From Design Draft to Real Attire: Unaligned Fashion Image Translation

Anonymous submission

Supplementary Material

As supplementary material of our paper, we present the following contents:

- Detailed network architectures.
- Comparison with state-of-the-art methods. (Figs. 1−6)
- Quantitative evaluation. (Tables 1−2)
- Ablation study. (Figs. 7−10)

1. Detailed network architectures

Our generator g and G_f in both D2RNet and R2DNet utilizes the fully convolutional UNet architecture as in [2]. While our discriminator follows the PatchGAN classifier as in [2]. The saliency-based sampling network S follows the [4] for structure alignment. For D_a^{label} in R2DNet, we add another Convolution layer for classification to replace the original layer which outputs 1 channel prediction map.

2. Comparisons with State-of-the-Art Methods

D2RNet. Figs. 1–3 present the qualitative comparison on design draft to real fashion item translation with four state-of-the-art imge translation models: CycleGAN [7], Pix2pix [2], Pix2pixHD [6] and SPADE [3]. The released SPADE uses semantic segmentation as input, so we modify its inputs to our deign drafts. Our method yields more clear shape of clothes and detailed patterns in design drafts.

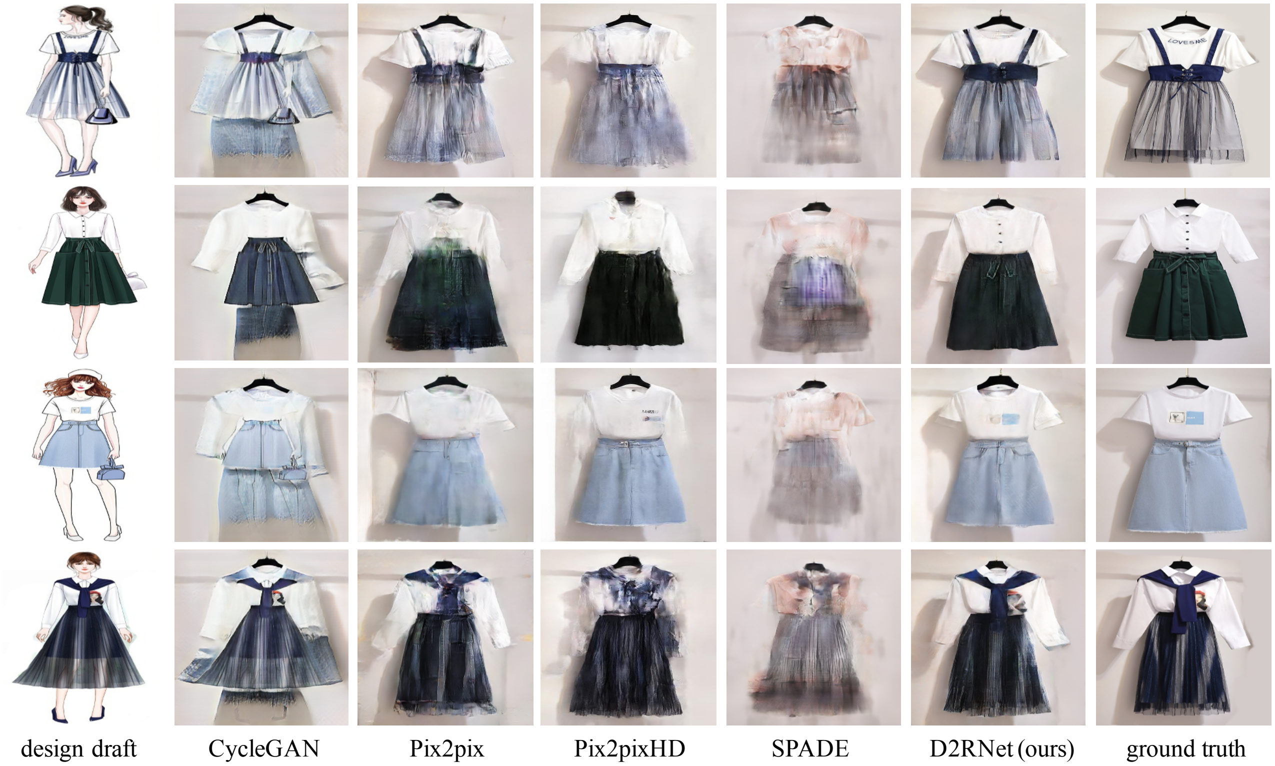

Figure 1. Our D2RNet compared with CycleGAN [7], Pix2pix [2], Pix2pixHD [6] and SPADE [3] (Part I).

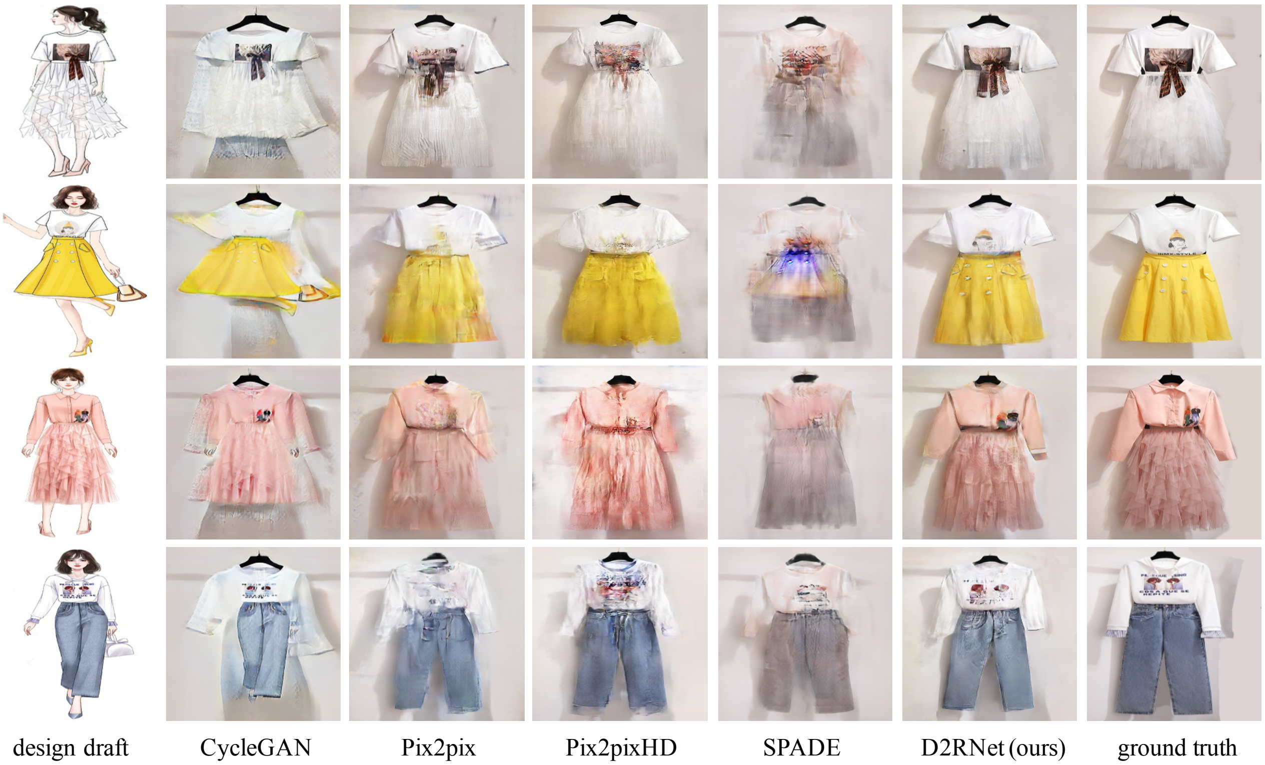

Figure 2. Our D2RNet compared with CycleGAN [7], Pix2pix [2], Pix2pixHD [6] and SPADE [3] (Part II).

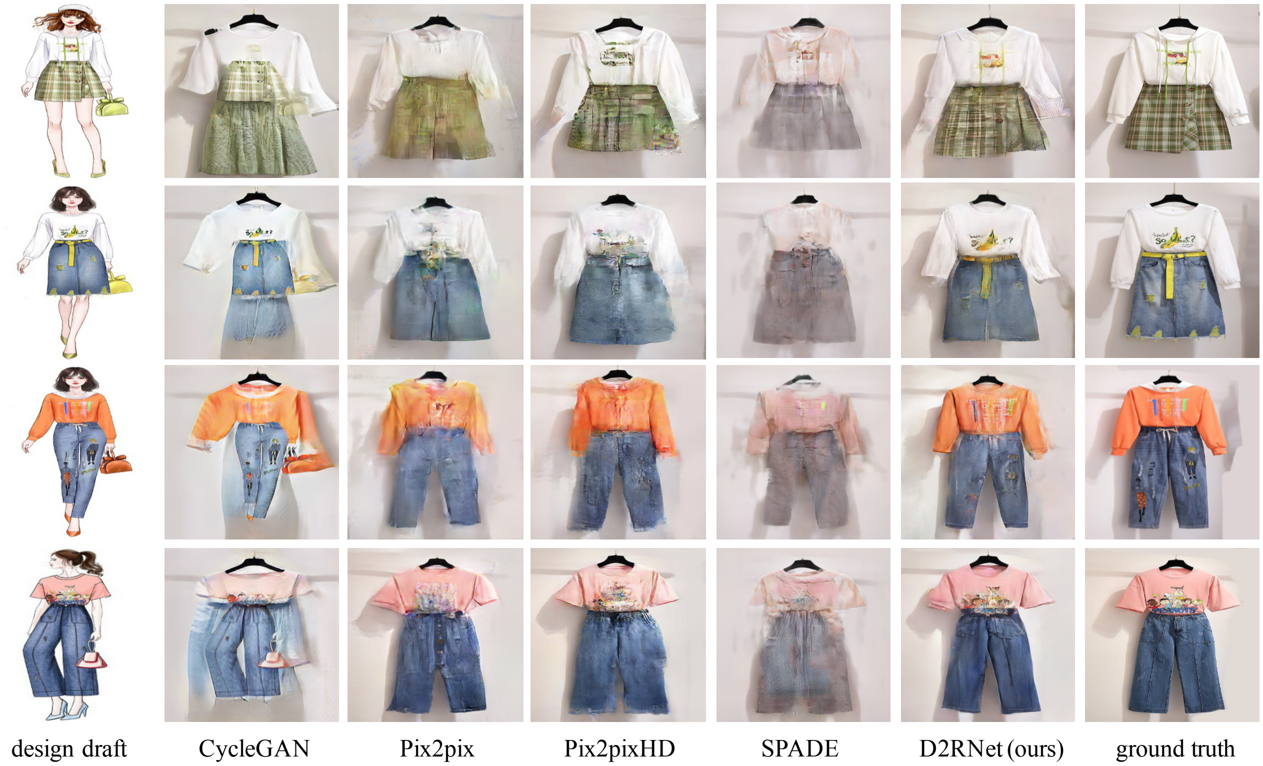

Figure 3. Our D2RNet compared with CycleGAN [7], Pix2pix [2], Pix2pixHD [6] and SPADE [3] (Part III).

R2DNet. Figs. 4–6 show the qualitative comparison on real fashion item to design draft translation with four state-of-the-art image-to-image translation models: CycleGAN [7], Pix2pix [2], StarGAN [1] and Pix2pixSC [5]. As expected, all models fail to accurately render the clothes on the target models, producing poor results. Our model fits the clothes to the pose of the model, producing the most satisfying design drafts.

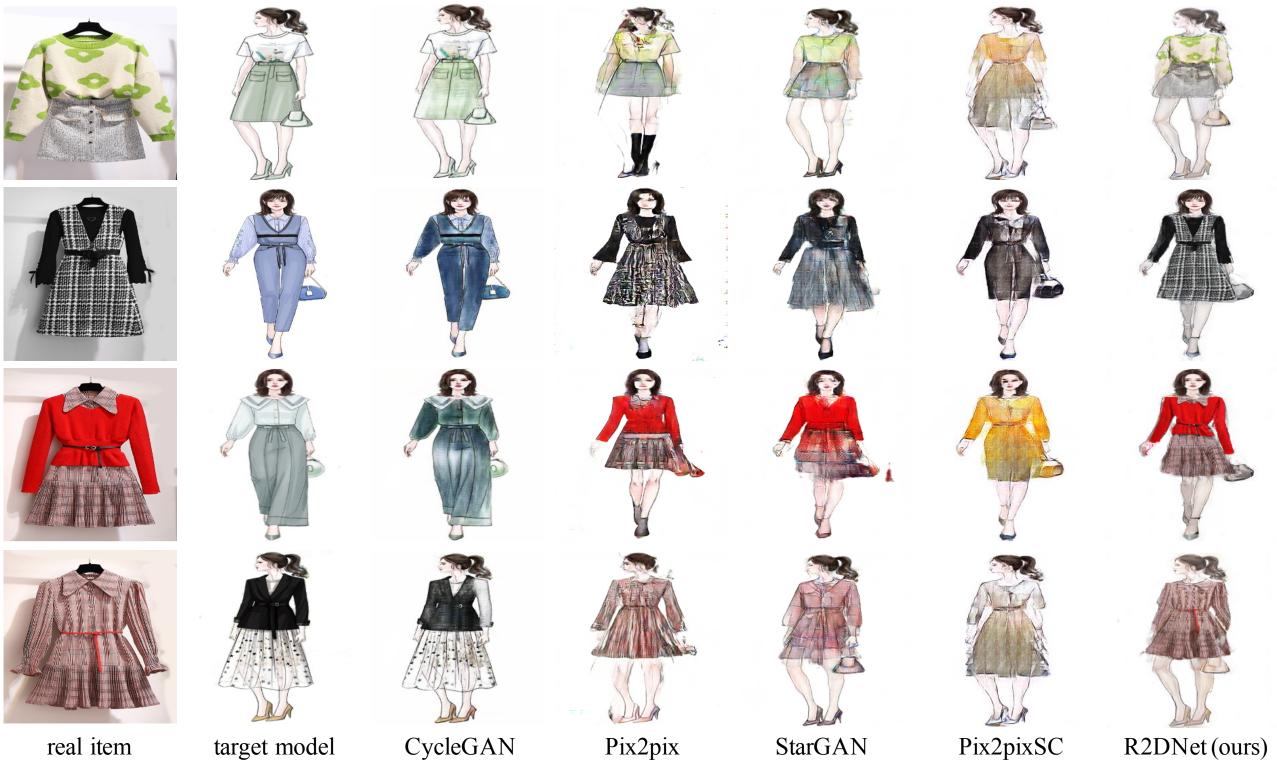

Figure 4. Our R2DNet compared with CycleGAN [7], Pix2pix [2], StarGAN [1] and Pix2pixSC [5] (Part I).

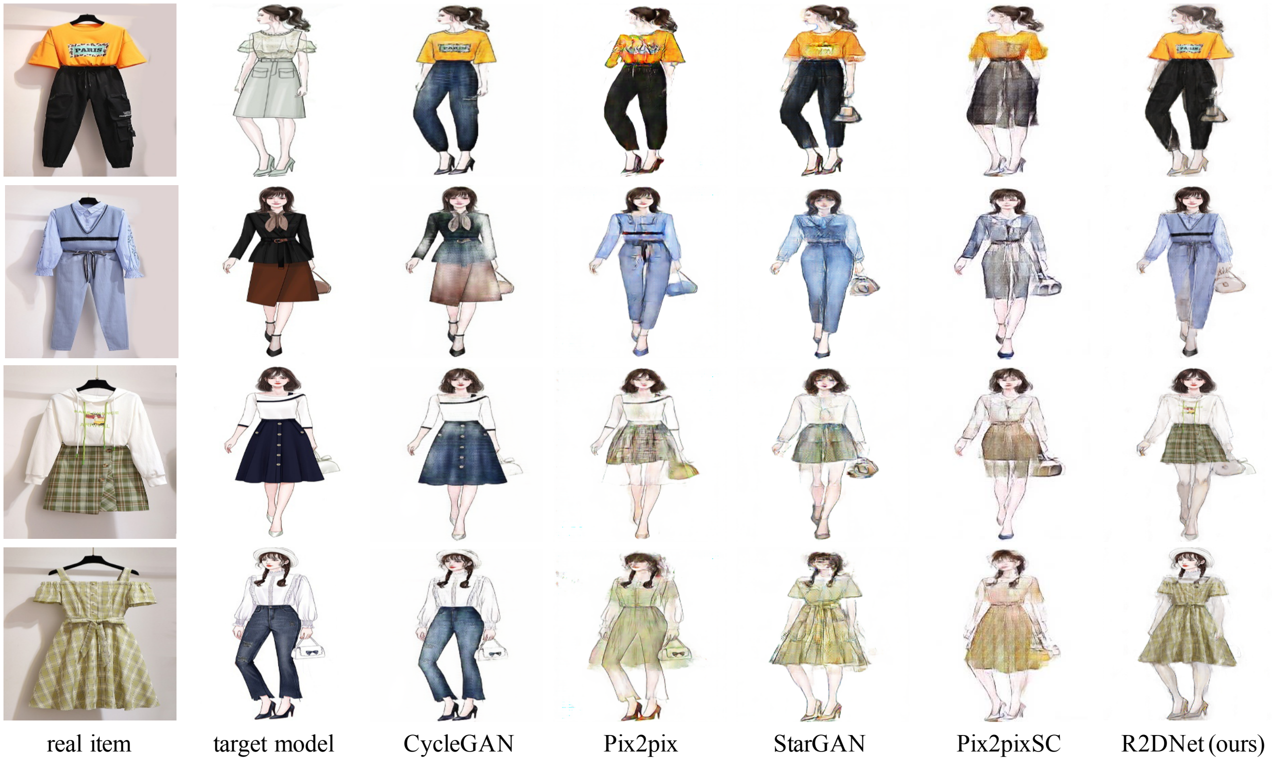

Figure 5. Our R2DNet compared with CycleGAN [7], Pix2pix [2], StarGAN [1] and Pix2pixSC [5] (Part II).

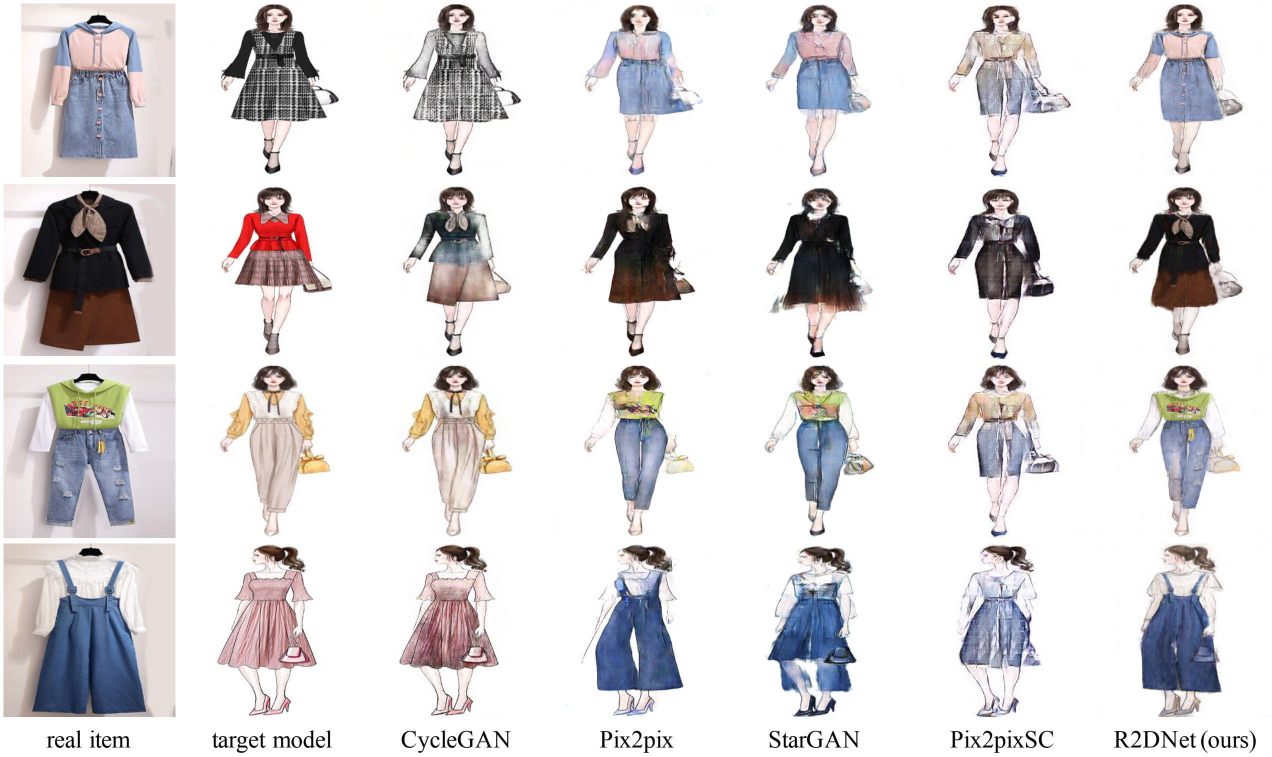

Figure 6. Our R2DNet compared with CycleGAN [7], Pix2pix [2], StarGAN [1] and Pix2pixSC [5] (Part III).

3. Quantitative Evaluation

To better understand the performance of the compared methods, we perform user studies for quantitative evaluations. Participants are shown fashion image translation cases in Figs. 1−6 and Fig. 5-6 in the main paper. For each task, ten groups of results are shown (for design draft to real fashion item translation task, the ground truth real item images are also shown for reference) and users are tasked to assign 1 to 5 scores to five results in each group based on the visual quality. A total of 22 users participated and 2,200 scores were collected. The average preference score is used as the evaluation metrics.

As shown in Table 1, for the task of design draft to real attire translation, the proposed method obtains the best average preference ratio of 4.94, while the average scores of CycleGAN [7], Pix2pix [2], Pix2pixHD [6] and SPADE [3] are 2.07, 3.29, 3.50 and 1.20, respectively. For the task of real attire, as shown in Table 2, the proposed method achieves preference ratios above 4 in all cases, which means our method is steadily preferred by the users. The proposed method obtains the best average preference ratio of 4.69, while the average scores of CycleGAN [7], Pix2pix [2], StarGAN [1] and Pix2pixSC [5] are 1.39, 3.35, 2.06 and 3.51, respectively. This user study quantitatively verifies the superiority of our method.

Table 1. User preference ratio of state-of-the-art methods compared with our D2RNet. The best score in each row is marked in bold.

	Design Draft to Real Attire				
ID	CycleGAN [7]	Pix2pix [2]	Pix2pixHD [6]	SPADE [3]	D2RNet
1	2.50	3.60	3.10	1.00	**4.80**
2	1.80	3.10	3.90	1.30	**4.90**
3	1.80	3.30	3.60	1.30	**5.00**
4	2.00	3.50	3.30	1.20	**5.00**
5	1.20	2.40	2.60	1.80	**5.00**
6	2.70	3.00	3.30	1.00	**5.00**
7	2.00	3.10	3.70	1.20	**5.00**
8	2.70	3.00	3.60	1.00	**4.70**
9	2.40	3.00	3.60	1.00	**5.00**
10	1.70	3.10	3.70	1.60	**5.00**
11	2.30	3.30	3.60	1.00	**4.80**
12	2.60	3.20	3.30	1.00	**4.90**
13	2.10	3.70	3.20	1.10	**4.90**
14	2.20	3.40	3.40	1.00	**5.00**
15	2.00	3.30	3.80	1.00	**4.90**
16	1.50	3.50	3.50	1.50	**5.00**
17	1.70	3.40	3.50	1.40	**5.00**
18	1.20	3.60	3.70	1.50	**5.00**
19	2.30	3.00	3.70	1.30	**5.00**
20	2.10	3.40	3.60	1.00	**4.90**
21	2.70	3.50	2.70	1.20	**4.90**
22	2.00	2.00	3.70	1.00	**5.00**
Average	2.07	3.29	3.50	1.20	**4.94**

Table 2. User preference ratio of state-of-the-art methods compared with our R2DNet. The best score in each row is marked in bold.

	Real Attire to Design Draft				
ID	CycleGAN [7]	Pix2pix [2]	Pix2pixSC [5]	StarGAN [1]	R2DNet
1	1.50	3.50	1.70	3.60	**4.70**
2	1.10	3.40	2.40	3.70	**4.40**
3	1.50	3.10	1.80	3.80	**4.80**
4	1.20	3.50	2.00	3.60	**4.70**
5	1.10	3.50	2.10	3.60	**4.70**
6	1.50	2.60	2.20	3.90	**4.80**
7	1.40	3.40	2.10	3.40	**4.70**
8	1.20	3.20	2.30	3.50	**4.80**
9	1.20	3.70	1.80	3.30	**5.00**
10	1.20	4.20	2.00	3.30	**4.30**
11	1.40	3.10	2.70	3.60	**4.20**
12	1.40	3.60	1.70	3.60	**4.70**
13	1.50	3.50	2.10	3.20	**4.70**
14	2.10	2.70	2.00	3.20	**5.00**
15	1.50	3.20	2.00	3.40	**4.90**
16	1.50	3.40	1.90	3.40	**4.80**
17	1.10	3.70	1.70	3.50	**5.00**
18	1.20	3.90	1.60	3.50	**4.80**
19	1.50	2.80	2.40	3.40	**4.90**
20	1.70	2.90	2.60	3.40	**4.40**
21	1.30	3.50	1.90	3.30	**5.00**
22	1.50	3.00	2.30	3.50	**4.70**
Average	1.39	3.35	2.06	3.51	**4.69**

4. Ablation Study

D2RNet. We examine the effectiveness of our two structure-aware streams with different conditional input, which is the key of our unaligned design draft to real fashion items translation. In Fig. 7-8, we perform a comparison between our full method, our uncompleted method and results using an additional pix2pix model for shape refinement. All ground truths are applied. Our two-stream framework effectively solves the missing or blurring texture details and unadjusted shape problem by combining G_d and G_s. If we use two-step coarse-to-fine networks, as the whole framework goes deeper, the details in the original draft are inevitably and more severely lost.

- **Double D**: A saliency-based sampling layer and a U-Net as the generator, which is trained with both D_d and D_s.
- **w/o Shape**: Results of our detail preservation network G_d.
- **w/o Detail**: Results of our shape generation network G_s.
- **Two Steps**: A G_d followed by an additional pix2pix model for shape refinement. The refinement network uses the distortion image and the output of G_d as input. Meanwhile, its discriminator aims to learn to judge the output of G_d as false. G_d is first trained, and is then fixed to train the subsequent refinement network.
- **Two Steps Plus**: A combination results of the output of two steps and the output of G_d using a fusion network.

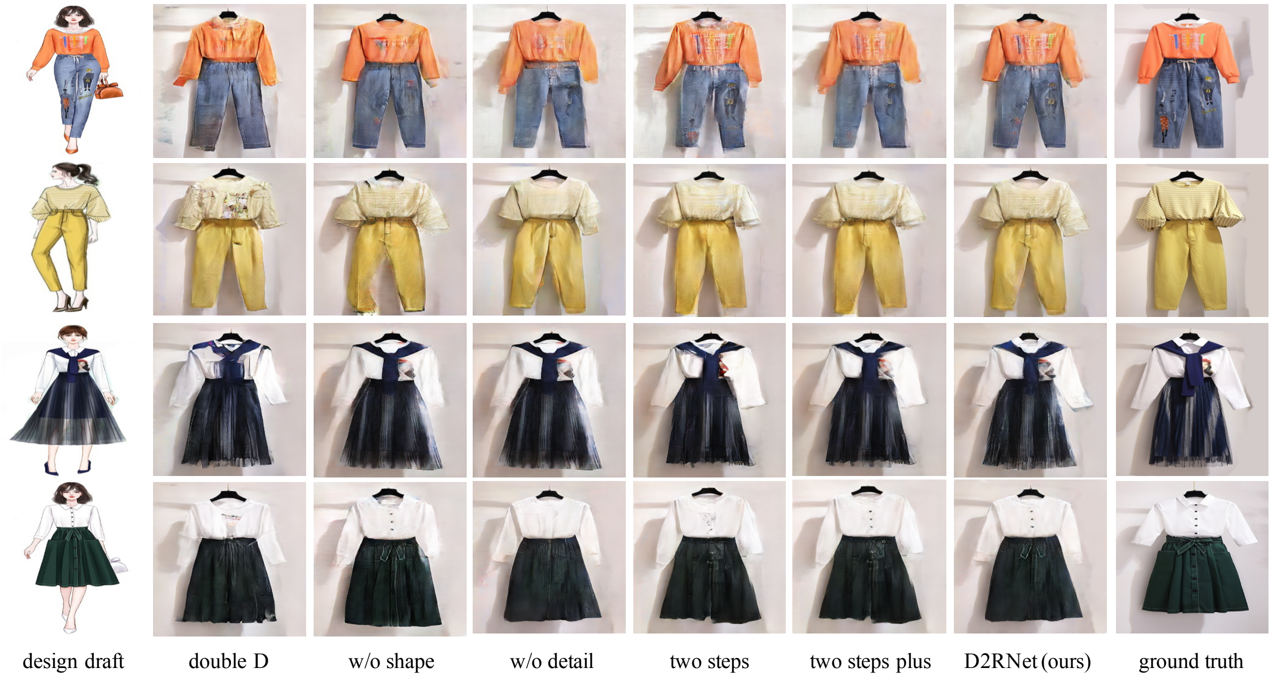

Figure 7. Effect of the dilation-based sketch modelling.

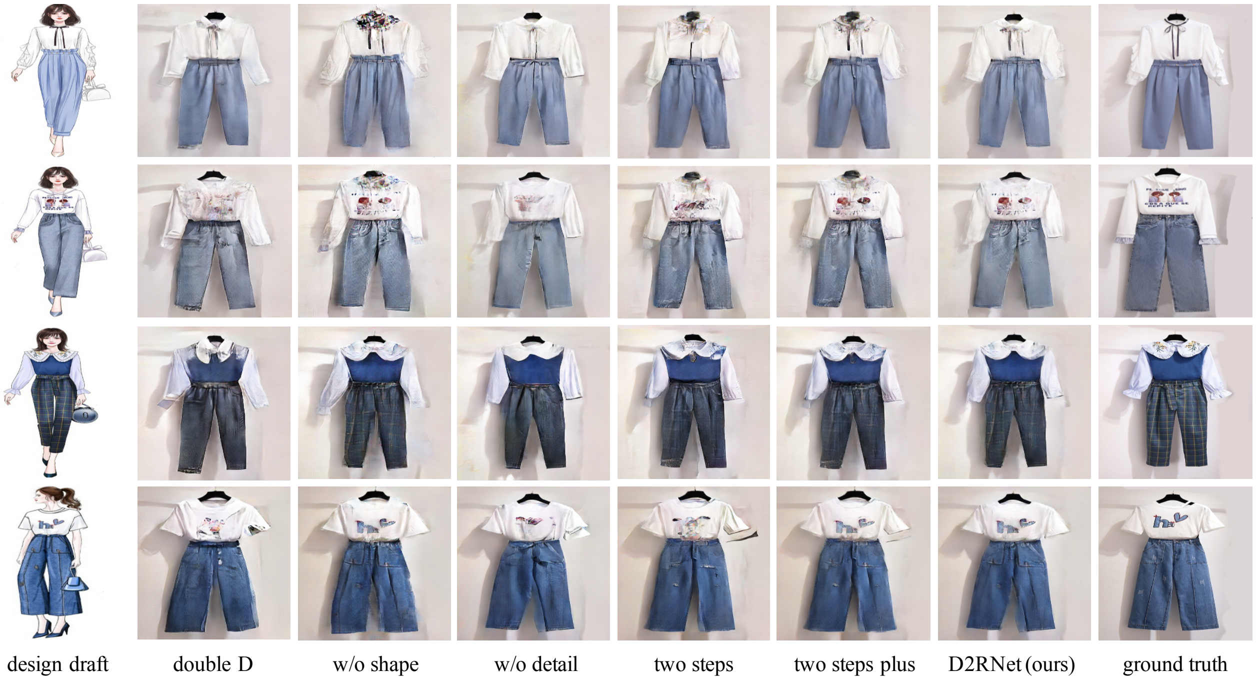

Figure 8. Effect of the dilation-based sketch modelling.

R2DNet. To analyze our R2DNet effectiveness, we design the following configurations:

- **w/o** G_f: Results of our appearance generation network G_a.
- **w/o** D_m: Results of our R2DNet without D_m.
- **w/o** D_a^{label}: Results of our R2DNet without D_a^{label}.

Fig. 9- 10 displays the outputs of these models. In the reverse task, while G_a can warp the clothes to fit the model, its results are rough with plain textures. Generally, when D_m is removed, the network cannot fit the clothes to model. For example, in the second row of Fig. 8, the rendered clothes are too loose. On the other hand, without D_a^{label}, the network fails to infer the covered body in the target model, yielding ghosting artifacts. By comparison, the proposed full R2DNet can well adjust the shape of clothes to fit the target model, and preserve the texture details in the design drafts, showing superior performance.

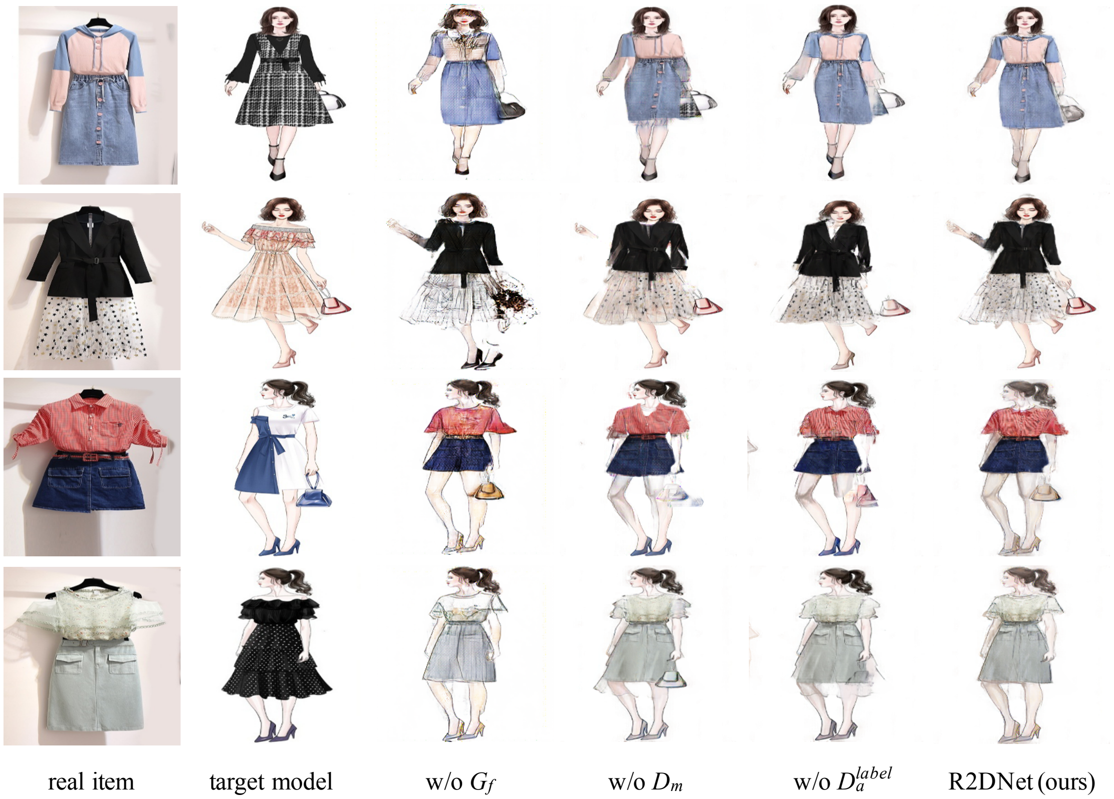

Figure 9. Effect of different training objectives.

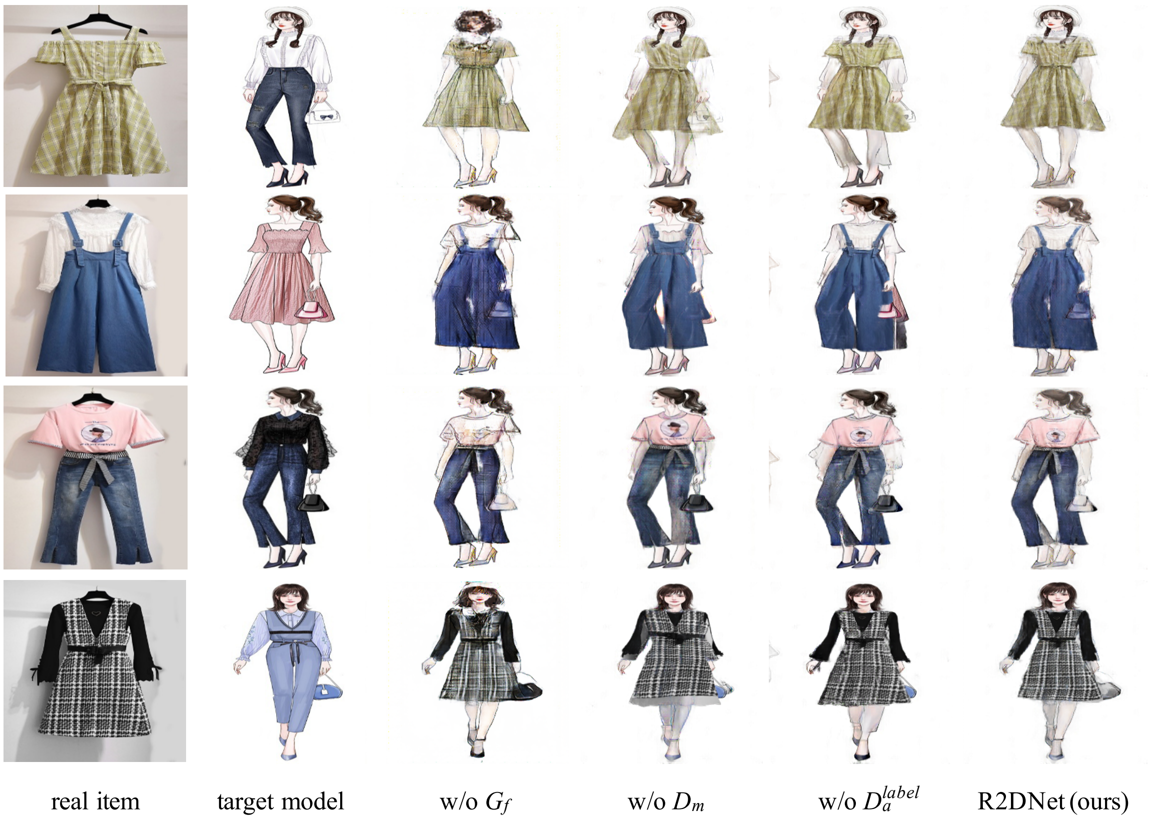

Figure 10. Effect of different training objectives.

References

[1] Yunjey Choi, Minje Choi, Munyoung Kim, Jung Woo Ha, Sunghun Kim, and Jaegul Choo. StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, pages 8789–8797, 2018. 5, 6, 7, 8

[2] Phillip Isola, Jun Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, pages 5967–5976, 2017. 2, 3, 4, 5, 6, 7, 8

[3] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, pages 2337–2346, 2019. 3, 4, 7

[4] Adria Recasens, Petr Kellnhofer, Simon Stent, Wojciech Matusik, and Antonio Torralba. Learning to zoom: a saliency-based sampling layer for neural networks. In *Proc. European Conf. Computer Vision*, pages 51–66, 2018. 2

[5] Miao Wang, Guo-Ye Yang, Ruilong Li, Run-Ze Liang, Song-Hai Zhang, Peter. M Hall, and Shi-Min Hu. Example-guided style-consistent image synthesis from semantic labeling. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, pages 1495–1504, 2019. 5, 6, 7, 8

[6] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, pages 8798–8807, 2018. 3, 4, 7

[7] Jun Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proc. Int'l Conf. Computer Vision*, pages 2242–2251, 2017. 3, 4, 5, 6, 7, 8