SELF-SUPERVISED PART AND VIEWPOINT DISCOVERY FROM IMAGE COLLECTIONS

Varun Jampani

Google Research

ECCV 2020 Tutorial on

New Frontiers for Learning with Limited Labels or Data

Image collections are quite common

- Image collection: Set of images with common object category of interest.
- Examples include image search results, photo collections, tourist pictures of a landmark etc.

Car image collection

Face image collection

Object understanding from image collections

Self-supervised learning of object properties from image collections

Object properties:

- *Geometry*: 3D shape, 3D viewpoint etc.
- *Semantics*: Keypoints, part segmentation, bounding boxes etc.
- *Material properties*: Diffuse albedo, specularities, roughness etc.

Object understanding from image collections

Self-supervised learning of object properties from image collections

Object properties:

- *Geometry*: 3D shape, 3D viewpoint etc.
- *Semantics*: Keypoints, part segmentation, bounding boxes etc.
- *Material properties*: Diffuse albedo, specularities, roughness etc.

Self-supervised Co-Part Segmentation

CVPR 2019

Wei-Chih Hung, Varun Jampani, Sifei Liu, Pavlo Molchanov, Ming-Hsuan Yang, Jan Kautz

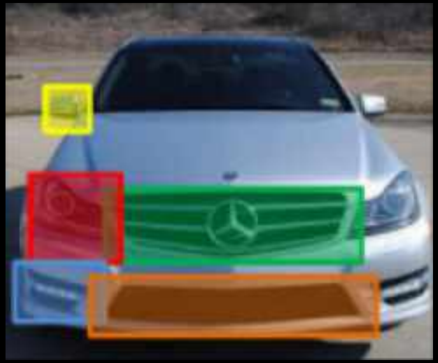
Some slides credit: Wei-Chih Hung

Our Goal

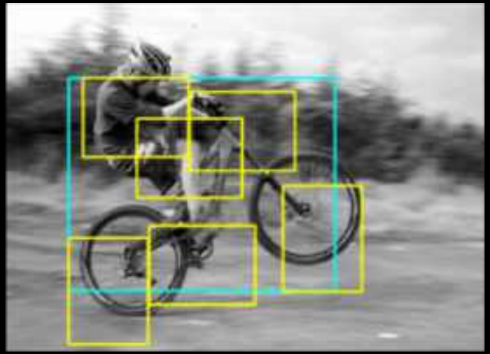
Learn part segmentation from image collection

Why Parts?

- Parts are relatively stable with respect to object deformations.
- Can provide reliable mid-level correspondences between images.
- Useful for several high-level vision tasks.



Fine-grain recognition

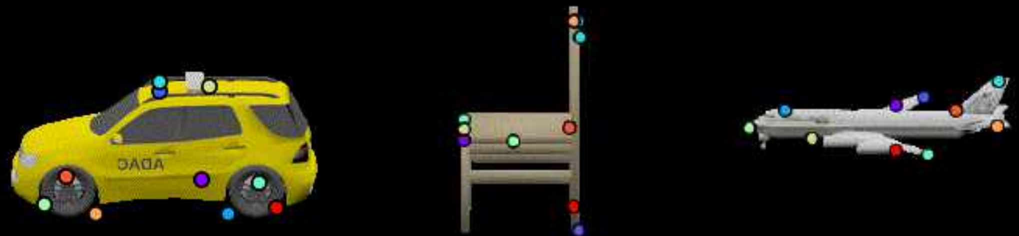


Object detection

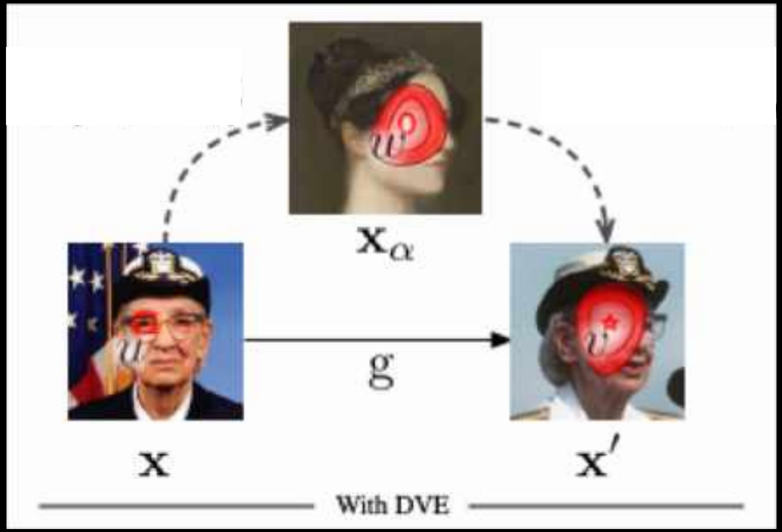
Pose estimation

3D reconstruction

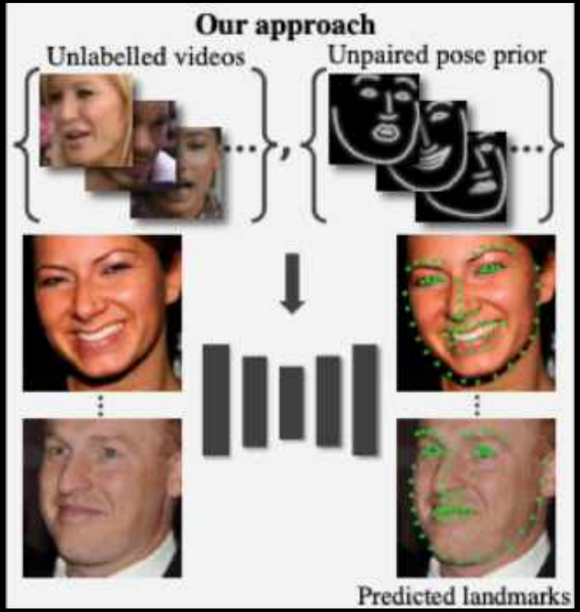
Keypoint Discovery



KeypointNet [1]



Descriptor vector exchange [2]



Learning from videos [3]

1. Suwajanakorn, S., et al. “Discovery of latent 3d keypoints via end-to-end geometric reasoning.” *NeurIPS 2018*
2. Thewlis, J., et al., “Unsupervised learning of landmarks by descriptor vector exchange.” *ICCV 2019*
3. Jakab, T., et al., “Self-supervised Learning of Interpretable Keypoints from Unlabelled Videos.” *CVPR 2020*

Part Segmentation vs. Keypoints

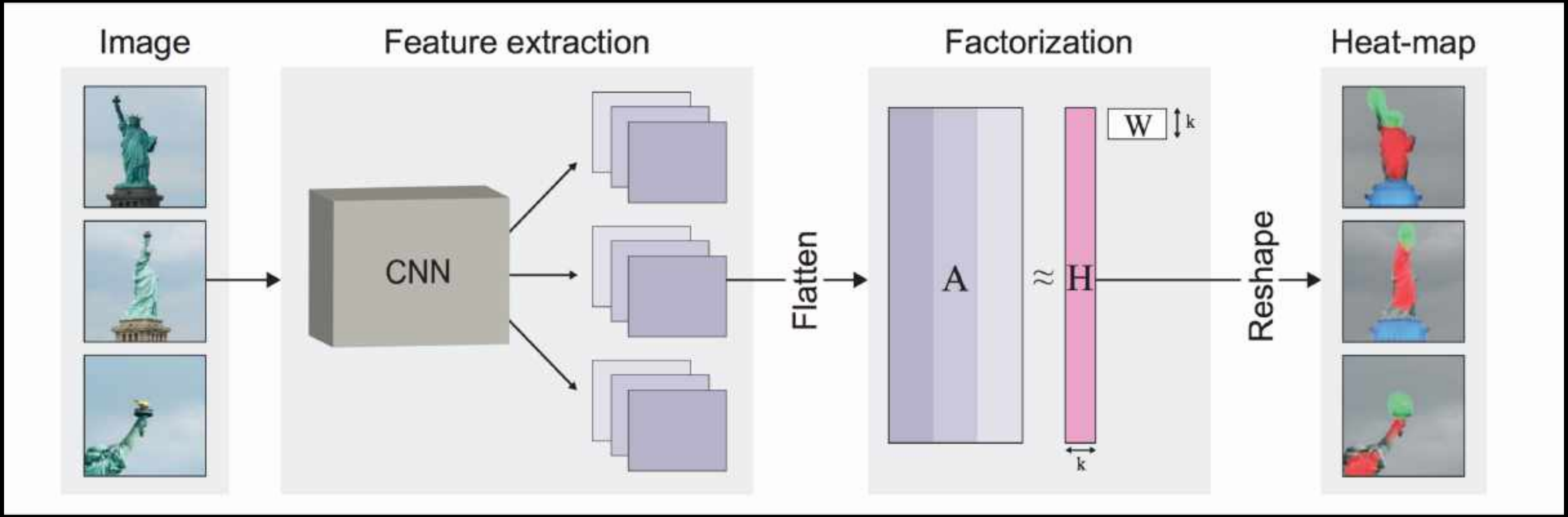
- Part segmentation
  - provides both localization and segmentation of parts
  - can represent disjoint regions as a single part

Parts from [1]

Keypoints from MAFL [2] and 300W [3]

1. https://ai.googleblog.com/2018/03/mobile-real-time-video-segmentation.html
2. Zhang et al. "Facial landmark detection by deep multi-task learning." *ECCV 2014*
3. Sagonas et al. "300 faces in-the-wild challenge: The first facial landmark localization challenge." *ICCV Workshops 2013*

Deep Feature Factorization (DFF) [1]



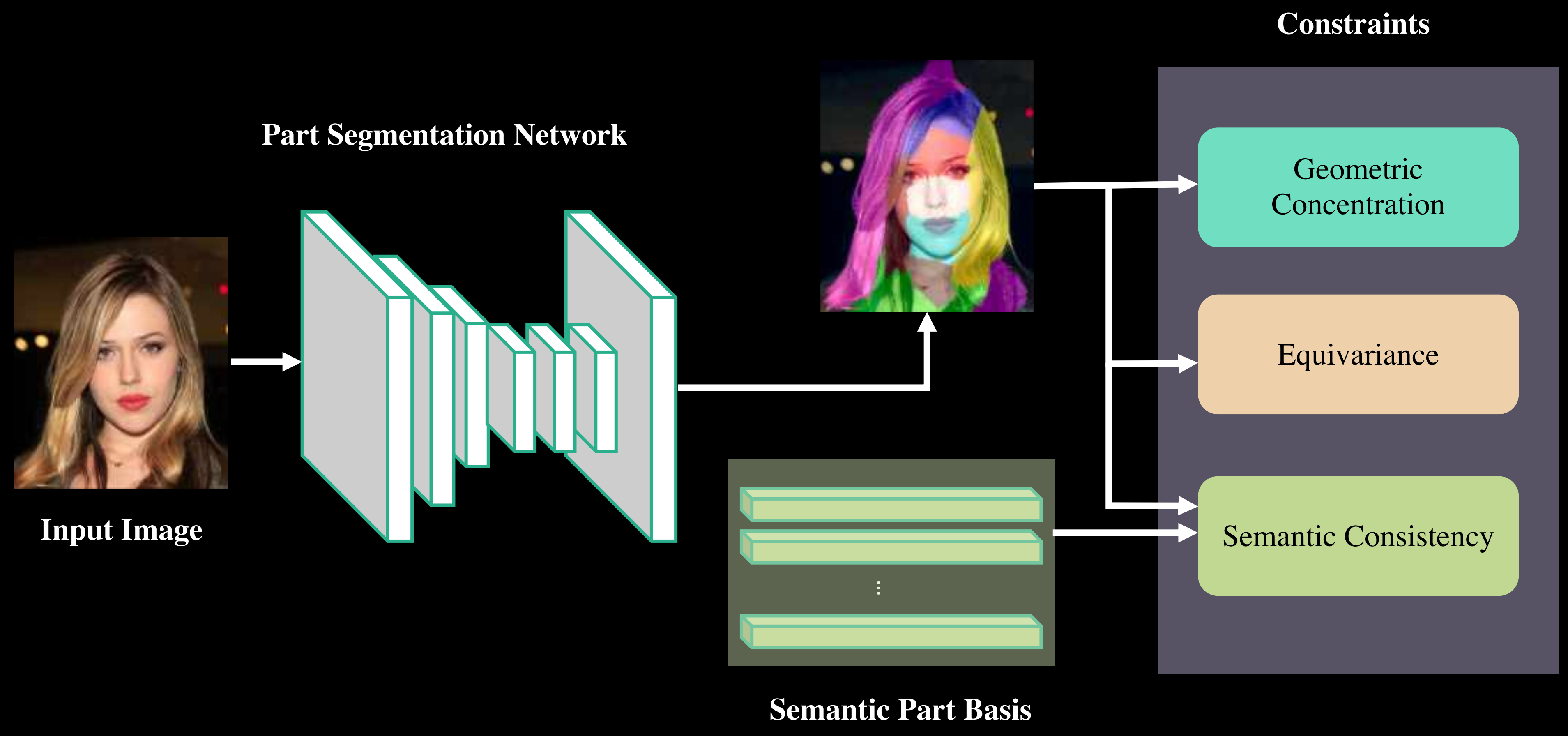
- Apply non-negative matrix factorization (NMF) solver on all images
  - Difficult to scale to large datasets
  - Not easy to apply other constraints

1. Collins et al. "Deep feature factorization for concept discovery." *ECCV* 2018.

Properties of Good Part Segmentation

- ***Geometric concentration***:
  - Parts are concentrated geometrically and form connected components
- ***Robustness to variations***:
  - Part segments are robust with respect to object deformations
- ***Semantic consistency***:
  - Part segments should be semantically consistent across different object instances with appearance and pose variations
- ***Objects as union of parts***:
  - Parts appear on objects (not background) and the union of parts forms an object

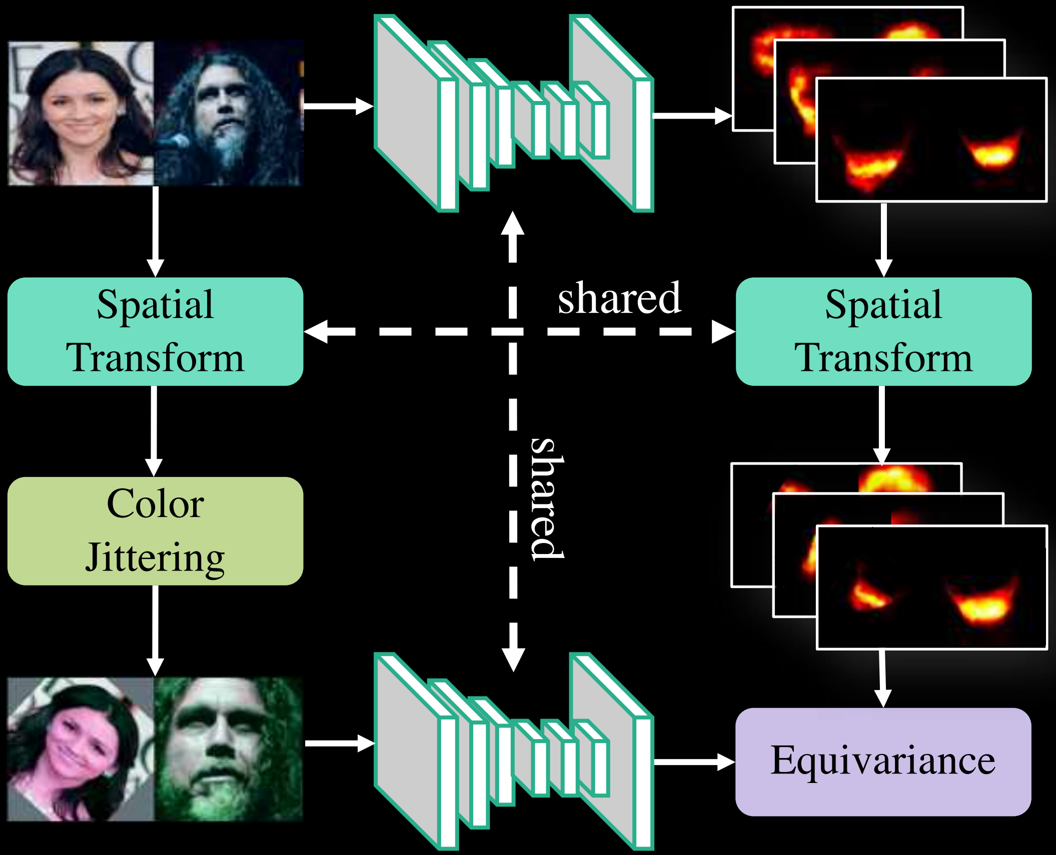
SCOPS Framework



Geometric Concentration

Most part pixels are locally concentrated

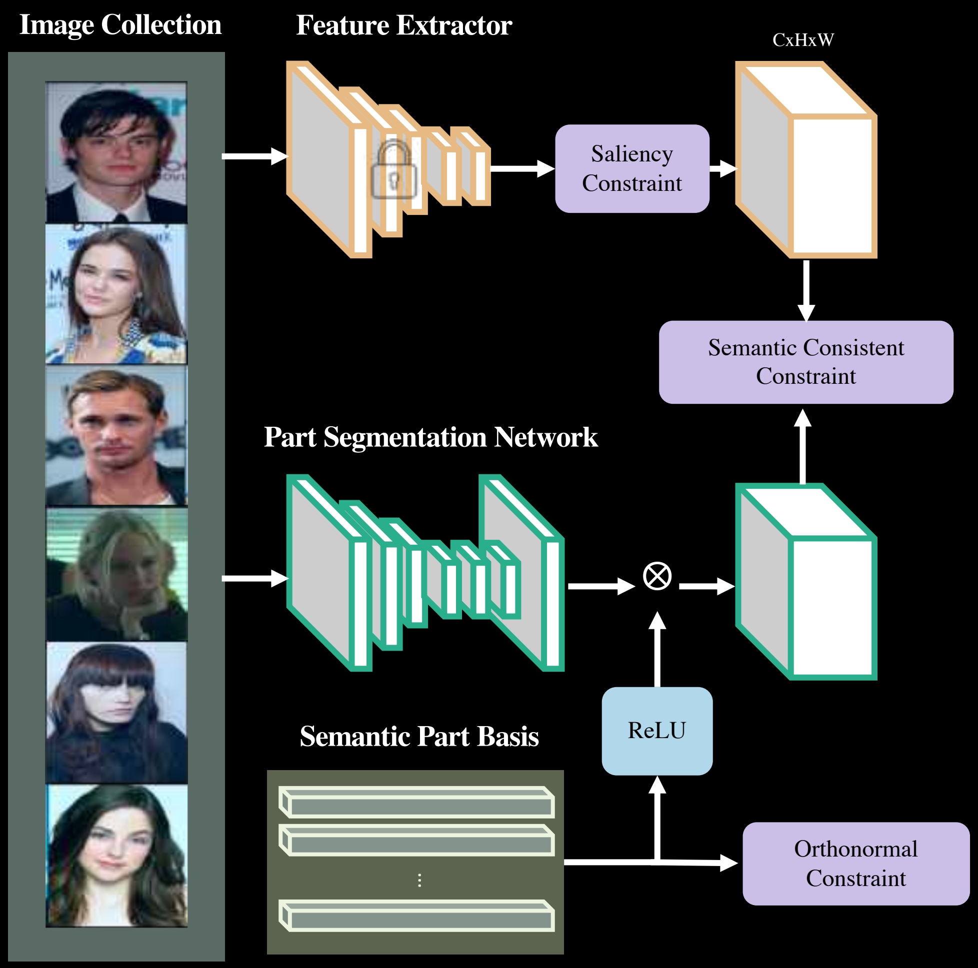
Equivariance



w/o Equivariance

w/ Equivariance

Semantic Consistency



w/o Sematic Consistency

w/ Sematic Consistency

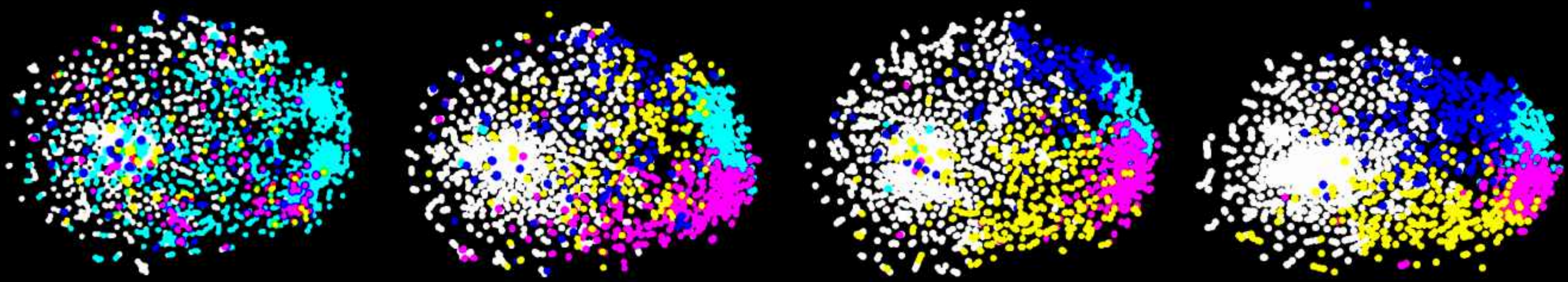
Progression through training

100 iterations

1000 iterations

3000 iterations

50000 iterations



Part features across all images – TSNE. visualization

Results on faces (Unaligned CelebA)

Landmark Estimation Error

Method	Error (%)
ULD (K=8)	40.82
DFF (K=8)	31.30
SCOPS (K=4)	21.76
SCOPS (K=8)	15.01

Image

Unsupervised Landmark Detection (ULD)

Deep Feature Factorization (DFF)

SCOPS (Ours)

Results on birds (CUB)

Image

Unsupervised Landmark Detection (ULD)

Deep Feature Factorization (DFF)

SCOPS (Ours)

Results on Pascal-Part (Horse)

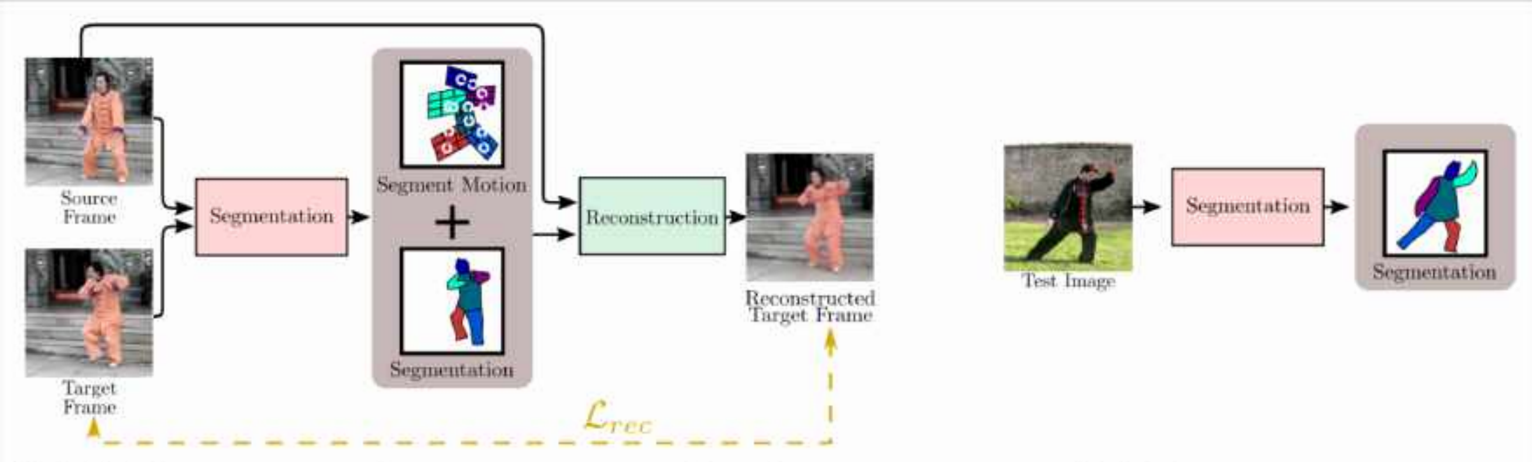
Image

Deep Feature Factorization (DFF)

SCOPS (Ours)


PRINCESS
DRAGON

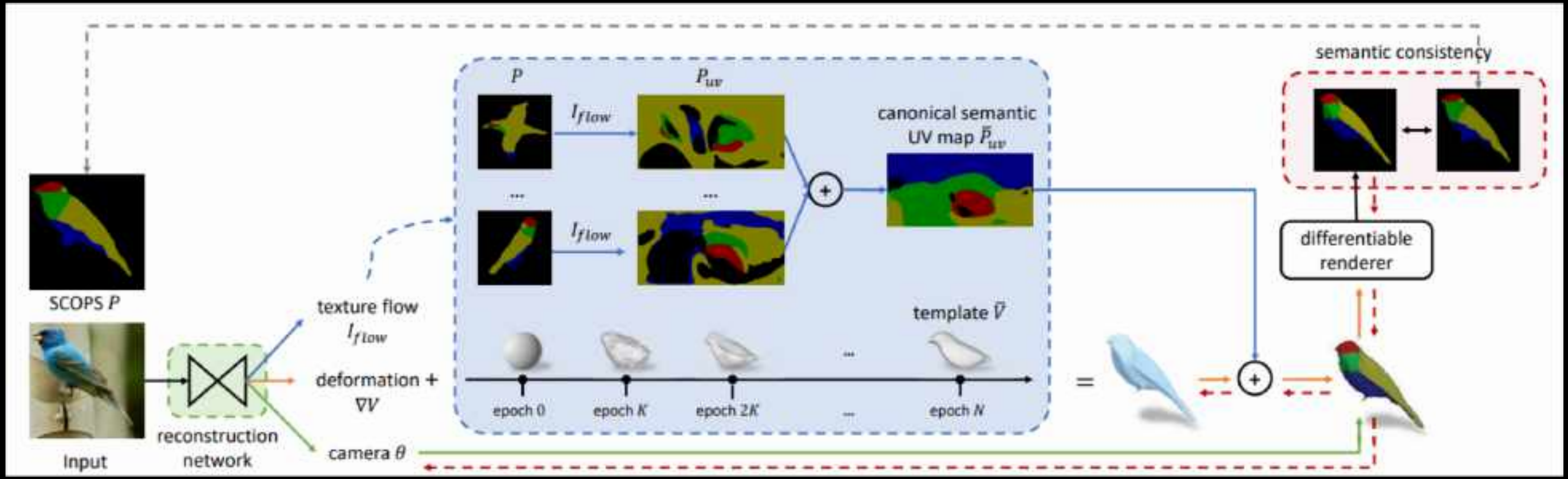
Motion-supervised Co-Part Segmentation [1]



(a) Self-supervised training on video frames (b) Test on images

1. Siarohin et al. "Motion-supervised Co-Part Segmentation." arXiv 2020

Learning 3D shapes via part discovery



1. Li et al. “Self-supervised Single-view 3D Reconstruction via Semantic Consistency.” *ECCV 2020*

Part discovery: Remarks

- Part discovery with self-supervised constraints
- To avoid degenerate solutions and to constrain solution space
  - Leverage part properties such equivariance and geometric concentration
  - Semantic consistency - Leverage hidden consistencies in classification features
- Useful for higher level vision tasks such as object reconstruction

Self-supervised viewpoint learning from image collections

CVPR 2020

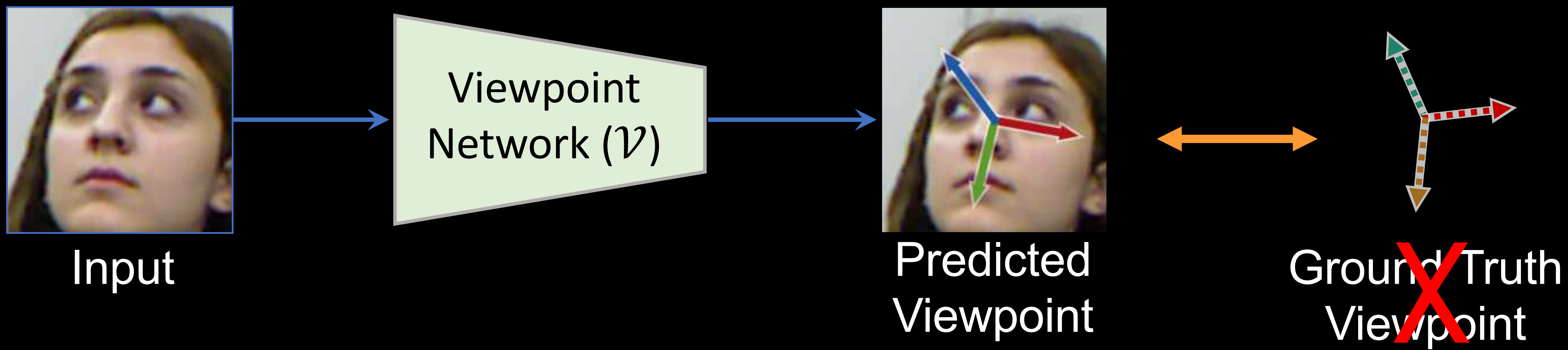
Siva Kumar Mustikovela, Varun Jampani, Shalini De Mello, Sifei Liu, Umar Iqbal, Carsten Rother, Jan Kautz

Some slides credit: Siva Kumar Mustikovela

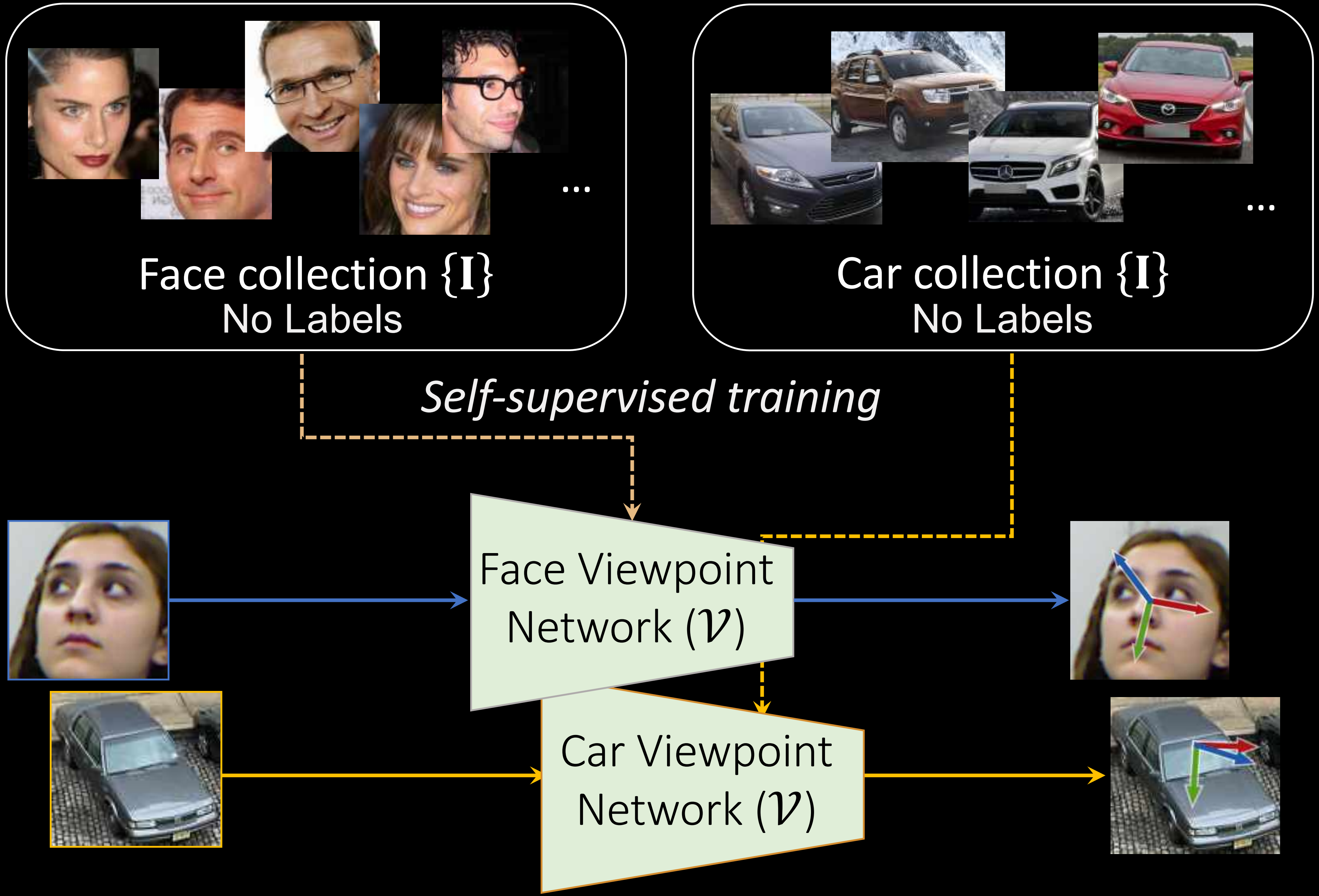
Viewpoint Annotation is Hard

- Hard to align 3D CAD models
- Error prone
- Time consuming and expensive

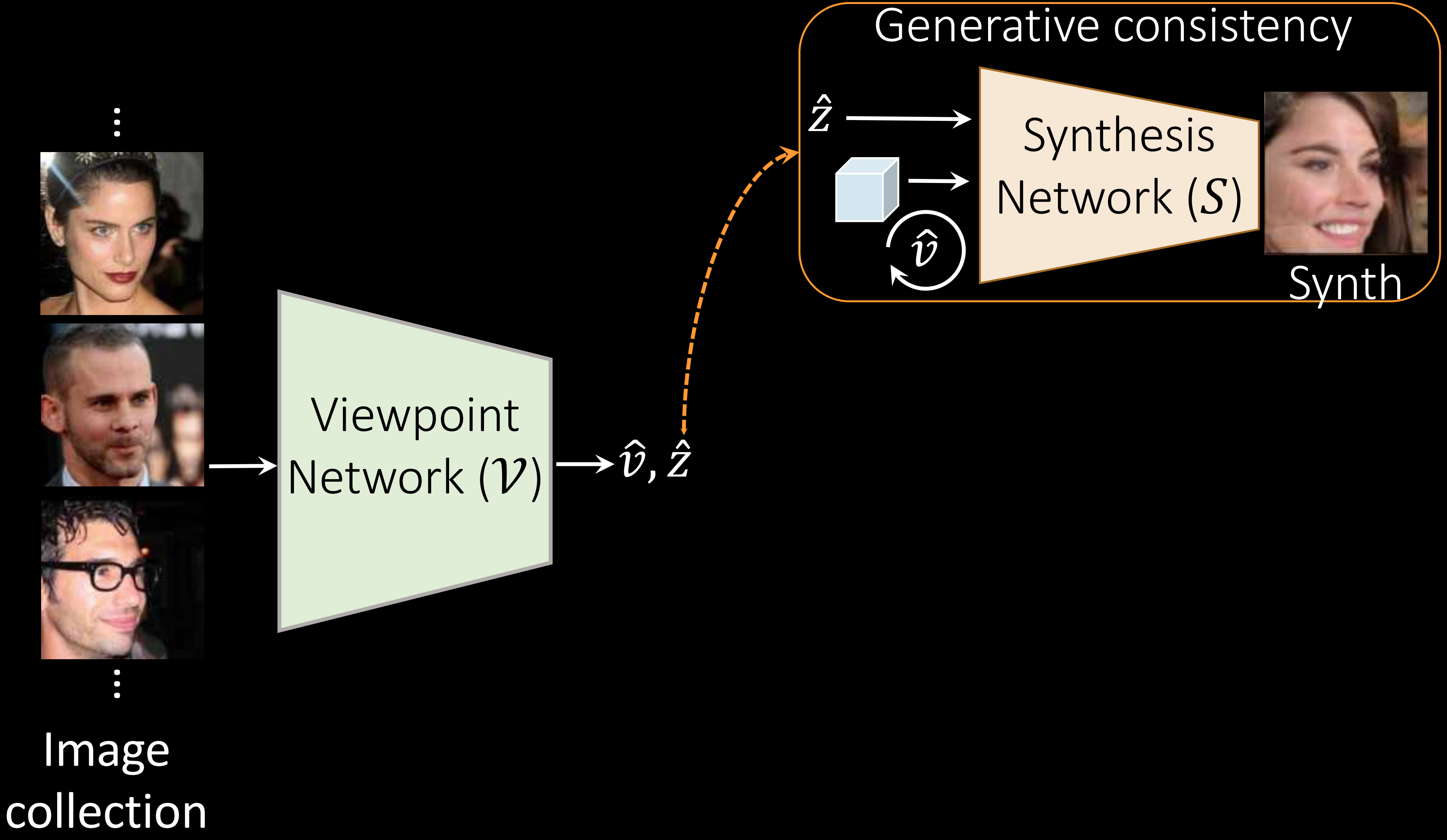
Align
CAD model



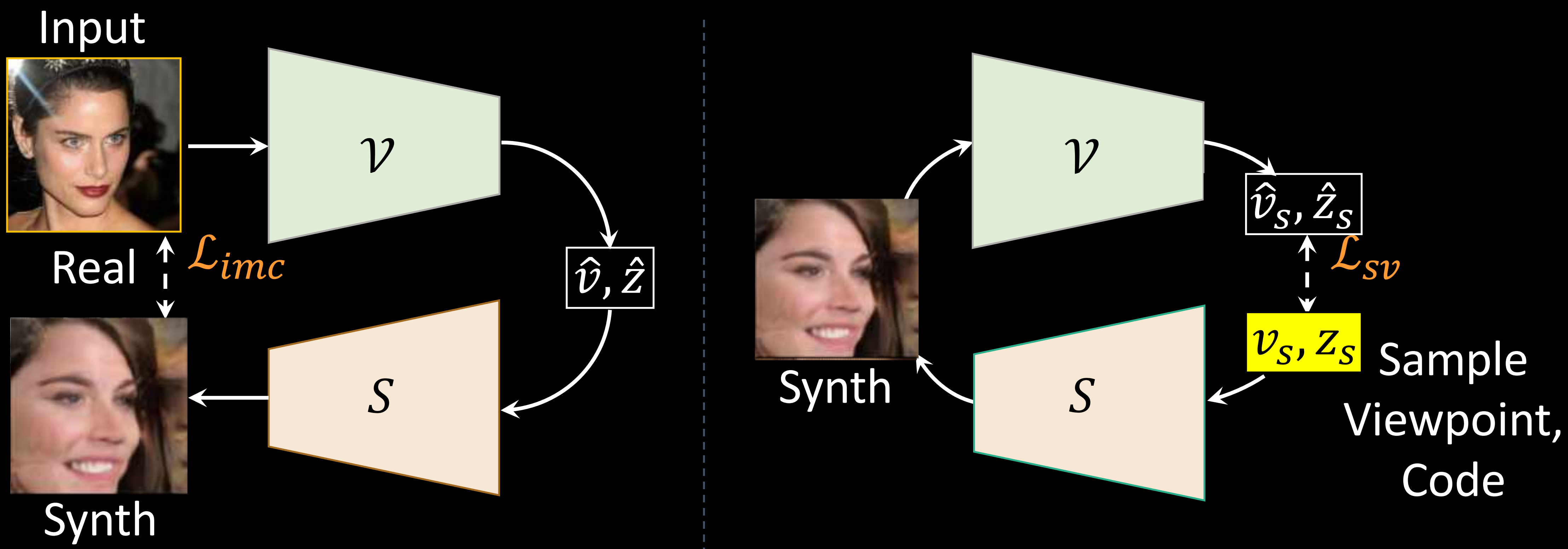
Viewpoint: (Azimuth, Elevation, Tilt)



Face collection {I}
No Labels
Car collection {I}
No Labels
Self-supervised training
Face Viewpoint
Network (𝒱)
Car Viewpoint
Network (𝒱)


Generative consistency
$\hat{z}$
Synthesis
Network (S)
$\hat{v}$
Synth
Viewpoint
Network ($\mathcal{V}$)
$\hat{v}, \hat{z}$
Image
collection

Generative consistency



(a) Image consistency

(b) Style and viewpoint consistency

Analysis by Synthesis

Synthesis for Analysis

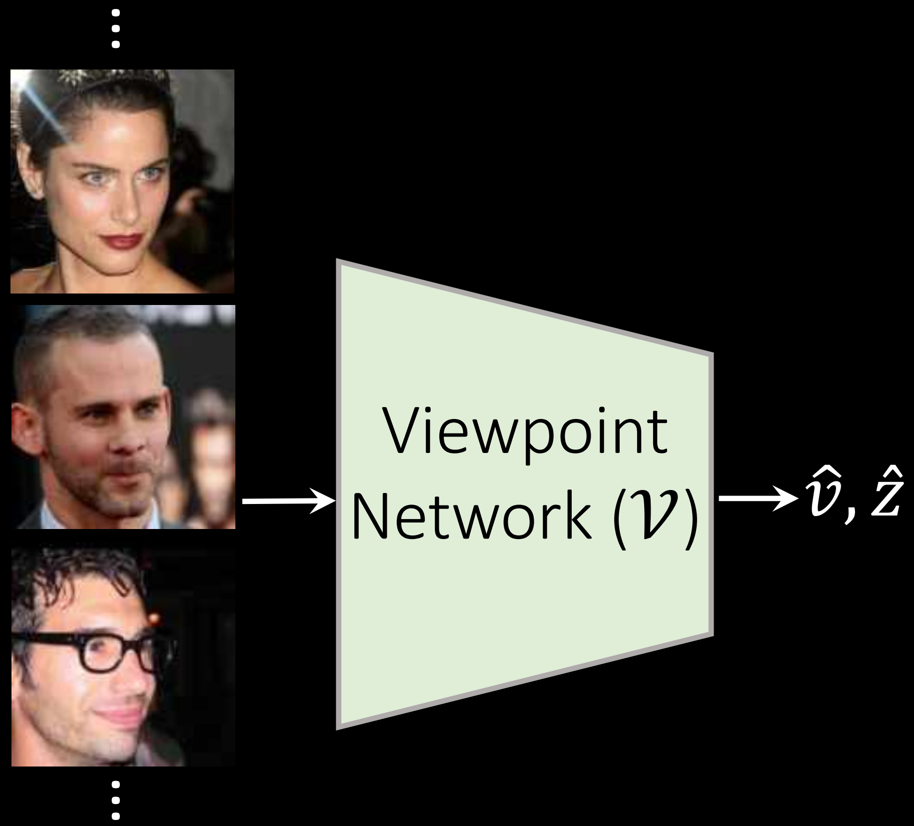
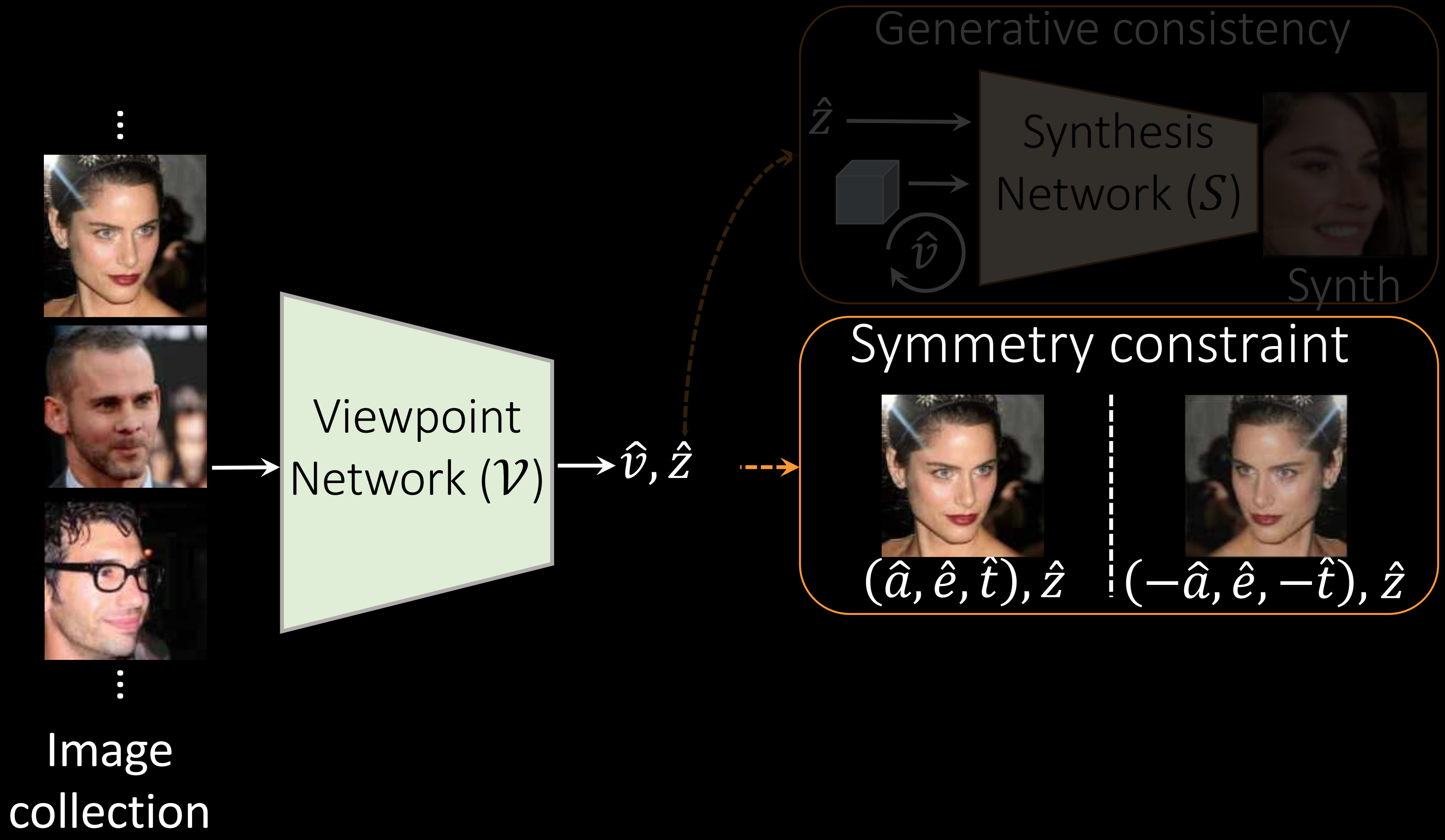
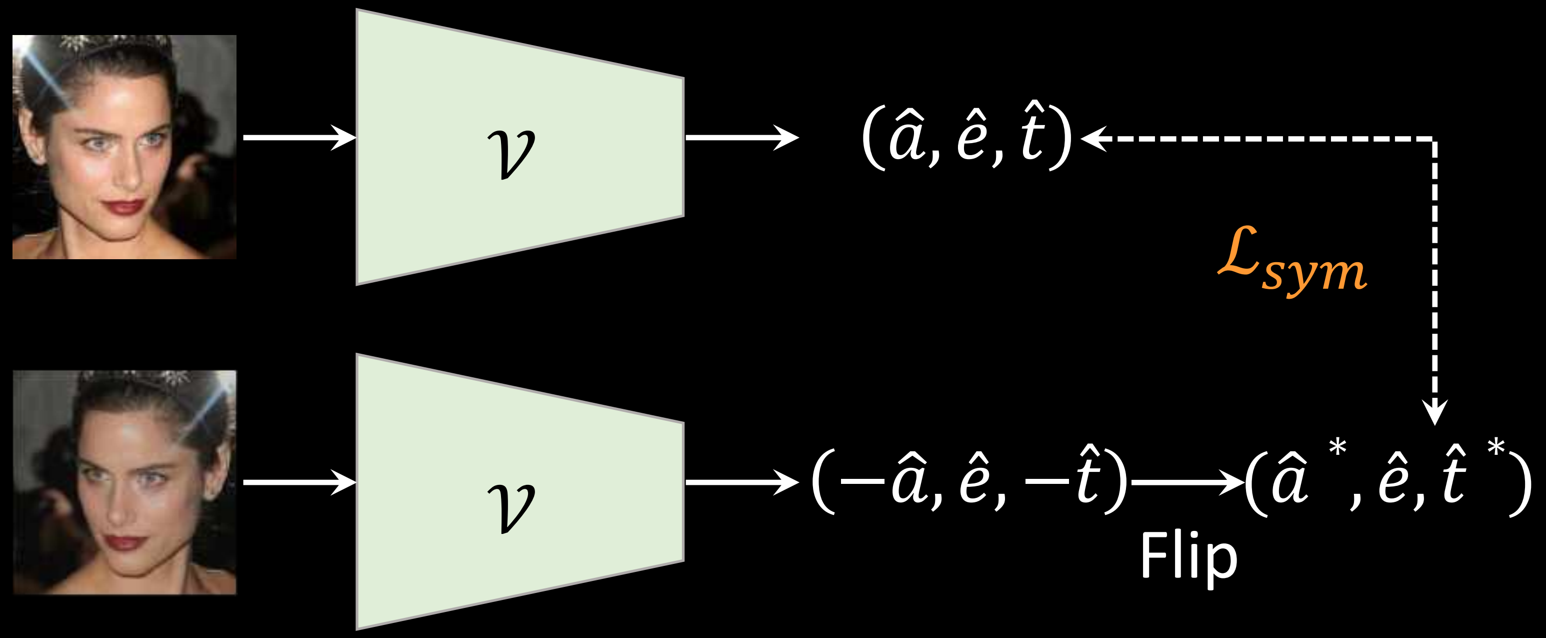


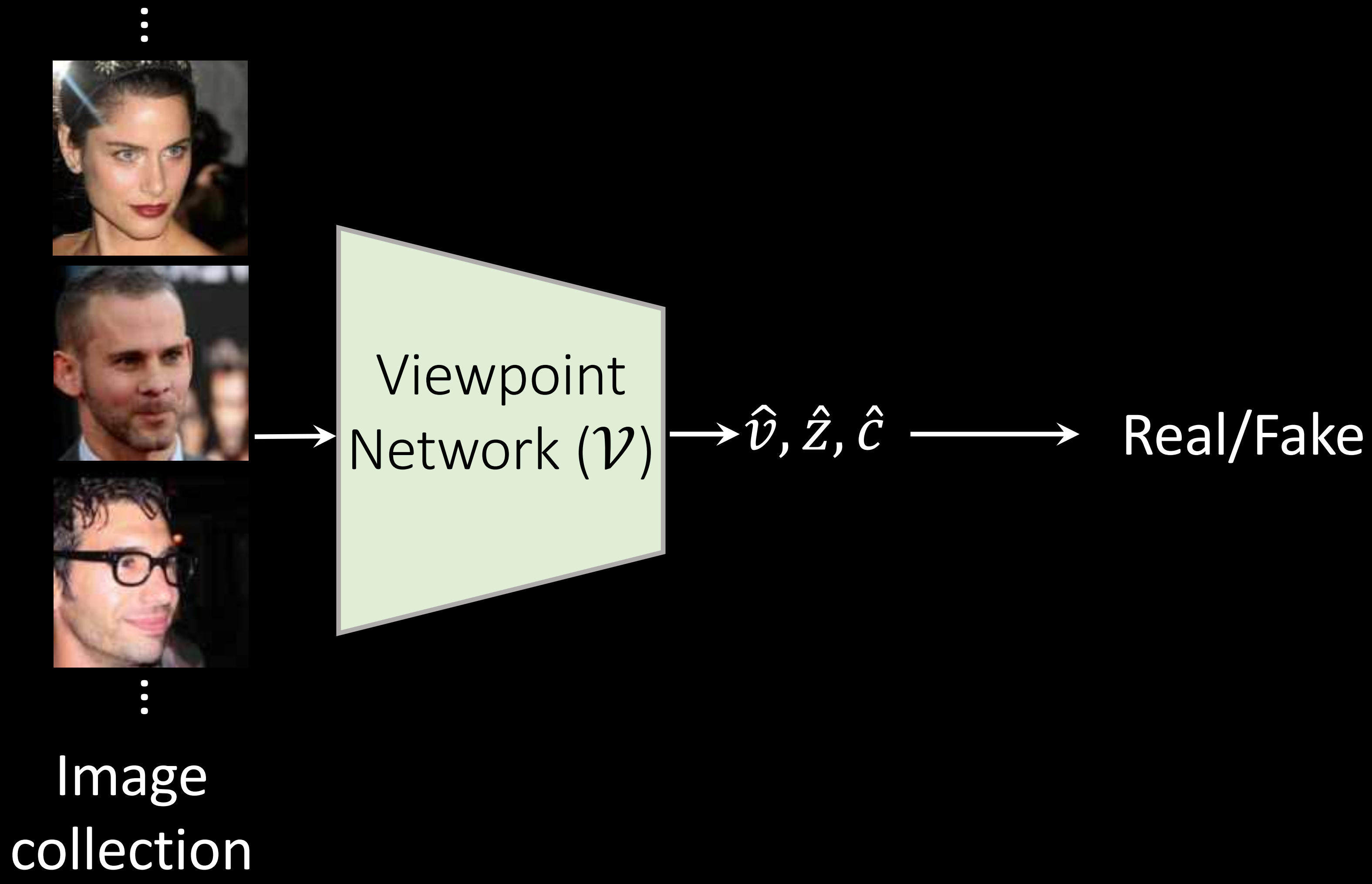
Image
collection

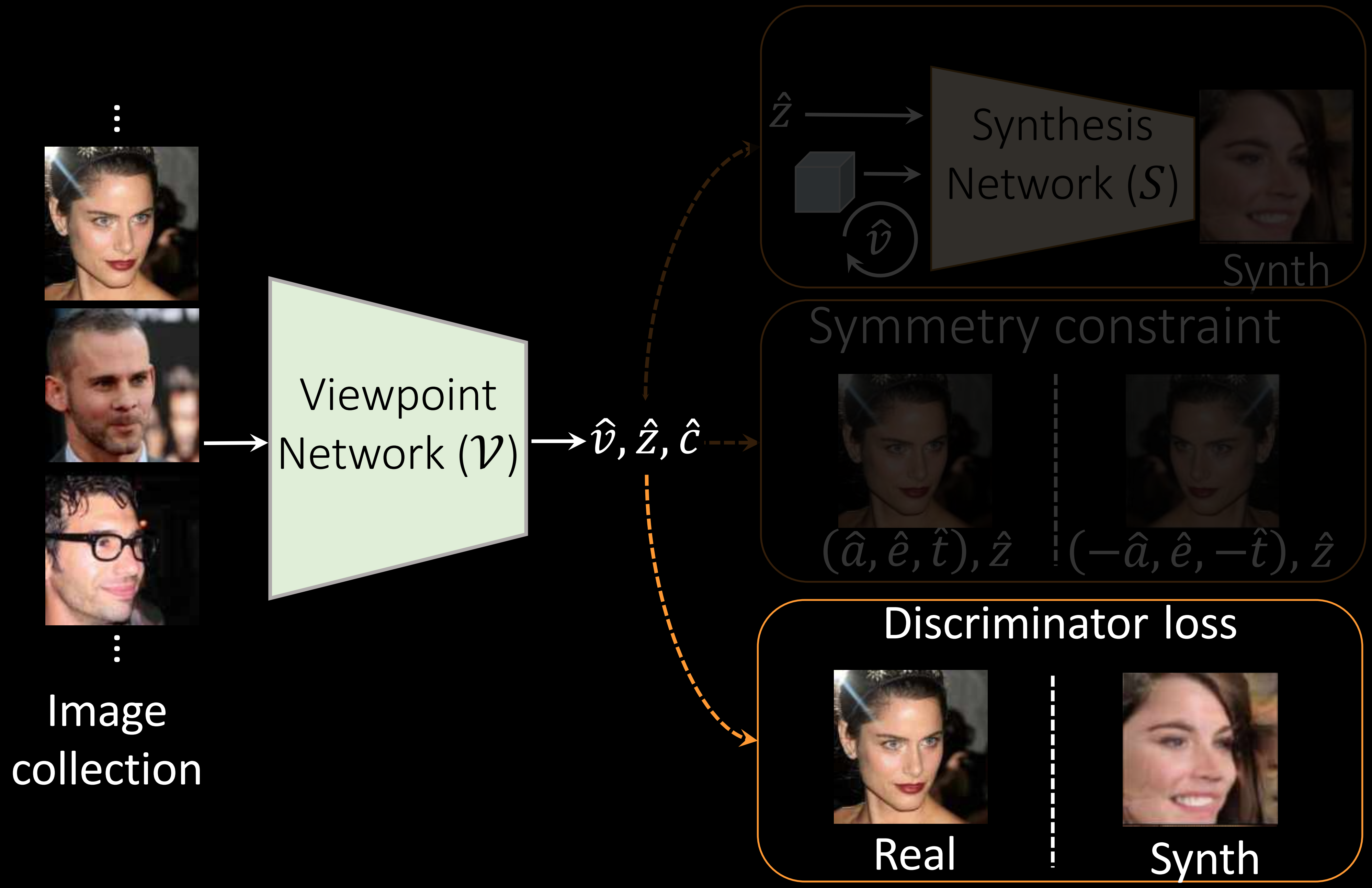


Generative consistency
$\hat{z}$
Synthesis Network (S)
$\hat{v}$
Synth
Symmetry constraint
$(\hat{a}, \hat{e}, \hat{t}), \hat{z}$
$(-\hat{a}, \hat{e}, -\hat{t}), \hat{z}$
Viewpoint Network ($\mathcal{V}$)
$\hat{v}, \hat{z}$
Image collection

Symmetry Constraint




Viewpoint
Network ($\mathcal{V}$)
$\hat{v}, \hat{z}, \hat{c}$
Real/Fake
Image
collection



$\hat{z}$
Synthesis Network (S)
$\hat{v}$
Synth
Symmetry constraint
$(\hat{a}, \hat{e}, \hat{t}), \hat{z}$
$(-\hat{a}, \hat{e}, -\hat{t}), \hat{z}$
Viewpoint Network ($\mathcal{V}$)
$\hat{v}, \hat{z}, \hat{c}$
Discriminator loss
Real
Synth
Image collection

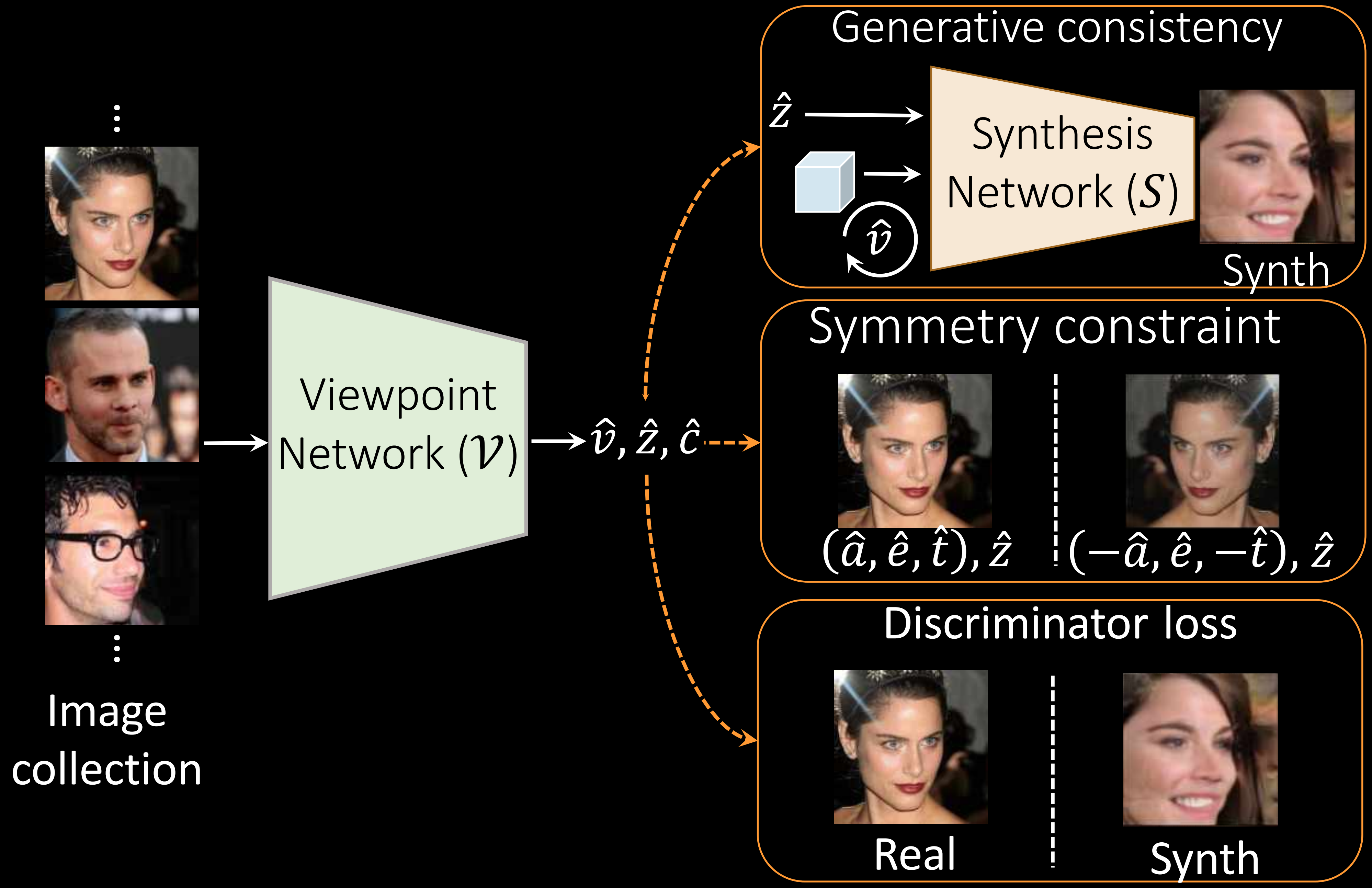
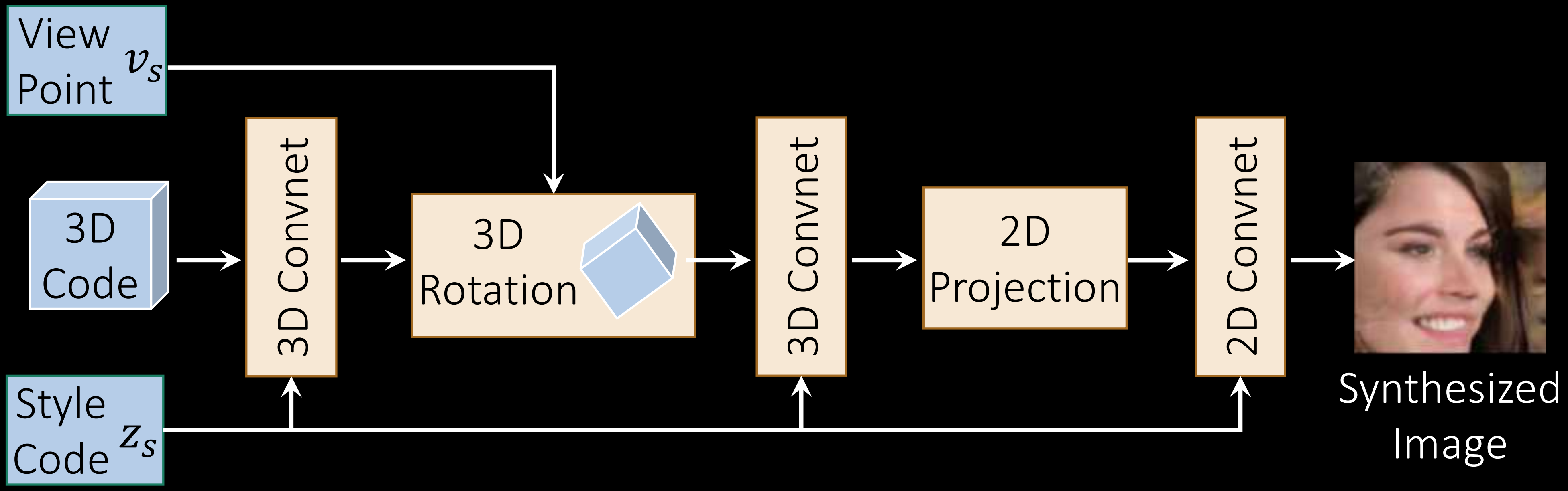


Image
collection
Viewpoint
Network ($\mathcal{V}$)
$\hat{v}, \hat{z}, \hat{c}$
Generative consistency
$\hat{z}$
$\hat{v}$
Synthesis
Network (S)
Synth
Symmetry constraint
$(\hat{a}, \hat{e}, \hat{t}), \hat{z}$
$(-\hat{a}, \hat{e}, -\hat{t}), \hat{z}$
Discriminator loss
Real
Synth

Viewpoint-aware synthesis network [1]



[1] Nguyen-Phuoc et al. “Hologan: Unsupervised learning of 3d representations from natural images.” In *ICCV 2019*

Synthesis Results - Varying Azimuth

Synthesis Results - Varying Elevation

Synthesis Results - Varying Tilt

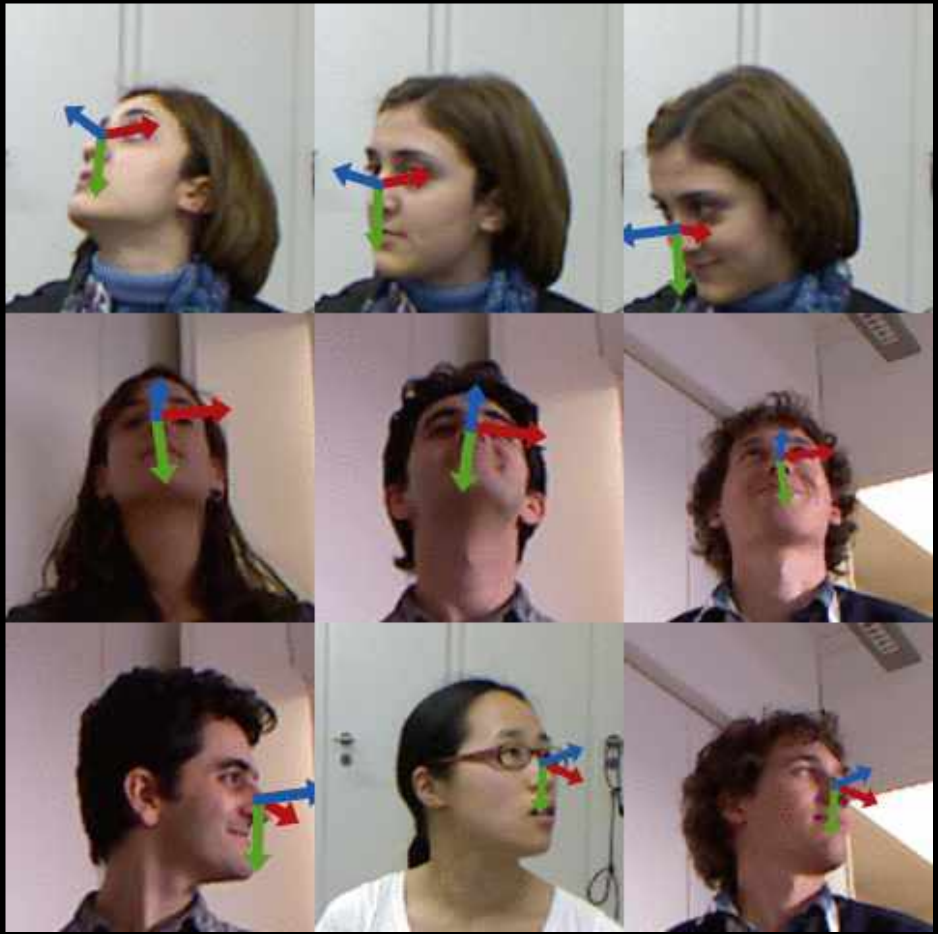
Head pose estimation

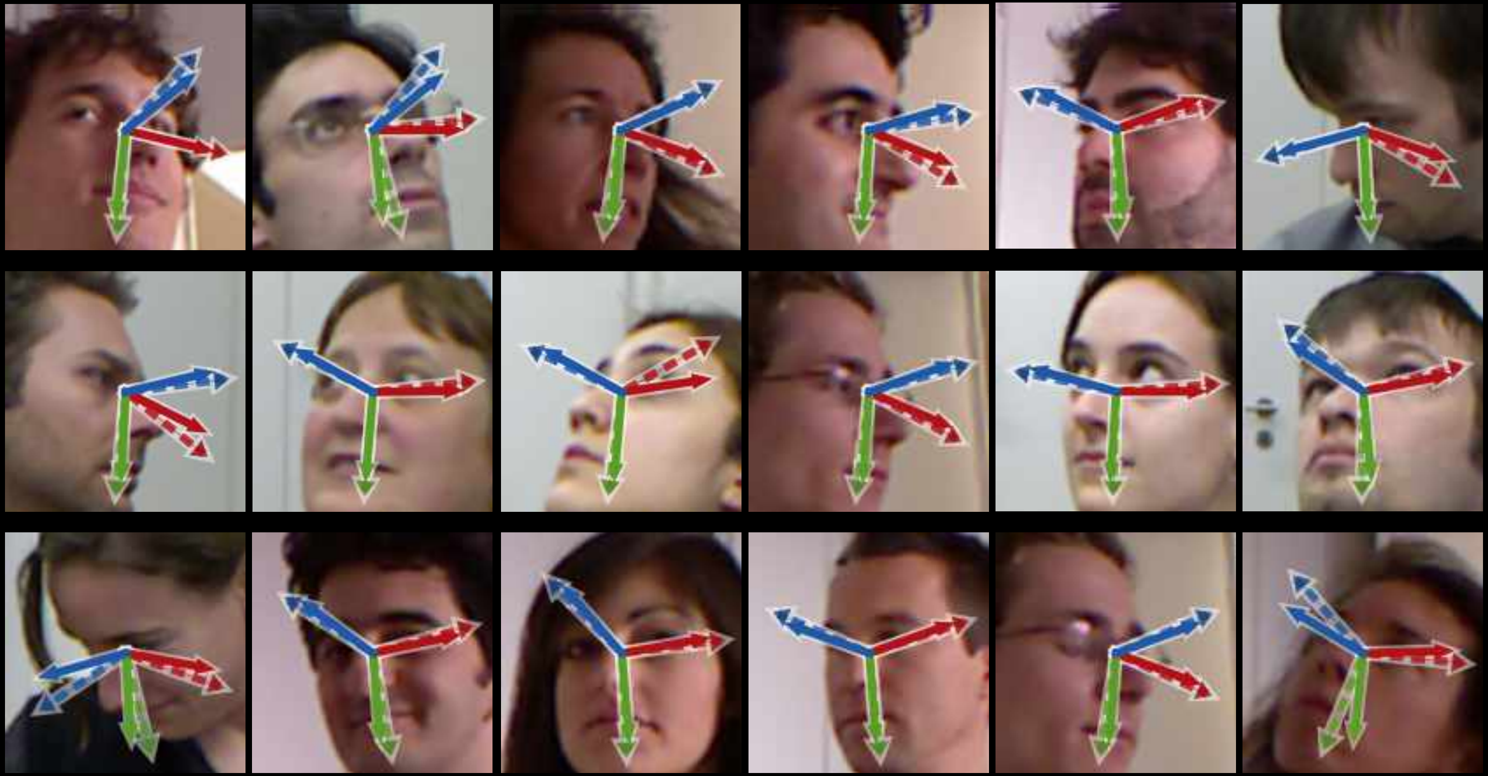
	Method	Azimuth	Elevation	Tilt	MAE
Self-Supervised	LMDIS [Zhang et al. CVPR 18] + PnP	16.8	26.1	5.6	16.1
	IMM [Jakab et al. Neurips 18] + PnP	14.8	22.4	5.5	14.2
	SCOPS [Hung et al. CVPR 19] + PnP	15.7	13.8	7.3	12.3
	HoloGAN [Nguyen-Phuoc et al. ICCV19]	8.9	15.5	5.0	9.8
	SSV (Ours)	6.0	9.8	4.4	6.7

Head pose estimation

	Method	Azimuth	Elevation	Tilt	MAE
Self-Supervised	LMDIS [Zhang et al. CVPR 18] + PnP	16.8	26.1	5.6	16.1
	IMM [Jakab et al. Neurips 18] + PnP	14.8	22.4	5.5	14.2
	SCOPS [Hung et al. CVPR 19] + PnP	15.7	13.8	7.3	12.3
	HoloGAN [Nguyen-Phuoc et al. ICCV19]	8.9	15.5	5.0	9.8
	SSV (Ours)	**6.0**	**9.8**	**4.4**	**6.7**
Supervised	3DDFA [Zhu et al. TPAMI 17]	36.2	12.3	8.7	19.1
	KEPLER [Kumar et al. FG 17]	8.8	17.3	16.2	13.9
	Dlib [Kazemi et al. CVPR 14]	16.8	13.8	6.1	12.2
	FAN [Bulat et al. CVPR 17]	8.5	7.4	7.6	7.8
	Hopenet [Ruiz et al. CVPRW 18]	5.1	6.9	3.3	5.1
	FSA [Yang et al. CVPR 19]	**4.2**	**4.9**	**2.7**	**4.0**

Sample Results





Other Objects

(PascalVOC 3D)

Median error
Lower is better

	Method	Car	Bus	Train
Self-Supervised	**SSV (Ours)**	**10.1**	**9.0**	**5.3**
	VGG-View	34.2	19.0	9.4
Supervised	Tulsiani *et al.* CVPR 15	9.1	5.8	8.7
	Mahendran *et al.* BMVC 18	8.1	4.3	7.3
	Liao *et al.* CVPR 19	5.2	3.4	**6.1**
	Grabner *et al.* CVPR 18	**5.1**	**3.3**	6.7

Inlier Count
Higher is better

Self-Supervised	**SSV (Ours)**	**0.67**	**0.82**	**0.96**
	VGG-View	0.43	0.69	0.82
Supervised	Tulsiani *et al.* CVPR 15	0.89	**0.98**	0.80
	Liao *et al.* CVPR 19	**0.93**	0.97	**0.84**
	Grabner *et al.* CVPR 18	**0.93**	0.97	0.80


FIAT
07·60·NR
UN
102
NYPD
TOYOTA
909
EXPRESS INC

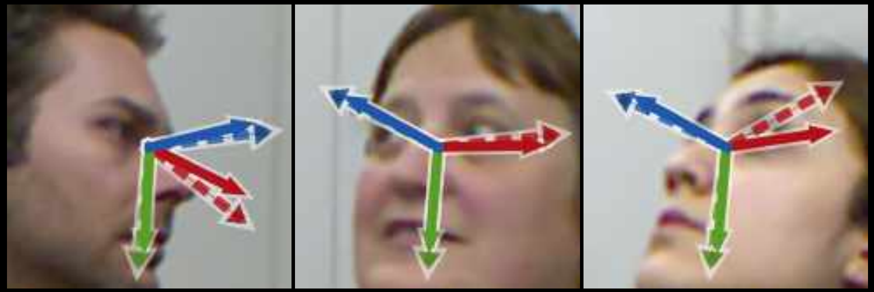
Viewpoint discovery: Remarks

- One of the first approaches for self-supervised viewpoint learning
- Works on several object categories
- Performance close to even fully-supervised approaches

- Key techniques
  - Viewpoint-aware GAN: *Analysis-by-synthesis* and *Synthesis-for-analysis*
  - Symmetry constraints

Conclusion

- Part and viewpoint discovery from image collections



- Useful for higher level tasks such as 3D object reconstruction
- Leverage prior-knowledge about the problem to design loss functions and to avoid degenerate solutions
- Future outlook: Self-supervised learning of other object attributes

Thank you

Comments and suggestions are most welcome

varunjampani@gmail.com

http://varunjampani.github.io