샘플링과 정규화를 통한 단일 이미지로부터 3D 인물 텍스처 생성*

차시헌⁰, 서광균, Amirsaman Ashtari, 노준용
KAIST Visual Media Lab
{chacorp, skg1023, a.s.ahtari, junyongnoh}@kaist.ac.kr

Generating 3D Human Texture from a Single Image with Sampling and Refinement

Sihun Cha⁰, Kwanggyoon Seo, Amirsaman Ashtari, Junyong Noh
KAIST Visual Media Lab

Abstract

Generating the texture map for a 3D human mesh from a single image is challenging. To generate a plausible texture map, the invisible part of the texture needs to be synthesized with relevance to the visible part and the texture should semantically align to the UV space of the template mesh. To overcome such challenges, we propose a novel method that incorporates *SamplerNet* and *RefinerNet*. *SamplerNet* predicts a sampling grid that enables sampling from the given visible texture information, and *RefinerNet* refines the sampled texture to maintain spatial alignment.

1. Introduction

Generating a 3D human avatar from a single image is attracting a lot of attention with the growing popularity of augmented and VR/AR applications. There have been numerous methods proposed to predict a 3D mesh from a single image [1,2]. While the methods can reconstruct the shapes and poses of the 3D mesh, few has studied for estimating a texture map from a single image [3].

Previous method [3] uses convolutional neural network (CNN) for generating texture map from a single image. However, the method often produced blurry texture and unable to synthesis the textural pattern of the cloth, where the parts are invisible in the source image. This is a critical because many clothes have repeated patterns.

In this paper, we propose a method to generate a texture map that synthesizes the textural pattern in the invisible parts of the source image. Our method consists of two CNNs: *SamplerNet* and *RefinerNet*. *SamplerNet* generates a sampling grid to fill the invisible part of the input partial texture map by referring to the pixels of the visible part. Given the sampled texture map from the sampling grid, *RefinerNet* generates a refinement texture map and a mask. The final texture map is produced by alpha blending the sampled and refinement texture map with the predicted mask. We compared our method with a previous method and show that our method outperforms qualitatively and quantitatively.

2. Method

Our goal is to produce the texture map for a 3D human mesh from a single image. We map the human appearance in the source image to the UV space of the template mesh to produce the partial texture map. Given the partial texture map, our method generates a texture map through two steps.

Sampler Network. *SamplerNet* produces a sampled texture map, T_{sample}, where invisible parts are synthesized by referring to visible parts in the partial texture map. *SamplerNet* is based on U-Net architecture, in which all layers are residual blocks except for the first and the last layer. *SamplerNet* is trained with the following loss terms:

$$\mathcal{L}_{\text{SamplerNet}} = \lambda_{\text{L1}}\mathcal{L}_{\text{L1}} + \lambda_{\text{LPIPS}}\mathcal{L}_{\text{LPIPS}}.$$

$\mathcal{L}_{L1}$ and $\mathcal{L}_{LPIPS}$ are L1 loss and perceptual loss,

* 포스터 발표논문
* 본 논문은 포스터 (Poster) 로서, 본 논문의 원본 논문은 Special Interest Group on Computer Graphics and Interactive Techniques Conference Posters, 2022에 게재확정 되었음.
* 본 논문은 2022년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.2020-0-00450, AR 실감 콘텐츠 제작을 위한 심층학습 기반 단일 이미지의 입체적 움직임 자동 생성 및 대화형 증강 기술 개발)

respectively. The losses are calculated between T_{sample} and the ground truth texture map. λ_{L1} and λ_{LPIPS} are set to 10 and 1, respectively.

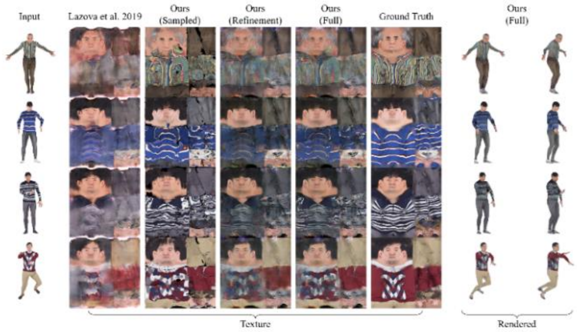

Figure 1: Visual comparison with the previous method.

Refine Network. Although DensePose robustly predicts the correspondence between the 2D image and the UV of the 3D mesh, there remains a prediction error, producing a misaligned texture map. As *SamplerNet* only considers the invisible part of the partial texture map, the visible part remains intact, maintaining misalignment. To alleviate this issue, we use *RefinerNet* that produces a refinement texture map T_{refine} and blending mask M. The final output is computed as follows:

$$T_{final} = T_{sample} \odot M + T_{refine} \odot (1 - M),$$

where $\odot$ is hadamard product.

The architecture of *RefinerNet* is also based on U-Net, with 9 residual blocks for the bottleneck. *RefinerNet* is trained with the following loss terms:

$$\mathcal{L}_{\mathrm{RefinerNet}} = \lambda_{\mathrm{L1}}\mathcal{L}_{\mathrm{L1}} + \lambda_{\mathrm{VGG}}\mathcal{L}_{\mathrm{VGG}} + \lambda_{\mathrm{FM}}\mathcal{L}_{\mathrm{FM}} + \lambda_{\mathrm{GAN}}\mathcal{L}_{\mathrm{GAN}},$$

where $\mathcal{L}_{\mathrm{VGG}}$ and $\mathcal{L}_{\mathrm{FM}}$ measure the perceptual difference using the VGG-19 network and a discriminator, respectively. $\mathcal{L}_{\mathrm{GAN}}$ is an adversarial loss. λ_{L1}, λ_{VGG}, λ_{FM} and λ_{GAN} are set to 10, 10, 10, and 1, respectively. All the losses are computed between the ground truth texture map and T_{final}.

3. Experiments

Training Details. We used a total of 1,229 texture maps for training and 212 texture maps for testing, which are in the size of 256×256. We used the Adam optimizer with a learning rate of 0.0002 and beta parameters (0.9, 0.999). The batch size is set to 8, and each model was trained separately for 30,000 iterations on a NVIDIA GTX 1080 Ti GPU.

Evaluation. To evaluate our method, we performed both qualitative and quantitative evaluations. For the quantitative evaluation, we use the Structural Similarity Index Measure (SSIM), Peak Signal-to-Noise Ratio (PSNR), and LPIPS to measure the quality of the generated texture map. We compared the intermediate and final outputs of our method to the final output of Lazova et al. [3].

As shown in Table.1, our method outperforms in all metrics. We can observe that the quality of the final texture map is improved by adding each network. The visual results of the generated textures as shown in Figure.1. The results from Lazova et al. lack detailed patterns and are generally blurry. Our intermediate result T_{refine} better preserve the details of the input image compared to those from Lazova et al. However, the patterns generated in invisible parts are still blurry. Our final results show significant improvement compared to those from Lazova et al. and T_{refine}.

Table 1: Quantitative Comparison.

Method	LPIPS↓	PSNR↑	SSIM↑
[Lazova et al. 2019]	0.2454	17.6744	0.5593
Ours (sampled)	0.2305	17.4234	0.5787
Ours (refinement)	0.2258	17.8700	0.5993
Ours (full)	**0.2061**	**18.1052**	**0.6032**

Discussion. In this work, we proposed a novel method for generating a 3D human texture map from a single image. Compared to the previous approach, our method generates texture maps with improved quality by synthesizing textural patterns in invisible parts. Our method relies on visible part for the texture synthesis, and therefore it is unable to generate non-existing information. An interesting future research direction would incorporate a reference texture to fill out the missing texture regions.

References

[1] Thiemo Alldieck, Marcus Magnor, Bharat Lal Bhatnagar, Christian Theobalt, and Gerard Pons-Moll. Learning to reconstruct people in clothing from a single RGB camera. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 1175–1186, 2019.
[2] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In Proceedings of the IEEE conference on computer vision and pattern recognition. 7122–7131, 2018.
[3] Verica Lazova, Eldar Insafutdinov, and Gerard Pons-Moll. 360-degree textures of people in clothing from a single image. In 2019 International Conference on 3D Vision (3DV). IEEE, 643–653, 2019.