ShapeFormer: Shape Prior Visible-to-Amodal Transformer-based Amodal Instance Segmentation

Minh Tran[1], Winston Bounsavy[1], Khoa Vo[1], Anh Nguyen[2], Tri Nguyen[3], Ngan Le[1]

***Abstract*—Amodal Instance Segmentation (AIS) presents a challenging task as it involves predicting both visible and occluded parts of objects within images. Existing AIS methods rely on a bidirectional approach, encompassing both the transition from amodal features to visible features (amodal-to-visible) and from visible features to amodal features (visible-to-amodal). Our observation shows that the utilization of amodal features through the amodal-to-visible can confuse the visible features due to the extra information of occluded/hidden segments not presented in visible display. Consequently, this compromised quality of visible features during the subsequent visible-to-amodal transition. To tackle this issue, we introduce ShapeFormer, a decoupled Transformer-based model with a visible-to-amodal transition. It facilitates the explicit relationship between output segmentations and avoids the need for amodal-to-visible transitions. ShapeFormer comprises three key modules: (i) Visible-Occluding Mask Head for predicting visible segmentation with occlusion awareness, (ii) Shape-Prior Amodal Mask Head for predicting amodal and occluded masks, and (iii) Category-Specific Shape Prior Retriever aims to provide shape prior knowledge. Comprehensive experiments and extensive ablation studies across various AIS benchmarks demonstrate the effectiveness of our ShapeFormer. The code is available at: https://github.com/UARK-AICV/ShapeFormer**

***Index Terms*—Amodal Instance Segmentation, Shape Prior, Transformer**

I. Introduction

Human perception grants us the remarkable ability to comprehend objects in their entirety i.e., an ability known as amodal perception [9]. Based on such observation, pioneering work by [10], [26] introduced the concept of *amodal instance segmentation (AIS)*, which focuses on determining the complete shape of an object, including its visible and occluded regions. Indeed, AIS holds significant potential in various applications, including robot manipulation [1] and autonomous driving [14]. As depicted in Fig. 1(a), AIS aims to produce the segment mask of the visible part of chocolate box (*visible*), the entire chocolate box (*amodal*), even when a part of it was occluded by a bag of tomatoes. The segment of this bag of tomatoes is considered as an*occluding* mask. The mask joined between the amodal and occluding region is considered as an *occluded* mask. In AIS, visible and amodal masks are obligated while occluding and occluded masks are supplemental outputs.

The literature has witnessed the emergence of numerous approaches [3], [6], [10], [12], [14], [19], [22] which address the AIS challenge across various benchmarks [3], [14], [26]. These methods typically utilize a bidirectional approach for feature learning, involving transitions between amodal and visible features in both directions – from amodal to visible (amodal-to-visible) and from visible to amodal (visible-to-amodal). As illustrated in Fig. 1, recent existing techniques endeavor to capture the interplay between visible-to-amodal and amodal-to-visible relationships through mechanisms like feature concatenation as in ASN [14] and VRSP-Net [23], or self-attention in AISFormer [19]. However, our examination of these approaches indicates that their predictions of visible masks fall short. This inadequacy is evident in Fig. 1 (b),

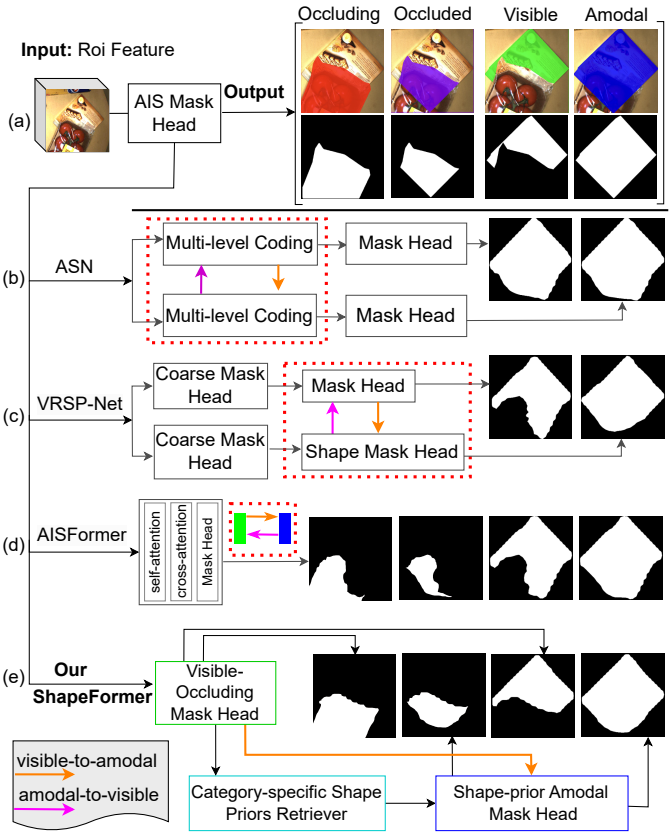


Fig. 1: Comparison between our ShapeFormer and existing SOTA approaches. (a) AIS setting, which takes a RoI feature as input and returns four masks including occluding, occluded, visible and amodal. (b) ASN [14]: bidirectional learning at multi-level coding via feature concatenation. (c) VRSP-Net [23]: bidirectional learning at mask head via feature concatenation. (d) AISFormer [19]: bidirectional learning at embeddings via self-attention. (e) Our **ShapeFormer** omits the amodal-to-visible transition, leverages the precise visible feature and shape prior knowledge to predict amodal mask.

[1]AICV Lab, Department of EECS, University of Arkansas, USA. minht@uark.edu
2 Department of CS, University of Liverpool, UK.
3 Cruise LLC, USA.

(c), and (d), where their visible masks result results exhibit notable deficiencies. We hypothesize that the amodal-to-visible relation introduces confusion into the visible masks prediction because unlike visible masks, amodal masks include regions that are not presented in the image display [2]. Consequently, the utilization of amodal features to enhance visible features could potentially compromise the precision of visible predictions. Moreover, when the visible mask itself is inadequate, the potential of visible features to enhance amodal mask predictions remains unrealized.

To tackle the aforementioned challenge, we introduce ShapeFormer, a novel approach that focuses exclusively on the visible-to-amodal transition, departing from the bidirectional-transition approach used in existing methods. Recent research [2], [6], [23], [24] underscores the efficacy of incorporating shape prior information during this transition. Building on this insight, we propose integrating shape prior knowledge into ShapeFormer. Traditional shape prior AIS methods typically employ vanilla or variational autoencoders to acquire shape priors, followed by refining coarse amodal masks. However, they often overlook the importance of object categories in shape retrieval, which can lead to overfitting the shape prior model to the training dataset. In contrast, our ShapeFormer employs a category-specific vector quantized variational autoencoder to retrieve shape priors based on the visible mask and the corresponding object category *id*. Furthermore, recent research AISFormer [19] highlights the superior effectiveness of transformer-based architectures over CNN-based ones in modeling relationships among AIS output masks. Therefore, ShapeFormer is defined as a transformer-based architecture, aligning with these advancements in modeling techniques.

In particular, Fig. 1 (e) provides an overview of our proposed ShapeFormer, which consists of three key modules: (i) Visible-Occluding (Vis-Occ) Mask Head: This module predicts the visible segmentation mask and its category *id* while acknowledging occluding segmentation. (ii) Category-Specific Shape Prior Retriever (Cat-SP Retriever): Utilizing category-specific vector quantized variational autoencoder, coupled with data augmentation, to retrieve shape priors based on the visible mask and the corresponding category *id*. (iii) Shape-prior Amodal (SPA) Mask Head: Instead of simply concatenating the retrieved shape prior with the visible feature and coarse amodal feature as done in previous approaches [6], [23], we leverage the shape prior knowledge as a mask within a transformer decoder's masked attention module and presents a novel shape-prior masked attention mechanism. This integration empowers the model to focus on specific regions when predicting the amodal mask, thus enhancing its accuracy and performance.

In summary, our contributions are as follows:

• We introduce ShapeFormer, a novel AIS framework with a decoupled transformer-based architecture that focuses on the visible-to-amodal transition. ShapeFormer explicitly models the relation among output segmentations while omitting the amodal-to-visible transition to prevent deficiencies in visible segmentation observed in prior works.

• We develop Cat-SP Retriever, a category-specific vector quantized autoencoder that leverages visible mask information and pretrains discrete codebooks for each object category to effectively retrieve shape priors. Additionally, we enhance the performance of shape prior retrieving by incorporating occlusion data augmentation, enabling better generalization to different shapes and preventing overfitting.

• We introduce the shape-prior masked attention to decode the amodal segmentation using the retrieved shape prior. This attention mechanism enables the model to focus on relevant parts of objects when predicting the amodal mask.

• Comprehensive experiments across four AIS benchmarks shows that our ShapeFormer consistently outperforms previous state-of-the-art methods. We also conduct an analysis on the effectiveness of the visible-to-amodal modeling in ShapeFormer compared to bidirectional modeling baseline. Finally, extensive ablation studies are carried out to examine the contributions of our proposed Cat-SP Retriever and shape-prior masked attention to the new state-of-the-art performance set by our ShapeFormer.

II. Related Work

Amodal instance segmentation involves predicting an object's shape, including both its visible and occluded parts. Li and Malik [10] first propose a method to tackle AIS by enlarging the modal bounding box following the direction of high heatmap values and synthetically adds occlusion. Subsequent to this pioneering work, numerous other methodologies have emerged in the literature.

Notably, ORCNN [3] introduces amodal and visible instance mask heads, along with an additional mask head for occluded mask prediction. Building upon ORCNN, ASN [14] incorporates a multi-level coding module for bidirectional modeling of visible and amodal features. BCNet [8] augments amodal mask prediction with an extra branch for occluding mask prediction within the bounding box. AISFormer [19] introduces a transformer-based mask head, showcasing the effectiveness of transformer modeling for generating AIS output masks. However, their model implicitly learn all the relationship between output masks in one transformer model. As we mentioned earlier, this modeling contains the bidirection relation between visible and amodal feature, making visible segmentation output defective, consequently impacting the quality of the amodal segmentation output.

Recent studies [6], [23] highlight the advantages of incorporating shape priors into AIS. These methods leverage mask shapes as prior knowledge to enhance amodal mask predictions. VRSP-Net [23] predicts coarse amodal masks, retrieves shape priors through a plain autoencoder, and refines final amodal mask predictions. AmodalBlastomere [6] uses a similar approach with a variational autoencoder for blastomere and cell segmentation. Despite their advancements, these methods tend to neglect the significance of object categories when retrieving prior shapes. Furthermore, their training procedures often result in overfitting the shape prior model to the training dataset. Additionally, these approaches

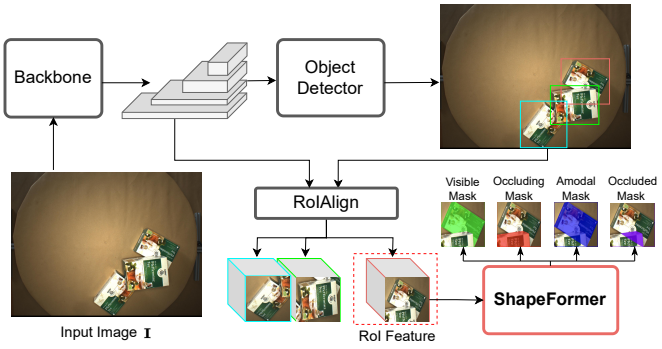


Fig. 2: The overview pipeline illustrates the integration of our ShapeFormer as the amodal mask head within an object detection framework. The input image $\mathbf{I}$ goes through a backbone followed by an object detector to predict the regions of interest (RoI) and extract their corresponding feature. These RoI features are then processed through the proposed **ShapeFormer** (Fig. 3) to obtain the desired output AIS masks.

simply employ the shape prior by concatenating it with the visible features for refining amodal masks.

Our proposed method, ShapeFormer, exploits the strengths of both transformers and shape priors in AIS while addressing their inherent challenges. Specifically, our approach tackles bidirectional feature learning in previous works (e.g. AISFormer [19], ASN [14], VRSP-Net [23]) by decoupling the model's transition from visible to amodal with the inclusion of shape priors. Furthermore, we mitigate previous issues associated with shape-prior-based methods by introducing a category-specific shape prior retriever, coupled with occlusion copy-paste augmentation to reduce overfitting. Additionally, the incorporation of shape-prior masked attention enables effective utilization of shape priors within a transformer-base model to predict amodal segmentation.

TABLE I: *Network architecture comparison* between our proposed ShapeFormer and existing AIS approaches. AE and VAE denote Autoencoder and Variational Autoencoder.

Methods	**Networks**	**visible -to-amodal**	**amodal-to -visible**	**Shape-prior**
ASN [14]	CNNs	✓	✓	✗
AISFormer [19]	Transformer	✓	✓	✗
VRSP-Net [23]	CNNs	✓	✓	Vanilla AE
ShapeFormer	Transformer	✓	✗	Conditional Vector Quantized VAE

III. Proposed ShapeFormer

We commence by providing an overview pipeline illustrating the integration of our ShapeFormer as the amodal mask head within an object detection framework. Subsequently, we introduce ShapeFormer, which incorporates a transformer-based approach for visible-to-amodal transition, along with shape prior modeling. Lastly, we outline the objective functions for optimizing the network during training.

A. Overall AIS Setup

Fig. 2 illustrates the integration of our ShapeFormer as the amodal mask head within an object detection framework. Given an input image $\mathbf{I}$, our framework follows most of previous AIS settings [3], [8], [23] by utilizing a pre-trained backbone network, such as ResNet [4], RegNet [17] to extract spatial visual representation. An object detector such as FCOS [18], or Faster-RCNN [4], can be subsequently adopted to obtain n regions of interest (RoI) predictions and their corresponding visual features $\{\mathbf{F}^i\}_{i=1}^n$. We also follow most of previous works [8], [19], [23], choosing Faster R-CNN as our object detector for fair comparison. Here, each RoI is presented by its visual feature $\mathbf{F}^i \in \mathbb{R}^{C_e \times H_r \times W_r}$, where C_e denotes the feature channel size and $H_r \times W_r$ represents the spatial shape of the pooling feature. In this context, given a RoI, our ShapeFormer takes $\mathbf{F}^i$ as input and aims to predict the amodal mask $\mathbf{M}_a^i$, the visible mask $\mathbf{M}_v^i$, the occluding mask $\mathbf{M}_o^i$ and the occluded mask $\mathbf{M}_p^i$.

B. ShapeFormer

Fig. 3 illustrates the key modules of our proposed ShapeFormer, which takes the RoI feature $\mathbf{F}^i$ as input. The first module, *Vis-Occ Mask Head*, is designed to precisely predict the visible mask while considering occlusion (i.e. occluding mask). The second module, *Cat-SP Retriever*, is responsible for retrieving a shape prior based on the visible mask and the instance's category *id*. The final module, *SPA Mask Head*, utilizes the shape prior and embeddings produced by the preceding modules to predict amodal mask and occluded mask.

*1) **Vis-Occ Mask Head**:* Operating on the RoI feature $\mathbf{F}_v^i$, this module aims to make precise predictions for visible segmentation while taking occlusions into consideration. To capture the relation between the visible and the occluding masks, we introduce a transformer-based mask predictor inspired by the previous work [19], which demonstrated the efficacy of relation modeling among object masks within an RoI. In fact, we first initialize two learnable per-segment query embeddings $\mathbf{q}_v \in \mathbb{R}^{C_e}$ and $\mathbf{q}_o \in \mathbb{R}^{C_e}$ that represent the embedding of the visible segmentation and the occluding segmentation, respectively. We also extract the attention feature $\mathbf{F}_v^i$ from $\mathbf{F}^i$ by a series of three 3×3 convolutional layers with a stride of 1, followed by the extraction of Vis-Occ feature $\mathbf{E}_v^i$ by a 2×2 transposed convolutional layer with a stride of 2 plus a 1×1 convolutional layer with a stride of 1. Here, $\mathbf{F}_v^i$ represents key-value cross attention feature for decoding the mask embeddings whereas $\mathbf{E}_v^i$ encapsulates the semantic feature concerning whether each pixel in the RoI belongs to the visible mask of the primary object or pertains to occluding objects. To decode the two query embeddings $\mathbf{q}_v, \mathbf{q}_o$ from the attention feature $\mathbf{F}_v^i$, we introduce the *Vis-Occ Transformer Decoder* $\mathcal{D}_v$ with L_v layers, as illustrated in Fig. 4 (a). Each layer contains one self-attention block, responsible for learning the relation between visible and occluding embeddings, followed by a cross attention block that learns the relation between the two embeddings with the attention feature $\mathbf{F}_v^i$. Formally, the decoded visible and occluding embeddings, denoted as $\tilde{\mathbf{x}}_v$ and $\tilde{\mathbf{x}}_o$, respectively, can be computed as $[\tilde{\mathbf{x}}_v, \tilde{\mathbf{x}}_o] = \mathcal{D}_v([\mathbf{q}_v, \mathbf{q}_o], \mathbf{F}_v^i)$. They are then correlated with everypixel embedding in $\mathbf{E}_v^i$ through a Vis-Occ Aware Mask Extraction to determine whether the pixel

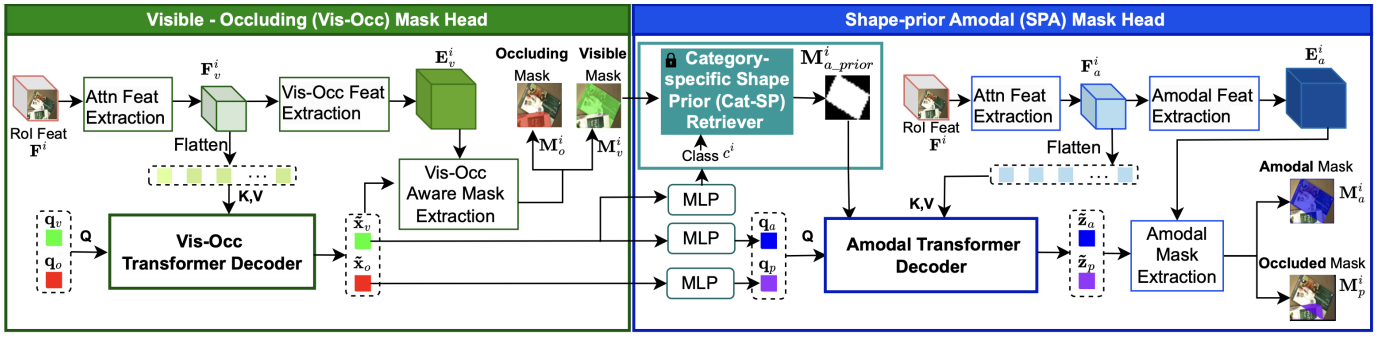
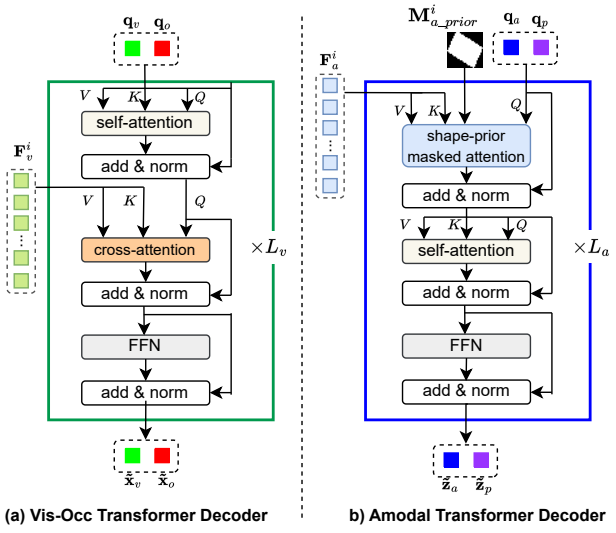


Fig. 3: The pipeline of our **ShapeFormer** consisting of three main components of Visible-Occluding (Vis-Occ) Mask Head, Shape-prior Amodal (SPA) Mask Head, and Category-specific Shape Prior (Cat-SP) Retriever. Feat denotes feature.



(a) Vis-Occ Transformer Decoder b) Amodal Transformer Decoder

Fig. 4: Detailed architecture. (a): **Vis-Occ Transformer Decoder** models the relation between visible mask and occluding mask. (b): **Amodal Transformer Decoder** with **shape-prior masked attention** models the relation between amodal mask and occluded mask.

belongs to the visible segment or the occluding segment. The Vis-Occ Aware Mask Extraction is designed as a dot product on the feature dimension C_e. The output visible mask (denoted as $\mathbf{M}_v^i$), the occluding mask (denoted as $\mathbf{M}_o^i$) are formally computed as follow:

$$[\mathbf{M}_v^i, \mathbf{M}_o^i] = \text{sigmoid}([\tilde{\mathbf{x}}_v, \tilde{\mathbf{x}}_o] \otimes \mathbf{E}_v^i) \quad (1)$$

*2) **Cat-SP Retriever**:* Fig. 5 illustrates the overall architecture of our Cat-SP Retriever, denoted as f_S. It takes a visible mask $\mathbf{M}_v^i$ along with its corresponding category *id* (denoted as c^i) as inputs. The purpose of our Cat-SP Retriever is to search for a category-specific shape prior denoted as $\mathbf{M}_{a\_prior}^i$, achieved through the operation $f_S(\mathbf{M}_v^i, c^i)$. To obtain category *id* c^i of $\mathbf{M}_v^i$, we employ a MLP consisting of two hidden layers followed by a softmax layer to transform the decoded visible embedding $\tilde{\mathbf{x}}_v$ into the category probability $\mathbf{p}^i \in \mathbb{R}^C$, where C presents the total number of categories. The category *id* is then obtained by applying $\arg max$ on $\mathbf{p}^i$ as below.

$$c^i = \arg max\ \mathbf{p}^i, \text{ where } \mathbf{p}^i = \text{softmax}(\text{MLP}(\tilde{\mathbf{x}}_v)) \quad (2)$$

We propose the utilization of a variational autoencoder with vector quantization mechanism to conduct f_S. We initialize C

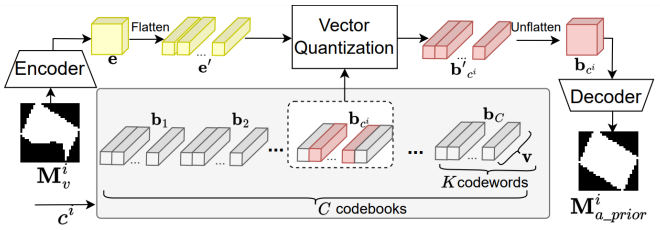


Fig. 5: Flowchart of **Cat-SP Retriever**. Input is visible mask $\mathbf{M}_v^i$ and its class label c^i. Output is category-specific shape prior $\mathbf{M}_{a\_prior}^i = f_S(\mathbf{M}_v^i, c^i)$.

codebooks representing C categories in a dataset. A codebook $\mathbf{b}_j \in \mathbb{R}^{K \times \mathbf{v}}, j \in \{1, 2, ..., C\}$ is a collection of K codewords (vectors) of size $\mathbf{v}$ representing the possible latent codes for object shape of each category. The process begins with encoding the input visible mask $\mathbf{M}_v^i$ into a encoded feature of $\mathbf{e} \in \mathbb{R}^{k \times k \times \mathbf{v}}$ using an encoder, structured similarly to the UNet encoder [15]. The encoded feature $\mathbf{e}$ is then flatten into a list of $m = k \times k$ latent vectors of size $\mathbf{v}$. These latent vectors $\mathbf{e}' \in \mathbb{R}^{m \times \mathbf{v}}$, is subjected to the quantization step. Here, the category-specific codebook $\mathbf{b}_{c^i}$ is determined based on the predicted category c^i from the previous step. Each latent vector from $\mathbf{e}'$ is compared to the codewords in $\mathbf{b}_{c^i}$. The nearest codeword, determined by cosine distance, is selected as the quantized representation for each latent vector. After quantization, a set of m codewords, denoted as $\mathbf{b}'_{c^i} \in \mathbb{R}^{m \times K}$, is selected from $\mathbf{b}_{c^i}$ to represent the m encoded latent vectors in $\mathbf{e}'$. The collection $\mathbf{b}'_{c^i}$ is then unflattened to form a spatial decoded feature $\mathbf{b}_{c^i}$, which is then passed through a Unet decoder to yield the corresponding $\mathbf{M}_{a\_prior}^i$.

*3) **SPA Mask Head**:* This final module is designed to predict amodal mask using occluding embedding $\tilde{\mathbf{x}}_o$, visible embedding $\tilde{\mathbf{x}}_v$ from the Vis-Occ Mask Head and shape prior $\mathbf{M}_{a\_prior}^i$ from Cat-SP Retriever. Similar to the Vis-Occ Mask Head, we also extract the amodal attention feature $\mathbf{F}_a^i$ from $\mathbf{F}^i$, followed by the extraction of amodal feature $\mathbf{E}_a^i$, both using the same convolutional operations as in Vis-Occ Mask Head. Note that $\mathbf{F}_a^i$ represents key-value cross attention feature for decoding the mask embeddings whereas $\mathbf{E}_a^i$ encapsulates the amodal semantic feature. It is important to note that amodal semantic also includes the occluded information, thus we also predict the occluded mask, learning from the occluding embedding $\tilde{\mathbf{x}}_o$. This enables the model to discern which

parts of the occluding object obscure the amodal portion. To accomplish this, we create learnable queries for both amodal and occluded masks, denoted as $\mathbf{q}_a$ and $\mathbf{q}_p$, respectively. Due to $\tilde{\mathbf{x}}_v$ and $\tilde{\mathbf{x}}_o$ are in the visible embedding space whereas $\mathbf{q}_a$ and $\mathbf{q}_p$ are in the amodal embedding space, we propose to use MLPs to transfer those two embedding spaces, i.e. $\mathbf{q}_a = MLP(\tilde{\mathbf{x}}_v)$ and $\mathbf{q}_p = MLP(\tilde{\mathbf{x}}_o)$. We then introduce Amodal Transformer Decoder $\mathcal{D}_a$, which incorporates shape prior $\mathbf{M}^i_{a\_prior}$. This decoder is responsible for decoding the amodal embedding $\tilde{\mathbf{z}}_a$ and the occluded embedding $\tilde{\mathbf{z}}_p$, as detailed in Fig. 4(b). In this process, $\mathbf{q}_a$ and $\mathbf{q}_p$ are treated as queries $\mathbf{Q}$, and the amodal attention feature $\mathbf{F}^i_a$ serves as $\mathbf{K}$ and $\mathbf{V}$.

Differing from a conventional transformer decoder that employs the traditional cross-attention mechanism, we introduce a *shape-prior masked attention* within the Amodal Transformer Decoder $\mathcal{D}_a$. This attention mechanism takes $\mathbf{Q}$, $\mathbf{K}$, $\mathbf{V}$, and the shape prior $\mathbf{M}^i_{a\_prior}$ as inputs. The incorporation of shape prior $\mathbf{M}^i_{a\_prior}$ in this masked attention allows the model to focus on specific regions, enhancing both the efficiency and effectiveness for predicting amodal segmentation. To elaborate, at each layer l of the decoder, given the intermediate output from the previous layer $l-1$, denoted as $\mathbf{Z}^l = [\mathbf{z}^{l-1}_a, \mathbf{z}^{l-1}_p]$, the output of the masked attention at layer l is as below:

$$\mathbf{Z}^l = \text{softmax}(\mathcal{M} + \mathbf{Q}\mathbf{K}^T)\mathbf{V} + \mathbf{Z}^{l-1} \tag{3a}$$

$$\mathbf{Q} = \mathbf{Z}^l \cdot \mathbf{W}^{\mathbf{Q}};\ \mathbf{K} = \mathbf{F}^i_a \cdot \mathbf{W}^{\mathbf{K}}\ ;\ \mathbf{V} = \mathbf{F}^i_a \cdot \mathbf{W}^{\mathbf{V}} \tag{3b}$$

$$\mathcal{M}(x,y) = \begin{cases} 0 & \text{if } \mathbf{M}_{a\_prior}(x,y) = 1 \\ -\infty & \text{otherwise} \end{cases} \tag{3c}$$

Here, $\mathbf{W}^{\mathbf{Q}}$, $\mathbf{W}^{\mathbf{K}}$, $\mathbf{W}^{\mathbf{V}}$ are learning parameters of query $\mathbf{Q}$, key $\mathbf{K}$ and value $\mathbf{V}$, respectively. Following the masked attention, the process continues with self-attention, which aims to capture the correlation between the amodal and occluded embeddings. The decoding process of $\mathcal{D}_a$ can be expressed as follow:

$$[\tilde{\mathbf{z}}_a, \tilde{\mathbf{z}}_p] = \mathcal{D}_a([\mathbf{q}_a, \mathbf{q}_p], \mathbf{F}^i_a, \mathbf{M}_{a\_prior}) \tag{4}$$

where $\tilde{\mathbf{z}}_a$ and $\tilde{\mathbf{z}}_p$ are then correlated with each pixel embedding in $\mathbf{E}^i_a$ through an Amodal Mask Extraction, which is designed as a dot product on the feature dimension C_e to derive the amodal and the occluded masks. In summary, the output amodal mask (denoted as $\mathbf{M}^i_a$), the occluded mask (denoted as $\mathbf{M}^i_p$) are computed as follow:

$$\mathbf{M}^i_a = \text{sigmoid}(\tilde{\mathbf{z}}_a \otimes \mathbf{E}^i_a); \mathbf{M}^i_p = \text{sigmoid}(\tilde{\mathbf{z}}_p \otimes \mathbf{E}^i_a) \tag{5}$$

C. Training Process

1) ***Training Cat-SP Retriever****:* To achieve representative codebooks for our Cat-SP Retriever, we employ the training process optimizing the following objective functions.

$$\begin{aligned} \mathcal{L}_{csp} &= \mathcal{L}_{rec} + \mathcal{L}_{vq} \\ \mathcal{L}_{rec} &= MSE(\mathbf{M}^i_{a\_prior}, \mathbf{M}^i_{a\_gt}) \\ \mathcal{L}_{vq} &= MSE(\mathbf{e}', \mathbf{b}'_{c^i}) \end{aligned} \tag{6}$$

Here, the reconstruction loss, denoted as $\mathcal{L}_{rec}$, is calculated by computing the mean square error (MSE) between the predicted shape prior $\mathbf{M}^i_{a\_prior}$ and the corresponding ground truth amodal mask $\mathbf{M}^i_{a\_gt}$. Meanwhile, the vector quantization loss, $\mathcal{L}_{vq}$ is optimized to learn the codewords in the selected codebook to better match the flatten encoded feature $\mathbf{e}'$. Additionally, during the training of f_S, we generate augmented visible masks from the ground-truth amodal mask. This augmentation helps enhancing the generalization of Cat-SP Retriever by covering more occlusion scenarios that can occur during testing. Examples showcasing our augmented data can be seen in Fig. 6. Our f_S is pretrained and remains fixed during during the training of ShapeFormer, serving as a consistent source of shape prior knowledge throughout the training process.

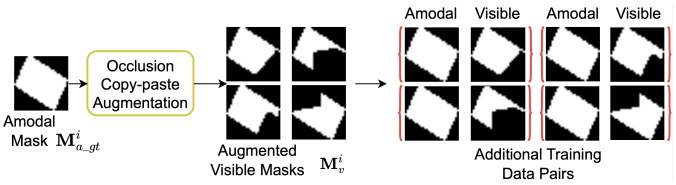


Fig. 6: Generation of augmented visible masks ($\mathbf{M}^i_v$) from ground-truth amodal masks ($\mathbf{M}^i_{a\_gt}$) during training Cat-SP Retriever.

2) ***Training ShapeFormer****:* Our ShapeFormer is trained in an end-to-end manner concurrently with the object detection framework. Our training follows AIS protocols as shown in Fig.2, employing a two-stage instance segmentation process similar to Mask R-CNN. This approach enables the concurrent training of both the bounding box and amodal mask prediction heads without the need for pre-bootstrapping in object detection. In other words, the training procedure optimizes a multi-task loss function $\mathcal{L}$ as follow:

$$\mathcal{L} = \mathcal{L}_{det} + \mathcal{L}_{cls} + \mathcal{L}_v + \mathcal{L}_o + \mathcal{L}_a + \mathcal{L}_p \tag{7}$$

where $\mathcal{L}_{det}$ is object detection loss, defined similarly to that in Faster R-CNN object detection. The occluding mask loss $\mathcal{L}_o$, the visible mask loss $\mathcal{L}_v$, the amodal mask loss $\mathcal{L}_a$, occluded mask loss $\mathcal{L}_p$ and the classification loss $\mathcal{L}_{cls}$ are computed as follow:

$$\begin{aligned} \mathcal{L}_o &= \mathcal{L}_{BCE}(\mathbf{M}^i_o, \mathbf{M}^i_{o\_gt}), \mathcal{L}_v = \mathcal{L}_{BCE}(\mathbf{M}^i_v, \mathbf{M}^i_{v\_gt}) \\ \mathcal{L}_a &= \mathcal{L}_{BCE}(\mathbf{M}^i_a, \mathbf{M}^i_{a\_gt}), \mathcal{L}_p = \mathcal{L}_{BCE}(\mathbf{M}^i_p, \mathbf{M}^i_{p\_gt}) \\ \mathcal{L}_{cls} &= \mathcal{L}_{CE}(\mathbf{p}^i, c^i_{gt}) \end{aligned} \tag{8}$$

Here, $\mathbf{M}^i_{o\_gt}, \mathbf{M}^i_{v\_gt}, \mathbf{M}^i_{p\_gt}$, and $\mathbf{M}^i_{a\_gt}$, are the ground truth of the occluding, visible, occluded and amodal masks, respectively. c^i_{gt} is the ground-truth category of the RoI. $\mathcal{L}_{BCE}$ denotes the binary cross entropy loss whereas $\mathcal{L}_{CE}$ denotes the cross entropy loss.

IV. Experiments

A. Datasets, Metrics and Implementation Details

Datasets: We benchmark our ShapeFormer on four AIS datasets, namely KINS [14], COCOA [26], COCOA-cls [3], and D2SA [3]. KINS is a large-scale traffic dataset with 95,311 training instances and 92,492 testing instances with 7 categories. COCOA is an AIS dataset that is derived from MSCOCO [11] with no categories, including 15,139 training

TABLE II: *Performance comparison on KINS test set* with various backbones. † indicates our reproduced results.

Backbones& Methods	Venue	Shape Prior	Visible		Amodal			
			$AP\uparrow$	$AR\uparrow$	$AP\uparrow$	$AP_{50}\uparrow$	$AP_{75}\uparrow$	$AR\uparrow$
ResNet-50 PCNet [25]	CVPR20	✗	-	-	29.1	51.8	29.6	18.3
ASBU [13]	ICCV21	✗	-	-	29.3	52.1	29.7	18.4
Mask R-CNN [7]	ICCV17	✗	28.0	19.2	30.0	54.5	30.1	19.4
ORCNN [3]	WACV19	✗	28.8	20.0	30.6	54.2	31.3	19.7
ASN [14]	CVPR19	✗	-	-	32.2	-	-	-
AISFormer [19]	BMVC22	✗	29.7	20.0	33.8	57.8	35.3	21.1
AmodalBlastomere [6]	TMI20	✓	-	-	30.3	-	-	-
VRSP-Net [23]	AAAI21	✓	29.9	19.9	32.1	55.4	33.3	20.9
ShapeFormer(Ours)	-	✓	**31.3**	**21.1**	**34.1**	**58.6**	**35.7**	**22.0**
ResNet-101 Mask R-CNN [4] †	ICCV17	✗	-	-	30.2	54.3	30.4	19.5
BCNet [8]	CVPR21	✗	-	-	28.9	-	-	-
BCNet [8] †	CVPR21	✗	-	-	32.6	57.2	35.4	21.5
AISFormer [19]	BMVC22	✗	30.9	20.1	34.6	58.2	36.7	21.9
ShapeFormer(Ours)	-	✓	**32.6**	**22.3**	**35.2**	**59.3**	**37.2**	**23.0**
RegNet ASPNet [12]	CVPR22	✗	-	-	35.6	-	-	-
AISFormer [19]	BMVC22	✗	31.9	21.1	35.6	59.9	37.0	22.5
ShapeFormer(Ours)	-	✓	**33.7**	**22.8**	**36.1**	**59.9**	**38.7**	**23.0**

TABLE III: *Performance comparison on COCOA test set* with various backbones. † indicates our reproduced results.

Backbones& Methods	Venue	Shape Prior	Visible		Amodal			
			$AP\uparrow$	$AR\uparrow$	$AP\uparrow$	$AP_{50}\uparrow$	$AP_{75}\uparrow$	$AR\uparrow$
ResNet-50 PCNet [25]	CVPR20	✗	-	-	22.6	46.8	19.7	6.3
ASBU [13]	ICCV21	✗	-	-	23.8	47.9	21.2	6.4
ORCNN [3] †	WACV19	✗	32.9	9.3	34.8	**62.9**	35.1	9.6
AISFormer [19] †	BMVC22	✗	32.3	9.2	35.6	62.5	36.3	9.7
ShapeFormer(Ours)	-	✓	**33.2**	**9.4**	**35.7**	62.7	**36.6**	**9.9**
ResNet-101 Amodal MRCNN [3]	WACV19	✗	29.4	-	35.6	-	-	-
ORCNN [3]	WACV19	✗	30.0	-	30.1	-	-	-
ORCNN [3] †	WACV19	✗	34.0	9.3	36.5	64.5	37.2	10.0
AISFormer [19] †	BMVC22	✗	33.7	9.3	37.3	64.7	38.6	10.3
ShapeFormer(Ours)	-	✓	**34.7**	**9.7**	**37.8**	**65.1**	**39.4**	**10.4**

instances and 8,279 testing instances. COCOA-cls is proposed to capture 80 categories object category in COCOA, however, it has much fewer annotation (6,763 training instances and 3,799 testing instances). D2SA is an AIS dataset with 60 categories of instances related to supermarket items with 13,066 training instances and 15,654 testing instances.

Metrics: Following existing AIS approaches [3], [19], [23], we adopt mean average precision (AP) and mean average recall (AR). To evaluate our Cat-SP retriever, we adopt Intersection over Union (IoU) metric between retrieved shape priors and ground-truth shape.

Implementation Details: We implement our ShapeFormer based on Detectron2 [21]. For the KINS dataset, we use an SGD optimizer [16] with a learning rate of 0.0025 and a batch size of 1 on 48000 iterations. For D2SA datasets, we also train with an SGD optimizer but with a learning rate of 0.005 and a batch size of 2 on 70000 iterations. For COCOA and COCOA-cls, we train on 10000 iterations with the learning rate of 0.0005 and a batch size of 2. All experiments have been conducted using an Intel(R) Core(TM) i9-10980XE 3.00GHz CPU and a Quadro RTX 8000 GPU.

B. Performance Comparison

*1) **Quantitative Results and Comparison**:* In the following tables, on each backbone, the best scores are in **bold** and the second best scores are in underlines. **KINS.** Table II presents a comparison between ShapeFormer and SOTA AIS methods on the KINS dataset. ShapeFormer demonstrates consistent improvements across various backbone architectures, including ResNet-50 [5], ResNet-101 [5] and RegNet [17]. Specifically, when compared to methods utilizing ResNet-50 as the backbone, our method outperforms both SOTA shape-based methods (e.g., and VRSP-Net [23] by 1.4 visible AP and 2.0 amodal AP) and non-shape-based methods (e.g., AISFormer [19] by 1.6 visible AP and 0.3 amodal AP), respectively. When ResNet-101 is utilized as the backbone, our method achieves a larger margins of improvement over AISFormer, outperforming it by 1.7 in terms of visible AP and 0.6 in terms of amodal AP. Furthermore, compared to APSNet [12] and AISFormer [19] on the RegNet backbone, our approach achieves SOTA performance by surpassing them in visible AP by 1.8 and amodal AP by 0.5.

TABLE IV: *Performance comparison on D2SA test set* with ResNet-50 as backbone. † indicates our reproduced results.

Methods	Venue	Shape Prior	Visible		Amodal			
			$AP\uparrow$	$AR\uparrow$	$AP\uparrow$	$AP_{50}\uparrow$	$AP_{75}\uparrow$	$AR\uparrow$
Mask R-CNN [4]	ICCV17	✗	68.98	70.11	63.57	83.85	68.02	65.18
ORCNN [3]	WACV19	✗	69.67	70.46	64.22	83.55	69.12	65.25
ASN [14] †	CVPR19	✗	-	-	63.94	84.35	69.57	65.20
BCNet [8] †	CVPR21	✗	-	-	65.97	84.23	72.74	66.90
AISFormer [19]	BMVC22	✗	71.60	71.59	67.22	84.05	72.87	68.13
VRSP-Net [23]	AAAI21	✓	72.28	71.85	70.27	85.11	75.81	69.17
ShapeFormer(Ours)	-	✓	**73.78**	**73.05**	**71.03**	**86.05**	**76.13**	**69.31**

TABLE V: *Performance comparison on COCOA-cls test set*, ResNet-50 as backbone. † indicates our reproduced results.

Methods	Venue	Shape Prior	Visible		Amodal			
			$AP\uparrow$	$AR\uparrow$	$AP\uparrow$	$AP_{50}\uparrow$	$AP_{75}\uparrow$	$AR\uparrow$
Mask R-CNN [4]	ICCV17	✗	30.10	31.52	33.67	56.50	35.78	34.18
ORCNN [3]	WACV19	✗	30.80	32.23	28.03	53.68	25.36	29.83
ASN [14] †	CVPR19	✗	-	-	35.33	58.82	37.10	35.50
BCNet [8] †	CVPR21	✗	-	-	35.14	**58.84**	36.65	35.80
AISFormer [19]	BMVC22	✗	34.00	**36.44**	35.77	57.95	38.23	36.71
VRSP-Net [23]	AAAI21	✓	34.58	36.42	35.41	56.03	38.67	37.11
ShapeFormer(Ours)	-	✓	**35.01**	36.42	**35.83**	58.82	**38.85**	**37.13**

COCOA. We also conduct experiments on COCOA test set in Table III. Our ShapeFormer achieves best performance on most metrics across backbones. ShapeFormer surpasses the SOTA AISFormer by 0.1 in amodal AP and 0.9 in visible AP when evaluated on ResNet 50. Additionally, it achieves a 0.5 improvement in amodal AP and a perfect 1.0 in visible AP when assessed on ResNet 101.

D2SA. Table IV further validates our approach on D2SA dataset. We achieve best results across all metrics. Specifically, we gains 1.5 on visible AP and 0.76 on amodal AP in comparison with the second best method, i.e. VRSP-Net.

COCOA-cls. Table V shows our results on COCOA-cls dataset. Our ShapeFormer outperform across other methods on visible and amodal AP metrics and show competitive results on AR metrics.

In summary, our experimental results across datasets demonstrate that our approach, which incorporates visible-to-amodal modeling with shape prior, delivers comprehensive and competitive performance in both visible and amodal AP.

*2) **Qualitative results and comparison**:* Fig. 7 illustrates the qualitative output of ShapeFormer. To explain where the network learns, we also visibly include attention maps cor-

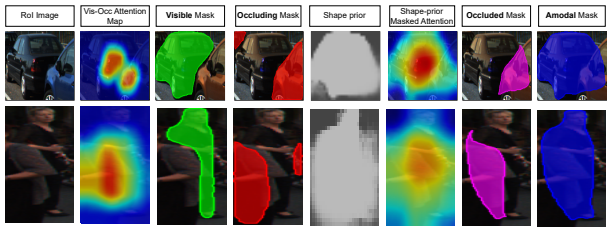


Fig. 7: Qualitative results of ShapeFormer. Left to right: Input RoI, Vis-Occ attention map, Visible masks, Occluding masks, Shape priors, Shape-prior masked attention, Amodal masks, and Occluded masks.

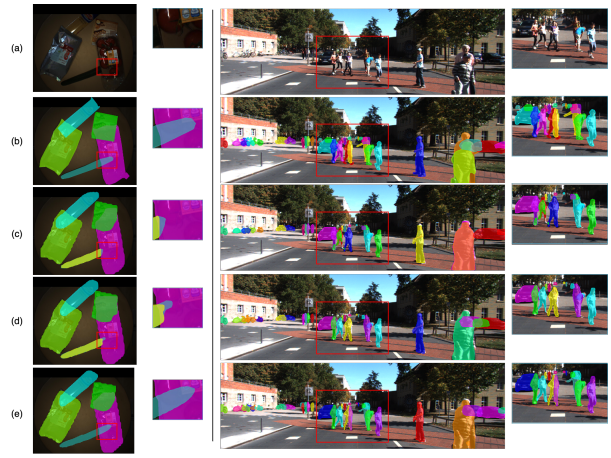


Fig. 8: Quantitative comparison between our ShapeFormer and state-of-the-art methods (e.g. AISFormer [19], VRPS-Net [23]) on amodal segmentation results. From top to bottom, (a) Image, (b) Ground truth, (c) VRSP-Net's predictions, (d) AISFormer's predictions, (e) Our ShapeFormer's prediction. Images are sampled from D2SA (left) and KINS (right) test sets. Best view in zoom and color.

responding to Vis-Occ attention map in Vis-Occ Mask Head module and Shape-prior Masked Attention in SPA Mask Head module. This visualization offers a comprehensive overview of both the output masks and the corresponding attention maps generated during the prediction process of our ShapeFormer model. The results are arranged from left to right, encompassing: input RoIs, Vis-Occ Attention Maps, Visible Masks, Occluding Masks, Shape priors, Shape-prior Masked Attention, Amodal masks, and Occluded masks. Fig. 8 shows qualitative comparison between our ShapeFormer and existing SOTA methods (e.g. AISFormer [19], VRPS-Net [23]). Images are sampled from D2SA and KINS test sets. As can be seen, our ShapeFormer accurately extracts the amodal mask of the occluded object (i.e. the cucumber) (left) and efficiently handles the dense group of pedestrians (right).

C. Ablation Experiments & Analysis

*1) **Effectiveness of Visible-to-Amodal Modeling**:* In Table VI, we assess the efficacy of visible-to-amodal transition compared to bidirectional learning baseline with ResNet-50 backbone. The baseline is implemented as in Fig. 9, which shares the same design with Vis-Occ Mask Head but includes the integration integration of the amodal embedding and amodal mask prediction to enable bidirectional relationship. The result of Vis-Occ Mask Head is obtained by training

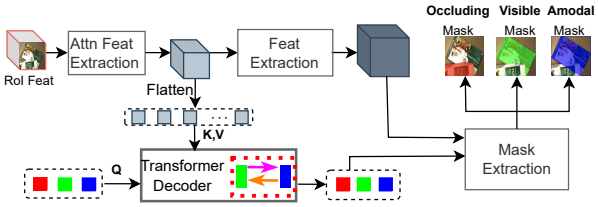


Fig. 9: Baseline with a bidirectional-transition. The baseline implementation shares the same design as the Vis-Occ Mask Head but includes the integration bidirectional relationships.

TABLE VI: The effectiveness of *visible-to-amodal modeling* without shape prior (w/o SP) compared to bidirectional modeling baseline (Fig. 9).

Models	D2SA				KINS			
	Visible		Amodal		Visible		Amodal	
	$AP\uparrow$	$AR\uparrow$	$AP\uparrow$	$AR\uparrow$	$AP\uparrow$	$AR\uparrow$	$AP\uparrow$	$AR\uparrow$
Bidirectional-baseline	71.6	71.6	67.2	68.1	29.7	20.0	33.5	21.1
Visible only	73.7	72.9	-	-	31.6	21.0	-	-
Visible-to-Amodal (w/o SP)	73.9	72.9	69.4	68.9	31.5	21.1	33.7	21.1

TABLE VII: Ablation study on *IoU performance* with *various configurations of the Cat-SP Retriever*, namely using augmented data for training (Aug.), using object category (Cat.)

Cat.	Aug.	KINS	D2SA	COCOA-cls	COCOA
✗	✗	93.34	93.42	85.24	85.95
✗	✓	94.08	94.51	**86.25**	**86.62**
✓	✗	94.01	94.32	85.17	-
✓	✓	**94.14**	**95.31**	86.12	-

it separately from ShapeFormer, showing that dropping the amodal-to-visible relation in the baseline results in better visible segmentation. Moreover, the performance of ShapeFormer-w/o SP (we remove the use of shape prior for fair comparison with the baseline) illustrates that our design of ShapeFormer does not affect the visible result produced by the Vis-Occ Mask Head, hence results in the enhanced visible-to-amodal feature and final amodal segmentation result.

*2) **Effectiveness of Cat-SP Retriever**:* In Table VII, we examine the benefits of category-specific input for retrieving the shape prior and generating augmented data to enhance training generalization. Our findings indicate that using augmented data during training improves IoU scores across all datasets: 0.74 on KINS, 1.15 on D2SA, 1.01 on COCOA-cls, and 0.67 on COCOA. Regarding the use of category information, we observe a performance improvement when incorporating category information for KINS (0.67 IoU) and D2SA (0.9 IoU). Using category information does not result in performance gains for COCOA-cls. This could be attributed to the variation of shapes within same category in COCOA-cls. In the case of COCOA, where we lack category annotations, we denote the corresponding values as (–). In the final row of the table, we incorporate using both object category and augmented data into the training process, which yields the best performance on KINS and D2SA, and the second-best performance on COCOA-cls.

*3) **Effectiveness of shape-prior masked attention**:* Table VIII showcases the impact of shape-prior masked attention in the Amodal Transformer Decoder $\mathcal{D}_a$. Herein, we evaluate

TABLE VIII: Impact of our *shape-prior masked attention* in amodal transformer decoder $\mathcal{D}_a$.

Datasets	Shape-prior masked attention	$AP\uparrow$	$AP_{50}\uparrow$	$AP_{75}\uparrow$	$AR\uparrow$
KINS	✗	33.72	57.80	34.74	21.10
	✓	**34.05**	**58.61**	**35.74**	**22.04**
D2SA	✗	69.44	84.25	74.87	68.92
	✓	**71.03**	**86.05**	**76.13**	**69.31**
COCOA	✗	34.92	62.21	35.47	9.60
	✓	**35.71**	**62.71**	**36.64**	**9.90**
COCOA -cls	✗	35.78	**59.25**	36.79	37.05
	✓	**35.83**	58.82	**38.85**	**37.13**

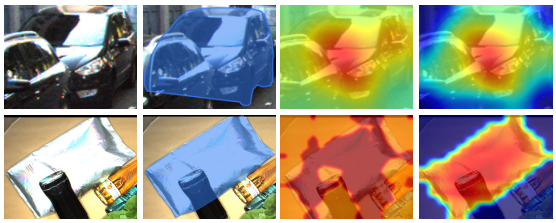


Fig. 10: A visual comparison between using cross attention [20] and **our shape-prior masked attention**. From left to right, (a) RoI image, (b) Amodal ground truth mask, (c) Cross attention's attention map, (d) Our shape-prior masked attention's attention map.

the amodal segmentation performance utilizing the ResNet-50 backbone, where we compare two scenarios: one without shape-prior masked attention (marked as ✗), and the other with shape-prior masked attention (marked as ✓). The results demonstrate that incorporating shape-prior masked attention yields consistent improvements across multiple datasets. These findings highlight the importance of shape-prior masked attention and prior knowledge in enhancing the performance of the Amodal Transformer Decoder $\mathcal{D}_a$ for amodal segmentation. Fig. 10 visualizes the shape-prior masked attention of the Amodal Transformer Decoder on RoIs. The attention maps are well-constrained to the object shape owing to the shape-prior masked attention. Moreover, we can see that the decoder typically attends to the visible parts of objects that are similar to the occluded regions when predicting the amodal mask.

V. Conclusion

In conclusion, our proposed ShapeFormer introduces a novel approach to Amodal Instance Segmentation (AIS) by prioritizing the visible-to-amodal transition over the traditional bidirectional method. We address the issue of compromised visible features and present a structured architecture that connects visible and amodal components through shape prior modeling. The transformer-based framework of ShapeFormer leverages advancements in AIS, incorporating a category-specific vector quantized autoencoder for shape prior knowledge. By first predicting visible segmentation while acknowledging occluded objects, and subsequently utilizing shape priors during amodal mask prediction, our model outperforms previous SOTA on AIS across KINS, COCOA, D2SA, COCO-cls datasets. We hope our work sheds light on future research in AIS aiming to further expand the amodal understanding domain.

Acknowledgments: This work is sponsored by the National Science Foundation (NSF) under Award No OIA-1946391 RII Track-1, NSF 2223793 EFRI BRAID, NSF 2119691 AI SUSTEIN, NSF 2236302, and the National Institutes of Health (NIH) 1R01CA277739-01.

References

[1] S. Back, J. Lee, T. Kim, S. Noh, R. Kang, S. Bak, and K. Lee. Unseen object amodal instance segmentation via hierarchical occlusion modeling. In *ICRA*, pages 5085–5092. IEEE, 2022.

[2] J. Duncan. Selective attention and the organization of visual information. *Journal of experimental psychology: General*, 113(4):501, 1984.

[3] P. Follmann, R. König, P. Härtinger, M. Klostermann, and T. Böttger. Learning to see the invisible: End-to-end trainable amodal instance segmentation. In *WACV*, pages 1328–1336. IEEE, 2019.

[4] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *ICCV*, pages 2961–2969, 2017.

[5] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[6] W.-D. Jang, D. Wei, X. Zhang, B. Leahy, H. Yang, J. Tompkin, D. Ben-Yosef, D. Needleman, and H. Pfister. Learning vector quantized shape code for amodal blastomere instance segmentation. *arXiv preprint arXiv:2012.00985*, 2020.

[7] L. Ke, M. Danelljan, X. Li, Y.-W. Tai, C.-K. Tang, and F. Yu. Mask transfiner for high-quality instance segmentation. In *CVPR*, pages 4412–4421, 2022.

[8] L. Ke, Y.-W. Tai, and C.-K. Tang. Deep occlusion-aware instance segmentation with overlapping bilayers. In *CVPR*, pages 4019–4028, 2021.

[9] P. J. Kellman and T. F. Shipley. A theory of visual interpolation in object perception. *Cognitive psychology*, 23(2):141–221, 1991.

[10] K. Li and J. Malik. Amodal instance segmentation. In *ECCV*, pages 677–693. Springer, 2016.

[11] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. Springer, 2014.

[12] R. Mohan and A. Valada. Amodal panoptic segmentation. In *CVPR*, pages 21023–21032, 2022.

[13] K. Nguyen and S. Todorovic. A weakly supervised amodal segmenter with boundary uncertainty estimation. In *ICCV*, pages 7396–7405, 2021.

[14] L. Qi, L. Jiang, S. Liu, X. Shen, and J. Jia. Amodal instance segmentation with kins dataset. In *CVPR*, pages 3014–3023, 2019.

[15] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241. Springer, 2015.

[16] S. Ruder. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*, 2016.

[17] N. Schneider, F. Piewak, C. Stiller, and U. Franke. Regnet: Multimodal sensor registration using deep neural networks. In *2017 IEEE intelligent vehicles symposium (IV)*, pages 1803–1810. IEEE, 2017.

[18] Z. Tian, C. Shen, H. Chen, and T. He. Fcos: Fully convolutional one-stage object detection. In *ICCV*, pages 9627–9636, 2019.

[19] M. Tran, K. Vo, K. Yamazaki, A. Fernandes, M. Kidd, and N. Le. Aisformer: Amodal instance segmentation with transformer. *arXiv preprint arXiv:2210.06323*, 2022.

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *NeurIPS*, 30, 2017.

[21] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019.

[22] Y. Xiao, Y. Xu, Z. Zhong, W. Luo, J. Li, and S. Gao. Amodal segmentation based on visible region segmentation and shape prior. *arXiv preprint arXiv:2012.05598*, 2020.

[23] Y. Xiao, Y. Xu, Z. Zhong, W. Luo, J. Li, and S. Gao. Amodal segmentation based on visible region segmentation and shape prior. In *AAAI*, volume 35, pages 2995–3003, 2021.

[24] J. Yao, Y. Hong, C. Wang, T. Xiao, T. He, F. Locatello, D. P. Wipf, Y. Fu, and Z. Zhang. Self-supervised amodal video object segmentation. *NeurIPS*, 35:6278–6291, 2022.

[25] X. Zhan, X. Pan, B. Dai, Z. Liu, D. Lin, and C. C. Loy. Self-supervised scene de-occlusion. In *CVPR*, pages 3784–3792, 2020.

[26] Y. Zhu, Y. Tian, D. Metaxas, and P. Dollár. Semantic amodal segmentation. In *CVPR*, pages 1464–1472, 2017.