Self-Supervised Viewpoint Learning From Image Collections

SK Mustikovela, CVPR 2020

VI Lab 2021 Summer Seminar

Jisu Kim

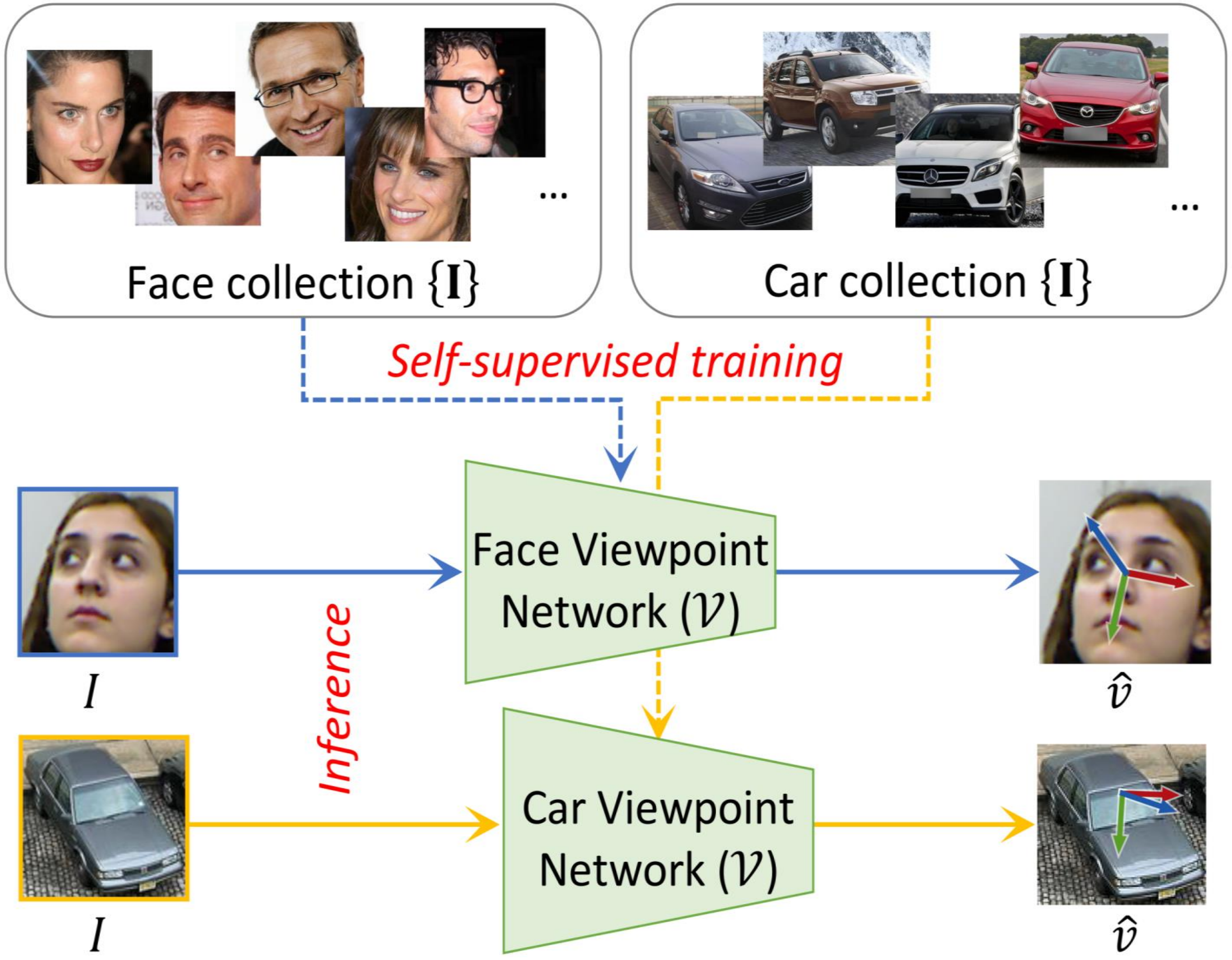
Abstract

Object viewpoint estimation
: Requires large labeled training dataset

Many unlabeled images from the internet

...

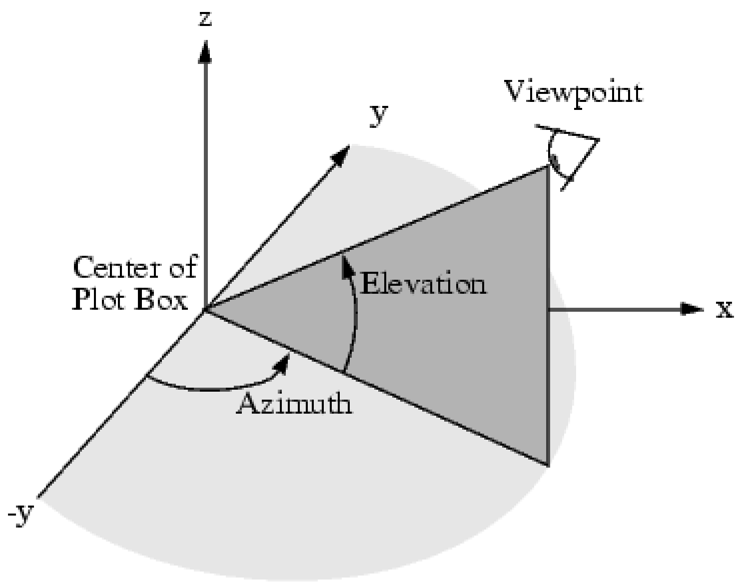
Abstract



Viewpoint estimation
from self-supervised learning

Reconstruct images
in a viewpoint aware manner

Introduction



Object Viewpoint

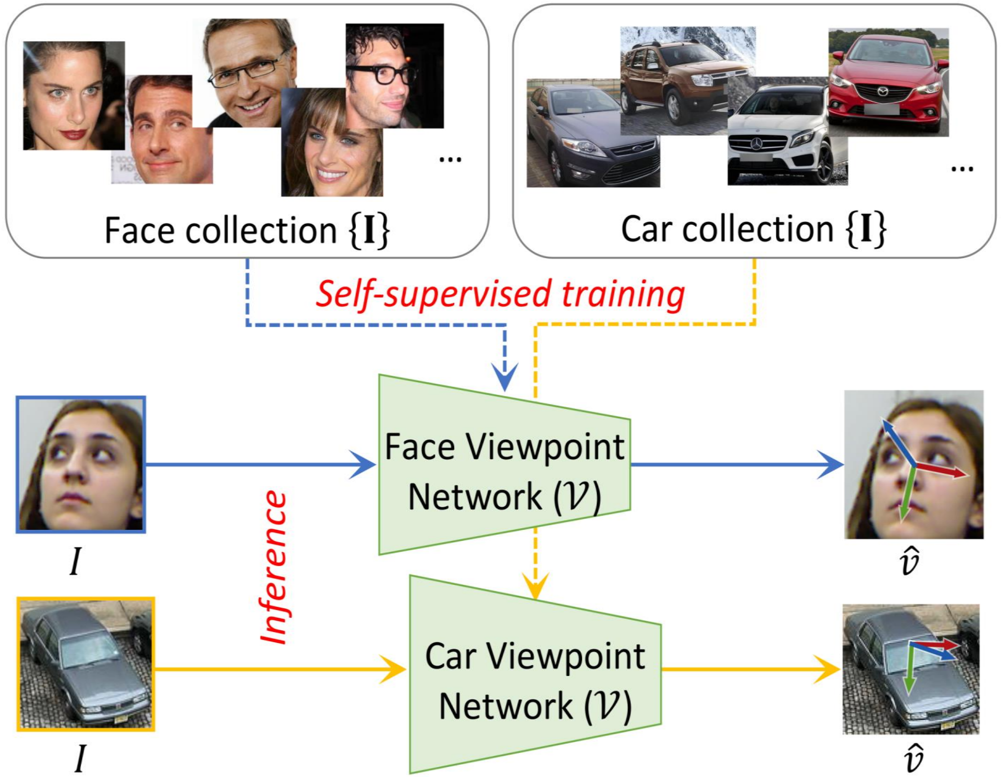
Object viewpoint
: Link between 2D to 3D

Viewpoint estimation
: Requires large-scale human annotated datasets

$\hat{a}$: Azimuth
$\hat{e}$: Elevation
$\hat{t}$: In-plane rotation

http://matlab.izmiran.ru/help/techdoc/visualize/chview3.html

Introduction



Self-supervised learning

No GT image

No human annotation

=> Viewpoint estimation network with self-supervise learning

=> Constraints to use only image collection

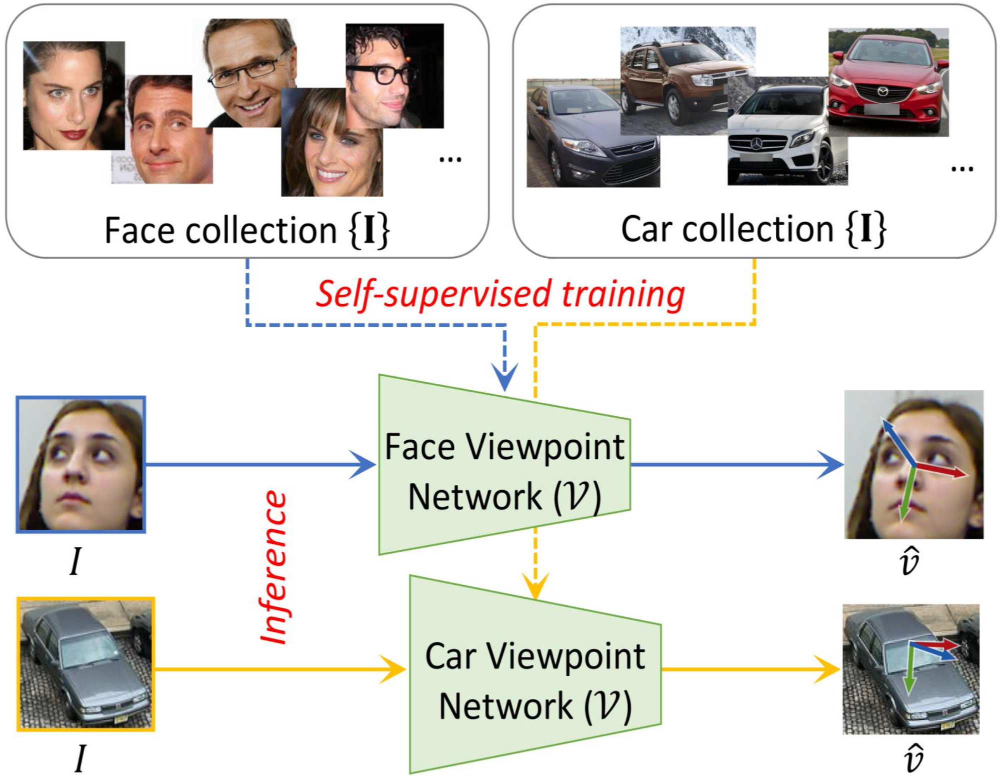
=> Comparison to fully-supervised approaches

Related Works

Viewpoint estimation
: Requires object viewpoint annotations during training
: Uses large annotated datasets
: Data augmentation with synthetic images

Self-supervised object attribute discovery
: Heavy reliance on differentiable rendering
: No prior works propose to learn viewpoint in self-supervised manner

Self-Supervised Viewpoint Learning

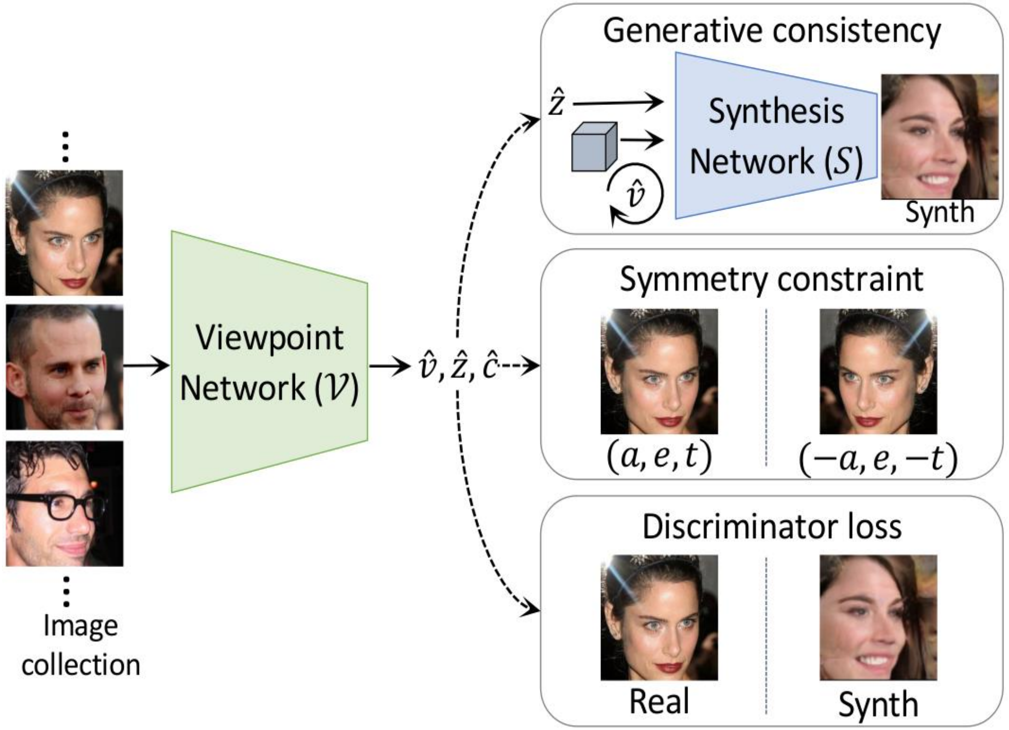


$\{I\}$: Image collection without annotation

V : Viewpoint estimation network

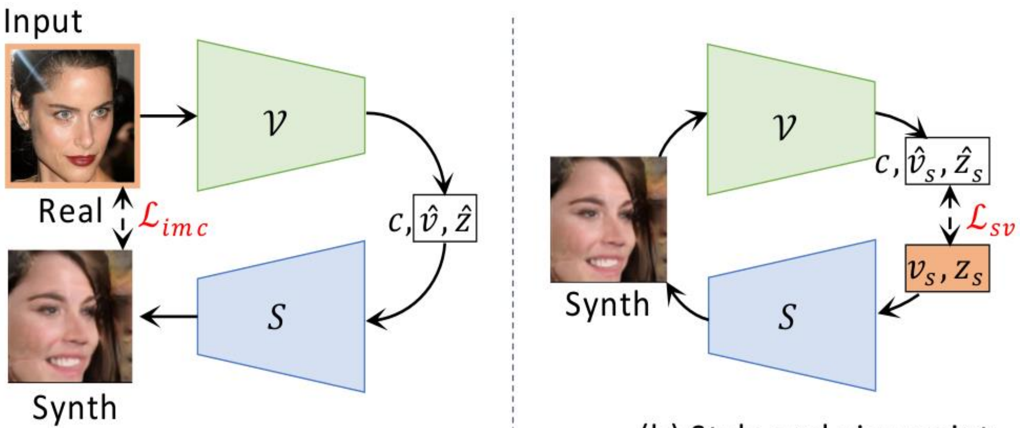
$\hat{v}$: Object viewpoint

Self-Supervised Viewpoint Learning



Learn the viewpoint network V with self-supervised manner

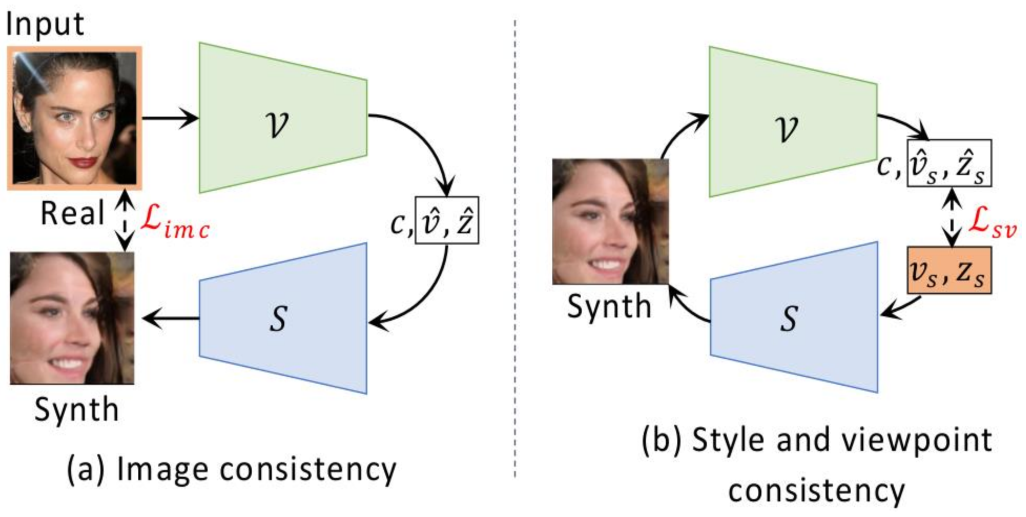
Use synthetic image as pseudo-GT (Cyclic structure)



(a) Image consistency

(b) Style and viewpoint consistency

Generative Consistency



$$\mathcal{L}_{imc} = 1 - \langle \Phi(I), \Phi(\hat{I}_s) \rangle$$
$$\mathcal{L}_{sv} = \|z_s - \hat{z}_s\|_2^2 + \mathcal{L}_v(v_s, \hat{v}_s)$$

(a) Ensure that V generalizes well to real images as well
(b) Learns correct viewpoints for synthetic images

Discriminator Loss & Symmetry Constraint

Discriminator

: Predicts a score $\hat{c}$ indicating whether an input image is real or synthetic

Symmetry Constraint

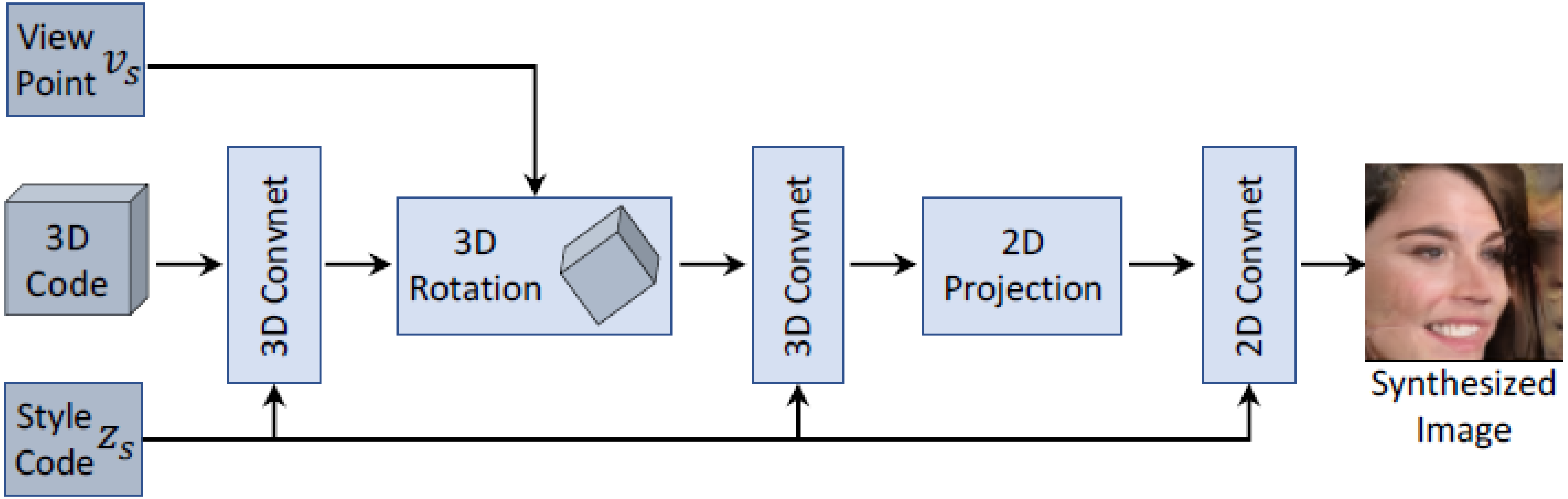
: Strong prior observed in many commonplace object categories

: $\hat{v} = (\hat{a}, \hat{e}, \hat{t}) \mid \hat{v}^* = (\hat{a}^*, \hat{e}^*, \hat{t}^*) \mid \hat{v}_f^{\wedge} = (-\hat{a}^*, \hat{e}^*, -\hat{t}^*)$

: $\mathcal{L}_{sym} = \mathcal{D}(\hat{v}, \hat{v}_f^*) + \|\hat{z} - \hat{z}^*\|_2^2$

: Enforces that the style of the flipped image pair is consistent

Viewpoint-Aware Synthesis Network



Loss functions

$$\mathcal{L}_{adv} = -\mathbb{E}_{\hat{x} \sim p_{\mathrm{synth}}}[\hat{c}]$$

Wasserstein GAN Loss

$$\mathcal{L}_{sv} = \|z_s - \hat{z}_s\|_2^2 + \mathcal{L}_v(v_s, \hat{v}_s)$$

Viewpoint & Style Consistency Loss

$$\mathcal{L}_{sym} = \mathcal{D}(\hat{v}, \hat{v}_f^*) + \|\hat{z} - \hat{z}^*\|_2^2$$

Flip Consistency Loss

Experiments

	Method	Azimuth	Elevation	Tilt	MAE
Self-Supervised	LMDIS [76] + PnP	16.8	26.1	5.6	16.1
	IMM [24] + PnP	14.8	22.4	5.5	14.2
	SCOPS [22] + PnP	15.7	13.8	7.3	12.3
	HoloGAN [42]	8.9	15.5	5.0	9.8
	HoloGAN [42] with v	7.0	15.1	5.1	9.0
	SSV w/o $\mathcal{L}_{sym} + \mathcal{L}_{imc}$	6.8	13.0	5.2	8.3
	SSV w/o $\mathcal{L}_{imc}$	6.9	10.3	**4.4**	7.2
	SSV-Full	**6.0**	**9.8**	**4.4**	**6.7**
Supervised	3DDFA [79]	36.2	12.3	8.7	19.1
	KEPLER [33]	8.8	17.3	16.2	13.9
	DLib [29]	16.8	13.8	6.1	12.2
	FAN [4]	8.5	7.4	7.6	7.8
	Hopenet [51]	5.1	6.9	3.3	5.1
	FSA [70]	**4.2**	**4.9**	**2.7**	**4.0**

300W-LP Dataset
without GT viewpoint annotation

Calculate absolute error
between prediction and GT

Image Results

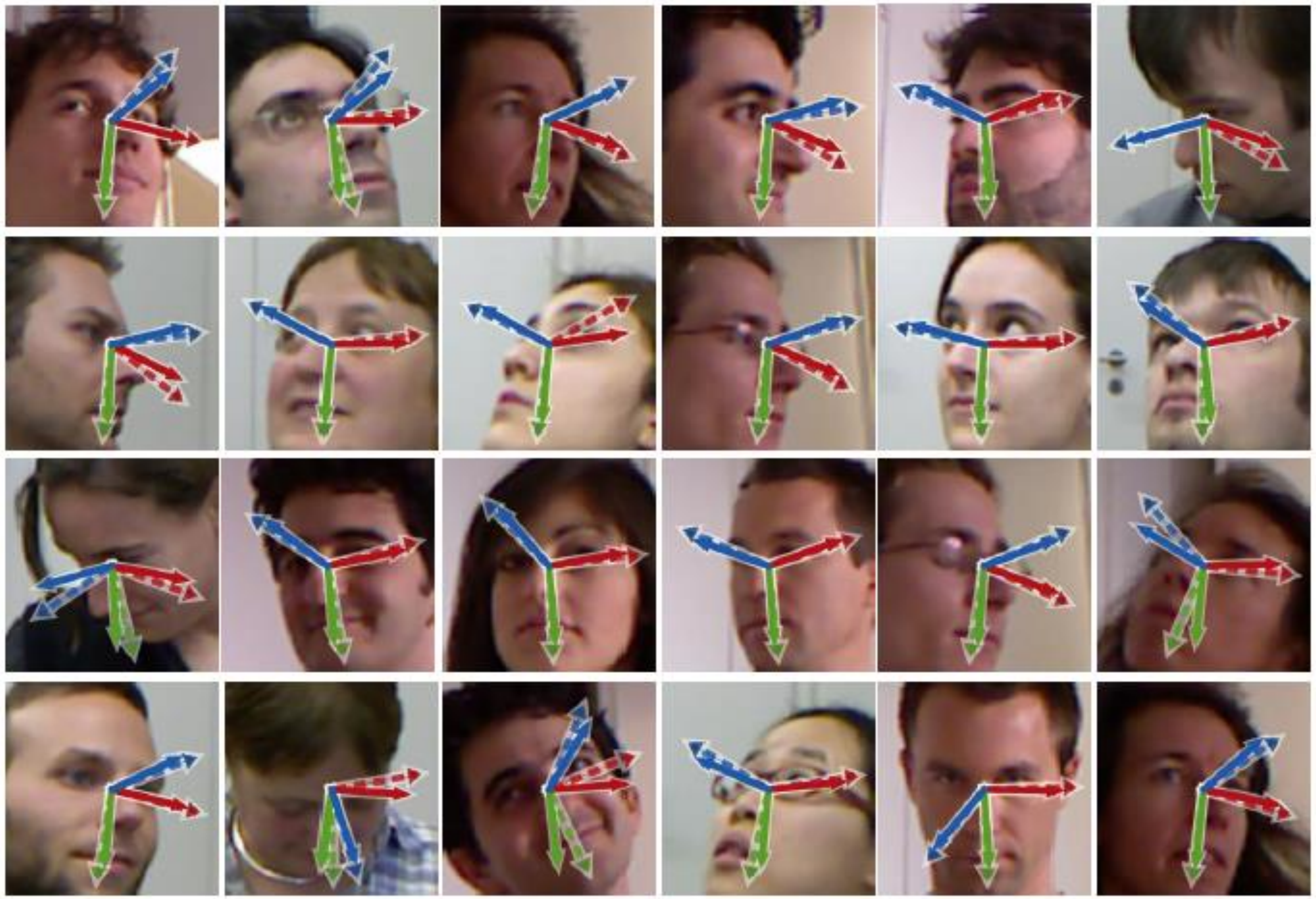


Image Results

Image Results

Image Results

Conclusion

Viewpoint estimation of unannotated object images with self-supervised learning

Viewpoint-aware synthesis network and additional constraints

Self-supervised techniques that performs competitively to fully-supervised ones