Liquid Warping GAN: An Exploratory Analysis

Puneet Gupta
University of Toronto
puneet@cs.toronto.edu

Ronak Patel
University of Toronto
rvp5281@cs.toronto.edu

Rahul Shekhawat
University of Toronto
rahul170492@cs.toronto.edu

Abstract

Liquid Warping Generative Adversarial Network (LWGAN) is a method that has been recently introduced to tackle human motion imitation, appearance transfer, and novel view synthesis. In their paper the method is shown to work very well on the Impersonator (iPER) dataset. As it is a relatively recent approach to motion imitation, no work has been done on understanding why the model performs well, its limitations and how it can be improved upon. In this paper, we run experiments on the model to determine the advantages as well as the limitations of the Liquid Warping GAN model. We update the hyperparameters, and incorporate state of the art stabilizing methods in an effort to address the limitations of the model as well as understand why the changes lead to a more robust model.

1 Introduction

Human image synthesis has applications in the domains of virtual cinematography, computer and video games and covert disinformation attacks. Human motion imitation is the ability to copy the poses from a video and the texture of the human from an image and generate a video. The only requirements for this would be an image of the person from the desired target output, and a source video of the action being performed. Recent advances in GAN have allowed Human Motion Imitation to be implemented using Liquid Warping GAN [11].

Liquid Warping GAN (LWGAN) [11] is a unified GAN-based framework that consists of three training modules: Body Mesh Recovery, Flow Composition, and Liquid Warping GAN. Once trained, the model is able to handle Human Motion Imitation, Appearance Transfer and Novel View Synthesis. The aim of this paper is to analyze the robustness of the current architecture, identify the strengths and limitations, and attempt to address some of the limitations through empirical exploration. For the purposes of this paper, we focus on the Human Motion Imitation when performing experiments and qualitative analysis.

The LWGAN model architecture (Figure 1) is shown to perform well on the Impersonator (iPER) dataset [11]. When other motion imitation models such as PG2 [12], SHUP [2], and DSC [20] are trained on the same dataset, LWGAN is shown to outperform them [11]. In this paper, we discuss various state-of-the-art methods that could bring an improvement in training the LWGAN model. While discussing the different methods, we provide hypotheses on what improvement the methods can bring to LWGAN. We then implement the aforementioned methods, analyze the results (both qualitative and quantitative) and propose modifications for the LWGAN model architecture.

2 Related Work

Human image synthesis can be viewed as a sequence of tasks including human motion imitation, and appearance transfer. Today, most methods use Conditioned GANs (CGANS) or Variational Auto-Encoders for human motion imitation. The aim is to combine target images with a source pose (2D key-points) to generate realistic output [12, 19]. Densepose, by Güler et al. [14] was another major development and replaced the sparse 2D key-points with the dense correspondences, However, these approaches need individually trained models for mapping from 2D pose (or parts) to an image of each person and thus, are not scalable.

In appearance transfer, typically, a detailed 3D human mesh is estimated, using graphics-based methods, after which the human appearance with clothes can be transferred from one person to

Preprint. Under review.

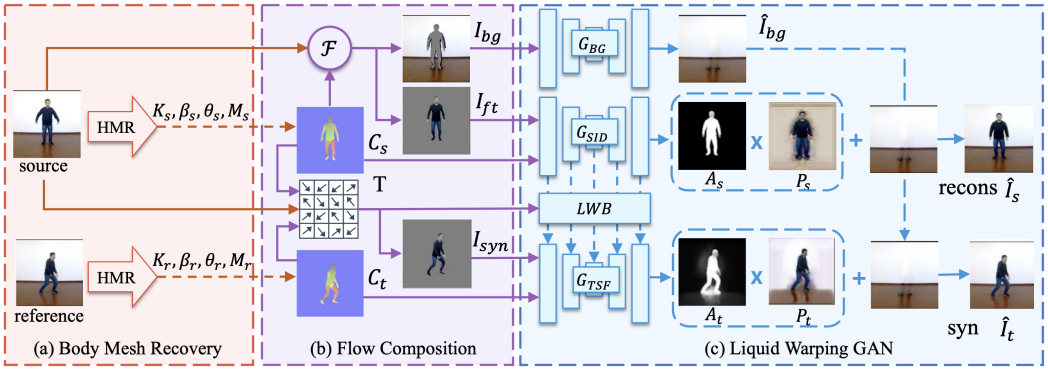

Figure 1: Training pipeline for LWGAN [11]. A source image I_s and reference image I_r are input into **(a)** The BMR module estimates the 3D pose. **(b)** The transformation flow T from the source to reference is calculated and the image is transformed. **(c)** The three generator architecture separately generates the background $\hat{I}_{bg}$, reconstructed source $\hat{I}_s$, and synthetic target images $\hat{I}_t$ under the reference condition by L^G_{tsf}.

another based on the 3D mesh. However, these models are computationally intensive despite their high-fidelity results. Raj et al. introduced SwapNet [15], a framework for single-view garment transfer in unconstrained images without solving the 3D reconstruction problem. However, SwapNet fails to handle large pose changes between source and target images. Zanfir et al. [21] proposed another solution aimed at automatic appearance modelling with minimal images and achieved a higher inception score [18]. However, the approach fails if a body part, such as a hand, is occluding the body shape.

Liu et al. proposed the LWGAN model architecture (Figure 1) for human image synthesis to tackle the limitation of the approaches discussed above [11]. The proposed architecture has three stages: body mesh recovery, flow composition, and Liquid Warping Block (LWB). Unlike previous approaches, which use 2D keypoints, the authors leverage Human Mesh Recovery (HMR) [9] for pose and shape estimation. In the flow composition module, a Neural Mesh Renderer (NMR) [7] is used to render source and reference correspondence maps. The paper proposes a novel Liquid Warping Block, that preserves source information such as texture, style, and colour that links the source with target streams. Thus, the output images synthesized are more realistic. The generator modules have similar architectures, names ResUnet, a combination of ResNet [5] and U-Net [16], without sharing parameters. The final synthesized output is between -1 and 1. The discriminator has the Pix2Pix [8] architecture.

3 Discussing Experiments and Results

Since the introduction of GANs by Goodfellow et al. in 2014, extensive research has been performed on how to improve their training. This section discusses the experiments performed on LWGAN and compares its performance to baseline. The iPER dataset [11] is used as training data for the experiments. The iPER dataset contains thirty subjects with differing shapes, height, weight, and gender. The subjects are wearing different clothes and they perform a 360°rotation in an A-pose. The dataset also contains a video for each subject with the subject performing random actions.

3.1 One-Sided Label Smoothing

During GAN training, if the discriminator depends on a small set of features, the generator may exploit it by producing only those features. The optimization may turn too greedy and generalize poorly. One-Sided Label smoothing aims to overcome this problem by penalizing the discriminator when it predicts above 0.9, reducing its propensity for overconfident predictions [3] and thus, improves generated image quality [18].

With the LWGAN, the generator losses decreased but image quality remained the same. Lower generator losses were expected, for the reasons stated above. The stagnation in image synthesis quality is due to label smoothing simply not being a powerful enough tool for knowledge distillation through the complex three generator architecture.

3.2 Gradient Penalty

GAN training has been known to be notoriously unstable [4]. Gradient Penalty (GP), an alternative to weight clipping, tackles the issues by penalizing the norm of the discriminator's gradient with

respect to its input. GP achieves stable training on a myriad of GAN architectures with almost no hyperparameter tuning [4]. GP implementation also improved training speed and sample quality over the weight clipping method. We believe that applying GP would prevent exploding gradients and lead to more stable training.

GP with weight of 10 led to better overall stability and fewer fluctuations for all of the losses. The transfer stream loss (L^G_{tsf}) and facial loss (L^G_{face}) for the generator went down by 15% and 13% respectively. Another experiment was performed with GP weight of 2 to see its effect on overall results. While the losses were approximately the same, weight of 2 had fewer fluctuations overall. For both, the mean discriminator output for real and fake images dropped to zero after a few epochs and did not change. Texture and style features weren't well preserved in the output images, compared to baseline.

3.3 Spectral Normalization

Spectral Normalization (SN) tackles exploding gradients or mode collapse to stabilize discriminator training [13]. Every weight is normalized by the largest singular value of the weight matrix. As the weights change slowly, only a single power iteration needs to be performed for each step of learning. Thus, SN is less computationally expensive. It achieves better or comparative inception scores than techniques such as weight normalization [17], weight clipping [1], and GP [4]. Unlike GP, SN imposes global regularization on the discriminator over local regularization[13]. SN has shown to be capable of generating images of better or at least equal quality compared to other training stabilization techniques.

Based on the above reasons, we hypothesized more stable training with less oscillations. This was confirmed by the loss curves (Appendix A4). The discriminator losses converged on 0.69 for real images, displaying a more balanced outlook, and not of one adversary overpowering the other. Upon inference, the colour, style, and shape of the clothing transferred well. This can be attributed to the SN method stabilizing training to avoid mode collapse, and allowing for more unique generalizations.

3.4 Two Time-Scale Update Rule (TTUR)

Another known issue for GANs is that they may not converge, TTUR would solve this issue as it has been proved to converge to a local Nash Equilibrium [6]. Different learning rates are used for the generator and discriminator to limit a network getting too far ahead of the other. When used in conjunction with Adam [10], mode collapse is avoided and Fréchet Inception scores are better than conventional training methods.

For the first experiment, we used a η_G =1e-04 and η_D =4e-04. As TTUR is a stabilizing agent, we increased our discriminator learning rate to take advantage of it. L^G_{tsf} and L^G_{face} went down by 22.5% and 3% respectively. At inference, the output was worse compared to baseline. The four times higher learning rate had lead to discriminator learning too quickly, and becoming to powerful for the generator.

Since the output quality was observed to be subpar, we realized the generator needs to be improved. For the next experiment, we used a η_G =4e-04 and η_D =1e-04. L^G_{tsf} and L^G_{face} were the same as the baseline but the training was more stable. Appearance transfer was slightly better, the texture of clothing was captured correctly. However, the detailed features such as creases and gradient colours of clothing were not captured. With such a high learning rate in comparison to the discriminator, the generator did not learn the variant granularity to fool its adversary.

3.5 Hyperparameter Tuning

After running the above experiments, we felt that the qualitative results could still be improved upon. The LWGAN uses eight weighted losses with each one having an associated hyperparameter λ_i.

For our first experiment, we increased the weight associated with L^G_{rec} by 50% (default=10). However, the loss values did not change drastically compared to baseline. The qualitative results at inference showed improvement: the model generalized better to the shape of the source image and retained intricate and minute features. This is on track with what we were expected, as increasing the λ_{rec}, increases the importance of reconstruction by penalizing deviation more.

Leading on from that, and the base displaying blurry facial features, for our second experiment, we increased the weight associated with L^G_{face} by a factor of 2 (default=5). Our results were not as satisfactory as the former experiment. The facial loss has increased, in line with an increase of λ_{face},

but no benefit was seen at inference. In fact, the focus on L^G_{face}, overpowered other losses, leading to a reduced image quality and articles across the board.

3.6 Ensemble Approaches

GANs are used for image synthesis and are notoriously unstable to train. We tailored the final experiments by selecting techniques that showed promising results from previous experiments. We hypothesized that these combinations will further stabilize the LWGAN training and improve generated image quality.

- $\boldsymbol{H_{E_1}}$ **:** Higher η_D leads to faster learning of the discriminator, but the GP will prevent it from overpowering the generator and allow for more stable training.
- $\boldsymbol{H_{E_2}}$ **:** Higher η_G leads to faster training of the generator. Learning will continue into the later epoch, but the higher decay rate over fewer epochs balances it with the discriminator.
- $\boldsymbol{H_{E_3}}$ **:** Higher η_G would leads to faster training of the generator and SN would ensure training is more stable with fewer oscillations. The higher λ_{rec} leads to better output at inference.
- $\boldsymbol{H_{E_4}}$ **:** Higher η_G would leads to faster training of the generator but without a stabilizer a decrease in training stability is expected. The higher λ_{rec} leads to better output at inference.

Exp	**SN**	**GP Weight**	**TTUR** ($\boldsymbol{\eta_G}$, $\boldsymbol{\eta_D}$, $\boldsymbol{\eta^*_F}$)	**Loss Weights**	$\boldsymbol{\Delta L^G_{face}}$	$\boldsymbol{\Delta L^G_{tsf}}$	$\boldsymbol{\Delta L^G_{rec}}$
E_1	-	2	(1E-4, 3E-4, -)	$\lambda^{\#}_{rec} = 10$	-12.80%	-5.75%	-16.54%
E_2	-	-	(4E-4, 1E-4, 1E-5)	$\lambda^{\#}_{rec} = 10$	-2.03%	-2.15%	-8.66%
E_3	✓	-	(4E-4, 1E-4, 1E-5)	$\lambda_{rec} = 15$	0.93%	-1.25%	32.28%
E_4	-	-	(4E-4, 1E-4, 1E-5)	$\lambda_{rec} = 15$	-3.37%	-2.49%	17.32%

Table 1: Overview of Ensemble Experiments and Results
* final learning rate with linear decay over 15 epochs; # default value of $\lambda_{rec} = 10$

Our E_1 loss curves (Appendix A7) confirmed our hypotheses, generator losses also dropped. The higher η_D balanced by the stabilizing GP, allowed for a smooth training process with less oscillations. The qualitative results were also positive, with great visual accuracy for the clothing style transfer. Though the 67% smaller η_G meant the generator did not learn enough, and thus produced grainy output.

E_2 oscillated more, without GP, but the inference output confirmed our hypothesis. The higher η_G and η_F allowed the generator to learn finer details. The model was able to produce relatively accurate hair, facial structure, and transfer of the clothing down to the button. The drawback was the edge detection on this model, with boundaries on the image, often hazy.

In E_3 training was relatively stable with fewer oscillations compared to E_2, which is expected. The higher η_G and η_F along with stabilizing Spectral Norm led to a smooth learning process. 32.28% increase in reconstruction loss can be attributed to the 50% increase in λ_{rec} offset by the improvement in reconstruction. The qualitative results were also positive, with great visual accuracy for the facial features and clothing style transfer, including intricate features such as buttons.

Training for E_4 is unstable with large deviation from E_3 loss curves, which was expected due to absence of a stabilizer. However, the model performs better than baseline and is able to learn facial features but performs poorly on clothing style transfer compared E_4.

4 Conclusion

In conclusion incorporating SN and TTUR stabilized the LWGAN training process, and improved the inference quality. Combined with an increased λ^G_{rec}, prioritizing reconstruction, inference quality improved as well. That is why we recommend an approach that applies SN to the convolution layers in the discriminator, uses TTUR that enhances the generator learning with a faster but shorter decay sequence to balance it with the discriminator, and prioritizing reconstruction loss.

While our recommendations do improve inference, there are still failure modes (Appendix A7.3.2). The detail to articulate hands is not present and the facial reconstruction lacks 3D photo-realism, it puts a generated texture map on a human model. The model uses a pre-trained SphereFaceNet50, and it would be valuable to explore other options. Moving forward using Fréchet Inception Distance (FID) to quantitatively measure the inference would improve on our qualitative analysis.

References

[1] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein gan, 2017.

[2] G. Balakrishnan, A. Zhao, A. V. Dalca, F. Durand, and J. Guttag. Synthesizing images of humans in unseen poses, 2018.

[3] I. Goodfellow. Nips 2016 tutorial: Generative adversarial networks, 2016.

[4] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville. Improved training of wasserstein gans, 2017.

[5] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. URL `http://arxiv.org/abs/1512.03385`.

[6] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium, 2017.

[7] Y. U. Hiroharu Kato and T. Harada. Neural 3d mesh renderer. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[8] P. Isola, J. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. *CoRR*, abs/1611.07004, 2016. URL `http://arxiv.org/abs/1611.07004`.

[9] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik. End-to-end recovery of human shape and pose. *CoRR*, abs/1712.06584, 2017. URL `http://arxiv.org/abs/1712.06584`.

[10] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization, 2014.

[11] W. Liu, Z. Piao, J. Min, W. Luo, L. Ma, and S. Gao. Liquid warping gan: A unified framework for human motion imitation, appearance transfer and novel view synthesis, 2019.

[12] L. Ma, X. Jia, Q. Sun, B. Schiele, T. Tuytelaars, and L. V. Gool. Pose guided person image generation, 2017.

[13] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida. Spectral normalization for generative adversarial networks, 2018.

[14] N. Neverova, R. A. Güler, and I. Kokkinos. Dense pose transfer. *CoRR*, abs/1809.01995, 2018. URL `http://arxiv.org/abs/1809.01995`.

[15] A. Raj, P. Sangkloy, H. Chang, J. Lu, D. Ceylan, and J. Hays. Swapnet: Garment transfer in single view images. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[16] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597, 2015. URL `http://arxiv.org/abs/1505.04597`.

[17] T. Salimans and D. P. Kingma. Weight normalization: A simple reparameterization to accelerate training of deep neural networks, 2016.

[18] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training gans, 2016.

[19] C. Si, W. Wang, L. Wang, and T. Tan. Multistage adversarial losses for pose-based human image synthesis. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 118–126, 2018.

[20] A. Siarohin, E. Sangineto, S. Lathuiliere, and N. Sebe. Deformable gans for pose-based human image generation, 2017.

[21] M. Zanfir, A.-I. Popa, A. Zanfir, and C. Sminchisescu. Human appearance transfer. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.