ORGANIZATION

SPONSOR

MAPR 2024

THE 7TH INTERNATIONAL CONFERENCE ON MULTIMEDIA ANALYSIS AND PATTERN RECOGNITION

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
{danhnt, ngaptb, taidvt, tiepnv, thanhnd}@uit.edu.vn, tamnguyen@udayton.edu, †corresponding author

Da Nang, August 15-16th, 2024

Content

1. Introduction
2. Related work
3. Our proposed InstSynth
   - Prompt-guided Masked Conditional Instance Synthesis (ProMCIS)
   - Instance-wise Urban Segmenter
4. Experiments
5. Conclusion

1. Introduction

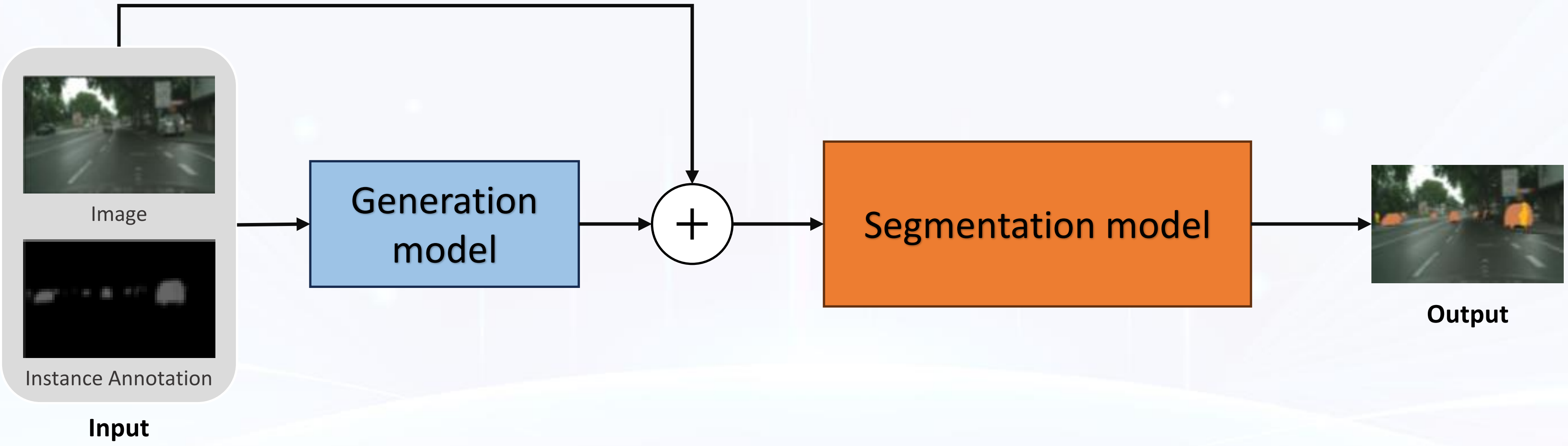
MAPR 2024

- **Instance-level Scene Understanding** is crucial in computer vision to support modern Advanced Driver Assistance Systems
- Abundant annotated training data is required to tackle this task.
- **Instance-level annotation is costly** due to significant manual effort required.

1. Introduction

Contribution:

- Introduced **InstSynth** for enhancing scene understanding with a novel data synthesis approach
- Constructed **IS-Cityscapes** - an instance-level synthesized dataset
- Significantly outperformed state-of-the-art models FastInst and OneFormer on the Cityscapes benchmark, achieving **increases in AP of 14.49% and 11.59%**, respectively.



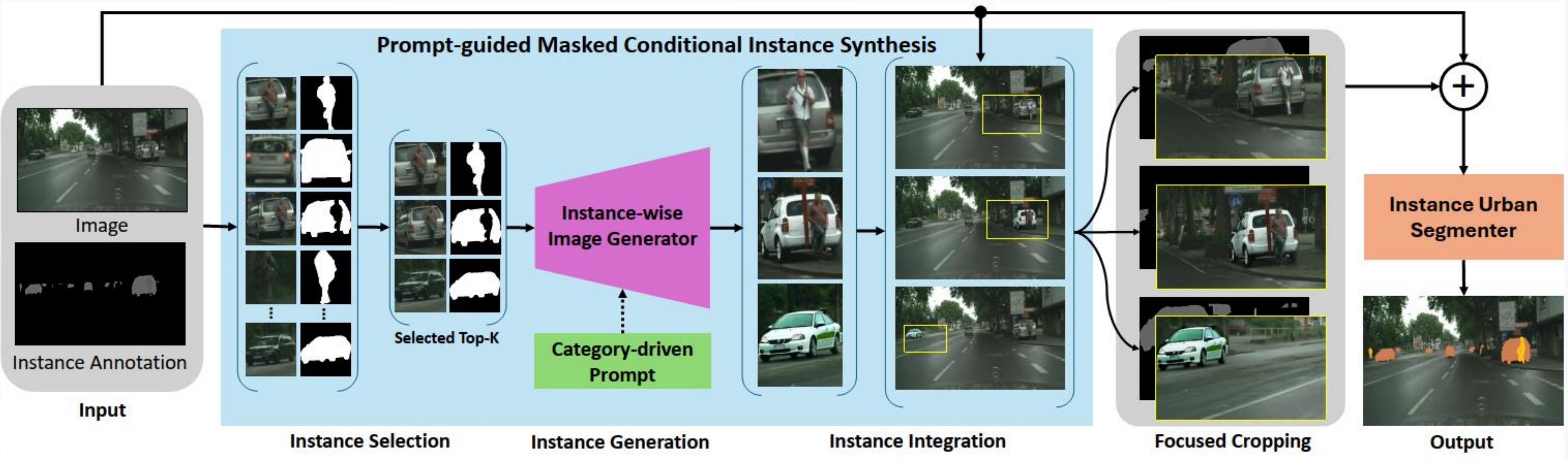
2. Related work

MAPR 2024

- **Urban Scene Understanding research:** gaining attention within the community due to its wide range of potential applications.
- **Instance segmentation models:** one-stage approach or two-stage approach.
- **Conditional Image Generation models:** Diffusion-based models demonstrate outstanding capabilities in generating and editing diverse and high-quality images guided by text prompts.
- **Data Augmentation:** using traditional augmentation techniques or using deep learning-based augmentation techniques.
- **Urban Scene Datasets:** Cityscapes, CamVid, Mapillary are among the potential high-resolution urban scene datasets featuring fine-grained annotations.

3. Method

InstSynth has 2 main components:
- ▶ Prompt-guided Masked Conditional Instance Synthesis
- ▶ Instance-wise Urban Segmenter

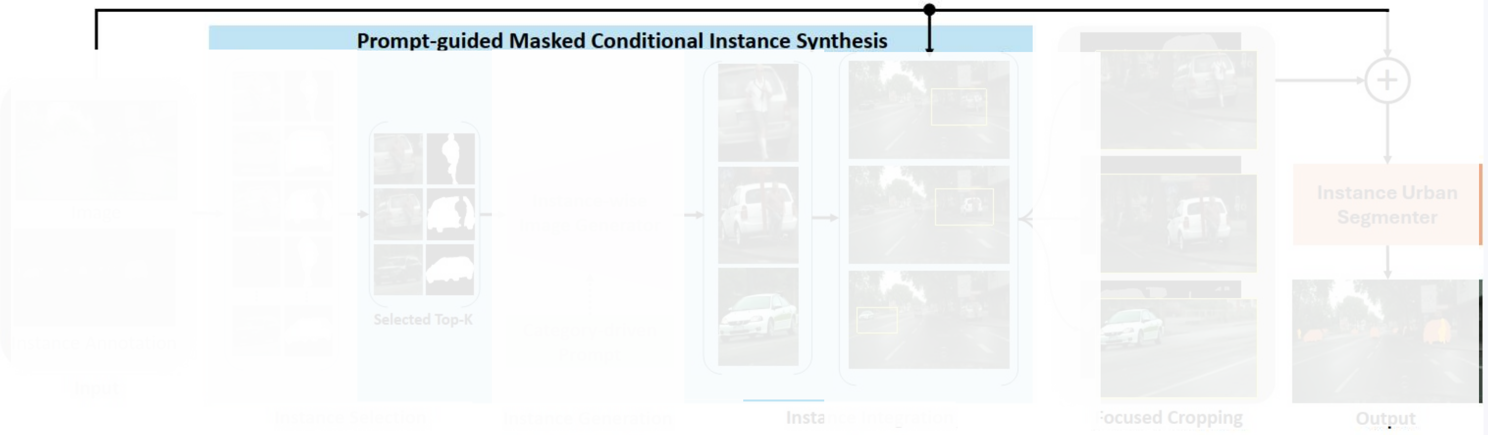


Overview of our InstSynth framework.

InstSynth **makes use of existing annotated data** to boost the performance of the instance segmentation model

3. Method

Prompt-guided Masked Conditional Instance Synthesis: generates realistic urban images in three phases using the Cityscape dataset, ensuring adherence to dataset regulations and reliability standards.

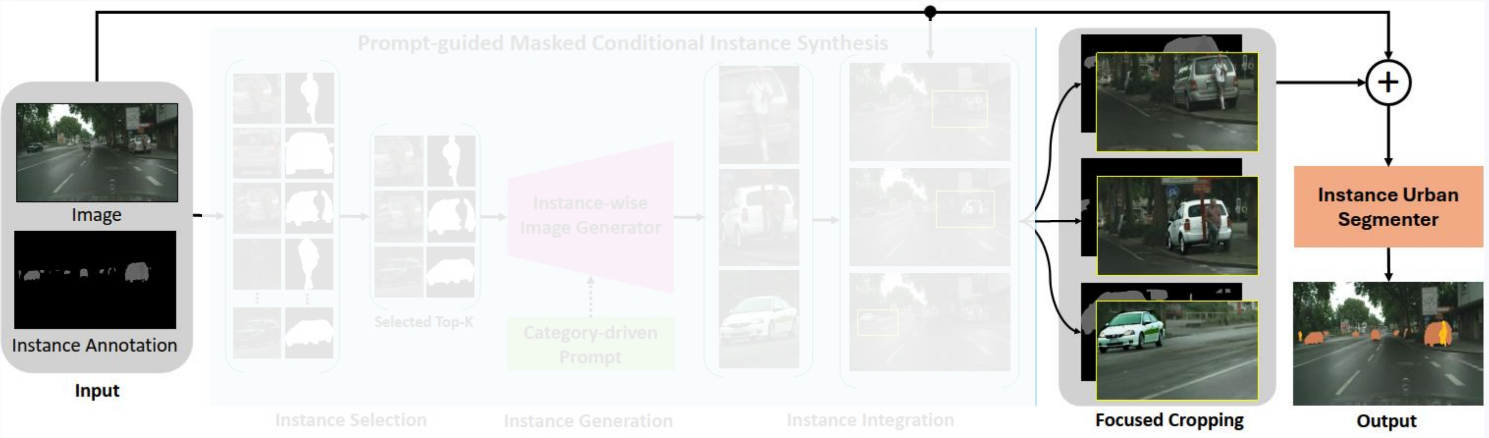


Focus: Prompt-guided Mask Confitional Instance Synthesis module

- Phase 1: Select and crop top-K prominent instances from mask annotations.
- Phase 2: Use pre-trained generation models (GLIGEN, DiffInpainting, BlendedDiff).
- Phase 3: Integrate the inpainted images back into the original images using our algorithm.

3. Method

Instance-wise Urban Segmenter: FastInst and OneFormer are employed to perform instance-wise urban scene understanding tasks.



Focus: Instance-wise Urban Segmenter

- The instance urban segmenter trains on annotated data from real and augmented images.

MAPR 2024

4. Experiments

- Our **InstSynth with BlendedDiff** helps FastInst and OneFormer enhance their performance, improving **AP scores from 35.5% and 21.75% to 36.52% and 38.93%,** respectively, on CityScapes.

Method	Backbone	Version	Crop size	PQ ↑	IoU ↑	AP ↑	AP50 ↑
CMT-DeepLab‡ [29]	MaX-S† [29]	-	1025 × 2049	64.60	81.40	-	-
Axial-DeepLab-L‡ [30]	Axial ResNet-L† [30]	-	1025 × 2049	63.90	81.00	35.80	-
Axial-DeepLab-XL‡ [30]	Axial ResNet-XL† [30]	-	1025 × 2049	64.40	80.60	36.70	-
Panoptic-DeepLab‡ [31]	SWideRNet† [32]	-	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
	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 [21]	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 [22]	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.

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
	R50-FPN-D3**	GLIGEN [4]	34.88	59.20
		DiffInpainting [21]	36.44	62.06
		BlendedDiff [22]	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. State-of-the-art comparison on CityScapes. Left: Comparison on OneFormer. Right: Comparison on FastInst

4. Experiments – Ablation Study

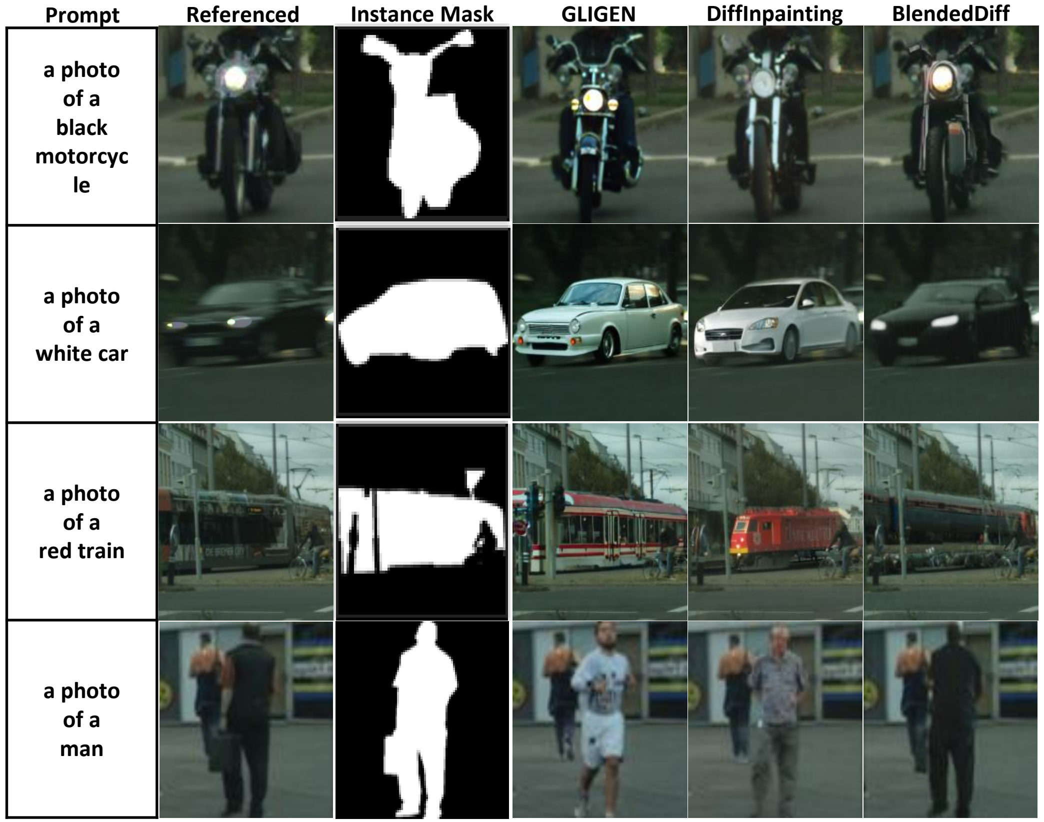
- **BlendedDiff** demonstrates its empowerfulness when it yields the highest performance over all four mentioned metrics.

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 [21]	0.81	115.33	0.72	15.95
BlendedDiff [22]	**0.87**	**93.43**	**0.90**	**25.23**

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

Tab. Ablation study on different image generation models



Exemplary instance image generation from three different models of GLIGEN, DiffInpainting, and BlendedDiff

5. Conclusion

MAPR 2024

In this work:

- We proposed **InstSynth** – a novel instance-wise prompt-guided synthetic data approach for instance-wise scene understanding.
- We constructed **IS-CityScapes** – a synthesized dataset that increase four times the number of instances to over 200K for training
- Experimental results proves our SOTA results on CityScapes

In the future:

- Improve the ability of our instance generation method to deal with various diversity to solve real-world intense situations while driving.

ORGANIZATION

SPONSOR

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

{danhnt, ngaptb, taidvt, tiepnv, thanhnd}@uit.edu.vn, tamnguyen@udayton.edu, †corresponding author

Acknowledgements

ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN
VNUHCM - UIT

MAPR 2024

Da Nang, August 15-16th, 2024