

Article

UVIRT—Unsupervised Virtual Try-on Using Disentangled Clothing and Person Features

Hideki Tsunashima ^{1,*}, Kosuke Arase ², Antony Lam ² and Hirokatsu Kataoka ¹¹ Computer Vision Research Team, Artificial Intelligence Research Center, National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba 305-8560, Japan; hirokatsu.kataoka@aist.go.jp² Mercari, Inc., Tokyo 106-6188, Japan; kosuke.arase@mercari.com (K.A.); antonylam@mercari.com (A.L.)

* Correspondence: h.tsunashima@asagi.waseda.jp; Tel.: +81-29-861-2267

Received: 5 August 2020; Accepted: 27 September 2020; Published: 2 October 2020



Abstract: Virtual Try-on is the ability to realistically superimpose clothing onto a target person. Due to its importance to the multi-billion dollar e-commerce industry, the problem has received significant attention in recent years. To date, most virtual try-on methods have been supervised approaches, namely using annotated data, such as clothes parsing semantic segmentation masks and paired images. These approaches incur a very high cost in annotation. Even existing weakly-supervised virtual try-on methods still use annotated data or pre-trained networks as auxiliary information and the costs of the annotation are still significantly high. Plus, the strategy using pre-trained networks is not appropriate in the practical scenarios due to latency. In this paper we propose Unsupervised VIRTual Try-on using disentangled representation (UVIRT). After UVIRT extracts a clothes and a person feature from a person image and a clothes image respectively, it exchanges a clothes and a person feature. Finally, UVIRT achieve virtual try-on. This is all achieved in an unsupervised manner so UVIRT has the advantage that it does not require any annotated data, pre-trained networks nor even category labels. In the experiments, we qualitatively and quantitatively compare between supervised methods and our UVIRT method on the MPV dataset (which has paired images) and on a Consumer-to-Consumer (C2C) marketplace dataset (which has unpaired images). As a result, UVIRT outperform the supervised method on the C2C marketplace dataset, and achieve comparable results on the MPV dataset, which has paired images in comparison with the conventional supervised method.

Keywords: virtual try-on; image-to-image translation; unsupervised learning; GAN; disentanglement

1. Introduction

Consumer-to-consumer (C2C) online apparel markets are a multi-billion dollar industry [1]. In C2C services, users can both buy and sell products to each other. Some of the key advantages of buying apparel through C2C services include availability of easy to find items, low price, and the large volume of merchandise. On the other hand, sellers do not need physical stores, it is easier to manage inventory, and oftentimes, they can convert unneeded goods into cash. Despite all these advantages, buyers have the problem that they cannot try on any of the apparels. Although some e-commerce (EC) websites provide services where consumers can try on apparels by delivery [2], such services are time-consuming for both the buyer and seller. Hence, virtual try-on approaches have received attention recently. Effective virtual try-on systems would enable consumers to make decisions on buying apparels more easily, improving sales.

State-of-the-art virtual try-on methods typically employ Deep Neural Networks (DNN) and 2D images rather than 3D models [3–13]. These virtual try-on methods also typically employ supervised learning. Supervised virtual try-on methods require ground truth data such as clothing parsing

semantic segmentation masks and paired images. However, the annotation of clothing parsing semantic segmentation masks and paired images is labor intensive. Moreover, in C2C applications, supervised virtual try-on methods cannot be applied because of the lack of existing ground truth data. In C2C markets, annotating clothing parsing semantic segmentation masks and paired images is impractical since a very high volume of item listings can be added daily [14]. Therefore, virtual try-on methods which do not require annotated data are desired (e.g., unsupervised approaches). Recently, weakly-supervised virtual try-on methods have been proposed [11,12]. Although these weakly-supervised methods do not use ground truth data, they still require some annotated data or pre-trained networks which generate human dense images or parsing semantic segmentation maps. The human dense images or parsing semantic segmentation masks are used as auxiliary input information. Specifically, [12] requires clothing-parsing semantic segmentation masks to achieve virtual try-on. As for Reference [11], the method does not support clothing-only image to person translation, which is a needed functionality in virtual try-on. This is because human-parsing semantic segmentation networks require images that contain at least one person.

In this paper, we propose Unsupervised VIRTUAL Try-on using disentangled representation (UVIRT). UVIRT does not need any supervision, namely it does not need any annotations and does not require pre-trained networks. After the clothing features and person features are disentangled from each domain image (i.e., an only clothing image, and a person image), we can achieve unsupervised virtual try-on by manipulating the disentangled features. In Figure 1, firstly we decompose each domain image into style and content features. Next, after we switch the content features, we decode each domain image (person images vs. clothing only images) using the switched content features and each image's original style features. In our setup, the content features contain the clothing features and the style features contain the person and background features, respectively. UVIRT can warp clothing onto a person by disentangling the clothing and the person features without annotated data or additional pre-trained networks, which would need to generate a human dense image or a human-parsing semantic segmentation map. Thus, UVIRT overcomes the problem of needing ground truth data and being difficult to apply to practical scenarios. Moreover, UVIRT significantly outperforms conventional methods in testing time because UVIRT does not use additional networks (e.g., human pose detector and human parser).

A key concept employed in our solution is *disentanglement*. Disentanglement is the idea that the inputs (e.g., images) can be broken down into factors understood intuitively by humans [15]. In particular, disentanglement has been broadly researched in unsupervised learning [16–22]. Inspired by this, our UVIRT approach achieves virtual try-on without ground truth data or paired images by statistically decomposing the person and the clothing features in the latent space and by being able to reconstruct each domain image from the disentangled features.

In the experiments section, we first qualitatively and quantitatively compare UVIRT with the conventional methods by evaluating on the MPV dataset [10]. We then qualitatively and quantitatively confirm that UVIRT can be applied to an in-the-wild C2C market image dataset collected by us.

We summarize the main contributions as follows:

- We are the first to propose a fully unsupervised virtual try-on method. Conventional virtual try-on methods use inputs such as paired images, clothing parsing semantic segmentation maps, and human dense image maps, as ground truth data or some kind of additional information. On the other hand, we do not need any additional data by employing the concept of disentanglement.
- We are the first to apply real-world C2C market data to virtual try-on. This is demonstrated through experiments on a real-world C2C market dataset. This is particularly challenging because the C2C market dataset has images that come from a wide variety of RGB camera types. We show that it is still possible to fuse disentangled components (person vs. clothing) features despite these challenges.
- Despite being completely unsupervised, our approach achieves competitive results with existing supervised approaches.

- Our approach is 13 times faster than the supervised method (i.e., Characteristic-Preserving Virtual Try-On Network (CP-VTON) [4]) because ours does not use additional networks, such as the human pose detector and human parser.

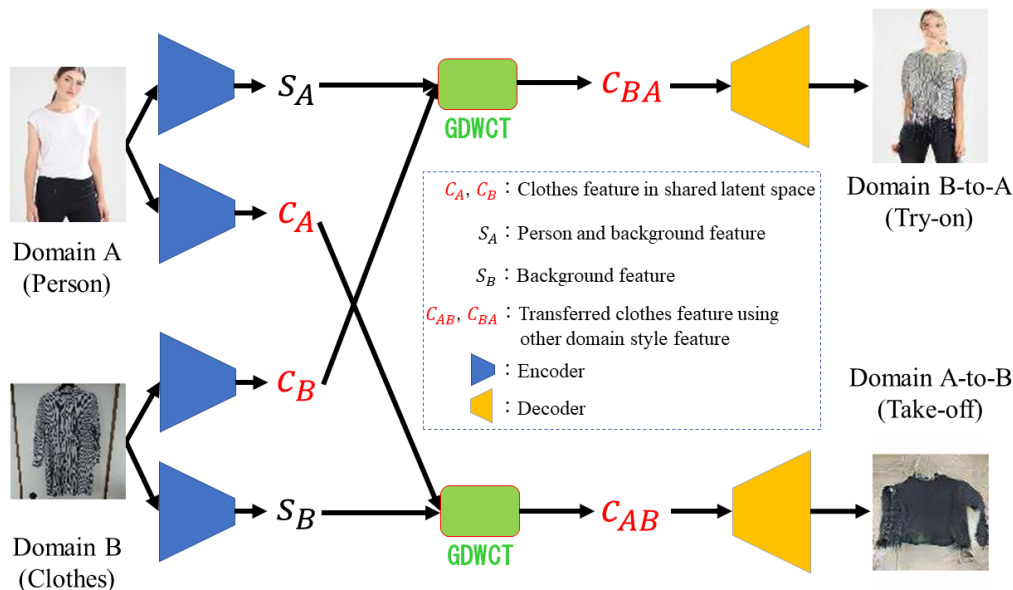


Figure 1. Overview of our system. Firstly, the person image and the clothing image are encoded into the content feature c and the style features s . Note that each domain's content features c_A, c_B captures the clothing features, the domain A style feature s_A captures the person and background feature and the domain B style feature s_B captures the background feature. Next, the exchanged content features c_B, c_A are combined with the style features of each domain s_A, s_B by the Group-wise Deep Whitening-and-Coloring Transformation (GDWCT). Finally, the transformed content features c_{BA}, c_{AB} are decoded into their new domain images. Note that this figure is subset of the whole architecture. The two discriminator, $x_{B,AB}$, and $x_{A,BA}$ streams are omitted.

The rest of the paper is organized as follows—Section 2 briefly reviews related work on virtual try-on and image-to-image translation. The proposed method is introduced in Section 3 and experimental results are presented in Section 4. Finally, conclusions are offered in Section 5.

2. Related Work

2.1. Supervised Virtual Try-on

Supervised virtual try-on methods can be mainly categorized into two general approaches, namely, GAN-based and AutoEncoders+geometrical transformation-based (hereafter, AE-based) methods.

GAN-based methods directly generate warped clothing onto people. However, GAN-based methods require paired images [13], which consist of the person wearing some target clothing and only the clothing by itself, or categorized apparel images [6]. The annotation cost of such datasets is high.

AE-based methods [3–5,7–12] achieve virtual try-on through a two-stage strategy. In the first stage, the image of the target clothing is aligned with the pose of the person. In the second stage, the aligned clothing image is warped onto the person. While there have been promising results, AE-based methods need paired images and clothing parsing semantic segmentation masks. Similar to GAN-based methods, their applicability is limited in practical scenarios because of the high annotation cost. For instance, the annotation cost of semantic segmentation is generally very high in comparison with other annotations, such as class label annotation. At the same time, applying supervised virtual try-on methods to C2C market data is impractical because it is difficult for end-users to prepare annotated data with clothing parsing semantic segmentation masks. In addition, users can list any number of items on C2C markets. This means that the inventory is continually changing and

has an immense variety of products. This makes application of a static training dataset for virtual try-on ineffective.

2.2. Image-to-Image Translation.

Image-to-image translation (I2I) has been broadly researched. I2I methods are mainly categorized into supervised and unsupervised methods. Supervised methods map images based on one-to-one (e.g., pix2pix [23]), or one-to-many strategies (e.g., bicycleGAN [24]) by using paired data, such as cityscapes [25]. On the other hand, unsupervised methods map based on not only one-to-one (e.g., UNIT [26]) and one-to-many (e.g., MUNIT [27]) but also many-to-many (e.g., StarGAN [28]) strategies by using unpaired data, such as what was done in horse2zebra [29]. Although conventional unsupervised methods are prevalent due to low annotation cost, they are difficult to translate with respect to desired image regions. In our experiments, although we first employed MUNIT, it could not warp clothing to people well. Therefore, we utilize the state-of-the-art I2I assistant method “Group-wise Deep Whitening-and-Coloring Translation (GDWCT)” [30]. GDWCT can warp a clothing feature to a person as it can transfer domain-specific features well.

2.3. Weakly-Supervised Virtual Try-on

To the best of our knowledge, recently, two weakly-supervised virtual try-on methods have been proposed [11,12]. We refer to them as “weakly-supervised virtual try-on methods” because these two methods still use additional data generated by pre-trained networks. The first method was proposed by Wang et al. [12]. Wang et al. achieved weakly-supervised virtual try-on by using clothing parsing semantic segmentation masks as attention for the model. While promising, annotated data was still required. The second method was proposed by Pumarola et al. [11]. They achieved weakly-supervised virtual try-on by using generated person-parsing dense images which represents 3D information and generated person parsing semantic segmentation masks. Concretely, these dense images are constructed in a UV coordinate system. In Computer Graphics (CG), the 2D orthogonal coordinate system is often used to map textures onto 3DCG models. In the 2D orthogonal coordinate system, the horizontal direction is called “U” and vertical direction is called “V”. DensePose [31] directly infers the dense image from the UV coordinate system. Concretely, the person parsing UV map is generated by a pre-trained PSPNet [32] and the person parsing semantic segmentation masks are generated by the pre-trained DensePose model [31]. However, the weakly-supervised virtual try-on method proposed by Pumarola et al. cannot be applied to model the task of applying clothing only image features to person images (dubbed clothing2person task). They can only apply their approach to transfer clothing between images that all contain people because the above two pre-trained networks only work on images containing people. We propose an unsupervised virtual try-on method (UVIRT) which achieves clothing2person generation and does not use any additional annotated data or pre-trained models. UVIRT can be broadly applied to not only aligned images of fashion e-commerce (EC) sites in general but as an added benefit, also to in-the-wild unaligned images (e.g., apparel images from C2C markets).

3. Methods

This section describes our proposed framework by explaining our virtual try-on strategy. We propose an unsupervised virtual try-on method (UVIRT) that does not require annotated data. Our goal is that we warp clothing onto a target person while retaining appearance features (e.g., long sleeves, characters). An overview of our model is illustrated in Figure 1. Moreover, the concrete algorithm is described in Algorithm 1.

Algorithm 1 Unsupervised Virtual Try-on (UVIRT)

Require: The weight of latent L_1 loss λ_{latent} , the weight of pixel L_1 loss λ_{pixel} , the weight of whitening regularization loss λ_w , the weight of coloring regularization loss λ_c , the batchsize m , Adam hyperparameters α, β_1, β_2

Require: Initial style and content encoder parameters in domain \mathcal{A} and \mathcal{B} $\phi_{s_A}, \phi_{s_B}, \phi_{c_A}, \phi_{c_B}$, initial decoder parameters in domain \mathcal{A} and \mathcal{B} θ_A, θ_B , initial discriminator parameters in domain \mathcal{A} and \mathcal{B} ψ_A, ψ_B

while ϕ, θ has not converged **do**

for $i = 1, \dots, m$ **do**

 Sample each domain real data $x_A \sim p_{data_A}(x), x_B \sim p_{data_B}(x)$

$s_A, s_B = \text{StyleEncoder}_{\phi_{s_A}}(x_A), \text{StyleEncoder}_{\phi_{s_B}}(x_B)$

$c_A, c_B = \text{ContentEncoder}_{\phi_{c_A}}(x_A), \text{ContentEncoder}_{\phi_{c_B}}(x_B)$

$c_{BA}, c_{AB} = \text{GDWCT}_{\theta_A}(c_B, s_A), \text{GDWCT}_{\theta_B}(c_A, s_B)$

$x_{BA}, x_{AB} = \text{Decoder}_{\theta_A}(c_{BA}), \text{Decoder}_{\theta_B}(c_{AB})$

$\mathcal{L}_{D_{adv}}^A, \mathcal{L}_{G_{adv}}^A \leftarrow \frac{1}{2} (D_{\psi_A}(x_A) - 1)^2 + \frac{1}{2} (D_{\psi_A}(x_{BA}))^2, \quad \frac{1}{2} (D_{\psi_A}(x_{BA}) - 1)^2$

$\mathcal{L}_{D_{adv}}^B, \mathcal{L}_{G_{adv}}^B \leftarrow \frac{1}{2} (D_{\psi_B}(x_B) - 1)^2 + \frac{1}{2} (D_{\psi_B}(x_{AB}))^2, \quad \frac{1}{2} (D_{\psi_B}(x_{AB}) - 1)^2$

$\psi \leftarrow \text{Adam} \left(\nabla_{\psi_A, \psi_B} \frac{1}{m} \sum_{i=1}^m \mathcal{L}_{D_{adv}}^A + \mathcal{L}_{D_{adv}}^B, \psi_A + \psi_B, \alpha, \beta_1, \beta_2 \right)$

$s_{BA}, s_{AB} = \text{StyleEncoder}_{\phi_{s_A}}(x_{BA}), \text{StyleEncoder}_{\phi_{s_B}}(x_{AB})$

$c_{BA}, c_{AB} = \text{ContentEncoder}_{\phi_{c_A}}(x_{BA}), \text{ContentEncoder}_{\phi_{c_B}}(x_{AB})$

$c_{ABA}, c_{BAB} = \text{GDWCT}_{\theta_A}(c_{AB}, s_{BA}), \text{GDWCT}_{\theta_B}(c_{BA}, s_{AB})$

$x_{ABA}, x_{BAB} = \text{Decoder}_{\theta_A}(c_{ABA}), \text{Decoder}_{\theta_B}(c_{BAB})$

$c_{AA}, c_{BB} = \text{GDWCT}_{\theta_A}(c_A, s_A), \text{GDWCT}_{\theta_B}(c_B, s_B)$

$x_{AA}, x_{BB} = \text{Decoder}_{\theta_A}(c_{AA}), \text{Decoder}_{\theta_B}(c_{BB})$

$\mathcal{L}_s^{BA}, \mathcal{L}_s^{AB} \leftarrow \|s_{BA} - s_A\|_1, \|s_{AB} - s_B\|_1$

$\mathcal{L}_c^{BA}, \mathcal{L}_c^{AB} \leftarrow \|c_{BA} - c_B\|_1, \|c_{AB} - c_A\|_1$

$\mathcal{L}_{cyc}^{ABA}, \mathcal{L}_{cyc}^{BAB} \leftarrow \|x_{ABA} - x_A\|_1, \|x_{BAB} - x_B\|_1$

$\mathcal{L}_i^{AA}, \mathcal{L}_i^{BB} \leftarrow \|x_{AA} - x_A\|_1, \|x_{BB} - x_B\|_1$

$\Sigma_c^B, \Sigma_c^A = \text{ComputeCovariance}(c_B), \text{ComputeCovariance}(c_A)$

$\mathcal{R}_w^B, \mathcal{R}_w^A \leftarrow \|\Sigma_c^B - I\|_1, \|\Sigma_c^A - I\|_1$

$UD^A, UD^B = s_A^{CT}, s_B^{CT}$

$D^A, D^B = \text{ComputeL2Norm}(s_A^{CT}), \text{ComputeL2Norm}(s_B^{CT})$

$U^A, U^B = \frac{UD^A}{D^A}, \frac{UD^B}{D^B}$

$\mathcal{R}_c^A, \mathcal{R}_c^B \leftarrow \|U_A^T U_A - I\|_1, \|U_B^T U_B - I\|_1$

$\phi, \theta \leftarrow \text{Adam} \left(\nabla_{\phi_{s_A}, \phi_{s_B}, \phi_{c_A}, \phi_{c_B}, \theta_A, \theta_B} \frac{1}{m} \sum_{i=1}^m \lambda_{latent} (\mathcal{L}_s^{BA} + \mathcal{L}_s^{AB} + \mathcal{L}_c^{BA} + \mathcal{L}_c^{AB}) + \right.$
 $\lambda_{pixel} (\mathcal{L}_{cyc}^{ABA} + \mathcal{L}_{cyc}^{BAB} + \mathcal{L}_i^{AA} + \mathcal{L}_i^{BB}) + \lambda_w (\mathcal{R}_w^B + \mathcal{R}_w^A) + \lambda_c (\mathcal{R}_c^B + \mathcal{R}_c^A) + \mathcal{L}_{D_{adv}}^B + \mathcal{L}_{G_{adv}}^B,$
 $\left. \phi_{s_A} + \phi_{s_B} + \phi_{c_A} + \phi_{c_B} + \theta_A + \theta_B, \alpha, \beta_1, \beta_2 \right)$

end for

end while

3.1. Model

Inspired by MUNIT [27], UVIRT first extracts the content (clothing) features c_A, c_B in a shared latent space C and the style features (person and background) s_A, s_B from the clothed person image x_A and the clothing only x_B , respectively. Next, although we want to reflect the style features to the content features, inspired by MUNIT, the content and style features highly are entangled when using Adaptive Instance Normalization (AdaIN) [33]. Hence, we apply the style features s_A, s_B to the exchanged content features c_B, c_A by using Group-wise Deep Whitening-and-Coloring Transformation (GDWCT) [30] in order to not entangle the content and style features. Briefly, after the content feature gets whitened, using the mean of the content features (Group-wise Deep Whitening Translation: GDWT), the style feature is applied to the whitened content feature (Group-wise Deep Coloring Translation: GDCT). Finally, UVIRT can perform virtual try-on by reconstructing the person image (i.e., $x_{B,A}$) from the switched content features transferred by GDWCT $c_{B,A}$ between the clothed person image x_A and the clothing only image x_B .

3.2. Loss Functions

Our model loss function is the same as in the GDWCT paper [30]. We describe the ten loss functions and two regularization terms in the following text. Note that, for the sake of simplicity, we describe things for only one domain, that is, the data flow of the transformation from domain B to domain A.

3.2.1. Adversarial Loss.

We introduce two adversarial losses. All adversarial losses employ the LSGAN loss [34]. We compute the LSGAN loss between an original image x_A and a virtual try-on image $x_{B,A}$. These adversarial losses encourage generating high fidelity images. The adversarial loss terms are as follows,

$$\mathcal{L}_{D_{adv}}^A = \frac{1}{2} \mathbb{E}_{x_A} \left[(D(x_A) - 1)^2 \right] + \frac{1}{2} \mathbb{E}_{x_{B,A}} \left[(D(x_{B,A}))^2 \right] \quad (1)$$

$$\mathcal{L}_{G_{adv}}^A = \frac{1}{2} \mathbb{E}_{x_{B,A}} \left[(D(x_{B,A}) - 1)^2 \right], \quad (2)$$

where $\mathcal{L}_{D_{adv}}^A$ is the adversarial loss of the discriminator, $\mathcal{L}_{G_{adv}}^A$ is the adversarial loss of the generator, $D(\cdot)$ is the output of discriminator.

3.2.2. Reconstruction Loss.

We introduce eight reconstruction losses. All reconstruction losses are based on the L1 norm. First, we compute the L1 loss between the encoded style feature s_A and the re-encoded style feature $s_{B,A}$ from the transferred image $x_{B,A}$. It encourages the model to warp the disentangled clothing features of the reference image s_A to the translated image $x_{B,A}$. The first reconstruction loss term is as follows,

$$\mathcal{L}_s^{B,A} = \mathbb{E}_{x_A} [\|s_{B,A} - s_A\|_1]. \quad (3)$$

Second, we compute the L1 loss between the encoded content feature c_A and the re-encoded content feature $c_{B,A}$ from the transferred image $x_{B,A}$. It encourages the model to maintain the disentangled person and background features c_A after being translated $c_{B,A}$. In the case of domain B, to maintain the disentangled background features, the second reconstruction loss term is as follows,

$$\mathcal{L}_c^{B,A} = \mathbb{E}_{x_B} [\|c_{B,A} - c_A\|_1]. \quad (4)$$

Finally, we compute the L1 loss between the original image x_A and the cyclic reconstructed image $x_{AB,A}$ from the re-encoded style feature $s_{B,A}$ and content feature c_{AB} . Moreover, we compute the L1 loss between the original image x_A and the reconstructed image $x_{A,A}$ from the encoded style feature

s_A and content feature c_A , vice versa. This loss is the cycle consistency loss [29] which encourages image-level reconstruction. The final reconstruction loss terms are as follows,

$$\mathcal{L}_{cyc}^{ABA} = \mathbb{E}_{x_A} [\|x_{ABA} - x_A\|_1] \quad (5)$$

$$\mathcal{L}_i^{AA} = \mathbb{E}_{x_A} [\|x_{AA} - x_A\|_1]. \quad (6)$$

3.2.3. Regularization.

We introduce two regularization terms for GDWCT. In the GDWT phase, we encourage the covariance matrix Σ_c^B of the content feature c_B to be the identity matrix I so that the whitened content feature c_w is encouraged to $c_w = c_B - c_\mu$ where c_μ is mean of the content feature. This procedure not only removes the need for an eigendecomposition but improves the backpropagation time. The first regularization term is as follows,

$$\mathcal{R}_w = \mathbb{E} [\|\Sigma_c - I\|_1]. \quad (7)$$

Next, in the GDCT phase, we want to apply the covariance matrix Σ_s^B of the style feature s_B to the whitened content feature c_w . However, this procedure incurs expensive processing time and long backpropagation time. Therefore, the covariance matrix Σ_s^B of style feature s_B is decomposed to UDU^T . The colored style feature s^{CT} processed by multi layer perceptron is decomposed to UD , where the i -th column vector u_i of U is the unit vector and D is the diagonal matrix, that is, $s^{CT} = UD$. D corresponds to the L_2 norm of each column vector of s^{CT} . Next, UD is divided by D . Consequently, we can acquire U . U of the colored style feature has to be encouraged to be the orthogonal matrix. The second regularization term is as follows,

$$\mathcal{R}_c = \mathbb{E}_s [\|U^T U - I\|_1]. \quad (8)$$

GDWCT employs a ResNet like decoder with multiple channels. To improve processing speed, we group these channels into batches of 32.

3.2.4. Total Loss.

Our model including generator (i.e., encoders and decoders) and discriminator is optimized in an end-to-end manner. The full objectives are as follows,

$$\mathcal{L}_D = \mathcal{L}_{D_{adv}}^A + \mathcal{L}_{D_{adv}}^B \quad (9)$$

$$\mathcal{L}_G = \mathcal{L}_{G_{adv}}^A + \mathcal{L}_{G_{adv}}^B + \lambda_{latent} (\mathcal{L}_s + \mathcal{L}_c) + \lambda_{pixel} (\mathcal{L}_{cyc} + \mathcal{L}_i) + \lambda_w \mathcal{R}_w + \lambda_c \mathcal{R}_c, \quad (10)$$

where λ_{latent} is the coefficient of latent-level reconstruction losses and λ_{pixel} is the coefficient of image-level reconstruction losses. We set the value of $\lambda_{latent} = 1$, $\lambda_{pixel} = 10$, $\lambda_w = 0.001$, $\lambda_c = 10$, respectively.

3.3. Collecting C2C Market Datasets

We propose a new virtual try-on dataset built from a real-world C2C marketplace, that is, from Mercari [35] (hereafter, C2C market dataset). Specifically, we collected in-the-wild clothing images consisting of 539,546 ladies top clothing 96×96 resolution images, organized into 3389 categories. These images come from a very wide variety of RGB camera types because the marketplace itself has millions of users. The final resolution of the C2C market images is 128×96 due to adding white margins to the images. For testing virtual try-on, we also couple our collected clothing only images with 10,322 front-view person images from the Multi-Pose Virtual try-on datasets [10]. In Figure 2,

we can see that the C2C market dataset images are not aligned (e.g., hanged, folded). In Section 4.6, we investigate such characteristics of the dataset.

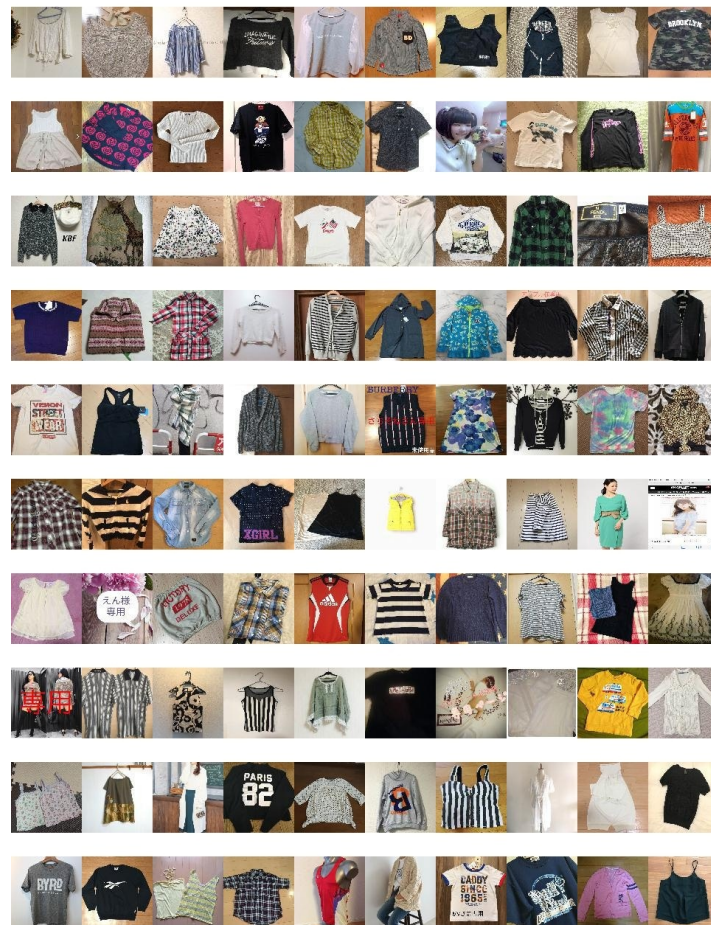


Figure 2. Sample images from the consumer-to-consumer (C2C) market dataset collected by us.

4. Experiments

4.1. Evaluation Criteria

Although conventional methods use the inception score [36] as their evaluation criterion, we do not use the inception score in our experiments. This is because the inception score measures the diversity and distinguishability of the generated images when using the Inception-v3 model. In virtual try-on evaluation, the inception score is not an appropriate evaluation criterion because it can only quantify how classifiable clothing are. Furthermore, our aim is not to randomly generate a set of images with a large class variation as is usually done in the Generative Adversarial Networks literature. Instead, we aim to generate virtual try-on images in a consistent manner. Given a person, we wish to apply a new article of clothing on the same person. Thus we choose more suitable metrics in our experiments.

4.1.1. Learned Perceptual Image Patch Similarity (LPIPS).

We use the Learned Perceptual Image Patch Similarity (LPIPS) [37] to measure the perceptual similarity between any given synthesized image and its ground truth image similar to how Structured Similarity Index Measure (SSIM) [38] is used. LPIPS is very similar to human perception. However, LPIPS cannot be applied to the C2C market dataset because LPIPS needs ground truth data. LPIPS is defined as

$$d(x, x_0) = \sum_l \frac{1}{H_l W_l} \sum_{h,w} \left\| w_l \odot (y_{hw}^l - y_{0hw}^l) \right\|_2^2, \quad (11)$$

where x is the ground truth image, x_0 is the generated image, l is the number of the layers of a pre-trained network (in our case, AlexNet [39]), H, W is the height and width of the images respectively, w_l are used in scaling the activations channel-wise and computing the ℓ_2 distance and y^l, y_0^l the are output of the layer of the pre-trained network (AlexNet).

4.1.2. Fréchet Inception Distance (FID).

We also use the Fréchet Inception Distance (FID) [40]. The generated images are high fidelity perceptually when the FID is low, because the FID has been found to be consistent with human judgement [40]. The FID measures the discrepancy between any given real data distribution and generated data distribution. The FID uses 2nd order information of the intermediate layer of the pre-trained Inception-v3 model. On its own, the Fréchet Distance [41] is the Wasserstein-2 distance between two distributions p_1 and p_2 assuming they are both multivariate Gaussian distributions. The FID is defined as

$$F(p_1, p_2) = \|\mu_{p_1} - \mu_{p_2}\|_2^2 + \text{trace} \left(C_{p_1} + C_{p_2} - 2(C_{p_1} C_{p_2})^{\frac{1}{2}} \right), \quad (12)$$

where μ_{p_1} and C_{p_1} are the mean and covariance of the samples from p_1 and similar for p_2 . In this paper, we compute the FID 10 times using random generated paired images because the FID of each generated image is different.

4.2. Implementation Details

4.2.1. Conventional Dataset

We also conduct experiments to compare conventional supervised virtual try-on methods and our UVIRT method on the Multi-Pose Virtual try-on (MPV) datasets collected by Dong et al. [10]. It contains 35,687 person and 13,524 clothing images at 256×192 resolution. Dong et al. split them into train/test sets of size 52,236 and 10,544, respectively. We chose the front-view person and top clothing image pairs from the MPV dataset, resulting in 10,322 train and 2171 test pairs.

4.2.2. Setting.

In our experiments for comparing between conventional methods and UVIRT on MPV dataset, we trained UVIRT for 500k iterations. UVIRT consists of four encoders, two decoders, and two discriminators. The four encoders and the two decoders make up the generator and we trained using the Adam optimizer [42], with $\alpha = 0.0001$ (i.e., learning rate), $\beta_1 = 0.5$, $\beta_2 = 0.999$, and a batch size of 4. The two discriminators and the one classifier were trained using the Adam optimizer with the same hyperparameters as the generator. Note that we trained UVIRT with the paired image setting for fair comparison because CP-VTON can be only trained with the paired image setting.

In experiments using the C2C market dataset, we trained UVIRT for 500 k iterations. The architecture, optimizers, and hyperparameters were the same as the above experiments except for a batch size of 8. Note that we trained UVIRT with the unpaired setting because the C2C market dataset does not have paired images.

4.2.3. Architecture.

As shown in Figure 1, the two encoders which extract the content features are structured as a ResNet-like network, which consists of one convolutional layer, two downsample convolutional layers, and eight ResBlocks, sequentially. The two encoders which extract the style features have a fully convolutional network, which consists of one convolutional layer, four downsample convolutional

layers, and a global average pooling layer, sequentially. The two decoders are structured as a ResNet-like network, which consists of the GDWCT module whose architecture is the same as in Reference [30]. The two discriminators have fully convolutional networks, which consist of three downsample convolutional layers and one point-wise convolutional layer, sequentially. Plus, the above discriminators are multi-scale discriminators [43]. Finally, the one classifier is a simple convolutional network, which consists of two downsample convolutional layers and one fully connected layer, sequentially. Note that in the C2C market dataset, we use 2171 test images which contain only clothing because we use the test person images of the MPV dataset as the person and background domain. We fix the number of the test images of the C2C market dataset because the number of person images of the MPV dataset is 2171.

4.3. Baseline

We adopt CP-VTON [4] as the baseline image-based virtual try-on method. First, CP-VTON estimates the parameters of geometrical transformations by concatenating the key points feature, a rough body segmentation mask and only the head image in the channel dimension, and using the only clothing images. In total, the concatenated feature has 23 channels and the only clothing images have 3 channels. The key points feature has 19 channels, the rough body segmentation mask has 1 channel and the head-only image has 3 channels. Second, CP-VTON generates a warped mask and a coarse rendered person image. Finally, CP-VTON concatenates the warped mask, coarse rendered person image, and warped clothing. As a result, CP-VTON generates a virtual try-on on image. CP-VTON employs supervised training using a clothing parsing semantic segmentation mask. Moreover, CP-VTON uses key points, a body rough segmentation mask and a head-only image as auxiliary information. On the contrary, our UVIRT directly generates a virtually tried on image using completely unsupervised training.

4.4. Comparison with Conventional Supervised Methods

4.4.1. Qualitative Results

We first perform visual comparisons of our proposed UVIRT method with conventional supervised methods, namely CP-VTON [4], illustrated in Figure 3, which shows that our UVIRT method generates comparable results. Although the pose of the people in the images generated by our UVIRT method is altered, UVIRT naturally warps the clothing to the people because UVIRT can adequately disentangle person and clothing features. As a result, UVIRT provides competitive performance without any annotation cost.

4.4.2. Quantitative Results.

We also compare our proposed UVIRT method with CP-VTON quantitatively based on LPIPS and FID. Note that we virtually try on the test clothing images to the test person images based only on predetermined pairings since LPIPS can be applied to the comparison of generated images only when ground truth images are available. In Table 1, although CP-VTON achieves a lower mean LPIPS (i.e., better score) than our UVIRT method, UVIRT can generate high fidelity images. The reason for the higher LPIPS score is that LPIPS measures how a generated image corresponds to a ground truth image. Next, we measured the FID of the images generated by CP-VTON and our UVIRT method. We warp the test clothing to the random test persons 10 times since we do not need ground truth images when computing FID. We computed the mean and standard deviations of FID after we computed FID 10 times on the all test images, that is, 2171 images. In Table 1, UVIRT achieves a worse score than CP-VTON. The reason of the former deterioration is that the images generated by UVIRT exhibit some reconstruction deficiencies and distorted body parts. These results are natural in LPIPS and FID because CP-VTON is trained in a supervised manner, in comparison with our UVIRT trained

in an unsupervised manner. Note that UVIRT is the first unsupervised virtual try-on method. In future work, we aim to improve with respect to these metrics.



Figure 3. Qualitative comparison of proposed UVIRT vs. CP-VTON on the test images of the MPV dataset. The left figure shows the results of reconstructing the reference images using the target clothing. The right figure shows the results of warping target clothing to reference images. Although our UVIRT method changes the person's poses, ours can warp clothing to the persons in each of the experiments. Thus, our UVIRT method achieves competitive performance in the virtual try-on task.

Table 1. Quantitative comparison of proposed UVIRT vs. CP-VTON on the test images of the MPV dataset. Lower is better in the LPIPS and FID metrics, means and standard deviations are shown. The bold characters indicate the highest score.

Method	LPIPS ↓	FID ↓
CP-VTON (Supervised)	0.191	25.23 (±0.17)
UVIRT (ours) (Unsupervised)	0.442	44.80(±0.21)

4.5. Virtual Try-on on C2C Market Dataset

4.5.1. Qualitative Results

We conduct experiments on the C2C market dataset and compare visual results of our UVIRT and CP-VTON in Figure 4. In this experiment, we use CP-VTON trained with the MPV dataset since CP-VTON cannot be trained in an unsupervised setting. In testing time, CP-VTON can be applied to novel dataset when anyone have person images, only clothing images, human-parsing semantic segmentation masks and pose features. Human-parsing semantic segmentation masks and pose features are generated by a pre-trained human-parser and pose estimator. In Figure 4, the images generated by CP-VTON are distorted and have color artifacts because CP-VTON warps geometrically transformed clothing onto the person. CP-VTON is affected by distribution shift as the MPV dataset is aligned by taking the white background and centering. On the other hand, our UVIRT can generate

high fidelity person images with tried on clothing without aligned images since ours exchanges the disentangled person and clothing features in the latent space. Thus, our UVIRT is an effective method when ground truth and aligned images are unavailable. Although our UVIRT indicates competitive results with the supervised method, ours does share some of the same problems. First, in Figure 5a, not only did the color of the target clothing change but also the face of the generated image is not reconstructed well. Moreover, in Figure 5b, not only did the color of the clothing change but also the person changed. The reason for the above problems is that UVIRT cannot disentangle between the person and the clothing features well enough. UVIRT encodes input images to the style and content features. It is possible that the style and content features are not semantically orthogonal. Thus, the style and content features are entangled. Therefore, we must tackle the entanglement problem. Concretely, we will try using the unsupervised disentanglement techniques of the generative models [16–22,44] in future work.

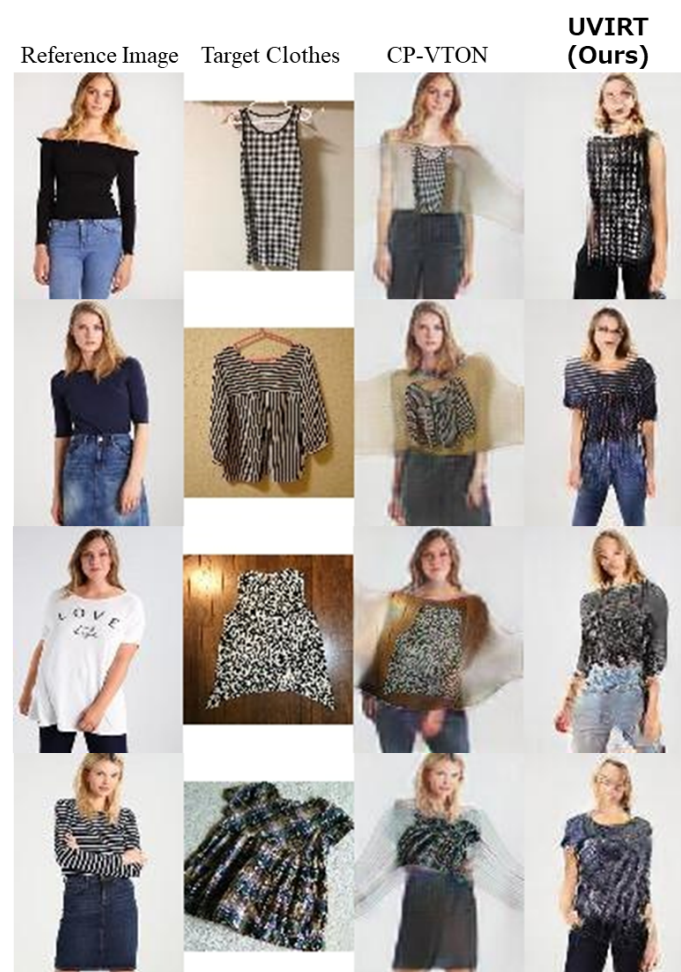


Figure 4. Qualitative comparison of Proposed UVIRT vs. CP-VTON on the test images of the C2C market dataset. CP-VTON cannot warp the unaligned clothing to the persons well. On the other hand, our UVIRT can warp the unaligned images to the person because of disentangling the person and the clothing features in the latent space.

4.5.2. Quantitative Results.

We compare the conventional supervised virtual try-on method CP-VTON and our unsupervised virtual try-on method UVIRT using the widely used metric, FID to verify the performance of image synthesis, summarized in Table 2. CP-VTON was trained with the MPV dataset because it is a supervised method and cannot be trained on the C2C market dataset. UVIRT was trained with the

MPV dataset and the C2C market dataset for fair comparison. When both methods were trained on the MPV dataset, CP-VTON showed better performance on the C2C market dataset. However, when UVIRT was trained on the C2C market dataset (something that cannot be done for CP-VTON), UVIRT demonstrated significantly better performance. The strong point of UVIRT is that it can be trained with in-the-wild datasets. Our UVIRT also achieves more high fidelity image generation than CP-VTON when trained with the C2C market dataset. Consequently, we have demonstrated our UVIRT is effective qualitatively and quantitatively in an in-the-wild dataset virtual try-on task.



Figure 5. Failure case of our UVIRT. (a): The color of the target clothing change. Moreover, the face of the generated image is not reconstructed well because the generated image entangles the person and the clothing features. (b): Not only do the color of the clothing change but also the person changes because of the same reasoning discussed in (a).

Table 2. Quantitative comparison of proposed UVIRT vs. CP-VTON on the test images of the C2C market dataset. Our UVIRT outperforms CP-VTON pre-trained with the MPV dataset.

Method	FID ↓
CP-VTON * (Supervised)	122.83 (± 0.51)
UVIRT * (ours) (Unsupervised)	245.25 (± 0.77)
UVIRT [†] (ours) (Unsupervised)	92.55 (± 0.29)

* methods were trained with the MPV dataset for fair comparison, namely the cross-dataset evaluation.

[†] method was trained with the C2C market dataset.

4.6. Ablation Study

In this section, we conduct an ablation study on the scale of the C2C market dataset. We confirm whether or not the scale of the dataset affects virtual try-on results in our unlabeled setting. In Table 3, we use the FID metric for comparison of the training data at different scales. We randomly extract the training data at 5%, 10%, 25% and 50% from the full training dataset. We compute FID in test images after we training UVIRT with each extracted training data. When we use the 50% scale setting of the training data, FID is best because UVIRT could capture the data distribution correctly from the increased amount of training data. However, when we use the 100% scale of the training data, the FID of UVIRT worsens compared with the 50% scale of the training data. We assume that UVIRT does not capture the data distribution well because more training epochs were needed in the 100% scale setting. Computational resources limited our ability to fully train the model on all the data. But we believe that even with limited resources, training with medium sized datasets can still give good results.

Table 3. Comparison of the impact of the training data scale on the C2C market dataset. In the unsupervised setting, the C2C market dataset at 50% scale is superior to other scale settings.

Training Data Scale	FID ↓
5% (33,721 images)	95.01 (± 0.26)
10% (67,443 images)	98.97 (± 0.41)
25% (134,886 images)	99.31 (± 0.60)
50% (269,773 images)	92.55 (± 0.29)
100% (539,546 images)	95.13 (± 0.24)

4.7. Testing Time Comparison

We compare the testing time between UVIRT and CP-VTON on the C2C market dataset. This comparison excludes the data loading time and counts only the testing time. In Table 4, the testing time is the mean forward propagation time on the 2171 test images. We use batchsize = 1. The testing time of CP-VTON includes the time needed to run a pre-trained human keypoints detector (OpenPose) [45–48] and the pre-trained human parser (JPPNet) [49,50]. CP-VTON needs the keypoints and the human-parsing semantic segmentation mask in its testing phase as well as training phase. In Table 4, UVIRT outperforms CP-VTON by about 13 times. UVIRT not only is a fully-unsupervised method but also has fast testing time.

Table 4. Testing time comparison of proposed UVIRT vs. CP-VTON on the test images of the C2C market dataset. We use an NVIDIA Tesla V100 in our tests.

Method	Testing Time [msec] ↓
CP-VTON (Supervised)	662.81
UVIRT (Unsupervised)	53.94

5. Conclusions

In this paper, we proposed a novel virtual try-on system, namely Unsupervised VIRTual Try-on system (UVIRT). The annotation cost of the training data for virtual try-on is high because the conventional supervised approaches need annotated data such as clothing-parsing semantic segmentation masks and paired images. Therefore, we resolved the problem of the annotation cost through our unsupervised approach and by using a state-of-the-art image-to-image translation method. In addition, we can apply UVIRT to practical scenarios for virtual try-on in a real-world C2C market. Conventional supervised virtual try-on methods cannot be applied to the C2C market dataset because the accumulation of annotated data for the C2C market is very difficult. In the C2C market dataset, we outperformed a pre-trained conventional supervised virtual try-on system. Moreover, we not only constructed a new C2C market dataset but also indicate the usefulness of unsupervised virtual try-on

by investigating how the size of the C2C market dataset affects performance. In future work, we will resolve the issues with distortion of the person and deficiencies in face reconstruction.

Author Contributions: Conceptualization, H.T.; methodology, H.T.; software, H.T.; validation, H.T.; formal analysis, H.T.; investigation, H.T.; resources, H.T.; data curation, H.T., K.A. and A.L.; writing—original draft preparation, H.T.; writing—review and editing, K.A., A.L. and H.K.; visualization, H.T.; supervision, H.K.; project administration, H.K.; funding acquisition, K.A., A.L. and H.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Statista. eBay: Annual Net Revenue 2013–2019. 2020. Available online: <https://www.statista.com/statistics/507881/ebays-annual-net-revenue/> (accessed on 4 August 2020).
2. Amazon. Prime Wardrobe. 2020. Available online: https://www.amazon.co.jp/b/?ie=UTF8&bbn=5429200051&node=5425661051&tag=googhydr-22&ref=pd_sl_pn6acna5m_b&adgrpid=60466354043&hvpon=&hvptwo=&hvadid=338936031340&hvpos=&hvnetw=g&hvrnd=9233286119764777328&hvqmt=b&hvdev=c&hvdvcmdl=&hvlocint=&hvlocphy=1009310&hvtargid=kwd-327794409245&hydadcr=3627_11172809&gclid=Cj0KCQjw6_vzBRCIARIsAOs54z7lvkoeDTmckFM2Vakdras4uWdqWSe2ESCzS6dRoX_fTNLNE8Y5p-8aAq1REALw_wcB (accessed on 4 August 2020).
3. Han, X.; Wu, Z.; Wu, Z.; Yu, R.; Davis, L.S. VITON: An Image-based Virtual Try-on Network. In Proceedings of the CVPR 2021: IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 19–21 June 2018; pp. 7543–7552.
4. Wang, B.; Zheng, H.; Liang, X.; Chen, Y.; Lin, L. Toward Characteristic-Preserving Image-based Virtual Try-On Network. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 589–604.
5. Hsieh, C.; Chen, C.; Chou, C.; Shuai, H.; Cheng, W. Fit-me: Image-Based Virtual Try-on With Arbitrary Poses. In Proceedings of the 2019 IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 22–25 September 2019; pp. 4694–4698.
6. Yildirim, G.; Jetchev, N.; Vollgraf, R.; Bergmann, U. Generating High-Resolution Fashion Model Images Wearing Custom Outfits. In Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops, Seoul, Korea, 27 October–2 November 2019; doi:10.1109/ICCVW.2019.00389.
7. Yu, R.; Wang, X.; Xie, X. VTNFP: An Image-Based Virtual Try-On Network With Body and Clothing Feature Preservation. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 10511–10520.
8. Jae Lee, H.; Lee, R.; Kang, M.; Cho, M.; Park, G. LA-VITON: A Network for Looking-Attractive Virtual Try-On. In Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops, Seoul, Korea, 27 October–2 November 2019; pp. 3129–3132.
9. Issenhuth, T.; Mary, J.; Calauzènes, C. End-to-End Learning of Geometric Deformations of Feature Maps for Virtual Try-On. *arXiv* **2019**, arXiv:cs.CV/1906.01347.
10. Dong, H.; Liang, X.; Shen, X.; Wang, B.; Lai, H.; Zhu, J.; Hu, Z.; Yin, J. Towards Multi-Pose Guided Virtual Try-On Network. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 9026–9035.
11. Pumarola, A.; Goswami, V.; Vicente, F.; De la Torre, F.; Moreno-Noguer, F. Unsupervised Image-to-Video Clothing Transfer. In Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops, Seoul, Korea, 27 October–2 November 2019; doi:10.1109/ICCVW.2019.00394.
12. Wang, K.; Ma, L.; Oramas, J.; Gool, L.V.; Tuytelaars, T. Unsupervised shape transformer for image translation and cross-domain retrieval. *arXiv* **2018**, arXiv:cs.CV/1812.02134.
13. Jetchev, N.; Bergmann, U. The Conditional Analogy GAN: Swapping Fashion Articles on People Images. In Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops, Venice, Italy, 22–29 October 2017; pp. 2287–2292.

14. Mercari. The Total Merchandise Number of C2C Market Application “Mercari” Goes Beyond 10 Billion (Japanese Website). 2018. Available online: https://about.mercari.com/press/news/article/20180719_billionitems/ (accessed on 4 August 2020).
15. Bengio, Y.; Courville, A.; Vincent, P. Representation Learning: A Review and New Perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 1798–1828.
16. Higgins, I.; Matthey, L.; Pal, A.; Burgess, C.; Glorot, X.; Botvinick, M.; Mohamed, S.; Lerchner, A. beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. In Proceedings of the International Conference on Learning Representations, Toulon, France, 24–26 April 2017.
17. Chen, X.; Duan, Y.; Houthoofd, R.; Schulman, J.; Sutskever, I.; Abbeel, P. InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets. In Proceedings of the Advances in Neural Information Processing Systems 29, Centre Convencions Internacional Barcelona, Barcelona, Spain, 5–10 December 2016.
18. Kim, H.; Mnih, A. Disentangling by Factorising. In Proceedings of the 35th International Conference on Machine Learning, Stockholmsmässan, Stockholm, Sweden, 10–15 July 2018.
19. Mathieu, E.; Rainforth, T.; Siddharth, N.; Teh, Y.W. Disentangling Disentanglement in Variational Autoencoders. In Proceedings of the 36th International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019.
20. Kumar, A.; Sattigeri, P.; Balakrishnan, A. Variational Inference of Disentangled Latent Concepts from Unlabeled Observations. In Proceedings of the International Conference on Learning Representations, Vancouver, BC, Canada, 30 April–3 May 2018.
21. Esmaili, B.; Wu, H.; Jain, S.; Bozkurt, A.; Siddharth, N.; Paige, B.; Brooks, D.H.; Dy, J.; van de Meent, J.W. Structured Disentangled Representations. In Proceedings of the Machine Learning Research, Long Beach, CA, USA, 9–15 June 2019.
22. Karras, T.; Laine, S.; Aila, T. A Style-Based Generator Architecture for Generative Adversarial Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019.
23. Isola, P.; Zhu, J.Y.; Zhou, T.; Efros, A.A. Image-To-Image Translation With Conditional Adversarial Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
24. Zhu, J.Y.; Zhang, R.; Pathak, D.; Darrell, T.; Efros, A.A.; Wang, O.; Shechtman, E. Toward Multimodal Image-to-Image Translation. In Proceedings of the Advances in Neural Information Processing Systems 30, Long Beach, CA, USA, 4–9 December 2017.
25. Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; Schiele, B. The Cityscapes Dataset for Semantic Urban Scene Understanding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016.
26. Liu, M.Y.; Breuel, T.; Kautz, J. Unsupervised Image-to-Image Translation Networks. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017.
27. Huang, X.; Liu, M.Y.; Belongie, S.; Kautz, J. Multimodal Unsupervised Image-to-image Translation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018.
28. Choi, Y.; Choi, M.; Kim, M.; Ha, J.; Kim, S.; Choo, J. StarGAN: Unified Generative Adversarial Networks for Multi-domain Image-to-Image Translation. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 8789–8797.
29. Zhu, J.Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2242–2251.
30. Cho, W.; Choi, S.; Park, D.K.; Shin, I.; Choo, J. Image-To-Image Translation via Group-Wise Deep Whitening-And-Coloring Transformation. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 10631–10639.
31. Güler, R.A.; Neverova, N.; Kokkinos, I. Densepose: Dense human pose estimation in the wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7297–7306.

32. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid Scene Parsing Network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
33. Huang, X.; Belongie, S. Arbitrary Style Transfer in Real-Time with Adaptive Instance Normalization. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017.
34. Mao, X.; Li, Q.; Xie, H.; Lau, R.Y.; Wang, Z.; Paul Smolley, S. Least Squares Generative Adversarial Networks. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017.
35. Mercari, I. Mercari (Japanese Website). 2020. Available online: <https://www.mercari.com/jp/> (accessed on 4 August 2020).
36. Salimans, T.; Goodfellow, I.; Zaremba, W.; Cheung, V.; Radford, A.; Chen, X.; Chen, X. Improved Techniques for Training GANs. In Proceedings of the Advances in Neural Information Processing Systems 29, Barcelona, Spain, 5–10 December 2016.
37. Zhang, R.; Isola, P.; Efros, A.A.; Shechtman, E.; Wang, O. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018.
38. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612.
39. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. In Proceedings of the Advances in Neural Information Processing Systems 25, Lake Tahoe, CA, USA, 3–8 December 2012.
40. Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; Hochreiter, S. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In Proceedings of the Advances in Neural Information Processing Systems 30, Long Beach, CA, USA, 4–9 December 2017.
41. Dowson, D.C.; Landau, B.V. The frechet distance between multivariate normal distributions. *J. Multivar. Anal.* **1982**, *12*, 450–455.
42. Kingma, D.; Ba, J. Adam: A Method for Stochastic Optimization. In Proceedings of the International Conference on Learning Representation (ICLR), San Diego, CA, USA, 7–9 May 2015.
43. Wang, T.; Liu, M.; Zhu, J.; Tao, A.; Kautz, J.; Catanzaro, B. High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs. In Proceedings of the 2018 the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 8798–8807.
44. Zhu, X.; Xu, C.; Tao, D. Learning Disentangled Representations with Latent Variation Predictability. In Proceedings of the European Conference on Computer Vision (ECCV), Glasgow, UK, 23–28 August 2020.
45. Cao, Z.; Hidalgo Martinez, G.; Simon, T.; Wei, S.; Sheikh, Y.A. OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, doi:10.1109/TPAMI.2019.2929257.
46. Cao, Z.; Simon, T.; Wei, S.E.; Sheikh, Y. Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
47. Simon, T.; Joo, H.; Matthews, I.; Sheikh, Y. Hand Keypoint Detection in Single Images using Multiview Bootstrapping. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
48. Wei, S.E.; Ramakrishna, V.; Kanade, T.; Sheikh, Y. Convolutional pose machines. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016.
49. Gong, K.; Liang, X.; Zhang, D.; Shen, X.; Lin, L. Look Into Person: Self-Supervised Structure-Sensitive Learning and a New Benchmark for Human Parsing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
50. Liang, X.; Gong, K.; Shen, X.; Lin, L. Look into Person: Joint Body Parsing & Pose Estimation Network and a New Benchmark. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *41*, 871–885.

