Deep Image Spatial Transformation for Person Image Generation

Yurui Ren[1,2] Xiaoming Yu[1,2] Junming Chen[1,2] Thomas H. Li[3,1] Ge Li ✉[1,2]
[1]School of Electronic and Computer Engineering, Peking University [2]Peng Cheng Laboratory
[3]Advanced Institute of Information Technology, Peking University
{yrren,xiaomingyu,junming.chen}@pku.edu.cn tli@aiit.org.cn geli@ece.pku.edu.cn

CVPR 2020

2020.04.16

Presented by Yonggyu Kim

Introduction

- **Task**

Pose-guided person image generation is to transform a source person image to a target pose.

- **Motivation**

CNN are limited by the lack of ability to spatially transform the inputs.

- **Solution**

Feature level에서 input을 재조합 하는 differentiable global-flow local-attention framework 제안

Contribution

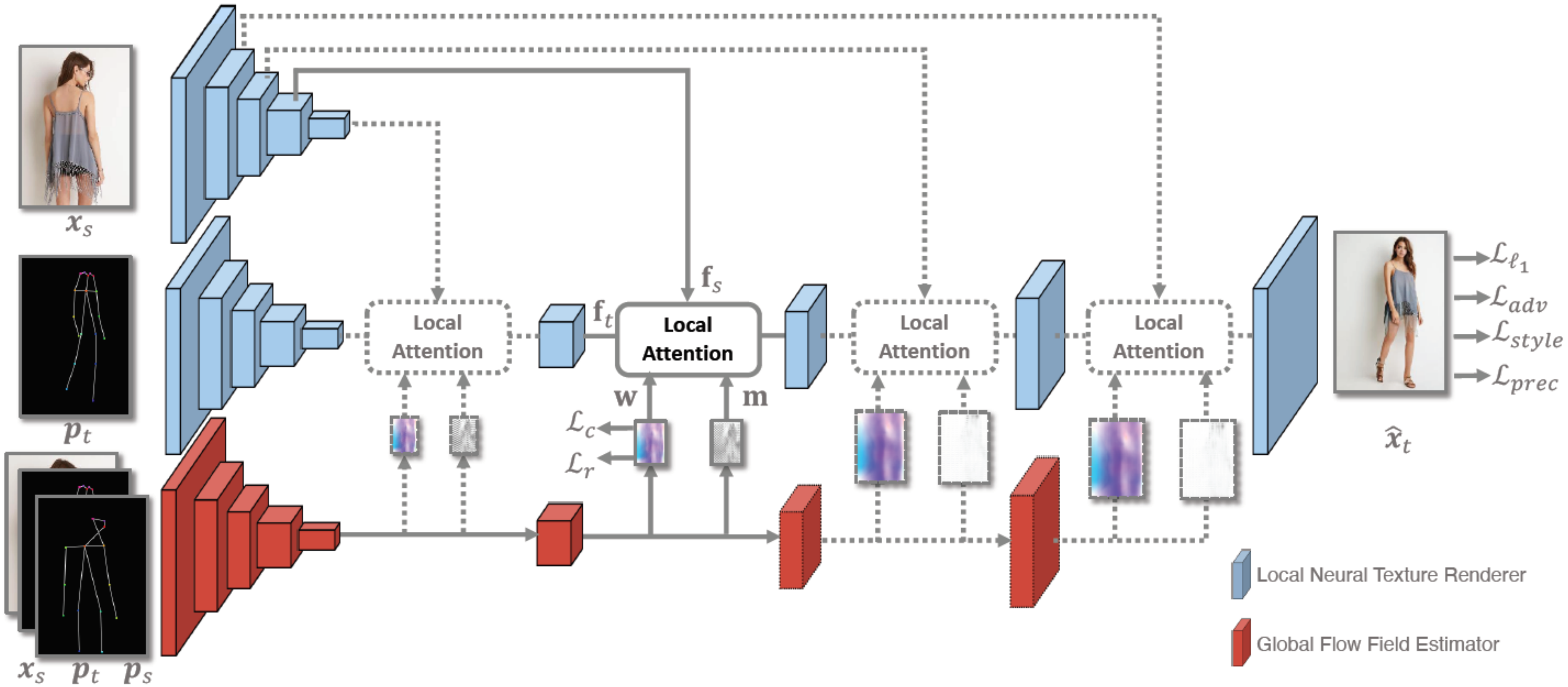
1. **A global-flow local-attention framework** is proposed for the pose-guided person image generation task. **Experiments** demonstrate the **effectiveness** of the proposed method.

2. The carefully-designed **framework and content-aware sampling operation** ensure that our model is able to warp and **reasonably reassemble** the input data at the **feature level**. This operation not only enables the model to generate **new contents**, but also reduces the difficulty of the **flow field estimation task**.

3. Additional experiments on **view synthesis** and **video animation** show that our model can be **flexibly applied** to different tasks requiring spatial transformation.

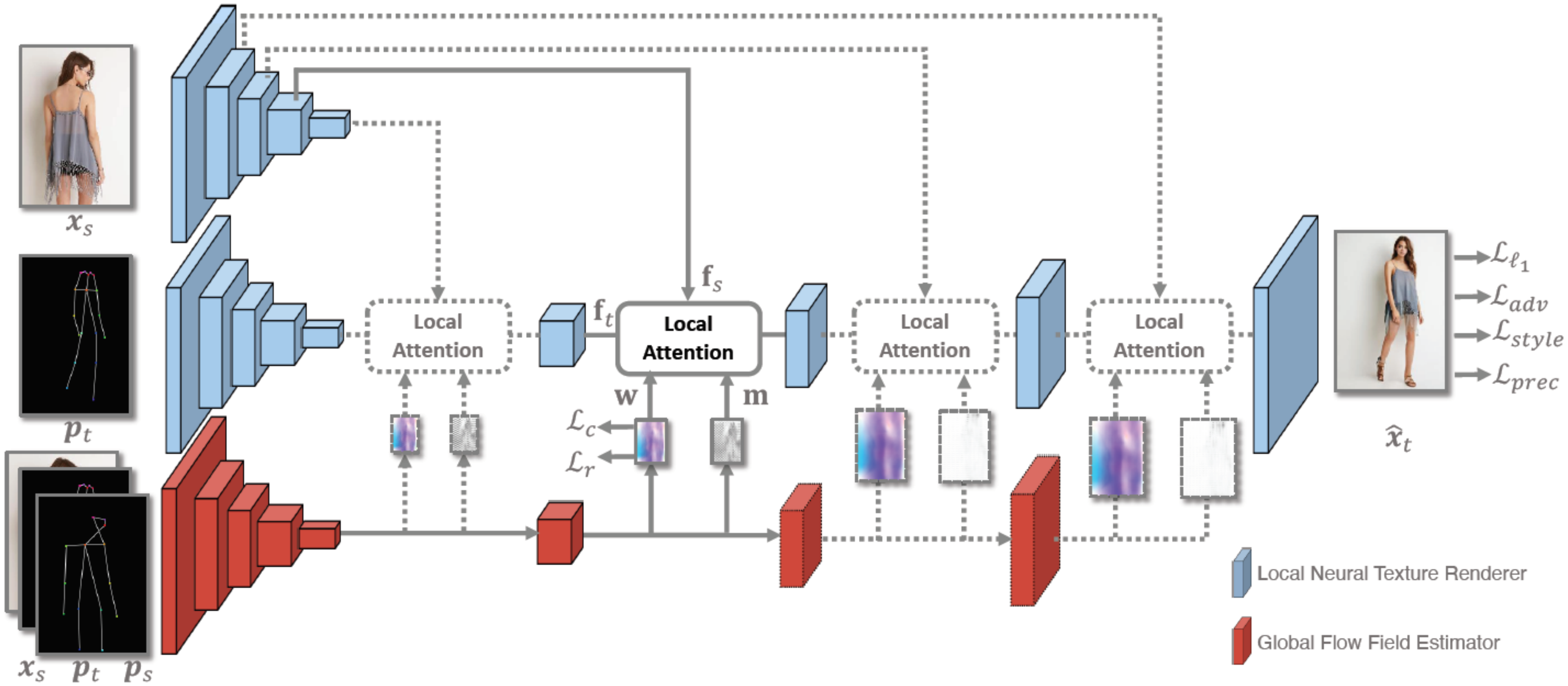
Method

- **Global-flow local-attention framework**
  - Global Flow Field Estimator F : Global flow field & Occlusion mask
  - Local Neural Texture Renderer G : local attention block을 활용해 source feature를 가져온 다음에 render



Method

- Global Flow Field Estimator F : Global flow field & Occlusion mask
  - **sampling correctness loss** & regularization term



- sampling correctness loss

$$\mathbf{v}_{s,\mathbf{w}} = \mathbf{w}(\mathbf{v}_s)$$

$$\mathcal{L}_c = \frac{1}{N}\sum_{l\in\Omega} exp(-\frac{\mu(\mathbf{v}^l_{s,\mathbf{w}}, \mathbf{v}^l_t)}{\mu^l_{max}})$$

$$\mu^l_{max} = \max_{l'\in\Omega} \mu(\mathbf{v}^{l'}_s, \mathbf{v}^l_t)$$

Y Ren, "Structureflow : Image inpainting via structure-aware appearance flow", ICCV2019

Method

- Global Flow Field Estimator F : Global flow field & Occlusion mask
  - sampling correctness loss & **regularization term**

This regularization term is used to punish local regions where the transformation is not an affine transformation.

$$(x_i, y_i) \in \mathcal{N}_n(\mathbf{c}_s, l) \qquad \mathbf{c}_s = \mathbf{c}_t + \mathbf{w}$$

$$(x_i, y_i) \in \mathcal{N}_n(\mathbf{c}_t, l) \qquad \mathbf{c}_t$$

$$\mathbf{T}_l = \mathbf{A}_l \mathbf{S}_l = \begin{bmatrix} \theta_{11} & \theta_{12} & \theta_{13} \\ \theta_{21} & \theta_{22} & \theta_{23} \end{bmatrix} \mathbf{S}_l$$

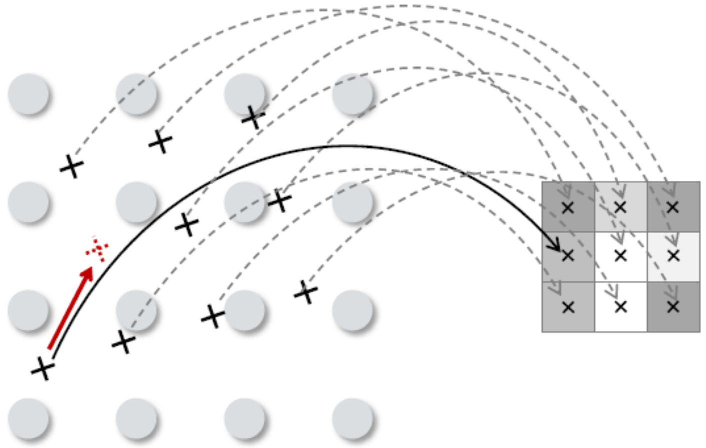
$$\mathbf{T}_l = \begin{bmatrix} x_1 & x_2 & \dots & x_{n\times n} \\ y_1 & y_2 & \dots & y_{n\times n} \end{bmatrix} \quad \mathbf{S}_l = \begin{bmatrix} x_1 & x_2 & \dots & x_{n\times n} \\ y_1 & y_2 & \dots & y_{n\times n} \\ 1 & 1 & \dots & 1 \end{bmatrix}$$

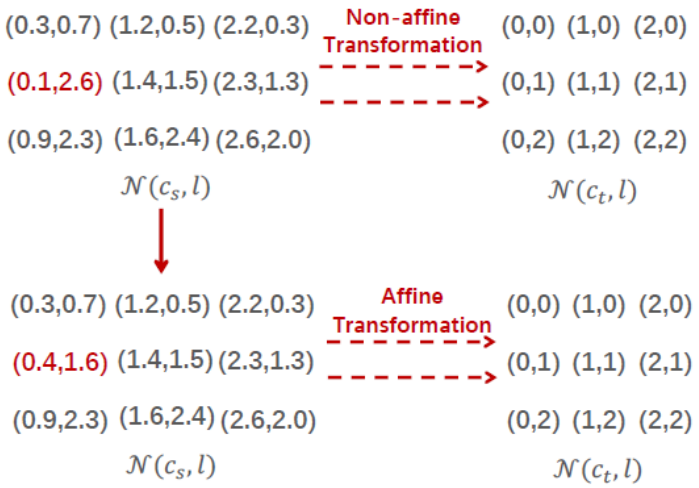
$$\hat{\mathbf{A}}_l = (\mathbf{S}_l^H \mathbf{S}_l)^{-1} \mathbf{S}_l^H \mathbf{T}_l$$

$$\mathcal{L}_r = \sum_{l \in \Omega} \left\| \mathbf{T}_l - \hat{\mathbf{A}}_l \mathbf{S}_l \right\|_2^2$$

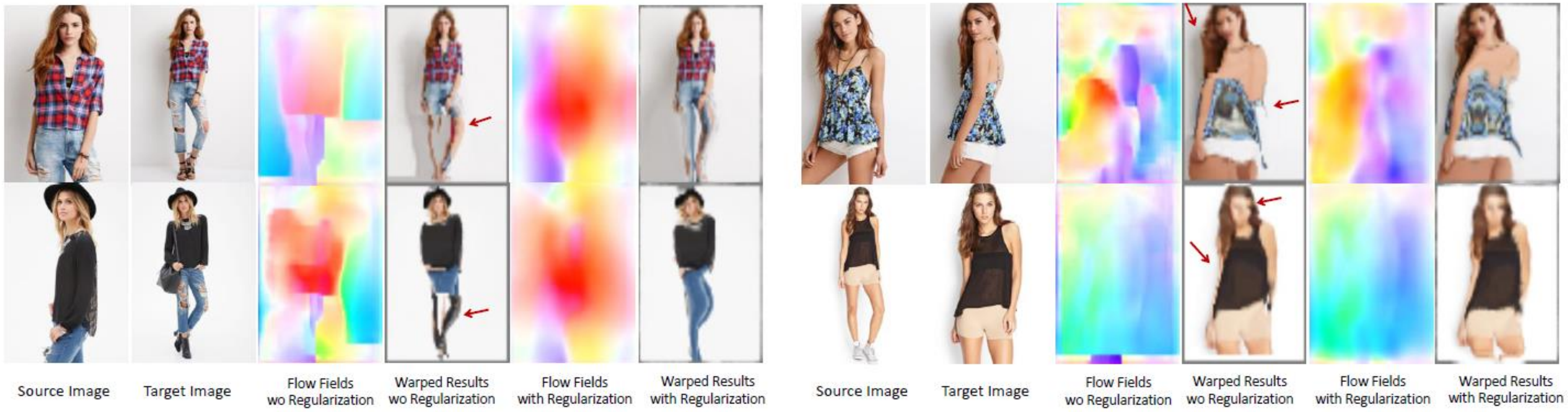
Method

- Global Flow Field Estimator F : Global flow field & Occlusion mask
  - sampling correctness loss & **regularization term**



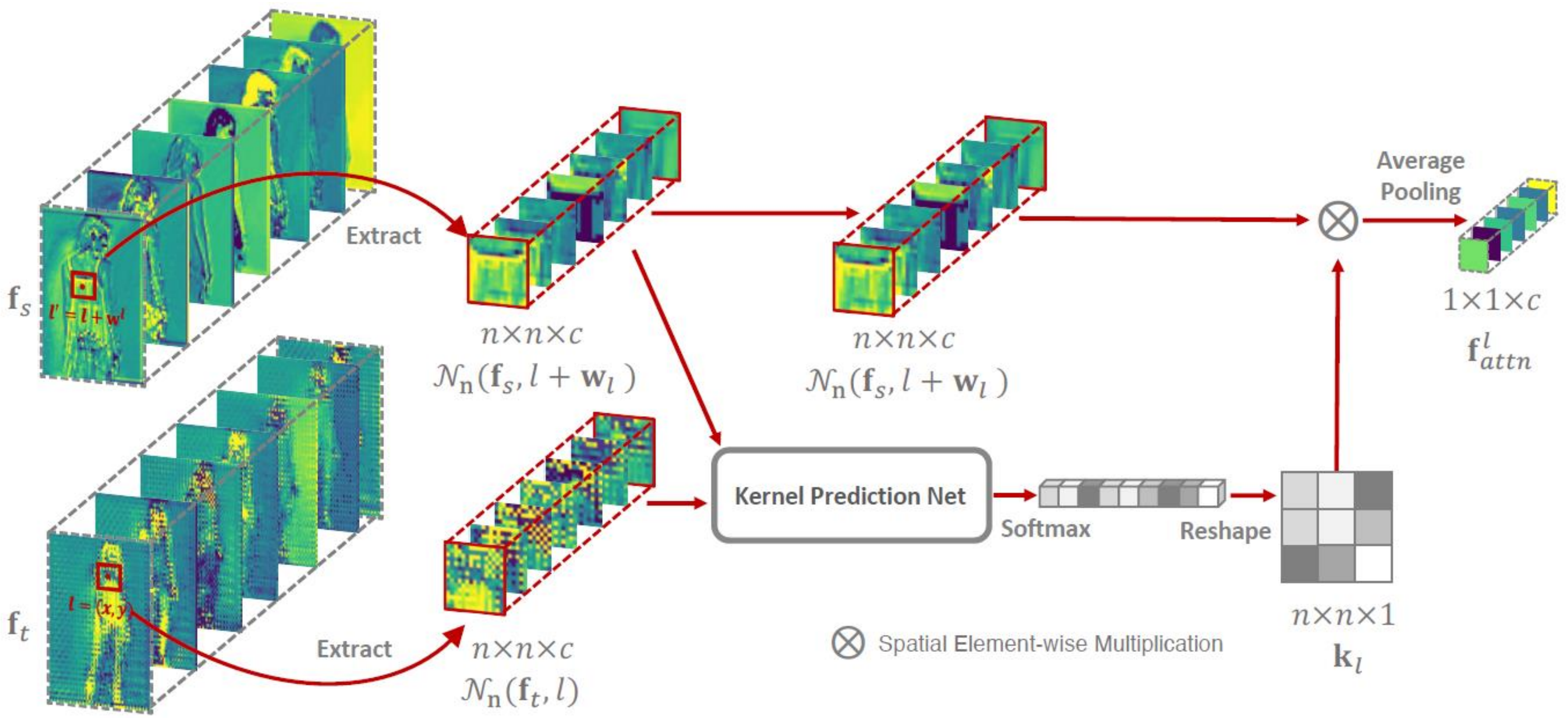


Affine transformation
: rotation, scale, translation



Method

- Local Neural Texture Renderer G



$$\mathbf{k}_l = M(\mathcal{N}_n(\mathbf{f}_s, l + \mathbf{w}^l), \mathcal{N}_n(\mathbf{f}_t, l))$$

$$\mathbf{f}^l_{attn} = P(\mathbf{k}_l \otimes \mathcal{N}_n(\mathbf{f}_s, l + \mathbf{w}^l))$$

$$\mathbf{f}_{out} = (\mathbf{1} - \mathbf{m}) * \mathbf{f}_t + \mathbf{m} * \mathbf{f}_{attn}$$

$$\mathcal{L}_{perc} = \sum_i \|\phi_i(\mathbf{x}_t) - \phi_i(\hat{\mathbf{x}}_t)\|_1$$

$$\mathcal{L}_{adv} = \mathbb{E}[\log(1 - D(G(\mathbf{x}_s, \mathbf{p}_t, \mathbf{w}, \mathbf{m})))] + \mathbb{E}[\log D(\mathbf{x}_t)]$$

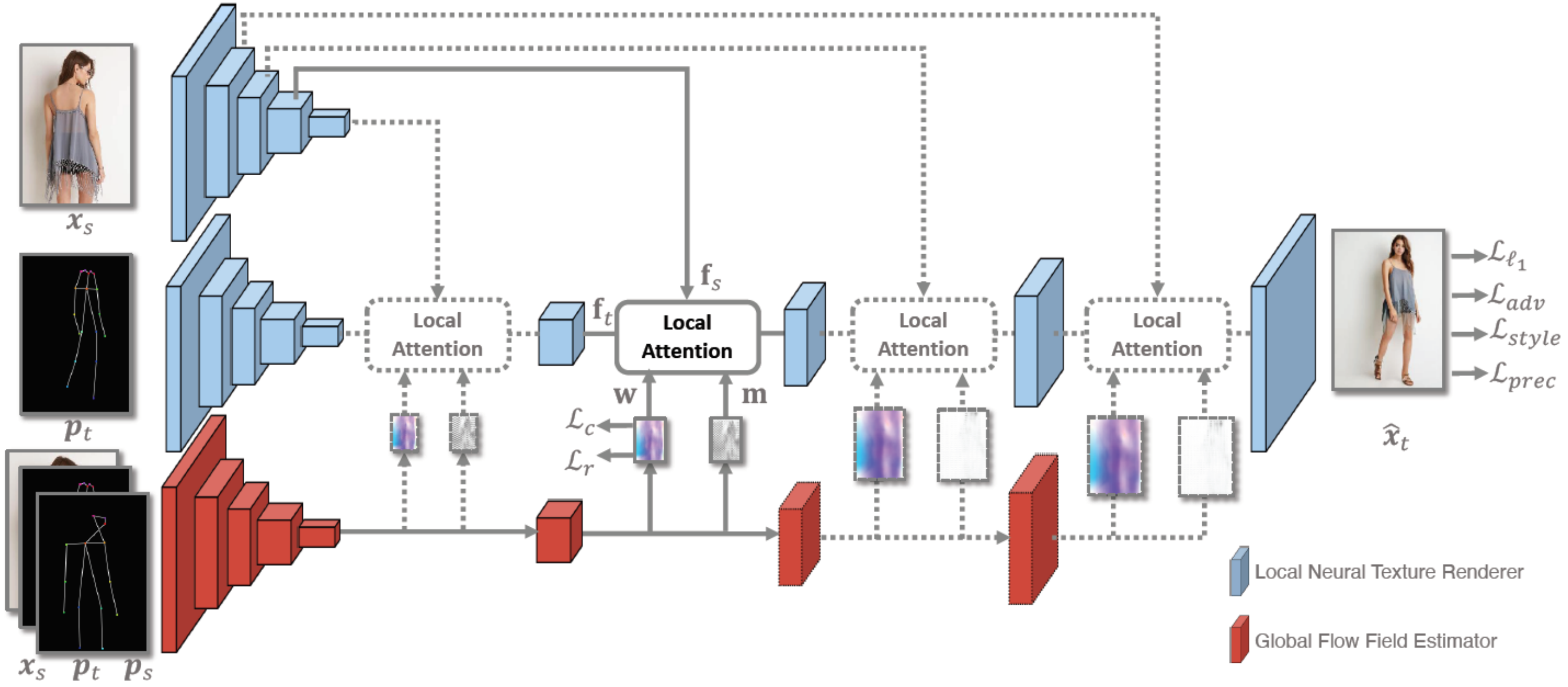
$$\mathcal{L}_{\ell_1} = \|\mathbf{x}_t - \hat{\mathbf{x}}_t\|_1$$

$$\mathcal{L}_{style} = \sum_j \left\| G_j^\phi(\mathbf{x}_t) - G_j^\phi(\hat{\mathbf{x}}_t) \right\|_1$$

$$\mathcal{L} = \lambda_c \mathcal{L}_c + \lambda_r \mathcal{L}_r + \lambda_{\ell_1} \mathcal{L}_{\ell_1} + \lambda_a \mathcal{L}_{adv} + \lambda_p \mathcal{L}_{prec} + \lambda_s \mathcal{L}_{style}$$

Method

- Total loss



$$\mathcal{L}_{perc} = \sum_i \|\phi_i(\mathbf{x}_t) - \phi_i(\hat{\mathbf{x}}_t)\|_1$$

$$\mathcal{L}_{adv} = \mathbb{E}[\log(1 - D(G(\mathbf{x}_s, \mathbf{p}_t, \mathbf{w}, \mathbf{m})))] + \mathbb{E}[\log D(\mathbf{x}_t)]$$

$$\mathcal{L}_{\ell_1} = \|\mathbf{x}_t - \hat{\mathbf{x}}_t\|_1$$

$$\mathcal{L}_{style} = \sum_j \left\| G_j^\phi(\mathbf{x}_t) - G_j^\phi(\hat{\mathbf{x}}_t) \right\|_1$$

$$\mathcal{L} = \lambda_c \mathcal{L}_c + \lambda_r \mathcal{L}_r + \lambda_{\ell_1} \mathcal{L}_{\ell_1} + \lambda_a \mathcal{L}_{adv} + \lambda_p \mathcal{L}_{prec} + \lambda_s \mathcal{L}_{style}$$

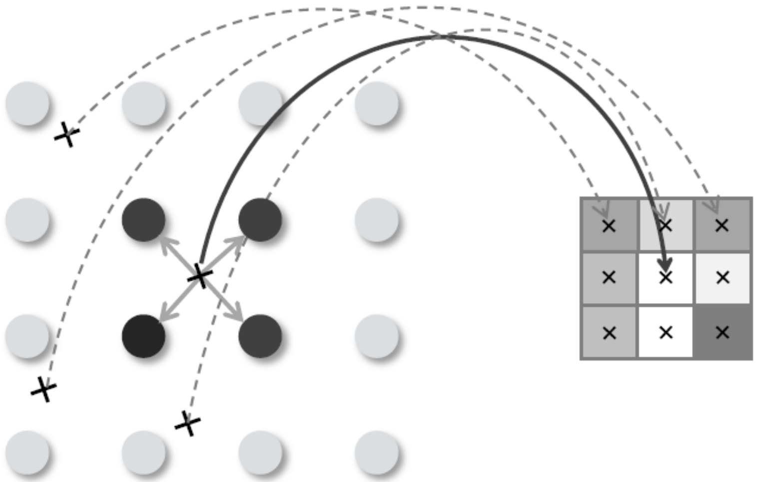
Method

- Bilinear interpolation sampling의 한계

$$\mathbf{f}_{out}^{x,y} = (\lceil\Delta y\rceil - \Delta y)(\lceil\Delta x\rceil - \Delta x)\mathbf{f}_{in}^{x',y'} + (\Delta y - \lfloor\Delta y\rfloor)(\Delta x - \lfloor\Delta x\rfloor)\mathbf{f}_{in}^{x'+1,y'+1}$$
$$+ (\Delta y - \lfloor\Delta y\rfloor)(\lceil\Delta x\rceil - \Delta x)\mathbf{f}_{in}^{x',y'+1} + (\lceil\Delta y\rceil - \Delta y)(\Delta x - \lfloor\Delta x\rfloor)\mathbf{f}_{in}^{x'+1,y'}$$

$$\frac{\partial \mathbf{f}_{out}^{x,y}}{\partial \Delta x} =(\lceil\Delta y\rceil - \Delta y)(\mathbf{f}_{in}^{x'+1,y'} - \mathbf{f}_{in}^{x',y'}) + (\Delta y - \lfloor\Delta y\rfloor)(\mathbf{f}_{in}^{x'+1,y'+1} - \mathbf{f}_{in}^{x',y'+1})$$

$$\frac{\partial \mathbf{f}_{out}^{x,y}}{\partial \mathbf{f}_{in}^{x',y'}} = (\lceil\Delta y\rceil - \Delta y)(\lceil\Delta x\rceil - \Delta x)$$

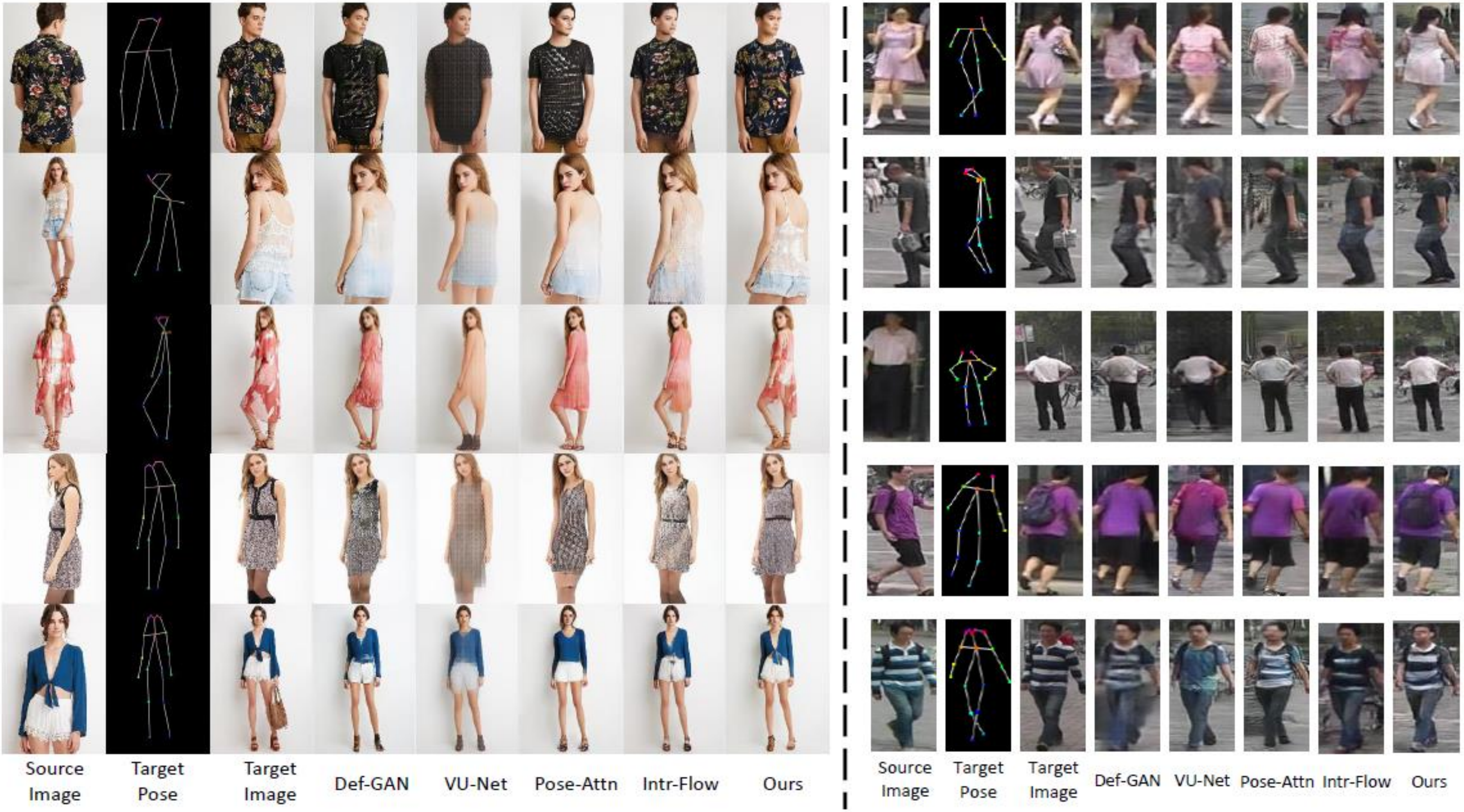


- Flow fields는 올바른 gradient를 얻기 위해서 reasonable input features가 필요함. 만약, input features 가 meaningless 하다면 correct flow fields를 얻을 수 없음.

- Input features 또한 마찬가지로 correct flow fields 없이는 reasonable gradients를 얻을 수 없음.

- 비록 pre-training을 통해 meaningful input feature를 뽑는다고 하더라도, 인접한 픽셀 간에 high-correlation ($\mathbf{f}_{in}^{x',y'} \approx \mathbf{f}_{in}^{x',y'+1}$)이 자주 발생한다. 따라서, gradients는 대부분의 위치에서 작고 large motion을 잡아내기 어려움.

Experiments



Experiments

	Flow-Based	Content-aware Sampling	FID	LPIPS
Baseline	N	-	16.008	0.2473
Global-Attn	N	-	18.616	0.2575
Bi-Sample	Y	N	12.143	0.2406
Full Model	Y	Y	**10.573**	**0.2341**

Table 2. The evaluation results of the ablation study.