



# Fashion U Want

의류 Segmentation 및 서비스화

프로젝트 주제

Preprocessing

1\_OpenPose

2\_Human Parse

3\_Dense Pose

4\_Cloth Mask

5\_Parse Agnostic

6\_Human Agnostic

TRYON

Clothing-Agnostic Person Representation

Segmentation Generation(GS/모델: U-Net)

Clothing Image Deformation

Try-On Synthesis via ALIAS normalization

기존 GAN 기반 Try-On 모델의 한계점

MV-VTON 프로젝트로의 전환 배경

MV-VTON의 기대 효과

## 의류 Segmentation 및 서비스화

◆ **프로젝트 소개**



## 프로젝트 주제

온라인 쇼핑몰은 제품 정보를 제공하는 데 있어 점점 더 혁신적인 방법을 찾고 있으며, 고객들이 온라인에서 제품을 선택하는 과정에서 **구매 결정에 대한 불확실성을 줄이는 것이 중요한 과제**가 되었습니다. 가상 피팅 서비스는 고객의 신체에 맞는 의류를 실시간으로 확인하게 하여 반품율을 줄이고 구매 전환율을 높이는 데 기여합니다. 본 프로젝트는 **VITON 데이터셋을 기반으로 한 의류 Segmentation 모델을 개발하여 가상 피팅 서비스의 기반을 마련하고자 합니다.**

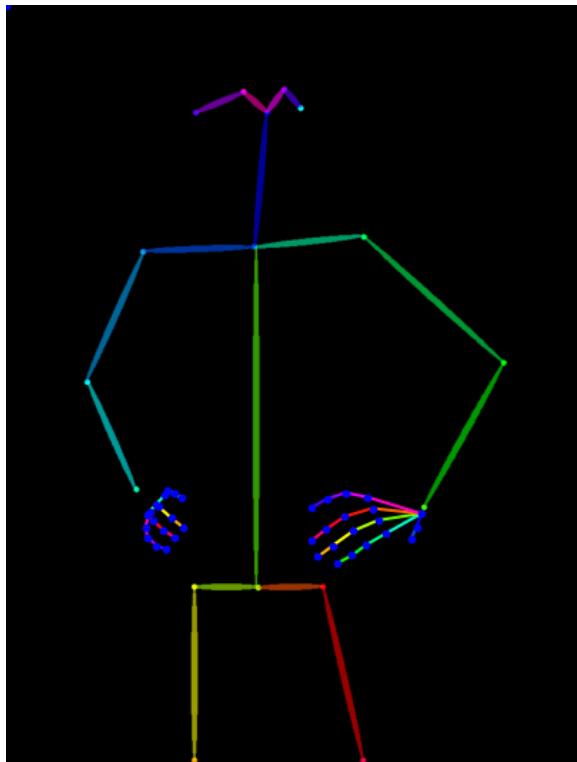
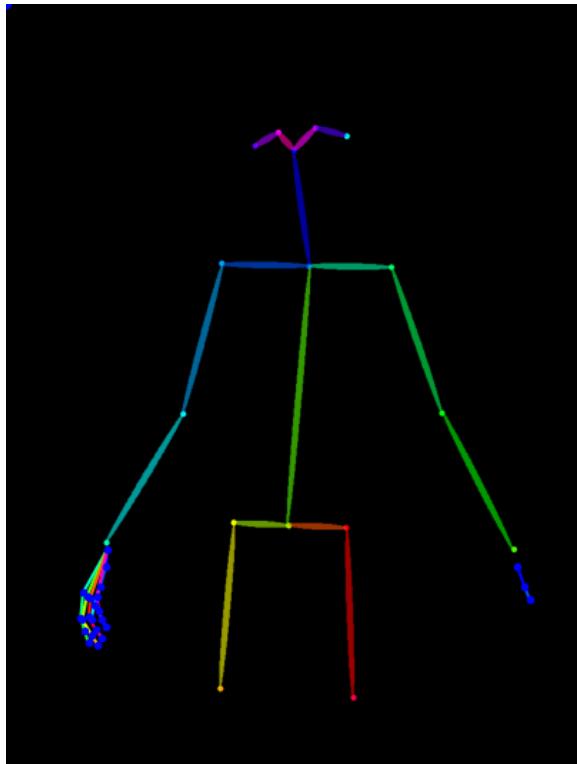
## Preprocessing

- TRYON 모델을 사용하기 위해서는 VITON-HD dataset만을 사용
  - 하지만, 범용성을 위해 실제로 VITON-HD와 같은 형식의 **옷 입은 모델의 이미지와 의류 이미지를 수집하여 Preprocessing과정을 통해 TRYON을 사용하고자 함**
- ⇒ 실제 데이터를 수집하고 정제하면 TRYON을 통해 가상 피팅 서비스 제공이 가능

### 1\_OpenPose

- **Body 25** : 사람의 포즈를 추정하기 위해 사용되는 포인트 기반 모델
- **Hand Keypoints** : 왼손과 오른손 각각 **21개의 키포인트**로 구성
- **openpose\_img, openpose\_json** 좌표



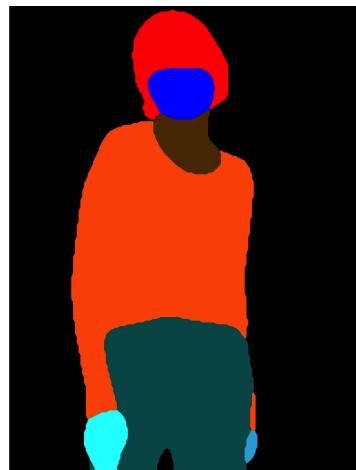


## 2\_Human Parse

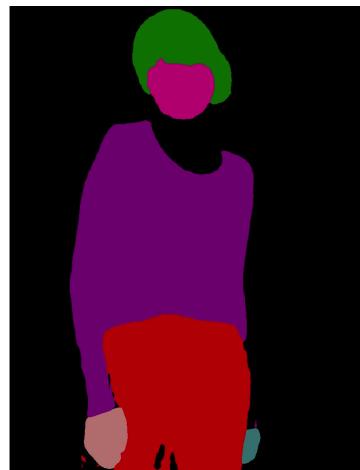
- 이미지에서 사람의 몸과 옷 등의 다양한 부위를 **픽셀 단위**로 segmentation
- **Self Correction for Human Parsing Model** 사용
  - LIP(Dataset)으로 train된 pretrained model 사용
  - label 20개로 1개빼고는 VITON-HD와 label이 모두 같아서 LIP 선택



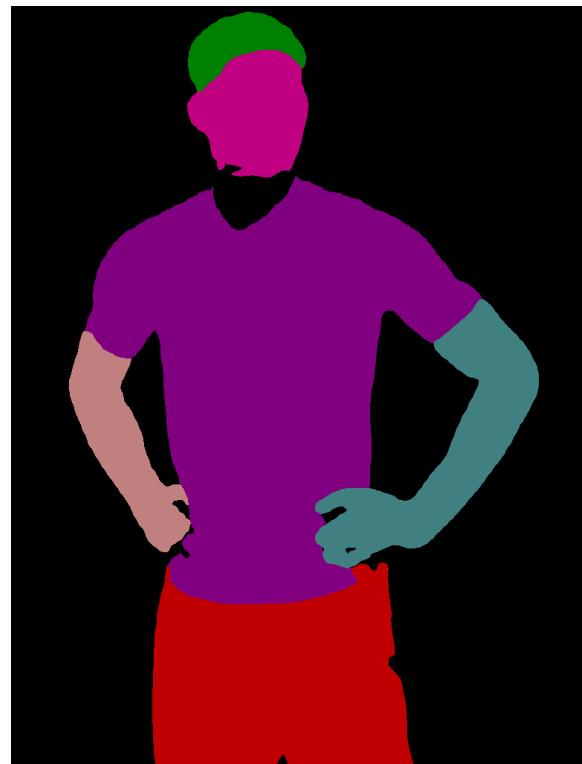
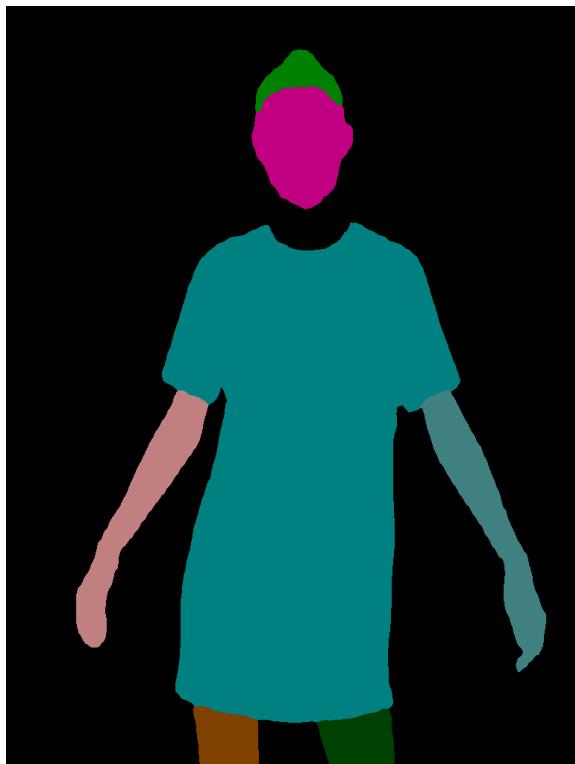
(a) original image



(b) viton-hd image



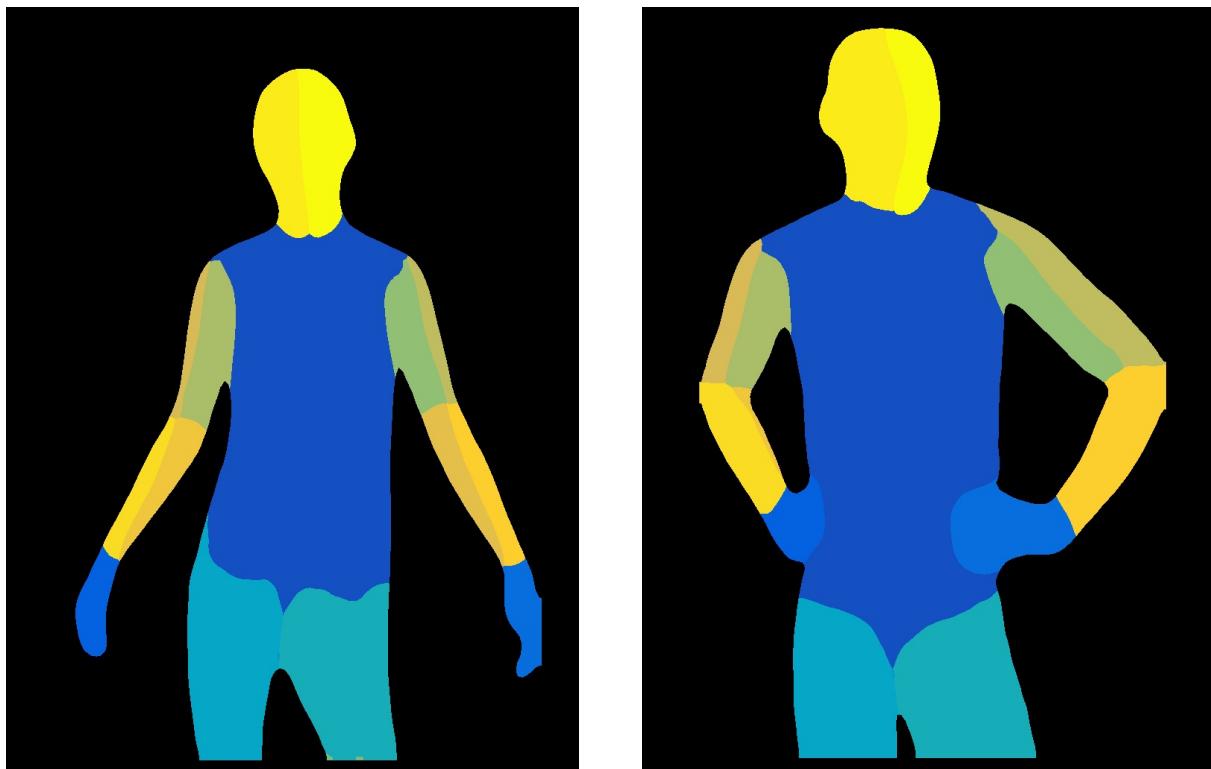
(c) self-correction image



### 3\_Dense Pose

- Detectron2의 DensePose를 사용하여 **DensePose Segmentation**
- 인체를 **24개 Segmentation**(머리, 팔, 다리 등)로 segmentation





## 4\_Cloth Mask

- 주어진 의류이미지 처리해 배경과 분리된 흑백 마스크 이미지를 생성
  - model : U2Net, Tracer-b7(EfficientNet B7)
- ⇒ [carvekit\\_colab](#) 라이브러리를 통해 **U2Net** 사용
- carveket 라이브러리는 pretrained model이 이미 경량화되어있기 때문에 새롭게 training시킬 필요 없다고 판단

	<b>U-Net</b>	<b>U2-Net</b>	<b>Tracer-B7</b>
<b>목적</b>	Segmentation	세밀한 경계 처리, 고해상도 Seg	복잡한 경계, 물체 분리용 고성능 Seg
<b>구조</b>	기본 Encoder-Decoder	Nested U-Net 구조 (Recursive)	EfficientNet-B7 백본 Model
<b>특징 학습 방식</b>	단일 해상도	다중 해상도 학습	고성능 특징 추출과 고해상도 경계 학습
<b>복잡한 경계 처리</b>	상대적으로 제한적	복잡한 경계 및 세밀한 구조에 강점	경계가 복잡하거나 얇은 물체 처리에 최적화

적용 사례	의료 영상, 일반 Seg	머리카락 분리, 배경 제거, 복잡한 물체 Seg	배경 제거, 고해상도 물체 분리, 세밀한 경계탐지
-------	---------------	-------------------------------	--------------------------------

- cloth image (768×1024)



## 5\_Parse Agnostic

- 옷(상의)을 입고 있는 사람의 이미지에서 옷과 신체를 분리하고 신체 구조를 기반으로 옷(상의)을 제거한 형태의 신체 마스크를 생성
- “Parse Agnostic” ⇒ 어떤 옷(상의)라도 입힐 수 있게 신체 구조만 남기므로, 특정 의류에 종속되지 않고, 어떤 의류를 넣어도 무관하게 작업할 수 있는 중립적인 상태의 이미지를 만들어 냄.

### INPUT `data_path` 구조

```

test/
└── image/                      # 입력 이미지 (사람 사진)
    ├── 0001.jpg
    ├── 0002.jpg
    └── ...
└── openpose_json/              # OpenPose에서 생성된 keypoints 데이터
    ├── 0001_keypoints.json
    ├── 0002_keypoints.json
    └── ...
└── image-parse-v3/             # 사람 파싱 결과 이미지 (PNG 형식)

```

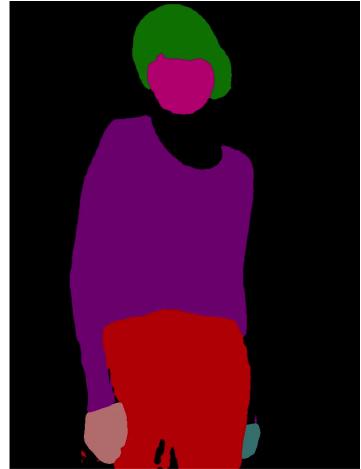
```
|   └── 0001.png  
|   └── 0002.png  
|   ...  
└── parse/                                # 최종적으로 생성될 agnostic 데이터 (ou
```



(a) original image

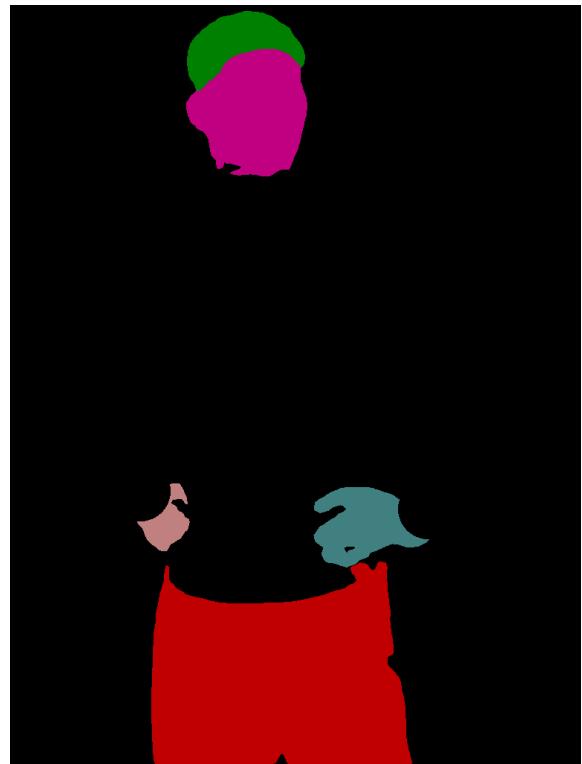
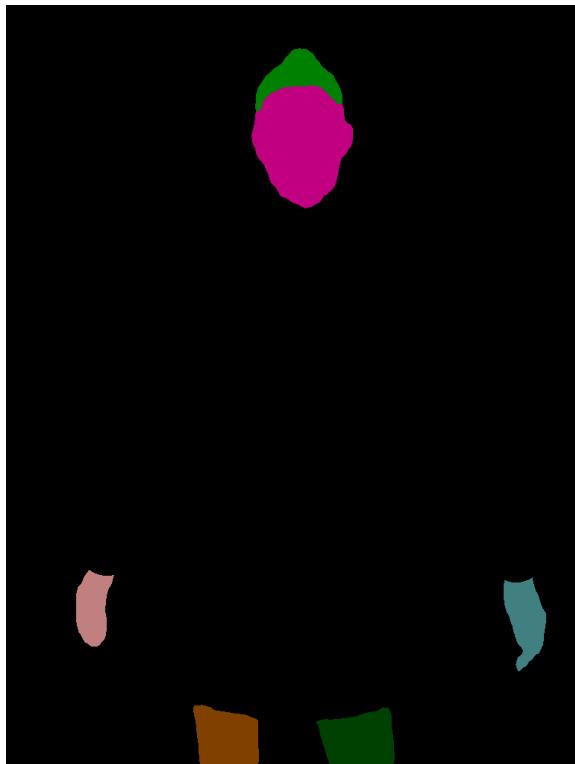


(b) pose json data



(c) self-correction image

- **Agnostic 신체 마스크**

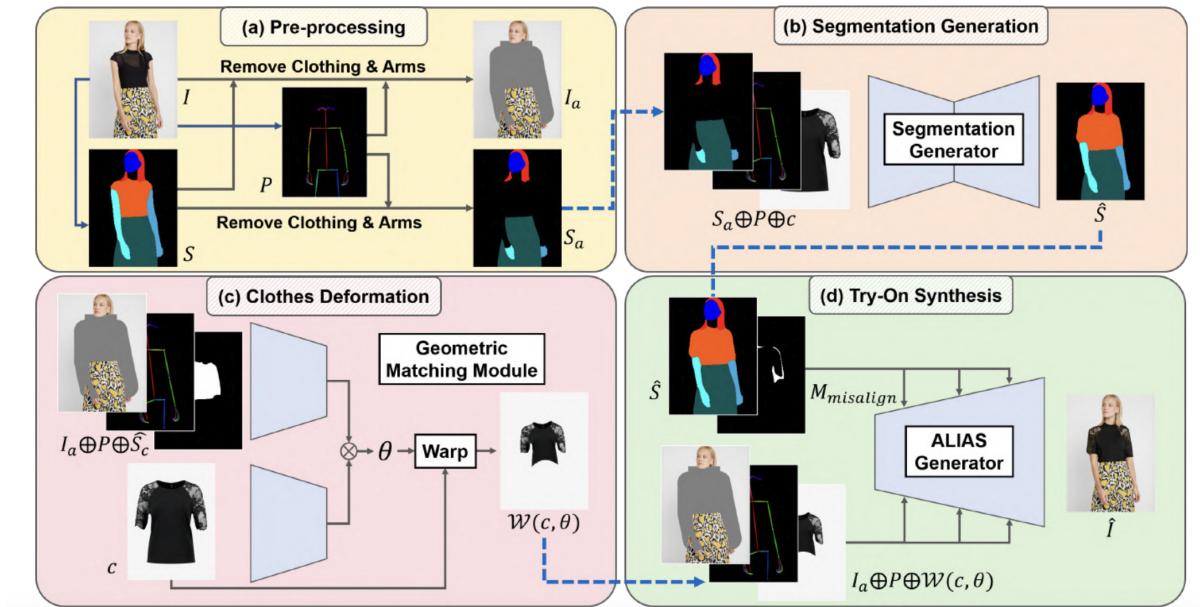


## 6\_Human Agnostic

- 사람의 신체 정보를 기반으로 특정 신체 부위나 의류를 제거한 뒤, 사람이 있는 이미지에서 신체와 관련된 중립적인 표현을 생성
- “Human Agnostic” ⇒ 특정 신체 부위나 의류에 구속되지 않는 상태



TRYON



Where :

$I$  : Input (모델 이미지)

$c$  : Clothing Image (의류 이미지)

**1\_OpenPose**  $\Rightarrow P$  : PoseMap (OpenPose 좌표)

**2\_Human Parse**  $\Rightarrow S$  : Human Parse (신체 클래스별로 Segmenatation)

**4\_Cloth Mask**  $\Rightarrow \hat{S}_c$  : Predicted Clothing Segmentation (의류 이미지 Segmentation)

**6\_Human Agnostic**  $\Rightarrow I_a$  : Human Agnostic (S와 P를 활용해서 특정 신체 부위나 의류를 제거된 이미지)

**5\_Parse Agnostic**  $\Rightarrow S_a$  : Parse Agnostic (S와 P를 활용해서 옷을 제거한 형태의 신체 마스크 이미지)

$\hat{S}$  : Synthetic Segmentation (인위적으로 합성한 이미지,  $S_a, P, c$  로 합성)

$W$  : Warped Clothing Image (TPS 변환으로 변환된 의류 이미지)

## Clothing-Agnostic Person Representation

- $S(\text{human\_parse})$ : 교체할 의류 영역을 제거하고 이미지의 나머지 부분을 보존하는데 사용
- $P(\text{openpose\_json})$ : 손이 아닌 팔은 재생산이 어렵기 때문에, 팔을 제거하는 데 활용

- 팔은 얼굴 피부에 맞게 생성해야 됨.
- $S$ (human parse) +  $P$ (openpose\_json)
- ⇒ 원래 의류 정보를 완전히 제거하고 나머지 이미지를 보존( $I_a, S_a$ )

## Segmentation Generation(GS/모델: U-Net)

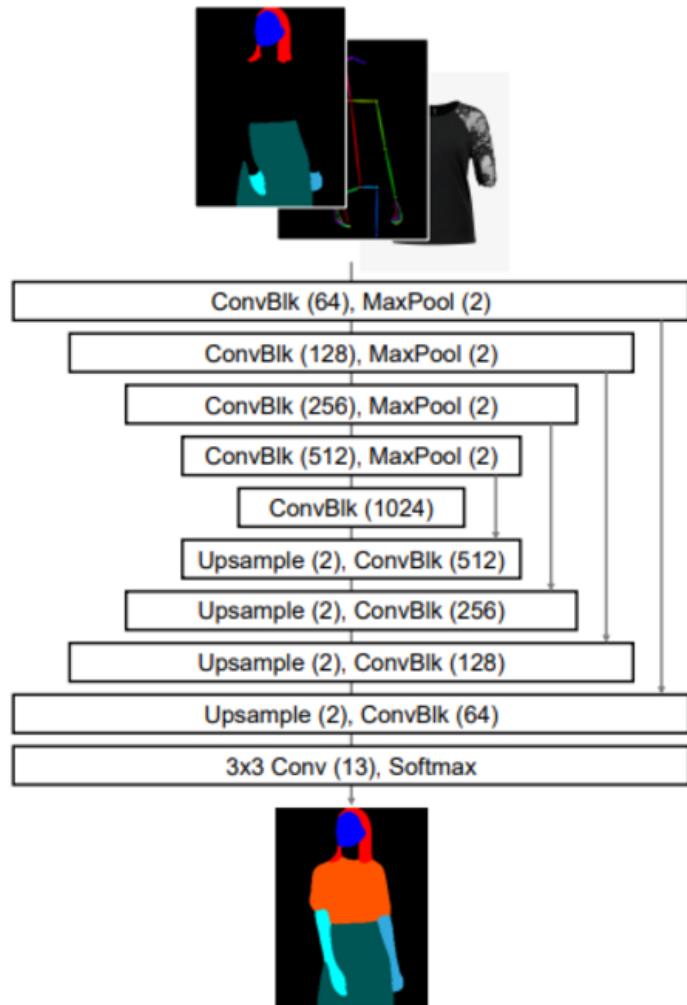


Figure 10: Segmentation Generator.  $k \times k$  Conv ( $x$ ) denotes a convolutional layer where the kernel size is  $k$  and the output channel is  $x$ . Also, ConvBlk ( $x$ ) denotes a block, which consists of two series of  $3 \times 3$  convolutional layer, instance normalization, and ReLU activation.

- $S_a + P + \text{target 의류 아이템 } c$   
⇒ (target 의류 아이템) $c$ 를 착용한 기준 이미지에서 사람의 **segmentation map**  
⇒  $\hat{S} = G_s(S_a, P, c)$

- **Loss:** 원래 의류 아이템 정보가 완전히 제거된  $S$ 와  $(S_a, P, c)$  사이의 매핑을 배울 수 있도록  $G_s$ 를 학습

( $\hat{S}$  과  $S$  사이의 **pixel-wise cross-entropy loss + conditional adversarial loss + hyper-parameter**)

$$\mathcal{L}_S = \mathcal{L}_{cGAN} + \lambda_{CE} \mathcal{L}_{CE} \quad (8)$$

$$\mathcal{L}_{CE} = -\frac{1}{HW} \sum_{k \in C, y \in H, x \in W} S_{k,y,x} \log(\hat{S}_{k,y,x}) \quad (9)$$

$$\mathcal{L}_{cGAN} = \mathbb{E}_{(X,S)} [\log(D(X, S))] + \mathbb{E}_X [1 - \log(D(X, \hat{S}))], \quad (10)$$

## Clothing Image Deformation

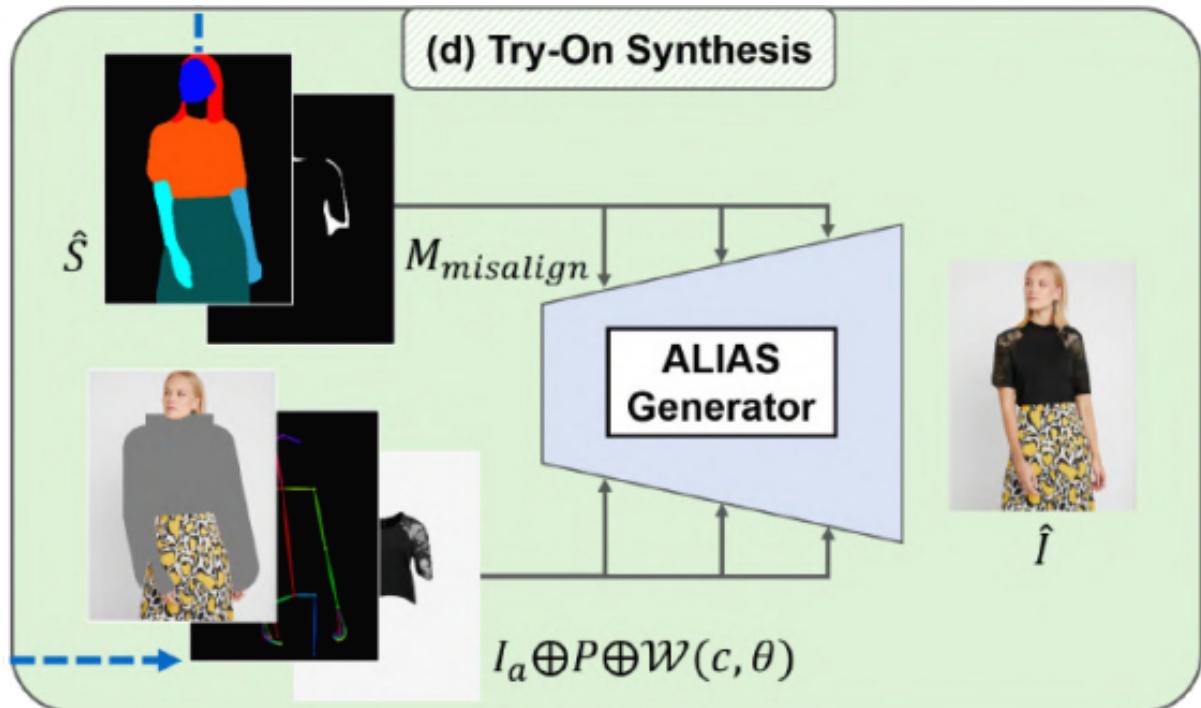


- target item  $c$ 를  $\hat{S}$ 의 의류 영역인  $\hat{S}_c$ 와 일치하도록 변형
- (3.1. Clothing-Agnostic Person Representation의 output)  $I_a + P + \hat{S}$  ( $\hat{S}_c$  대신  $S$ ) +  $c$   
⇒ warp된 의류이미지  $W(c, \theta)$
- **Loss**

$$\mathcal{L}_{warp} = ||I_c - W(c, \theta)||_1, 1 + \lambda_{const} \mathcal{L}_{const} \quad (11)$$

$$\begin{aligned} \mathcal{L}_{const} &= \sum_{p \in P} |(||pp_0||_2 - ||pp_1||_2| + ||pp_2||_2 - ||pp_3||_2|) \\ &\quad + (|S(p, p_0) - S(p, p_1)| + |S(p, p_2) - S(p, p_3)|), \end{aligned}$$

## Try-On Synthesis via ALIAS normalization



- 이전 단계의 output을 기반으로  $I$  (모델 이미지)에서 주어진 사람이 해당 의류 아이템을 입은 합성 이미지  $\hat{I}$ 를 생성하는 것이 목표
- $(I_a, P, W(c, \theta))$ 는 generator의 각 layer에 input
- $\hat{S}$ 의 경우, **ALIgntment-Aware Segment (ALIAS) normalization** 사용
  - $\hat{S}$ 와 해당 영역의 mask를 활용하여 semantic 정보를 보존하고 misaligned 영역에서의 misleading information를 제거

### ▼ Alignment-Aware Segment Normalization

$h_i \in R^{(N \times C_i \times H_i \times W_i)}$ 는  $i$ 번째 layer의 activation이다 ( $N$ : batch,  $H$ : height,  $W$ : width,  $C$ : channel)

- $\hat{S}$ : 합성 segmentation map

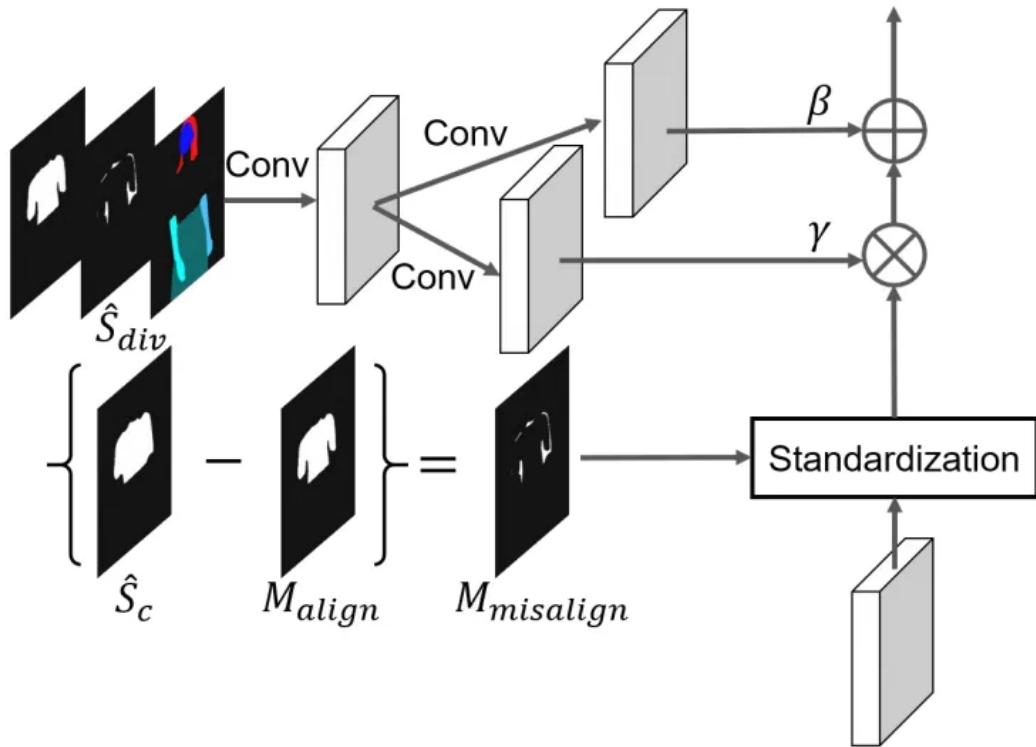
상의를 제거한 형태의 신체 마스크( $S_a$ )에 openpose 좌표( $P$ )로 맞게 의류 이미지( $C$ )를 합성

- $M_{align}$  :  $\hat{S}$ 에서 대상 의류 이미지의 warped mask  $W(M_c, \theta)$   
 $\hat{S}$ 에서 warp된 의류 이미지 부분만을 의미
- $M_{misalign}$  : the misalignment binary mask  
 $\hat{S}$ 에서 warp된 의류 이미지 부분을 뺀 나머지 부분을 의미

$$M_{misalign} \in L^{(H \times W)} (\hat{S}_c \text{에서 } W(M_c, \theta) \text{는 제외})$$

$$M_{align} = \hat{S}_c \cap \mathcal{W}(M_c, \theta) \quad (3)$$

$$M_{misalign} = \hat{S}_c - M_{align}. \quad (4)$$



- Alignment-Aware Segment Normalization의 input
    - $\hat{S}_{div} : (M_{align} + M_{misalign} + S_a)$
    - $M_{misalign}$  영역
    - $h_i$  (Nomarlization 하기 이전의 값)
      - N개의 샘플 배치에 대한 네트워크의 i번째 layer activation
  - 네트워크는 학습을 통해 affine 변환 파라미터( $\beta, \gamma$ )를 최적화된 값으로 결정
- ⇒ 특정 영역에 맞게 정규화된 활성화 값을 조정하는 과정
- $$\beta, \gamma = f(\hat{S}_{div}) \text{ (f} \Rightarrow \text{Conv 연산)}$$
- $$h'_i = \gamma * \hat{h}_i + \beta$$

$\gamma$  : standardization된 activation 값의 스케일 조정(강조할 특성 키우거나 줄임)

$\beta$  : standardization된 activation 값의 시프트 추가(특정 특성의 중요도 보정)

⇒ Alias Normalization의 과정을 통해 공간적 위치와 의미적 영역에 따라 파라미터 조정해서 semantic 정보를 보존하고 misaligned 영역에서의 misleading information를 제거한다.

### ▼ ALIASED Segment (ALIAS) Generator

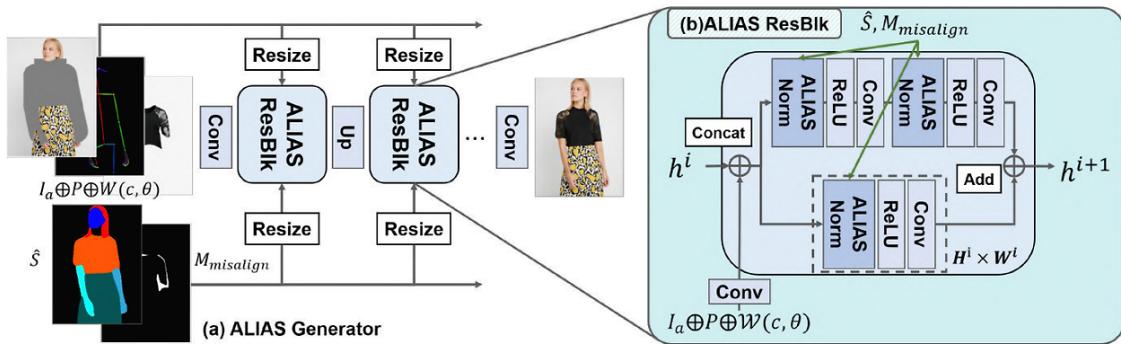


Figure 4: ALIAS generator. (a) The ALIAS generator is composed of a series of ALIAS residual blocks, along with up-sampling layers. The input ( $I_a, P, \mathcal{W}(c, \theta)$ ) is resized and injected into each layer of the generator. (b) A detailed view of a ALIAS residual block. Resized ( $I_a, P, \mathcal{W}(c, \theta)$ ) is concatenated to  $h^i$  after passing through a convolution layer. Each ALIAS normalization layer leverages resized  $\hat{S}$  and  $M_{misalign}$  to normalize the activation.

- 각 ALIAS ResBlk은 3개의 conv layer와 3개의 ALIAS Normalization conv layer로 구성
- encoder-decoder network의 encoder 부분을 제거
- Generator는 upsampling layer가 있는 residual blocks (ResBlk)을 사용
- ResBlks가 작동하는 해상도가 다르기 때문에 각 계층에 주입하기 전에 표준화 계층인  $\hat{S}$  및 Mimsalign의 입력 크기를 조정
 

⇒ 마찬가지로, generator의 입력 ( $I_a, P, \mathcal{W}(c, \theta)$ )은 다른 해상도으로 크기가 조정
- 각 ResBlk 전에 크기 조정된 입력( $I_a, P, \mathcal{W}(c, \theta)$ )은 컨볼루션 레이어를 통과한 후 이전 레이어의 활성화에 연결되며, 각 ResBlk는 연결된 입력을 활용하여 활성화를 세분화

픽셀 수준에서 단일 refinement보다 의류 디테일을 더 잘 보존하는 특징 수준에서 multi-scale refinement를 수행한다. → 기존의 모델의 문제점 해결

$L_{cGAN} \rightarrow$  conditional adversarial loss,  $L_{FM} \rightarrow$  feature matching loss,  
 $L_{percept} \rightarrow$  perceptual loss

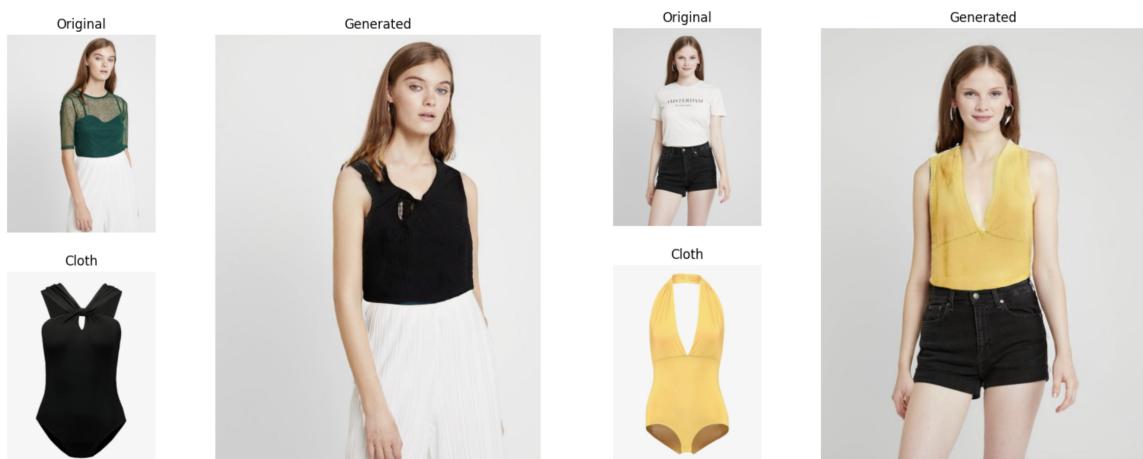
$$\mathcal{L}_I = \mathcal{L}_{cGAN} + \lambda_{FM} \mathcal{L}_{FM} + \lambda_{percept} \mathcal{L}_{percept} \quad (13)$$

$$\mathcal{L}_{cGAN} = \mathbb{E}_I[\log(D_I(S_{div}, I))] + \mathbb{E}_{(I,c)}[1 - \log(D_I(S_{div}, \hat{I}))] \quad (14)$$

$$\mathcal{L}_{FM} = \mathbb{E}_{(I,c)} \sum_{i=1}^T \frac{1}{K_i} [\|D_I^{(i)}(S_{div}, I) - D_I^{(i)}(S_{div}, \hat{I})\|_{1,1}] \quad (15)$$

$$\mathcal{L}_{percept} = \mathbb{E}_{(I,c)} \sum_{i=1}^V \frac{1}{R_i} [\|F^{(i)}(I) - F^{(i)}(\hat{I})\|_{1,1}], \quad (16)$$

## TRYON Result



## 기존 GAN 기반 Try-On 모델의 한계점

### 1. 정면 포즈 한정

- 기존 모델은 **정면 포즈의 사람**을 대상으로만 합성이 가능하며, 다양한 각도의 시점에서 자연스러운 합성이 어렵다.
- **멀티뷰(Multi-View)** 시점의 사람(ex. 옆 모습, 뒷 모습)에 대해서는 합성 품질이 급격히 저하된다.

### 2. Human Parsing의 부정확성

- **Human Parsing**이 부정확할 경우, 의상 정보가 제대로 추출되지 않아 아래 그림과 같이 왜곡된 결과를 초래한다.
- 옷의 경계선이나 포즈에 맞는 의상 변형이 실패하여 비현실적인 결과를 생성한다.
- 특히, 소매나 디테일이 많은 의상의 경우 오류가 심화된다.

### 3. 고화질 합성의 한계

- HR(High-Resolution) Try-On 모델에서도 고화질의 세부 디테일을 재현하는 데 실패하는 경우가 빈번하다.
- 텍스처 및 디테일 정보가 왜곡되거나, 경계선 부분에서 artifact(왜곡 현상)가 발생 한다.

### 4. 포즈 기반 옷 변형의 실패

- 옷의 워핑 과정에서 포즈에 따라 옷을 변형하는 작업이 불완전하여, 특정 포즈에서는 옷의 형태가 비현실적이거나 어색해질 수 있다.
- 이는 특히 **GAN 기반의 제한된 표현 능력**으로 인해 발생한다.



## MV-VTON 프로젝트로의 전환 배경

MV-VTON은 기존 GAN 기반 Try-On 모델의 한계를 극복하고, 보다 **다양한 시점에서의 자연스러운 옷 합성**을 목표로 한다. 이를 위해 다음과 같은 혁신적인 기술을 도입했다.

### 1. Diffusion 모델 기반

- **Diffusion 모델**을 백본으로 채택하여, 기존 GAN의 한계를 극복.
- 반복적인 노이즈 제거를 통해 **더욱 정교하고 자연스러운 이미지 합성**을 가능하게 한다.

### 2. View-Adaptive Selection

- 사람의 포즈와 옷의 포즈를 기반으로 전면 또는 후면 뷰 의상을 선택하는 하드 셀렉션.
- 포즈 유사도를 분석하여 세부적인 텍스처 정보를 보완하는 소프트 셀렉션.

### 3. Global + Local Feature Fusion

- **Global Features**를 통해 전반적인 의미론적 정보를 유지.
- **Local Features**로 세부 디테일을 보완.
- Joint Attention Blocks를 사용해 두 특징을 융합, artifact를 최소화하고 **현실적인 Try-On 결과**를 보장.

### 4. 멀티뷰 데이터셋 활용

- 다양한 각도(전면, 45도, 측면, 135도, 후면)의 이미지를 포함한 **MVG(Multi-View Garment)** 데이터셋을 수집하여, 보다 다양한 시점에서의 합성 결과를 학습.

## MV-VTON의 기대 효과

MV-VTON은 기존 GAN 기반 Try-On 모델과 HR Try-On 모델의 한계를 극복하고, 다음과 같은 성과를 기대한다:

- 다양한 시점에서의 자연스러운 합성
- 고화질의 세부 디테일 표현
- 포즈와 의상 간의 정교한 매칭
- 사용자 맞춤형 Try-On 경험 제공

MV-VTON은 최신 AI 기술을 활용해 가상 의류 합성의 새로운 패러다임을 제시하며, 실제 의류 산업 및 온라인 쇼핑 분야에서의 활용 가능성을 대폭 확대할 것으로 기대됩니다.