DatasetDM: Synthesizing Data with Perception Annotations Using Diffusion Models

Weijia Wu, Yuzhong Zhao, Hao Chen , Yuchao Gu, Rui Zhao, Yefei He, Mike Zheng Shou, Hong Zhou, Chunhua Shen

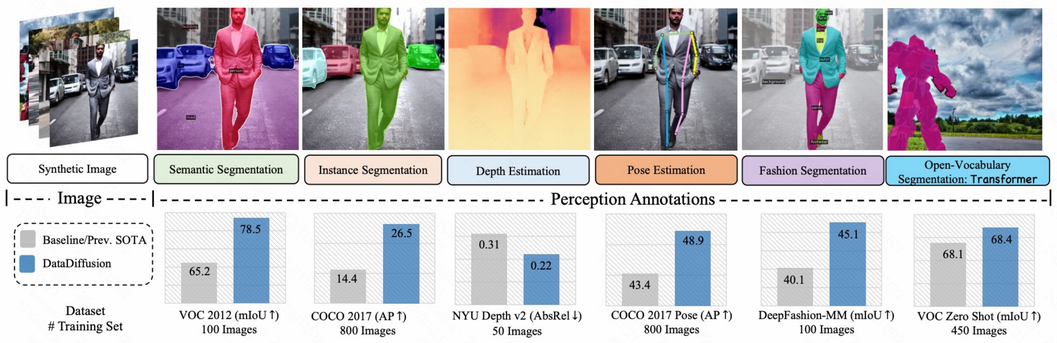
Code and Weight are available at:
https://github.com/showlab/DatasetDM

Motivation

Data Challenge: Existing data-hungry deep-learning models for perception tasks usually require a large amount of data with labor-intensive and expensive pixel-level annotations to achieve significant progress.

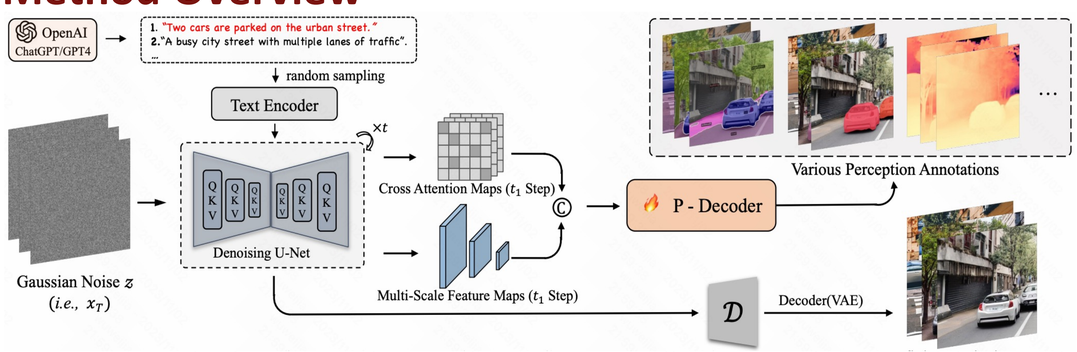
Background: Recent DALL-E , and Stable Diffusion , have shown phenomenal generative semantic and compositional power for image generation with cross attention map.

Insight: *Can we leverage the diffusion models to generate perception annotations?*



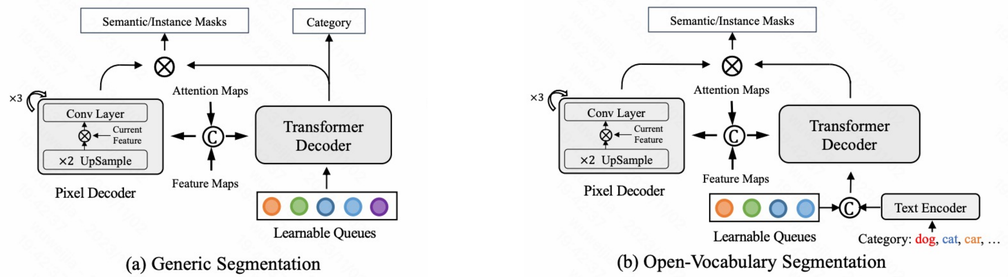
DatasetDM can produce **synthetic images**, along with various perception annotations, including depth, segmentation, and human pose estimation.

Method Overview



One P-Decoder is proposed to produce corresponding perception annotations such as masks and depth maps.

P-Decoder



The proposed decoder is a generalized architecture for the six supported tasks, with only minor variations required for different downstream applications, i.e., determining whether to activate certain layers.

Experiments

method	VOC (Semantic Seg.)/% # real	# synth.	mIoU	COCO2017 (Instance Seg.)/% # real	# synth.	AP	NYU Depth V2 (Depth Est.) # real	# synth.	REL ↓	COCO2017 (Pose Est.)/% # real	# synth.	AP
Baseline	100	-	65.2	400	-	14.4	50	-	0.31	800	-	42.4
DatasetDM	100	40k	78.5	400	80k	26.5	50	35k	0.21	800	80k	47.5

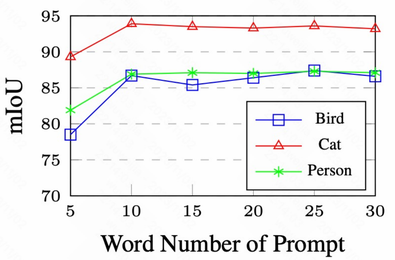
Performance for Downstream Tasks

method	backbone	# real image	# synthetic image	AP	AP^{50}	AP^{75}	AP^{S}	AP^{M}	AP^{L}
Baseline	R50	400	-	4.4	9.5	3.5	1.1	3.3	12.1
DatasetDM	R50	-	80k (R:400)	12.2	24.3	10.9	1.6	11.3	30.9
DatasetDM	R50	400	80k (R:400)	14.8	29.7	13.0	2.3	15.1	36.0
Baseline	Swin-B	400	-	11.3	23.0	9.6	3.2	10.1	27.1
DatasetDM	Swin-B	-	80k (R:400)	17.6	34.1	15.8	3.4	17.8	39.5
DatasetDM	Swin-B	400	80k (R:400)	23.3	43.0	22.2	7.7	26.1	48.7
Baseline	Swin-B	800	-	14.4	28.8	12.7	5.6	15.7	29.2
DatasetDM	Swin-B	800	80k (R:800)	26.5	46.9	25.8	7.7	29.8	53.3

Instance segmentation on COCO val2017. 'R: ' denotes the real data used to train.

Method	Zero-Shot Setting seen	unseen	harm.
Baseline(no Syn.)	61.3	10.7	18.3
Li *et al.* [38]	62.8	50.0	55.7
DiffuMask [62]	71.4	65.0	68.1
DatasetDM	78.8	60.5	68.4

Zero Shot Segmentation on VOC 2012



Zero Shot Segmentation on VOC 2012

Please Scan Code for Project Page