MISA: MIning Saliency-Aware Semantic Prior for Box Supervised Instance Segmentation

Hao Zhu[1,2], **Yan Zhu**[1,2], **Jiayu Xiao**[1,2],
Yike Ma[1], **Yucheng Zhang**[1], **Jintao Li**[1] and **Feng Dai**[1*]
[1]Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
[2]University of Chinese Academy of Sciences, Beijing, China
{zhuhao22z, zhuyan22s, ykma, zhangyucheng, jtli, fdai}@ict.ac.cn, {jiayu.xiao}@vipl.ict.ac.cn

Abstract

Box supervised instance segmentation (BSIS) aims to achieve an effective trade-off between annotation costs and model performance by solely relying on bounding box annotations during training process. However, we observe that BSIS model is bottlenecked by the intricate objective under limited guidance, and tends to sacrifice segmentation capability in order to effectively recognize multiple instances. To boost the BSIS model's perceptual ability for object shape and contour, we introduce MISA, that is, MIning Saliency-Aware semantic prior from a well-optimized box supervised semantic segmentation (BSSS) network, and incorporating cross-model guidance into the learning process of BSIS. Specifically, we first design a Frequency-Space Distillation (FSD) module to extract assorted salient prior knowledge from BSSS model, and perform cross-model alignment for transfering the prior to BSIS model. Furthermore, we introduce Semantic-Enhanced Pairwise Affinity (SEPA), which borrows the object perceptual ability of BSSS model to emphasize the contribution of salient objects for pairwise affinity, providing more accurate guidance for the BSIS network. Extensive experiments show that our proposed MISA consistently surpasses the existing state-of-the-art methods by a large margin in the BSIS scenario.

1 Introduction

Instance segmentation, which aims to locate different objects and assign them pixel-wise masks, is a fundamental task in vision. The efficacy of instance segmentation methods operating in various paradigms (*e.g*, *query-based* [Wang *et al.*, 2020b], *top-down* [Chen *et al.*, 2019a], and *bottom-up* [Gao *et al.*, 2019]) has experienced a swift evolution in the vision community. Nonetheless, these methods heavily rely on labour intensive manual mask annotations, which incur substantial time costs. In contrast, coarse object annotations significantly reduce annotation costs, *e.g*, the average time cost of point-level and box-level annotations for each instance is 9s and 7s respectively, which is much lower than that of mask-level annotation (79.2s) [Cheng *et al.*, 2022b].

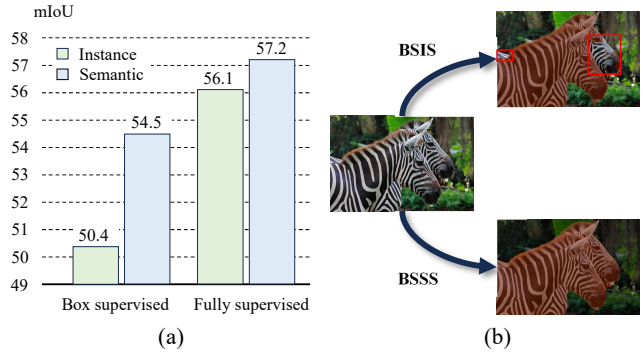

Figure 1: Segmentation performance when instance discrimination is ignored. (a) The mIoU between semantic segmentation and instance segmentation under both box supervised and fully supervised settings. Instance segmentation demonstrates a lower mIoU, especially under weak supervision. (b) Visualization results of BSIS and BSSS. BSIS exhibits a remarkable decrease in mask accuracy due to the complicated objective.

For a trade-off between annotation costs and segmentation performance, our work focuses on box-supervised instance segmentation (BSIS). Given the limited information from bounding boxes, most current methods model the pairwise affinity at the image level to obtain extra prior knowledge [Hsu *et al.*, 2019; Tian *et al.*, 2021], and propagate the pairwise affinity globally and locally [Li *et al.*, 2022b; Li *et al.*, 2023b]. Although these methods mitigate the challenges arising from the lack of mask supervision, there still exists a significant gap compared to fully supervised approaches. The main reason is that the objective of instance segmentation task is inherently more complicated, the network is supposed to not only segment objects but also perform instance discrimination. Consequently, the model may sacrifice the perceptual ability for object shape and contour in order to recognize multiple instances, leading to suboptimal segmentation results. This issue becomes even more serious due to the lack of accurate mask supervision. An intuitive comparison is shown in Figure 1, the performance of BSIS network is obviously inferior to box-supervised semantic segmentation (BSSS) when neglecting instance discrim-

*Corresponding author.

ination. By straightforwardly focusing on the segmentation task, the model demonstrates better segmentation capability and may provide valuable guidance for addressing complex segmentation tasks (*i.e*, BSIS) in weakly supervised setting.

In this work, we present MISA to address box supervised instance segmentation task. We MIne the Saliency-Aware semantic prior knowledge within dense feature of a well-optimized BSSS model, and perform cross-model guidance onto the query-based BSIS model to facilitate its representation learning process, thereby enhancing model's segmentation capability. The proposed MISA constitutes of two components: (1) a Frequency-Space Distillation (FSD) module, and (2) Semantic Enhanced Pairwise Affinity (SEPA).

As for the Frequency-Space Distillation (FSD) module, we extract assorted salient semantic prior knowledge from BSSS model and perform cross-model alignment aiming to transfer the salient prior to BSIS model. Specifically, we decompose the dense feature representations of segmentation networks into high-frequency and low-frequency components respectively. Then we acquire the high-frequency features of object regions and the low-frequency features of background regions according to the ground-truth bounding boxes. Next, we align the aforementioned features between BSIS and BSSS networks to facilitate the BSIS model in capturing essential semantic prior by minimizing feature level Mean Squared (L2) loss. Rich textural knowledge in the high-frequency component of the object features drives the network towards accurately segmenting edges, while the low-frequency component of background features implies global structural information, which prevents false positive predictions.

As for the Semantic Enhanced Pairwise Affinity (SEPA), we borrow the object perceptual ability of the BSSS network to emphasize the contribution of salient foreground objects for pairwise affinity, providing more accurate guidance for the BSIS network. In particular, inspired by class activation maps (CAMs) [Zhou *et al.*, 2016], we extract the classifier weights corresponding to different object categories and utilize them to reweight the semantic features, highlighting the contributions of different channels in discriminating foreground objects. The modulated saliency-aware features then cooperated with the low-level cues to model the pairwise affinity, which will propagate on the predicted mask to effectively guide the BSIS model.

Overall, our contributions can be summarized as follows:

- We propose MISA to guide the query-based box supervised instance segmentation during learning process via MIning Saliency-Aware semantic prior from BSSS model. To the best of our knowledge, this is the first work to introduce cross-model guidance to boost the performance of BSIS.
- Two novel technologies are presented: (1) a Frequency-Space Distillation (FSD) module, and (2) Semantic Enhanced Pairwise Affinity (SEPA). Both technologies explore the salient semantic prior within the dense feature of the BSSS model from different perspectives, so as to provide valuable guidance that enhances the segmentation capability of BSIS model.
- Extensive experimental results on various benchmarks demonstrate that our proposed method remarkably outperforms existing state-of-the-art BSIS methods. Our work narrows down the gap in performance between full mask and box supervised instance segmentation.

2 Related Works

2.1 Instance Segmentation

Instance segmentation aims to segment and recognize each foreground object from an image. Existing instance segmentation methods can be roughly categorized into query-based, top-down and bottom-up paradigms [Wang *et al.*, 2022]. Top-down approaches [He *et al.*, 2017; Chen *et al.*, 2019a] detect target objects and further utilize semantic segmentation algorithms to achieve pixel-level results. Bottom-up methods [Gao *et al.*, 2019] accomplish segmentation by clustering similar pixels and grouping them into distinct objects. Recently, query-based methods [Wang *et al.*, 2020b; Cheng *et al.*, 2022a] have emerged as state-of-the-art paradigm, which introduce a set of learnable query vectors to decode predicted masks from dense feature representations. Our work builds upon the query-based paradigm.

2.2 Box Supervised Segmentation

Box supervised segmentation aims to accomplish semantic or instance segmentation tasks using only bounding box annotations.It has gained growing attention due to its balance between segmentation performance and annotation costs [Shen *et al.*, 2023]. For box supervised semantic segmentation (BSSS), Box2Seg [Kulharia *et al.*, 2020] utilizes GrabCut [Rother *et al.*, 2004] to refine CAMs [Zhou *et al.*, 2016] and combines those with attention maps to jointly guide the segmentation network. BAP [Oh *et al.*, 2021] aims to acquire high-quality pseudo masks by excluding background regions from the bounding boxes. For box supervised instance segmentation (BSIS), BoxInst [Tian *et al.*, 2021] assumes that adjacent pixels with similar colors are likely belong to the same instance. Recently, several approaches [Li *et al.*, 2023b; Li *et al.*, 2022b] utilize low-level image cues to model pairwise affinity, and propagate it on the predicted masks to generate pseudo labels. In our work, we focus on mining salient semantic prior from a well-optimized BSSS network to boost the segmentation ability of BSIS network.

2.3 Knowledge Distillation

Knowledge distillation [Hinton *et al.*, 2015] aims to compress the complexity of the model by aligning soft labels between a cumbersome model and a compact model. This paradigm has been widely adopted in object detection and semantic segmentation. SSTKD [Ji *et al.*, 2022] transfers the low-level structural and statistical texture information from teacher to student model individually. SKD [Liu *et al.*, 2019] proposes to capture structured information between pixels by constraining both pairwise similarity and holistic correlations. DeFeat [Guo *et al.*, 2021] highlights the significance of both foreground and background during the learning process, and distilling them separately to yield more prominent results. In contrast, our method focuses on performing a cross-model distillation to boost the segmentation performance of BSIS.

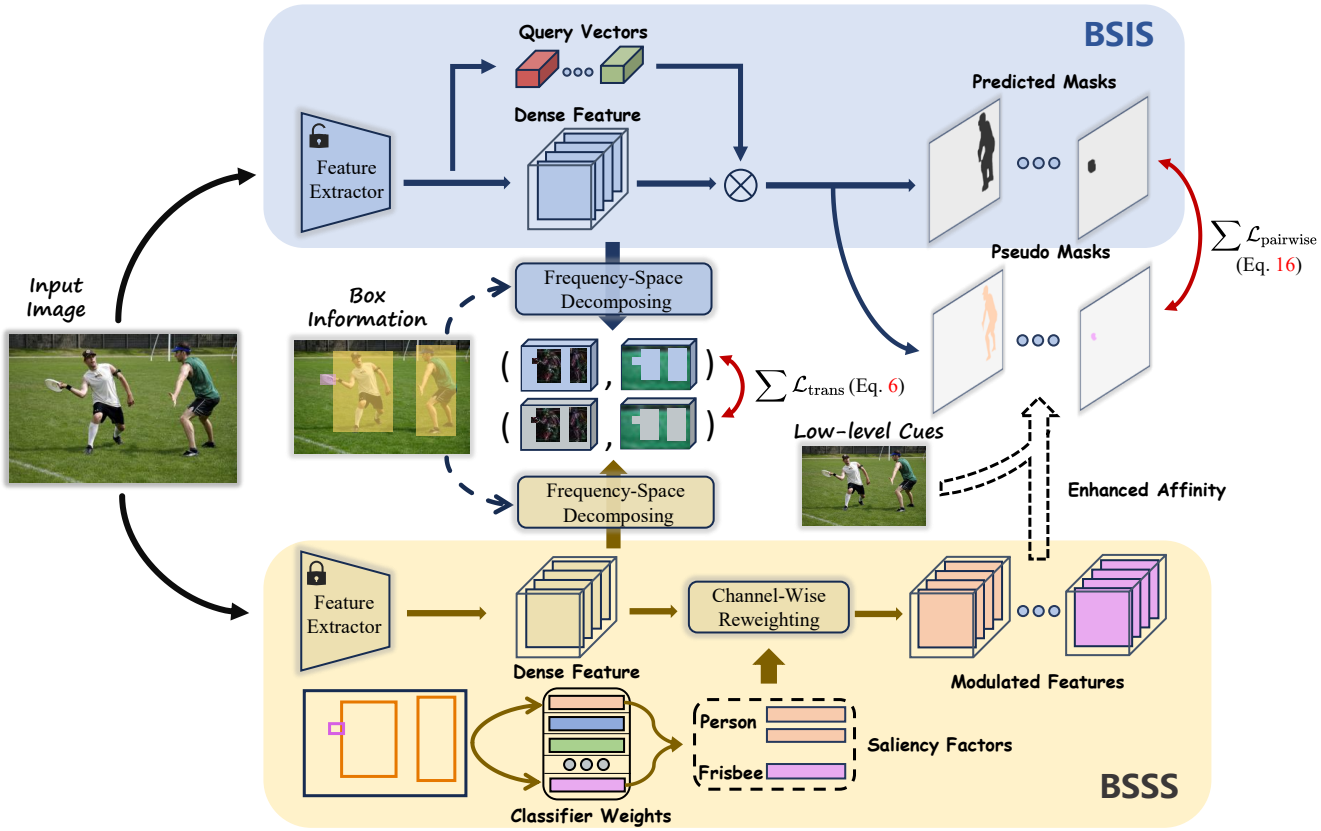

Figure 2: The overview architecture of our proposed MISA. The images are fed into well-optimized BSSS model and BSIS model to extract dense features. Then we decompose the high-frequency features of object regions and the low-frequency features of background regions in Frequency-Space Distillation (FSD) module. These features are aligned to facilitate salient knowledge transfer from BSSS to BSIS models. Subsequently, we reweight the dense feature of BSSS model at channel dimension based on saliency factors. The modulated dense features are incorporated with low-level cues to model the Semantic Enhanced Pairwise Affinity (SEPA) for generating more accurate pseudo masks.

3 Methodology

In this section, we introduce our proposed MIning Saliency-Aware semantic prior (MISA) for box supervised instance segmentation. The overall architecture of MISA is shown in Figure 2. In Section 3.1, we review the semantic segmentation and query-based instance segmentation methods. In Section 3.2, we present the Frequency-Space Distillation (FSD) module to effectively extract and transfer the salient prior knowledge within dense feature of BSSS model. In Section 3.3, we introduce the Semantic Enhanced Pairwise Affinity (SEPA) to emphasize the contribution of salient objects for modeling more accurate pairwise affinity.

3.1 Preliminary

Semantic Segmentation

Most semantic segmentation methods employ an encoder-decoder architecture. A training image is fed into an encoder to obtain low-resolution feature maps, then the feature are continuously upsampled in decoder, generating the semantic dense feature representation $\boldsymbol{S} \in \mathbb{R}^{H \times W \times D}$ to obtain prediction mask:

$$\boldsymbol{M}_{\text{sem}} = \text{softmax}(\boldsymbol{S} * \boldsymbol{K}), \tag{1}$$

where $\boldsymbol{M}_{\text{sem}} \in [0,1]^{H \times W \times C}$ denotes the prediction semantic probability map, D refers to the channels of semantic feature map, C denotes the number of categories. '$*$' represents a convolution operation and $\boldsymbol{K} = \{\boldsymbol{k}^1, ..., \boldsymbol{k}^C\} \in \mathbb{R}^{C \times D}$ refers to a 1×1 conv with C convolutional kernels.

Query-based Instance Segmentation

Query-based instance segmentation methods often feed the extracted image feature into both the query branch and mask feature branch, yielding corresponding instance-aware embeddings $\boldsymbol{Q} = \{\boldsymbol{q}^1, ..., \boldsymbol{q}^N\} \in \mathbb{R}^{N \times D'}$ and dense feature representations $\boldsymbol{I} \in \mathbb{R}^{H \times W \times D'}$ for generating predicted masks:

$$\boldsymbol{M}^n = \text{sigmoid}(\boldsymbol{I} \cdot \boldsymbol{q}^n), \quad n \in \{1, \cdots, N\}, \tag{2}$$

where N denotes the number of positive sample grid points in the training image (*i.e*, SOLOv2), $\boldsymbol{M}^n \in [0,1]^{H \times W \times 1}$ is the prediction mask corresponding to the n^{th} grid point.

3.2 Frequency-Space Distillation

The purpose of proposed Frequency-Space Distillation module is to explore the salient prior knowledge within dense feature of a well-optimized BSSS model, and imbue the BSIS network with extracted valuable salient prior. As shown in

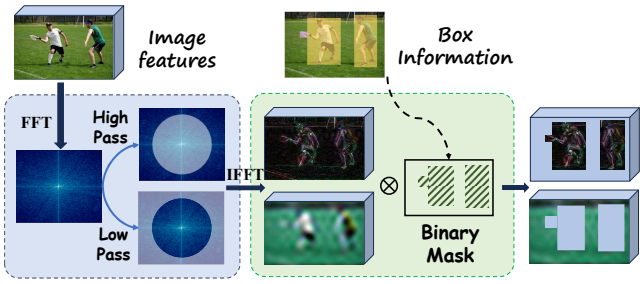

Figure 3: Details of frequency-space decomposing approach. The dense feature is decomposed into high-frequency and low-frequency components through Fourier transformation. Next, the foreground regions in the high-frequency feature and the background regions in the low-frequency feature are extracted based on the ground-truth.

Figure 3, to capture texture knowledge and global structural information within the dense feature representations, we first apply spectral modulation in Fourier domain [Si *et al.*, 2023] to extract the high-frequency and low-frequency components of the dense feature separately:

$$\begin{aligned} \mathcal{T}(\boldsymbol{F}) &= \mathrm{FFT}\left(\boldsymbol{F}\right), \\ \mathcal{T}'(\boldsymbol{F}) &= \mathcal{T}\left(\boldsymbol{F}\right) \cdot \boldsymbol{\alpha}, \\ \boldsymbol{F}' &= \mathrm{IFFT}\left(\mathcal{T}'\left(\boldsymbol{F}\right)\right), \end{aligned} \tag{3}$$

where $\boldsymbol{F}$ is the dense feature presentations (*i.e*, $\boldsymbol{S}$ and $\boldsymbol{I}$), $\mathrm{FFT}(\cdot)$ and $\mathrm{IFFT}(\cdot)$ denote Fourier transform and inverse Fourier transform respectively. $\boldsymbol{\alpha} \in [0,1]^{H \times W}$ is the Fourier mask, such as for high-pass filtering:

$$\boldsymbol{\alpha}(i,j) = \begin{cases} \theta & \text{if } \left(i - \frac{H}{2}\right)^2 + \left(j - \frac{W}{2}\right)^2 \leq t^2, \\ 1 & \text{otherwise,} \end{cases} \tag{4}$$

here θ is the frequency-dependent scaling factor and t denotes the threshold frequency radius.

Texture knowledge in the high-frequency component of the foreground regions is beneficial for recognizing object contours, while global structural information in the low-frequency component of the background regions helps reduce false positive predictions. Based on this, we decompose the high-frequency foreground regions and the low-frequency background regions of the dense feature representations separately. Let $\boldsymbol{B} \in \{0,1\}^{H \times W}$ be a binary mask, where $\boldsymbol{B}(i,j) = 1$ if the location (i,j) belongs to any bounding boxes in a training image and $\boldsymbol{B}(i,j) = 0$ otherwise. The decomposed dense feature representations can be calculated as follows:

$$\begin{aligned} \boldsymbol{F}'_{\text{high}} &= \boldsymbol{F}'_{\text{high}} \cdot \boldsymbol{B}, \\ \boldsymbol{F}'_{\text{low}} &= \boldsymbol{F}'_{\text{low}} \cdot (\boldsymbol{1} - \boldsymbol{B}). \end{aligned} \tag{5}$$

Finally, the decomposed feature representations between BSIS and BSSS network will be aligned via minimizing the Mean Squared (L2) loss to accomplish salient prior transfer:

$$\begin{aligned} \mathcal{L}_{\text{trans}} = &\frac{\lambda_{\text{bg}}}{2N_{\text{bg}}} \sum_{h=1}^{H} \sum_{w=1}^{W} \sum_{d=1}^{D} (\boldsymbol{S}'_{\text{low}}(h,w,d) - \boldsymbol{I}'_{\text{low}}(h,w,d))^2 \\ &+ \frac{\lambda_{\text{fg}}}{2N_{\text{fg}}} \sum_{h=1}^{H} \sum_{w=1}^{W} \sum_{d=1}^{D} (\boldsymbol{S}'_{\text{high}}(h,w,d) - \boldsymbol{I}'_{\text{high}}(h,w,d))^2, \end{aligned} \tag{6}$$

where λ_{bg} and λ_{fg} denote the loss coefficients in background and foreground regions, N_{bg} and N_{fg} refer to the number of pixels in background and foreground regions. We adopt adaptation and projector layers [Yang *et al.*, 2022] to adjust the dense feature of BSIS before frequency-space decomposing.

3.3 Semantic Enhanced Pairwise Affinity

Since ground-truth masks are not available, most previous research leverage the color consistency assumption that pixels with similar colors correspond to shared labels [Lin *et al.*, 2016; Tian *et al.*, 2021]. In this context, some studies [Araslanov and Roth, 2020; Li *et al.*, 2023b] propagate low-level pixel color affinity on predicted masks $\boldsymbol{M}$ to generate pseudo labels $\hat{\boldsymbol{M}}$, formulated as below:

$$\hat{\boldsymbol{M}}^n(i,j) = \boldsymbol{w}_{ij} \sum_{(k,l) \in \mathcal{N}(i,j)} \boldsymbol{A}_{ij,kl} \cdot \boldsymbol{M}(k,l), \tag{7}$$

$$\boldsymbol{A}_{ij,kl} = \exp\left(-\left(\frac{|\boldsymbol{V}(i,j) - \boldsymbol{V}(k,l)|}{\sigma_{\boldsymbol{V}}}\right)^2\right), \tag{8}$$

where $\boldsymbol{A}_{ij,kl}$ is the low-level pairwise affinity between pixels at locations (i,j) and (k,l), $\boldsymbol{V}(i,j)$ and $\boldsymbol{V}(k,l)$ represent RGB vectors, $\sigma_{\boldsymbol{V}}$ denotes the standard deviation of $\boldsymbol{V}$, $\mathcal{N}$ represents the scope of the various receptive fields (*e.g*, 8-way local neighbors [Ru *et al.*, 2022] and global pixels [Li *et al.*, 2023b]). $\boldsymbol{w}_{ij}$ is the normalization coefficient:

$$\boldsymbol{w}_{ij} = \frac{1}{\sum_{(k,l) \in \mathcal{N}(i,j)} \boldsymbol{A}_{ij,kl}} \tag{9}$$

However, in cases where the object and background have similar color intensities, the color consistency assumption become inapplicable [Li *et al.*, 2023a]. To acquire more robust prior knowledge, we employ dense feature $\boldsymbol{S}$, obtained from a well-optimized BSSS model (from Equation 1), to establish high-level feature affinity, denoted as $\boldsymbol{A}^f$:

$$\boldsymbol{A}^f_{ij,kl} = \exp\left(-\left(\frac{|\boldsymbol{S}(i,j) - \boldsymbol{S}(k,l)|}{\sigma_{\boldsymbol{S}}}\right)^2\right). \tag{10}$$

Nevertheless, the formulation of Eq. (10) indicates that each feature vector contributes equally to the pairwise affinity. This lacks concerns on the impact of salient foreground objects, which can incorporate undesired noise from the background regions. To provide more accurate cross-model guidance, we further mine deep semantic prior from the BSSS network and modulate the deep feature to strengthen the awareness of target categories. To specify, inspired by CAMs [Zhou *et al.*, 2016], we first extract the classifier parameters $\boldsymbol{K} \in \mathbb{R}^{C \times D}$, which map D-dimensional dense feature to C-dimensional score map, encapsulating the contributions of different feature channels to different categories.

We view $\boldsymbol{K}$ as the saliency factors and then use it to reweight the channel of dense feature for each foreground category c to effectively reduce noise and amplify the distance between foreground and background vectors:

$$\boldsymbol{S}^c = \boldsymbol{K}^c \cdot \boldsymbol{S}, \tag{11}$$

where $\boldsymbol{K}^c$ represents the c^{th} entry of the $\boldsymbol{K}$. Consequently, the high-level feature pairwise affinity for each foreground category c can be calculated as below:

$$\boldsymbol{A}^c_{ij,kl} = \exp\left(-\left(\frac{|\boldsymbol{S}^c(i,j) - \boldsymbol{S}^c(k,l)|}{\sigma_{\boldsymbol{S}^c}}\right)^2\right). \quad (12)$$

The ultimate Semantic Enhanced Pairwise Affinity $\hat{\boldsymbol{A}}^c$, corresponded to category c, is defined as:

$$\hat{\boldsymbol{A}}^c{}_{ij,kl} = \boldsymbol{A}_{ij,kl} + \boldsymbol{A}^c_{ij,kl}, \quad (13)$$

Therefore, the propagated pseudo labels $\hat{\boldsymbol{M}}^n$ in Equation 7 with ground-truth category c can be adjusted as:

$$\hat{\boldsymbol{M}}^n(i,j) = \hat{\boldsymbol{w}}_{ij} \sum_{(k,l)\in\mathcal{N}(i,j)} \hat{\boldsymbol{A}}^c{}_{ij,kl} \cdot \boldsymbol{M}(k,l). \quad (14)$$

Finally, we adopt the Mean Absolute (L1) loss between predicted masks $\boldsymbol{M}$ and propagated pseudo labels $\hat{\boldsymbol{M}}$ to guide the BSIS network effectively.

$$\mathcal{L}_{\text{pairwise}} = \sum_{n=1}^{N}\sum_{i=1}^{H}\sum_{j=1}^{W}\left|\hat{\boldsymbol{M}}^n(i,j) - \boldsymbol{M}^n(i,j)\right|. \quad (15)$$

4 Experiments

4.1 Datasets

PASCAL VOC 2012

The PASCAL VOC 2012 dataset [Everingham *et al.*, 2010] includes 20 object categories. This dataset is divided into training and validation subsets, with 10,582 images for training and 1,449 images for validation.

COCO

The COCO dataset [Lin *et al.*, 2014] is widely used in image segmentation task. It comprises 80 different object categories and contains a training set of $110k$ images, a validation set of $5k$ images, and a testing set of $20k$ images.

4.2 Evaluation Metrics

We use the standard COCO style metrics to evaluate the proposed method. It contains mask AP (averaged over IoU thresholds), AP_{50}, AP_{75} (mask AP at different IoUs), and AP_S, AP_M, AP_L (mask AP at different scales). All AP is calculated using mask IoU [He *et al.*, 2017].

4.3 Implementation Details

Our proposed method is implemented in Pytorch [Paszke *et al.*, 2017] with `mmcv/mmdet` [Chen *et al.*, 2019b] repository. We choose SOLOv2 [Wang *et al.*, 2020b] and Mask2Former [Cheng *et al.*, 2022a] as the query-based instance segmentation frameworks, and we default to using SOLOv2 unless specified. The different backbones of the model (*i.e*, ResNet [He *et al.*, 2016] and Swin Transformer [Liu *et al.*, 2021]) are pretrained on ImageNet [Russakovsky *et al.*, 2015]. We train our model on 8 GPUs with a batch size of 16, and adopt AdamW as the optimizer with the initial learning rate set to 1.2×10^{-4} and weight decay set to 0.05.

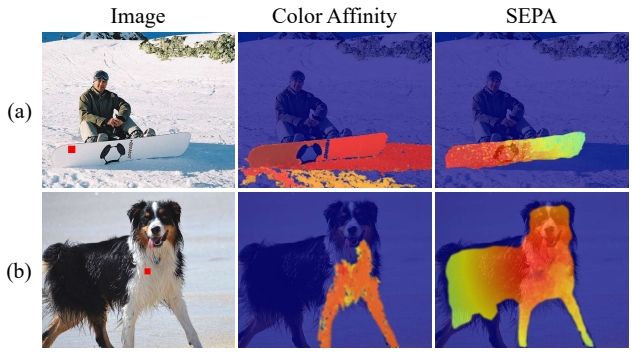

Figure 4: Examples of different pairwise affinity. The propagation of color affinity is erroneously expanded or terminated prematurely. Utilizing SEPA yields more accurate results than color affinity.

We also apply the projection loss term [Tian *et al.*, 2021; Li *et al.*, 2022a] to constrain the predicted mask. Following Apro[Li *et al.*, 2023b], we propagate the Semantic Enhanced Pairwise Affinity (SEPA) locally and globally to obtain pseudo labels. We set $\theta = 0.2$ and $t = 10$ in Equation 4. The affinity balanced coefficient in Equation 13 is set to 0.7. In the objective function of Equation 6, we set $\lambda_{\text{bg}} = 6$ and $\lambda_{\text{fg}} = 10$ by default. The data augmentation strategies follow the default settings recommended in `mmdet`. We adopt the SeMask [Jain *et al.*, 2023] as the segmentation framework of BSSS, and we align the backbone of the BSSS and BSIS models to better comprehend the impact of our approach in comparative experiment. The BSSS model is used solely for auxiliary training and is not utilized during inference.

4.4 Main Results

Results on COCO

We compare MISA with the state-of-the-art BSIS methods on the COCO `test-dev` dataset in Table 1. Our proposed method demonstrates remarkable superiority over other methods across various backbones. Specifically, MISA outperforms BoxInst [Tian *et al.*, 2021] and DiscoBox [Lan *et al.*, 2021] by 2.8% and 2.9% mask AP with SOLOv2 framework and ResNet-50 backbone. It obtains 36.0% mask AP, which is higher than BoxLevelSet and Box2Mask by 2.6% and 1.8% mask AP with ResNet-101 backbone. With more powerful backbone and query-based framework (*i.e*Swin Transformer [Liu *et al.*, 2021] and Mask2Former [Cheng *et al.*, 2022a]), MISA exhibits outstanding performance with 42.0% mask AP, surpassing existing state-of-the-art models such as SIM [Li *et al.*, 2023a], BoxTeacher [Cheng *et al.*, 2023], and Apro [Li *et al.*, 2023b]. This is attributed to MISA's ability to mine and leverage salient semantic prior knowledge as guidance, effectively mitigating the influence of noise and boosting the the perceptual ability for object shape and contour.

Results on PASCAL VOC 2012

We also conduct experiments on PASCAL VOC 2012 `val` dataset to further validate the effectiveness and generalization capability of our method. As shown in Table 2, our proposed MISA demonstrates superior performance compared to recent BSIS methods based on different architectures.

Method	Backbone	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l
Mask-supervised.							
Mask R-CNN [He *et al.*, 2017]	ResNet-101	37.5	59.3	40.2	21.1	39.6	48.3
PolarMask [Xie *et al.*, 2020]	ResNet-101	32.1	50.6	32.8	12.1	33.3	47.1
SOLO [Wang *et al.*, 2020a]	ResNet-101	37.8	59.5	40.4	16.4	40.6	54.2
SOLOv2 [Wang *et al.*, 2020b]	ResNet-101	39.7	60.7	42.9	17.3	42.9	57.4
CondInst [Tian *et al.*, 2020]	ResNet-101	39.1	60.9	42.0	21.5	41.7	50.9
Mask2Former [Cheng *et al.*, 2022a]	ResNet-101	44.2	–	–	23.8	47.7	66.7
Box-supervised.							
BoxInst [Tian *et al.*, 2021]	ResNet-50	32.1	55.1	32.4	15.6	34.3	43.5
DiscoBox [Lan *et al.*, 2021]	ResNet-50	32.0	53.6	32.6	11.7	33.7	48.4
Box2Mask [Li *et al.*, 2022b]	ResNet-50	32.6	55.4	33.4	14.7	35.8	45.9
Apro [Li *et al.*, 2023b]	ResNet-50	33.4	56.1	34.2	15.5	36.3	46.7
MISA	ResNet-50	**34.9**	**56.8**	**35.0**	**16.2**	**37.9**	**48.4**
BoxInst [Tian *et al.*, 2021]	ResNet-101	33.2	56.5	33.6	16.2	35.5	45.1
BoxLevelSet [Li *et al.*, 2022a]	ResNet-101	33.4	56.8	34.1	15.2	36.8	46.8
Box2Mask [Li *et al.*, 2022b]	ResNet-101	34.2	57.8	35.2	16.0	37.7	48.3
Apro [Li *et al.*, 2023b]	ResNet-101	34.6	57.9	35.6	16.3	37.8	48.6
MISA	ResNet-101	**36.0**	**58.7**	**36.9**	**16.8**	**39.3**	**50.9**
Box2Mask*[Li *et al.*, 2022b]	ResNet-101	38.3	65.1	38.8	19.3	41.7	55.2
Apro*[Li *et al.*, 2023b]	ResNet-101	38.5	64.7	**39.4**	20.6	42.4	54.3
MISA*	ResNet-101	**39.7**	**66.3**	39.2	**21.3**	**43.0**	**55.8**
SIM[Li *et al.*, 2023a]	Swin-B	40.2	66.9	41.3	21.1	43.5	56.0
BoxTeacher [Cheng *et al.*, 2023]	Swin-B	40.6	65.0	42.5	23.4	44.9	54.2
Apro* [Li *et al.*, 2023b]	Swin-B	40.9	67.2	41.8	23.6	45.4	55.3
MISA*	Swin-B	**42.0**	**68.1**	**42.7**	**23.8**	**46.1**	**56.8**

Table 1: Comparisons with state-of-the-art methods on the COCO `test-dev` dataset. Methods annotated with "$*$" utilize the Mask2Former framework. Our proposed method consistently outperforms existing methods on various backbones and network architectures.

Method	Backbone	AP	AP_{50}	AP_{75}
BoxInst [Tian *et al.*, 2021]	ResNet-50	34.3	59.1	34.2
SIM [Li *et al.*, 2023a]	ResNet-50	36.7	65.5	35.6
BoxLevelSet [Li *et al.*, 2022a]	ResNet-50	36.3	64.2	35.9
BoxTeacher [Cheng *et al.*, 2023]	ResNet-50	38.6	66.4	38.7
Apro [Li *et al.*, 2023b]	ResNet-50	38.3	65.1	39.4
Apro* [Li *et al.*, 2023b]	ResNet-50	42.3	70.6	44.5
MISA	ResNet-50	**39.9**	**67.4**	**40.6**
MISA*	ResNet-50	**43.5**	**71.8**	**45.2**
BoxInst [Tian *et al.*, 2021]	ResNet-101	36.5	61.4	37.0
SIM [Li *et al.*, 2023a]	ResNet-101	38.6	67.1	38.3
BoxLevelSet [Li *et al.*, 2022a]	ResNet-101	38.3	66.3	38.7
BoxTeacher [Cheng *et al.*, 2023]	ResNet-101	40.2	67.6	40.8
Apro [Li *et al.*, 2023b]	ResNet-101	40.3	67.2	41.9
Apro* [Li *et al.*, 2023b]	ResNet-101	43.6	72.0	45.7
MISA	ResNet-101	**41.5**	**68.4**	**42.2**
MISA*	ResNet-101	**44.7**	**73.5**	**46.1**

Table 2: Performance comparisons with state-of-the-art methods on the PASCAL VOC 2012 `val` dataset. Our method achieves the state-of-the-art mask AP.

Qualitative Results

We show the visualization results of the different pairwise affinity in Figure 4. The examples provide an intuitive illustration of the limitations of color affinity. Specifically, the color affinity in Figure 4(a) erroneously propagates from the snowboard to background due to their similar appearances. In Figure 4(b), the propagation of color affinity is incorrectly interrupted at the black-and-white boundary of the dog's fur. In contrast, our proposed SEPA mitigates the issues arising from misleading color prior and yields more robust propagation results.

Figure 5 presents visualization results that demonstrate the performance of our proposed MISA on the COCO `val` split. It should be emphasized that our method consistently demonstrates robust object segmentation capabilities even in intricate scenarios and multiple-instance images.

4.5 Ablation Study

We conduct ablation studies on PASCAL VOC 2012 `val` split to verify the effectiveness of each module in MISA. We adopt ResNet-50 and SOLOv2 as the baseline, and the mask AP at different IoU thresholds are reported. Additionally, we discuss the sensitivity of hyperparameters and report the relevant experiments in the supplementary materials.

Contributions of FSD and SEPA

Table 3 shows the contributions of the two components introduced in our work (*i.e*, FSD and SEPA). Relying solely on low-level cues to model pairwise affinity [Li *et al.*, 2023b], the method achieves mask AP of only 38.3%. After incorporating the FSD module, the performance is improved to 39.5% mask AP. When propagating the SEPA to obtain pseudo labels, our method reaches 39.3% mask AP. When both the FSD and SEPA modules are employed, the performance is significantly boosted to 39.9% mask AP. This is due to the combination of FSD and SEPA provides a valuable prior guidance for BSIS network.

Distillation Schemes for FSD

We investigate the impact of frequency domain decomposition and foreground-background decoupling schemes on the

Figure 5: Qualitative results of MISA with ResNet-101 backbone and SOLOv2 framework on COCO `val` dataset. The model demonstrates commendable performance in object segmentation without any mask supervision.

Method	FSD	SEPA	AP	AP_{50}	AP_{75}
baseline			38.3	65.1	39.4
MISA	✓		39.5	67.0	40.4
		✓	39.3	66.5	40.1
	✓	✓	**39.9**	**67.4**	**40.6**

Table 3: Effects of two semantic prior modules.

Method	Distillation	Frequency	Space	AP	AP_{50}	AP_{75}
baseline				38.3	65.1	39.4
MISA	✓			38.7	66.0	39.3
	✓	✓		38.8	66.4	39.5
	✓		✓	39.0	66.5	39.9
	✓	✓	✓	**39.5**	**67.0**	**40.4**

Table 4: Effects of distillation schemes.

performance. As shown in Table 4, using only feature level distillation yields a limited 0.4% mask AP improvement. After incorporating the frequency domain decomposition and foreground-background decoupling modules, there are 0.1% and 0.3% mask AP improvements respectively. When the aforementioned modules are combined, the performance is further enhanced to 39.5% mask AP. This suggests that FSD enables BSIS to capture salient prior knowledge from BSSS.

Different Pairwise Affinities for Propagation

We validate the efficacy of SEPA using two propagation modes (*i.e*, PAMR [Araslanov and Roth, 2020] and Apro [Li *et al.*, 2023b]. PAMR employs the 8-way neighbors propagation scope, whereas Apro performs affinity propagation in both global and local contexts. As show in Table 5, the use of semantic feature improves the performance by 0.2% and 0.4% mask AP for PAMR and Apro respectively. After using the saliency factors to modulate the semantic feature, the results are largely improved by 0.6% and 1.0% mask AP. This indicates that SEPA can generate more accurate pseudo labels in different propagation modes.

Method	Color	Feature	Saliency	AP	AP_{50}	AP_{75}
PAMR	✓			37.7	65.1	38.7
	✓	✓		37.9	65.4	39.1
	✓	✓	✓	**38.3**	**65.9**	**39.3**
Apro	✓			38.3	65.1	39.4
	✓	✓		38.7	65.9	39.9
	✓	✓	✓	**39.3**	**66.5**	**40.1**

Table 5: Effects of different pairwise affinities.

5 Conclusion

In this work, we indicate that owing to the intricate objective and limited guidance, the box supervised instance segmentation (BSIS) model often seeks a trade-off by significantly sacrificing its segmentation capability to recognize multiple instances. To this end, we present MIning Saliency-Aware semantic prior (MISA) from a well-optimized box supervised semantic segmentation (BSSS) network, and incorporating cross-model guidance into the training process of BSIS, so as to boost the segmentation ability of BSIS model. The proposed MISA consists of two technologies, (1) a Frequency-Space Distillation (FSD) module, and (2) Semantic Enhanced Pairwise Affinity (SEPA). Both technologies make efforts to mine the salient semantic prior within the dense feature of the BSSS model from different perspectives and provide valuable guidance for BSIS model. The experimental results on different benchmarks demonstrate the superior performance of our proposed method. Taking a broader view, our proposed MISA shows a novel perspective for exploring cross-model guidance to boost the model capability of BSIS.

Acknowledgments

This work is supported by National Key R&D Program of China (2023YFD2000303) and National Natural Science Foundation of China (62372433, 62072438).

Contribution Statement

Hao Zhu and Yan Zhu contributed equally to this paper.

References

[Araslanov and Roth, 2020] Nikita Araslanov and Stefan Roth. Single-stage semantic segmentation from image labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4253–4262, 2020.

[Chen *et al.*, 2019a] Kai Chen, Jiangmiao Pang, Jiaqi Wang, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jianping Shi, Wanli Ouyang, et al. Hybrid task cascade for instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4974–4983, 2019.

[Chen *et al.*, 2019b] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019.

[Cheng *et al.*, 2022a] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299, 2022.

[Cheng *et al.*, 2022b] Bowen Cheng, Omkar Parkhi, and Alexander Kirillov. Pointly-supervised instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2617–2626, 2022.

[Cheng *et al.*, 2023] Tianheng Cheng, Xinggang Wang, Shaoyu Chen, Qian Zhang, and Wenyu Liu. Boxteacher: Exploring high-quality pseudo labels for weakly supervised instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3145–3154, 2023.

[Everingham *et al.*, 2010] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88:303–338, 2010.

[Gao *et al.*, 2019] Naiyu Gao, Yanhu Shan, Yupei Wang, Xin Zhao, Yinan Yu, Ming Yang, and Kaiqi Huang. Ssap: Single-shot instance segmentation with affinity pyramid. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 642–651, 2019.

[Guo *et al.*, 2021] Jianyuan Guo, Kai Han, Yunhe Wang, Han Wu, Xinghao Chen, Chunjing Xu, and Chang Xu. Distilling object detectors via decoupled features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2154–2164, 2021.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[He *et al.*, 2017] 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.

[Hinton *et al.*, 2015] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[Hsu *et al.*, 2019] Cheng-Chun Hsu, Kuang-Jui Hsu, Chung-Chi Tsai, Yen-Yu Lin, and Yung-Yu Chuang. Weakly supervised instance segmentation using the bounding box tightness prior. *Advances in Neural Information Processing Systems*, 32, 2019.

[Jain *et al.*, 2023] Jitesh Jain, Anukriti Singh, Nikita Orlov, Zilong Huang, Jiachen Li, Steven Walton, and Humphrey Shi. Semask: Semantically masked transformers for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 752–761, 2023.

[Ji *et al.*, 2022] Deyi Ji, Haoran Wang, Mingyuan Tao, Jianqiang Huang, Xian-Sheng Hua, and Hongtao Lu. Structural and statistical texture knowledge distillation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16876–16885, 2022.

[Kulharia *et al.*, 2020] Viveka Kulharia, Siddhartha Chandra, Amit Agrawal, Philip Torr, and Ambrish Tyagi. Box2seg: Attention weighted loss and discriminative feature learning for weakly supervised segmentation. In *European Conference on Computer Vision*, pages 290–308. Springer, 2020.

[Lan *et al.*, 2021] Shiyi Lan, Zhiding Yu, Christopher Choy, Subhashree Radhakrishnan, Guilin Liu, Yuke Zhu, Larry S Davis, and Anima Anandkumar. Discobox: Weakly supervised instance segmentation and semantic correspondence from box supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3406–3416, 2021.

[Li *et al.*, 2022a] Wentong Li, Wenyu Liu, Jianke Zhu, Miaomiao Cui, Xian-Sheng Hua, and Lei Zhang. Box-supervised instance segmentation with level set evolution. In *European conference on computer vision*, pages 1–18. Springer, 2022.

[Li *et al.*, 2022b] Wentong Li, Wenyu Liu, Jianke Zhu, Miaomiao Cui, Risheng Yu, Xiansheng Hua, and Lei Zhang. Box2mask: Box-supervised instance segmentation via level-set evolution. *arXiv preprint arXiv:2212.01579*, 2022.

[Li *et al.*, 2023a] Ruihuang Li, Chenhang He, Yabin Zhang, Shuai Li, Liyi Chen, and Lei Zhang. Sim: Semantic-aware instance mask generation for box-supervised instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7193–7203, 2023.

[Li *et al.*, 2023b] Wentong Li, Yuqian Yuan, Song Wang, Wenyu Liu, Dongqi Tang, Jian Liu, Jianke Zhu, and Lei

Zhang. Label-efficient segmentation via affinity propagation. *arXiv preprint arXiv:2310.10533*, 2023.

[Lin *et al.*, 2014] 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.

[Lin *et al.*, 2016] Di Lin, Jifeng Dai, Jiaya Jia, Kaiming He, and Jian Sun. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3159–3167, 2016.

[Liu *et al.*, 2019] Yifan Liu, Ke Chen, Chris Liu, Zengchang Qin, Zhenbo Luo, and Jingdong Wang. Structured knowledge distillation for semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2604–2613, 2019.

[Liu *et al.*, 2021] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[Oh *et al.*, 2021] Youngmin Oh, Beomjun Kim, and Bumsub Ham. Background-aware pooling and noise-aware loss for weakly-supervised semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6913–6922, 2021.

[Paszke *et al.*, 2017] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[Rother *et al.*, 2004] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. " grabcut" interactive foreground extraction using iterated graph cuts. *ACM transactions on graphics (TOG)*, 23(3):309–314, 2004.

[Ru *et al.*, 2022] Lixiang Ru, Yibing Zhan, Baosheng Yu, and Bo Du. Learning affinity from attention: End-to-end weakly-supervised semantic segmentation with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16846–16855, 2022.

[Russakovsky *et al.*, 2015] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

[Shen *et al.*, 2023] Wei Shen, Zelin Peng, Xuehui Wang, Huayu Wang, Jiazhong Cen, Dongsheng Jiang, Lingxi Xie, Xiaokang Yang, and Q Tian. A survey on label-efficient deep image segmentation: Bridging the gap between weak supervision and dense prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[Si *et al.*, 2023] Chenyang Si, Ziqi Huang, Yuming Jiang, and Ziwei Liu. Freeu: Free lunch in diffusion u-net. *arXiv preprint arXiv:2309.11497*, 2023.

[Tian *et al.*, 2020] Zhi Tian, Chunhua Shen, and Hao Chen. Conditional convolutions for instance segmentation. In *European Conference on Computer Vision*, pages 282–298. Springer, 2020.

[Tian *et al.*, 2021] Zhi Tian, Chunhua Shen, Xinlong Wang, and Hao Chen. Boxinst: High-performance instance segmentation with box annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5443–5452, 2021.

[Wang *et al.*, 2020a] Xinlong Wang, Tao Kong, Chunhua Shen, Yuning Jiang, and Lei Li. Solo: Segmenting objects by locations. In *European Conference on Computer Vision*, pages 649–665. Springer, 2020.

[Wang *et al.*, 2020b] Xinlong Wang, Rufeng Zhang, Tao Kong, Lei Li, and Chunhua Shen. Solov2: Dynamic and fast instance segmentation. *Advances in Neural Information Processing Systems*, 33:17721–17732, 2020.

[Wang *et al.*, 2022] Wenguan Wang, James Liang, and Dongfang Liu. Learning equivariant segmentation with instance-unique querying. *Advances in Neural Information Processing Systems*, 35:12826–12840, 2022.

[Xie *et al.*, 2020] Enze Xie, Peize Sun, Xiaoge Song, Wenhai Wang, Xuebo Liu, Ding Liang, Chunhua Shen, and Ping Luo. Polarmask: Single shot instance segmentation with polar representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12193–12202, 2020.

[Yang *et al.*, 2022] Zhendong Yang, Zhe Li, Mingqi Shao, Dachuan Shi, Zehuan Yuan, and Chun Yuan. Masked generative distillation. In *European Conference on Computer Vision*, pages 53–69. Springer, 2022.

[Zhou *et al.*, 2016] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2921–2929, 2016.