哈爾濱工業大學 HARBIN INSTITUTE OF TECHNOLOGY
THE HONG KONG POLYTECHNIC UNIVERSITY 香港理工大學
TAL 好未来
JUNE 18-22, 2023 CVPR VANCOUVER, CANADA
WED-AM-287

Inferring and Leveraging Parts from Object Shape for Improving Semantic Image Synthesis

Yuxiang Wei[1,2], Zhilong Ji[3], Xiaohe Wu[1], Jinfeng Bai[3], Lei Zhang[2], Wangmeng Zuo[1]

[1]Harbin Institute of Technology, [2]The Hong Kong Polytechnic University, [3]Tomorrow Advancing Life

Paper: http://arxiv.org/abs/2305.19547 Code: https://github.com/csyxwei/iPOSE

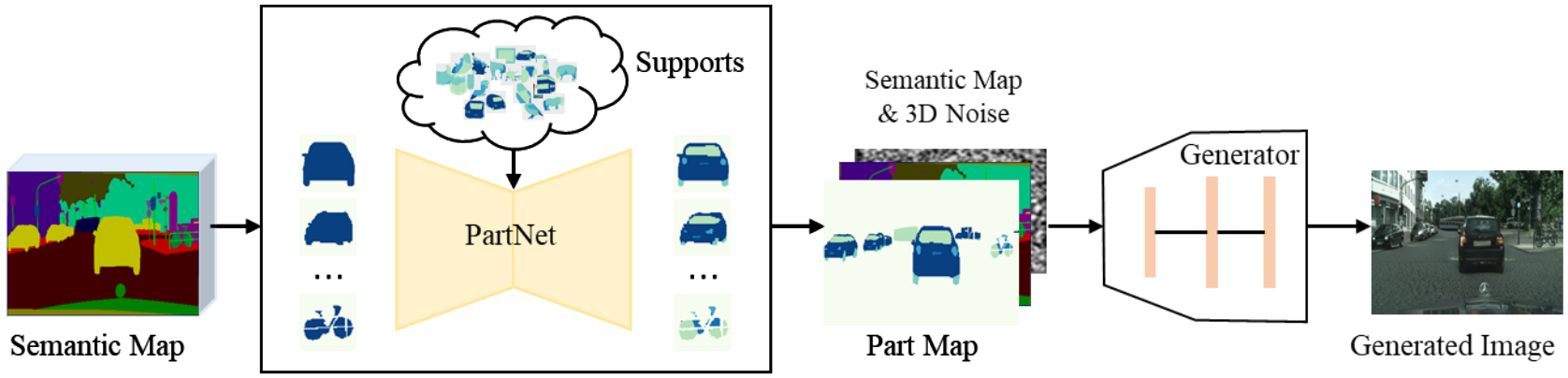
Semantic Image Synthesis

- Semantic Image Synthesis
  - Generating semantically aligned and photo-realistic images with the given semantic maps.

- Our Goal
  - Generating images with realistic object parts

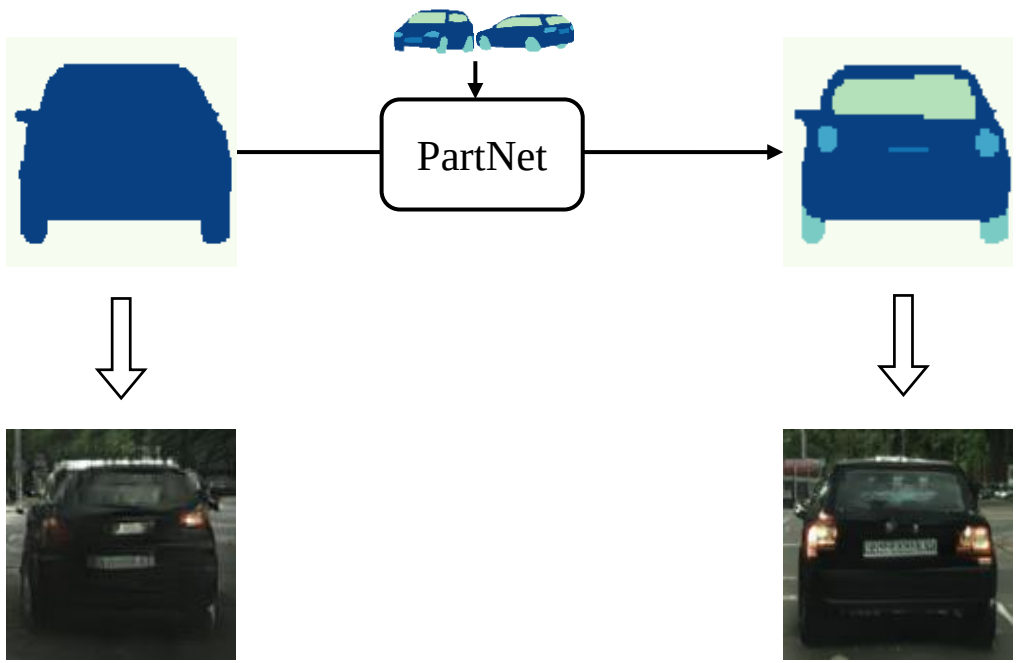
Our iPOSE

- We propose to infer parts from the object shape and leverage it to improve semantic image synthesis.

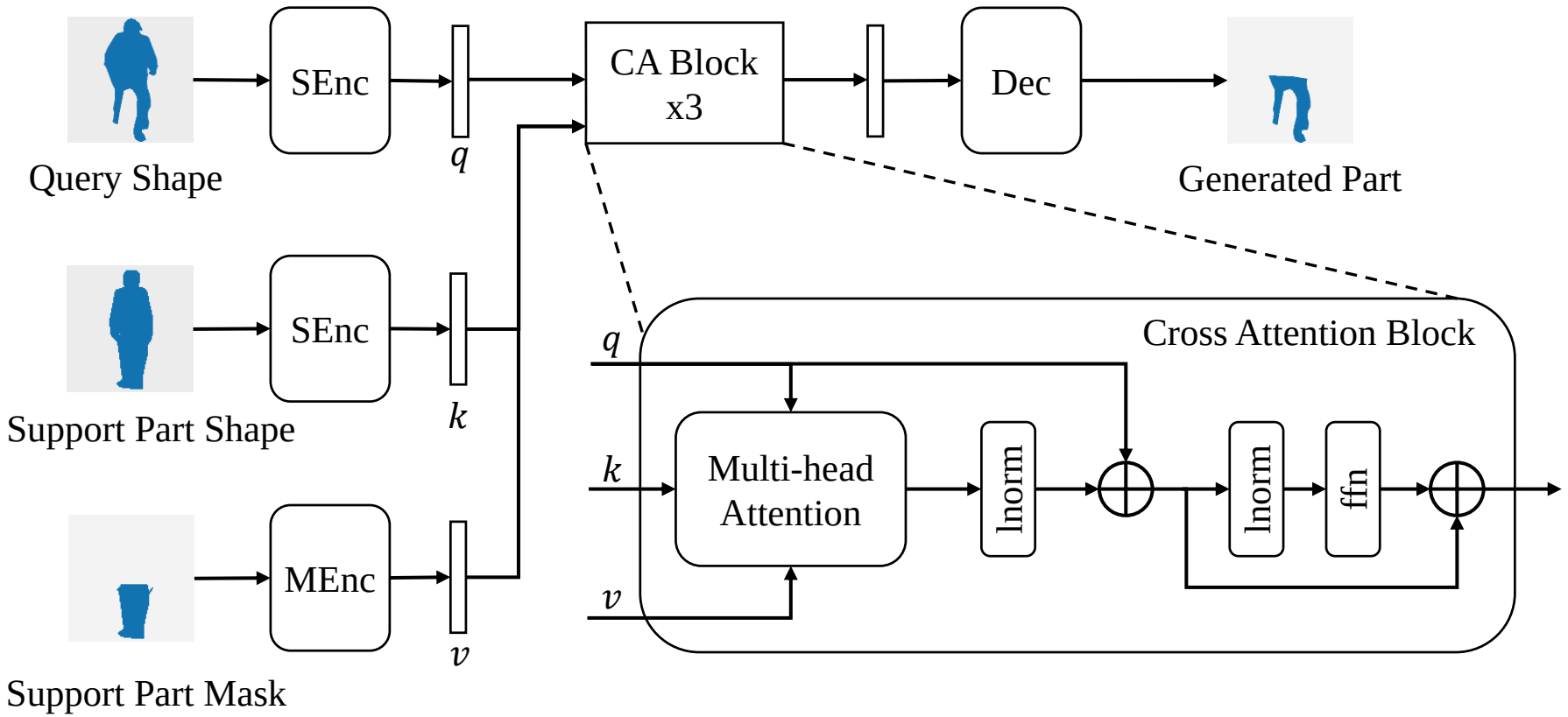


Motivation

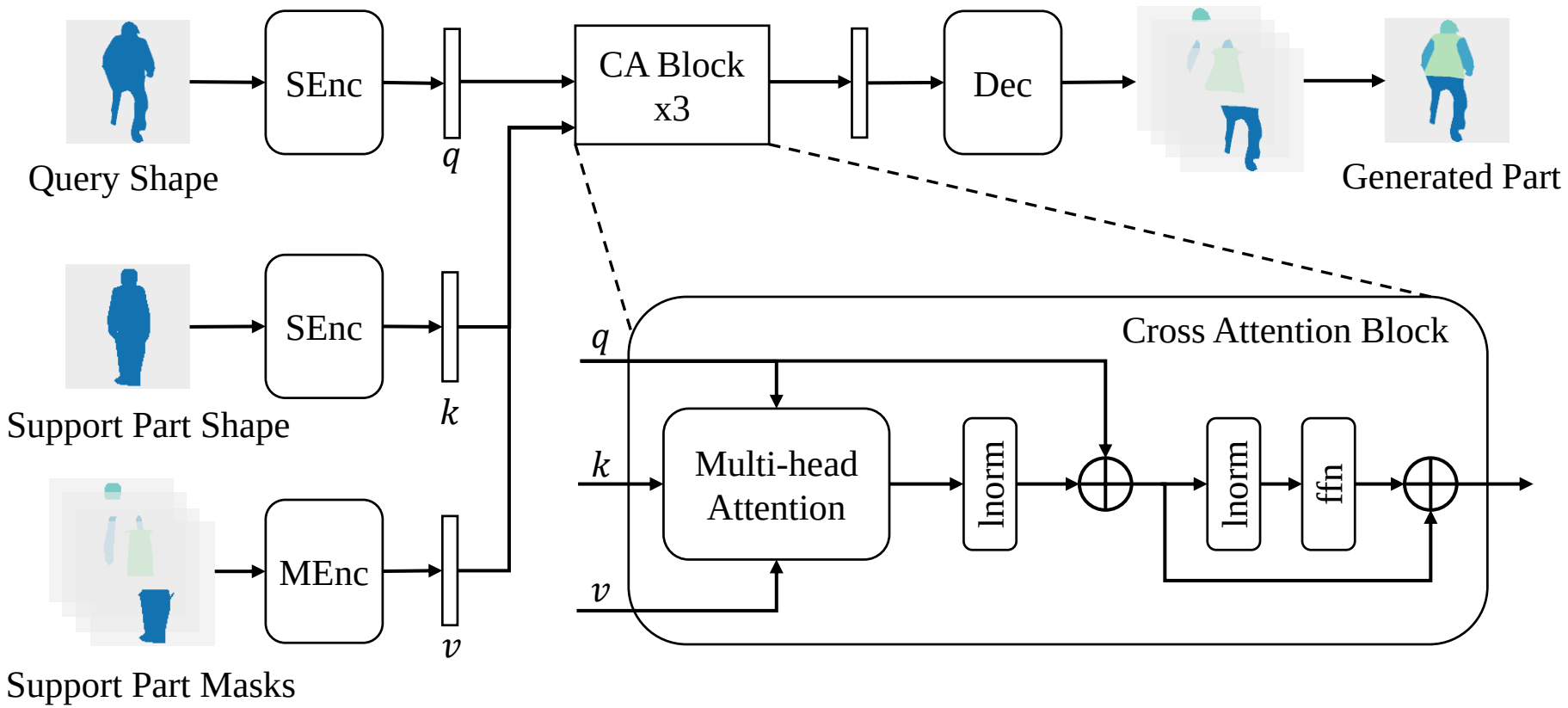
- Existing methods only exploited the object-level semantic information for image synthesis, and usually failed to generate photo-realistic object parts.



PartNet


Query Shape
SEnc
q
CA Block
x3
Dec
Generated Part
Support Part Shape
SEnc
k
Support Part Mask
MEnc
v
Cross Attention Block
q
k
v
Multi-head
Attention
lnorm
lnorm
ffn

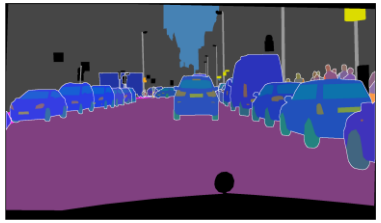
PartNet



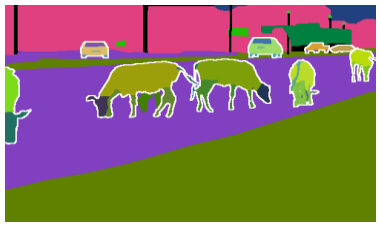
$$\mathcal{L}_{pre} = \text{BCE}(\text{PartNet}(O_q, S_{y_q}), P_q^{gt}), \qquad \mathcal{L}_{rec} = \text{BCE}(\text{PartNet}(O_q, P_q^{gt}), P_q^{gt}).$$

Part Dataset

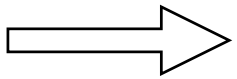
Cityscapes PPS



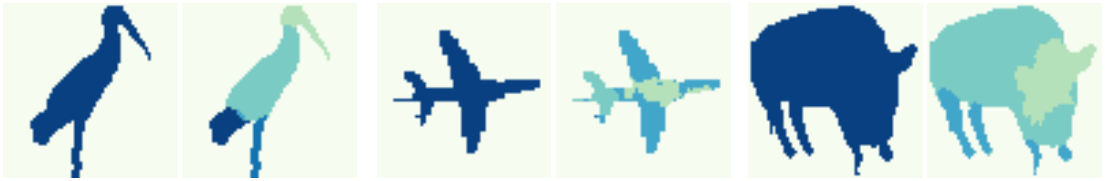
Pascal VOC PPS



Decompose



Horse Motorbike Plant



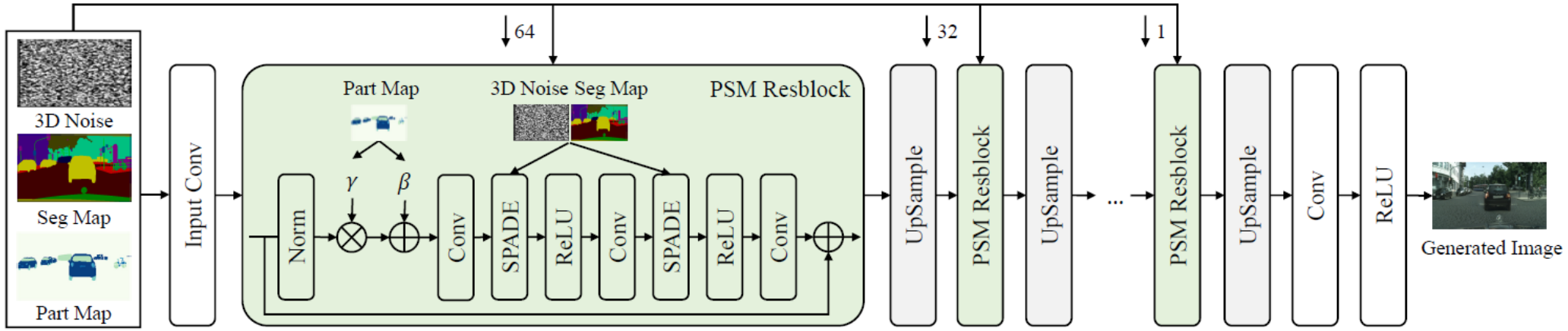
Bird Aeroplane Sheep

...

de Geus, Daan, et al. "Part-aware panoptic segmentation." CVPR 2021.

Part Semantic Modulation

- To achieve a disentangled synthesis, we inject the part and semantic sequentially.



$$\mathcal{L}_D^{global} = \mathbb{E}_I[\log D(I)] + \mathbb{E}_{M,Z}[\log(1 - D(G(M, P, Z))],$$

$$\mathcal{L}_G^{global} = \mathbb{E}_{M,Z}[\log(1 - D(G(M, P, Z))],$$

$$\mathcal{L}_{style} = \max\left(\frac{1}{n}\sum_i \min_j \boldsymbol{C}_{i,j}, \frac{1}{m}\sum_j \min_i \boldsymbol{C}_{i,j}\right)$$

Qualitative Comparison

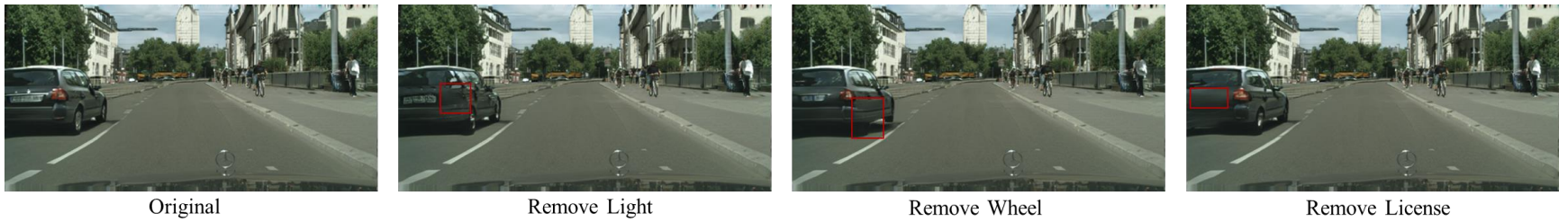
(a) Qualitative comparison on Cityscapes

(b) Qualitative comparison on ADE20K (top two rows) and COCO-Stuff (bottom two rows)

Quantitative Comparison

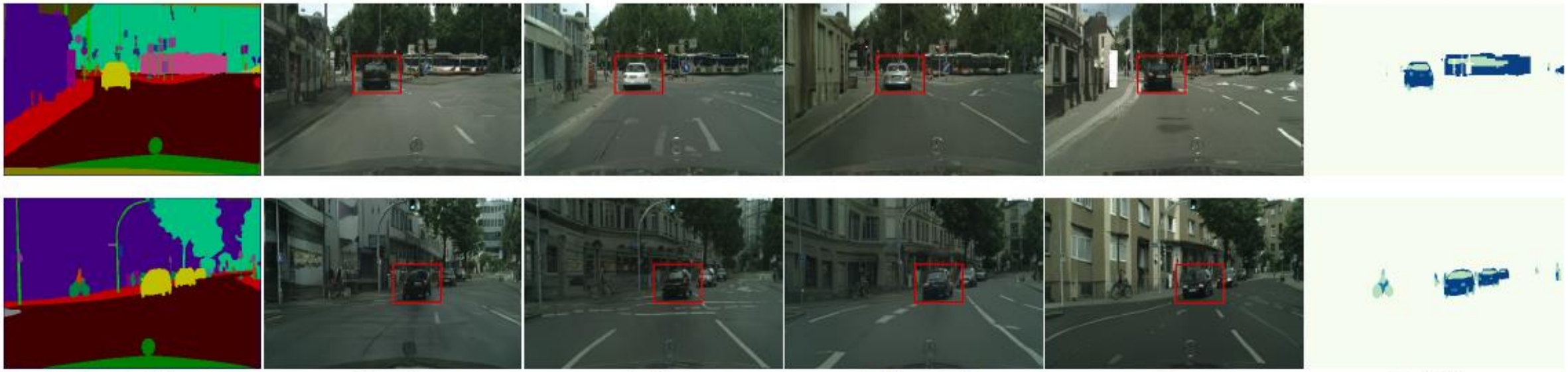
Method	Cityscapes			ADE20K			COCO-Stuff		
	FID (↓)	AC (↑)	mIOU (↑)	FID (↓)	AC (↑)	mIOU (↑)	FID (↓)	AC (↑)	mIOU (↑)
SIMS [30]	49.7	75.5	47.2	n/a	n/a	n/a	n/a	n/a	n/a
SPADE [28]	71.8	81.9	62.3	33.9	79.9	38.5	22.6	67.9	37.4
CC-FPSE [23]	54.3	82.3	65.5	31.7	82.9	43.7	19.2	70.7	41.6
SC-GAN [49]	49.5	82.5	66.9	29.3	83.8	45.2	18.1	72.0	42.0
OASIS [34]	47.7	n/a	69.3	28.3	n/a	48.8	<u>17.0</u>	n/a	44.1
RESAIL [37]	45.5	**83.2**	69.7	30.2	84.8	49.3	18.3	73.1	<u>44.7</u>
SAFM [25]	49.5	<u>83.1</u>	70.4	32.8	<u>86.6</u>	<u>50.1</u>	24.6	<u>73.4</u>	43.3
SDM [48]	<u>42.1</u>	n/a	**77.5**	<u>27.5</u>	n/a	39.2	n/a	n/a	n/a
Ours	**41.3**	82.2	<u>70.6</u>	**26.9**	**87.1**	**53.8**	**15.7**	**74.8**	**45.1**

Multi-modal Synthesis



Ablation Study

Part Inject	$\mathcal{L}^{g}_{G/D}$	$\mathcal{L}_{style}$	FID(↓)	mIOU(↑)	AC(↑)	obj FID (↓)
			47.7	66.9	81.5	44.1
	✓		43.6	66.7	81.9	39.2
	✓	✓	42.8	70.5	82.1	37.5
Ours PSM	✓	✓	**41.3**	**70.6**	**82.2**	**30.4**
SPADE	✓	✓	42.9	69.9	81.9	31.2
Concat	✓	✓	42.7	70.6	82.0	31.5



Take Home Message

- We propose a method iPOSE to infer parts from object shape and leverage them to improve semantic image synthesis.
- A PartNet is proposed to predict the part map based on a few support part maps, which can be easily generalized to new object categories.
- A part semantic modulation Resblock is presented to incorporate the predicted part map and semantic map for image synthesis.
- Global adversarial and object-level CLIP style losses are further introduced to generate photo-realistic images

哈爾濱工業大學 HARBIN INSTITUTE OF TECHNOLOGY

THE HONG KONG POLYTECHNIC UNIVERSITY 香港理工大學

TAL 好未来

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Thank you!

Paper: http://arxiv.org/abs/2305.19547

Code: https://github.com/csyxwei/iPOSE