GAN-based 3D Manipulation of Car Models

(Masters or last year Engineering internship)

George Drettakis, GRAPHDECO, Inria Sophia Antipolis (France)
http://team.inria.fr/graphdeco

George.Drettakis@inria.fr
http://www-sop.inria.fr/members/George.Drettakis/

In collaboration with Thomas Leimkuehler of the Max-Planck Institut for Informatics:
https://people.mpi-inf.mpg.de/~tleimkue/

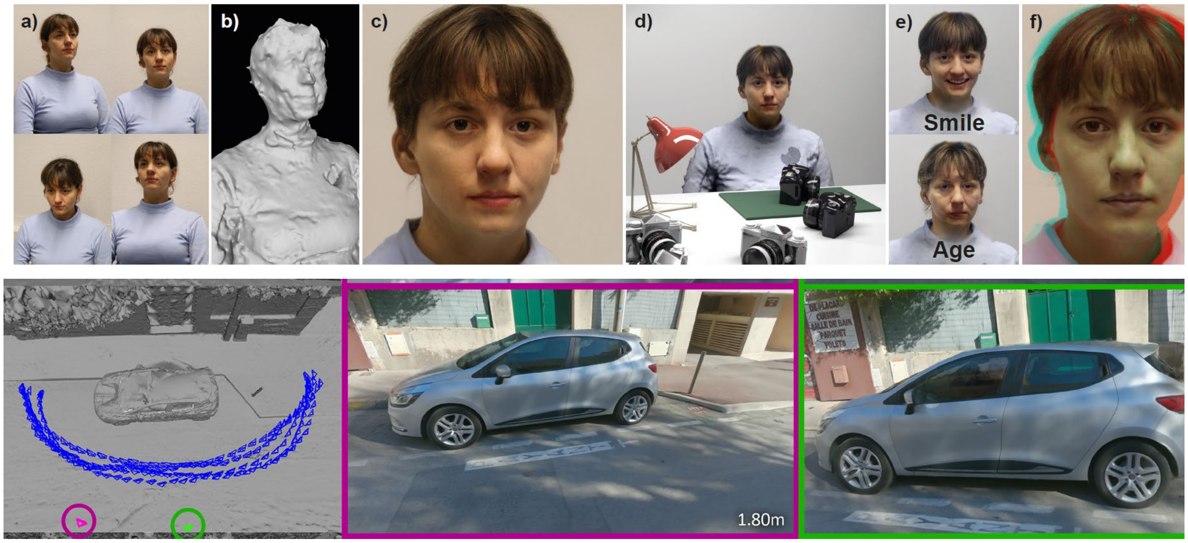

Figure 1: Top row: We will build on the results obtained in FreeStyleGAN [1] where multi-view images a) are used to provide a coarse geometric proxy b). We can render the closest view with StyleGAN to generate a very realistic image c). This allows insertion into 3D virtual scenes d) and semantic manipulation e) or stereo viewing f). In this internship we will extend these ideas to models of cars captured with multiple images (left), allowing free-viewpoint rendering (center and right)

Context and goal

FreeStyleGAN [1] extends the capabilities of StyleGAN [2, 3, 4] to enable free-viewpoint renderings of faces using multi-view data. A novel camera manifold formulation together with a multi-view embedding strategy [5, 6] for the first time allows GAN-generated images to be used in virtual environments and enables new applications like generative stereoscopic image synthesis, while preserving semantic editing capabilities, such as changes of lighting or facial expression [7, 8, 9]. In this internship we plan to extend these ideas to other classes of objects, and in particular cars.

Methodology

Our paper [1] focused on faces, while the underlying strategies can be adapted to a variety of different domains. A particularly promising domain are cars: The pre-trained model from the original StyleGAN papers based on the LSUN car dataset [10] appears to be a good starting point, but a joint GAN fine-tuning and embedding step might be necessary to achieve high-quality results. Multi-view data is available, but geometry reconstruction might have to be complemented by custom solutions [11] or guided by geometric priors [12] to account for errors arising from highly specular appearance, which violate typical photo-consistency assumptions. A new camera manifold needs to be derived [13], which accounts for full 360° views. Convincing integration of the generated images into synthetic or real scenes with moving cameras will require a dedicated solution for reflection synthesis using manipulations in latent space [14]. This can be complemented by localized activation map interventions [15, 16], likely guided by semantic labels [17], or GAN dissection [18]. More details and hands-on suggestions on all those points are given below.

Approach

We next describe the possible steps of the internship; the exact set of tasks will depend on the duration of the internship and the background of the student.

1. Projection

We will start by performing experiments to project a single car image onto the StyleGAN latent space using [5]. We will observe result quality and then explore semantic editing capabilities using [7], which should be exactly what is shown in GANSpace [14]. To facilitate this step, we will build on the datasets of car images available from [11].

2. Define Camera Manifold

Based on the above observations, we will have an initial idea what the camera manifold might look like. GANSpace [14] show considerable camera variations, including 360° views of the car. Therefore, the manifold will likely be a lot less restricted than the one for faces. We will use images with different viewing angles, and perform global transformations (e.g., translation, rotation, scaling) to the images before projecting onto the GAN and observe result quality to get a feeling for the boundaries of the camera manifold (for example, we would expect a car rotated 90° and shifted to the image corner to look bad after projection). Full control and dense viewpoint coverage can be obtained using synthetic data, but the background might heavily influence projection quality [1]. Based on all these observations, we will devise a first image alignment procedure. This will likely be guided by a car detector and/or semantic segmentation of the input images and can be complemented by an analysis akin to Sec. 4.3 in [1].

3. Generating Views

Once we have a manifold formulation and the corresponding alignment we will use it to project multi-view images onto the GAN using a parameterized embedding [1] and try to generate arbitrary manifold views. Training data generation requires a rough camera calibration and geometric proxy obtained using SfM/MVS, and it may be worth considering completely synthetic training data, even though it might be difficult to match the appearance of the multi-view dataset. Special care will be given to avoid problems with the background (e.g., masking it out as a first step).

4. Fine-Tuning

We will then experiment with fine-tuning the GAN to improve result quality: Instead of only finding latents, we will also fine-tune car-specific GAN weights, by enabling GAN weights to change when training.

5. Better Geometry

We will investigate how to improve 3D reconstruction [11, 12] to get a better geometry estimate. We will use the reconstruction to leave the camera manifold using warping [1]. At this point we should be able to synthesize free viewpoints of a car.

6. Reflections

Car reflections will likely not look convincing or coherent, especially when compositing the synthesized car into a 3D scene [19]. As reflections are encoded by the latents as well (see manipulations in [14]), it might be worth identifying corresponding subspaces in latent space via learning, with the goal of approximate control over the reflections. Training data could comprise synthetic images, maybe real-time on-the-fly reflection synthesis using simple environment mapping. We don't expect the reflections to be completely disentangled from other image properties (for example, changing a reflection might alter the size of the tires). To attack this problem, restricting the effect of latent manipulations to specific image regions [15,16,18] identified via semantic labels [11] might be worth considering.

Work environment and requirement

The internship will take place at Inria Sophia Antipolis in the GRAPHDECO group (http://team.inria.fr/graphdeco). Inria will provide a monthly stipend of around 1100 euros for EU citizens in their final year of masters, and 400 euros for other candidates.

Candidates should have strong programming and mathematical skills as well as knowledge in computer graphics, geometry processing and machine learning, with experience in C++, OpenGL and GLSL on the graphics side, and tensorflow/pytorch for learning.

References

[1] T. Leimkühler, G. Drettakis. FreeStyleGAN: Free-viewpoint Portrait Rendering with StyleGAN using Multi-view Images. ACM Trans. On Graphics. 2021. https://repo-sam.inria.fr/fungraph/freestylegan/
[2] T. Karras, S. Laine, T. Aila. A style-based generator architecture for generative adversarial networks. CVPR 2019.
[3] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, T. Aila. Analyzing and improving the image quality of StyleGAN. CVPR 2020.
[4] T. Karras, M. Aittala, J. Hellsten, S. Laine, J. Lehtinen, T. Aila. Training Generative Adversarial Networks with Limited Data. NeurIPS 2020.
[5] R. Abdal, Y. Qin, P. Wonka. Image2StyleGAN: How to Embed Images Into the StyleGAN Latent Space? ICCV 2019.
[6] A. Tewari, M. Elgharib, M. BR, F. Bernard, H.-P. Seidel, P. Pérez, M. Zöllhofer, C. Theobalt. PIE: Portrait Image Embedding for Semantic Control. Siggraph Asia 2020.
[7] E. Härkönen, A. Hertzmann, J. Lehtinen, S. Paris. 2020. GANSpace: Discovering Interpretable GAN Controls. NeurIPS 2020.
[8] R. Abdal, P. Zhu, N. Mitra, P. Wonka. StyleFlow: Attribute-conditioned exploration of StyleGAN-generated images using conditional continuous normalizing flows. ToG 2021.
[9] A. Tewari, M. Elgharib, G. Bharaj, F. Bernard, H.-P. Seidel, P. Pérez, M. Zöllhofer, C. Theobalt. StyleRig: Rigging StyleGAN for 3D Control over Portrait Images. CVPR 2020.
[10] F. Yu, Y. Zhang, S. Song, A. Seff, J. Xiao. LSUN: Construction of a large-scale image dataset using deep learning with humans in the loop. ArXiv 2015.

[11] S. Rodriguez, S. Prakash, P. Hedman, G. Drettakis. Image-Based Rendering of Cars using Semantic Labels and Approximate Reflection Flow. I3D 2020. https://repo-sam.inria.fr/fungraph/ibr-cars-semantic/
[12] K. Rematas, T. Ritschel, M. Fritz, T. Tuytelaars. Image-based Synthesis and Re-Synthesis of Viewpoints Guided by 3D Models. CVPR 2014.
[13] C. Lino, M. Christie. Intuitive and efficient camera control with the Toric space. ToG 2015.
[14] Supplemental video of [7]: https://youtu.be/jdTICDa_eAI
[15] D. Bau, H. Strobelt, W. Peebles, J. Wulff, Bolei. Zhou, J.-Y. Zhu, A. Torralba. Semantic Photo Manipulation with a Generative Image Prior. Siggraph 2019.
[16] R. Abdal, Y. Qin, P. Wonka. Image2StyleGAN++: How to Edit the Embedded Images? CVPR 2020.
[17] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. PAMI 2017.
[18] D. Bau, J.-Y. Zhu, H. Strobelt, B. Zhou, J. B. Tenenbaum, W. T. Freeman, A. Torralba. GAN Dissection: Visualizing and Understanding Generative Adversarial Networks. ICLR 2019.
[19] T. Bertel, Y. Tomoto, S. Rao, R. Ortiz-Cayon, S. Holzer, C. Richardt. Deferred neural rendering for view extrapolation. Siggraph Asia Poster 2020.
[20] Y. Zhang, W. Chen, H. Ling, J. Gao, Y. Zhang, A. Torralba, S. Fidler. Image GANs meet differentiable rendering for inverse graphics and interpretable 3D neural rendering. ICLR 2021.
[21] X. Pan, B. Dai, Z. Liu, C. C. Loy, P. Luo. Do 2D GANs know 3D shape? Unsupervised 3D shape reconstruction from 2D image GANs. ICLR 2021.
[22] V. Blanz, T. Vetter. A morphable model for the synthesis of 3D faces. Siggraph 1999.
[23] W. Smith, A. Seck, H. Dee, B. Tiddeman, J. B. Tenenbaum, B. Egger. A morphable face albedo model. CVPR 2020.
[24] B. Meier. Painterly Rendering for Animation. Siggraph 1996.
[25] R. Sayeed, T. Howard. State of the Art Non-Photorealistic Rendering (NPR) Techniques. TPCG 2006.
[26] P. Bénard, A. Lagae, P. Vangorp, S. Lefebvre, G. Drettakis, J. Thollot… A Dynamic Noise Primitive for Coherent Stylization. EGSR 2010.
[27] P. Bénard, A. Bousseau, J. Thollot. Dynamic solid textures for real-time coherent stylization. I3D 2009.