Edit Like A Designer: Modeling Design Workflows for Unaligned Fashion Editing

Qiyu Dai
Wangxuan Institute of Computer Technology, Peking University
School of Software and Microelectronics, Peking University
qiyudai@pku.edu.cn

Shuai Yang
Wangxuan Institute of Computer Technology, Peking University
williamyang@pku.edu.cn

Wenjing Wang
Wangxuan Institute of Computer Technology, Peking University
daooshee@pku.edu.cn

Wei Xiang
Bigo
Beijing, China
xiangwei1@bigo.sg

Jiaying Liu*
Wangxuan Institute of Computer Technology, Peking University
liujiaying@pku.edu.cn

ABSTRACT

Fashion editing has drawn increasing research interest with its extensive application prospect. Instead of directly manipulating the real fashion item image, it is more intuitive for designers to modify it via the design draft. In this paper, we model design workflows for a novel task of unaligned fashion editing, allowing the user to edit a fashion item through manipulating its corresponding design draft. The challenge lies in the large misalignment between the real fashion item and the design draft, which could severely degrade the quality of editing results. To address this issue, we propose an Unaligned Fashion Editing Network (UFE-Net). A coarsely rendered fashion item is firstly generated from the edited design draft via a translation module. With this as guidance, we align and manipulate the original unedited fashion item via a novel alignment-driven fashion editing module, and then optimize the details and shape via a reference-guided refinement module. Furthermore, a joint training strategy is introduced to exploit the synergy between the alignment and editing tasks. Our UFE-Net enables the edited fashion item to have semantically consistent geometric shape and realistic details to the edited draft in the edited region, as well as to keep the unedited region intact. Experiments demonstrate our superiority over the competing methods on unaligned fashion editing.

CCS CONCEPTS

• **Computing methodologies** → **Computer vision**; **Neural networks**; • **Applied computing** → **Fine arts**.

*Corresponding Author. This work was supported by the Fundamental Research Funds for the Central Universities and the National Natural Science Foundation of China under Contract No.61772043. This is a research achievement of Key Laboratory of Science, Technology and Standard in Press Industry (Key Laboratory of Intelligent Press Media Technology).

MM '21, October 20–24, 2021, Virtual Event, China

ACM ISBN 978-1-4503-8651-7/21/10…$15.00
https://doi.org/10.1145/3474085.3475511

(a) The workflow of editing like designers

(b) Result of our unaligned fashion editing

Figure 1: Our UFE-Net models real-world design workflows, allowing users to edit the fashion item through manipulating the design draft. Given an input real fashion item image, our model generates its corresponding design draft automatically. After performing flexible editing on the design draft, the network then renders a realistic fashion editing result.

KEYWORDS

Fashion editing, Image-to-image translation, Image alignment

ACM Reference Format:
Qiyu Dai, Shuai Yang, Wenjing Wang, Wei Xiang, and Jiaying Liu. 2021. Edit Like A Designer: Modeling Design Workflows for Unaligned Fashion Editing. In *Proceedings of the 29th ACM International Conference on Multimedia (MM '21), October 20–24, 2021, Virtual Event, China.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3474085.3475511

1 INTRODUCTION

With an explosively growing demand for fashion and individuality in modern society, an increasing number of people tend to make fashion customization or design fashion items to fit in with their

personal expressions, which drives fashion editing as a hot field, receiving a lot of attention. Fashion editing is tasked to interactively manipulate the fashion images by users, including virtual make-up, human pose and clothing transfer, virtual try-on, *etc.* In this work, we pay attention to the manipulation on real clothing fashion items.

The recent development of Generative Adversarial Networks (GANs) has made fashion editing easier for normal users. The generative models make it possible for users to render realistic fashion item images that meet their needs, which has applications with great potentials including fashion design, clothing customization, and online shopping. Previous generative fashion editing methods focus on directly manipulating real fashion images, conditioned on coarse-level user-provided input, such as attribute labels [1], sketches and color strokes [5], and fine-level outfits for style reference [6, 10, 18, 26, 28]. However, in actual design workflows, designers are more inclined to reflect their ideas with the design draft, so that they can flexibly choose the granularity of manipulation (*e.g.*, coarsely change the color of the whole clothing, or make fine-grained adjustments to the collar) and easily modify their clothing design via the design draft rather than directly modifying the real item[1]. For people with no design experience, **it is a more natural and feasible way to follow such design workflows: perform editing on the design drafts at first, and then render the real fashion items with corresponding modifications.** This has motivated our work. By modeling design workflows, we investigate a new task of design-draft-driven fashion editing, enabling users to freely modify the fashion item with the help of the design draft as shown in Fig. 1.

Design-draft-driven fashion editing aims at manipulating real fashion item photos by performing flexible editing on the corresponding design drafts. Our goal is that the edited fashion item should accurately reflect the modification of the design draft within the editing area and have a good visual quality such as clear outline and rich details, while keeping consistent with the original real photo in the other area.

However, a major challenge lies in the huge misalignment between the fashion item and design draft. Fig. 1 gives some examples of the fashion item and its design draft from the only related design draft dataset [8] to our best knowledge. As can be seen, there is an evident structural discrepancy between the two domains. Moreover, the pliable clothes are deformed according to the poses of the models, making it even hard to find their pixel-wise correspondence. Most supervised image-to-image translation methods, such as Pix2pix [12], Pix2pixHD [27], and SPADE [20], require strict alignment at the pixel level, while unsupervised methods such as CycleGAN [31] focus on unpaired translation, which are not suitable for our translation problem with paired but unaligned data. To the best of our knowledge, D2RNet [8] is the first attempt to address the problem of unaligned real fashion item and design draft translation. However, it is designed for translation rather than editing, thus its cycle translation results suffer from serious shape distortion and texture artifacts within both the editing region and the rest, which has not reached the goal of fashion editing.

In this paper, we put forward a new draft-driven fashion editing framework to model real-world design workflows and propose a novel Unaligned Fashion Editing Network (UFE-Net) to address the aforementioned challenges. Our UFE-Net contains three modules: a fashion item and design draft translation module, an alignment-driven fashion editing module, and a reference-guided refinement module. Our key idea is to learn accurate alignment and robust editing jointly, so as to exploit the synergy between the two tasks to improve the performance of each other. Specifically, we first map a real fashion item to its corresponding design draft to allow user editing. Then, instead of arduously seeking a perfect cycle-consistent translation, we get the coarsely edited fashion item rendered from the edited design draft via the translation module and only use it as guidance. With the guidance, next, our key alignment-driven fashion editing module aligns and edits the original fashion item, where we will show later that joint learning can significantly improve the performance. Finally, in order to remove the artifacts and polish appearance, the manipulated fashion item is further refined via a reference-guided refinement module, so we can get the ultimate high-quality photorealistic editing results. Experimental results demonstrate the superiority of the proposed method over other state-of-the-art baselines in design-draft-driven fashion editing. We summarize our main contributions as follows:

- We propose a new draft-driven fashion editing framework to model real-world design workflows, enabling the user to conveniently and naturally edit fashion items by modifying their corresponding design drafts like professional designers.
- We present a novel Unaligned Fashion Editing Network, which combines key processes of design draft and fashion item translation, coarse-to-fine alignment, feature-based editing, as well as shape and appearance refinement, to progressively render the photo-realistic editing fashion item results. To the best of our knowledge, our UFE-Net is the first to address the problem of draft-driven unaligned fashion editing.
- We propose an alignment-driven editing module, which exploits the synergy between the alignment and editing tasks to improve the performance of each other, making the edited fashion item well aligned with the original one, while synthesizing plausible content in the editing region.

2 RELATED WORK

2.1 Image-to-Image Translation

Image-to-image translation [12] aims to map images from one domain to another, and has demonstrated promising results in image editing [2, 3, 13]. Based on the data, image-to-image translation can be categorized into supervised and unsupervised methods. Supervised image-to-image translation requires paired data. The pioneering Pix2pix [12] shows good results in generating realistic photos from sketches and the segmentation. Follow-ups improve pix2pix in high-resolution image synthesis [27], multi-modal image translation [32, 34] and SPADE for robust input condition [20]. However, most supervised methods require pixel-wise alignment between the training image pair, which is not satisfied in our task.

Meanwhile, unsupervised methods [2, 11, 15, 17, 30] focus on unpaired data. CycleGAN [31] proposes a novel cycle consistency constraint to build the pixel-wise relationship between the two domains. Recently, CUT [19] proposes to use contrastive learning to adaptively learn a more robust cross-domain correspondence.

[1] https://en.wikipedia.org/wiki/Fashion_design#History

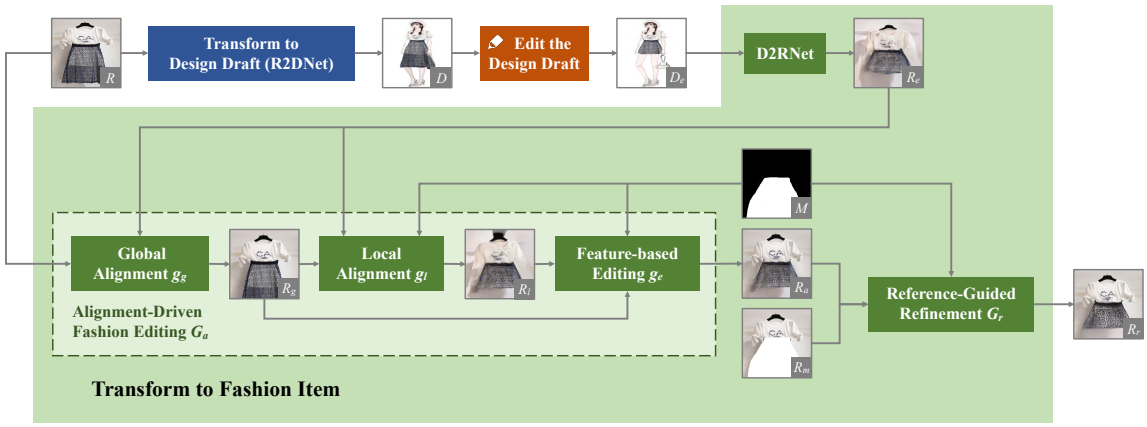

Figure 2: Illustration of the proposed UFE-Net in the test phase. The input fashion item R is first transformed to the design draft D via R2DNet. Then, users conduct editing on D and provide a mask M to indicate the editing region. Finally, we transform the edited D_e back to the ultimate photo-realistic edited fashion item R_r. Specifically, during the transformation, we first obtain a coarse result R_e via D2RNet. Then, alignment-driven editing module G_a globally aligns R to R_e via g_g, followed by R_e locally aligned and edited via g_l and g_e, respectively. The overall shape and appearance of the resulting R_a are finally refined via reference-guided refinement module G_r.

Different from unsupervised tasks, our problem handles paired but unaligned data, where the core lies in robust alignment.

2.2 Fashion Editing

The recent rapid advancement of deep learning has motivated many human-centric image generation and manipulation researches. Some researchers focus on facial images. For example, StarGAN [2] flexibly edits facial attributes by learning the mappings between multiple attribute labels. Based on the image-to-image translation framework, Gu *et al.* [7] use facial masks to modify the local facial regions. SEAN [33] is built upon SPADE [20] to incorporate the style features in each facial region for fine-grained facial style transfer.

Besides facial images, works have been done on the whole human body for fashion editing. Compared with the regular faces, human poses and clothes are more diverse, thus are much more challenging. To deal with the diverse pose, Dong *et al.* propose WarpingGAN [4] to learn the warping from the source pose to the target. CoCosNet [30] further refines the coarsely warped results by using the SPADE framework as post-processing. Zhu *et al.* [34] exploit human parsing results for the manipulation of multi-modal clothing style transfer. HumanGAN [22] combines VAE and GAN to better model the clothing styles.

Recently, attention have been paid towards virtual try-on. CP-VTON [26] warps the real fashion item to fit the pose of a target person through geometric matching. ACGPN [28] and DCTON [6] improve CP-VTON by additionally considering the occlusion problem of the arms to synthesize more natural sleeves. Outfit-VTON [18] learns mappings from dense pose to parsing and further to the color images based on the shape and appearance reconstruction. The aforementioned methods are mostly based on coarse-grained parsing or pose information, which is inconsistent with the design workflows of designers based on design drafts in the real world. As a result, they rely on existing clothes for try-on, and cannot achieve fine-grained clothes editing such as modifying the collar.

To the best of our knowledge, D2RNet [8] is the first work to handle the design draft for fashion synthesis. It learns a bi-directional mapping between the design drafts and real fashion items. However, this method suffers from structure distortion and texture blur problems, and fails to keep the unedited region intact. In this work, we propose a novel UFE-Net for coarse-to-fine shape alignment and appearance refinement, which effectively solves the distortion and blur problem.

3 MODELING DESIGN WORKFLOWS FOR UNALIGNED FASHION EDITING

Motivated by the creation process of designers in the real world, in this paper, we focus on modeling the fashion editing workflow, enabling users to edit real fashion item images by modifying design drafts on a region of interest. Specifically, given a fashion item image R, our goal is to infer its corresponding design draft D at first, so that easy and flexible editing can be conducted on it. After that, the edited design draft D_e is used as a guidance to edit R and synthesize the corresponding final edited fashion item image R_e.

However, there is a great challenge that design drafts and real items are not strictly aligned at pixel level, which increases the difficulty of translation between these two image domains, leading to serious degradation of the edited item results.

To this end, we present a novel unaligned fashion editing network UFE-Net. As illustrated in Fig. 2, we first perform real fashion item to design draft translation, and the reverse translation to get the roughly edited item (Section 3.1). Then, a novel alignment-driven fashion editing module is proposed to jointly align and manipulate the real fashion item guided by the rough edited item image (Section 3.2). Finally, a reference-guided refinement module is designed to further optimize the edited image to obtain the final editing result with refined details (Section 3.3).

3.1 Fashion Item and Design Draft Translation

At this stage, we aim at finding a cycle translation between the fashion item photo and the design draft, allowing users to easily edit the translated design draft rather than the original photo, and

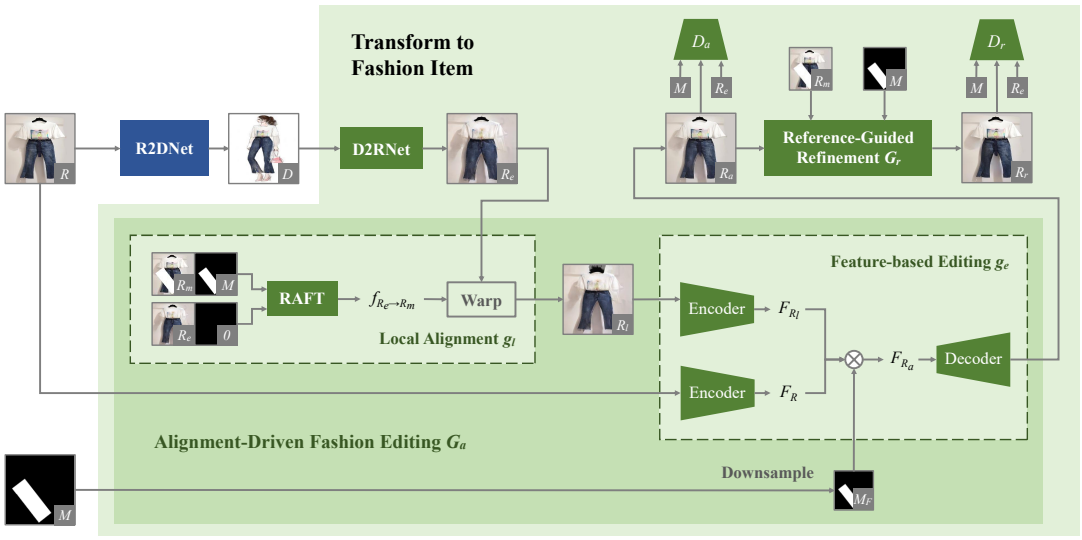

Figure 3: Illustration of the proposed UFE-Net in the train phase. First, the network conducts a cycle translation from the input fashion item photo R to the design draft D, and reversely from D to the roughly edited R_e. Then we jointly train g_l and g_e to warp R_e to align with the masked original image R_m for local alignment, and to combine the feature map of the warped image R_l and R for feature-based editing, respectively. Finally, G_r is trained to refine the aligned editing result R_a with reference of R_m. Note that during training, our task is to reconstruct R rather than modifying R.

to obtain the roughly edited photo from the edited draft. Among the existing image-to-image translation algorithms, we adopt the state-of-the-art photo-draft two-way translation framework D2RNet [8]. It consists of a submodule R2DNet for real fashion item to design draft translation, and a submodule D2RNet for the reverse translation. Given a target fashion item image R, we obtain its design draft $D = \text{R2DNet}(R)$. Then, users can perform different types of editing on D to get the edited design draft D_e. Finally, D2RNet synthesizes the edited real item $R_e = \text{D2RNet}(D_e)$. Please refer to [8] for the implementation details of R2DNet and D2RNet.

Although D2RNet [8] alone can accomplish the task of fashion editing to some extent, its result is far from satisfactory, as shown in Fig. 4. First, it only conducts an image-to-image translation process to generates R_e, rather than referencing R as in image editing, thus resulting in undesired inconsistency between R_e and the input R in the unedited region. Second, due to the huge structural discrepancy between the fashion item and design draft, it is hard for D2RNet to keep reasonable shape and appearance during the cycle translation, thus yielding low-quality content in the edited region. Therefore, the roughly synthesized result R_e can just be used as a reference. It is quite essential to further enhance the quality of R_e, which we will introduce in detail in the following sections.

3.2 Alignment-Driven Fashion Editing

In this section, we propose a novel alignment-driven fashion editing module G_a to jointly align R_e and R, and edit R in reference of R_e to obtain the editing result R_a with distinctive improved quality. To maintain the unedited region unchanged, we introduce an editing mask M to indicate the edited region in R, which could be easily provided by the user. Therefore, $R_a = G_a(R, R_e, M)$.

Problem analysis. There are three challenging technical roadblocks to be cleared: 1) Invalidity of R, 2) Unknown optical flow in the masked region, and 3) Large editing region. Specifically:

First, R is not always a valid alignment target as in standard image alignment problems. For example, if we increase the length of the skirt, the whole outfit is shrunk in R_e to provide space for the extended part. Since R does not provide such space, we cannot simply align R_e to R. To solve this problem, we propose a novel coarse-to-fine alignment scheme, where a global alignment g_g first adjusts R to a valid version, followed by a local alignment g_l to align R_e to the adjusted R.

Second, the edited content in R and R_e does not match, making predicting the accurate correspondence between the two images in the mask region nontrivial. Thus, it is not straightforward to exploit conventional image alignment algorithms, which yields heavy distortions within the edited area.

Third, different from the small and thin occlusions in image alignment problem, in our task, the editing is often conducted over a large region such as the sleeves and even the whole skirt. Simply driven by image alignment, it is extremely difficult to robustly estimate the accurate correspondence in such a large masked area.

To this end, we design a novel alignment-driven fashion editing framework, where optical-flow-based local image alignment g_l and feature-based image editing g_e are jointly trained to stabilize training and improve the performance of both tasks. Intuitively, the gradient feedback from g_e can help update g_l to perceive global and semantic information for robust optical flow prediction. Therefore, our alignment-driven fashion editing module G_a consists of a global alignment submodule g_g, a local alignment submodule g_l, and an editing submodule g_e, which we will detail in the following.

Global alignment. In our coarse-to-fine alignment scheme, we first apply a coarse-level global alignment to R to adjust the fashion items into appropriate scale and position that better displays the clothing as in R_e. Such rough alignment to eliminate large displacements also benefits the following fine-level local alignment.

In specific, for cases that the roughly edited fashion item R_e and the original R have significant differences in layout, we would like to align R to R_e using only scaling and translation to avoid any structure and detail distortions. We first estimate the landmarks of R_e and R respectively, using the pretrained fashion landmark detector [24][2]. Four landmark points are chosen: the highest two points (p_{hl}, p_{hr}) on the collar, and the middle two points (p_{ml}, p_{mr}) on the boundary of upper and lower clothes. For scaling transformation, we scale R with the width scaling ratio α and the height scaling ratio β:

$$\alpha = \frac{x_{p_{mr}} - x_{p_{ml}}}{x_{p'_{mr}} - x_{p'_{ml}}},\ \beta = \frac{y_{p_{hl}} - y_{p_{ml}} + y_{p_{hr}} - y_{p_{mr}}}{y_{p'_{hl}} - y_{p'_{ml}} + y_{p'_{hr}} - y_{p'_{mr}}}, \tag{1}$$

where (x_p, y_p) are the coordinates of point p, and we use p and p' to indicate landmarks in R_e and R, respectively. After scaling, for translation transformation, we translate R to match its middle two points to those in R_e. So far, we get the transformed R_g from R through the global alignment submodule: $R_g = g_g(R, R_e)$. Note that we adopt the training-free global alignment only on the testing phase.

Local alignment. Given globally aligned R_g, R_e, and the mask M, our local alignment submodule g_l aims to further locally calibrate the edited region of R_e to R_g based on their unedited regions. We employ the optical-flow-based image alignment framework, which is skilled in the accurate estimation of fine-level displacements between a pair of images, and modify the vanilla framework to handle the mismatch issue in the editing region.

Specifically, in order to prevent being affected by the unrelated content of R_g in the editing area, we masked it out: $R_m = R_g \otimes (1-M)$, where $\otimes$ denotes the element-wise multiplication. Then, we design g_l as a modified RAFT [25] to predict the optical flow $f_{R_e \to R_m}$ from R_e to R_m. Different from vanilla RAFT that takes an RGB image pair $\{R_e, R_m\}$ as input, our g_l receives a pair of four-channel tensors: $\{R_e||\mathbf{0}, R_m||M\}$, where $||$ is the channel-wise concatenation operation, and $\mathbf{0}$ is an all-zero mask. The additional mask inputs help g_l locate the modified region and pay attention to its surrounding information for flow estimation. Finally, R_e is warped onto R_m to get the local alignment result : $R_l = g_l(R_e, R_g, M)$, where $g_l(R_e, R_g, M) = \text{Warp}(R_e, f_{R_e \to R_m})$

Feature-based editing. It is a great challenge to learn an alignment between an edited image and its original one, because there is no correspondence in the edited region. Our main idea is to learn alignment and editing simultaneously, where these two tasks can benefit each other: only with accurate alignment can the editing submodule improve the quality of edited results, which in turn positively encourages the alignment submodule to find a better correspondence. Therefore, we propose to jointly train g_l and a feature-based editing submodule g_e.

Specifically, we construct the feature-based editing submodule g_e in an encoder-decoder architecture, containing an encoder to extract features, several residual blocks, and a decoder to render images from the features. The key idea is to edit in the feature domain to improve the robustness of the manipulation. For the globally aligned real fashion item R_g to be edited and the locally aligned editing reference R_l, a shared encoder extracts their feature maps, denoted by F_{R_g} and F_{R_l}, respectively. The mask M is downsampled to the size of F_{R_g}, denoted by M_F. We apply feature-based editing by copying the edited information in the mask region of F_{R_l} to the corresponding region of F_{R_g}:

$$F_{R_a} = F_{R_l} \otimes M_F + F_{R_g} \otimes (1 - M_F). \tag{2}$$

The resulting fused feature F_{R_a} is fed into the residual layers and the decoder to reconstruct the edited results $R_a = g_e(R_g, R_l, M)$.

In summary, the proposed alignment-driven fashion editing can be formulated as

$$R_a = G_a(R, R_e, M) := g_e(g_g(R, R_e), g_l(R_e, g_g(R, R_e), M), M). \tag{3}$$

Network training and loss functions. As illustrated in Fig. 3, we jointly train alignment along with editing in an adversarial way, encouraging two tasks to facilitate each other. To stabilize training, a progressive training strategy is proposed, which consists of the following three main steps:

- First, we pretrain local alignment submodule g_l with pseudo labels for supervised training. To be specific, we distort the real fashion item image R with random optical flow f_r to obtain the warped image R_{rw}, which simulates the coarsely edited image R_e from the unaligned translation module with the distorted shape[3]. Then, we generate a random mask M. In this way, g_l receives pair $\{R, R_{rw}\}$ and M as input. Following [25], we calculate the $L1$ loss between the estimated flow and the ground truth f_r. Denote $\{f_1, ..., f_K\}$ as the flow sequence estimated by g_l, where K denotes the numbers of iteration, the training objective is:

$$L_{pseudo} = \sum_{i=1}^{K} \lambda_f^{K-i} \|f_i - f_r\|_1 \tag{4}$$

  where λ_f denotes the weighting factor.
- Second, to speed up network convergence, we fix g_l to pretrain the feature-based editing submodule g_e with $\{R, R_{rw}\}$ and the random mask M.
- Finally, we jointly train alignment and editing with the real data $\{R, R_e\}$ and the random mask M.

In the last two steps, we apply a discriminator D_a following SN-PatchGAN [29] to improve the editing results through adversarial learning. We apply $L1$ loss and perceptual loss L_{perc} [14] to measure the distance at the pixel level and the feature level, respectively. We further add higher weight to the loss within the edited area, pushing the network to concentrate more on that region, which is formulated as:

$$L_1 = \mathbb{E}_{R,M}[\|(R_a - R) \otimes (\mathbf{1} + \lambda_M M)\|_1], \tag{5}$$

$$L_{perc} = \mathbb{E}_{R,M}[\sum_i \lambda_i \|(\Psi_i(R_a) - \Psi_i(R)) \otimes (\mathbf{1} + \lambda_M M)\|_2^2], \tag{6}$$

where $R_a = G_a(R, R_e, M)$, Ψ_i is the feature maps of layer i in VGG [23], λ_i is the weight of layer i, and $\lambda_M > 0$ denotes the weight of the edited region.

For adversarial loss, we adopt Hinge loss [29]:

$$\begin{aligned} L_{D_a} = &\mathbb{E}_{R,M}[ReLU(\Gamma - D_a(R, R_e, M))] \\ &+ \mathbb{E}_{R,M}[ReLU(\Gamma + D_a(R_a, R_e, M))], \end{aligned} \tag{7}$$

[2]https://github.com/svip-lab/HRNet-for-Fashion-Landmark-Estimation.PyTorch

[3]Here we directly use R instead of R_g since our simulated R_{rw} and R do not have global misalignment, *i.e.*, $R_g = R$.

$$L_{G_a} = -\mathbb{E}_{R,M}[D_a(R_a, R_e, M))], \tag{8}$$

where Γ denotes the margin parameter. And the whole loss function of joint training is:

$$L_{joint} = \lambda_1 L_1 + \lambda_2 L_{perc} + L_{G_a} + L_{D_a}. \tag{9}$$

3.3 Reference-Guided Refinement

Although the proposed alignment-driven editing module can calibrate the layout of the clothing and modulate the target content in the feature domain to achieve certain robustness, it is inevitable to produce artifacts due to the low quality of the editing reference R_e.

To this end, we design the reference-guided refinement module G_r for post-processing, which aims at solely removing the artifacts and optimizing the overall shape and appearance.

Our refinement network utilizes the encoder-decoder framework, and follows DeepFillv2 [29] to use gated convolutions [29] to select features dynamically according to the reference image, and use dilated gated convolutions in the bottleneck to cover a large receptive field. G_r is fed with the masked real fashion item R_m, the mask M, and the edited result R_a from the proposed alignment-driven editing module G_a, and learns to refine the editing region of R_a based on the clean information of R_m. Without direct copy-paste operations as in G_a, the proposed G_r is capable of synthesizing realistic editing results which keep consistency between pixels of the edited region and that of the rest, so the geometry and appearance of fashion item within the edited region are refined, as well as artifacts are eliminated. We get the refined $R_r = G_r(R, R_a, M)$[4], as the ultimate result of our whole fashion editing framework.

As with G_a, we train G_r in an adversarial manner. We first pretrain G_r on the large-scale fashion dataset [16] to learn general fashion priors. To construct the pseudo reference images, we apply random structural distortion, Gaussian blur, and color jitters to the original fashion item R. Then we fine-tune the model on the real data. To encourage the network to focus on refining particular parts of fashion item, we add masks of several predefined shapes during training, which cover the upper clothes, the lower clothes and the sleeves.

As for the loss function, we use vanilla $L1$ loss and perceptual loss L_{perc} for supervising the quality of final edited results, as well as Hinge loss for adversarial training:

$$\begin{aligned} L_{D_r} = &\mathbb{E}_{R,M}[ReLU(1 - D_r(R, R_a, M))] \\ &+ \mathbb{E}_{R,M}[ReLU(1 + D_r(R_r, R_a, M))], \end{aligned} \tag{10}$$

$$L_{G_r} = -\mathbb{E}_{R,M}[d_r(R_r, R_a, M))], \tag{11}$$

where D_r is the discriminator of G_r. The whole loss is:

$$L_{refine} = \lambda_3 L_1 + \lambda_4 L_{perc} + L_{G_r} + L_{D_r}. \tag{12}$$

4 EXPERIMENTAL RESULTS

4.1 Implementation Details

Dataset. We utilize D2R dataset [8], containing pairs of real fashion item images and design drafts with the size of 256×256. For fashion item and design draft translation, we use 818 images for training and 65 for testing. For alignment-driven editing and reference-guided refinement, we train our network using 5,850 pairs of coarsely edited real fashion item images and the corresponding original images. Moreover, DeepFashion [16], including 52,712 images of people dressed in fashion clothes, are additionally used for pretraining the reference-guided refinement module.

[4]Note that global misalignment is only applied during testing, and during training we directly use $R_g = R$.

Table 1: Quantitative results on unaligned fashion editing. The best FID score is highlighted.

Method	D2RNet	D2RNet w/ PB	Pix2pix	DeepFillv2	UFE-Net
FID Scores	101.95	98.51	89.34	91.30	**84.80**

Table 2: User study on unaligned fashion editing. The best user score is highlighted.

Method	D2RNet	D2RNet w/ PB	Pix2pix	DeepFillv2	UFE-Net
User Scores	3.41	2.71	1.84	2.39	**4.65**

Network training. For all experiment, we set $\lambda_f = 0.8$, $\lambda_M = 2$, $\lambda_1 = 50$, $\lambda_2 = 5$, $\lambda_3 = 10$, $\lambda_4 = 10$, and $\Gamma = 10$. For more network details and experimental results, please refer to the supplementary material.

4.2 Comparisons with Baseline Methods

Since there are few other works exactly handling our task, we choose the following most related benchmark methods for comparison with the proposed UFE-Net:

- **D2RNet** [8]: The state-of-the-art network for cycle translation between real fashion item and design draft.
- **D2RNet with Poisson Blending** [21]: Poisson blending is applied to merge the edited content by D2RNet into the original real fashion item image.
- **Pix2pix** [12]: The fundamental image-to-image translation network. To fit our task, Pix2pix receives a masked real fashion item image, and its corresponding edited design draft as a reference to infer the missing region.
- **DeepFillv2** [29]: The coarse-to-fine network for reference-guided image inpainting and editing.

Qualitative evaluation. As is shown in Fig. 4, we present visual comparisons with the baseline methods and our UFE-Net on the task of unaligned fashion editing.

D2RNet generates roughly edited fashion item results with seriously distorted structure and texture. D2RNet also fails to keep the unedited region intact, which is due to the misalignment between the real item and the design draft. After applying Poisson blending to the results from D2RNet, there is an evident inconsistency between the edited region and the rest, with blurry outline of clothes. The method also cannot handle the cases of changing the length of lower clothes, because the image fusion method is incapable of refining the shape and detail, as well as inferring the proper position for the edited clothes parts.

Besides, we compare with reference-guided methods. For Pix2pix, we adapt the network and task it to fill in the original fashion item image with the edited region masked, guided by the design draft. For DeepFillv2, previous one-channel sketch guidance is replaced with three-channel design draft image. We observe that both Pix2pix and DeepFillv2 fail to infer the clear shape and detailed texture conditioned on the design draft, because of the huge misalignment

Figure 4: Our UFE-Net compared with D2RNet [8], D2RNet with Poisson Blending [21], Pix2pix [12], and DeepFillv2 [29]

of the reference design draft and the original fashion item, making it difficult to establish accurate correspondence between them.

By comparison, thanks to the alignment-driven editing and reference-guide refinement, our UFE-Net outperforms all these baselines, which is superior in accurately rendering modifications on design drafts to photorealistic editing fashion items, as well as keeping the unedited region intact.

Quantitative evaluation. To quantitatively evaluate our method, we use the Fréchet Inception Distance (FID) score [9] to measure the distance between real and generate editing images. Lower FID scores indicate higher image quality. As is shown in Table 1, our UFE-Net achieves the lowest FID score, demonstrating that our method has the best generation visual quality.

Subjective evaluation. We perform a user study to measure subjective image quality. Specially, we randomly select 10 groups of results, each of which contains 5 results from 5 methods for comparison. Participants are required to assign 1 to 5 scores to 5 results in terms of the comprehensive image quality. Higher scores indicate the higher quality. We collect from 15 participants. The average preference scores are shown in Table 2. Our UFE-Net obtains the highest score, verifying our superiority over the other methods.

4.3 Fashion Editing with Cartoon Image

Since most normal users have no experience in fashion design, they may have trouble drawing on design drafts like professional designers. So it would be more feasible to modify the design draft using existing material, such as online cartoon clothing images. In Fig. 5, we show an example of editing real fashion items with

Figure 5: Result of editing real fashion item with cartoon image. Our UFE-Net can generalize to other fashion examples outside the dataset.

cartoon images on the Internet. As can be seen, a red skirt from the cartoon image is pasted to the design draft. The proposed UFE-Net effectively maps the cartoon garment to the vividly editing fashion item result, with clear structure and details.

4.4 Ablation Study

To analyze each component in the proposed UFE-Net, we conduct ablation studies with different configurations. Figs. 6– 7 present the corresponding fashion editing results.

Analysis of alignment-driven editing. We first examine the effect of each component in the alignment-driven editing module:

- **w/o mask**: Our UFE-Net without masking out the original fashion item R in the local alignment submodule g_l, directly aligning the edited image R_e with the unedited R.
- **w/o g_g**: Our UFE-Net without the global alignment submodule g_g, where the coarsely edited result R_e is directly fed into the local alignment submodule g_l in the test phase.

Figure 6: Effect of the mask-guided local alignment in handling the mismatch issue.

Figure 7: Performance analysis for the network configurations with and without different modules.

- **w/o** g_l: Our UFE-Net without the local alignment submodule g_l, where the transformed R_g from g_g is directly sent to the feature-based editing submodule g_e.
- **w/o** g_e: Our UFE-Net without the feature-based editing submodule g_e, where the reference-guided refinement module G_r directly takes the aligned result R_l as input for guidance.

Since the edited region between the original R and the edited R_e does not match, it is difficult to directly estimate an accurate correspondence between them. So we introduce an editing mask to handle the mismatch issue in the editing region. As is shown in Fig. 6, in the local alignment g_l, if we directly predict the optical flow between R_e and unmasked R, a heavy distortion of shape and texture occurs throughout the fashion item R_l, which further affects the final editing output R_r. We observe that masking out the edited region of the original image can help predict an accurate flow to generate a proper structure in the edited region, as well as keeping the rest region intact.

As shown in Fig. 7, without g_g in the test phase, it is evident that the UFE-Net fails to align the part of the garment in the edited region when changing the length of lower clothes. In our network, g_g is responsible for adjusting the fashion items into proper scale and position at the global level, while the following local alignment submodule g_l is skilled in accurate alignment, and it is impossible for g_l to make adjustments within such a large range.

When removing g_l or g_e in the training phase, we observe that the shape of clothes in the edited region is heavily distorted, as well as the details are seriously blurred, which verifies that only by jointly learning accurate alignment and robust editing can both tasks benefit each other, so the network can find better correspondence and generate high-quality editing results. Without any one of the two modules, the UFE-Net would not work.

Analysis of the framework. We further analyze our model design from the perspective of the whole framework.

- **w/o** G_a: Our UFE-Net without the whole alignment-driven editing module G_a, where G_r directly refines the result with the guidance of R_e.
- **w/o** G_r: Directly use the output of the alignment-driven editing module G_a.

When G_a is removed, the reference-guided refinement module G_r directly infers the masked region of the input fashion item image guided by the coarsely edited result without any alignment. We can see that the outline of the garment in the edited region is not appropriately adjusted, and details are not polished. Since there is a heavy distortion of shape and texture within the edited region of the coarsely edited result, it is difficult for G_r to learn a better refinement guided by such a low-quality result.

Without G_r, the outline of the garment is unclear and the appearance is not satisfactory, which means that the joint alignment and editing in the previous G_a pay more attention to the structures, having limited ability to polish the appearance of the editing result, so it is necessary to further employ G_r.

5 CONCLUSIONS

In this paper, modeling design workflows in the real world, a novel draft-driven fashion editing framework is proposed, which allows users to edit the real fashion items by manipulating their corresponding design drafts in a convenient and natural way, working like experienced designers. We present a new Unaligned Fashion Editing Network, progressively performing coarse-to-fine alignment, feature-based editing and further refinement to get photorealistic editing fashion item results, which semantically corresponds with design drafts within the edited region, and keeps the rest intact. Moreover, we propose an alignment-driven editing method, which jointly learns the alignment and editing tasks to facilitate each other, and get better results with both accurate alignment and robust editing. In the future, we will apply our editing framework to more general unaligned image-to-image translation tasks.

REFERENCES

[1] Kenan E Ak, Joo Hwee Lim, Jo Yew Tham, and Ashraf A Kassim. 2019. Attribute manipulation generative adversarial networks for fashion images. In *Proc. Int'l Conf. Computer Vision*. 10541–10550.

[2] Yunjey Choi, Minje Choi, Munyoung Kim, Jung Woo Ha, Sunghun Kim, and Jaegul Choo. 2018. StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*.

[3] Wenqing Chu, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, and Rongrong Ji. 2020. SSCGAN: Facial Attribute Editing via Style Skip Connections. In *Proc. European Conf. Computer Vision*. Springer.

[4] Haoye Dong, Xiaodan Liang, Ke Gong, Hanjiang Lai, Jia Zhu, and Jian Yin. 2018. Soft-gated warping-gan for pose-guided person image synthesis. In *Advances in Neural Information Processing Systems*.

[5] Haoye Dong, Xiaodan Liang, Yixuan Zhang, Xujie Zhang, Xiaohui Shen, Zhenyu Xie, Bowen Wu, and Jian Yin. 2020. Fashion editing with adversarial parsing learning. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*. 8120–8128.

[6] Chongjian Ge, Yibing Song, Yuying Ge, Han Yang, Wei Liu, and Ping Luo. 2021. Disentangled Cycle Consistency for Highly-realistic Virtual Try-On. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*.

[7] Shuyang Gu, Jianmin Bao, Hao Yang, Dong Chen, Fang Wen, and Lu Yuan. 2019. Mask-guided portrait editing with conditional gans. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*. 3436–3445.

[8] Yu Han, Shuai Yang, Wenjing Wang, and Jiaying Liu. 2020. From Design Draft to Real Attire: Unaligned Fashion Image Translation. In *Proc. ACM Int'l Conf. Multimedia*. 1533–1541.

[9] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *arXiv preprint arXiv:1706.08500* (2017).

[10] Chia-Wei Hsieh, Chieh-Yun Chen, Chien-Lung Chou, Hong-Han Shuai, Jiaying Liu, and Wen-Huang Cheng. 2019. FashionOn: Semantic-guided image-based virtual try-on with detailed human and clothing information. In *Proc. ACM Int'l Conf. Multimedia*. 275–283.

[11] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. 2018. Multimodal unsupervised image-to-image translation. In *Proc. European Conf. Computer Vision*. Springer, 172–189.

[12] Phillip Isola, Jun Yan Zhu, Tinghui Zhou, and Alexei A. Efros. 2017. Image-to-Image Translation with Conditional Adversarial Networks. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*. 5967–5976.

[13] Youngjoo Jo and Jongyoul Park. 2019. SC-FEGAN: face editing generative adversarial network with user's sketch and color. In *Proc. Int'l Conf. Computer Vision*. 1745–1753.

[14] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. 2016. Perceptual losses for real-time style transfer and super-resolution. In *Proc. European Conf. Computer Vision*. Springer, 694–711.

[15] Ming Yu Liu, Thomas Breuel, and Jan Kautz. 2017. Unsupervised Image-to-Image Translation Networks. In *Advances in Neural Information Processing Systems*. 700–708.

[16] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. 2016. DeepFashion: Powering Robust Clothes Recognition and Retrieval with Rich Annotations. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*.

[17] Roey Mechrez, Itamar Talmi, and Lihi Zelnik-Manor. 2018. The contextual loss for image transformation with non-aligned data. In *Proc. European Conf. Computer Vision*. 768–783.

[18] Assaf Neuberger, Eran Borenstein, Bar Hilleli, Eduard Oks, and Sharon Alpert. 2020. Image based virtual try-on network from unpaired data. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*. 5184–5193.

[19] Taesung Park, Alexei A Efros, Richard Zhang, and Jun-Yan Zhu. 2020. Contrastive learning for unpaired image-to-image translation. In *Proc. European Conf. Computer Vision*. Springer, 319–345.

[20] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. 2019. Semantic image synthesis with spatially-adaptive normalization. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*. 2337–2346.

[21] Patrick Pérez, Michel Gangnet, and Andrew Blake. 2003. Poisson image editing. In *Proc. ACM SIGGPRAH*. 313–318.

[22] Kripasindhu Sarkar, Lingjie Liu, Vladislav Golyanik, and Christian Theobalt. 2021. HumanGAN: A Generative Model of Humans Images. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*.

[23] Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *Proc. Int'l Conf. Learning Representations*.

[24] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. 2019. Deep high-resolution representation learning for human pose estimation. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*. 5693–5703.

[25] Zachary Teed and Jia Deng. 2020. Raft: Recurrent all-pairs field transforms for optical flow. In *Proc. European Conf. Computer Vision*. Springer, 402–419.

[26] Bochao Wang, Huabin Zheng, Xiaodan Liang, Yimin Chen, Liang Lin, and Meng Yang. 2018. Toward characteristic-preserving image-based virtual try-on network. In *Proc. European Conf. Computer Vision*. 589–604.

[27] Ting Chun Wang, Ming Yu Liu, Jun Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. 2018. High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*.

[28] Han Yang, Ruimao Zhang, Xiaobao Guo, Wei Liu, Wangmeng Zuo, and Ping Luo. 2020. Towards photo-realistic virtual try-on by adaptively generating-preserving image content. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*. 7850–7859.

[29] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. 2019. Free-form image inpainting with gated convolution. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*. 4471–4480.

[30] Pan Zhang, Bo Zhang, Dong Chen, Lu Yuan, and Fang Wen. 2020. Cross-domain correspondence learning for exemplar-based image translation. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*. 5143–5153.

[31] Jun Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. 2017. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. In *Proc. Int'l Conf. Computer Vision*. 2242–2251.

[32] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. 2017. Toward multimodal image-to-image translation. In *Advances in Neural Information Processing Systems*. 465–476.

[33] Peihao Zhu, Rameen Abdal, Yipeng Qin, and Peter Wonka. 2020. SEAN: Image synthesis with semantic region-adaptive normalization. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*. 5104–5113.

[34] Zhen Zhu, Zhiliang Xu, Ansheng You, and Xiang Bai. 2020. Semantically multi-modal image synthesis. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*. 5467–5476.