DisPositioNet: Disentangled Pose and Identity in Semantic Image Manipulation

Azade Farshad[1,2], Yousef Yeganeh[1], Helisa Dhamo[3], Federico Tombari[1,4], and Nassir Navab[1]

[1]Computer Aided Medical Procedures, Technical University of Munich, Germany
[2]Munich Center for Machine Learning (MCML)
[3]Huawei Noah's Ark Lab, UK
[4]Google, Zurich

Problem Statement

Problem: Semantic Image Manipulation using Scene Graphs

Current Solutions

- Learn Entangled Features for Pose and Appearance
- Complicated and Unpredictable Image Manipulation

Our Contributions:

- **Self-Supervised Disentanglement** of **Pose** and **Appearance**
- **Disentangled Scene Graph** Neural Network
- **Higher Diversity** in Image Manipulation
- **Outperforming SOTA** in Image Generation from Scene Graphs

Our Solution

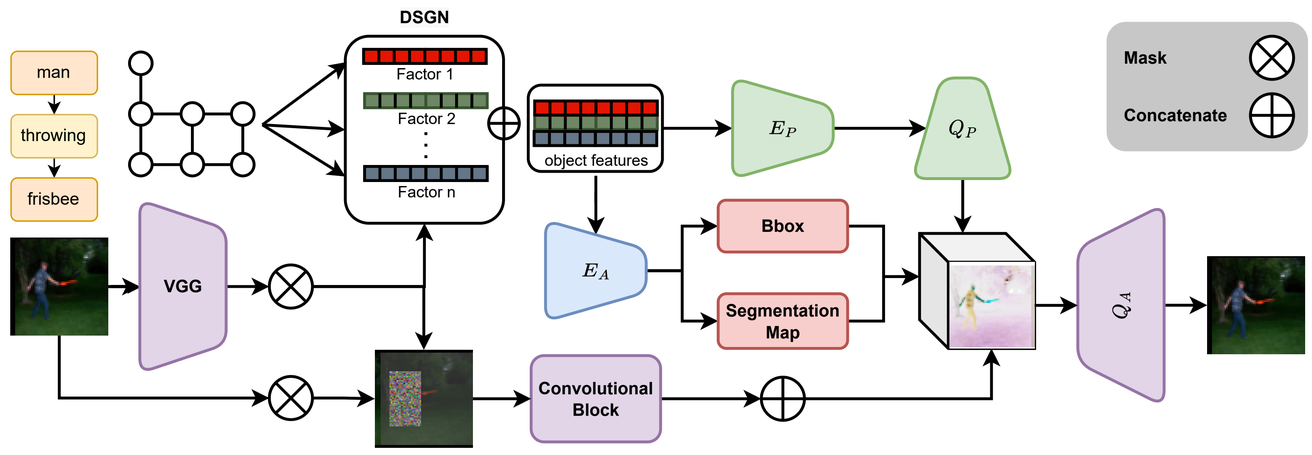

Figure 1: An overview of DispositioNet

Experiments and Results

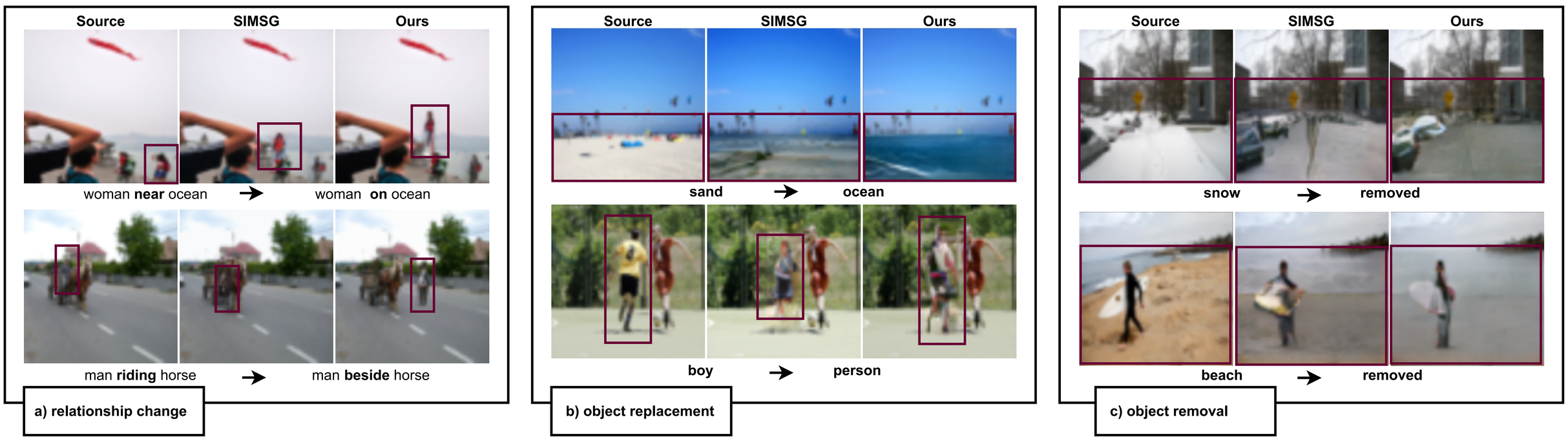

Figure 2: A comparison of our method against SIMSG [1] on VG dataset.

Disentanglement		All pixels			RoI only	
Embeddings	Graph	MAE ↓	SSIM ↑	LPIPS ↓	MAE ↓	SSIM ↑
Generative						
–	–	41.88	34.89	0.27	N/A	N/A
✓	–	41.80	35.18	0.26	N/A	N/A
✓	✓	**41.62**	**35.30**	**0.26**	N/A	N/A
GT Graphs						
–	–	8.61	87.55	0.050	**21.62**	**58.51**
✓	–	8.47	87.53	0.048	21.77	58.30
✓	✓	**8.41**	**87.56**	**0.048**	21.76	58.18
Predicted Graphs						
–	–	13.82	83.98	0.077	28.82	49.34
✓	–	9.65	86.68	0.054	25.62	51.19
✓	✓	**9.39**	**86.91**	**0.052**	**25.40**	**51.85**

Tab. 1: **Ablation Study on Visual Genome.**

Method	Decoder	All pixels					RoI only	
		MAE ↓	SSIM ↑	LPIPS ↓	FID ↓	IS ↑	MAE ↓	SSIM ↑
Generative, GT Graphs								
ISG [2]	Pix2pixHD	46.44	28.10	0.32	58.73	6.64±0.07	N/A	N/A
SIMSG [1]	SPADE	41.88	34.89	0.27	44.27	7.86±0.49	N/A	N/A
DispositioNet (Ours)	SPADE	**41.62**	**35.30**	**0.26**	**40.75**	**7.93**±0.36	N/A	N/A
GT Graphs								
Cond-sg2im [3]	CRN	14.25	84.42	0.081	13.40	11.14±0.80	29.05	52.51
SIMSG [1]	SPADE	8.61	87.55	0.050	**7.54**	**12.07**±0.97	**21.62**	**58.51**
DispositioNet (Ours)	SPADE	**8.41**	**87.56**	**0.048**	7.66	11.65±0.58	21.76	58.18
Predicted Graphs								
SIMSG [1]	SPADE	13.82	83.98	0.077	16.69	10.61±0.37	28.82	49.34
DispositioNet (Ours)	SPADE	**9.39**	**86.91**	**0.052**	**14.42**	**10.69**±0.33	**25.40**	**51.85**

Tab. 2: **Image reconstruction on Visual Genome.**

Conclusion

Our extensive experiments show that DispositioNet:

- improves image generation and manipulation quality.
- generates more diverse images due to the variational representation for object features.
- provides more meaningful and useful features for the disentangled latent embedding using a disentangled GNN for extracting the scene graph features

References

[1] Helisa Dhamo, Azade Farshad, Iro Laina, Nassir Navab, Gregory D Hager, Federico Tombari, and Christian Rupprecht. "Semantic image manipulation using scene graphs". In: *CVPR*. 2020, pp. 5213–5222.

[2] Oron Ashual and Lior Wolf. "Specifying object attributes and relations in interactive scene generation". In: *ICCV*. 2019, pp. 4561–4569.

[3] Justin Johnson, Agrim Gupta, and Li Fei-Fei. "Image Generation from Scene Graphs". In: *CVPR*. 2018.

Check out the Source Code!