DiffusionInst: Diffusion Model for Instance Segmentation

Zhangxuan Gu[1] , **Haoxing Chen**[1,2] , **Zhuoer Xu**[1] , **Jun Lan**[1]
Changhua Meng[1] , **Weiqiang Wang**[1]
[1]Tiansuan Lab, Ant Group Inc.
[2]Nanjing University
{guzhangxuan.gzx,xuzhuoer.xze,yelan.lj,changhua.mch,weiqiang.wwq}@antgroup.com
haoxingchen@smail.nju.edu.cn

Abstract

Recently, diffusion frameworks have achieved comparable performance with previous state-of-the-art image generation models. Researchers are curious about its variants in discriminative tasks because of its powerful noise-to-image denoising pipeline. This paper proposes DiffusionInst, a novel framework that represents instances as instance-aware filters and formulates instance segmentation as a noise-to-filter denoising process. The model is trained to reverse the noisy groundtruth without any inductive bias from RPN. During inference, it takes a randomly generated filter as input and outputs mask in one-step or multi-step denoising. Extensive experimental results on COCO and LVIS show that DiffusionInst achieves competitive performance compared to existing instance segmentation models. We hope our work could serve as a simple yet effective baseline, which could inspire designing more efficient diffusion frameworks for challenging discriminative tasks. Our code is available in *https://github.com/chenhaoxing/DiffusionInst*.

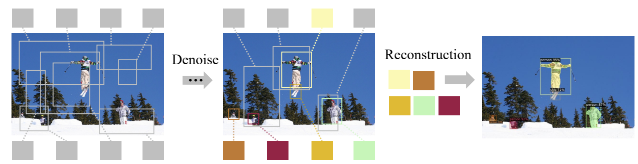

Figure 1: **Diffusion model for instance segmentation.** We propose to regard instance segmentation as a denoising diffusion process from noisy filters to instance masks with a dynamic mask head for mask reconstruction.

1 Introduction

Instance segmentation aims to represent objects with binary masks, which is a finer-grained representation compared to the bounding boxes of object detection. Standard instance segmentation approaches can be divided into two groups, *i.e.,* two-stage[He *et al.*, 2017; Liu *et al.*, 2018; Huang *et al.*, 2019], and single-stage[Chen *et al.*, 2020; Bolya *et al.*, 2019; Wang *et al.*, 2020a; Wang *et al.*, 2020b; Tian *et al.*, 2020]. Two-stage methods first detect objects, then crop their region features with RoI alignment to further classify each pixel. At the same time, the framework of single-stage instance segmentation is usually based on anchors and is thus much simpler. However, they all have dense prediction heads, requiring the non-maximum suppression (NMS) technique during inference.

Recently, QueryInst[Fang *et al.*, 2021] and Mask2Former[Cheng *et al.*, 2022] proposed end-to-end instance segmentation frameworks with the help of learnable queries and bipartite matching. Specifically, they extend DETR[Carion *et al.*, 2020] by feeding the instance-aware RoI features to the mask head for predicting instance masks. Unlike existing anchor-based and anchor-free methods, query-based approaches use random queries to replace the RPN and anchors, reducing the inductive bias in localization instances and improving the segmentation performance by increasing the training difficulty.

Considering that query-based approaches[Cheng *et al.*, 2022; Fang *et al.*, 2021] formulate like a noise-to-mask scheme, we believe they are a special case of diffusion models[Ho *et al.*, 2020; Song *et al.*, 2021; Song and Ermon, 2019]. To be exact, they directly denoise random queries to objects with their decoders only by one forward pass, while diffusion models can additionally perform multi-step denoising gradually during inference.

However, how to adapt the diffusion model in instance segmentation is still an open problem. Recently, DiffusionDet[Chen *et al.*, 2022a] has been proposed to tackle the object detection task by casting detection as a generative task over the space of bounding boxes in the image. At the training stage, it adds Gaussian noise to groundtruth boxes to obtain noisy boxes. Then the RoI features of noisy boxes are fed to the decoder to predict groundtruth boxes as a denoising pipeline. DiffusionDet iteratively generates bounding boxes during inference by reversing the diffusion process.

According to CondInst[Tian *et al.*, 2020], instance masks of one image can be represented by instance-aware filters (vectors) with a common mask feature. Inspired by it, this paper proposes DiffusionInst, a novel instance segmentation framework from a noise-to-filter diffusion view. By reusing the pipeline of DiffusionDet, we have made two changes to the instance segmentation task. Besides bounding boxes, we also generate noisy filters during the diffusion process. Secondly, we introduce a mask branch to obtain multi-scale information from FPN for global mask reconstruction. We show the denoising diffusion process of DiffusionInst in Figure 1.

Besides the ability to perform multi-step inference, another advantage of DiffusionInst compared to query-based models during training is that our noisy generated filters may contain different distribution noises conditioned on the randomly chosen time $t \in \{0, 1, \cdots, T\}$. Moreover, T denoising times/steps can be viewed as T different distribution noises in some cases, which significantly increases the difficulty of model learning and contributes much to model robustness and performance.

In summary, our main contributions are:

- We propose DiffusionInst, the first work of diffusion model for the instance segmentation by regarding it as a generative noise-to-filter diffusion process.
- Instead of predicting local masks, we utilize instance-aware filters and a common mask branch feature to represent and reconstruct global masks.
- Comprehensive experiments are conducted on the COCO and LVIS benchmarks. DiffusionInst achieves competitive results compared with existing approaches, showing the promising future of diffusion models in discriminative tasks.

2 Related Works

2.1 Instance Segmentation

Instance segmentation aims to predict pixel-wise instance masks with class labels for each instance presented in each image. The existing methods can be roughly summarized into some categories. Top-down methods[Li *et al.*, 2017; He *et al.*, 2017; Liu *et al.*, 2018; Chen *et al.*, 2020] detect the object first and then segment the object in the box. Bottom-up methods[Liu *et al.*, 2017; Gao *et al.*, 2019; Newell *et al.*, 2017] learn the pixel-wise embeddings and then cluster them into groups. Direct methods[Wang *et al.*, 2020a; Wang *et al.*, 2020b] perform instance segmentation directly without box detection or embedding learning. More recently, QueryInst[Fang *et al.*, 2021] and Mask2Former[Cheng *et al.*, 2022] proposed to decode random queries to objects for end-to-end instance segmentation frameworks by extending DETR[Carion *et al.*, 2020]. Unlike the above methods, we formulate instance segmentation as a generative denoising process.

2.2 Diffusion Model

Diffusion model[Ho *et al.*, 2020; Song *et al.*, 2021; Song and Ermon, 2019] is a parameterized Markov chain, which starts from the sample in random distribution and reconstructs the data sample via a gradual denoising process. Recently, diffusion models have made remarkable achievements in many fields, e.g., computer vision[Ho *et al.*, 2022; Rombach *et al.*, 2022; Yu *et al.*, 2022; Zhou *et al.*, 2021], language understanding[Li *et al.*, 2022; Austin *et al.*, 2021; Gong *et al.*, 2022], robust learning[Wang *et al.*, 2022; Nie *et al.*, 2022] and temporal data modeling[Park *et al.*, 2022; Kong *et al.*, 2021].

2.3 Diffusion Model for Visual Understanding.

Diffusion models have achieved great success in image generation and synthesis[Dhariwal and Nichol, 2021; Ho *et al.*, 2020; Song and Ermon, 2019]. However, their potential for visual understanding has yet to be fully explored. Recently, Chen *et al.*[Chen *et al.*, 2022b] adopted analog bits based diffusion model[Chen *et al.*, 2022c] to model panoptic masks. Chen *et al.*[Chen *et al.*, 2022a] formulated object detection as a noise-to-box task. In this paper, we further broaden the application of the diffusion model by formalizing instance segmentation as a denoising process. To the best of our knowledge, this is the first work that adopts a diffusion model for the instance segmentation task.

3 Methodology

In this section, we first briefly review the pipeline of diffusion models and DiffusionDet[Chen *et al.*, 2022a]. Then, we introduce different instance mask representation methods. Next, we present the architecture of DiffusionInst and its training and inference process. At last, we provide some discussions about employing the diffusion model in instance segmentation.

3.1 Preliminaries

Diffusion Model: Recent diffusion models usually use two Markov chains: a forward chain that perturbs the image to noise and a reverse chain that refines noise back to the image. Formally, given a data distribution $\mathbf{x}_0 \sim q(\mathbf{x}_0)$, the forward noise perturbing process at time t is defined as $q(\mathbf{x}_t|\mathbf{x}_{t-1})$. It gradually adds Gaussian noise to the data according to a variance schedule $\beta_1, \cdots, \beta_T$:

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}). \tag{1}$$

Given $\mathbf{x}_0$, we can easily obtain a sample of $\mathbf{x}_t$ by sampling a Gaussian vector $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and applying the transformation as follows:

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + (1-\bar{\alpha}_t)\epsilon, \tag{2}$$

where $\bar{\alpha}_t = \prod_{s=0}^{t}(1-\beta_s)$.

During training, a U-Net-like neural network is trained to predict $\mathbf{x}_0$ from $\mathbf{x}_t$ for different $t \in \{1, \cdots, T\}$. While performing inference, we start from a random noise $\mathbf{x}_T$ and iteratively apply the reverse chain to obtain $\mathbf{x}_0$. We refer the readers to [Yang *et al.*, 2022] for more details.

DiffusionDet: It is the first diffusion model in the object detection task. In their setting, data samples are a set of bounding boxes $\mathbf{x}_0 = \mathbf{b}$, where $\mathbf{b} \in \mathcal{R}^{N\times 4}$ is a set of N boxes.

3.2 Mask Representation

Intuitively, instance masks are usually represented by binary figures. However, according to PolarMask[Xie *et al.*, 2020] and BlendMask[Chen *et al.*, 2020], there are various representation methods for an instance mask. For example, PolarMask formulates an instance mask with polar coordinates.

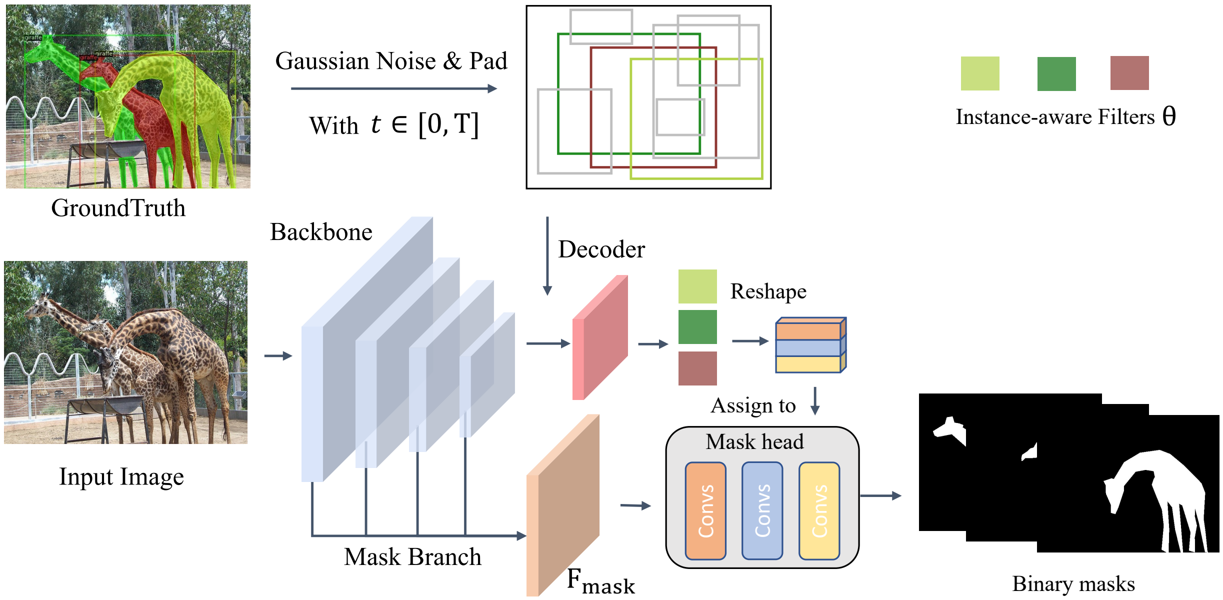

Figure 2: **The overview of our DiffusionInst.** The backbone with FPN extracts multi-scale features from an input image. During training, we add random t step noise to the groundtruth boxes and pad them to predefined numbers. Instance-aware noisy filters are constructed by combining features and noisy boxes. We additionally develop a mask branch to keep multi-scale information in $\mathbf{F}_{mask}$. By applying convolutions whose weights are assigned from noisy filters $\boldsymbol{\theta}$ to $\mathbf{F}_{mask}$, we can obtain instance masks. In inference, the noisy filters are randomly sampled from the Gaussian distribution. Note that input images will only go through the backbone and mask branch once while the multi-step denoising process is performed on the filters.

It represents one mask with a 36-dim vector from the center point by dividing 360° into 36 directions, and each value indicates the half-line length. In some cases, boxes (4-dim vectors) can even be viewed as very coarse masks.

As a result, we use the dynamic mask head to predict instance masks following CondInst[Tian *et al.*, 2020]. Specifically, instance mask can be generated by convolving an instance-agnostic mask feature map $\mathbf{F}_{mask}$ from mask branch and instance-specific filter $\boldsymbol{\theta} \in \mathcal{R}^d$, which is calculated as follows:

$$\mathbf{m} = \phi(\mathbf{F}_{mask}; \boldsymbol{\theta}), \tag{3}$$

where $\mathbf{F}_{mask}$ is multi-scale fused feature map from FPN features $\{P_3, P_4, P_5\}$. $\mathbf{m} \in \mathbb{R}^{H\times W}$ is the predicted binary mask. The ϕ indicates the mask head, which consists of three 1×1 convolutional layers with filter $\boldsymbol{\theta}$ as convolution kernel weights. For example, if the mask head ϕ have three convs with channel $\{8, 8, 1\}$, then the dimension of filters $\boldsymbol{\theta}$ is $d = (8 \times 8 + 8 \times 8 + 8) + (8 + 8 + 1) = 153$.

There are two advantages to using filters to represent instance masks in the diffusion process. One is directly denoise random noise to a whole mask figure is much more complicated than a vector. While DiffusionDet has shown marvelous results in the noise-to-box setting, it is natural to propose a noise-to-filter process with its success. Another benefit is that we replace the widely used box to mask prediction scheme, *i.e.*, decoding RoI features to local masks, with the dynamic mask head for predicting global masks. Unlike bounding boxes, we believe instance masks need larger receptive fields due to the higher requirements on instance edges. The RoI features are usually cropped from downsampled feature maps, in which sizes the details of instance edges are all missing. To this end, representing masks as a combination of filters and the multi-scale feature can solve this weakness.

3.3 DiffusionInst

With this mask representation method, we can regard a data sample in DiffusionInst as a filter $\mathbf{x}_0 = \boldsymbol{\theta}$ for instance segmentation. The overall framework of the DiffusionInst is illustrated in Figure 2. The whole architecture mainly contains the following components: (1) A CNN (*e.g.* ResNet-50[He *et al.*, 2016]), or Swin (*e.g.* Swin-B[Liu *et al.*, 2021]) backbone is utilized to extract compact visual feature representations with FPN[Lin *et al.*, 2017]. (2) A mask branch is utilized to fuse different scale information from FPN, which outputs a mask feature $\mathbf{F}_{mask} \in \mathcal{R}^{c\times H/4\times W/4}$. These two components work like an encoder, and the input image will only pass them once for feature extraction. (3) As for the decoder, we take a set of noisy bounding boxes associated with filters as input to refine boxes and filters as a denoise process. This component can be iteratively called. (4) Finally, we reconstruct the instance mask with the help of mask feature $\mathbf{F}_{mask}$ and denoised filters. Like DiffusionDet, we keep its optimization targets but omit them here for better understanding.

Training: During training, we tend to construct the diffusion process from groundtruth filters to noise filters relying on the corresponding bounding boxes. Then we train the model to reverse this process. Assuming an input image has N instances ($\boldsymbol{\theta}_0^{gt}$) need to be detected. We randomly choose

Methods	Backbone	Sched.	COCO AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Mask RCNN	ResNet-50	1x	34.4	55.1	36.7	18.1	37.5	47.4
Cascade Mask RCNN	ResNet-50	1x	35.9	56.6	38.4	19.4	38.5	49.3
SOLO	ResNet-50	1x	33.9	54.1	35.9	12.6	37.1	51.4
CondInst	ResNet-50	1x	35.9	56.9	38.3	19.1	38.6	46.8
DiffusionInst	ResNet-50	1x	30.4	55.2	29.9	14.4	32.7	45.3
DiffusionInst	ResNet-50	3x	31.3	56.2	31.1	15.6	33.7	46.4
DiffusionInst	ResNet-50	5x	35.1	59.0	36.4	17.6	38.0	51.8
DiffusionInst(4-step)	ResNet-50	5x	35.5	59.7	36.6	18.2	38.2	52.2
Mask RCNN	ResNet-101	3x	38.5	60.0	41.6	19.2	41.6	55.8
Cascade Mask RCNN	ResNet-101	3x	39.6	61.0	42.8	19.6	42.7	56.8
SOLO	ResNet-101	3x	37.8	59.5	40.4	16.4	40.6	54.2
SOLOv2	ResNet-101	3x	39.7	60.7	42.9	17.3	42.9	57.4
CondInst	ResNet-101	3x	39.1	60.9	42.0	21.5	41.7	50.9
DiffusionInst	ResNet-101	5x	36.3	61.5	37.3	18.9	39.6	52.9
DiffusionInst(4-step)	ResNet-101	5x	36.5	62.0	37.6	19.0	39.6	52.6
Mask RCNN	Swin-S	3x	43.3	67.3	46.6	**28.1**	46.7	58.6
Mask RCNN	Swin-B	3x	43.4	66.8	**46.9**	-	-	-
Cascade Mask RCNN	Swin-T	3x	43.7	66.6	47.3	**28.8**	**48.7**	60.6
DiffusionInst	Swin-B	5x	**44.0**	**69.5**	46.4	24.3	47.6	**64.4**
DiffusionInst(4-step)	Swin-B	5x	**44.2**	**69.8**	**46.7**	24.8	**47.7**	**64.4**

Table 1: **Results (AP%) of instance segmentation on COCO.** We list the performance of existing popular instance segmentation approaches on different backbones. For a fair comparison, models are trained using only COCO training data. Among them, our DiffusionInst achieves competitive performances, especially on large instances. Top-2 results are in bold.

a time t to perturb these groundtruth boxes and filters with Equation 2 and Equation 3. The details for box padding and corruption can be found in DiffusionDet. In conclusion, we can obtain the predicted instance masks as (the denoise process of the decoder is denoted as $f(\boldsymbol{\theta}, t)$):

$$\begin{aligned} \boldsymbol{\theta}_t &= \sqrt{\bar{\alpha}_t}\boldsymbol{\theta}_0^{gt} + (1-\bar{\alpha}_t)\epsilon, \\ \mathbf{m} &= \phi(\mathbf{F}_{mask}; f(\boldsymbol{\theta}_t, t)). \end{aligned} \tag{4}$$

With the dice loss[Milletari *et al.*, 2016] used in CondInst, we can obtain the training objective function as:

$$L_{overall} = L_{det} + \lambda L_{dice}(\mathbf{m}, \mathbf{m}^{gt}), \tag{5}$$

where L_{det} is the training loss of DiffusionDet and λ being 5 in this work is used to balance the two losses.

Inference: The inference pipeline of DiffusionInst is a denoising sampling process from noise to instance filters. Starting from filters $\boldsymbol{\theta}_T$ sampled in Gaussian distribution, the model progressively refines its predictions as follows:

$$\begin{aligned} \boldsymbol{\theta}_0 &= f(\cdots(f(\boldsymbol{\theta}_{T-s}, T-s))) \quad s = \{0, \cdots, T\}, \\ \mathbf{m} &= \phi(\mathbf{F}_{mask}; \boldsymbol{\theta}_0). \end{aligned} \tag{6}$$

3.4 Discussion

Although we have successfully introduced the diffusion model into the instance segmentation task, some aspects still require improvements. The first thing is that our noise-to-filter process still relies on the bounding boxes. In the future, we would like to see whether we can directly train our DiffusionInst with filters added by randomly generated noises. We are now working on one possible solution, *e.g.*, attention mechanism.

Secondly, a more significant performance gain of multi-step denoising is needed compared to the loss in inference speed. More specifically, when performing 4-step denoising, it only improves about 1% AP but is 3x slower than the one-step process. In the future, we would like to know if there is a more efficient and effective way for multi-step denoising.

Thirdly, since diffusion models are naturally proposed to tackle generative tasks, the noise-to-filter process in discriminative tasks needs more accurate instance contexts as the condition. Relying heavily on the instance-aware features, DiffusionInst takes more epochs to get satisfactory performances than standard instance segmentation approaches such as SOLO and Mask RCNN.

4 Experiments

4.1 Datasets

We conducted extensive experiments on two standard instance segmentation datasets: COCO[Lin *et al.*, 2014] and LVIS[Gupta *et al.*, 2019]. For all two datasets, we used the standard mask AP metric[Lin *et al.*, 2014] as the evaluation metric.

COCO. COCO is an 80-category label set with instance-level annotations. Following[Kirillov *et al.*, 2020], we use the COCO train2017 (118K training images) for training, and the ablation study is carried out on the val2017 (5K validation images). We also report our main results on test-dev (20k images) for comparison.

Methods	Backbone	COCO AP	AP_S	AP_M	AP_L
Mask RCNN	ResNet-50	36.8	17.1	38.7	52.1
YoLACT	ResNet-50	28.2	9.2	29.3	44.8
PolarMask	ResNet-50	29.1	12.6	31.8	42.3
SOLOv2	ResNet-50	38.2	16.0	41.2	55.4
CondInst	ResNet-50	37.8	18.2	40.3	52.7
Mask RCNN	ResNet-101	38.3	18.2	40.6	54.1
YoLACT	ResNet-101	31.2	12.1	33.3	47.1
SOLO	ResNet-101	37.8	16.4	40.6	54.2
SOLOv2	ResNet-101	39.7	17.3	42.9	57.4
PolarMask	ResNet-101	32.1	14.7	33.8	45.3
EmbedMask	ResNet-101	37.7	17.9	40.4	53.0
CondInst	ResNet-101	39.1	21.5	41.7	50.9
QueryInst	Swin-L	**48.9**	**30.8**	**52.6**	**68.3**
DiffusionInst	Swin-B	**43.0**	**22.8**	**45.6**	**60.0**

Table 2: **Results (AP%) of instance segmentation on COCO test-dev dataset.** Top-2 results are in bold.

Methods	Backbone	LVIS AP	AP_r	AP_c	AP_f
Mask RCNN	ResNet-50	16.1	0.0	12.0	27.4
+EQL	ResNet-50	18.6	2.1	17.4	27.2
+RFS	ResNet-50	22.2	11.5	21.2	28.0
+EQLv2	ResNet-50	25.5	17.7	24.3	30.2
DiffusionInst	ResNet-50	22.3	14.8	21.4	26.6
Mask RCNN	ResNet-101	21.7	1.6	20.7	**31.7**
+RFS	ResNet-101	25.7	17.5	24.6	30.6
+EQLv2	ResNet-101	27.2	**20.6**	**25.9**	31.4
DiffusionInst	ResNet-101	24.6	17.6	24.0	28.3
Mask RCNN	Swin-T	**28.6**	-	-	-
DiffusionInst	Swin-B	**34.8**	**29.2**	**34.9**	**37.1**

Table 3: **Results (AP%) of instance segmentation on LVIS.** We report the performances of our DiffusionInst (schedule 3x, one-step), as well as three advanced models built on Mask RCNN. Top-2 results are in bold.

LVIS. We also perform experiments on a more challenging LVIS dataset[Gupta *et al.*, 2019]. LVIS is a long-tail instance segmentation dataset containing 1203 categories, having more than 2 million high-quality instance mask annotations. LVIS contains 100k, 19.8k, and 19.8k images for training, validation, and testing, respectively. According to the frequency of occurrence in the training set, the categories are divided into three groups: rare (1-10 images), common (11-100 images), and frequent (>100 images).

4.2 Implement Details

In our experiments, we choose the ResNet-50[He *et al.*, 2016], ResNet-101 and Swin-Base[Liu *et al.*, 2021] with FPN[Lin *et al.*, 2017] as the backbone in the proposed method. The backbones are pretrained on ImageNet. We implement the proposed method with PyTorch[Paszke *et al.*, 2019] and it takes 26 hours to train a DiffusionInst on 8 A100 GPUs with batch size 32. The optimizer of the proposed method is SGD, with a learning rate of 2.5e-5 and a weight decay of 1e-4.

Architectures		AP
# Mask Feature Channel	1	31.6
	4	34.7
	8	**35.1**
	16	35.0
# Mask Head Layer	1	31.0
	2	34.8
	3	**35.1**
	4	**35.1**
# Predefined Filter	100	31.6
	300	34.7
	500	35.1
	1000	**35.2**

Table 4: **Architecture variants of DiffusionInst on COCO.** We evaluate different architecture variants of DiffusionInst from several views, including the number of channels, layers and kernels. Note that we use ResNet-50 as the backbone (schedule 5x, one-step).

4.3 Comparison with State-of-the-art

In this section, we experiment with our DiffusionInst in three datasets: COCO validation, COCO test-dev and LVIS. We also compare our model with popular existing methods such as Mask RCNN, YoLACT, SOLO, CondInst and QueryInst.

COCO validation set. In Table 1, we list five performances of the instance segmentation model with various backbones and schedules. To better understand the results, we highlight the top-2 APs. We can get the following observations from this table: Firstly, as the backbone complexity and capacity increase, the performance gains are enlarged, especially for our DiffusionInst (about 9% AP when changing ResNet-50 to Swin-Base). Our model needs more representative features as the condition for the diffusion process since we do not have inductive bias, while other approaches rely on RPN or anchor boxes by densely label assignments. In other words, instance-aware features are essential for denoise as the condition for filter refinement. Secondly, removing the RPN also leads to slower convergence (usually 5x schedule) since the model has to find instance locations by itself. Thirdly, multi-step denoise has incremental benefits but loses the inference speed. Finally, DiffusionInst performs better on large instances (AP_L) but sometimes misses small ones, which indicates that the diffusion model needs larger receptive fields on features in the instance segmentation.

COCO test-dev set. Note that COCO test-dev only evaluates the top 100 predicted instances. Thus, we only employ 100 predefined filters in DiffusionInst in Table 2. Similar conclusions can be drawn as the COCO validation set, except for the higher performance of QueryInst with Swin-Large as the backbone.

LVIS dataset. This dataset uses the same images with COCO, but pays more attention to long-tail instances. Existing approaches on this dataset main built on Mask RCNN, such as EQL[Tan *et al.*, 2020], RFS[Gupta *et al.*, 2019] and

Figure 3: **Visualization of our DiffusionInst on COCO validation set.** Note that the model is based on ResNet-50 and one-step denoising.

EQLv2[Tan *et al.*, 2021]. Among them, our DiffusionInst achieves the best performance with Swin-Base as the backbone.

4.4 Ablation Studies

We conduct experiments on the COCO validation set with the ResNet-50 backbone, 5x schedule and one-step denoising for the ablation studies as shown in Table 4.

Mask Branch. To enhance the expressiveness of the mask feature, we further explore the channel number of the mask branch. Unlike CondInst, we ignore its relative coordinate map and still get good performance. Among different choices, the 8-channel mask feature achieves 35.1% AP, and extra channels cannot improve performance. We set the channel number of the mask feature to 8 by default.

Mask Head. The dynamic mask head plays a critical role in our method. Thus, we conduct ablation studies to show the impact of parameters in the mask head. As presented in the table, with the number of convolutions in the mask head increasing, the performance improves steadily and achieves the peak of 35.1% AP with three stacked convolutions. More convolutions will not contribute to the final results.

Predefined filter numbers. The number of predefined filters has a similar function to the number of proposals in standard instance segmentation approaches. In this table, we vary filter numbers from 100 to 1000, and we choose to use 500 as a good balance of performance and inference speed except for the results on COCO test-dev set, which is forced to be 100.

4.5 Visualizations

The visualization of the proposed method on the COCO validation dataset is shown in Figure 3. In conclusion, our DiffusionInst can successfully segment instances with an accurate boundary but still miss some instances in bad cases. For example, some persons are ignored when they only occur in a small region of the whole image. Moreover, occluded instances are also easy to be misclassified.

5 Conclusion

We introduced a novel diffusion framework for the instance segmentation task in this work. By regarding instance segmentation as a noise-to-filter process, our DiffusionInst achieves comparable single-model results (*i.e.*, ResNet-50 backbone) on the COCO and LVIS datasets. However, as mentioned in Section 3.4, we also conclude three weaknesses of DiffusionInst, namely relies on the bounding boxes resulting in slightly poor performance on small instances, the unsatisfactory performance gain from multi-step denoising and the longer training time. Except that, we believe the diffusion model is a new possible solution to this task in the future and hope our work could serve as a simple yet effective baseline, which could inspire designing more efficient frameworks and rethinking the learning targets for the challenging instance segmentation task.

References

[Austin *et al.*, 2021] Jacob Austin, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured denoising diffusion models in discrete state-spaces. In *NeurIPS*, pages 17981–17993, 2021.

[Bolya *et al.*, 2019] Daniel Bolya, Chong Zhou, Fanyi Xiao, and Yong Jae Lee. YOLACT: real-time instance segmentation. In *ICCV*, pages 9156–9165, 2019.

[Carion *et al.*, 2020] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, volume 12346, pages 213–229, 2020.

[Chen *et al.*, 2020] Hao Chen, Kunyang Sun, Zhi Tian, Chunhua Shen, Yongming Huang, and Youliang Yan. Blendmask: Top-down meets bottom-up for instance segmentation. In *CVPR*, pages 8570–8578, 2020.

[Chen *et al.*, 2022a] Shoufa Chen, Peize Sun, Yibing Song, and Ping Luo. Diffusiondet: Diffusion model for object detection. *arXiv preprint arXiv:2211.09788*, 2022.

[Chen *et al.*, 2022b] Ting Chen, Lala Li, Saurabh Saxena, Geoffrey E. Hinton, and David J. Fleet. A generalist framework for panoptic segmentation of images and videos. *arXiv preprint arXiv:2210.06366*, 2022.

[Chen *et al.*, 2022c] Ting Chen, Ruixiang Zhang, and Geoffrey Hinton. Analog bits: Generating discrete data using diffusion models with self-conditioning. *arXiv preprint arXiv:2208.04202*, 2022.

[Cheng *et al.*, 2022] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *CVPR*, pages 1280–1289, 2022.

[Dhariwal and Nichol, 2021] Prafulla Dhariwal and Alexander Quinn Nichol. Diffusion models beat gans on image synthesis. In *NeurIPS*, pages 8780–8794, 2021.

[Fang *et al.*, 2021] Yuxin Fang, Shusheng Yang, Xinggang Wang, Yu Li, Chen Fang, Ying Shan, Bin Feng, and Wenyu Liu. Instances as queries. In *ICCV*, pages 6890–6899, 2021.

[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 *ICCV*, pages 642–651, 2019.

[Gong *et al.*, 2022] Shansan Gong, Mukai Li, Jiangtao Feng, Zhiyong Wu, and LingPeng Kong. Diffuseq: Sequence to sequence text generation with diffusion models. *arXiv preprint arXiv:2210.08933*, 2022.

[Gupta *et al.*, 2019] Agrim Gupta, Piotr Dollár, and Ross B. Girshick. LVIS: A dataset for large vocabulary instance segmentation. In *CVPR*, pages 5356–5364, 2019.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[He *et al.*, 2017] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask R-CNN. In *ICCV*, pages 2980–2988, 2017.

[Ho *et al.*, 2020] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020.

[Ho *et al.*, 2022] Jonathan Ho, Chitwan Saharia, William Chan, David J. Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *J. Mach. Learn. Res.*, 23:47:1–47:33, 2022.

[Huang *et al.*, 2019] Zhaojin Huang, Lichao Huang, Yongchao Gong, Chang Huang, and Xinggang Wang. Mask scoring R-CNN. In *CVPR*, pages 6409–6418, 2019.

[Kirillov *et al.*, 2020] Alexander Kirillov, Yuxin Wu, Kaiming He, and Ross B. Girshick. Pointrend: Image segmentation as rendering. In *CVPR*, pages 9796–9805, 2020.

[Kong *et al.*, 2021] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. In *ICLR*, 2021.

[Li *et al.*, 2017] Yi Li, Haozhi Qi, Jifeng Dai, Xiangyang Ji, and Yichen Wei. Fully convolutional instance-aware semantic segmentation. In *CVPR*, pages 4438–4446, 2017.

[Li *et al.*, 2022] Xiang Lisa Li, John Thickstun, Ishaan Gulrajani, Percy Liang, and Tatsunori B Hashimoto. Diffusion-lm improves controllable text generation. *arXiv preprint arXiv:2205.14217*, 2022.

[Lin *et al.*, 2014] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *ECCV*, volume 8693, pages 740–755, 2014.

[Lin *et al.*, 2017] Tsung-Yi Lin, Piotr Dollár, Ross B. Girshick, Kaiming He, Bharath Hariharan, and Serge J. Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017.

[Liu *et al.*, 2017] Shu Liu, Jiaya Jia, Sanja Fidler, and Raquel Urtasun. SGN: sequential grouping networks for instance segmentation. In *ICCV*, pages 3516–3524, 2017.

[Liu *et al.*, 2018] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *CVPR*, pages 8759–8768, 2018.

[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 *ICCV*, pages 9992–10002, 2021.

[Milletari *et al.*, 2016] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. IEEE, 2016.

[Newell *et al.*, 2017] Alejandro Newell, Zhiao Huang, and Jia Deng. Associative embedding: End-to-end learning for joint detection and grouping. In *NeurIPS*, pages 2277–2287, 2017.

[Nie *et al.*, 2022] Weili Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat, and Anima Anandkumar. Diffusion models for adversarial purification. *arXiv preprint arXiv:2205.07460*, 2022.

[Park *et al.*, 2022] Sung Woo Park, Kyungjae Lee, and Junseok Kwon. Neural markov controlled SDE: stochastic optimization for continuous-time data. In *ICLR*, 2022.

[Paszke *et al.*, 2019] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Z. Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, pages 8024–8035, 2019.

[Rombach *et al.*, 2022] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10674–10685, 2022.

[Song and Ermon, 2019] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *NeurIPS*, pages 11895–11907, 2019.

[Song *et al.*, 2021] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2021.

[Tan *et al.*, 2020] Jingru Tan, Changbao Wang, Buyu Li, Quanquan Li, Wanli Ouyang, Changqing Yin, and Junjie Yan. Equalization loss for long-tailed object recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11662–11671, 2020.

[Tan *et al.*, 2021] Jingru Tan, Xin Lu, Gang Zhang, Changqing Yin, and Quanquan Li. Equalization loss v2: A new gradient balance approach for long-tailed object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1685–1694, 2021.

[Tian *et al.*, 2020] Zhi Tian, Chunhua Shen, and Hao Chen. Conditional convolutions for instance segmentation. In *ECCV*, volume 12346, pages 282–298, 2020.

[Wang *et al.*, 2020a] Xinlong Wang, Tao Kong, Chunhua Shen, Yuning Jiang, and Lei Li. SOLO: segmenting objects by locations. In *ECCV*, volume 12363, pages 649–665, 2020.

[Wang *et al.*, 2020b] Xinlong Wang, Rufeng Zhang, Tao Kong, Lei Li, and Chunhua Shen. Solov2: Dynamic and fast instance segmentation. In *NeurIPS*, 2020.

[Wang *et al.*, 2022] Jinyi Wang, Zhaoyang Lyu, Dahua Lin, Bo Dai, and Hongfei Fu. Guided diffusion model for adversarial purification. *arXiv preprint arXiv:2205.14969*, 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 *CVPR*, pages 12190–12199, 2020.

[Yang *et al.*, 2022] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Yingxia Shao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. *arXiv preprint arXiv:2209.00796*, 2022.

[Yu *et al.*, 2022] Peiyu Yu, Sirui Xie, Xiaojian Ma, Baoxiong Jia, Bo Pang, Ruiqi Gao, Yixin Zhu, Song-Chun Zhu, and Ying Nian Wu. Latent diffusion energy-based model for interpretable text modelling. In *ICML*, volume 162, pages 25702–25720, 2022.

[Zhou *et al.*, 2021] Linqi Zhou, Yilun Du, and Jiajun Wu. 3d shape generation and completion through point-voxel diffusion. In *ICCV*, pages 5806–5815, 2021.