MAPR 2024 THE 7TH INTERNATIONAL CONFERENCE ON MULTIMEDIA ANALYSIS AND PATTERN RECOGNITION

IEEE

InstSynth: Instance-wise Prompt-guided Style Masked Conditional Data Synthesis for Scene Understanding

Thanh-Danh Nguyen[1,2], Bich-Nga Pham[1,2], Trong-Tai Dam Vu[1,2], Vinh-Tiep Nguyen[†1,2], Thanh Duc Ngo[1,2] and Tam V. Nguyen[3]

[1]University of Information Technology, Ho Chi Minh City, Vietnam, [2]Vietnam National University, Ho Chi Minh City, Vietnam, [3]University of Dayton, Dayton, OH 45469, United States († corresponding author)

Introduction

Overview: The scene understanding at the instance level is such an intense task that the models should have the ability to recognize each individual semantic instance. It plays an importance role in contributing to Advanced Driver Assistance Systems (ADAS) as a scene-understanding component.

Motivation: The need for extensive annotated training data to achieve accurate instance-level scene understanding. Manual annotation of such data is costly and requires significant effort.

Main contribution: InstSynth framework which employs prompt-guided style masked conditional data synthesis, utilize the existing annotated data to boost the performance of the instance segmentation models.

Method

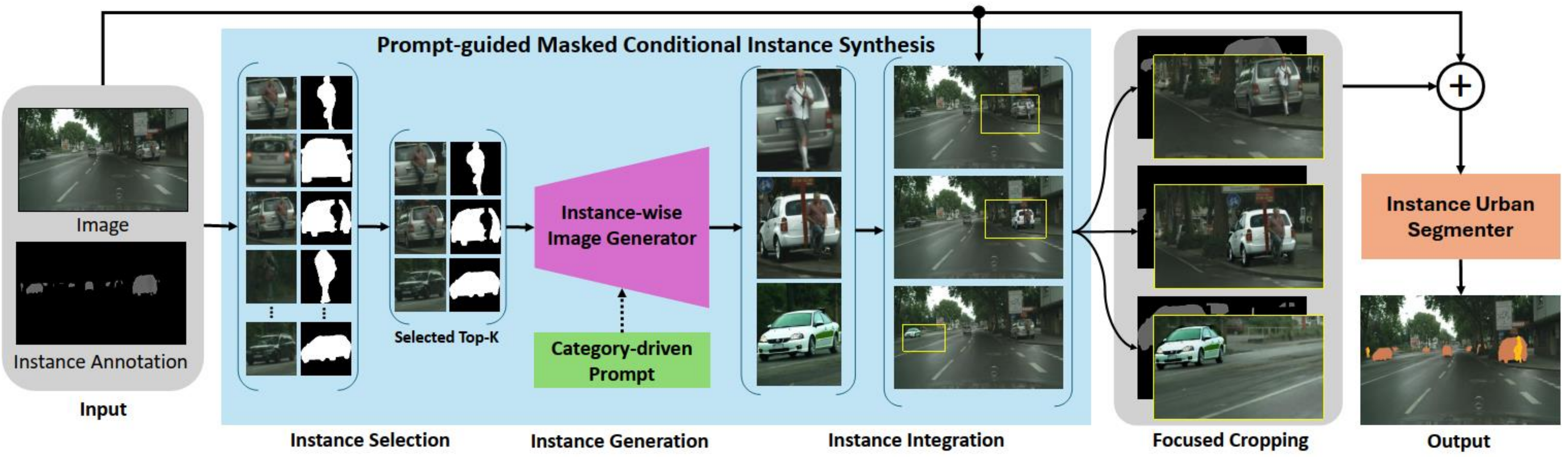

Fig. 1 Overview of our InstSynth framework. The pipeline allows a pair of image-annotation to be augmented into various variations with category-driven text prompts in temrs of boosting the data diversity to serve instance urban scene understanding

The **InstSynth** framework has two main components (Fig. 1): Prompt-guided Masked Conditional instance synthesis to diversify image-instance annotation pairs to strengthen the generalization of the segmentation model and the Instance Urban Segmenter for instance segmentation training.

Prompt-guided Masked Conditional Instance Synthesis:

- Utilizing the pre-trained generation models to diversify top-K prominent instances.
- Designed an algorithm for integrating the diversified instances back into the original images.

Instance Urban Segmenter:

- FastInst and OneFormer are employed to perform instance-wise urban scene understanding tasks.
- Trained on annotated data derived from real and augmented images.

Results

Method	Backbone	Version	Crop size	PQ ↑	IoU ↑	AP ↑	AP50 ↑
CMT-DeepLab‡ [30]	MaX-S† [30]	-	1025 × 2049	64.60	81.40	-	-
Axial-DeepLab-L‡ [31]	Axial ResNet-L† [31]	-	1025 × 2049	63.90	81.00	35.80	-
Axial-DeepLab-XL‡ [31]	Axial ResNet-XL† [31]	-	1025 × 2049	64.40	80.60	36.70	-
Panoptic-DeepLab‡ [32]	SWideRNet† [33]	-	1025 × 2049	66.40	82.20	40.10	-
OneFormer [9]	Mapillary-ConvNext-L	Original	360 × 720	48.84	72.58	21.75	40.94
	Swin-L		360 × 720	51.52	74.53	25.68	45.90
InstSynth* (Ours)	Mapillary-ConvNext-L	GLIGEN [4]	360 × 720	62.90	80.55	38.46	64.73
	Swin-L		360 × 720	60.33	79.18	35.67	61.09
	Mapillary-ConvNext-L	DiffInpainting [22]	360 × 720	62.90	80.96	38.66	64.69
	Swin-L		360 × 720	60.13	77.88	35.40	60.50
	Mapillary-ConvNext-L	BlendedDiff [23]	360 × 720	63.33	80.88	38.93	64.91
	Swin-L		360 × 720	60.47	79.10	35.75	61.01

All of our reproduced results of OneFormer are w/o CLIP, and w/ smaller crop size
The first, second, and third best results are marked in red, blue, and green, respectively.
** denotes our methods based on OneFormer instance segmentation architecture.*

Tab. 1 State-of-the-art comparison on CityScapes - OneFormertInst

Method	Backbone	Generation Base	AP	AP50
Mask2Former† [19]	R50-FPN-D3†	-	31.40	55.90
FastInst [8]	R50-FPN-D3†	-	35.50	59.00
	R50-FPN-D3*	-	24.93	45.69
	R50-FPN-D3**	-	27.65	49.21
InstSynth (Ours)	FastInst-R50-FPN-D3**	GLIGEN [4]	34.88	59.20
		DiffInpainting [22]	36.44	62.06
		BlendedDiff [23]	36.52	62.21

† denotes the published results of [8]
** denotes our reproduced results of FastInst w/o CLIP*
*** denotes our reproduced results of FastInst w/o CLIP, and w/ customized image sizes*
The first, second, and third best results are marked in red, blue, and green, respectively.

Tab. 2 State-of-the-art comparison on CityScapes - FastInst

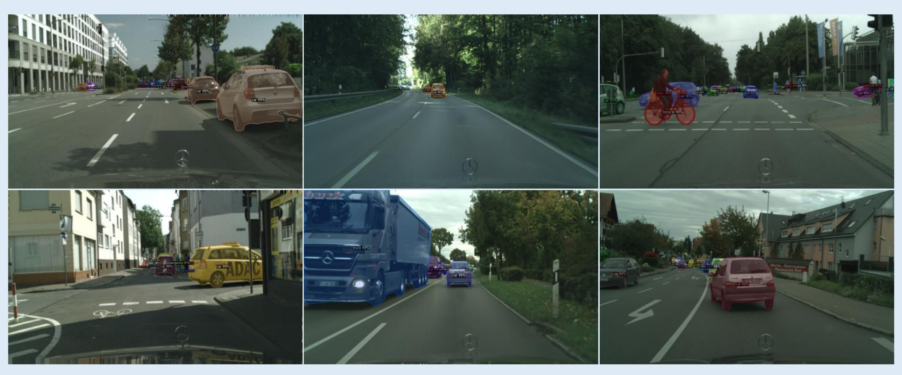

Fig. 2 Visualization results on CityScapes val-set with our FastInst R50-FPN-D3. The confidence threshold is 0.8

Method	CLIPScore ↑	FID ↓	SSIM ↑	PSNR ↑
GLIGEN [4]	0.79	125.51	0.67	14.39
DiffInpainting [22]	0.81	115.33	0.72	15.95
BlendedDiff [23]	**0.87**	**93.43**	**0.90**	**25.23**

The best results are marked in ***bold.***

Tab. 2 Ablation study on different image generation models

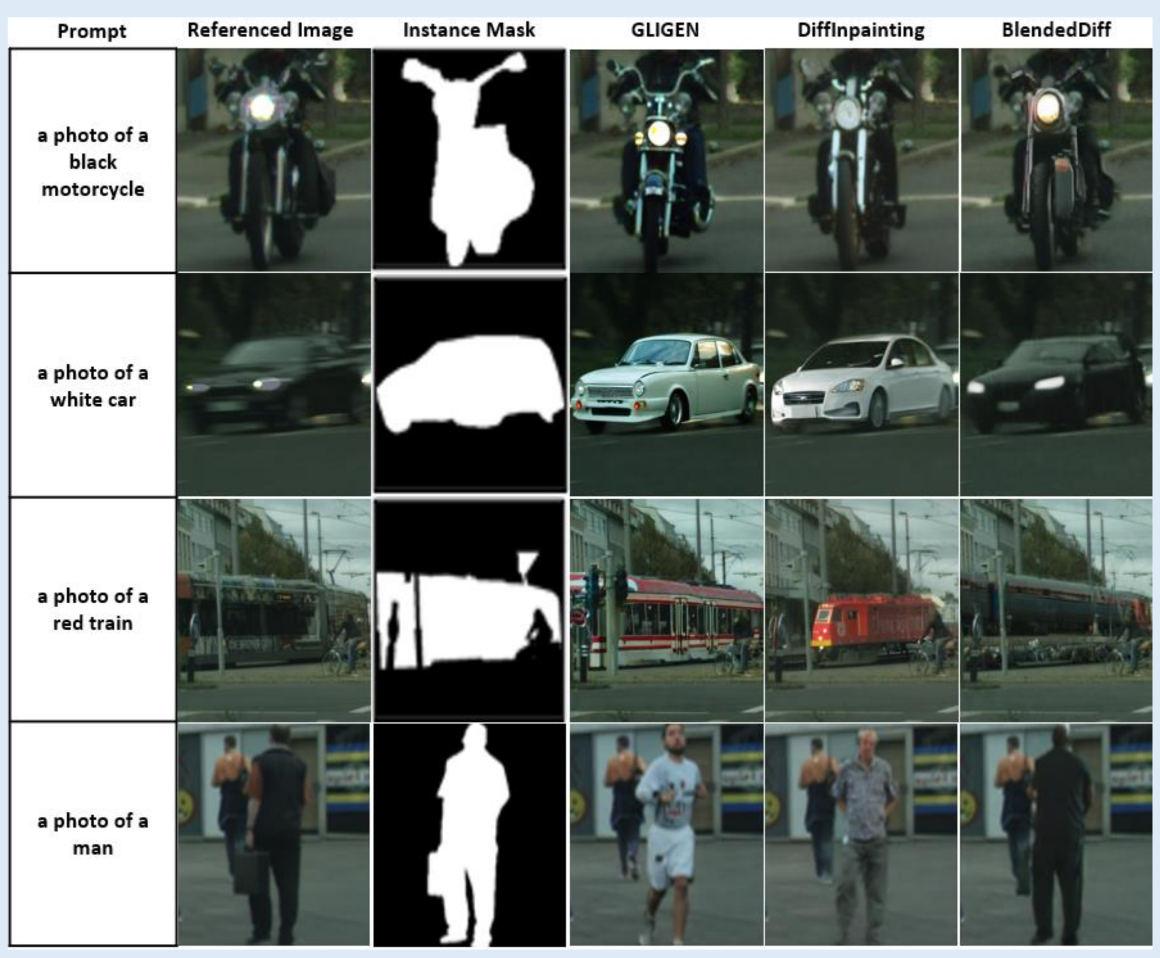

Fig. 3 Exemplary instance image generation from three different models

ORGANIZERS

DA NANG, VIETNAM, ON AUGUST 15-16, 2024

SPONSORS

KYANON One-Stop Digital Services House