Self-Learning Transformations for Improving Gaze and Head Redirection

Yufeng Zheng[1], Seonwook Park[1], Xucong Zhang[1], Shalini De Mello[2], Otmar Hilliges[1]

ETH zürich[1] NVIDIA[2]

NEURAL INFORMATION PROCESSING SYSTEMS

Overview

Many computer vision tasks benefit from the disentanglement of multiple factors, among which

- some are labelled and task-relevant (e.g. gaze direction, pose)
- others are unlabelled and extraneous (e.g. lighting condition)

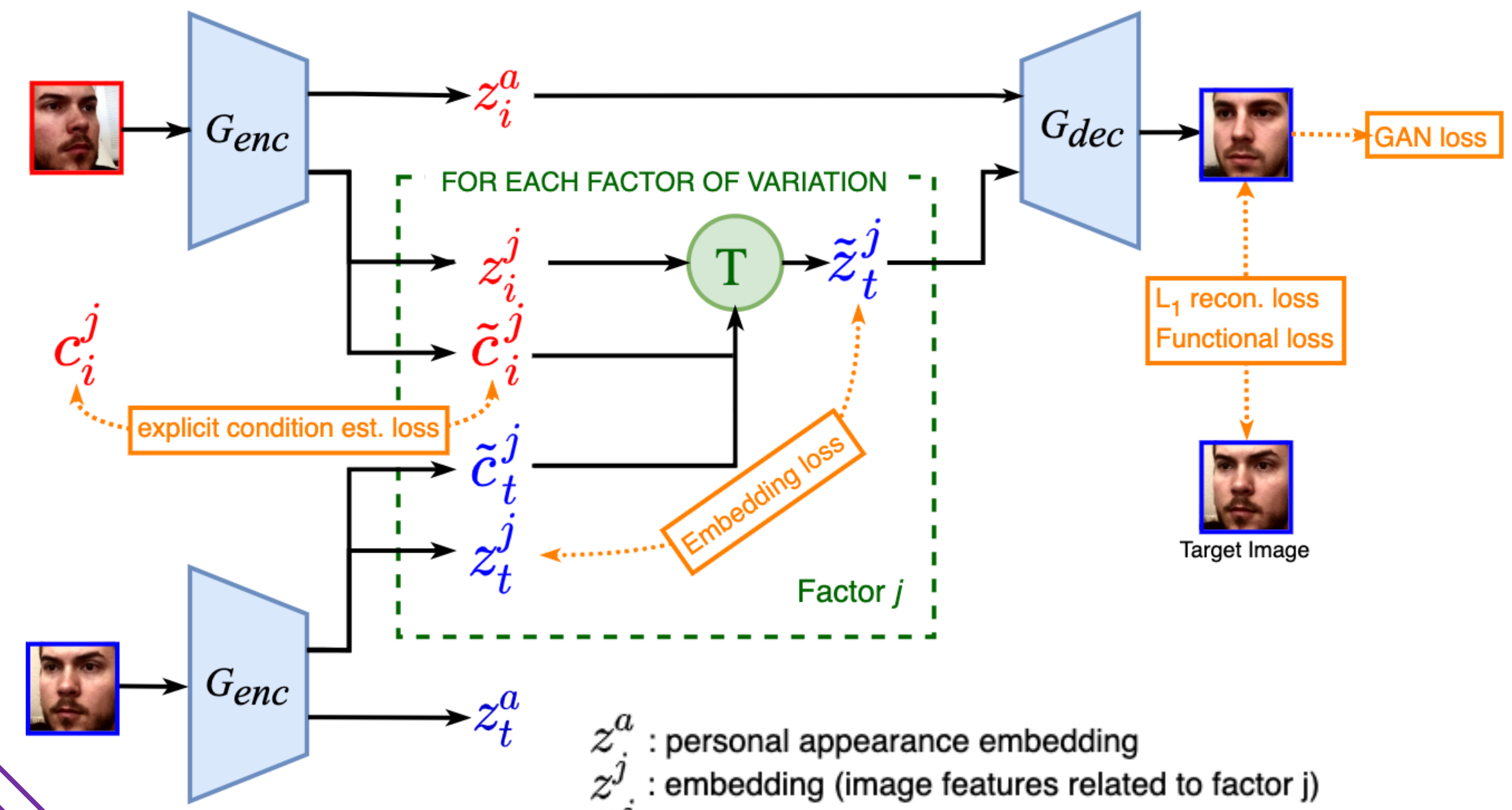
We propose the **Self-Transforming Encoder-Decoder (ST-ED)**, which

- learns to encode extraneous factors in a self-supervised manner,
- disentangles them from task-relevant factors.

We apply our method to gaze and head redirection, and show

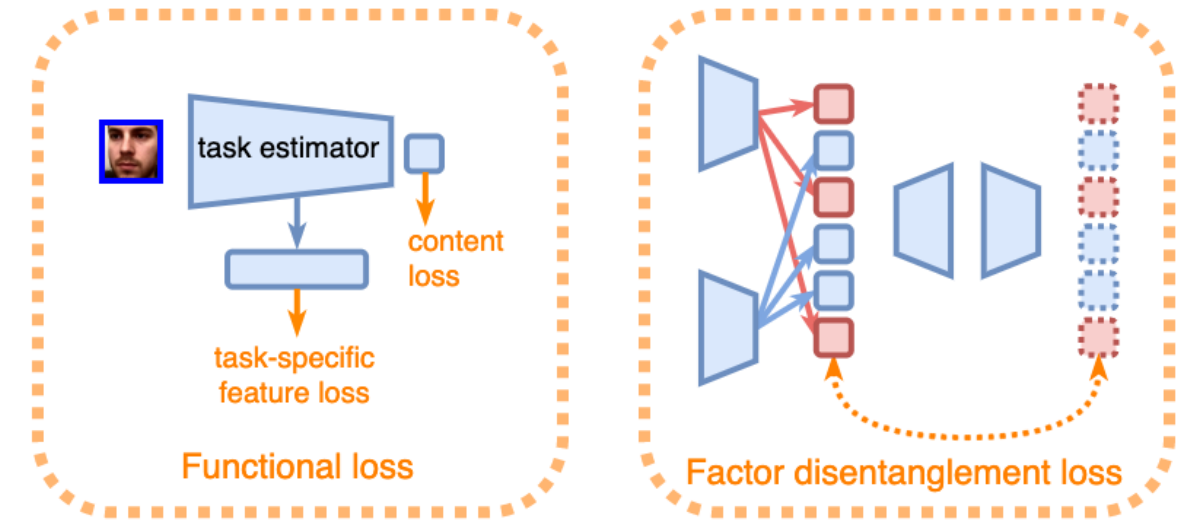
- better perceptual quality and redirection fidelity versus SOTA,
- improved downstream performance when used as data augmentation.

Self-Transforming Encoder-Decoder (ST-ED)



z^a : personal appearance embedding
z^j : embedding (image features related to factor j)
$\tilde{c}^j$: pseudo conditions (low dimensional control of factor j)

Novel Objectives



Rotation-based Transformation

$$\tilde{z}_t^j = T(z_i^j, \tilde{c}_i^j, \tilde{c}_t^j) = R_t^j \left(R_i^j\right)^{-1} z_i^j$$

1 DoF example: $R = \begin{pmatrix} \cos c & -\sin c \\ \sin c & \cos c \end{pmatrix}$

2 DoF example: $R = \begin{pmatrix} \cos\phi & 0 & \sin\phi \\ 0 & 1 & 0 \\ -\sin\phi & 0 & \cos\phi \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & -\sin\theta \\ 0 & \sin\theta & \cos\theta \end{pmatrix}$

Ablation Studies

Approach	Gaze Direction			Head Orientation			LPIPS	
	Re-dir.	$u \rightarrow g$	$h \rightarrow g$	Re-dir.	$u \rightarrow h$	$g \rightarrow h$	$g + h$	all
T-ED Base Model [1]	7.114	-	4.882	2.470	-	0.542	0.279	0.279
ST-ED Model	5.107	-	3.639	1.479	-	0.660	0.272	0.271
$+\boldsymbol{f_u}$	4.716	0.814	3.404	1.434	0.314	**0.385**	0.257	0.215
$+\mathcal{L}_F + \mathcal{L}_D$	**2.195**	**0.507**	**2.072**	**0.816**	**0.211**	0.388	**0.248**	**0.205**

- ST-ED Model predicts and controls with pseudo conditions for explicit factors
- $+\boldsymbol{f}_u$ additionally learns to discover and represent extraneous factors
- $+\mathcal{L}_F + \mathcal{L}_D$ additionally uses functional and factor disentanglement loss.

Comparison to SOTA

	(a) GazeCapture					(b) MPIIFaceGaze				
	Gaze Redir.	Head Redir.	$g \rightarrow h$	$h \rightarrow g$	LPIPS	Gaze Redir.	Head Redir.	$g \rightarrow h$	$h \rightarrow g$	LPIPS
StarGAN	4.602	3.989	0.755	3.067	0.257	4.488	3.031	0.786	2.783	0.260
He *et al.*	4.617	1.392	0.560	3.925	0.223	5.092	1.372	0.684	3.411	0.241
Ours	**2.195**	**0.816**	**0.388**	**2.072**	**0.205**	**2.233**	**0.884**	**0.365**	**1.849**	**0.203**

Qualitative Results

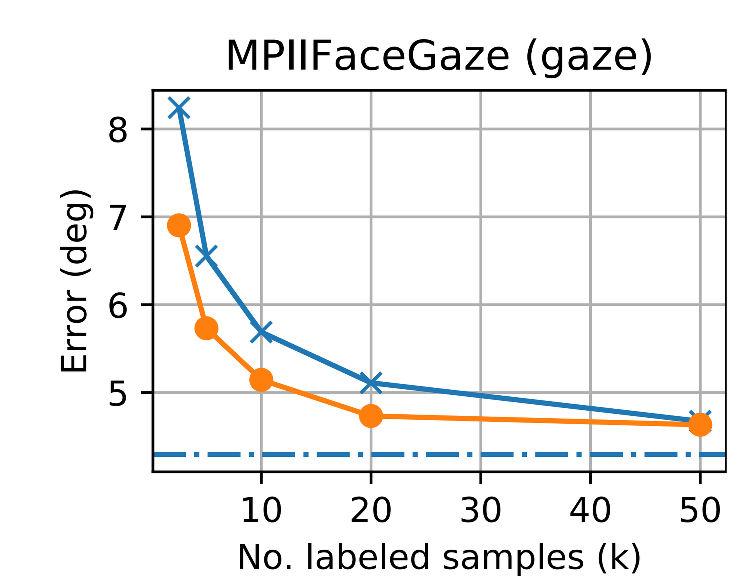
Input image StarGAN[3] He et al.[2] Ours Ours(all) Target image

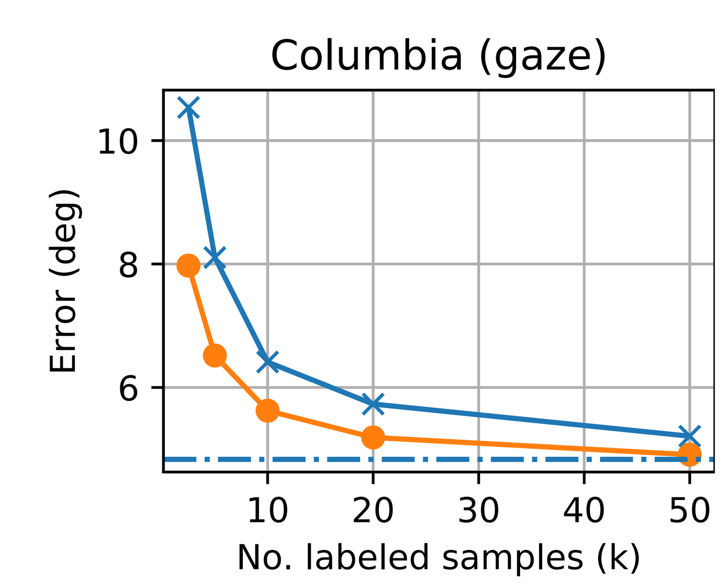
Data Augmentation for Downstream Tasks

Given limited labelled samples,

- Train semi-supervised redirector
- Augment labelled dataset with redirection
- Compare downstream performance

With ½ labelled data, our method achieves similar performance.





Conclusion

- Our ST-ED model learns extraneous factors of variation (unlabeled) from in-the-wild images.
- Our functional and disentanglement losses help to learn more accurate and disentangled factors.
- For the *first time*, we show that augmenting training data with gaze redirector results in improved downstream task performance.

Acknowledgement

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program grant agreement No. StG-2016-717054.

References

[1] Seonwook Park et al. Few-shot adaptive gaze estimation. In ICCV, 2019.
[2] Zhe He et al. Photo-realistic monocular gaze redirection using generative adversarial networks. In ICCV, 2019.
[3] Yunjey Choi et al. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In CVPR, 2018.
[4] Richard Zhang et al. The unreasonable effectiveness of deep features as a perceptual metric. In CVPR, 2018.