JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

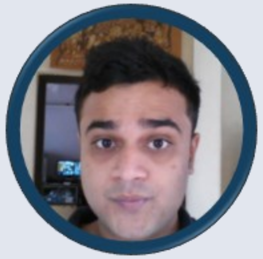
DejaVu: Conditional Regenerative Learning to Enhance Dense Prediction

THU-AM-284

Qualcomm AI Research
Qualcomm AI Research is an initiative of Qualcomm Technologies, Inc.

Shubhankar Borse
Senior Engineer

Hyojin Park
Senior Engineer



Debasmit Das
Staff Engineer

Hong (Herbert) Cai
Staff Engineer

Risheek Garrepalli
Senior Engineer

Fatih Porikli
Senior Dir, Technology

Overview

Introduction

Using regeneration as an auxiliary task for dense supervision

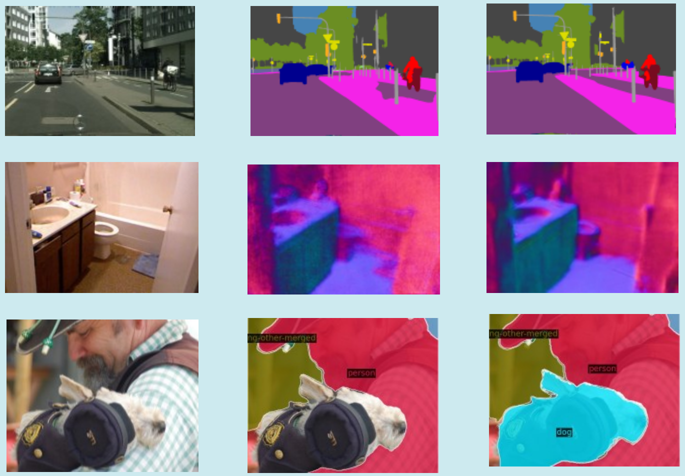
Approach

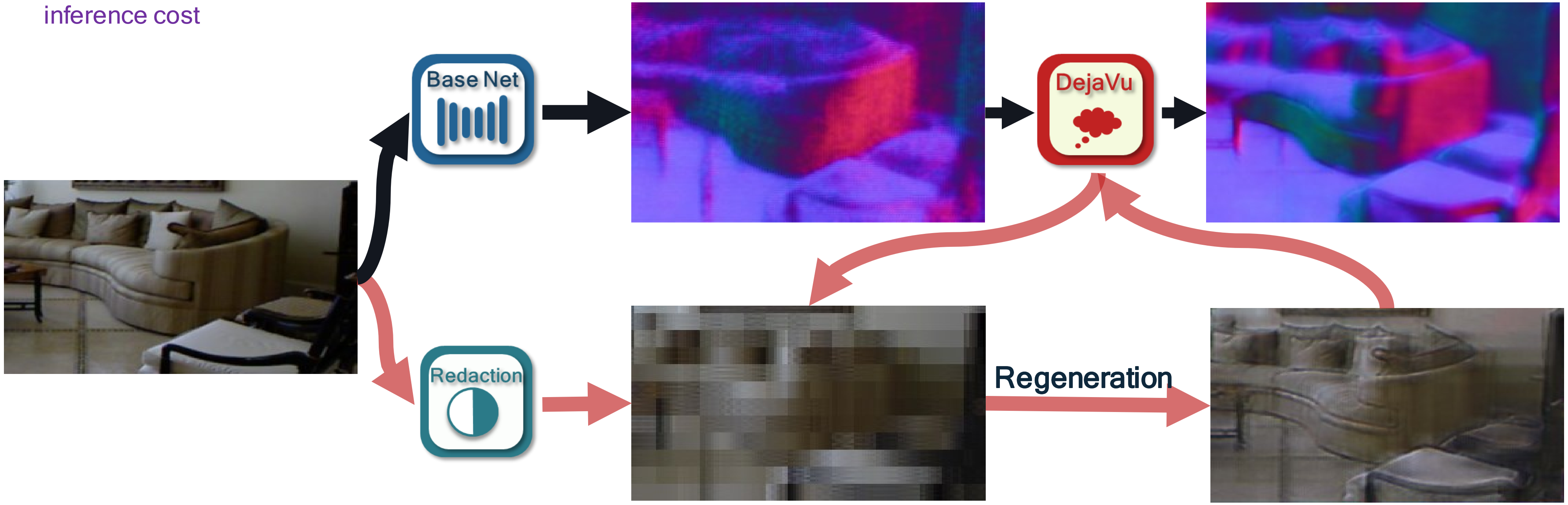
Image redaction and regeneration given an initial dense task

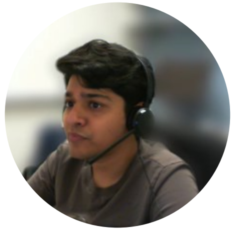
Results

Improved dense prediction at no inference cost

Visual Results







Related Work

- Generative models as an auxiliary task
- Generative models for pre-training

Motivation

- Regeneration as an auxiliary task

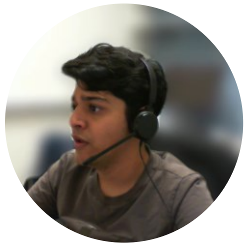
Method

- Redaction types
- Conditional Regeneration module
- DejaVu loss
- DejaVu Shared Attention Module (DV-SA)

Results

- Qualitative/Quantitative analysis
- Accuracy v/s Computation analysis

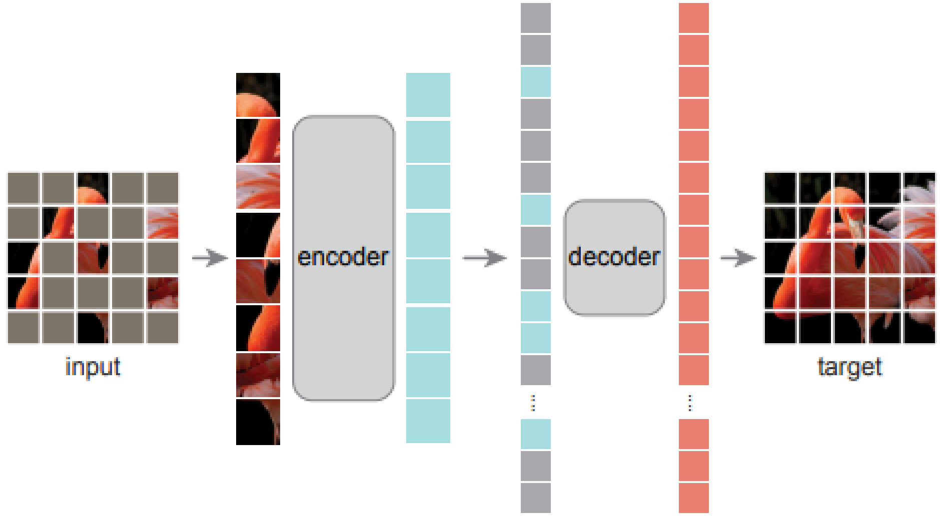
Related Work

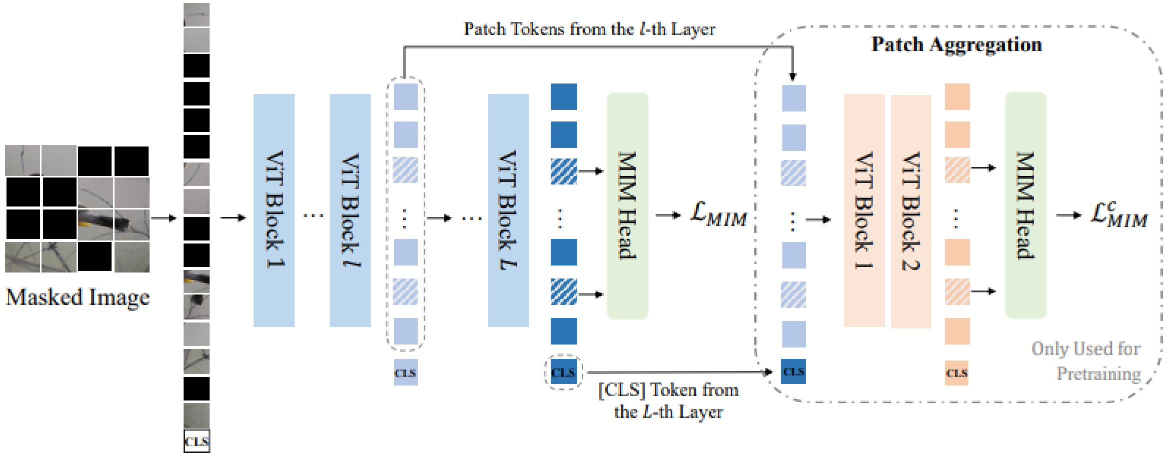


Source sample text

Regeneration as a pretraining task

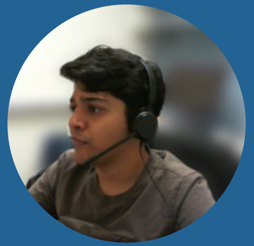
- Masked image modeling - Masked autoecoders
- Using masked autoencoders for specific tasks (segmentation, depth estimation, object detection)

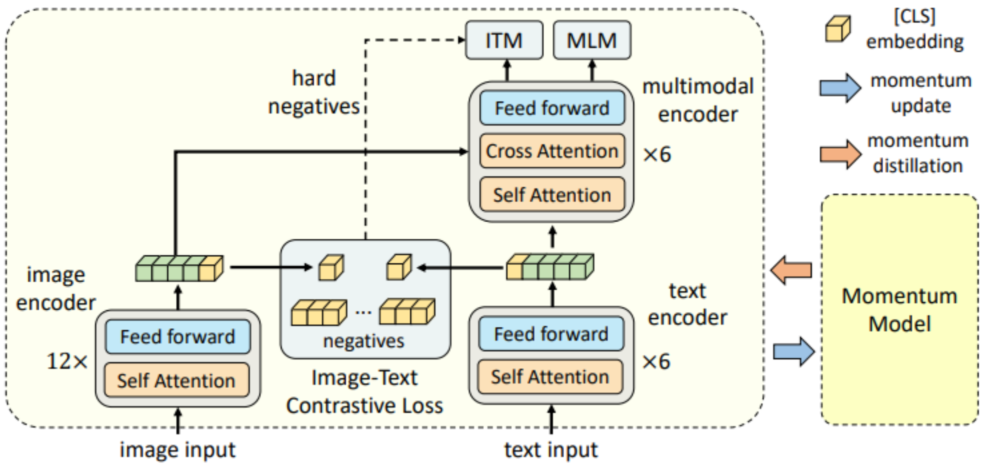


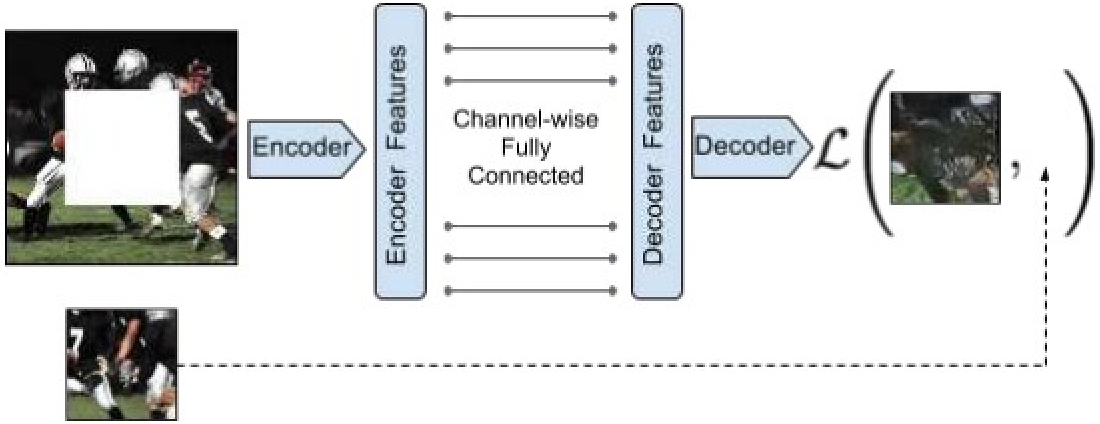


Sources:

- He, Kaiming, et al. "Masked autoencoders are scalable vision learners. " *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022.
- Peng, Zhiliang, et al. "Beit v2: Masked image modeling with vector-quantized visual tokenizers." *arXiv preprint arXiv:2208.06366* (2022).





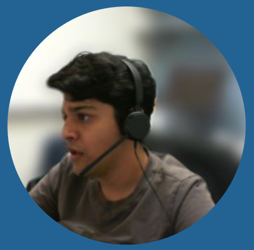


Train generative models using a reconstruction scheme

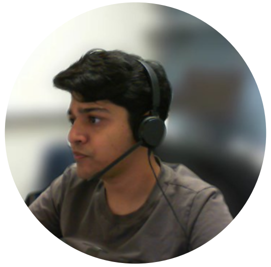
- Drop certain regions or mask several pixels
- Using reconstruction loss to enhance the representation power.
- Train the models with unsupervised manner as like adversarial loss

Sources:

- Li, Junnan, et al. "Align before fuse: Vision and language representation learning with momentum distillation." *Advances in neural information processing systems* 34 (2021): 9694-9705.
- Pathak, Deepak, et al. "Context encoders: Feature learning by inpainting." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.



Motivation



Source sample text

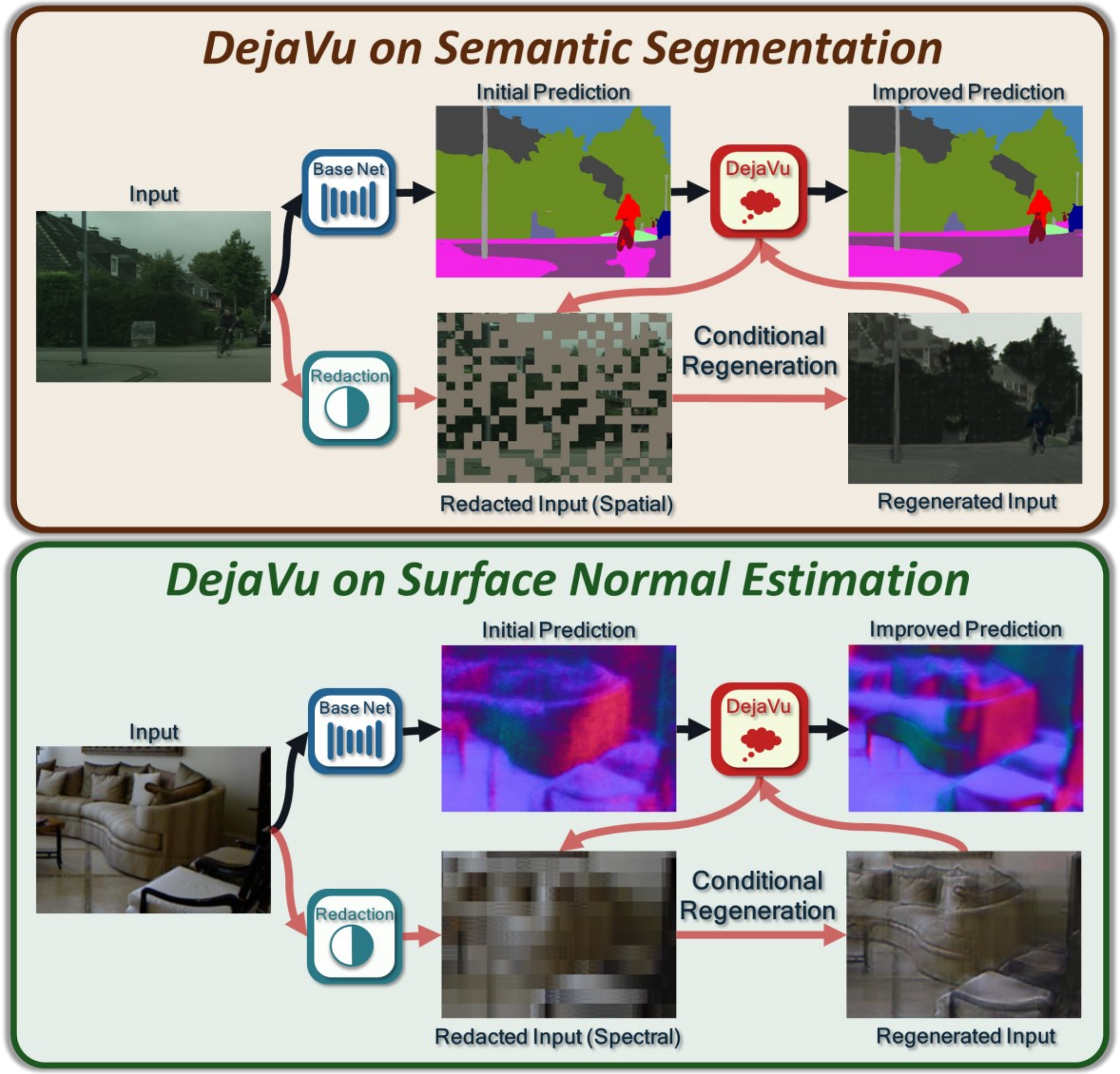


Image Regeneration as an auxiliary task

- Using image regeneration as an auxiliary task to improve dense predictions.
- We control the types of redactions based on the dense task.
- Inpainting-based regeneration is a good base for segmentation, whereas spectral regeneration improves tasks such as surface normal or depth estimation.

Motivation

- Regeneration as an auxiliary task

Related Work

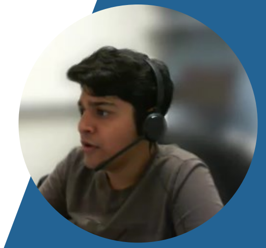
- Generative models as an auxiliary task
- Generative models for pre-training

Method

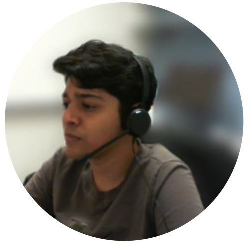
- Redaction types
- Conditional Regeneration module
- DejaVu loss
- DejaVu Shared Attention Module (DV-SA)

Results

- Qualitative/Quantitative analysis
- Accuracy v/s Computation analysis



Method



Method : Types of redaction

Various redaction types based on the task

- Spectral redaction: Redacting frequencies
- Spatial redaction: Redacting pixels

Input image

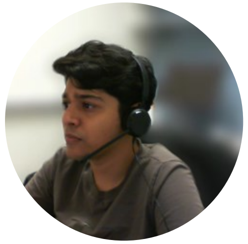
Spectral (Lowpass)

Spectral (Bandstop)

Spatial (Random)

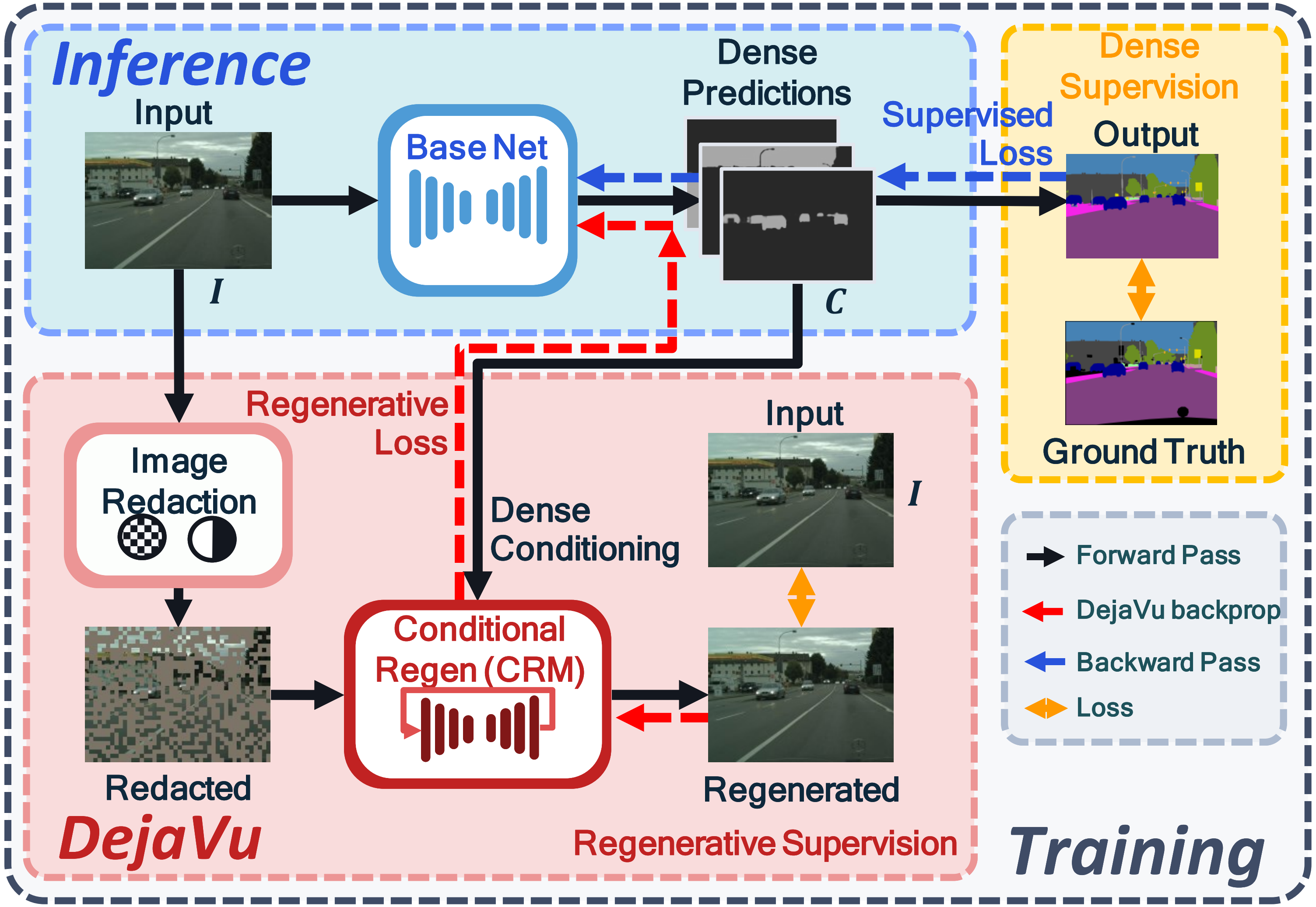
Spatial (Checkerboard)

Spatial (Random Blocks)

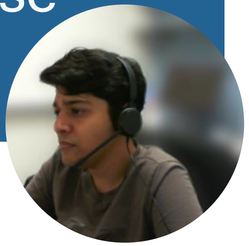


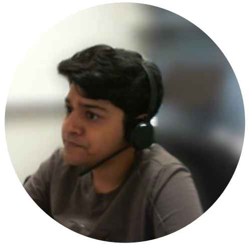
Method : The DejaVu loss function

Key Idea and implementation of the DejaVu loss



- Training: base model
- Dense Supervision
- Image Redaction
- Conditional regeneration
- Regenerative supervision
- Inference: using only base model
- DejaVu pushes the network to learn accurate dense predictions

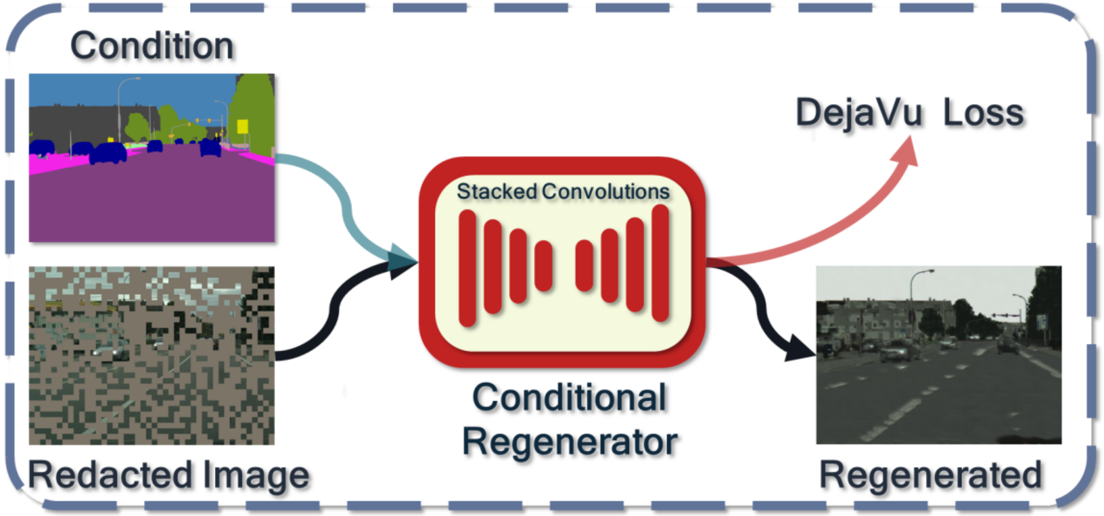




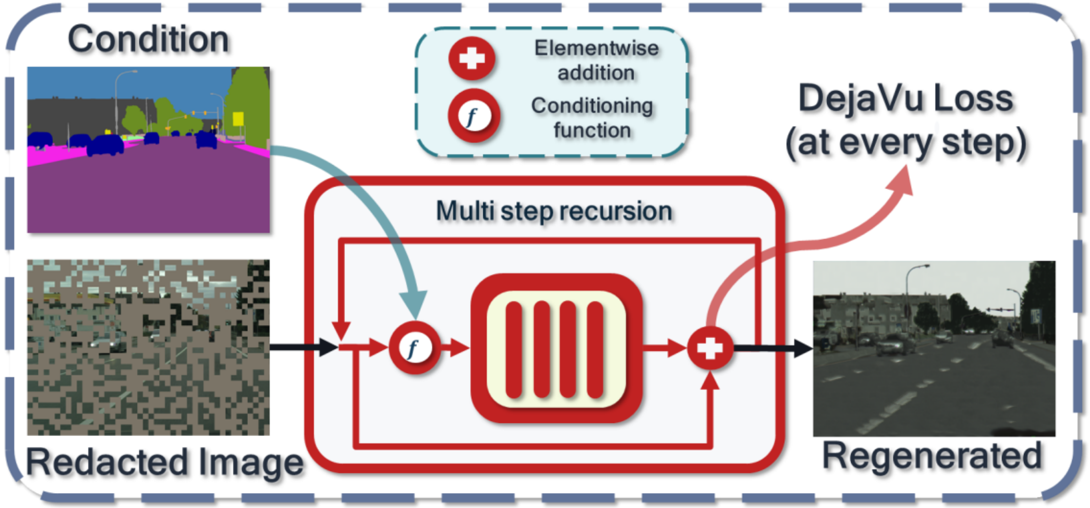
Method : Conditional Regeneration Module (CRM)

Two types of CRM modules which were studies

- Forward Mode
- Recursive Mode



(a) Conditional Regeneration Module: Forward Mode (CRM-F)

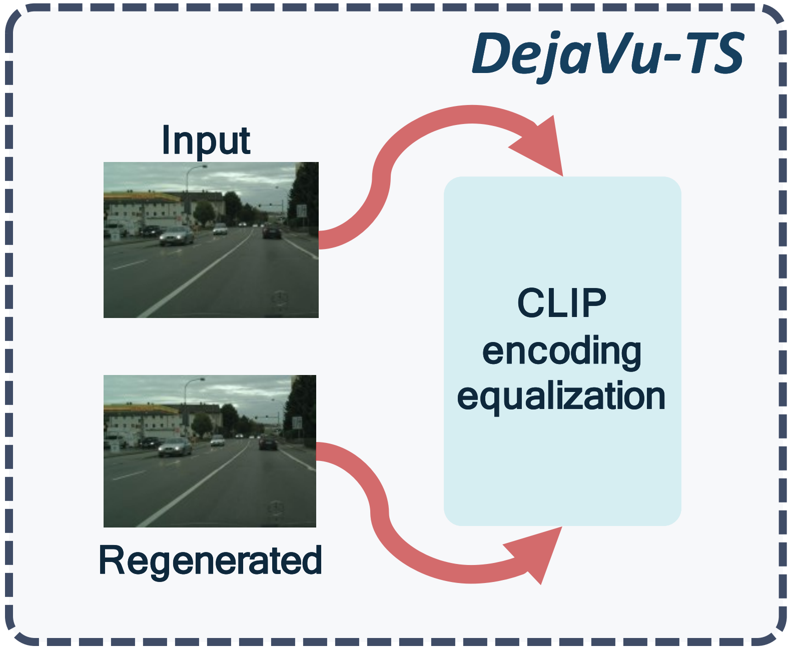


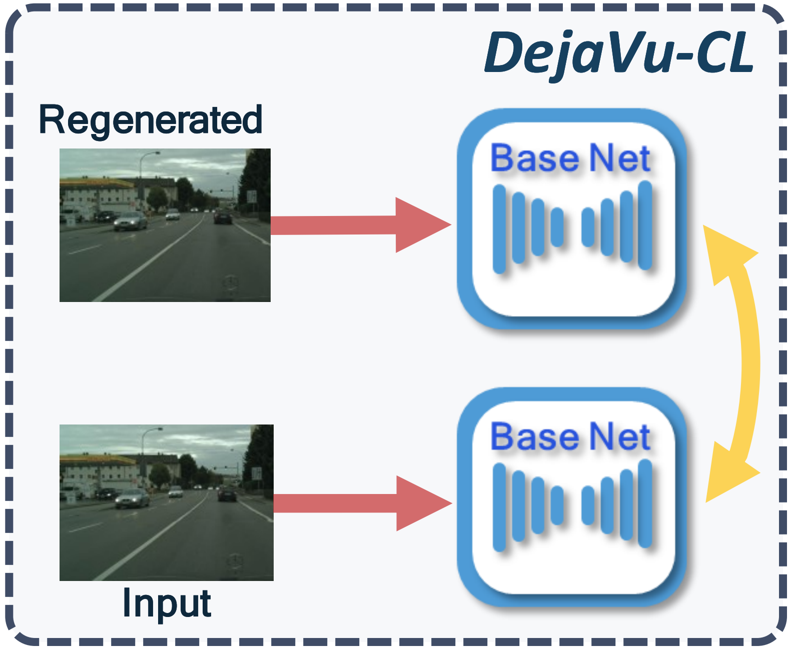
(b) Conditional Regeneration Module : Recursive Mode (CRM-R)

Method : Further supervision with DejaVu

Two types of objectives once the regenerated images are obtained

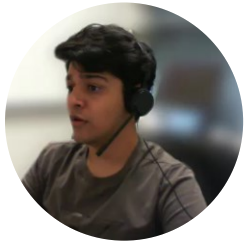
- Text supervision loss with CLIP (DejaVu-TS)
- Cyclic consistency loss (DejaVu-CL)





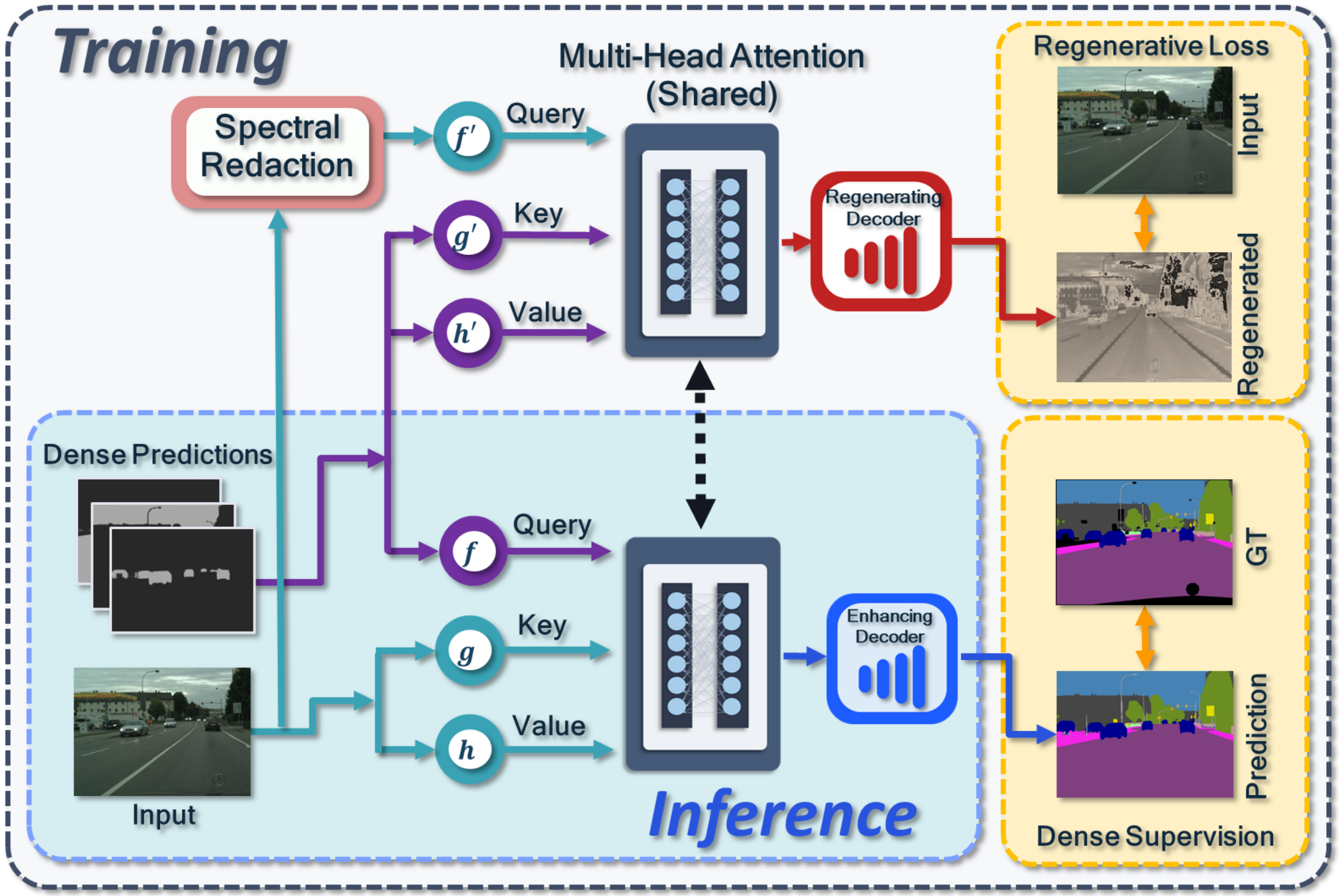
Source:

- Radford, Alec, et al. "Learning transferable visual models from natural language supervision." *International conference on machine learning*. PMLR, 2021.



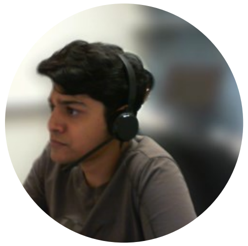
Method : The DejaVu Shared Attention Module (DejaVu-SA)

Key Idea and implementation of the DejaVu loss

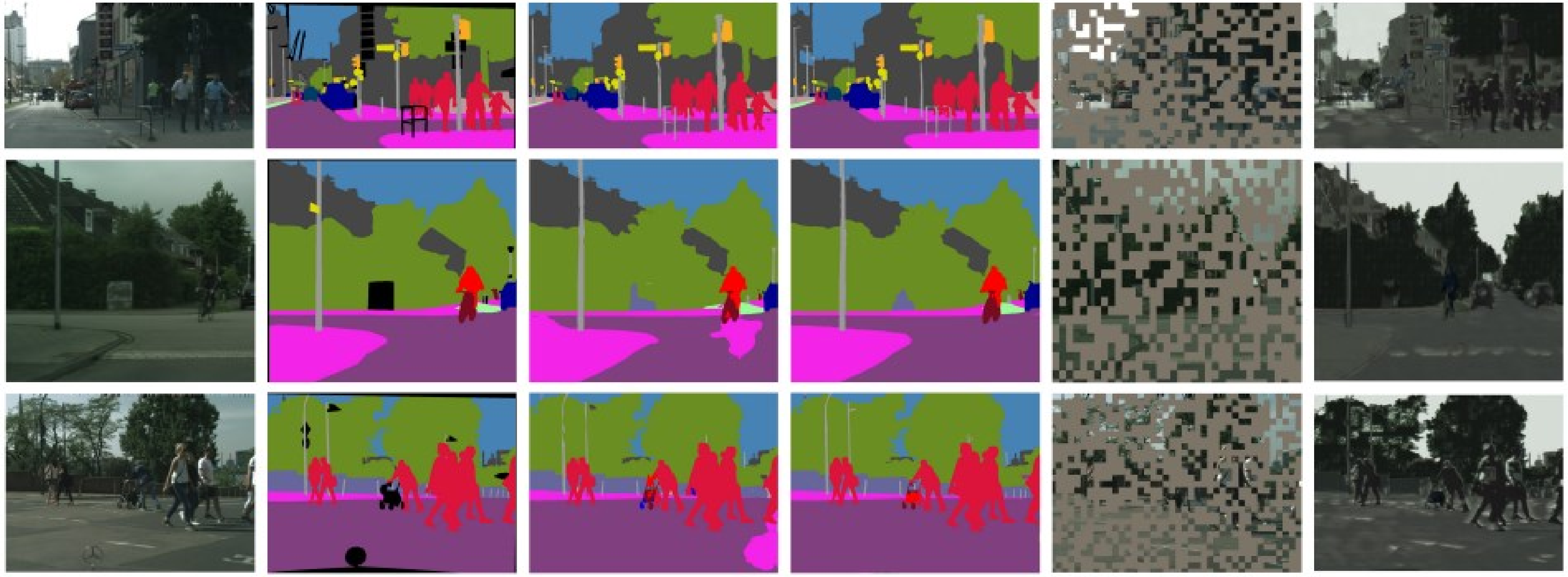


- Shared attention module to capture regeneration context
- Adding computations to the baseline model
- Shared computations to perform generation and the dense task.

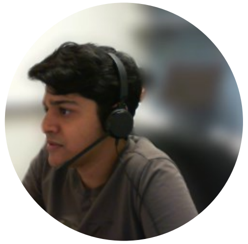
Results



Results: Semantic Segmentation



(a) Input (b) Ground Truth (c) Baseline Prediction (d) Our Prediction (e) Redacted Image (f) Regenerated Image



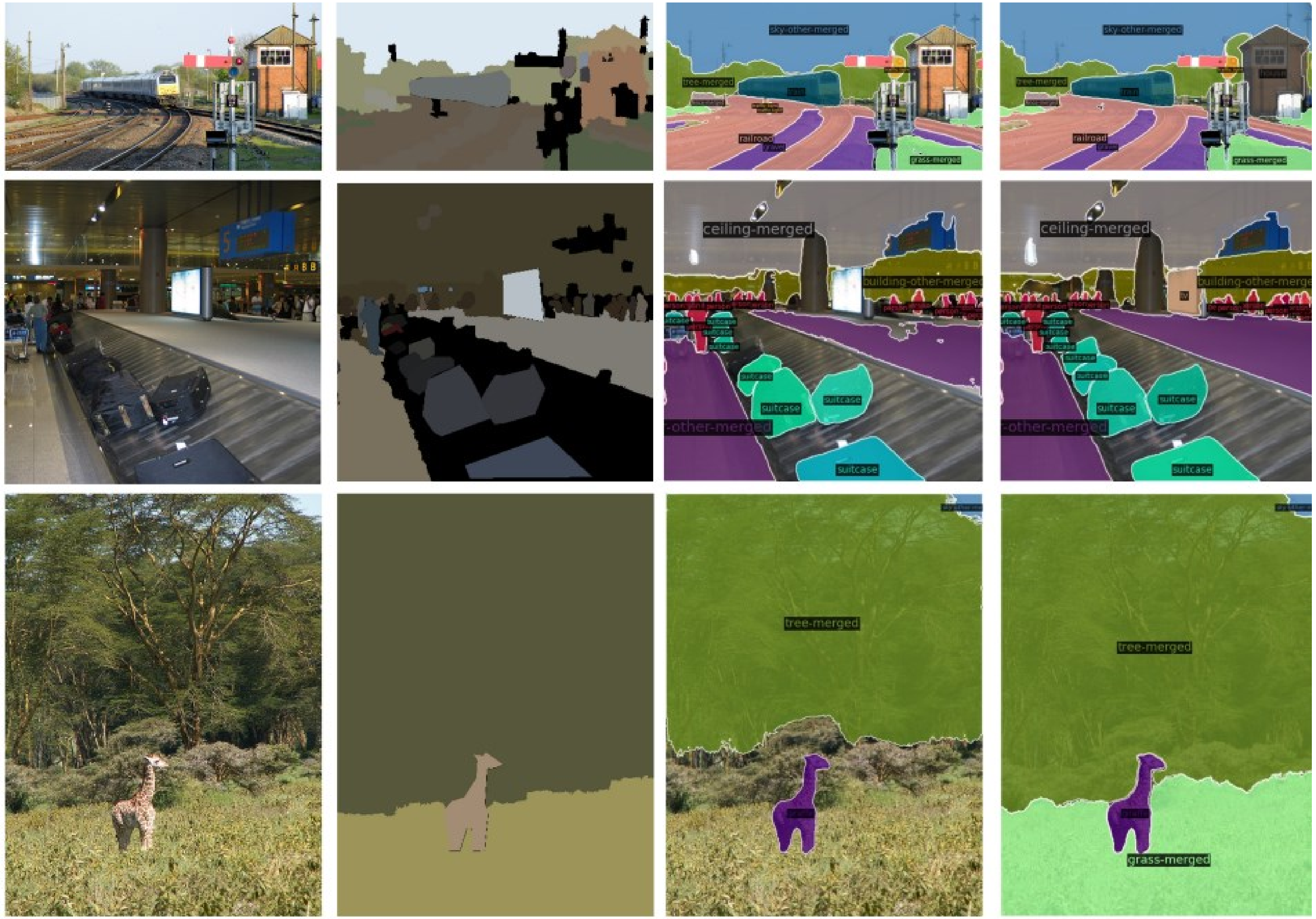
ADE20K

Method	*Backbone*	*mIoU↑*
Semantic FPN [40]	PoolFormer-M48	42.4
+DejaVu	PoolFormer-M48	**43.3**
UperNet [82]	ViT-B [21]	47.4
+DejaVu	ViT-B	48.2
SETR-MLA-DeiT [96]	ViT-B	46.2
Semantic FPN [40]	ViT-B	48.3
DenseCLIP [63]	ViT-B	49.8
+DejaVu	ViT-B	**50.3**
Mask2Former [16]	Swin-L	56.0
+DejaVu	Swin-L	**56.5**

Cityscapes

Backbone	*Method*	*mIoU↑*	*GMacs↓*
HRNet18	HRNet [75]	77.6	19
	+DejaVu	78.8	19
	HS3 [4]	78.1	19
	HS3-Fuse [4]	81.4	39
	OCR [80]	80.7	39
	+DejaVu	**82.0**	39
MiT-B5	Segformer [83]	84.0	362
Swin-L [52]	Mask2Former [16]	83.3	251
	SeMask [34]	84.0	258
ViT	ViT Adapter [14]	84.9	1089
HRNet48	HRNet	84.7	175
	+DejaVu	85.4	175
	OCR	86.1	348
	+DejaVu	86.5	348
	HMS [70]	86.7	893
	+DejaVu	**87.1**	893

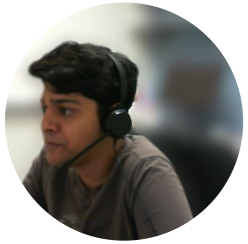
Results: Panoptic Segmentation



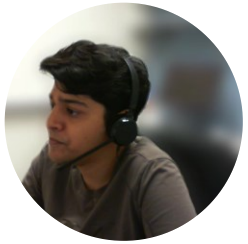
(a) Input (b) Ground Truth (c) Baseline Prediction (d) Our Prediction

MS-COCO

Method	*Backbone*	$PQ\uparrow$	$PQ^{st}\uparrow$	$PQ^{th}\uparrow$
MaX-Deeplab [67]	Max-S	48.4	53.0	41.5
MaskFormer [16]	Swin-T	47.7	51.7	41.7
Mask2Former [15]	Swin-T	53.2	59.3	44.0
+DejaVu	Swin-T	**54.3**	**60.5**	**44.9**
MaX-Deeplab [67]	Max-L	51.1	57.0	42.2
K-Net [82]	Swin-L	54.6	60.2	46.0
MaskFormer [16]	Swin-L	52.7	58.5	44.0
Mask2Former [15]	Swin-L	57.6	64.2	47.5
+DejaVu	Swin-L	**58.0**	**64.4**	**48.3**



Results: Multi-Task Learning



(a) Input (b) Ground Truth (c) Baseline Prediction (d) Our Prediction (e) Redacted Image (f) Regenerated Image

NYUD-v2

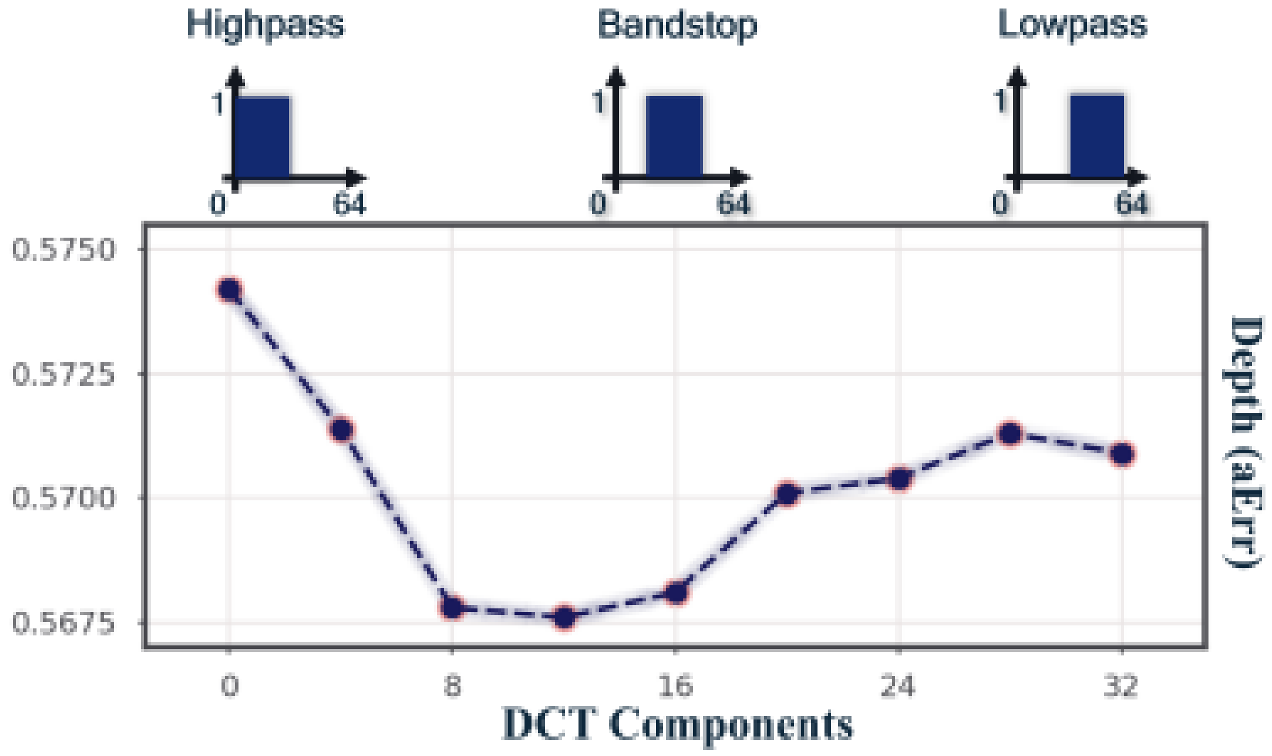
Method	*Seg.(mIoU)↑*	*Depth(aErr)↓*	*Norm(mErr)↓*
MTL [9]	36.95	0.5510	29.51
+DejaVu	**37.40**	**0.5426**	28.74
DWA [50]	36.46	0.5429	29.45
GradNorm [13]	37.19	0.5775	**28.51**
MTAN [50]	39.39	0.5696	28.89
MGDA [66]	38.65	0.5572	28.89
XTC [43]	41.00	0.5148	28.58
+DejaVu	**42.69**	**0.4996**	**27.49**

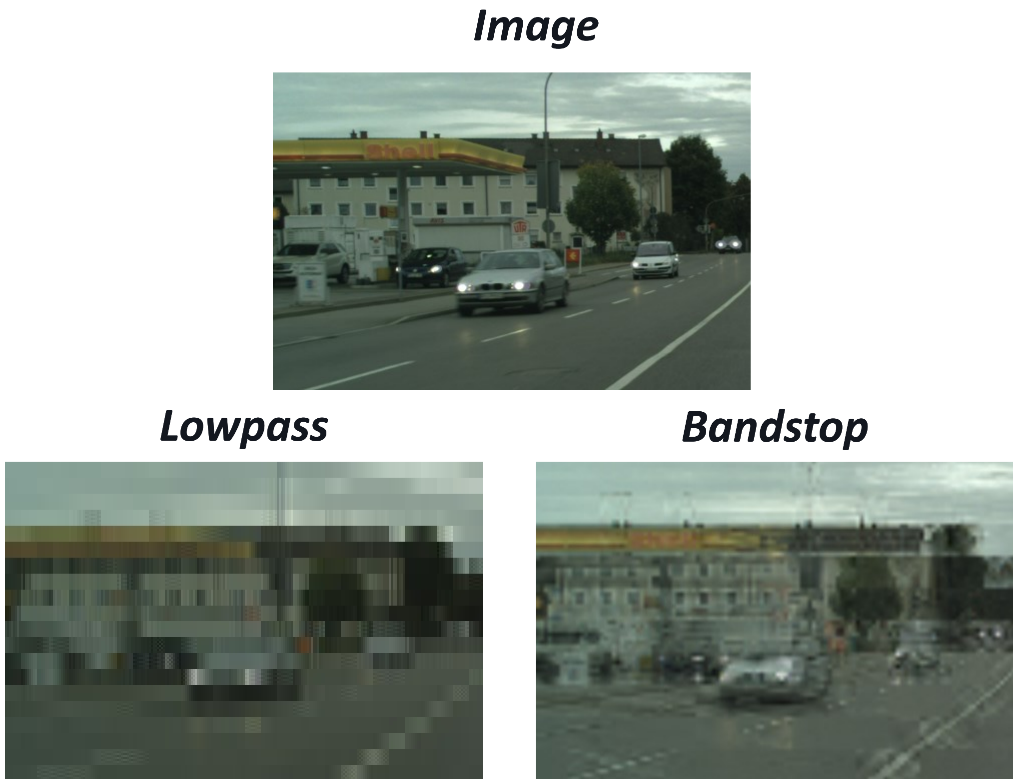
Source:

- Hong Li, Xialei Liu, and Hakan Bilen. Learning multiple dense prediction tasks from partially annotated data. In Proceedings of the IEEE/CVF Conference on Computer Vi sion and Pattern Recognition, pages 18879–18889, 2022.

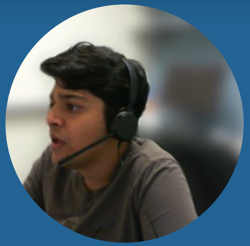
Redacting various bands of spectra

Studying the effect of redacting varying band of spectra for NYUD-v2 depth estimation



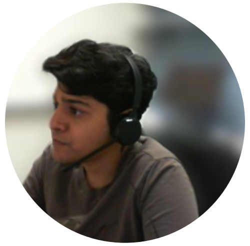


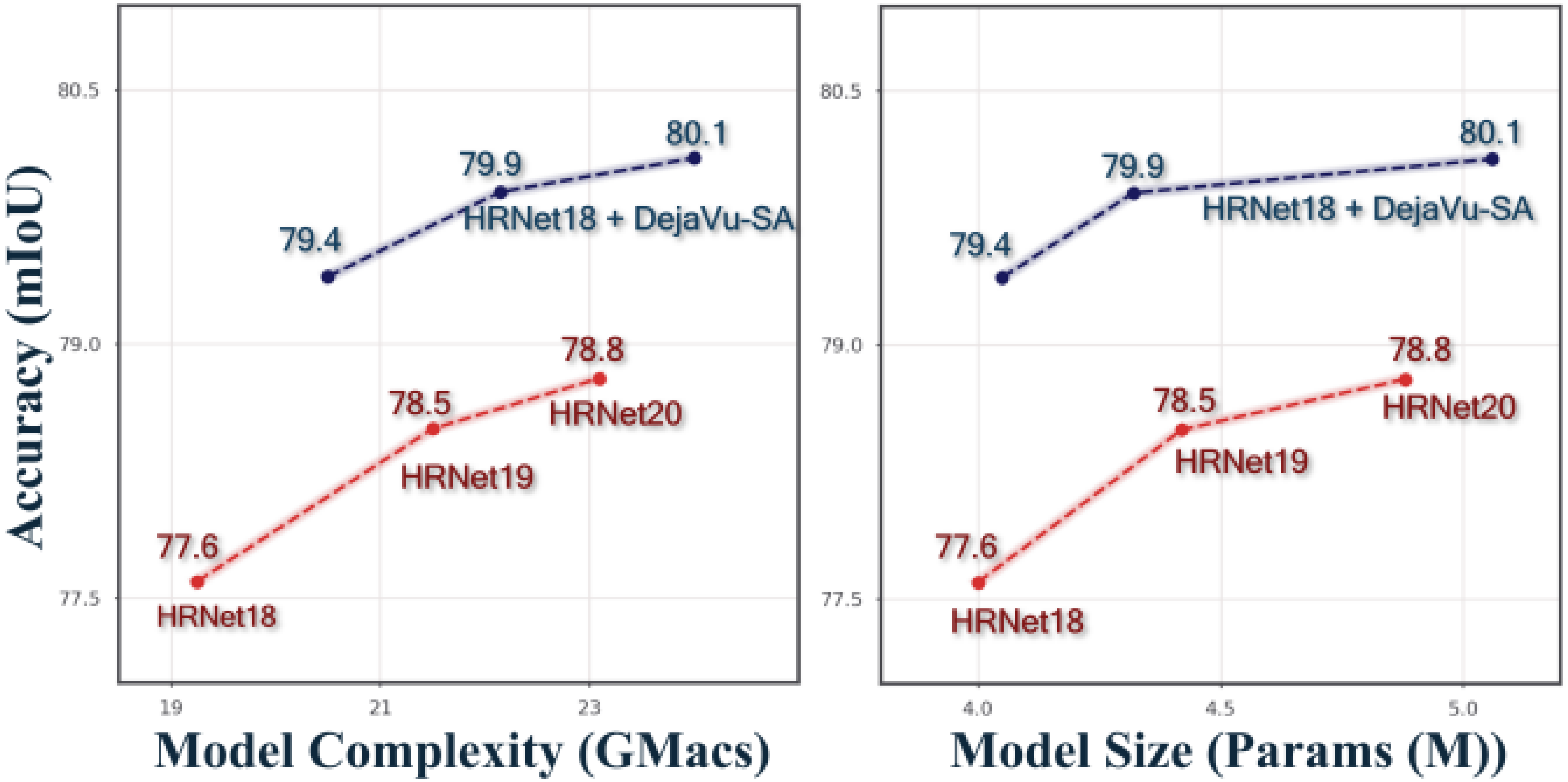
The error is lowest for middle-band redaction as most of the shape information is stored in the middle band



DejaVu-Shared Attention module

Varying the base model size v/s adding DejaVu-SA module





The DejaVu-SA module does increase computations, but provides a better accuracy-computation tradeoff compared to simply scaling up the base model

Motivation

- Regeneration as an auxiliary task

Related Work

- Generative models as an auxiliary task
- Generative models for pre-training

Method

- Redaction types
- Conditional Regeneration module
- DejaVu loss
- DejaVu Shared Attention Module (DV-SA)

Results

- Qualitative/Quantitative analysis
- Accuracy v/s Computation analysis

ArXiv: https://arxiv.org/pdf/2303.01573.pdf

QR:

Qualcomm

Thank you

Follow us on:

For more information, visit us at:

www.qualcomm.com & www.qualcomm.com/blog