A. Preliminary Knowledge

A.1. Detail of Vector-Quantization Module

This module draws inspiration from the well-known VQ-GAN [2]. 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:

$$\mathcal{Q}^* = \underset{E,G,\mathcal{Z}}{\arg\min}\max_{D} \mathbb{E}_{x\sim p(x)}[\mathcal{L}_{\mathrm{VQ}}(E,G,\mathcal{Z}) + \lambda\mathcal{L}_{\mathrm{GAN}}(\{E,G,\mathcal{Z}\},D)],$$

where

$$\mathcal{L}_{\mathrm{VQ}}(E,G,\mathcal{Z}) = \mathcal{L}_{\mathrm{rec}} + \|\mathrm{sg}[E(x)] - z_{\mathbf{q}}\|_2^2 + \beta\,\|\mathrm{sg}\,[z_{\mathbf{q}}] - E(x)\|_2^2$$

and

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

The adaptive weight λ is computed as:

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

In this context, $\mathcal{L}_{\mathrm{rec}}$ represents the perceptual reconstruction loss [7]. The symbol $\mathrm{sg}[\cdot]$ indicates the stop-gradient operation, while $\|\mathrm{sg}\,[z_{\mathbf{q}}] - E(x)\|_2^2$ is referred to as the commitment loss and has a weighting factor of β [5]. 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 [1], 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 $Y_{\mathbf{M}}^{(0)}$ (where $Y_{\mathbf{M}}$ represents the result after applying mask $\mathbf{M}$ to Y). For iteration t, our transformer operates as:

1. **Parallel Prediction**: Starting with the current set of masked tokens, $Y_{\mathrm{M}}^{(t)}$, 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 $Y_{\mathbf{M}}^{(t)}$ 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.

†: Co-corresponding authors.

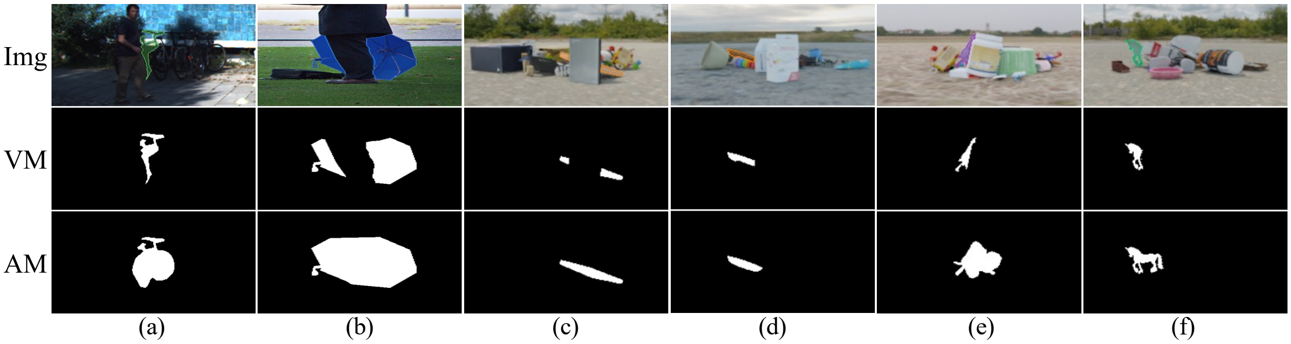

Figure 1. Supports for shape prior.

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

C. Further Ablation Studies

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

C.1. 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 [3] 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 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 2. Results indicate the effectiveness of the ST module as well as our introduced roll mechanism.

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 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.

C.2. 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 3 shows the mIoU metrics for the two datasets. The results indicate our attention mechanism improves the quality of amodal masks.

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

D. Supports for the claim of shape prior

Our claim of shape prior is based on a common phenomenon, which is supported by Fig. 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.

E. 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

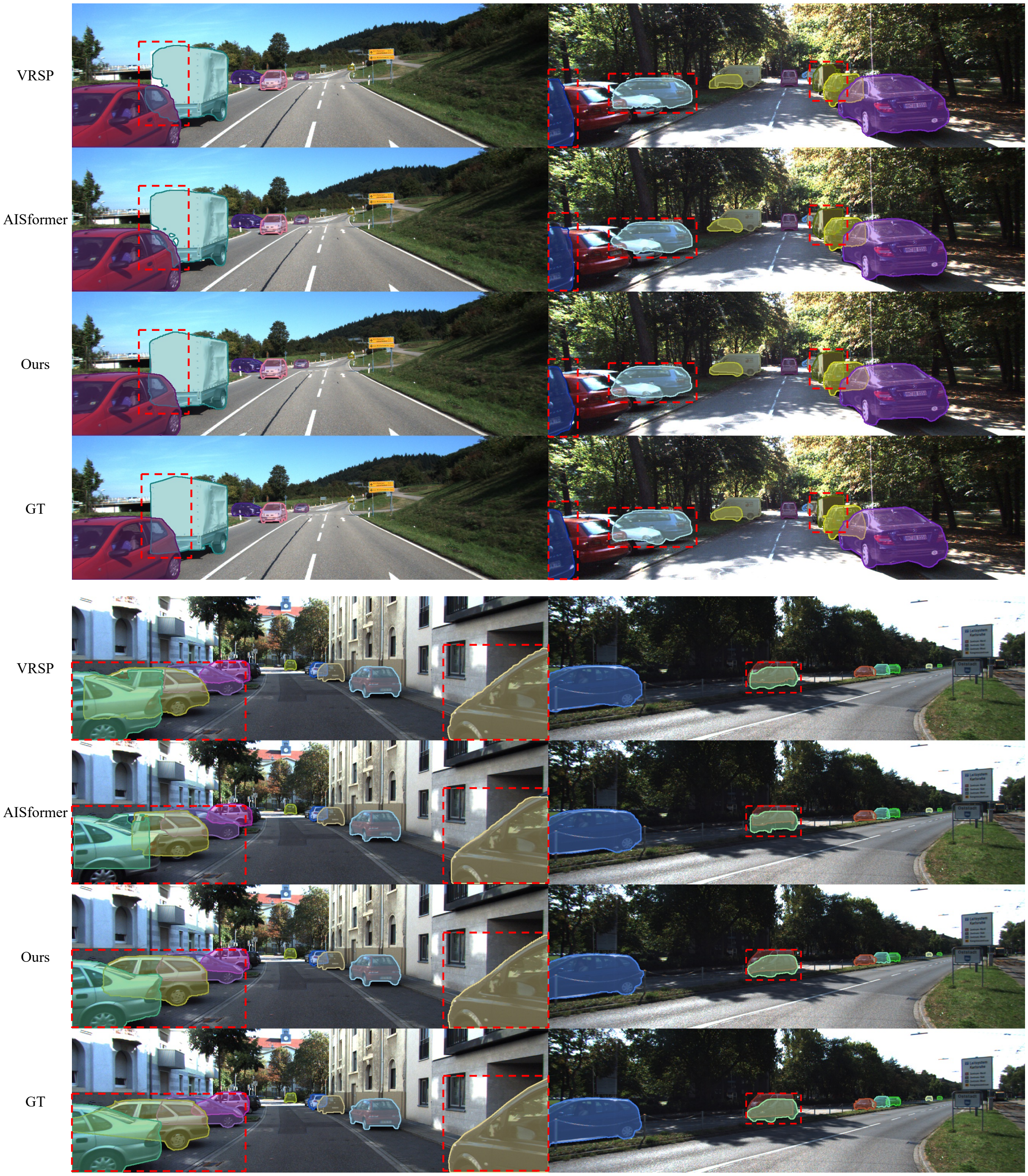

Figure 2. The qualitative results estimated by VRSP, AISFormer, and our method. GT indicates ground-truth amodal mask.

demonstrate the effectiveness of our method.

E.1. Visualization on KINS Dataset

To show the performance of our method on real scenarios, we show more results from KINS in Figure 2. In these images, for fair comparison, we select the intersection of the amodal masks predicted by VRSP [6] and AISFormer [4]. 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.

E.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 3. Our method predicts the invisible masks excellently by extracting valid spatio-temporal features and outperforms all the baselines.

F. 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.

References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

[6] 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.

[7] 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.

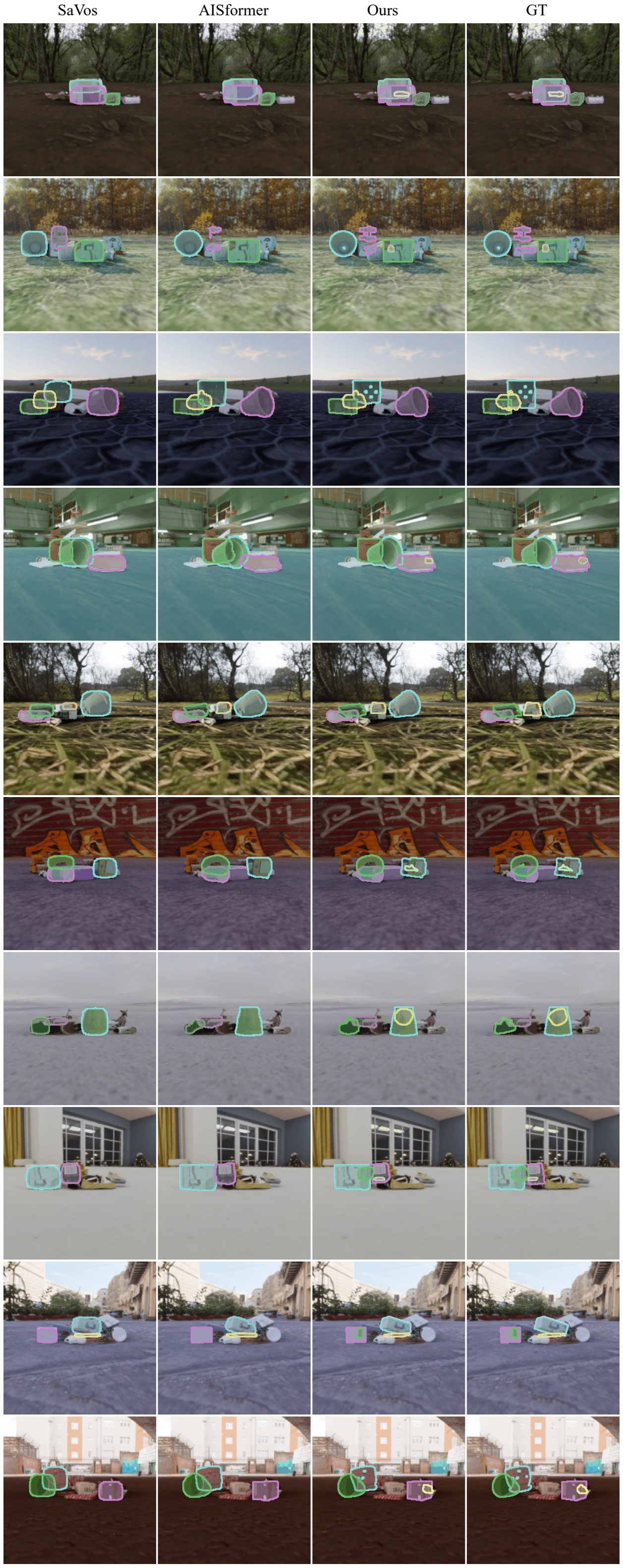

Figure 3. The qualitative results estimated by SaVos, AISFormer, and our method. GT indicates ground-truth amodal mask.