Spatial Content Alignment For Pose Transfer

Wing-Yin Yu

Lai-Man Po

Yuzhi Zhao

Jingjing Xiong

Kin-Wai Lau

Friday Seminar

2021-10-15

CityU

Department of Electrical Engineering

香港城市大學

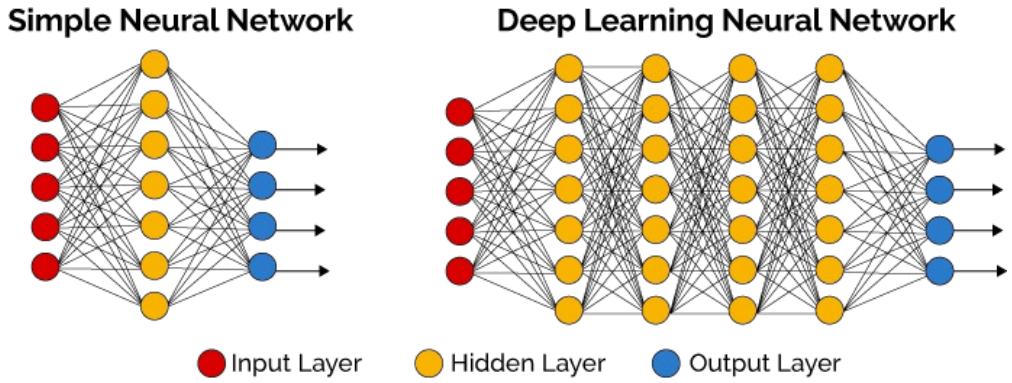
City University of Hong Kong

Content

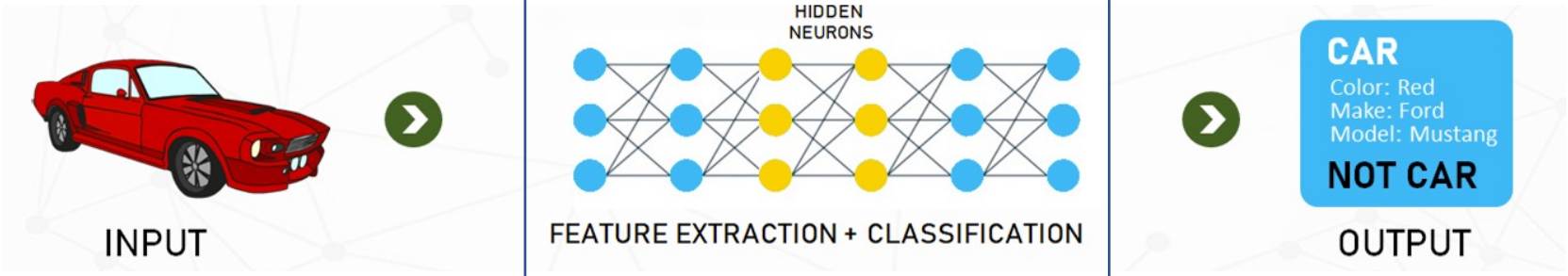
- **Background knowledge**
- **The task I have done**
- **Method**
- **Experiment result**

Deep Learning (DL)

- **Deep Learning** is a subfield of machine learning
- Automatically extract features by a neural network



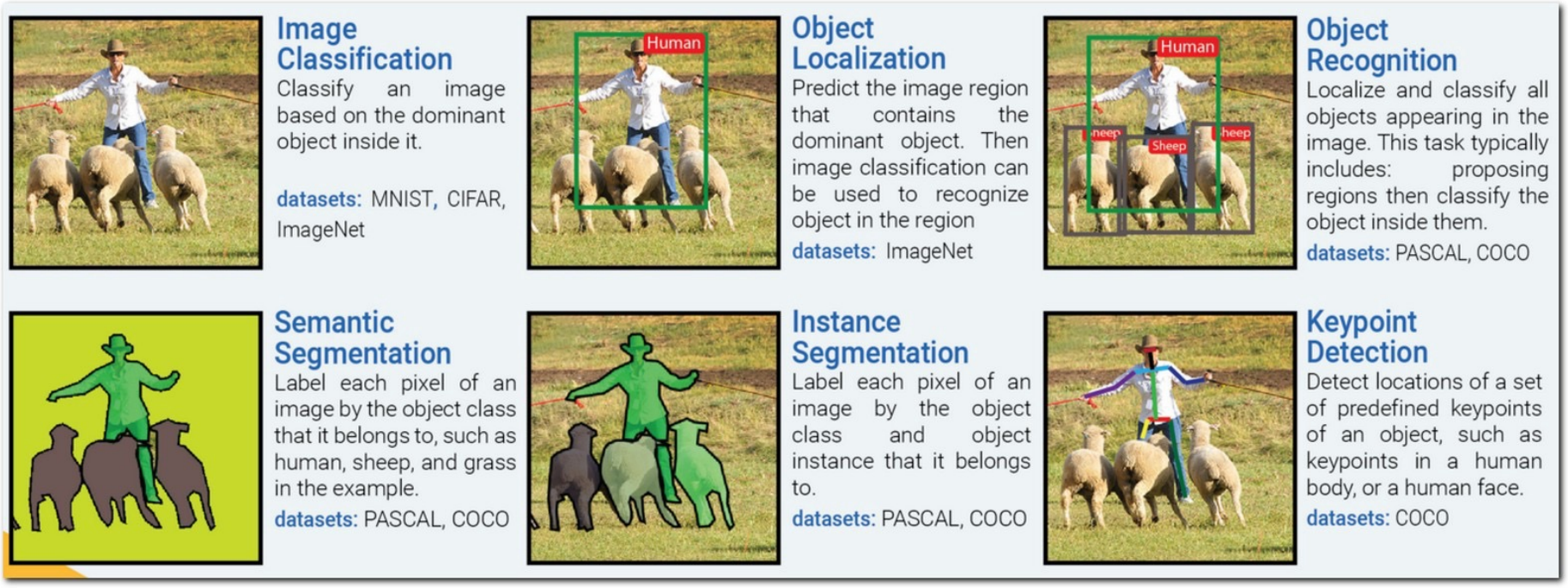
Example



Magic block

Computer Vision (CV)

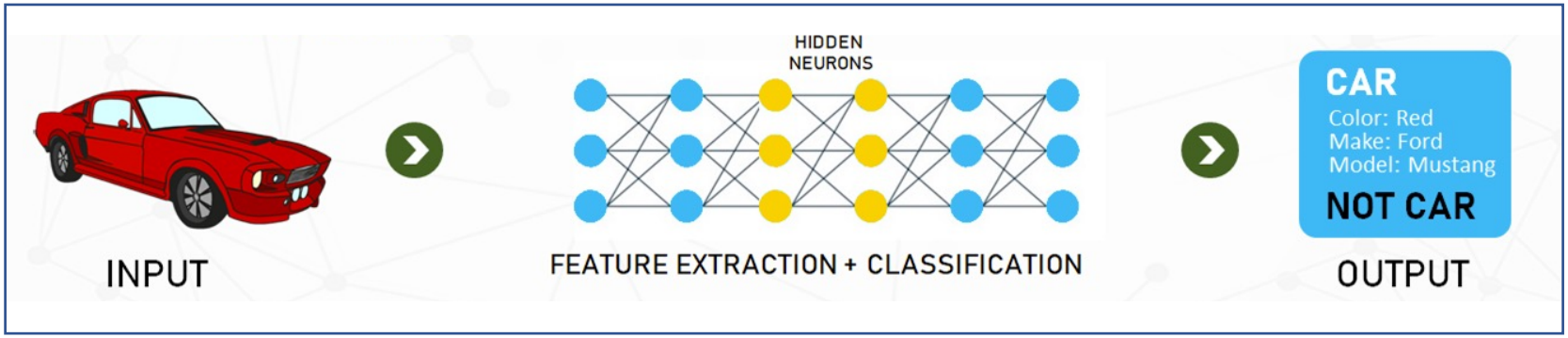
General computer vision tasks



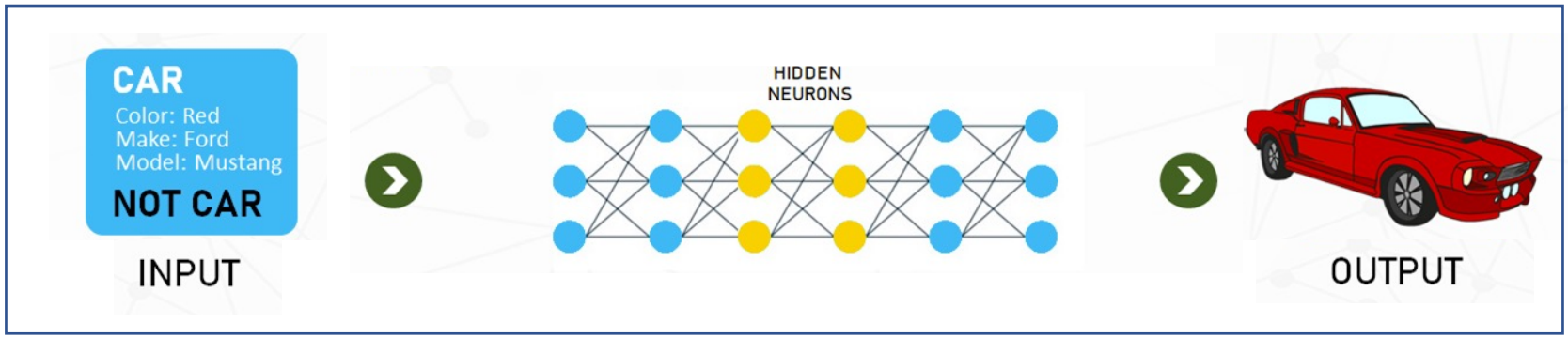
Credit: https://medium.com/@nikasa1889/the-modern-history-of-object-recognition-infographic-aea18517c318

What if…

Image → Label

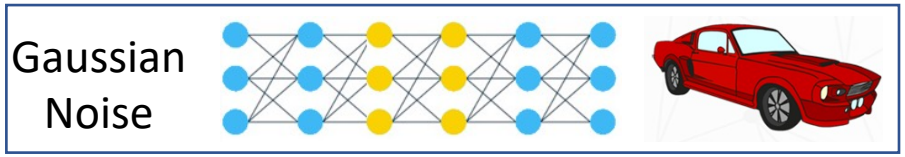


Label → Image?

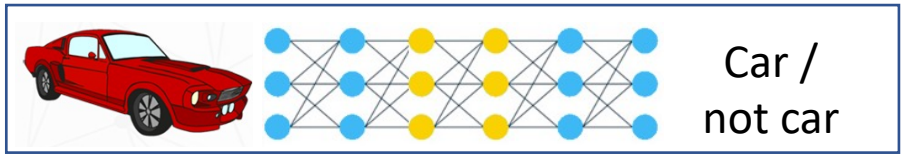


Generative Adversarial Networks (GANs)

- A framework of Deep Learning model
  - Capture **distribution** from training data
  - Generate new data from same **distribution**
- Major components:
  - **Generator**: generate real-looking examples
  - **Discriminator**: classify examples as real or fake



Generator



Discriminator

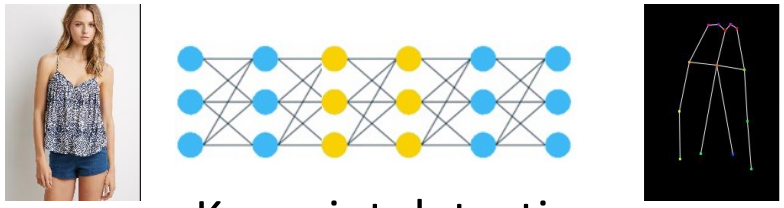
$$\min_G \max_D V(D,G)$$

$$V(D,G) = \mathbb{E}_{x\sim p_{data}(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[\log(1-D(G(z))]$$

Goodfellow, Ian, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. "Generative adversarial networks." *Communications of the ACM* 63, no. 11 (2020): 139-144.

Problem overview

- **Task 1:** Detect the pose keypoints from the image



Keypoint detection

- **Task 2:** Generate image from the keypoints

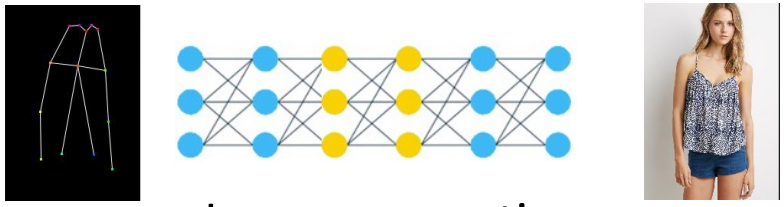


Image generation

- **Our task:** Generate image with other poses

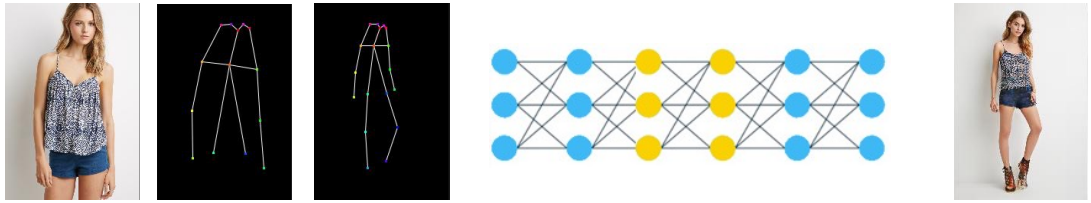
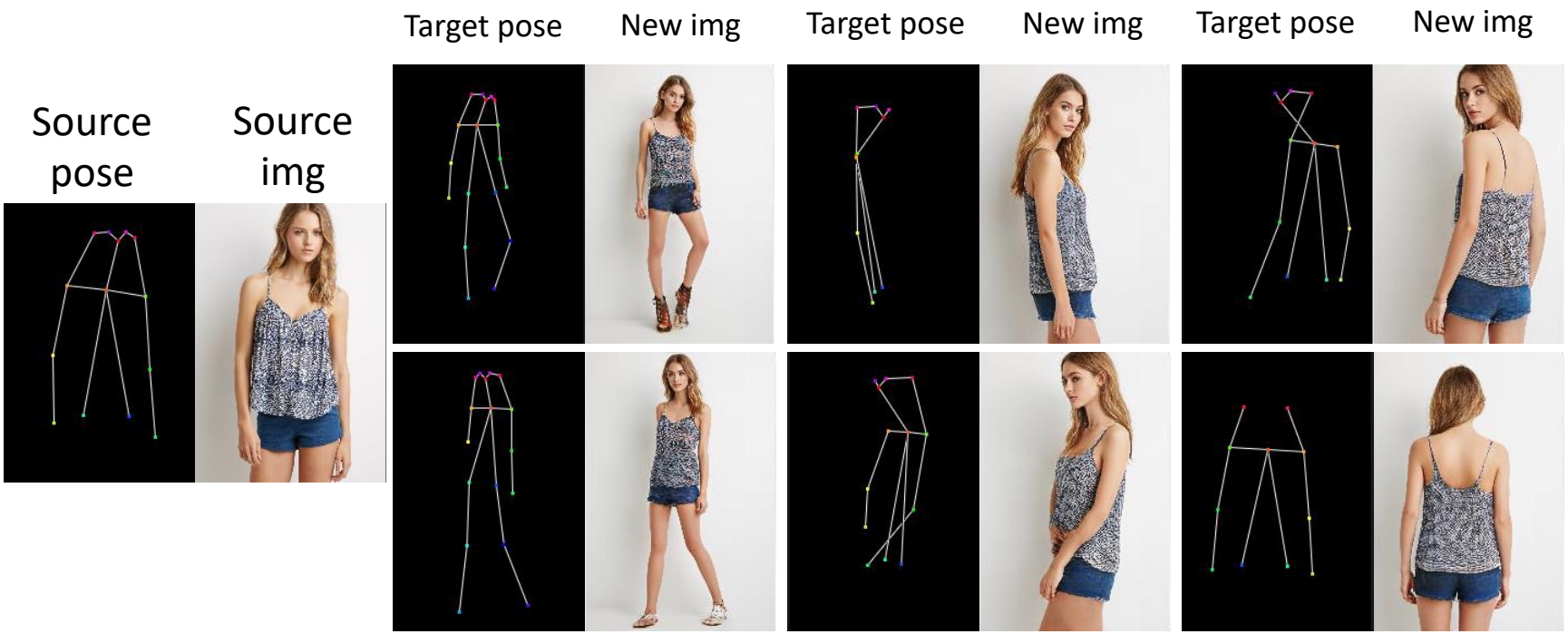


Image translation

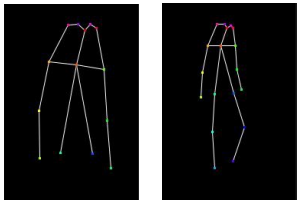
Pose transfer

- Input
  - Source person image
  - Source pose
  - Target pose
- Output
  - A new person image
    - Target pose
    - Style and Texture of source person image

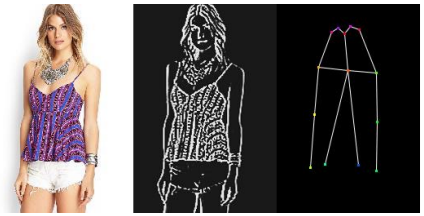


Motivation

- **Spatial misalignment problem**
  - Sparse correspondence between pose landmarks
    - Warp-based method: blurry effect
    - Attention-based method: lack of excitation
  - Insufficient content guidance
    - Person identity
    - Garment reconstruction
    - Incomplete texture



- **Content-driven guidance**
  - Leverage edge map as an extra constraint with pose heatmap
  - Guide the network to produce more photo-realistic person images through texture enhancement

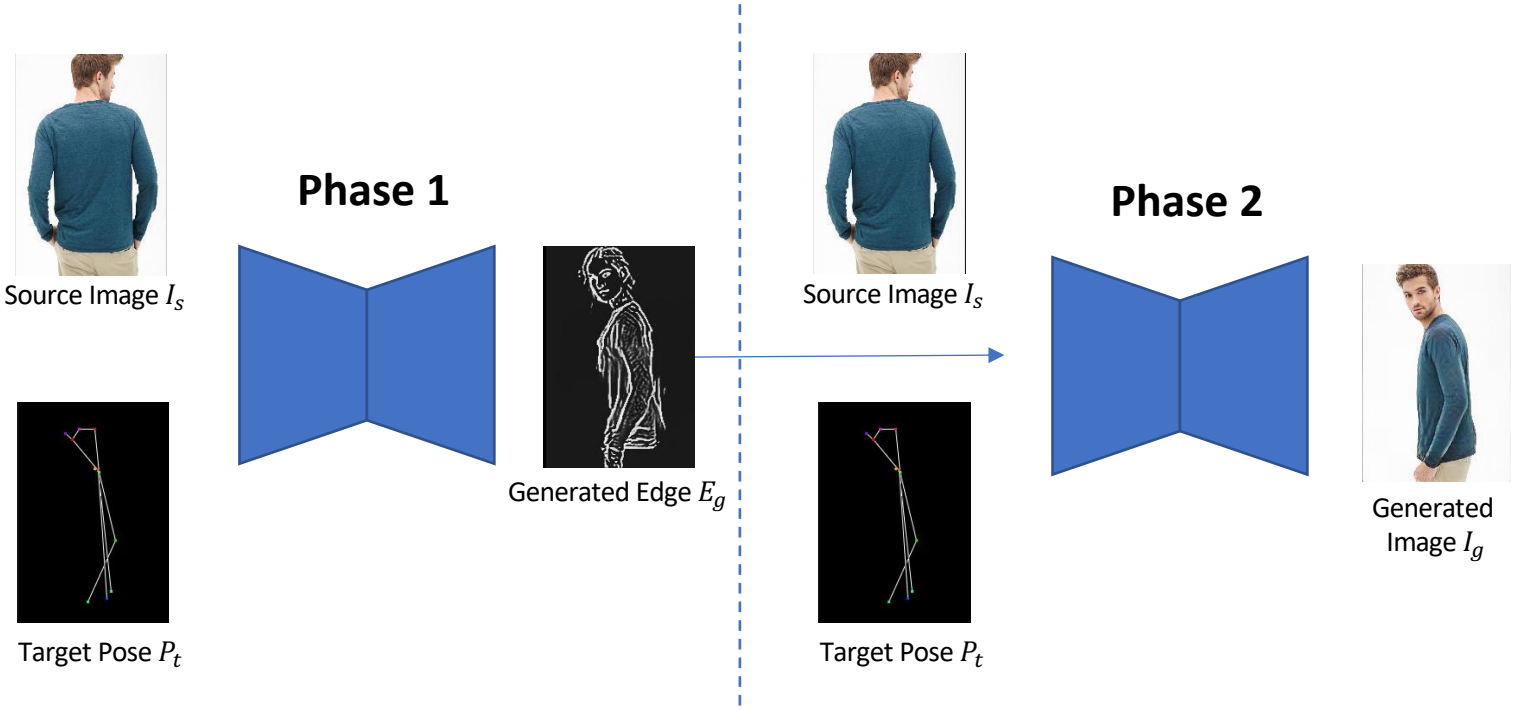


Source

Target

Methodology

- **Two-phase approach**
  - Phase 1: Prior Content Transfer Network (PCT-Net)
  - Phase 2: Image Synthesis Network (IS-Net)

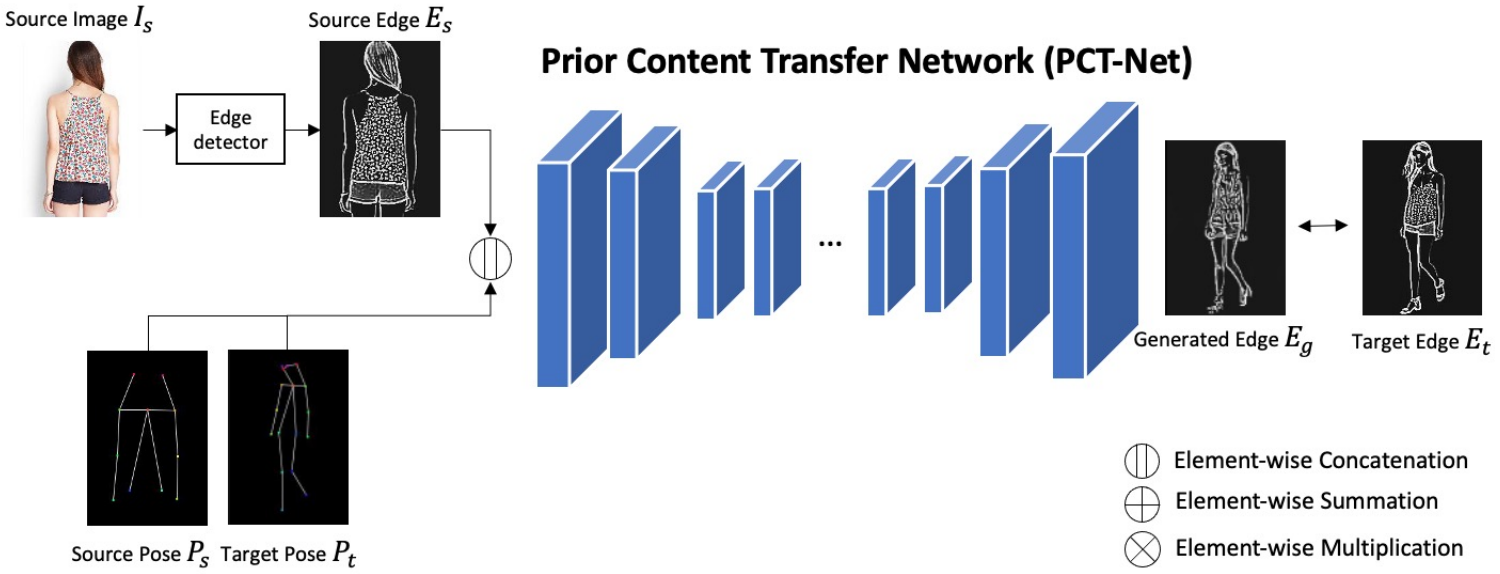


Methodology

Phase 1: Prior Content Transfer Network (PCT-Net)

- **Objective:** generate the content in advance
  - Leverage pixel-level edge map to highlight high-frequency signals
  - Dominate a wide range of spectrum for content information

$$E_g = \mathcal{F}(E_s, P_s, P_t)$$

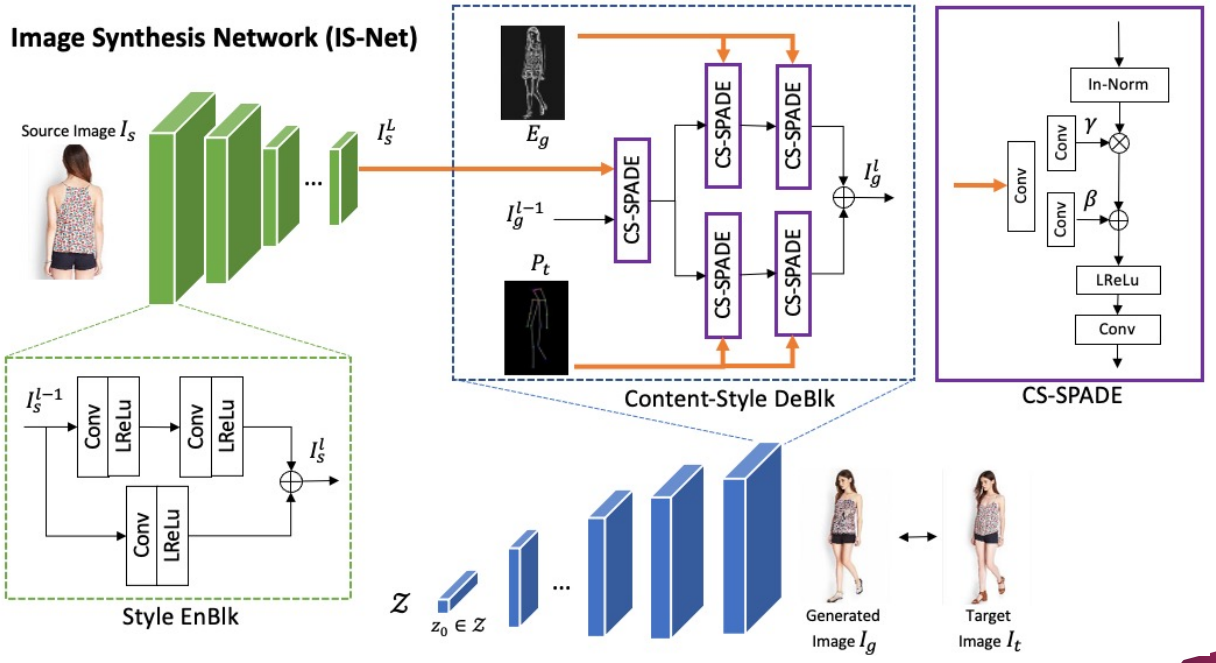


Methodology

Phase 2: Image Synthesis Network (IS-Net)

- **Objective:** render a realistic-looking person image
  - Encoder-decoder architecture
  - Appearance style of I_s , prior content E_g , spatial content P_t

$$I_g = G(I_s, E_g, P_t)$$



Methodology

- **Content-Style Spatially-Adaptive Normalization (CS-SPADE)**
  - Inspired by SPADE
  - Accept aligned content and non-aligned style features as **Source**
  - Inject content and style information with learnable parameters

$$\mathcal{W}(\mathrm{S}, f) = \delta(\gamma^i_{c,h,w}(\mathrm{S}) \cdot \frac{f^i_{n,c,h,w} - u^i_{nc}}{\sigma^i_{nc}} + \beta^i_{c,h,w}(S))$$

$\delta =$ LeakyReLU

Mean of $f^i_{n,c,h,w} = u^i_{nc} = \frac{1}{H^i W^i} \sum_{h,w} f^i_{n,c,h,w}$

SD of $f^i_{n,c,h,w} = \left(\sigma^i_{nc}\right)^2 = \frac{1}{H^i W^i} \sum_{h,w} {f^i_{n,c,h,w}}^2 - {u^i_{nc}}^2$

- **Content-Style Decoder Block (Content-Style DeBlk)**

$$I^l_g = \mathcal{U}\left(E_g, \mathcal{W}\left(I^{l-1}_g, I^L_s\right)\right) + \mathcal{V}\left(P_t, \mathcal{W}\left(I^{l-1}_g, I^L_s\right)\right)$$

Park, Taesung, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. "Semantic image synthesis with spatially-adaptive normalization." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2337-2346. 2019.

Objective Loss Function

$$\mathcal{L}_{full} = \lambda_{adv}\mathcal{L}_{adv} + \lambda_1\mathcal{L}_1 + \lambda_{per}\mathcal{L}_{per} + \lambda_{cx}\mathcal{L}_{cx}$$

Adversarial loss:

$$\mathcal{L}_{adv} = \mathbb{E}_{I_s,I_t,I_g}\left[\log\left(D_s(I_s,I_t)\right) + \log\left(1 - D_s\left(I_s,I_g\right)\right)\right] + \mathbb{E}_{I_t,P_t,I_g}\left[\log\left(D_c(P_t,I_t)\right) + \log\left(1 - D_c\left(P_t,I_g\right)\right)\right]$$

Appearance loss:

$$\mathcal{L}_1 = \frac{1}{CHW}\sum_{c,h,w}\left\|I_g\,|_{c,h,w} - I_t|_{c,h,w}\right\|_1$$

Perceptual loss:

$$\mathcal{L}_{per} = \frac{1}{C_\ell H_\ell W_\ell}\sum_{c,h,w}\left\|\theta_\ell\left(I_g\right)|_{c,h,w} - \theta_\ell(I_t)|_{c,h,w}\right\|_1$$

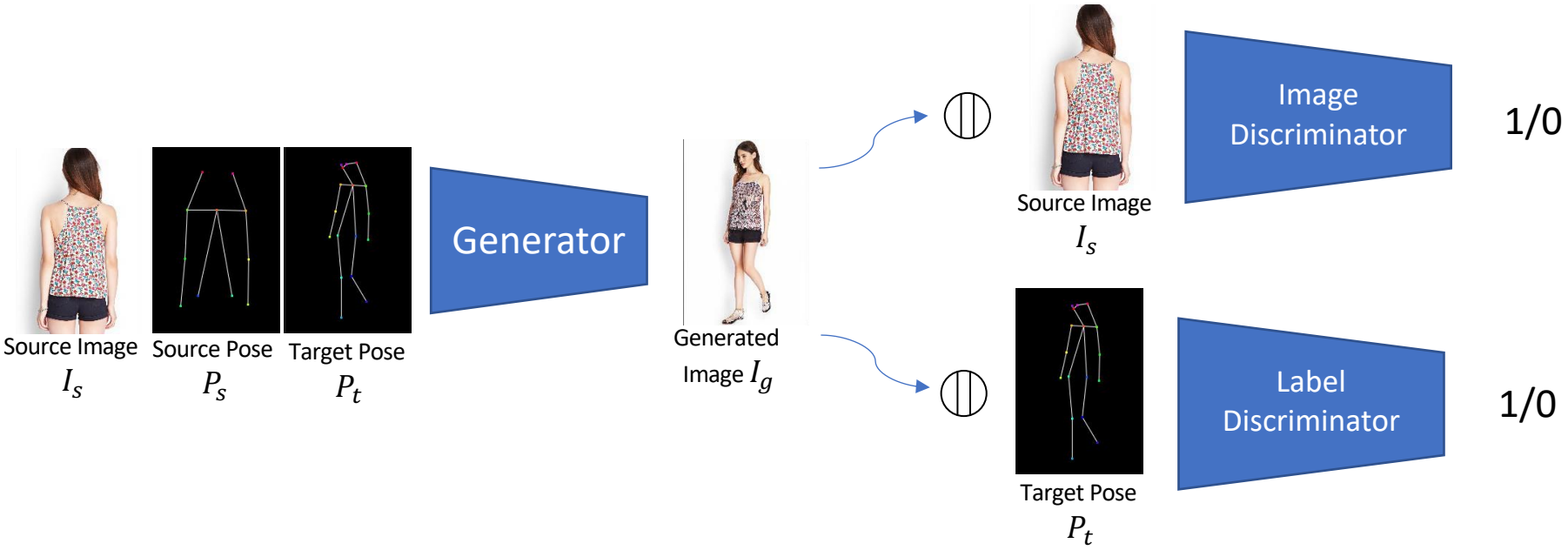
Contextual loss:

$$\mathcal{L}_{cx} = -\frac{1}{C_\ell H_\ell W_\ell}\sum_{c,h,w}\log\left[CX\left(\theta_\ell\left(I_g\right)|_{c,h,w}, \theta_\ell(I_t)|_{c,h,w}\right)\right]$$

Objective Loss Function

- **Adversarial loss:**
  - Minimize the KL divergence between two different domains

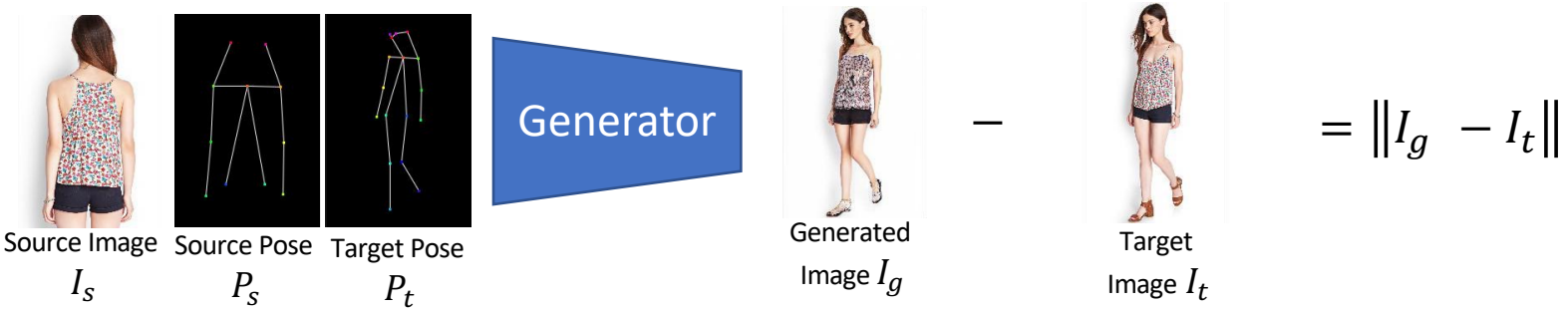
$$\mathcal{L}_{adv} = \mathbb{E}_{I_s,I_t,I_g}\left[\log\left(D_s(I_s,I_t)\right)+\log\left(1-D_s\left(I_s,I_g\right)\right)\right]+ \mathbb{E}_{I_t,P_t,I_g}\left[\log\left(D_c(P_t,I_t)\right)+\log\left(1-D_c\left(P_t,I_g\right)\right)\right]$$



Objective Loss Function

- **Appearance loss:**
  - Minimize the pixel-wise least absolute distance

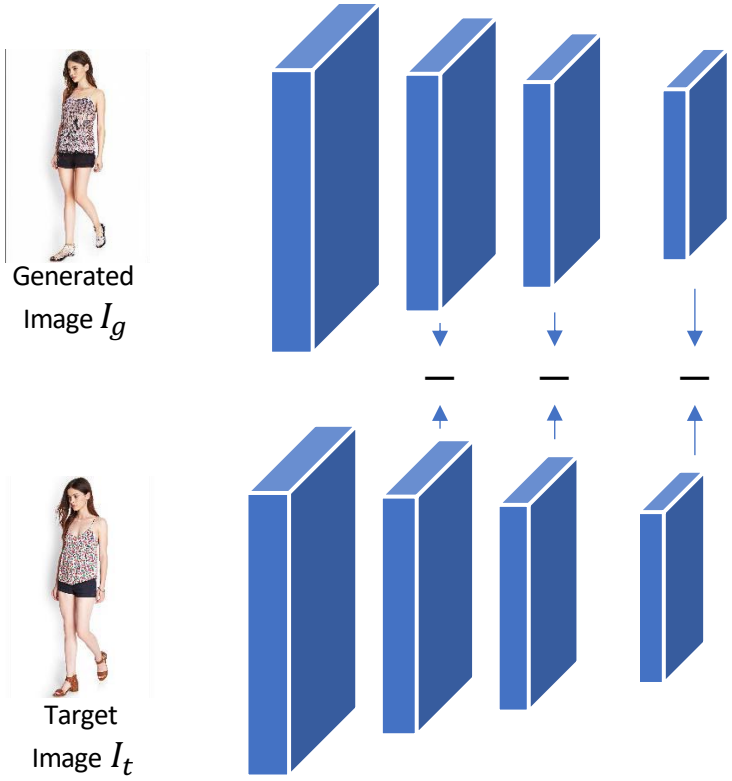
$$\mathcal{L}_1 = \frac{1}{CHW}\sum_{c,h,w}\left\| I_g\ |_{c,h,w} - I_t|_{c,h,w} \right\|_1$$



Objective Loss Function

- **Perceptual loss:**
  - minimize the distance in feature space,

$$\mathcal{L}_{per} = \frac{1}{C_\ell H_\ell W_\ell} \sum_{c,h,w} \left\| \theta_\ell(I_g)|_{c,h,w} - \theta_\ell(I_t)|_{c,h,w} \right\|_1$$



Minimize the L1 distance for the features

Johnson, Justin, Alexandre Alahi, and Li Fei-Fei. "Perceptual losses for real-time style transfer and super-resolution." In *European conference on computer vision*, pp. 694-711. Springer, Cham, 2016.

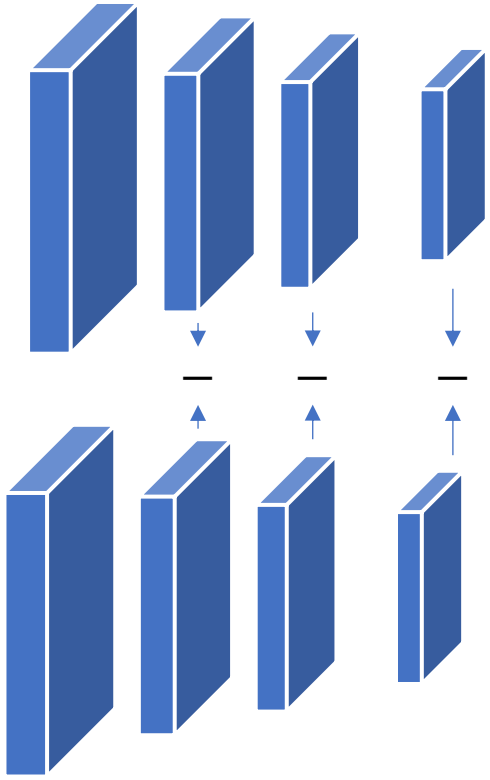
Objective Loss Function

- **Contextual loss:**
  - minimize the distance in feature space,

$$\mathcal{L}_{cx} = -\frac{1}{C_\ell H_\ell W_\ell}\sum_{c,h,w}\log\left[CX\left(\theta_\ell\left(I_g\right)|_{c,h,w}, \theta_\ell(I_t)|_{c,h,w}\right)\right]$$

Generated Image I_g

Target Image I_t



Maximize the similarity of the features

Roey Mechrez, Itamar Talmi, and Lihi Zelnik-Manor, "The contextual loss for image transformation with non-aligned data," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 768–783.

Comparison with SOTA

Table 1: Quantitative analysis of comparison against the current state-of-art methods. The best scores are highlighted with bold format.

Methods	IS↑	SSIM↑	FID↓	LPIPS↓
PG^2 [8]	3.202	0.773	-	-
Def-GAN [2]	3.362	0.760	18.457	0.233
RATE-Net [3]	3.125	0.774	14.611	0.218
PATN [4]	3.209	0.773	19.816	0.253
APS [5]	3.295	**0.775**	15.017	0.178
ADGAN [6]	3.364	0.772	13.224	0.176
SCA-GAN(Ours)	**3.497**	**0.775**	**11.676**	**0.167**

- Our method outperforms the current SOTA methods
- 4% enhancement of IS score compared to ADGAN
- Generate high-quality images with natural texture synthesis
- Beneficial to the edge content guidance for supporting the geometric transformation of fine-gained appearance details

Fig. 2: Qualitative results of comparison against current state-of-art methods. Please zoom in for details.

Ablation Study

Table 2: Quantitative results of ablation study. The best scores are highlighted with bold format.

Methods	IS↑	SSIM↑	FID↓	LPIPS↓
with semantic content	3.388	0.762	18.634	0.195
w/o prior transfer	3.365	0.774	12.402	**0.166**
w/o content branch	3.274	0.756	19.503	0.194
SPADE ResBlk	3.343	0.772	13.888	0.170
Batch-Norm (encoder)	3.340	**0.775**	12.197	**0.166**
In-Norm (encoder)	3.301	0.773	12.738	0.167
Full model	**3.497**	**0.775**	**11.676**	0.167

- Edge content is better than semantic content
- Provide sufficient guidance to perform detailed spatial alignment
- Preserve better boundary reconstruction between some ambiguous objects
- Content-Style DeBlk outperforms the original SPADE method
- Better image synthesis quality

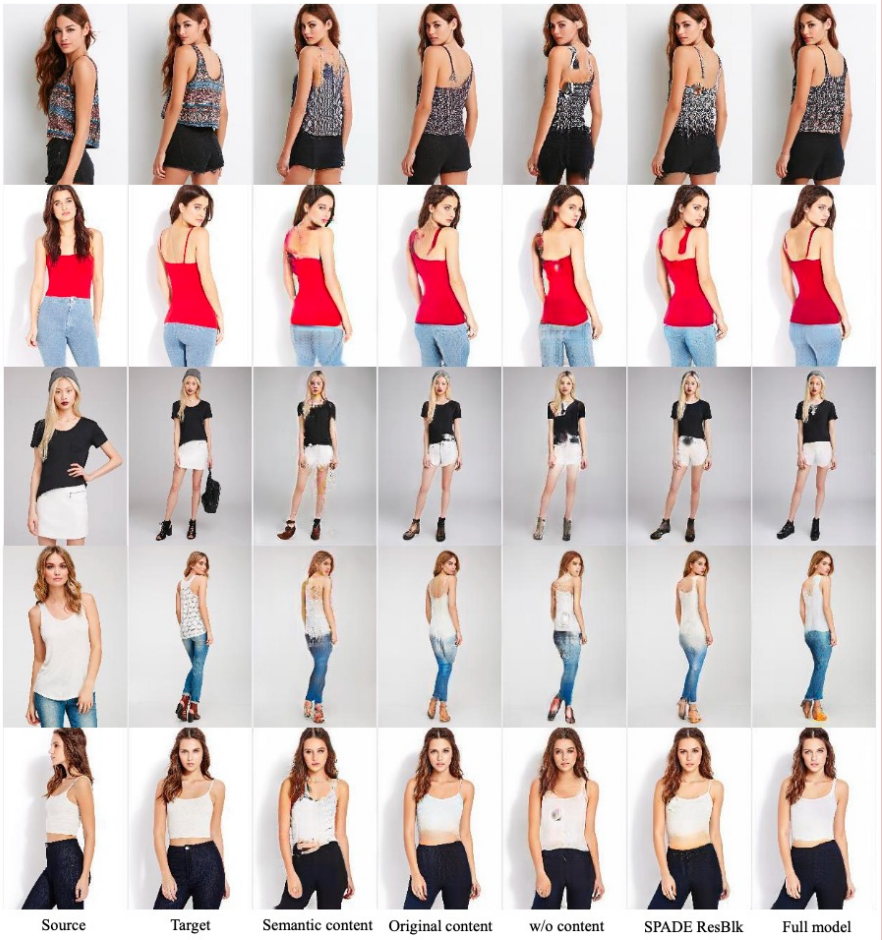


Fig. 3: Qualitative results of ablation study. Please zoom in for details.

Open source

Yu, Wing-Yin, Lai-Man Po, Yuzhi Zhao, Jingjing Xiong, and Kin-Wai Lau. "Spatial Content Alignment for Pose Transfer." In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1-6. IEEE, 2021.

https://github.com/rocketappslab/SCA-GAN

Thank You
Q & A