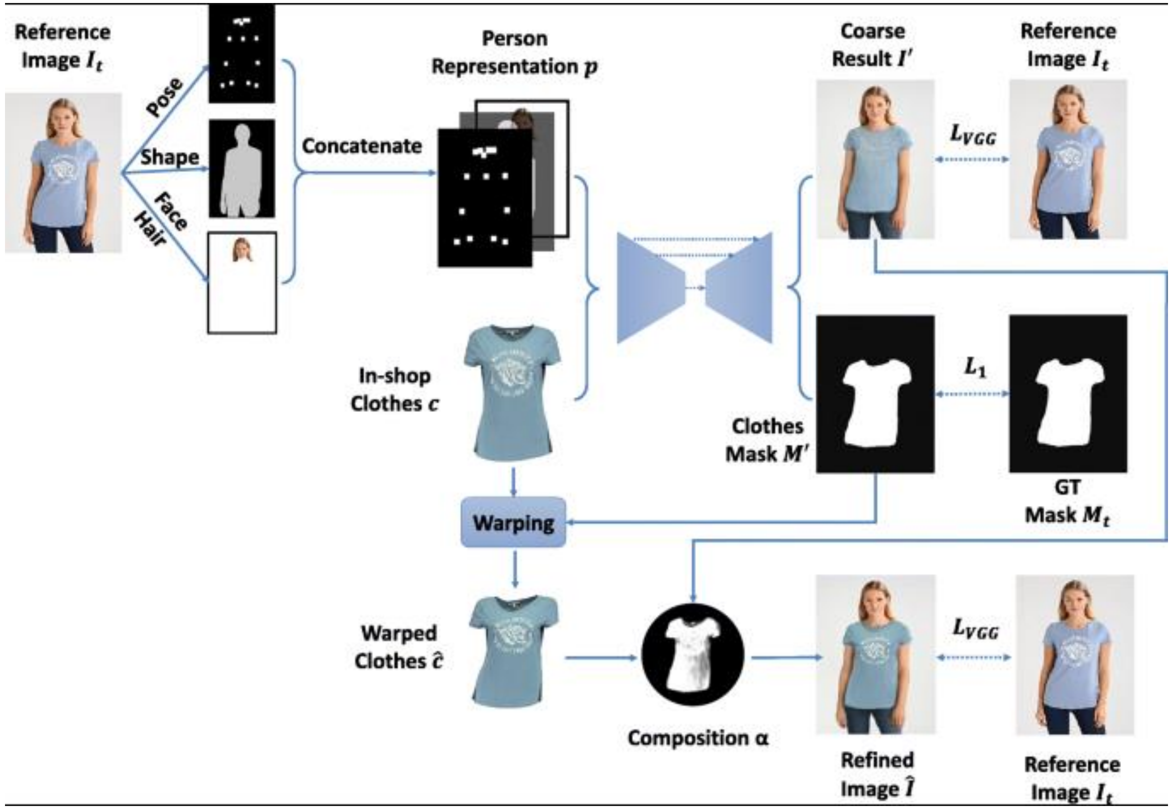
Weakly Supervised High-Fidelity Clothing Model Generation

Ruili Feng, Cheng Ma, Chengji Shen, Xin Gao, Zhenjiang Liu ,Xiaobo Li, Kairi Ou, Zhengjun Zha

University of Science and Technology of China   Zhejiang University   Alibaba Group

CVPR 2022

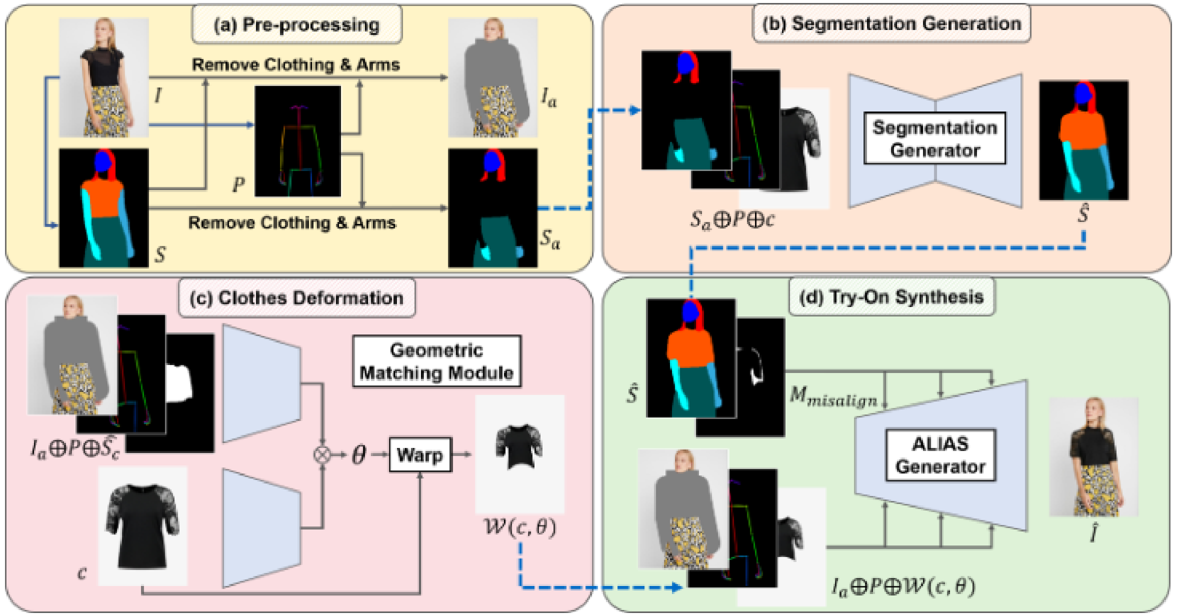
Virtual Try-on

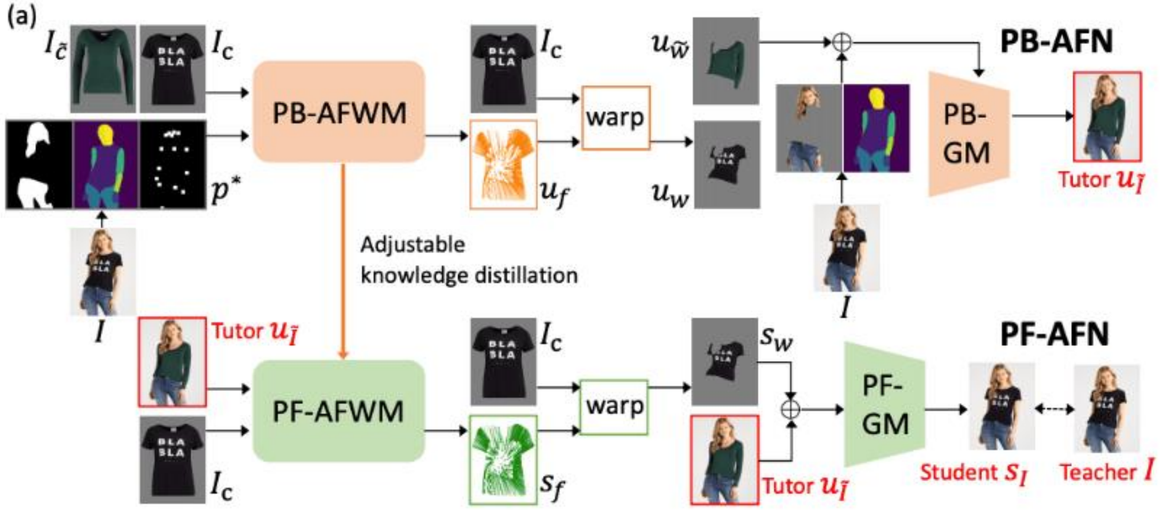


- VITON needs paired data
- Supervised learning

Weakly Supervised High-Fidelity Clothing Model Generation, Feng et al., CVPR 2022.

Virtual Try-on

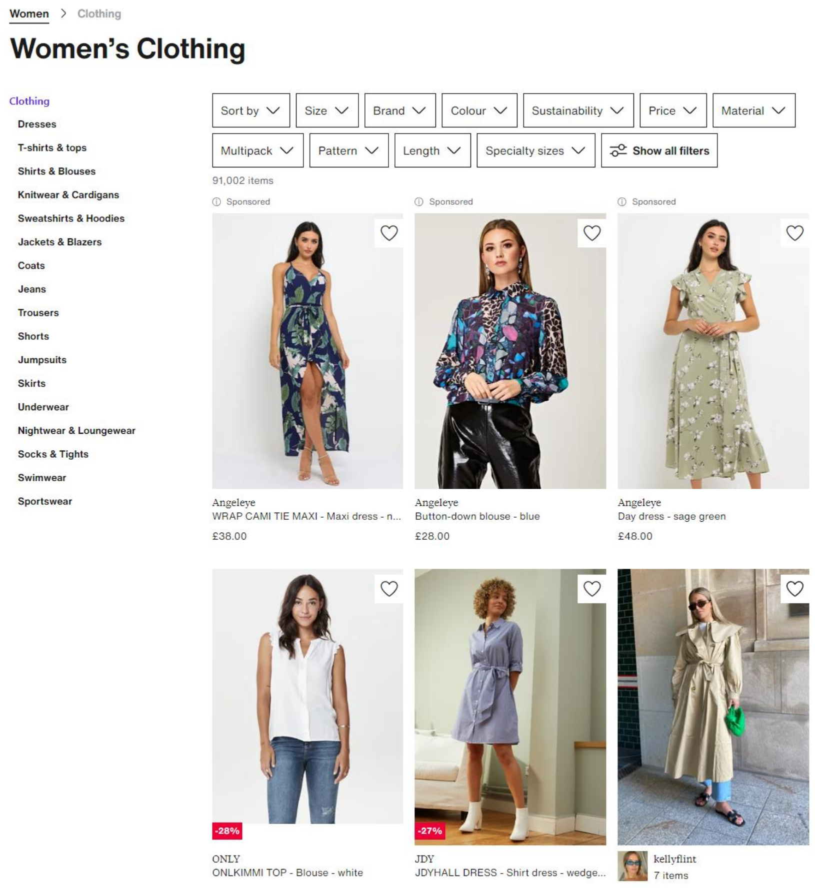




- “Parser-based” or “Parser-free”
- Warping + Synthesis

Motivation

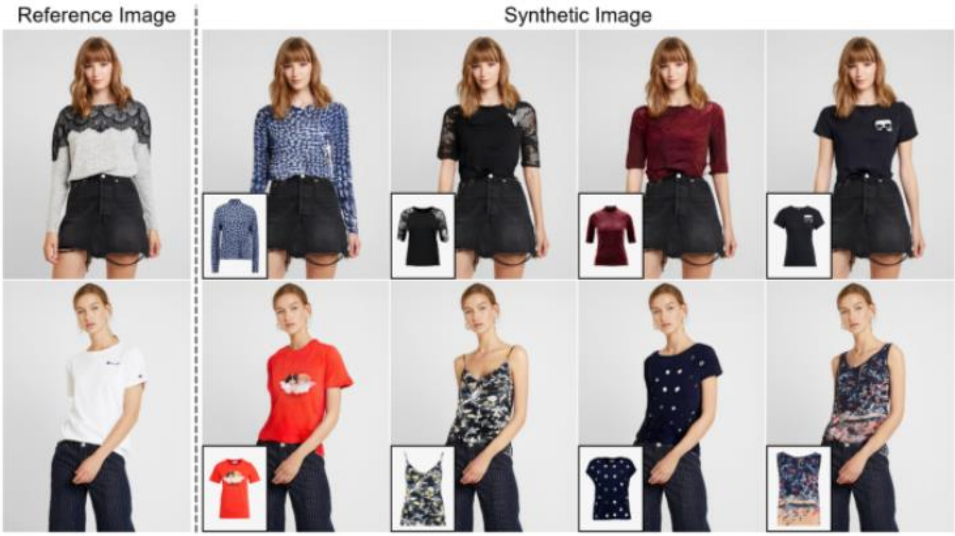
- Proprietary model images are expensive!
- Training VITON model with large dataset -> high-fidelity
- Weakly supervised learning



Zalando website

Task setting

- Conventional VITON setting

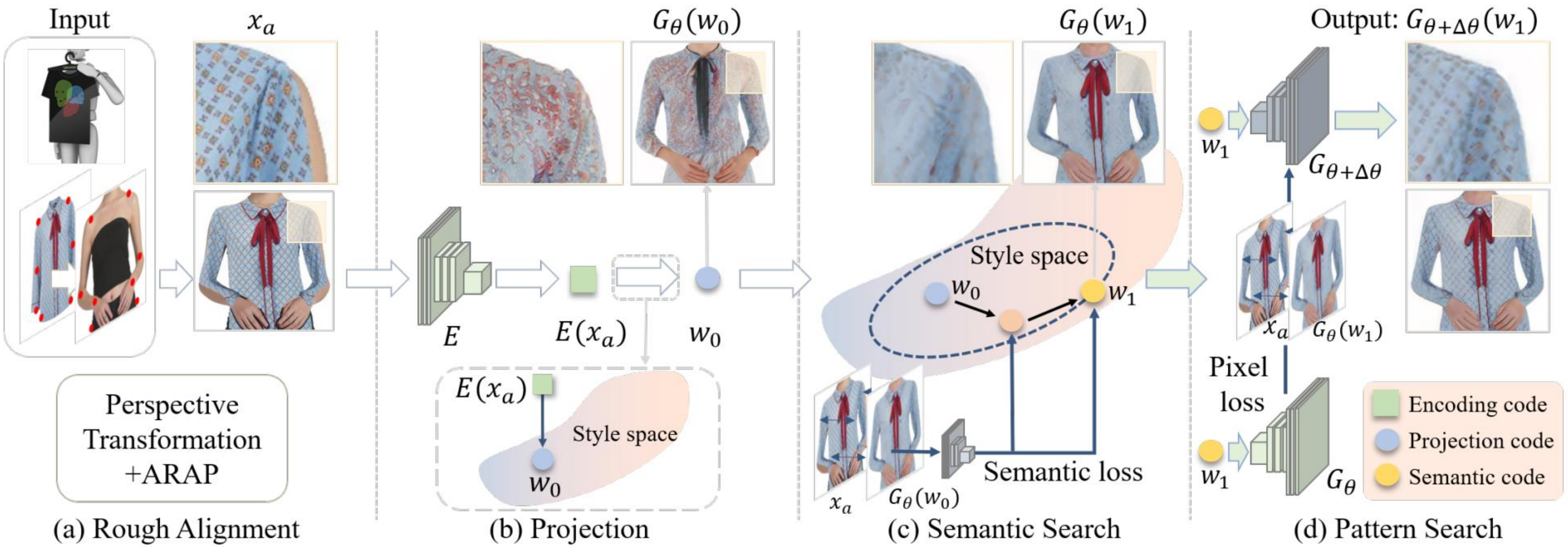


- Proposed setting

- The procedure of people predicting how they will look like while picking clothes
- Commercial Model Image dataset (CMI) – 2,348 images of models on underwear or sleeveless

Weakly Supervised High-Fidelity Clothing Model Generation, Feng et al., CVPR 2022.

Overview



Input
x_a
Perspective Transformation +ARAP
(a) Rough Alignment
$G_\theta(w_0)$
E
$E(x_a)$
w_0
Style space
(b) Projection
$G_\theta(w_1)$
Style space
w_0
w_1
x_a
$G_\theta(w_0)$
Semantic loss
(c) Semantic Search
Output: $G_{\theta+\Delta\theta}(w_1)$
w_1
$G_{\theta+\Delta\theta}$
x_a
$G_\theta(w_1)$
Pixel loss
G_θ
Encoding code
Projection code
Semantic code
(d) Pattern Search

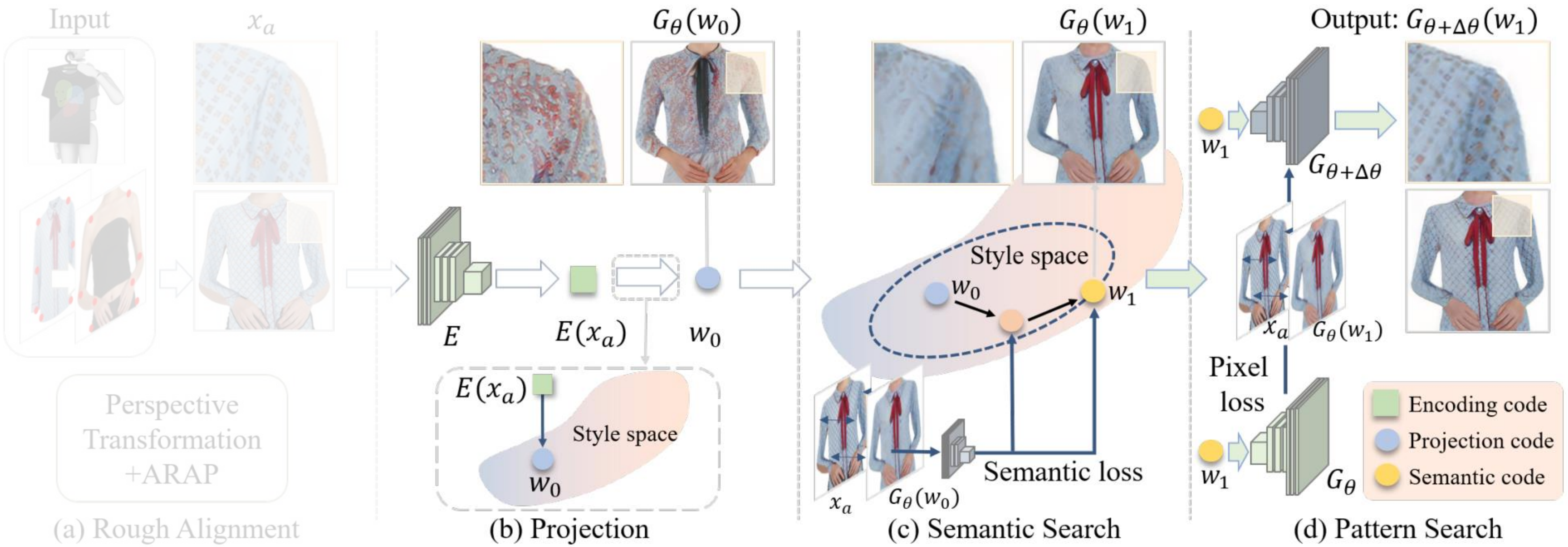
Rough Alignment

(a) Perspective transformation

(b) ARAP

- Get a rough alignment image x_a

Deep Generative Projection

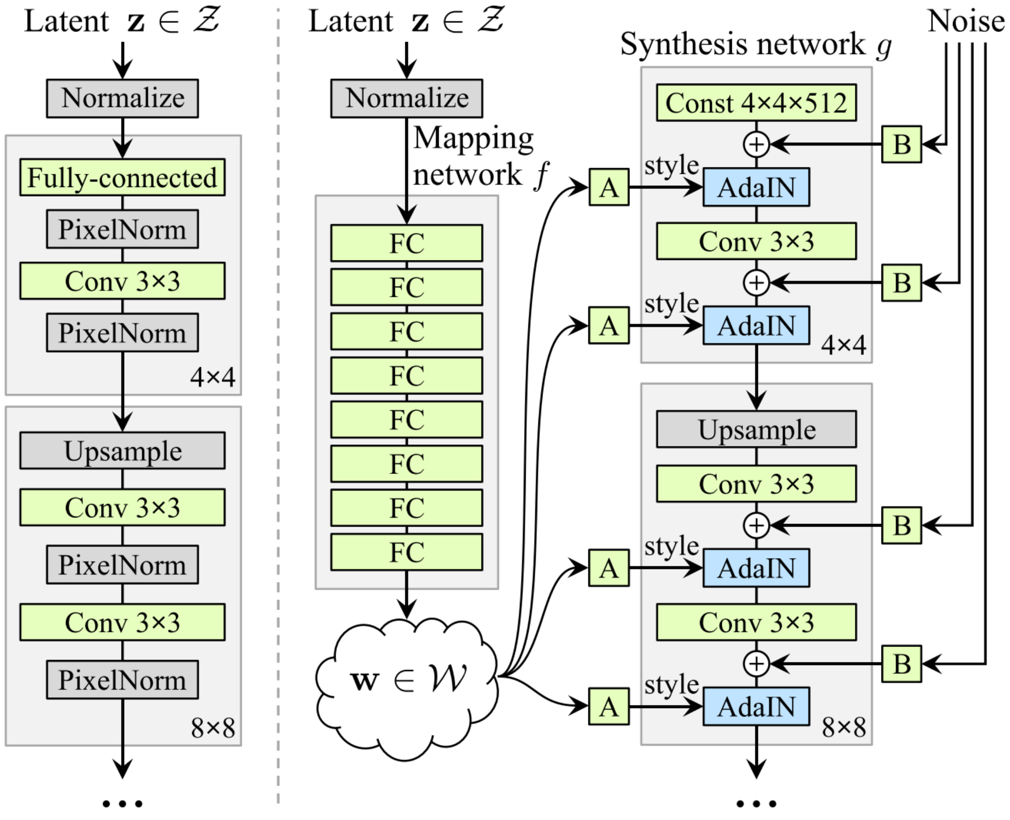


Projection

- Rough alignment -> Image with high-level semantics (clothing category, model pose)
- Refine flaw of the rough alignment

Projection

- Train pre-trained StyleGAN
- E-Shop Fashion(ESF) dataset
  - 180,000 clothing model images
  - 512 x 512
- FID 2.16

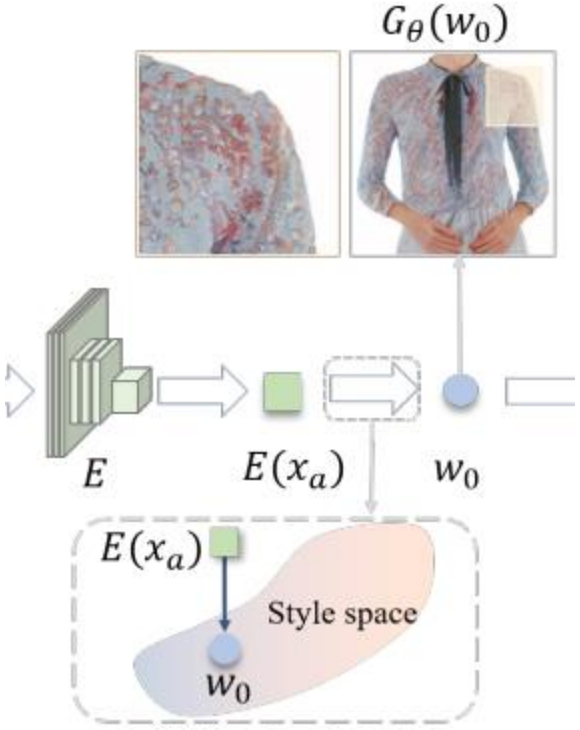


(a) Traditional (b) Style-based generator

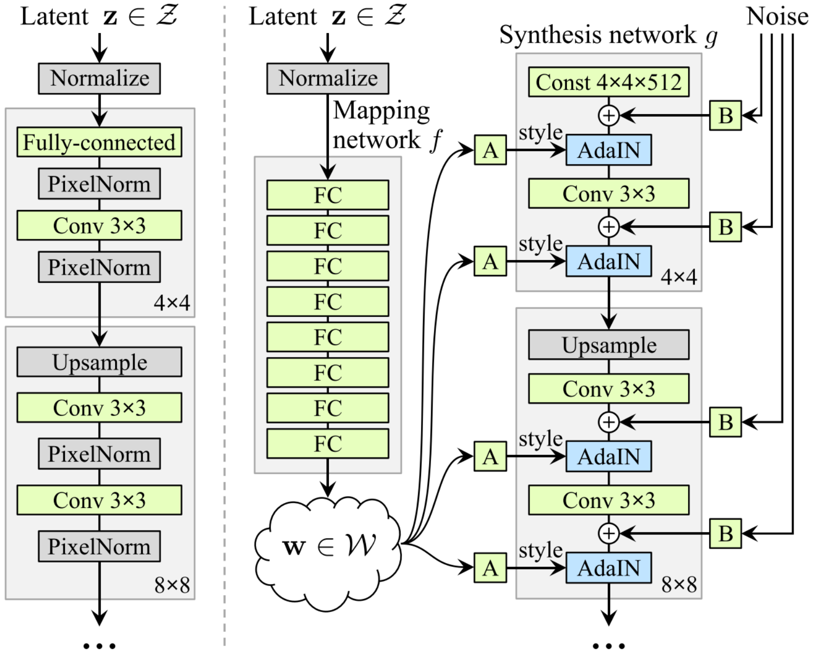
Projection

- Style space $\mathcal{W}+$
- Embedding images into high-density region of style space

  -> To yield much more plausible synthesis



(b) Projection



Projection

- Encoder $\boldsymbol{E}$

1. Compute PCA decomposition of 5M points on $\mathcal{W}+$
2. Get the mean value, covariance matrix, set of principal components

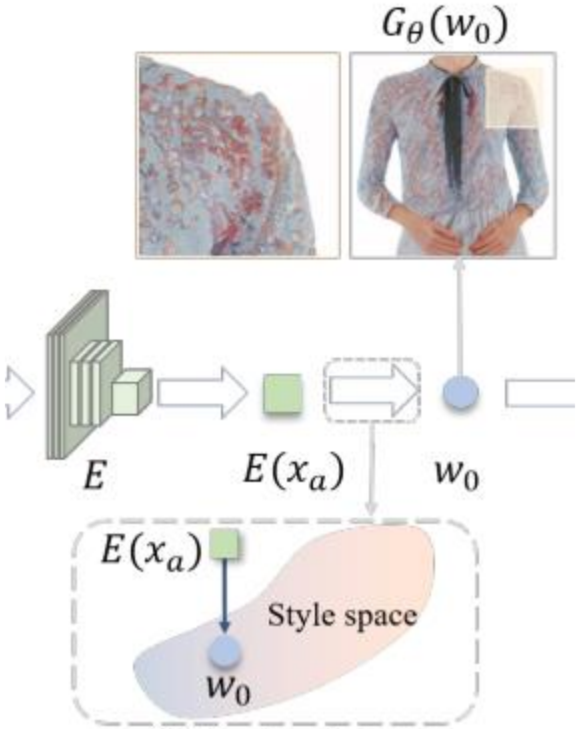
   $\boldsymbol{\mu}$ $\boldsymbol{\Sigma}$ $\boldsymbol{Q} = (\boldsymbol{q}_1, ..., \boldsymbol{q}_n)$

3. Predicts series of principal strengths $\boldsymbol{s} = (s_1, s_2, ..., s_n)^T$

$$\begin{aligned} \boldsymbol{s} &= \boldsymbol{E}(\boldsymbol{x}_a), \\ \boldsymbol{w}_0 &= Tr(\boldsymbol{q_1} s_1 \sqrt{\sigma_1} + ...\boldsymbol{q}_n s_n \sqrt{\sigma_n}) + \boldsymbol{\mu} \\ &= \boldsymbol{Q}\boldsymbol{\Lambda}^{\frac{1}{2}} Tr(\boldsymbol{s}) + \boldsymbol{\mu}, \end{aligned}$$

$$Tr(\boldsymbol{v}) = \begin{cases} \boldsymbol{v}, \|\boldsymbol{v}\|_2 < \psi, \\ \frac{\boldsymbol{v}}{\|\boldsymbol{v}\|_2}\psi, \|\boldsymbol{v}\|_2 \geq \psi. \end{cases}$$

$$\boldsymbol{w} = \boldsymbol{P}(\boldsymbol{x}_a)$$

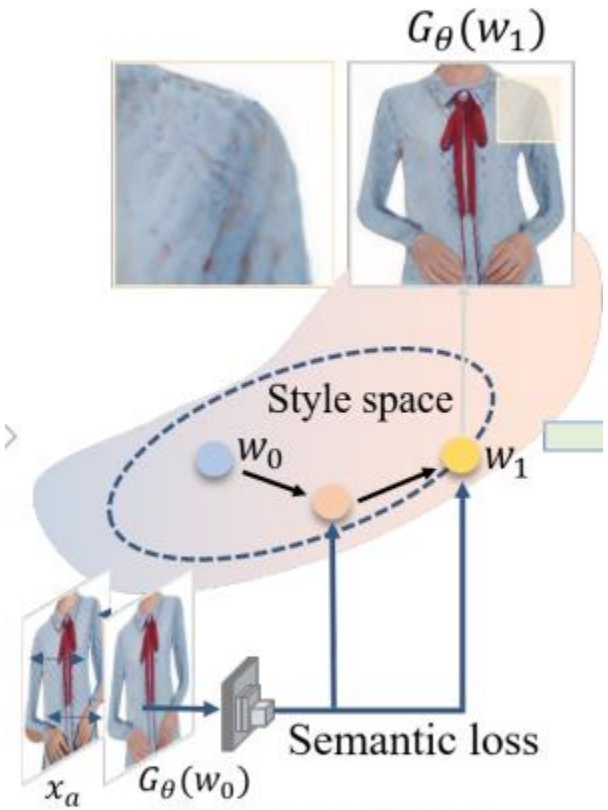


(b) Projection

Semantic search

- Optimization problem

$$\min_{\boldsymbol{w} \in \mathcal{C}} \eta_p l_p + \eta_f l_f + \eta_{attr} l_{attr} + \eta_{adv} l_{adv},$$
$$l_p = \|W * \boldsymbol{G}(\boldsymbol{w}) - W * \boldsymbol{x}_a\|_2^2,$$
$$l_f = \|\boldsymbol{V}(W * \boldsymbol{G}(\boldsymbol{w})) - \boldsymbol{V}(W * \boldsymbol{x}_a)\|_2^2,$$
$$l_{attr} = \|\boldsymbol{R}(W * \boldsymbol{G}(\boldsymbol{w})) - \boldsymbol{R}(W * \boldsymbol{x}_a)\|_2^2,$$
$$l_{adv} = \log[1 - \boldsymbol{D}(\boldsymbol{G}(\boldsymbol{w}))],$$



(c) Semantic Search

Semantic search

- Optimization problem

$$
\begin{aligned}
\min_{\boldsymbol{w}\in\mathcal{C}} \eta_p l_p + \eta_f l_f + \eta_{attr} l_{attr} + \eta_{adv} l_{adv}, \\
l_p = \|W * \boldsymbol{G}(\boldsymbol{w}) - W * \boldsymbol{x}_a\|_2^2, \\
l_f = \|\boldsymbol{V}(W * \boldsymbol{G}(\boldsymbol{w})) - \boldsymbol{V}(W * \boldsymbol{x}_a)\|_2^2, \\
l_{attr} = \|\boldsymbol{R}(W * \boldsymbol{G}(\boldsymbol{w})) - \boldsymbol{R}(W * \boldsymbol{x}_a)\|_2^2, \\
l_{adv} = \log[1 - \boldsymbol{D}(\boldsymbol{G}(\boldsymbol{w}))],
\end{aligned}
$$

Rough Alignment　　Projection

$$
W_{ij} = \begin{cases} 1 - \exp(-d((i,j), \partial I)^2), & I(ij) = 1, \\ 0, & I(ij) = 0. \end{cases}
$$

Semantic search

- Optimization problem

$$
\begin{aligned}
\min_{\boldsymbol{w}\in\mathcal{C}} \eta_p l_p + \eta_f l_f + \eta_{attr} l_{attr} + \eta_{adv} l_{adv}, \\
l_p = \|W * \boldsymbol{G}(\boldsymbol{w}) - W * \boldsymbol{x}_a\|_2^2, \\
l_f = \|\boldsymbol{V}(W * \boldsymbol{G}(\boldsymbol{w})) - \boldsymbol{V}(W * \boldsymbol{x}_a)\|_2^2, \\
l_{attr} = \|\boldsymbol{R}(W * \boldsymbol{G}(\boldsymbol{w})) - \boldsymbol{R}(W * \boldsymbol{x}_a)\|_2^2, \\
l_{adv} = \log[1 - \boldsymbol{D}(\boldsymbol{G}(\boldsymbol{w}))],
\end{aligned}
$$

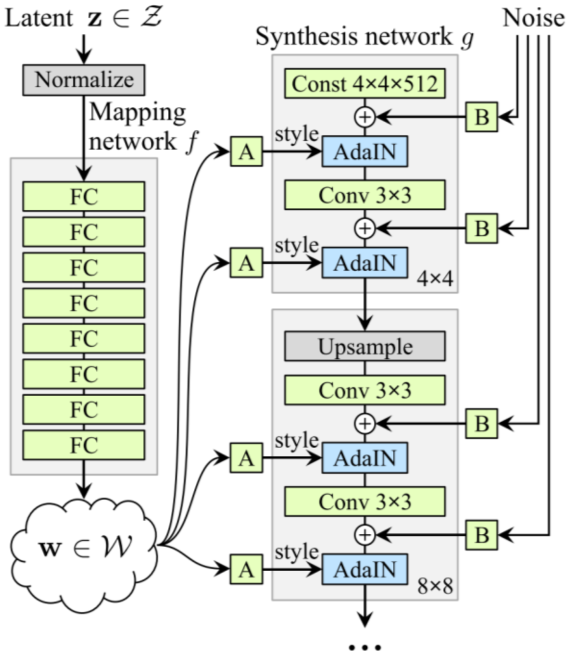
- Constrain $\mathcal{C}$

$$
\begin{aligned}
\boldsymbol{w}_{k+1} &= \arg\min_{\boldsymbol{w}\in\mathcal{C}} \|\boldsymbol{w}_{k+1} - \boldsymbol{w}\| \\
&= \begin{cases} \boldsymbol{w}_0 + 4\frac{\boldsymbol{w}_{k+1}-\boldsymbol{w}}{\|\boldsymbol{w}_{k+1}-\boldsymbol{w}\|_2}, & \|\boldsymbol{w}_{k+1} - \boldsymbol{w}_m\|_2 > 4, \\ \boldsymbol{w}_{k+1}, & \|\boldsymbol{w}_{k+1} - \boldsymbol{w}\|_2 \leq 4. \end{cases}
\end{aligned}
$$

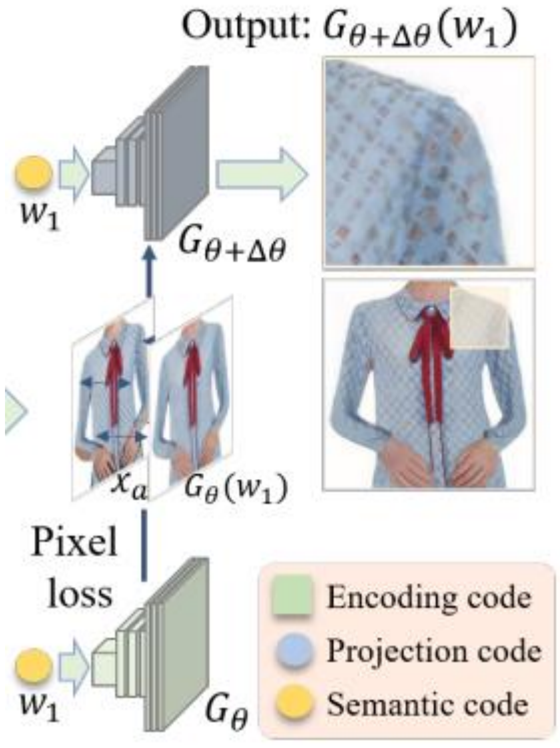
Pattern search

- Optimization problem

$$\min_{\boldsymbol{\theta}\in B(\boldsymbol{\theta}_0,4)} \eta_p \|W * \boldsymbol{G}_{\boldsymbol{\theta}}(\boldsymbol{w}) - W * \boldsymbol{x}_a\|_2 + \log(1 - \boldsymbol{D}(\boldsymbol{G}_{\boldsymbol{\theta}}(\boldsymbol{w}))),$$



(b) Style-based generator



(d) Pattern Search

Experiments

Table 1. Numerical metrics of DGP, ACGPN, PF-AFN, and VITON-HD on CMI and MPV datasets. ↓ indicates lower is better.

Methods	**CMI**		**MPV**	
	FID↓	**SWD↓**	**FID↓**	**SWD↓**
ACGPN	137.9	121.3	81.1	90.4
PF-AFN	97.3	76.7	67.8	67.1
VITON-HD	87.5	56.1	**40.6**	52.7
DGP (Ours)	**51.6**	**22.4**	48.4	**36.7**

Experiments

Figure 8. Comparison on the CMI and MPV datasets. The supervised competitor methods are basically less appealing, and perform especially poorly on complicated clothing like coats.

Experiments

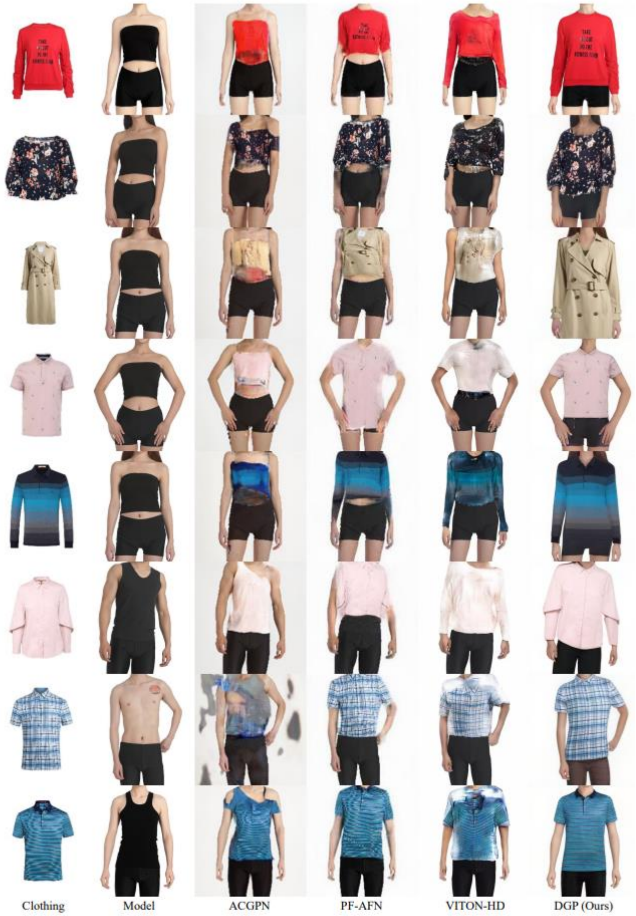


Figure S10. More visual results of qualitative comparison on the CMI dataset.