

# ShineOn: Illuminating Design Choices for Practical Video-based Virtual Clothing Try-on

Gaurav Kuppa<sup>1</sup>  
San Jose State University  
gaurav.kuppa@sjsu.edu

Andrew Jong<sup>1</sup>  
San Jose State University  
Carnegie Mellon University  
ajong@andrew.cmu.edu

Xin Liu  
Nanyang Technological University  
veralau9425@gmail.com

Ziwei Liu  
Nanyang Technological University  
ziwei.liu@ntu.edu.sg

Teng-Sheng Moh  
San Jose State University  
teng.moh@sjsu.edu

## Abstract

*Virtual try-on has garnered interest as a neural rendering benchmark task to evaluate complex object transfer and scene composition. Recent works in virtual clothing try-on feature a plethora of possible architectural and data representation choices. However, they present little clarity on quantifying the isolated visual effect of each choice, nor do they specify the hyperparameter details that are key to experimental reproduction. Our work, ShineOn, approaches the try-on task from a bottom-up approach and aims to shine light on the visual and quantitative effects of each experiment. We build a series of scientific experiments to isolate effective design choices in video synthesis for virtual clothing try-on. Specifically, we investigate the effect of different pose annotations, self-attention layer placement, and activation functions on the quantitative and qualitative performance of video virtual try-on. We find that DensePose annotations not only enhance face details but also decrease memory usage and training time. Next, we find that attention layers improve face and neck quality. Finally, we show that GELU and ReLU activation functions are the most effective in our experiments despite the appeal of newer activations such as Swish and Sine. We will release a well-organized code base, hyperparameters, and model checkpoints to support the reproducibility of our results. We expect our extensive experiments and code to greatly inform future design choices in video virtual try-on. Our code may be accessed at <https://github.com/andrewjong/ShineOn-Virtual-Tryon>.*

<sup>1</sup>These authors equally contributed to this paper.

## 1. Introduction

The age of the internet has led to an unprecedented consumer shift to online marketplaces. With the convenience and wide selection provided by online stores, shopping is easier than ever. However, purchasing clothes online remains a difficult proposition because end-consumers cannot accurately judge clothing fit and appearance on themselves. Constant returns of ill-fitted, unflattering clothing negatively impact the environment through wasted shipping and packaging. Fortunately, this problem has the potential to be alleviated with emerging virtual try-on technology. Virtual try-on aims to let users quickly visualize themselves in different outfits through a camera and digital display. Being able to achieve a high-quality, real-time fashion virtual try-on system may boost online retail sales, as well as cut down on the carbon footprint produced by packaging and returns.

Beyond the product application, the academic neural rendering community also has a vested interest in virtual try-on, as it serves as a benchmark task to evaluate complex object transfer and scene composition. This complexity is embodied in the criteria to accurately maintain the user's identity, render the cloth product at the appropriate location, follow the user's body proportions, preserve cloth texture detail, exhibit smooth temporal dynamics, establish temporal consistency, and blend well with the scene's lighting.

Many virtual try-on works that emerged in the past three years explore deep learning methods. These mostly focus on image try-on [33, 36, 5, 39, 25], and only recently has there been an exploration of video-based try-on [17, 6].

There has been substantial work towards virtual try-on, including image-image translation, perceptual loss, and the improvement of human parsing techniques. Image-image translation [14] serves to pivotal in transferring a cloth to a

target image. Common losses like perceptual loss [7, 16] are highly effective in retaining quality transfer. The proliferation of fashion datasets [20, 24, 8, 9] and human parsing techniques such as cloth segmentation, body segmentation [24, 8], pose annotation [10], and more have increasingly improved in quality. These techniques are critical for robust virtual try-on.

We want to develop a deep, scientific understanding of video virtual try-on and display transparency in our methods to enable more great work in the field of neural rendering, and specifically virtual try-on. We aim to shed light on the workings of virtual try-on and lead to meaningful insights on how to improve it. In particular, we test the effectiveness of DensePose pose annotations, self-attention, activation functions, and optical flow. We accumulate the results of our studies and compare our resulting ShineOn approach with existing try-on methods.

The specific contributions of this paper are (1) we present transparent, comprehensive bottom-up experiments testing pose annotations, self-attention, activation functions, and optical flow, (2) demonstrate a decrease in memory usage, training time, and improved face detail transfer by using DensePose pose annotations rather than CocoPose pose annotations, and (3) propose *ShineOn*, the accumulation of our most effective methods for practical video virtual try-on. At the time of writing, ShineOn is the only public video virtual try-on codebase readily available for scientific verification and reproducibility, which is critical for the future of try-on works.

## 2. Related Works

### 2.1. Image Virtual Try-on

There has been a large proliferation of publicly available datasets [20, 24, 8, 9] that have pushed forward the multiple-component tasks of try-on mechanics. These datasets are of great importance to advanced tasks, such as human parsing, cloth segmentation, body segmentation, pose estimation. These annotations led to more robust human parsing models, such as SSL and JPPNet [9, 22].

Quite a few proposed network architectures have been developed to accomplish body and cloth segmentation [24, 8]. Pose information has been embedded through CocoPose, and more recently through DensePose [10].

Realistic virtual try-on requires transferring of high-level structures such as cloth pattern, design, text, and texture. In addition to L1 loss, perceptual and feature losses [7, 16, 43, 4] consist of important components to address this issue.

The history of virtual try-on methods and deep learning dates back to 2017. These methods are dependent on effective and generalizable human parsing methods. From Jetchev and Bergmann’s Conditional Analogy GAN [15], virtual try-on has seen significant growth in the years

[11, 33, 1, 23, 36, 5]. These methods were iteratively improved by introducing perceptual loss, removing adversarial loss, and experiment with different network architectures to synthesize more detailed virtual try-on outputs. The general architecture for these methods is a two-stage approach that involves cloth warping and person rendering. Some virtual try-on networks employ more than two stages, which we found unnecessary. A two-stage approach has enough expressiveness for the model to learn how to do virtual try-on.

### 2.2. Video Virtual Try-on

Given the significant development for virtual try-on with images, the next natural step is to investigate virtual try-on for video. Video try-on would allow a user to easily examine the clothing’s appearance on their body at multiple angles, instead of processing individual images. However, video try-on raises new challenges, such as how to handle temporal consistency [21, 42] between video frames.

FW-GAN [6] requires a video of a reference person wearing the desired clothes. This was collected in a new dataset, VVT [6], which was obtained from scraping fashion walk videos on fashion websites. The image of the user is then warped to match the pose of the reference, and the desired clothes from the reference are composited with the warped user.

### 2.3. Attention and Activations Functions

Attention and transformers [38, 40, 32] have shown to work plenty in achieving world-class performance in the machine translation task. The notion of self-attention demonstrated the ability to model long-range dependencies in an interpretable way. Image transformers [26] and self-attention for convolution neural networks and generative adversarial networks [3, 2] began to popularize. It gave way to self-attention in image-image translation. However, self-attention has not been used for the virtual try-on task yet, our work will verify its power to model the longer-range dependencies when transferring the garment to the model person.

ReLU networks were suggested to have a bias towards learning low-frequency information [27]. Similarly, literature [30, 37, 30, 31] show that smoother activation functions are more effective at achieving robust results for representing and reconstructing media.

## 3. Problem Formulation and Challenges

### 3.1. Problem Formulation

**Try-on Task for Inference.** Given a desired cloth product image  $c$ , a video of a user with  $n$  frames  $V(v_1, \dots, v_n)$ , video annotation type  $a_t \in A$ , and each annotation type generated for every frame  $A(a_{t_1}, \dots, a_{t_n})$ , we develop a model to synthesize a new video  $V'$ , in which the user from

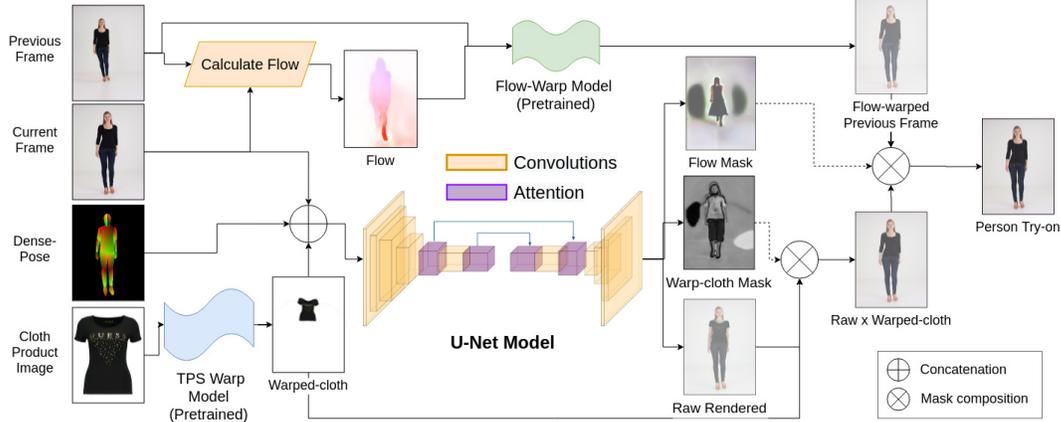


Figure 1: **Visualization of ShineOn Architecture.** The input person representation  $p$  and warped cloth  $w$  are fed into the U-Net model with self-attention. The output of the U-Net, after masking, is the person and warped cloth composed together. For experiments with Flow, we add the top branch to finally compose with a flow-warped previous frame.

video  $V$  is realistically wearing cloth  $c$ . See Figure 2 for details.

It is important to note that our formulation is fundamentally different that that posed in our closest related work, *FW-GAN* [6]. *FW-GAN* proposes to input only a single user image  $u$  instead of inputting a video of the user. The output  $V'$  then synthesizes the user to follow an arbitrary keypoint pose sequence that is parsed from a separate video. See Figure 2 for this comparison.

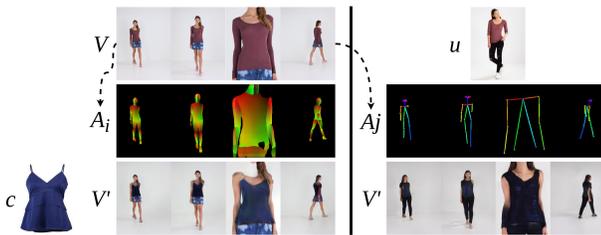


Figure 2: Problem formulation of *ShineOn* (ours, left) compared to *FW-GAN*'s [6] (right). Dashed arrows show that the annotations  $A_t$  are parsed from the source frames in  $V$ . The key difference is that our formulation directly treats the video as the user input, whereas [6] uses a separate still user image and reposes the user to the video's key points. This means the compared results feature different users (Left: blonde hair, Caucasian, and blue shorts. Right: black hair, Asian, and black jeans).

We argue that our formulation is more practical from an applications standpoint. In application, a user may want to see the clothes at different angles on their body and move their bodies as they wish to instantaneously change the view. When applied in real-time, this level of control corresponds to looking in a real-life mirror. We adopt this in our

approach because it permits direct user control, rather than reposing a still user image to arbitrary keypoints. This formulation involving direct control lets us assume that video frames of the user are available.

In either formulation, the synthesized video  $V'$  aims to meet the criteria described in Section 1.

**Reconstruction Task for Training.** It is challenging to collect large amounts of data with ground truth try-on pairs. Such ground truth data would require the same person to hold the exact same pose across two or more different outfits. This is already a challenge for a single person, much less a whole dataset, as people naturally to try on different clothes. Therefore, training for virtual try-on is often framed as a self-supervised task.

We follow the same self-supervised reconstruction task as our base experiment *CP-VTON* [33]. In the reconstruction task, the target garment that the user is wearing (e.g. shirt) in the training video is masked out. We then train the model to synthesize the missing garment in each frame given its corresponding isolated cloth product image  $c$ . This way we may calculate a loss between the synthesized and ground truth train sample. We use several common losses,  $\mathcal{L}_{L1}$ ,  $\mathcal{L}_{mask}$ , and  $\mathcal{L}_{VGG}$ , for the reconstruction task. As stated by Pix2Pix [14],  $\mathcal{L}_{L1}$  is essential to achieve image translation, and is the reason that many modern virtual try-on methods use these core components. Additionally, we utilize  $\mathcal{L}_{mask}$  to retain characteristic details of cloth and alignment with the target person's body shape. Lastly, realistic virtual try-on requires transferring of high-level structures such as cloth pattern, design, text, and texture;  $\mathcal{L}_{VGG}$  is a vital component to address this issue.

### 3.2. Dataset

We use the *Video Virtual Try-on* (VVT) dataset supplied by [6] for our video virtual try on task. The VVT dataset contains 791 videos recorded at 30 frames-per-second at  $192 \times 256$  resolution, and each video has a duration between eight and ten seconds. This results in roughly 250-300 frames per video. The train and test set contain 159,170 and 30,931 frames respectively (equivalent to 84% and 16% of the total). Each video has a corresponding isolated cloth product image. The VVT dataset only contains product images of upper clothes (shirts, blouses, sweaters, etc).

Because the final evaluation task of try-on is distinct from the reconstruction training task, we choose to treat the given test set as validation for each of our successive experiments. We then examine try-on as the true test in our final evaluation in Section 5 after all design choices are finalized.

### 3.3. Architecture

Our experiments start from the sequential two-module architecture proposed by our baseline [33]. The first module, *Warp*, warps the cloth product image  $c$  to the body shape and pose of the user at each video frame in  $V$ . The second module, *Try-on Composition*, produces both a raw synthesized try-on image and a mask. It uses the mask to compose the raw try-on with the previously warped cloth to produce the final try-on result. We refer the reader to [33] for details of the base architecture, and to Figure 1 for details of how we modify the architecture in our experiments.

Following [6], we use a pre-trained Warp module. We focus on improving the Try-on module as we observed it was responsible for textural artifacts in preliminary experiments.

U-Net [29] adds skip connections between corresponding encoder and decoder layers by concatenating their respective outputs. More generally, U-Net has shown to render smoothly synthesized images. This makes the U-Net architecture a good fit for image and video virtual try-on.

The U-Net model outputs a rendered person  $p$ , and composition mask  $m$ . The person try-on,  $\hat{p}$ , is obtained as shown below:

$$\hat{p} = w * m + p * (1 - m) \tag{1}$$

where  $w$  is the warped cloth.

### 3.4. Challenges

The U-Net architecture is effective in doing image virtual try-on. It is able to synthesize a cloth onto a single person. There are significant issues with generalizing image virtual try-on to video virtual try-on. We detail several challenges with these methods, that we address with our experiments.

The purpose of the person-representation is to feed in information to the neural network to learn the most effective

way to render a person’s image with the warped cloth. Current methods of person representation involve using a large tensor of images including 18-keypoint pose-annotation, body shape image, and reserved regions annotations. These methods depend on this large person representation to aid the network’s learning.

The most significant issue of video virtual try-on is the lack of high-quality cloth transfer while retaining strong texture features of the cloth. Methods can work on trained datasets, but fail to generalize at test time. At test time, there is a significant lack of detail. Virtual try-on suffers from flickering details and cloth texture distortion.

## 4. Experiments

We detail several experiments that we created, to judge the efficacy and impact on video virtual try-on. The nature of our experiments is that we test one experimental variable, while not changing other variables. At each experiment, we determine the result of each experiment and recommend the most effective variable as a part of our design choices in ShineOn.

### 4.1. Training Setup

In all experiments, we train the U-Net model to 10 epochs with an accumulated batch size of 64. We use the Adam optimizer [18] with an initial learning rate of 0.0001 that decays linearly after 5 epochs. Additionally, we utilize 16-bit precision training to increase training speed and reduce GPU memory consumption.

### 4.2. DensePose

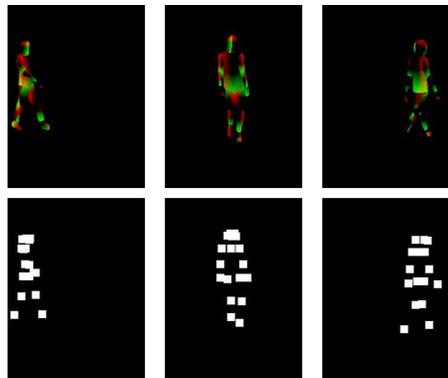


Figure 3: **Visual Comparison between DensePose and CocoPose** DensePose annotations (top row) contain dense 3D body information in the form of UV coordinates, whereas CocoPose annotations (bottom row) are sparse and limited to 2D keypoints.

DensePose has been shown to model 3D body information in the UV field, and it is an effective representation

of body shape, body parts, and includes additional information. We found it beneficial to replace the 18-keypoint CocoPose pose coordinates with the DensePose pose annotation, which show to be just as accurate for pose information. The visual comparison between DensePose and CocoPose, as shown in Figure 3, illustrates the increased detail of information compared to the larger CocoPose pose annotation. The benefit of DensePose is that it decreases the size of pose representation by 6 times. Since data loading is the bottleneck of the training pipeline, this decrease significantly decreases the training time of virtual try-on networks.

The 18-keypoint pose representation is a sparse representation of pose information and severely slows down training. We show a 6 times decrease in size of pose-annotation by using DensePose [10] to represent pose-information. Additionally, we also show that DensePose retains far more face details than using CocoPose as a pose annotation. Figure 6a provides clear quantitative analysis of the effects of using DensePose annotations.

### 4.3. Self-Attention

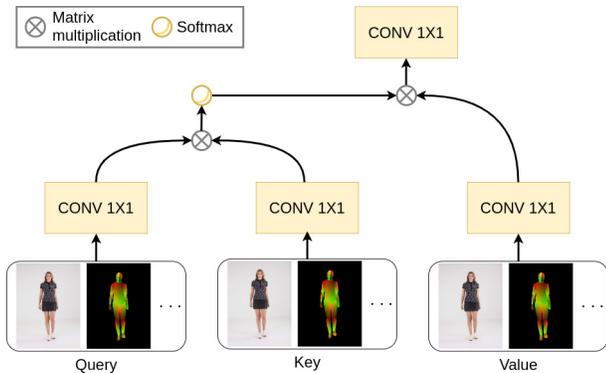


Figure 4: **Self-Attention Layer for Video Virtual Try-on** Self-attention layers is implemented by feeding in the input representation  $i$  or a convolved feature map derived from  $i$ . Self-attention layers have been shown to model long-range dependencies, and attend to important spatial regions [41].

Self-attention has been repeatedly used to model long-range dependencies, express interpretable results, and more recently, attend to visual and spatial regions of importance. There have been object detection, image generation, and many more computer vision tasks that use self-attention to achieve state-of-the-art performance, respectively.

Within our UNet model, we utilize self-attention layers within our encoder-decoder architecture. The self-attention layers are strategically placed where the input feature map depths are greatest so that the attention module has the capacity to learn the most important features to attend to. Self-attention is computed as:

$$f(q, k, v) = softmax(qk)' * v \quad (2)$$

where  $q$ ,  $k$  and  $v$  are query, key and value, respectively.

Figure 1 shows our application of self-attention to virtual try-on. We feed in a person representation  $p$ , or a downsampled representation of  $p$ .

As shown in Figure 6b, self-attention does not enable the network to retain more detail of the transferred cloth. There is an improvement of stronger features of and neck details. As opposed to some recommendations [41], we use self-attention where the feature map depth is greatest. This design choice retains similar quality for virtual try-on and reduces the size of the model.

### 4.4. Frequency Robustness

Recent literature suggests that ReLU networks are ineffective at rendering and recovering images with high-frequency information. In the garment transfer and virtual try-on domain, preserving high-frequency information corresponds to preserving clothing details and texture information. SIREN and Fourier Features [31, 30] use sine annotations and activation functions, respectively, to retain high-frequency information. Smooth Adversarial Learning [37] shows that smooth gradients for activation functions will lead to more robust models. We hypothesize that smoother activation functions, as shown in Figure 5, will lead to more high-frequency information being transferred to the target person.

In this paper, we test several activation functions, ReLU, GELU, Swish, and Sine, [28, 19, 12, 31], to see the effect on virtual try-on video synthesis. Through our experiments, we determine that ReLU, GELU, and Swish perform similarly on quantitative metrics as per Figure 6c. However,

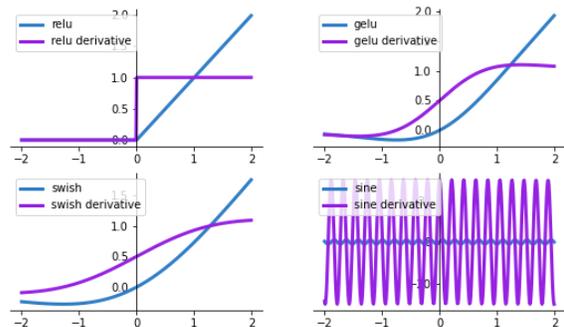


Figure 5: **Activation Functions and Their Derivatives** From left to right and top to bottom, the activation functions are ReLU, GELU, Swish, and Sine. These activation functions and their derivatives visualize the discontinuity of the ReLU function’s gradient, smooth nature of GELU and Swish activation function and their gradients, and erratic nature of Sine function’s gradient.

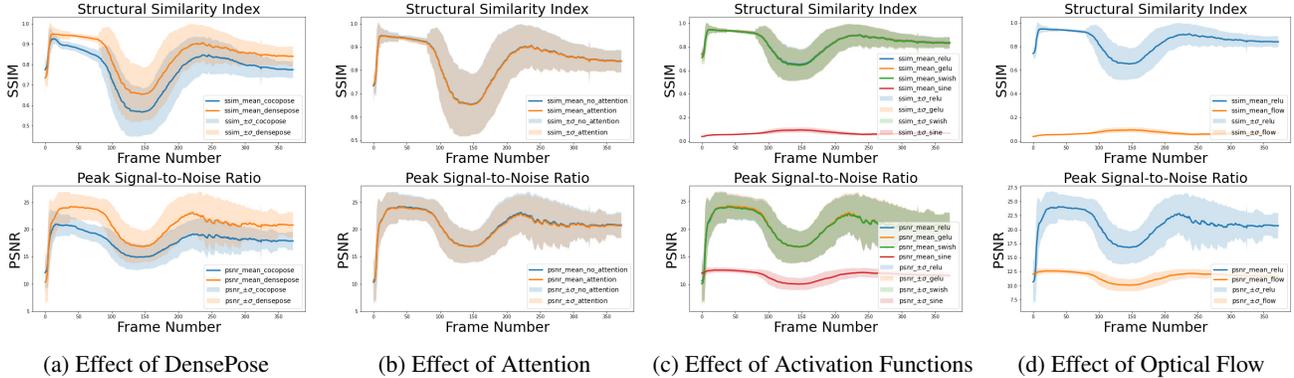


Figure 6: **Quantitative Comparisons of DensePose, Self-Attention, Activation Functions, and Optical Flow.** We show the mean and standard deviation of metrics across all videos in the VVT dataset for the reconstruction task. We observe that DensePose significantly improves performance over CocoPose. Self-Attention and GELU activation function improve performance over their alternatives. Lastly, Optical Flow causes video virtual try-on to be worse.

qualitative inspection reveals that GELU synthesizes more accurate face and neck features. Contrary to recent literature [27], ReLU also synthesizes similar quality results. On the other hand, the Sine activation function performs poorly in the virtual try-on task.

#### 4.5. Optical Flow

To improve temporal consistency, we take inspiration from [34, 6] and experiment with optical flow to improve temporal consistency between frames. In particular, we use flow directional annotations to warp the previous generated frame into the current timestep. A mask is then used to compose the warped frame with the originally synthesized result. We utilize a mask penalty,  $\mathcal{L}_F$  weighted by  $\lambda_F = 1 \times 10^4$  to constrain the learned flow mask. Our flow annotations are generated from FlowNet2 [13].

During experiments, we find that flow causes issues with the general quality of generated images; specifically, flow introduces undesired artifacts to the video. The decreased quality is quantified by Figure 6d. For this reason, we omit flow in our reported best-performing result.

### 5. Results

We present the results of our bottom-up experiments. Our experiments show that the best-performing and most compatible model uses DensePose pose annotations, self-attention layers, and GeLU or ReLU activation function. We use this model to qualitatively and quantitatively compare with existing baselines.

#### 5.1. Evaluation Metrics

In the field of virtual try-on, evaluation metrics are of high importance. Given the visual nature of virtual try-on, quantitative metrics do not tell the whole story. However,

quantitative metrics are good at judging the overall quality of the output.

**Structural Similarity Index (SSIM).** SSIM is a perceptual metric that quantifies image quality caused by image processing. SSIM is often used in analyzing quality between videos. SSIM [35] is defined by

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \quad (3)$$

where  $\mu$  is the average,  $\sigma^2$  is the variance, and  $c_1$  and  $c_2$  are constants used to stabilize the calculation. For our case, we use the multiscale SSIM metric, which applies SSIM across all input image channels, because it has been shown to perform better than vanilla SSIM.

**Peak Signal-to-Noise Ratio (PSNR).** PSNR is used to measure the quality of image reconstruction. Generally, this is used in the context of quality of image compression. We find that PSNR is a useful quantitative metric for judging the quality of a reconstructed image. PSNR is defined by

$$PSNR(x, y) = 10\log\left(\frac{MAX_i^2}{MSE(x, y)}\right) \quad (4)$$

where MAX is largest pixel value in the input images, and MSE is defined as the mean square error between the  $x$  and  $y$ . Similar to SSIM, PSNR cannot be the standalone metric to judge the quality of try-on. That being said, PSNR is a sufficient approximation to human perception of image and video quality. Therefore, it is part of the comparison tools we use, for its general ability to judge and compare between good-quality and bad-quality videos.

#### 5.2. Quantitative Results

Through our bottom-up experiment structure, we determine the best method as per quantitative and qualitative

analysis and determine the best-performing model as ShineOn. We used ShineOn methods and compared to existing try-on methods, CP-VTON and FW-GAN [33, 6], to synthesize video virtual-try on results and compare. Using the SSIM and PSNR metrics, we calculate the mean and standard deviation at each frame across the three methods that we compare. Given that our problem statement is different from FW-GAN, Figure 7 reports these metrics and demonstrates that our method quantitatively outperforms existing try-on methods.

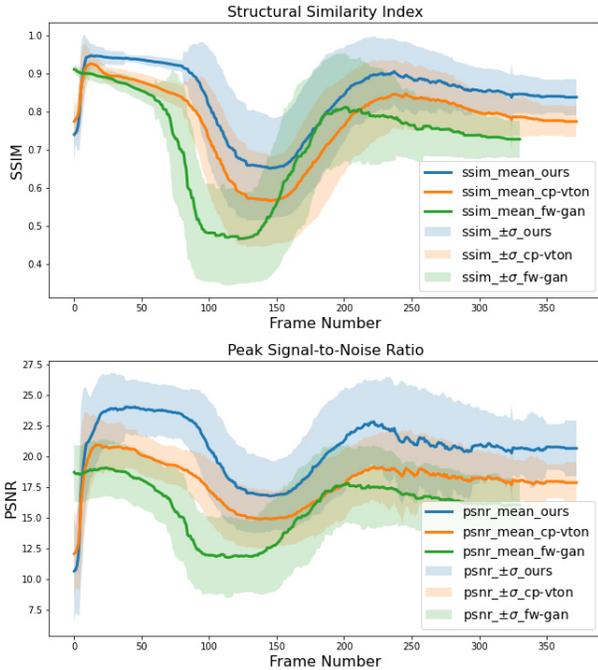


Figure 7: **Quantitative Comparison with Existing Try-On Methods.** We use SSIM and PSNR as our quantitative metrics to compare with CP-VTON and FW-GAN [33, 6]. We note, for completeness, that our problem statement differs from FW-GAN. While that makes our approach more applicable for practical use, the quantitative comparison is incomplete.

### 5.3. Qualitative Results

Qualitative inspection and comparison are absolutely vital to judging the effectiveness of video virtual try-on. Due to the importance of visual inspection in virtual try-on, qualitative criteria must be properly defined to evaluate the effectiveness of our presented methods. These criteria are quality of body parts, quality of cloth synthesis, and temporal consistency. We present visual comparisons in Table 1. Our ShineOn methods outperform the CP-VTON baseline. Our methods demonstrate strong body and face synthesis, better cloth transfer and texture, and temporal consistency.

Using the qualitative criteria to compare our methods with FW-GAN, we find that our method retains stronger face and body detail. It should be noted that our model’s higher quality face detail arises from the differing problem formulations. Insofar as ShineOn and FW-GAN are video virtual try-on methods, this comparison stands. ShineOn preserves shirt designs more consistently to the product image, while the shirt designs in FW-GAN tend to be scattered. The weaknesses of our ShineOn are that it fails to generate video virtual try-on results with complete neck details. Furthermore, cloth try-on is increasingly misaligned during the frames where the camera is zoomed in on the person subject. The failure case of zoomed-in images is not specific to our network. FW-GAN also fails to effectively synthesize cloth for zoomed-in frames. Even though FW-GAN provides temporal smoothness due to temporal discriminators, it struggles with temporal consistency.

## 6. Conclusion

We methodically illuminate the effect of several design choices for practical video virtual try-on. Our series of bottom-up experiments examine the outcomes of pose annotation choice, self-attention, activation functions, and optical flow. The strengths and weaknesses of these design outcomes are compared to existing work. Importantly, we release our code, hyperparameters, and model checkpoints to the public, not only to support experiment reproducibility but also as a framework for future works.

Within our experimental scope, we identify the design choices that result in the highest quality try-on. Ordered by importance: (1) DensePose annotations improve face detail and decrease required memory and training time, (2) self-attention slightly benefits face and body feature quality, and (3) ReLU and GELU activations perform equally well, but not Swish nor Sine. We also identify that our design choice using optical flow improves temporal smoothness but introduced artifacts. Our methodical approach to analyzing simple design choices results in significant improvement over the CP-VTON image try-on baseline. Though our best result struggles with neck details, it preserves user identity and shirt design better than FW-GAN.

For future work, we recommend investigating issues with synthesizing the neck. We also suggest improving the cloth warping module, as it fails to handle 3D orientation and geometric information of the cloth (e.g. differentiating between inside and outside). Future work may consider improving global temporal consistency instead of only local smoothness. Finally, as demonstrated here, we encourage reproducibility via methodical experiments supported by public code.

ID & Frame Num.	Cloth (Input)	FW-GAN User Image (Input)	FW-GAN	VVT User Video (Input)	CP-VTON	ShineOn: DensePose, Attn, GELU
4he21d00f-g11: frame 040						
4he21d00f-g11: frame 125						
4he21d00f-k11: frame 040						
4he21d00f-k11: frame 125						
an621da9d-q11: frame 040						
an621da9d-q11: frame 125						
g1021d05g-k11: frame 040						
g1021d05g-k11: frame 150						

Table 1: Qualitative try-on results comparing FW-GAN, CP-VTON (retrained on VVT), and the best ShineOn model that uses DensePose, Attention, and GeLU. As explained in Section 3.1, FW-GAN reposes a still user image (column 3), while we directly use the user video frame (column 5). ShineOn better preserves user identity (face quality and other body parts), target cloth color, and target cloth texture design. However, ShineOn suffers from neck synthesis on zoomed views; we recommend exploring this issue in future work.

## References

- [1] Huiwen Chang, James Hays, Duygu Ceylan, Amit Raj, Pat-sorn Sangkloy, and Jingwan Lu. Swapnet: Image based garment transfer, 2018.
- [2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. *arXiv preprint arXiv:2005.12872*, 2020.
- [3] Jean-Baptiste Cordonnier, Andreas Loukas, and Martin Jaggi. On the relationship between self-attention and convolutional layers. *arXiv preprint arXiv:1911.03584*, 2019.
- [4] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *arXiv preprint arXiv:2004.07728*, 2020.
- [5] Haoye Dong, Xiaodan Liang, Xiaohui Shen, Bochao Wang, Hanjiang Lai, Jia Zhu, Zhiting Hu, and Jian Yin. Towards multi-pose guided virtual try-on network. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9026–9035, 2019.
- [6] Haoye Dong, Xiaodan Liang, Xiaohui Shen, Bowen Wu, Bing cheng Chen, and Jian Yin. Fw-gan: Flow-navigated warping gan for video virtual try-on. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1161–1170, 2019.
- [7] L. A. Gatys, A. S. Ecker, and M. Bethge. Image style transfer using convolutional neural networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2414–2423, 2016.
- [8] Yuying Ge, Ruimao Zhang, Lingyun Wu, Xiaogang Wang, Xiaoou Tang, and Ping Luo. Deepfashion2: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images, 2019.
- [9] Ke Gong, Xiaodan Liang, Dongyu Zhang, Xiaohui Shen, and Liang Lin. Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing, 2017.
- [10] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7297–7306, 2018.
- [11] Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S. Davis. Viton: An image-based virtual try-on network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.
- [12] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus), 2020.
- [13] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. FlowNet 2.0: Evolution of optical flow estimation with deep networks, 2016.
- [14] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks, 2018.
- [15] Nikolay Jetchev and Urs Bergmann. The conditional analogy gan: Swapping fashion articles on people images, 2017.
- [16] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution, 2016.
- [17] A. Jong and T. S. Moh. Short video datasets show potential for outfits in augmented reality. In *2019 International Conference on High Performance Computing Simulation (HPCS)*, pages 201–208, 2019.
- [18] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.
- [19] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.
- [20] Ting-Yen Chen, Kuan-Hsien Liu, and Chu-Song Chen. MvC: A dataset for view-invariant clothing retrieval and attribute prediction. In *Proceedings of the ACM International Conference on Multimedia Retrieval*, 2016.
- [21] Wei-Sheng Lai, Jia-Bin Huang, Oliver Wang, Eli Shechtman, Ersin Yumer, and Ming-Hsuan Yang. Learning blind video temporal consistency. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 170–185, 2018.
- [22] Xiaodan Liang, Ke Gong, Xiaohui Shen, and Liang Lin. Look into person: Joint body parsing pose estimation network and a new benchmark, 2018.
- [23] Y. Liu, W. Chen, L. Liu, and M. S. Lew. Swapgan: A multi-stage generative approach for person-to-person fashion style transfer. *IEEE Transactions on Multimedia*, 21(9):2209–2222, 2019.
- [24] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [25] Assaf Neuberger, Eran Borenstein, Bar Hilleli, Eduard Oks, and Sharon Alpert. Image based virtual try-on network from unpaired data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.
- [26] Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Łukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. Image transformer. *arXiv preprint arXiv:1802.05751*, 2018.
- [27] Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *International Conference on Machine Learning*, pages 5301–5310, 2019.
- [28] Prajit Ramachandran, Barret Zoph, and Quoc V. Le. Searching for activation functions, 2017.
- [29] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015.
- [30] Vincent Sitzmann, Julien NP Martel, Alexander W Bergman, David B Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. *arXiv preprint arXiv:2006.09661*, 2020.
- [31] Matthew Tancik, Pratul P Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan T Barron, and Ren Ng. Fourier features

- let networks learn high frequency functions in low dimensional domains. *arXiv preprint arXiv:2006.10739*, 2020.
- [32] A Vaswani, N Shazeer, N Parmar, J Uszkoreit, L Jones, AN Gomez, L Kaiser, and I Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.
- [33] Bochao Wang, Huabin Zheng, Xiaodan Liang, Yimin Chen, Liang Lin, and Meng Yang. Toward characteristic-preserving image-based virtual try-on network, 2018.
- [34] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-video synthesis. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2018.
- [35] Z. Wang, E. P. Simoncelli, and A. C. Bovik. Multiscale structural similarity for image quality assessment. In *The Thirty-Seventh Asilomar Conference on Signals, Systems Computers, 2003*, volume 2, pages 1398–1402 Vol.2, 2003.
- [36] Zhonghua Wu, Guosheng Lin, Qingyi Tao, and Jianfei Cai. M2e-try on net: Fashion from model to everyone, 2018.
- [37] Cihang Xie, Mingxing Tan, Boqing Gong, Alan Yuille, and Quoc V Le. Smooth adversarial training. *arXiv preprint arXiv:2006.14536*, 2020.
- [38] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057, 2015.
- [39] Han Yang, Ruimao Zhang, Xiaobao Guo, Wei Liu, Wangmeng Zuo, and Ping Luo. Towards photo-realistic virtual try-on by adaptively generating-preserving image content. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.
- [40] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 1480–1489, 2016.
- [41] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *International Conference on Machine Learning*, pages 7354–7363, 2019.
- [42] Haokui Zhang, Chunhua Shen, Ying Li, Yuanzhouhan Cao, Yu Liu, and Youliang Yan. Exploiting temporal consistency for real-time video depth estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1725–1734, 2019.
- [43] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.