Virtual Fitting Room

Junwen Bu (junwenbu), Jie Chen (jiechen8), Zhiling Huang (zhiling)

CS231n Convolutional Neural Networks for Visual Recognition, Stanford University

Introduction

Virtual Fitting Room is a challenging task yet useful feature for e-commerce platforms and fashion designers.

The goal of our project is to provide a virtual try on experience, where the user can see how he/she will look wearing different pants, skirt, dress, etc.

Input

- A fashion model portrait image
- A texture image
- A fashion item type

Output

- A transformed image, where the selected fashion item is changed to the new style.

Method and Model

High-level Architecture

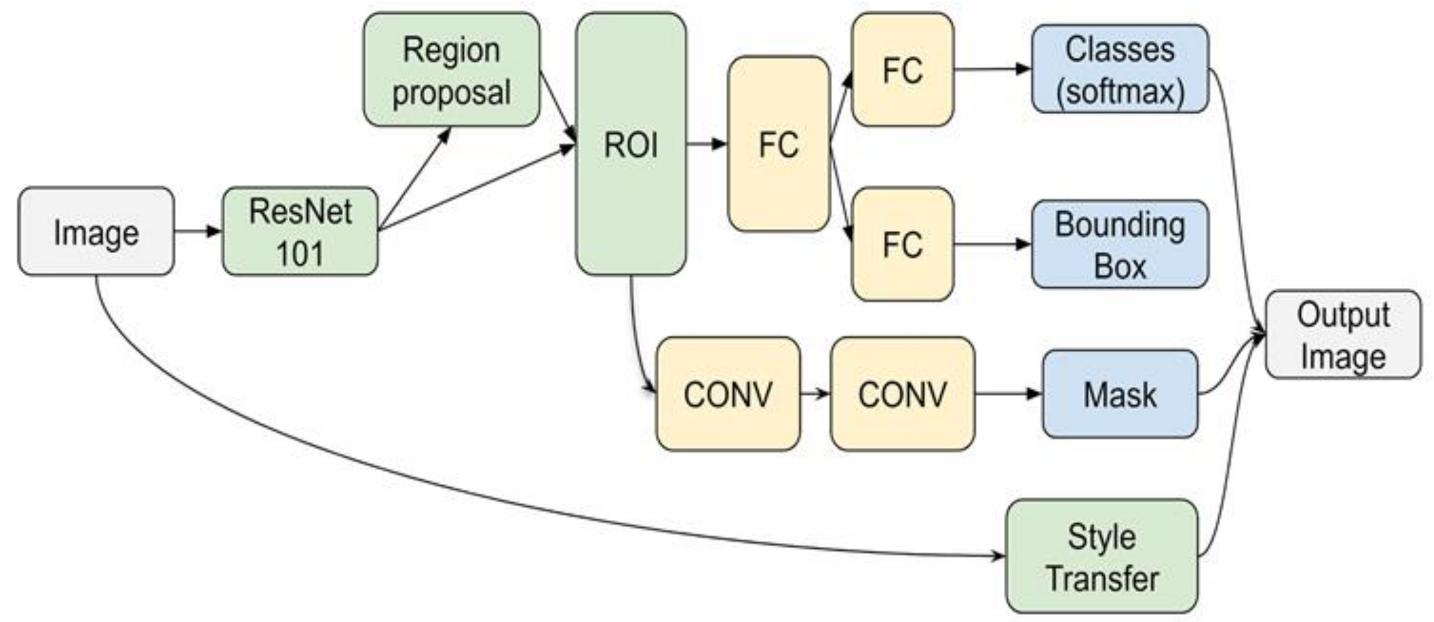

Figure 1: The method we proposed has two main components: 1) ***Mask R-CNN [1]***; 2) ***Neural Style Transfer [2]***.

Firstly we used Mask R-CNN to find the regions of different fashion items, and secondly used Neural Style Transfer to change the style of the selected fashion items.

Segmentation (Mask R-CNN)

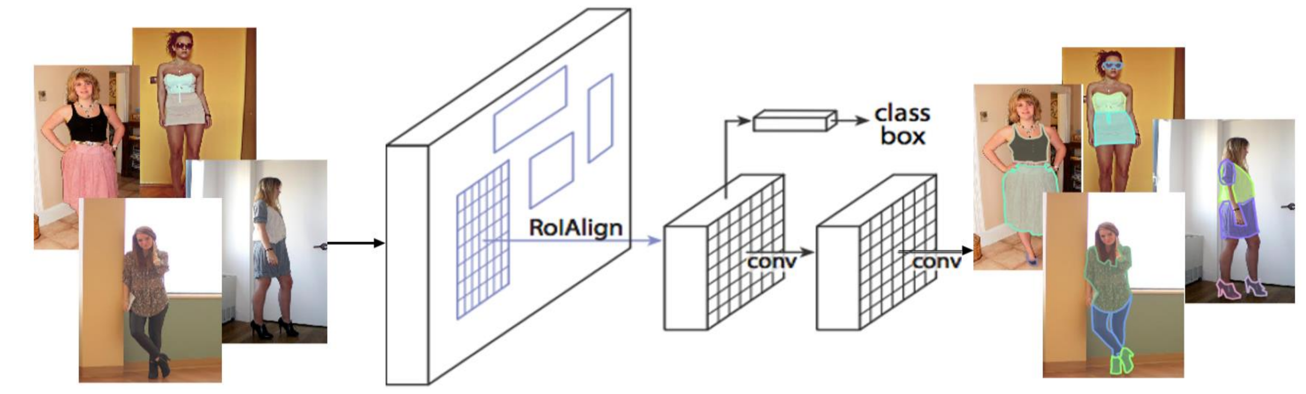

Figure 2: Mask R-CNN Architecture

Loss: Regional Proposal Network head loss + Mask head loss

- **Regional Proposal** Network head loss:

$$L(\{p_i\},\{t_i\}) = \frac{1}{N_{cls}}\sum_i L_{cls}(p_i, p_i^*) + \lambda \frac{1}{N_{reg}}\sum_i p_i^* L_{reg}(t_i, t_i^*)$$

p_i is the predicted classification; t_i is the predicted RoI.Mask head loss: classification and box regression loss will be same as in RPN head loss.

- **Mask head loss:** Classes + bounding Box + Mask loss *(loss of blue components in Figure 1)*

Style Transfer

Input: fashion images (content) + textures (style)

Output: fashion images content with artistic style of given texture.

We implemented the style transfer by performing gradient descent on the pixel values of our original image. The loss function is a weighted sum of three terms: **content loss**, **style loss** and **total variation loss**. *For **notations**, please refer report [5] Section 3.2.*

$$L_c = w_c \times \sum (F_{ij}^{\ell} - P_{ij}^{\ell})^2$$

$$L_S = \sum_{\ell \in \mathcal{L}} w_\ell \left(\sum_{i,j} (G_{ij}^{\ell} - A_{ij}^{\ell})^2\right)$$

$$L_{tv} = w_t \times \Big(\sum_{c=1}^{3}\sum_{i=1}^{H-1}\sum_{j=1}^{W}(x_{i+1,j,c} - x_{i,j,c})^2$$

$$+ \sum_{c=1}^{3}\sum_{i=1}^{H}\sum_{j=1}^{W-1}(x_{i,j+1,c} - x_{i,j,c})^2\Big)$$

Dataset

Source and format

- Raw image data from PaperDoll [3].
- Annotations from ModaNet [4].
- Labels are formatted in COCO style.

Training/Validation/Test

- Training: 20k, Validation: 2k, Test: 1k.

Preprocess

- Resized to 256 x 256
- Handle Grayscale images
- Channel-level normalization
- Data augmentation: horizontally flipping

Training

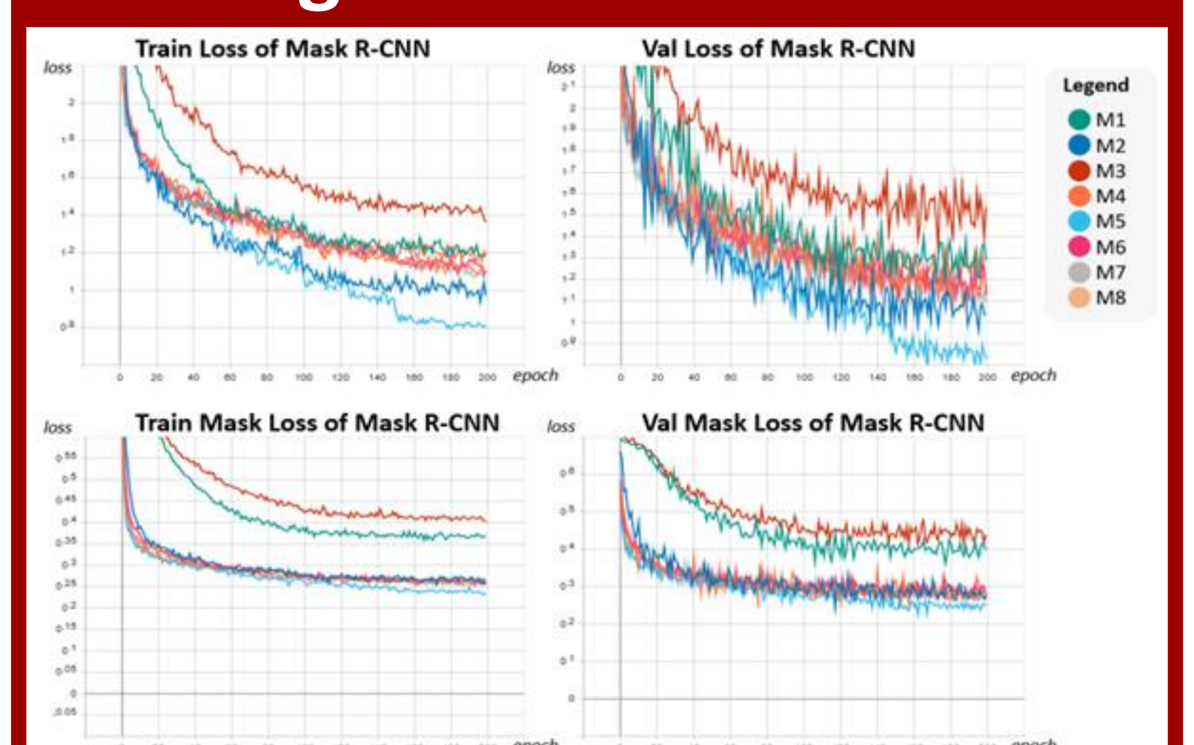

Figure 3 Loss Curve: Trained 8 models. Validation total and mask losses keep decreasing. The gap between training and validation loss is small.

Results

Model	Preloaded	Epoch 1-50		Epoch 51-100		Epoch 101-150		Epoch 151-200		mAP(%)
		Layers1	LR1	Layers2	LR2	Layers3	LR3	Layers4	LR4	
M1	ImageNet	All	5e-4	All	5e-4	All	5e-5	All	5e-5	41.62
M2	COCO	All	5e-4	All	5e-4	All	5e-5	All	5e-5	56.75
M3	ImageNet	Heads	5e-4	Heads	5e-4	All	5e-5	All	5e-5	48.61
M4	COCO	Heads	1e-3	Heads	1e-3	All	1e-4	All	1e-4	60.28
M5	COCO	Heads	1e-3	C4, C5, Heads	1e-3	C4, C5, Heads	1e-3	All	1e-4	**68.72**
M6	COCO	Heads	1e-3	Heads	1e-3	C5, Heads	5e-4	C5, Heads	2e-4	58.61
M7	COCO	Heads	1e-3	Heads	1e-3	C5, Heads	1e-4	All	1e-4	64.78
M8	COCO	Heads	1e-3	Heads	1e-3	Heads	1e-4	Heads	1e-4	50.09
FCN-CRF	-	-	-	-	-	-	-	-	-	**66.70**
PaperDoll	-	-	-	-	-	-	-	-	-	**33.34**

Table 1 Our Models and Baseline: **Layers** describes the trained layers in each step of the training process. **LRs** represents learning rate. **Heads**: Mask R-CNN, Regional proposal and Feature Pyramid Network heads. **C4, C5**: the 4th and 5th component in RestNet-101-FPN. **All**: all layers in our network.

Figure 5 Final Results of our models: 1st row, dress to leather; 2nd row outer to jeans, bag to muse.

Figure 4 Neural Style Transfer Results: We used jeans, leather, cloud, composition and muse as style input. See small boxes on the top left corner of each image.

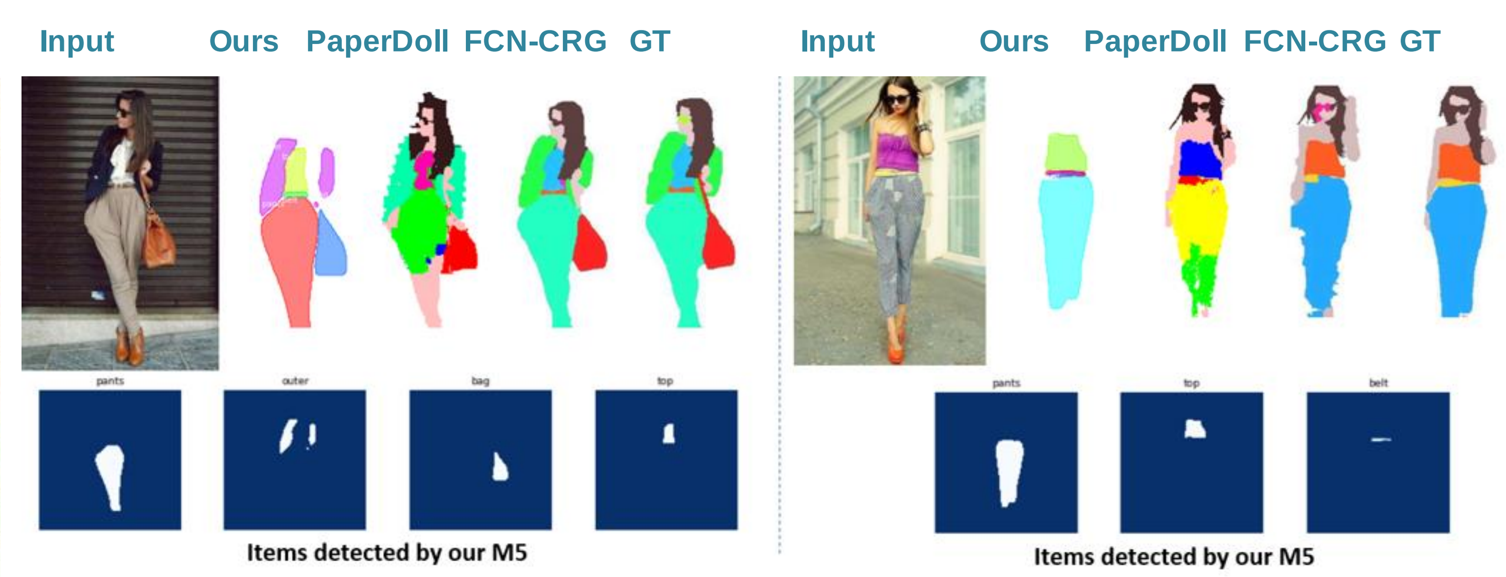

Figure 6(above) Comparison of segmentation between our model and two Baselines. **GT** represents *Ground Truth*.

Figure 7(left) Comparison between our models and others: Our results reserve more texture information.

Figure 8(above): Failed to identify left part of outer.

Figure 9(left): In the 1st case, model with uncommon pose holds outer in her hand. In the 2nd case, the dress looks like outer + skirt. In the 3rd case, the scarf looks like an outer.

Acknowledgements: CS231n Teaching Staff

Conclusions

1. We addressed the problem of virtual try on in two steps: Mask R-CNN and Neural Style Transfer.
2. Used Mask R-CNN to find the regions of different fashion items.
3. Used Neural Style Transfer to change the style of the selected fashion items.
4. Our model outperforms baseline both qualitatively and quantitatively (mAP 68.72%).

Future Work

- Increase the number of types of detectable fashion items and transferable textures.
- Introduce color loss to neural style transfer so that the color of the transformed item is closer to the target texture.

Reference

[1] K. He et al., Mask R-CNN. CoRR, abs/1703.06870, 2017.
[2] L. Gatys et al., A neural algorithm of artistic style. CoRR, abs/1508.06576, 2015.
[3] K.Yamaguchi, Paperdoll GitHub Repository https://github.com/kyamagu/paperdoll
[4] S. Zheng et al., Modanet: A large-scale street fashion dataset with polygon annotations. CoRR, abs/1807.01394, 2018.
[5] Virtual Fitting Room CS231n Project Final Report.