GC-VTON: Predicting Globally Consistent and Occlusion Aware Local Flows with Neighborhood Integrity Preservation for Virtual Try-on

Hamza Rawal[1], Muhammad Junaid Ahmad[1], Farooq Zaman[2]

[1]Motive
[2]Information Technology University, Lahore, Pakistan
{hamzarawal, farooqzaman20}@gmail.com, junaidahmad1998@outlook.com

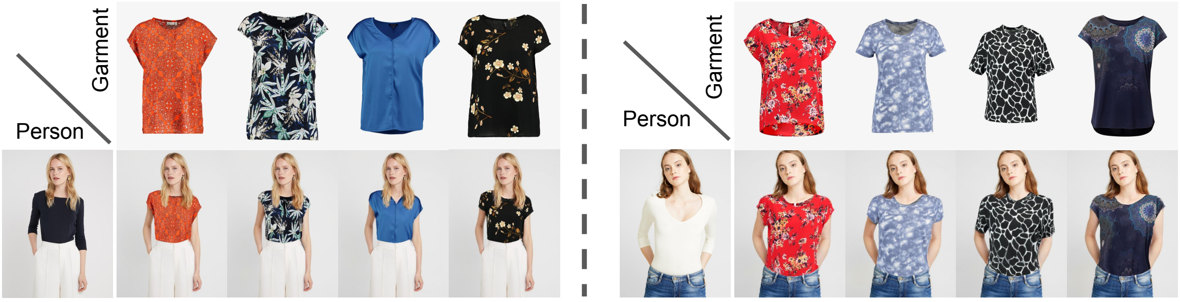

Figure 1. Sample try-on images generated by our method over a set of garments with complex textures. Zoom in for details.

Abstract

Flow based garment warping is an integral part of image-based virtual try-on networks. However, optimizing a single flow predicting network for simultaneous global boundary alignment and local texture preservation results in sub-optimal flow fields. Moreover, dense flows are inherently not suited to handle intricate conditions like garment occlusion by body parts or by other garments. Forcing flows to handle the above issues results in various distortions like texture squeezing, and stretching. In this work, we propose a novel approach where we disentangle the global boundary alignment and local texture preserving tasks via our GlobalNet and LocalNet modules. A consistency loss is then employed between the two modules which harmonizes the local flows with the global boundary alignment. Additionally, we explicitly handle occlusions by predicting body-parts visibility mask, which is used to mask out the occluded regions in the warped garment. The masking prevents the LocalNet from predicting flows that distort texture to compensate for occlusions. We also introduce a novel regularization loss (NIPR), that defines a criteria to identify the regions in the warped garment where texture integrity is violated (squeezed or stretched). NIPR subsequently penalizes the flow in those regions to ensure regular and coherent warps that preserve the texture in local neighborhoods. Evaluation on a widely used virtual try-on dataset demonstrates strong performance of our network compared to the current SOTA methods.

1. Introduction

Image-based virtual try-on aims at generating natural, distortion and artifacts-free images of a person wearing a selected garment. Image synthesis via GANs [7] has been widely used in applications like image editing [16, 23, 26], style-transfer [2, 8, 33] and image generation [13, 30, 34]. However, simply using synthesis methods that holistically change the image does not result in the desired quality in virtual try-on setting. Existing methods adopt a scheme where the garment is first warped to meet the target person pose requirements. A GAN based generator network then fuses the warped garment and the target person images to generate a final try-on image. Traditionally, the warping is either done by a Thin Plate Spline (TPS) warp [3, 5, 10, 12, 15, 25, 29], or a dense flow fields based warp [1, 4, 9, 11, 14], or a combination of both [28]. In any case, the warping is inherently not capable of modeling all the changes that a garment undergoes (e.g occlusions) when it fits on a target person. And forcing it to do so, results in artifacts such as texture squeezing, stretching, and garment tear, etc.

Recently, dense flow networks have shown good warping performance compared to the TPS based warping networks. However, for dense flows, simultaneous alignment of the global boundaries and local texture preservation is still a challenge. While improvement attempts have been made, such as conditioning the flow on global style [11] and sparse/dense body pose [1, 14], the results are still far from perfect. We believe that this is a product of overcomplicating the job of a flow predicting network and setting unrealistic goals to optimize both the problems simultaneously. To this end, we propose to disentangle the two jobs by dedicating a separate module to each task. A consistency loss is then utilized between the outputs of the two networks to ensure harmony between local and global flows.

Furthermore, garment-body and garment-garment interplay introduces various occlusions. For example half sleeves, sleeveless, full sleeves and full neck shirts tend to reveal or conceal certain body parts. The visible body parts can potentially occlude certain garment regions. Forcing flow to model occlusions is an ill-posed problem as the flow can only transform pixel location, but not conceal them. The only way a flow can model occlusions is by tearing the garment and squeezing it around the occlusions. Additional occlusion may be introduced by garment styles e.g upper garment (shirt) tucked into the bottom garment (pants). Flow predicting networks match global boundaries of this style by predicting high values for the lower portion of the upper garment, hence introducing extreme squeezing artifacts. Although handling each of the above occlusion types is equally crucial to get a good warp, existing works either ignore them altogether [10–12], or settle for just handling some of them [3, 9, 29]. This sets unrealistic goals for the flow predicting network and subsequently results in sub-par try-on results. To address these issues, we propose to predict a body-parts visibility mask that explicitly masks out the occluded regions from the warped garment. This setup prevents the flow predicting network from predicting high flow values in the occluded regions, which in turn prevents distortion and artifacts. We input the visibility masks to the generator module as well, which serves as an added guidance for synthesizing visible skin regions.

Additionally, in virtual try-on methods, different losses such as TV-loss [17] and second-order smoothness loss [6] are utilized to prevent irregular flows by enforcing smooth distances between pixels in local neighborhoods. However, changes in local neighborhoods are governed by global changes in the garment, which these losses fail to take into consideration. Additionally, they cannot identify the types of the artifacts (squeezing and stretching), thereby fail to apply appropriate penalties on the network. We show that these serious limitations allow artifacts to slip away and result in unrealistic try-on synthesis. To address these limitations, we propose a novel Neighborhood Integrity Preserving Regularization loss (NIPR). The penalty term in NIPR is globally-informed and artifact-specific, thus penalizing bad warps appropriately. Our contributions can be summarized as:

- We disentangle the global boundaries alignment and local flow adjustment tasks to achieve globally consistent local warps. Utilizing a consistency loss between the outputs of the two modules enforces the necessary harmony.
- We propose to estimate body-parts visibility mask to take care of occlusions in the warped garment.
- We show the limitations of the existing flow smoothening losses and propose NIPR to effectively guard against texture integrity violation.

2. Related Work

2.1. Garment Warping in Virtual Try-on

Virtual try-on methods using 3D information [11, 18, 20, 22, 32] have limited applications as the data annotation cost is high. Recently, 2D image based virtual try-on [10, 11, 29] has gained traction because of its simplicity and less input information required. Generally, a two-stage paradigm is adopted where the first step is to warp the given garment to match the target person pose and then a generator based network fuses the warped garment and the target person image. Initial methods [3, 5, 15] applied a TPS based warping technique where a network predicts a set of sparse control points which can be used to warp the garment. Nowadays, dense flow based warping is the method of choice because a flow field has more degree of freedom, thus suited for warping garments with rich textural details [4, 9, 14]. A good warp is characterized by two properties: the textural details that it preserves and the global boundary alignment (with the target pose) that it achieves. Although, the two requirements have distinct goals, most of the existing methods expect a single flow predicting network to handle them simultaneously. This leads to unrealistic warps which subsequently results in unnatural try-on results. We propose to handle this by disentangling the two tasks via two separate modules, the outputs of which are harmonized via a consistency loss.

2.2. Occlusion Handling

Occlusion is another major complication in virtual try-on which leads to artifacts around the occluded regions. Potential sources of occlusion include body parts and other

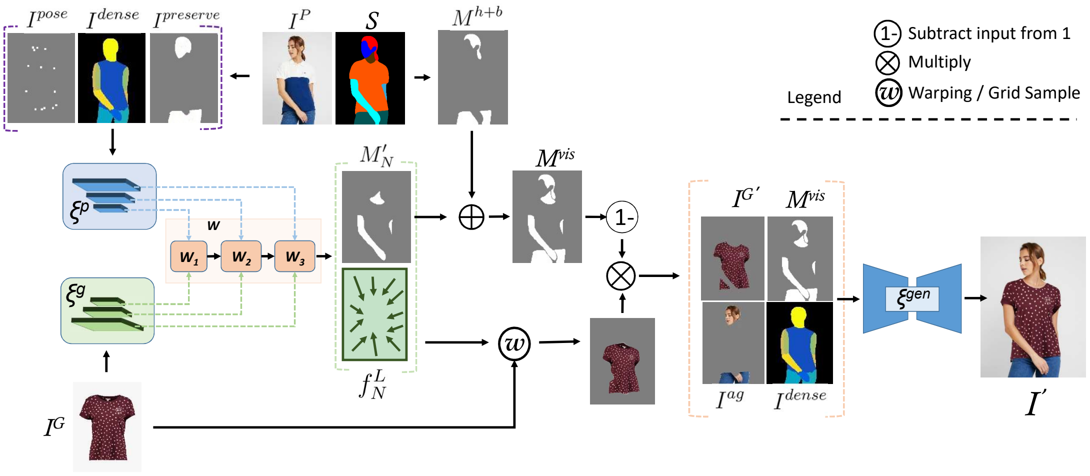

Figure 2. An overview of GC-VTON architecture. Person representation maps and garment image are encoded and then fed to the warping module $W = \{W_1, W_2, W_3\}$ which predicts flow fields and body-parts visibility masks. Each warping module W_i is composed of a LocalNet and GlobalNet modules. Visibility masks handle occlusions in the warped garment to give $I^{G'}$. $I^{G'}$ and other person representations are input to a generator module to get the try-on image $I^{'}$.

garments, all of which must be attended to, in order to achieve an artifact-free warp. The existing literature either ignores [10–12] the occlusion sources altogether or partially attend to some of the sources [3, 9, 29] thereby predicting sub-bar warps. In our work, we predict body-parts visibility mask which directly handles the occlusions and prevent flow prediction network from causing distortion.

2.3. Flow Regularization

Predicted flows must be smooth (without irregularities) in order to achieve realistic warping results. Multiple losses have been proposed in the existing works to enforce smoothness and coherence in the flow. For instance, TV-Loss [4, 17] minimizes the total variance in a flow field, hence encouraging global spatial smoothness. Authors in [6] proposed a second-order smoothness loss, which minimizes a generalized charbonnier loss function [24] between the distances of a pixel and its two vertical and horizontal neighbors. The intention is to encourage co-linearity in the predicted flow and prevent pixels from falling apart as a result of a bad flow. Authors in [29] proposed to equalize the distance of the horizontal and vertical neighbors from a selected point p, where p is a member of the set of the predicted control points used to warp the garment. Furthermore, they equalize the slopes of the lines between the neighbors. The loss encourages the warp to maintain co-linearity, parallelism, and immutability properties of the affine warp.

While these losses have certainly proved to be useful in the context of virtual try-on, they still come with some limitations. We show that the criteria defined in these losses to penalize a bad flow can still be satisfied even when a true (good) warp is not achieved. This can potentially result in certain artifacts being overlooked. We argue that this behavior can be attributed to two discrepancies. First, the criteria they define is strictly local in nature, whereas the distance between the pixel neighbors is a function of the global changes (in height and width) of the garment. Second, they cannot identify the nature of the artifacts i-e they do not know if the artifact is caused by stretching or squeezing. Both the artifacts are distinct in nature and need appropriate penalties. In our work, we propose a new regularization loss (NIPR) that effectively mitigates these issues by applying globally-compliant and artifact-specific penalties to the flows.

3. Methodology

The overall architecture of our method is given in Fig 2. In this section we detail the working of each module of our approach, its purpose and the intuition behind it.

3.1. Problem Setting

Given a target person image $\{I^P \epsilon \mathbb{R}^{3\times W\times H}\}$ wearing an original garment $\{I^{PG}\}$, and a garment image $\{I^G \epsilon \mathbb{R}^{3\times W\times H}\}$, the goals of the virtual try-on are to (i) remove $\{I^{PG}\}$ from $\{I^P\}$, (ii) modify $\{I^G\}$ to generate try-on garment $\{I^{G'}\}$, and (iii) generate an output image $\{I' \epsilon \mathbb{R}^{3\times W\times H}\}$, where the person in $\{I'\}$ is wearing the modified garment $\{I^{G'}\}$. $\{I^{G'}\}$ should conform to the person's pose and body settings and simultaneously preserve the texture and design details from $\{I^G\}$. And any traces of the person's original garment $\{I^{PG}\}$ should not be present in $\{I'\}$. $\{I'\}$ should also preserve the non-garment regions of the person $\{I^P\}$. Simply put, the generated image

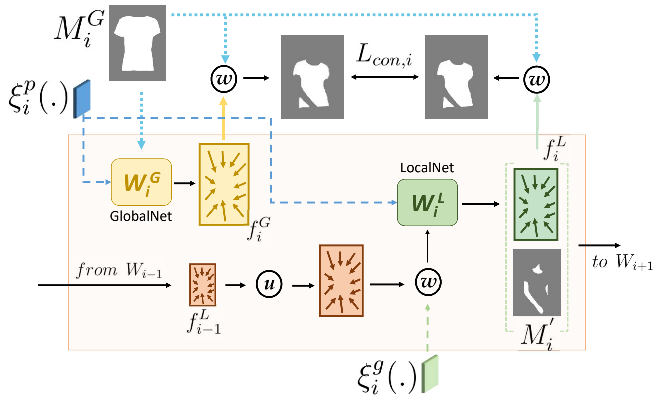

Figure 3. Warping block design. Local flow from previous block $\{f_{i-1}^L\}$ warps garment features at current scale $\{\xi_i^g(.)\}$. Conditioned on the warped garment feats and person feats $\{\xi_i^p(.)\}$, the LocalNet predicts local flow $\{f_i^L\}$ and body-parts visbility mask $\{M_i^{'}\}$(which are output to next block). Given garment mask and person representations, GlobalNet predicts global flow $\{f_i^G\}$. A consistency loss is employed between warped garment masks to harmonize local flows with global alignment.

$\{I'\}$ should look natural and plausible. Like other methods [11, 14], we assume that we have access to person pose $\{I^{pose}\}$, densepose $\{I^{dense}\}$ and person body segmentation map $\{S\}$.

3.2. Network Overview

In GC-VTON we use two encoders: (i) $\{\xi^g\}$ to encode $\{I^G\}$, and (ii) $\{\xi^p\}$ to encode person representations $\{I^{pose}$, I^{dense}, and $I^{preserve}\}$. $\{I^{preserve}\}$ is the non-garment region extracted from S. The encoded information is then fed to the warping module $\{W_i\}_{i=1}^N$. Each block of $\{W_i\}$ is composed of a LocalNet $\{W_i^L\}$ and a GlobalNet $\{W_i^G\}$ module (Fig 2). LocalNet consumes outputs from $\{\xi^g\}$ and $\{\xi^p\}$ to infer fine-grained local flows $\{f^L\}$ and body parts visibility mask $\{M^{'}\}$. We add hair and bottom garment mask $\{M^{h+b}\}$ from S to $\{M^{'}\}$ to form the full body visibility mask $\{M^{vis}\}$. GlobalNet takes a garment mask $\{M^G\}$ and output of $\{\xi^p\}$ to infer global alignment flow $\{f^G\}$ (Fig 3). GlobalNet is only needed at train time and is dropped at test time. Local flow $\{f^L\}$ is used to warp the original garment $\{I^G\}$ and the visibility mask $\{M^{vis}\}$ is used to mask out occluded regions to generate final warped garment $\{I^{G'}\}$. Warped garment $\{I^{G'}\}$, densepose $\{I^{dense}\}$, garment agnostic person image I^{ag} [3] and visibility mask $\{M^{vis}\}$ are then input to the generator, which generates try-on image $\{I'\}$.

3.3. Globally Consistent Local Warping

In a virtual try-on network, garment warping must take care of two tasks: local textures should be preserved and global boundaries of the garment must conform to the target person pose and body. We argue that local warping is a garment dependent problem where garment texture must be taken into account to prevent undesired artifacts. While global boundary alignment is inherently garment agnostic and does not care about local textures as the only goal is to match global boundaries. Therefore, using the same network to solve both the problems together creates a tug-of-war scenario. Generally, this results in unrealistic warps as the network aligns the global boundaries at the expense of bad local warps or vice versa. We propose to solve this problem by dedicating separate modules for the local warps and global boundaries alignment. This removes the competition between the two goals and results in smooth and coherent warps. The LocalNet exclusively focuses on generating flows that prevent local texture from artifacts. The local flow is then aligned to the global flow using a consistency loss between the warped garment masks of the local and global flows. This disentangling takes cares of local adjustments before aligning them to global boundaries, thus resulting in distortion-free warped garments. Formally,

$$f_i^L = W_i^L(\xi_i^g(I^g) \odot \xi_i^p(I^{pose} \odot I^{dense} \odot I^{preserve})) \quad (1)$$

$$f_i^G = W_i^G(\xi_i^{gm}(M^G) \odot \xi_i^p(I^{pose} \odot I^{dense} \odot I^{preserve})) \quad (2)$$

where i represents i-th feature scale, and f_i^L, f_i^G are the predicted local flow, global flow respectively. W^L and W^G are the local and global warping modules, while ξ_i^{gm} is a garment mask encoder inside W^G. The $\odot$ represents concat operation. Note that in line with the above discussion, our GlobalNet is a function of garment mask (thus texture agnostic) and the LocalNet on the other hand takes original garment image as input (thus texture and details dependent). Using the predicted flows, warping is achieved as:

$$y_i = warp(f_i, x_i) \quad (3)$$

where $warp$ is the grid sampling operator, $\{x_i\}$ is the input tensor at i-th scale, $\{f_i\}$ is the flow and $\{y_i\}$ is the warped output. We warp $\{I_i^G\}$ using $\{f_i^L\}$ to obtain $\{I_i^{G'}\}$. Similarly, garment mask $\{M_i^G\}$ is warped using local flow $\{f_i^{loc}\}$ and global flow $\{f_i^G\}$ to obtain $\{M_i^{G'/L}\}$ and $\{M_i^{G'/L}\}$ respectively. The consistency is then enforced by employing an L1 loss between the warped garment masks:

$$\mathcal{L}_{con,i} = L1(M_i^{G'/L}\ ,\ M_i^{G'/G}) \quad (4)$$

3.4. Occlusion Handling

As discussed, garment occlusions by body parts, e.g hands, cause distortions in the warped garment as flows are inherently not suited to handle the occlusions. To this end, we propose to predict body parts visibility mask which is

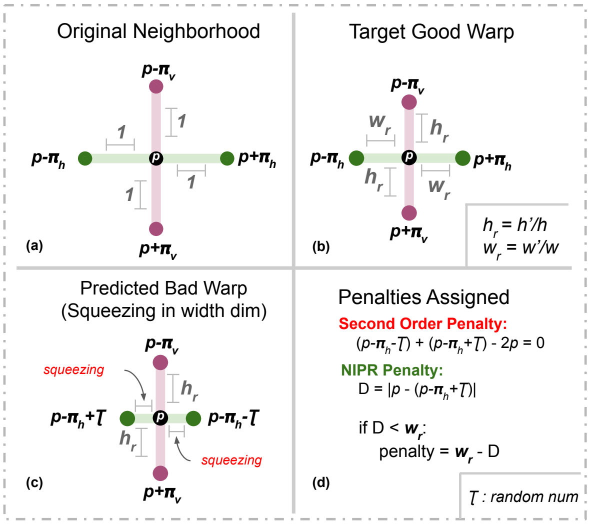

Figure 4. NIPR loss illustration. (a) An original local neighborhood. (b) A target true warp that is to be achieved. (c) Predicted bad warp with squeezing artifact in width dimension. (d) Second-order constraint [6] is satisfied and fails to correct the bad warp. NIPR appropriately penalizes the flow.

used to explicitly handle occlusions. Since body parts visibility mask also depends upon the garment $\{I^G\}$ and person pose, we task the LocalNet to predict an additional output $M_i^{'}$ as:

$$f_i^L, M_i^{'} = W_i^L(.) \tag{5}$$

where $W_i^L(.)$ is LocalNet and its inputs from Eq 1. Hair and bottom garments can also contribute to occlusion, but their visibility does not depend on the upper garment. Therefore, we directly add hair and bottom garment mask $\{M^{h+b}\}$ to $\{M_i^{'}\}$ to form $\{M_i^{vis}\}$. $\{M^{h+b}\}$ is obtained from person segmentation map $\{S\}$. The occluded regions are masked out from warped output as:

$$y_i = y_i * (1 - M_i^{vis}) \tag{6}$$

where the visibility mask is used to mask out the occluded regions. In the backpropagation step, this setup masks out the gradients for LocalNet, thereby preventing garment tear artifacts.

3.5. Neighborhood Integrity Preserving Regularization (NIPR)

Unconstrained dense flows try to align the garment to the global pose at the expense of bad local warping. As discussed, in challenging cases such as tucked-in style, the global alignment necessitates predicting high flow values in the bottom region of the garment. This causes unwanted texture squeezing, which must be guarded against to produce plausible warps. Existing regularization losses such as Second-Order Smoothness loss [6] try to ensure smooth flows by minimzing an objective that encourages equal distances between vertical neighbor pairs of a pixel and similarly for horizontal neighbor pairs (Fig 4). But as shown in Fig 4(d), the configuration is ill-posed as the constraints can be met even when correct warping is not achieved. This limitation is owed to ignoring global garment changes when equalizing distance in local neighborhoods. This issue is more evident in extreme squeezing and stretching scenarios. And since the only aim of these losses is to ensure equalized distances between neighbors, they simply cannot identify and correct extreme squeezing and stretching cases. Formally, the second order smoothness constraint is given as:

$$\mathcal{L}_{so} = \sum_{i=1}^{N} \sum_{p \epsilon P} \sum_{\pi \epsilon [\pi_v, \pi_h]} \left| f_i^{p-\pi} + f_i^{p+\pi} - 2 f_i^p \right| \tag{7}$$

where f_i^p is the p-th point on the flow field at i-th scale, and $[\pi_v, \pi_h]$ are the vertical and horizontal neighbors of p (Fig 4a). The $|.|$ operator represents absolute function.

In this work, we introduce a novel regularization loss NIPR with the goal of preserving texture integrity in local neighborhoods of the warped garment. NIPR is designed to enforce a simple yet effective principle: distance between pixels in a neighborhood before and after a warp should be consistent and must conform to the global changes. If distance between two adjacent vertical pixels in the original garment was 'one', then the distance between the same two pixels in the warped garment should roughly be $h_r = h'/h$. Where h' and h are the heights of the warped garment and the original garment respectively. If the distance is greater than h_r then it indicates stretching, if its less, it indicates squeezing (Fig 4c), and if its equal, it means roughly a perfect warp. Similarly, for two adjacent horizontal pixels, the distance in the warped garment should be $w_r = w'/w$.

For a particular location, if the distance between two neighbors exceeds r ($r = h_r$ for vertical neighbors and w_r for horizontal), we simply add this distance to the total loss. This encourages the network to reduce the distance in order to reduce the total loss. If the distance is less than r, we subtract it from r and add the difference to the total loss. The objective of the loss function now becomes to increase the distance to be equal to r. When applied to every location of a flow field, NIPR ensures a smooth and coherent warp that is free of the aforementioned artifacts. Specifically, to formulate NIPR, we add a new component to $\mathcal{L}_{so}$:

$$\mathcal{L}_{nipr} = \sum_{i=1}^{N} \sum_{p \epsilon P} \left(\sum_{\pi \epsilon [\pi_v, \pi_h]} \left| f_i^{p-\pi} + f_i^{p+\pi} - 2 f_i^p \right| + \sum_{u \epsilon U} \mathcal{L}_i^{preserve}(p, u) \right) \tag{8}$$

where,

$$\mathcal{L}_i^{preserve}(p,u) = \begin{cases} D_i(p,u) & if\, D_i(p,u) > r \\ r - D_i(p,u) & if\, D_i(p,u) < r \\ 0 & otherwise \end{cases} \tag{9}$$

is the integrity preservation component at i-th scale responsible for aligning flows in the local neighborhoods such that they conform to the global changes. $D_i(p,u)$ is the absolute distance between a point p on the flow field of i-th scale and its neighbor at a location $u \,\epsilon\, U$. Where U is a set of four neighbors $\{p - \pi_h,\, p + \pi_h,\, p - \pi_v,\, p + \pi_v\}$. Please note that the notation for neighbors in $\mathcal{L}_{so}$ is different than $\mathcal{L}^{preserve}$ because the former works on neighbor pairs and the later deals with a single neighbor at a time. And,

$$r = \begin{cases} h_r & if\;\; l \,\epsilon\, \pi_v \\ w_r & if\;\; l \,\epsilon\, \pi_h \end{cases} \tag{10}$$

is the ratio of the warped to the original garment's height or width depending upon the vertical or horizontal neighbor. Conditioning the penalty on the height and width ratios ensures that the penalty is in line with the global requirements. While conditioning it on the artifact types enables an appropriate response to each artifact, thus alleviating the artifacts effectively.

3.6. Try-on Generator

We synthesize the final try-on image I' by utilizing a ResUNet [21] based encoder-decoder generator ξ^{gen}. ξ^{gen} takes the warped garment $I_N^{G'}$, dense pose I^{dense}, visibility mask M_N^{vis}, and a clothing agnostic person representation I^{ag} [3]. Here N represents the outputs from the last warping block. The visibility mask guides the generator to synthesize skin in the visible regions and at the same time prevents synthesizing garment in the occluded regions.

3.7. Objective Functions

To train our warping module, we follow previous works [6, 11] and utilize $L1$ loss $\{\mathcal{L}_1^{garment}\}$ and perceptual loss $\{\mathcal{L}_p^{garment}\}$ between the warped and the target garments. We also apply $L1$ loss $(\mathcal{L}_1^{mask})$ between warped and target garment masks. In addition, we also use our custom consistency loss $\{\mathcal{L}_{con}\}$ and NIPR loss $\{\mathcal{L}_{nipr}\}$. Since the fine-grained local flows are susceptible to extreme squeezing and stretching, we apply $\{\mathcal{L}_{nipr}\}$ only to the local flow. To learn body-parts visibility masks, we use binary cross entropy $\{\mathcal{L}_{BCE}\}$ between the target and predicted masks. So the total objective for the warping module becomes:

$$\begin{aligned} \mathcal{L}_{warp} = & \{\alpha_1\mathcal{L}_1 + \alpha_2\mathcal{L}_p\}^{garment} \\ & + \{\alpha_3\mathcal{L}_1\}^{mask} \\ & + \alpha_4\mathcal{L}_{BCE} + \alpha_5\mathcal{L}_{con} \\ & + \alpha_6\mathcal{L}_{nipr} \end{aligned} \tag{11}$$

Method	SSIM↑	FID↓	LPIPS↓	Human↑
VITON-HD	0.868	9.970	0.1141	6.72%
HR-VITON	0.873	9.819	0.0954	12.78%
FS-VTON	0.879	7.904	0.0951	26.2%
GC-VTON (Ours)	**0.887**	**7.888**	**0.0831**	**54.3%**

Table 1. Quantitative comparison to existing methods on VITON-HD [3] dataset.

where $\{\alpha_i\}_{i=1}^{6}$ are the loss balancing hyper-params for the warping network. To train our generator network, we again utilize $L1$ and perception losses for the generated try-on image I'. We also employ an adversarial loss which has proved to be effective in GANs to generate realistic results. The total objective for the Generator network becomes:

$$\mathcal{L}_{gen} = \{\beta_1\mathcal{L}_1 + \beta_2\mathcal{L}_p\}^{try-on} + \beta_3\mathcal{L}_{adv} \tag{12}$$

where $\{\beta_i\}_{i=1}^{3}$ are the loss balancing hyper-params for the generator network.

4. Experiments

4.1. Dataset

For our work, we use the widely used virtual try-on dataset VITON-HD [3]. It contains 13,769 frontal-view female and upper garment pairs for training. The testing set contains 2,032 person and garment pairs. The garment in a training pair is the same that the person is wearing. For a testing pair, the garment is different than what the person is wearing. The original dataset has an image resolution of 1024x768 and we resize these images to 256x192 for our warping module and 512x384 for the generator module.

4.2. Implementation details

Our model is trained with a single RTX 3090 12 GB GPU. Our warping module is a set of N Conv-LeakyReLU blocks ($N = 5$ in this work). We train the warping module for 100 epochs with a batch size of 14 at 256x194 resolution. The output flow from warping module is upscaled by a factor of two to warp the garment at a resolution of 512x384, which is the input and output resolution for the generator network. The generator module is trained for 100 epochs with a batch size of four and is based on the implementation of [11]. We use Adam optimizer with an initial learning of $5e-4$ which linearly decays after 50 epochs. The loss mixing hyper-params for warping modules training are empirically found out and set to $\{\alpha_i\}_{i=1}^{6} = \{1, 0.2, 2, 2, 1, 1\}$. The generator loss mixing hyper-params are set to $\{\beta_i\}_{i=1}^{3} = \{1, 5, 1\}$.

4.3. Evaluation metrics and Baseline Methods

As per standard practice, we use Structural Similarity Index (SSIM) [27], Perceptual Distance (LPIPS) [31] and

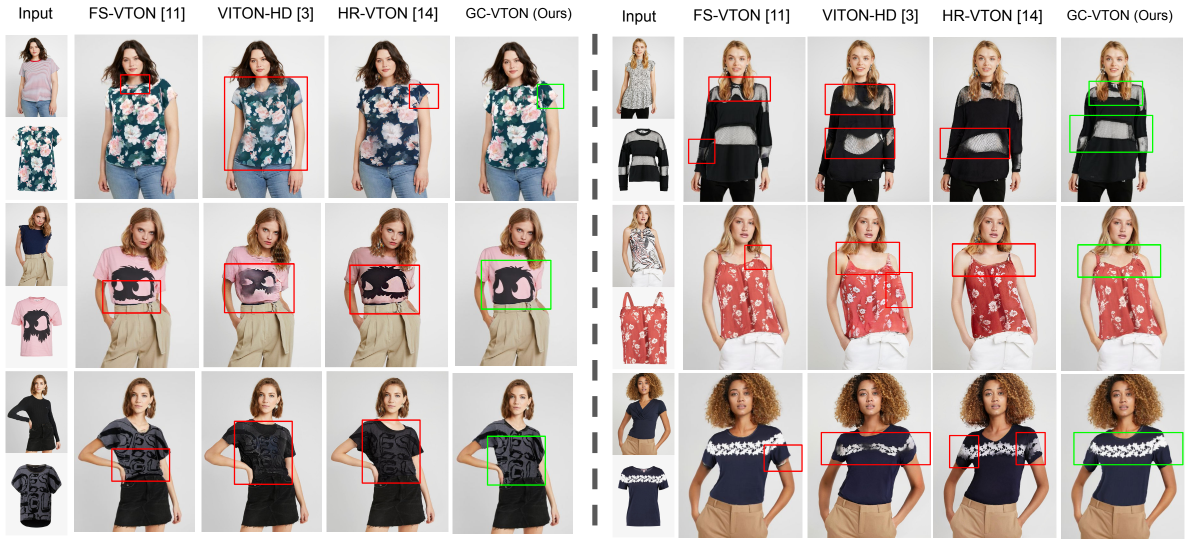

Figure 5. Qualitative comparison with other benchmarks. GC-VTON (ours) is able to produce natural looking and artifacts-free try-on images whereas other methods struggle in areas such as texture preservation, texture sharpness, preserving texture linearity and protection against squeezing in tucked-in style. Red boxes show errors in other method and green indicate corrections by our model.

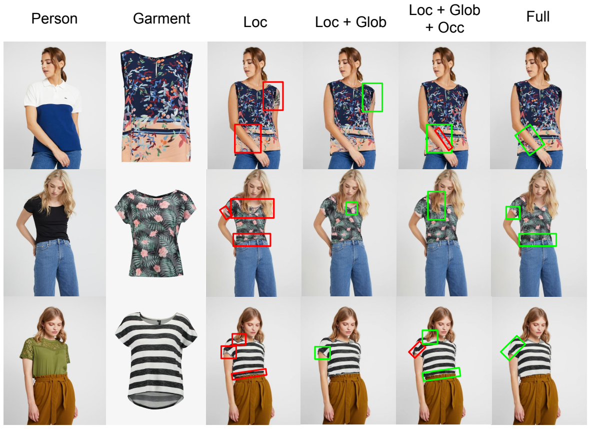

Figure 6. Qualitative comparison of different components of our pipeline. Green boxes show progressive improvement as each component is added.

Fréchet Inception Distance (FID) [19] to evaluate the quality of the generated images. SSIM and LPIPS are employed in a paired fashion whereas the FID is used in unpaired settings. Additionally, we also conduct a human study where we show 20 random images from each baseline method to each participant in a 10 person pool. The participants rate the images produced by each method in order of realism and we report the percentage of the preference for each method.

We compare our results to the current SOTA methods including FS-VTON [11], HR-VTON [14], and VITON-HD [3]. For all the methods we directly use their official codes to generate try-on images. For FS-VTON [11], we train their network from scratch as they don't have results available on VITON-HD dataset. We re-calculate the scores for each method using the obtained results.

Method	SSIM↑	FID↓	LPIPS↓
Loc	0.8729	8.236	0.1032
Loc + Glob	0.8754	8.168	0.0912
Loc + Glob + Occ	0.882	8.003	0.0899
GC-VTON (Full)	**0.887**	**7.888**	**0.0831**

Table 2. Quantitative comparison of all the components of our method.

4.4. Results

The quantitative results of our model are given in Table 1. Our GC-VTON clearly outperforms all the previous works on all the evaluation metrics. This confirms that our method consistently produces realistic try-on images compared to the baselines. Additionally, human evaluators also prefer the output of our method 54.3% of the times when presented side-by-side with other methods. A detailed qualitative comparison of GC-VTON (ours) against the existing works is presented in Figure 5. Existing models run into various artifacts in the generated images like missing key details from original garment (Lr1c4, Rr2c2, Rr2c3, Rr2c4) $\{Lr2c2 = Left\ Row2\ Col2\}$, unrealistic hair fusion (Lr2c3, Rr1c3), failing to preserve garment texture (Lr2c3, Lr3c3, Lr3c4, Rr3c3), unable to preserve texture around occlusions from body parts (Rr1c2), texture squeezing and stretching (Lr2c3, Lr3c2, Lr3c3), misalignment in the sleeves texture (Lr1c4, Rr3c4) and failing to maintain linearity of the texture (Rr1c2, Rr2c3, Rr1c4). In contrast our model does not suffer from these issues. We have to thank Local-Global disentanglement for aligning global boundaries and preserving local texture (Lr1c5, Lr3c5). Occlusion handling through visibility mask efficiently handles artifacts around occlusions (Rr1c5). Employing NIPR brings texture linearity preservation (Rr1c5) and preventing garments from extreme squeezing in tucked-in style (Lr2c5, Lr3c5). Additional results can be seen in supplementary materials.

4.5. Ablation Studies

In this section we present a qualitative (Figure 6) and quantitative analysis (Table 2) of our method and show how each component of our method plays an integral part in obtaining realistic try-on results. Multiple experiments are conducted to progressively add and evaluate the components. Here $\{\text{Loc}\}$ refers to using LocalNet only, $\{\text{Loc} + \text{Glob}\}$ refers to LocalNet + GlobalNet, $\{\text{Loc} + \text{Glob} + \text{Occ}\}$ represents LocalNet + GlobalNet + Occlusion Handling and finally $\{\text{Full}\}$ refers to full GC-VTON network with NIPR loss. In Figure 6, red boxes show errors and green boxes show how addition of each component helped mitigate the errors one by one.

Use of global consistency: demonstrates improvements in the SSIM ($0.8729 \rightarrow 0.8754$), a drop in FID ($8.236 \rightarrow 8.168$) and LPIPS ($0.1032 \rightarrow 0.0912$) as shown in Tab 2. Qualitative evidence of GlobalNet's positive role is shown in Fig 6 col-4, where in the row-1 the global alignment of left shoulder is corrected. Row-2 indicates the neck region misalignment in $\{\text{Loc}\}$ which is effectively resolved by the global consistency. In row-3, adding GlobalNet helped fix the right sleeve misalignment. The effective resolution of global boundary alignment at multiple garment locations is indicative of the strong abilities of the global consistency loss.

Occlusion Handling: Consistent improvements in SSIM, LPIPS and FID scores are observed when occlusion is explicitly handled. As indicated previously, handling all possible occlusion sources is vital for a distortion-free realistic warp. Results in Fig 6 col-5 are a visual confirmation that our occlusion handling mitigates occlusions induced by various sources. Specifically in row-1, distortions caused by hands are mitigated. In row-2 occlusion by hair is handled while in row-3 the squeezing due to tucked-in style (occlusion by another garment) is handled.

NIPR: is an integral component of our pipeline and quantitative results in Table 2 confirm the perceptual improvements it brings. Visual analysis in Fig 6 column 6 confirms the intuition behind the loss. In row-1 and row2, the left-over squeezing around occluded regions is alleviated. In row-3, the linearity of the the lines on the right sleeve is preserved, confirming the co-linearity preservation of local neighborhoods.

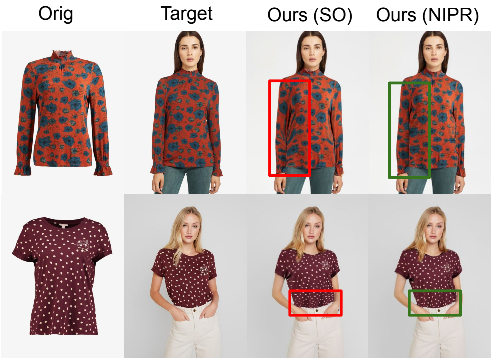

Figure 7. NIPR vs Second-Order constraint qualitative comparison. NIPR is a better guard against artifacts than the existing regularization losses. Please see supplementary for more results.

To dig a bit deeper, we train a model $\{\text{Ours(SO)}\}$ with all our components, except the NIPR is replaced with the second-order (SO) smoothness loss. In order to have access to a target warp, we compare the performance of this model in a paired setting to our model with NIPR. Fig 7 row-1 suggests that the model trained with Second Order constraint suffers from extreme texture stretching compared to the target warp. In the second row, the model with SO constraint suffers from extreme squeezing due to tucked-in style while NIPR effectively guards against the artifacts in both the cases. This is due to the fact that NIPR can identify the artifacts and appropriately assign penalties to the predicted flows. While the other losses only focus on local windows, NIPR is aware of global changes and tames the local neighborhoods to be in line with the global changes.

5. Conclusion

In this work, we present a novel method to generate artifacts-free and natural try-on images. For warping the garment via a flow field, we disentangle the global boundary alignment task from local texture preservation, thus allowing the network to equally focus on both. The outputs of the two tasks are matched via a consistency loss, thus harmonizing the local and global flows. To prevent artifacts around occluded regions, we predict a body-parts visibility mask, which masks out the occluded regions in the warped garment to prevent flows from creating distortions. Lastly, we propose a novel flow regularization loss NIPR that penalizes bad flows in local neighborhoods. NIPR applies a penalty that is carefully designed to appropriately tackle artifacts and ensure that local neighborhoods conform to global changes in the garment. Evaluation on VITON-HD dataset shows strong performance compared to the baselines both qualitatively and quantitatively.

References

[1] Shuai Bai, Huiling Zhou, Zhikang Li, Chang Zhou, and Hongxia Yang. Single stage virtual try-on via deformable attention flows. In *European Conference on Computer Vision*, pages 409–425. Springer, 2022. 1, 2

[2] Haibo Chen, Lei Zhao, Zhizhong Wang, Huiming Zhang, Zhiwen Zuo, Ailin Li, Wei Xing, and Dongming Lu. Dualast: Dual style-learning networks for artistic style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 872–881, June 2021. 1

[3] Seunghwan Choi, Sunghyun Park, Minsoo Lee, and Jaegul Choo. Viton-hd: High-resolution virtual try-on via misalignment-aware normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14131–14140, 2021. 1, 2, 3, 4, 6, 7

[4] Xin Dong, Fuwei Zhao, Zhenyu Xie, Xijin Zhang, Daniel K Du, Min Zheng, Xiang Long, Xiaodan Liang, and Jianchao Yang. Dressing in the wild by watching dance videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3480–3489, 2022. 1, 2, 3

[5] Chongjian Ge, Yibing Song, Yuying Ge, Han Yang, Wei Liu, and Ping Luo. Disentangled cycle consistency for highly-realistic virtual try-on. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16928–16937, 2021. 1, 2

[6] Yuying Ge, Yibing Song, Ruimao Zhang, Chongjian Ge, Wei Liu, and Ping Luo. Parser-free virtual try-on via distilling appearance flows. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8485–8493, 2021. 2, 3, 5, 6

[7] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 1

[8] Agus Gunawan, Soo Ye Kim, Hyeonjun Sim, Jae-Ho Lee, and Munchurl Kim. Modernizing old photos using multiple references via photorealistic style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12460–12469, June 2023. 1

[9] Xintong Han, Xiaojun Hu, Weilin Huang, and Matthew R Scott. Clothflow: A flow-based model for clothed person generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10471–10480, 2019. 1, 2, 3

[10] 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*, pages 7543–7552, 2018. 1, 2, 3

[11] Sen He, Yi-Zhe Song, and Tao Xiang. Style-based global appearance flow for virtual try-on. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3470–3479, 2022. 1, 2, 3, 4, 6, 7

[12] Thibaut Issenhuth, Jérémie Mary, and Clément Calauzenes. Do not mask what you do not need to mask: a parser-free virtual try-on. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*, pages 619–635. Springer, 2020. 1, 2, 3

[13] Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park. Scaling up gans for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10124–10134, 2023. 1

[14] Sangyun Lee, Gyojung Gu, Sunghyun Park, Seunghwan Choi, and Jaegul Choo. High-resolution virtual try-on with misalignment and occlusion-handled conditions. In *European Conference on Computer Vision*, pages 204–219. Springer, 2022. 1, 2, 4, 7

[15] Kedan Li, Min Jin Chong, Jeffrey Zhang, and Jingen Liu. Toward accurate and realistic outfits visualization with attention to details. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15546–15555, 2021. 1, 2

[16] Huan Ling, Karsten Kreis, Daiqing Li, Seung Wook Kim, Antonio Torralba, and Sanja Fidler. Editgan: High-precision semantic image editing. *Advances in Neural Information Processing Systems*, 34:16331–16345, 2021. 1

[17] Jiaming Liu, Yu Sun, Xiaojian Xu, and Ulugbek S Kamilov. Image restoration using total variation regularized deep image prior. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7715–7719. Ieee, 2019. 2, 3

[18] Sahib Majithia, Sandeep N Parameswaran, Sadbhavana Babar, Vikram Garg, Astitva Srivastava, and Avinash Sharma. Robust 3d garment digitization from monocular 2d images for 3d virtual try-on systems. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3428–3438, 2022. 2

[19] Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On aliased resizing and surprising subtleties in gan evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11410–11420, 2022. 7

[20] Chaitanya Patel, Zhouyingcheng Liao, and Gerard Pons-Moll. Tailornet: Predicting clothing in 3d as a function of human pose, shape and garment style. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7365–7375, 2020. 2

[21] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 6

[22] Igor Santesteban, Nils Thuerey, Miguel A Otaduy, and Dan Casas. Self-supervised collision handling via generative 3d garment models for virtual try-on. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11763–11773, 2021. 2

[23] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing.

In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1

[24] Deqing Sun, Stefan Roth, and Michael J Black. A quantitative analysis of current practices in optical flow estimation and the principles behind them. *International Journal of Computer Vision*, 106:115–137, 2014. 3

[25] Bochao Wang, Huabin Zheng, Xiaodan Liang, Yimin Chen, Liang Lin, and Meng Yang. Toward characteristic-preserving image-based virtual try-on network. In *Proceedings of the European conference on computer vision (ECCV)*, pages 589–604, 2018. 1

[26] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1

[27] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 6

[28] Fan Yang and Guosheng Lin. Ct-net: Complementary transfering network for garment transfer with arbitrary geometric changes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9899–9908, 2021. 1

[29] 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*, pages 7850–7859, 2020. 1, 2, 3

[30] Bowen Zhang, Shuyang Gu, Bo Zhang, Jianmin Bao, Dong Chen, Fang Wen, Yong Wang, and Baining Guo. Styleswin: Transformer-based gan for high-resolution image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11304–11314, 2022. 1

[31] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 6

[32] Fuwei Zhao, Zhenyu Xie, Michael Kampffmeyer, Haoye Dong, Songfang Han, Tianxiang Zheng, Tao Zhang, and Xiaodan Liang. M3d-vton: A monocular-to-3d virtual try-on network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13239–13249, 2021. 2

[33] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 1

[34] Zhen Zhu, Tengteng Huang, Baoguang Shi, Miao Yu, Bofei Wang, and Xiang Bai. Progressive pose attention transfer for person image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 1