Influential Article Review- Participant Learning Through Functional Neural Models

Ray Lynch

Arturo Hansen

Dixie Stokes

This paper examines technology. We present insights from a highly influential paper. Here are the highlights from this paper: Several recent works have shown how highly realistic human head images can be obtained by training convolutional neural networks to generate them. In order to create a personalized talking head model, these works require training on a large dataset of images of a single person. However, in many practical scenarios, such personalized talking head models need to be learned from a few image views of a person, potentially even a single image. Here, we present a system with such few-shot capability. It performs lengthy meta-learning on a large dataset of videos, and after that is able to frame few- and one-shot learning of neural talking head models of previously unseen people as adversarial training problems with high capacity generators and discriminators. Crucially, the system is able to initialize the parameters of both the generator and the discriminator in a person-specific way, so that training can be based on just a few images and done quickly, despite the need to tune tens of millions of parameters. We show that such an approach is able to learn highly realistic and personalized talking head models of new people and even portrait paintings. For our overseas readers, we then present the insights from this paper in Spanish, French, Portuguese, and German.

SUMMARY

- We perform comparison with baselines in three different setups, with 1, 8 and 32 frames in the fine-tuning set 32 hold-out frames for each of the 50 test video sequences.
- We argue that this is intrinsic to the methods themselves X2Face uses L2 loss during optimization, which leads to a good SSIM score. On the other hand, Pix2pixHD maximizes only perceptual metric, without identity preservation loss, leading to minimization of FID, but has bigger identity mismatch, as seen from the CSIM column.
- Moreover, these metrics do not correlate well with human perception, since both of these methods produce uncanny valley artifacts, as can be seen from qualitative comparison Figure 3 and the user study results.
- We have presented a framework for meta-learning of adversarial generative models, which is able to train highly realistic virtual talking heads in the form of deep generator networks. Crucially, only a handful of photographs is needed to create a new model, whereas the model trained on 32 images achieves perfect realism and personalization score in our user study.

- Currently, the key limitations of our method are the mimics representation and the lack of landmark adaptation.
- Using landmarks from a different person leads to a noticeable personality mismatch. So, if one wants to create «fake» puppeteering videos without such mismatch, some landmark adaptation is needed.
- We note, however, that many applications do not require puppeteering a different person and instead only need the ability to drive one's own talking head. For such scenarios, our approach already provides a high-realism solution.

HIGHLY INFLUENTIAL ARTICLE

We used the following article as a basis of our evaluation:

Zakharov, E., Shysheya, A., Burkov, E., & Lempitsky, V. (2019). Few-shot adversarial learning of realistic neural talking head models. In Proceedings of the IEEE International Conference on Computer Vision (pp. 9459-9468).

This is the link to the publisher's website:
https://arxiv.org/pdf/1905.08233v1.pdf

INTRODUCTION

In this work, we consider the task of creating personalized photo realistic talking head models, i.e. systems that can synthesize plausible video-sequences of speech expressions and mimics of a particular individual. More specifically, we consider the problem of synthesizing photorealistic personalized head images given a set of face landmarks, which drive the animation of the model. Such ability has practical applications for telepresence, including videoconferencing and multi-player games, as well as special effects industry. Synthesizing realistic talking head sequences is known to be hard for two reasons. First, human heads have high photometric, geometric and kinematic complexity. This complexity stems not only from modeling faces (for which a large number of modeling approaches exist) but also from modeling mouth cavity, hair, and garments. The second complicating factor is the acuteness of the human visual system towards even minor mistakes in the appearance modeling of human heads (the so-called uncanny valley effect). Such low tolerance to modeling mistakes explains the current prevalence of non-photorealistic cartoon-like avatars in many practically-deployed teleconferencing systems. To overcome the challenges, several works have proposed to synthesize articulated head sequences by warping a single or multiple static frame. Both classical warping algorithms and warping fields synthesized using machine learning (including deep learning) can be used for such purposes. While warping-based systems can create talking head sequences from as little as a single image, the amount of motion, head rotation, and disocclusion that they can handle without noticeable artifacts is limited.

Direct (warping-free) synthesis of video frames using adversarially-trained deep convolutional networks (ConvNets) presents the new hope for photorealistic talking heads. Very recently, some remarkably realistic results have been demonstrated by such systems. However, to succeed, such methods have to train large networks, where both generator and discriminator have tens of millions of parameters for each talking head. These systems, therefore, require a several-minutes-long video or a large dataset of photographs as well as hours of GPU training in order to create a new personalized talking head model. While this effort is lower than the one required by systems that construct photo-realistic head models using sophisticated physical and optical modeling, it is still excessive for most practical telepresence scenarios, where we want to enable users to create their personalized head models with as little effort as possible.

In this work, we present a system for creating talking head models from a handful of photographs (so-called few shot learning) and with limited training time. In fact, our system can generate a reasonable result based on a single photograph (one-shot learning), while adding a few more photographs increases the

fidelity of personalization. Similarly, the talking heads created by our model are deep ConvNets that synthesize video frames in a direct manner by a sequence of convolutional operations rather than by warping. The talking heads created by our system can, therefore, handle a large variety of poses that goes beyond the abilities of warping-based systems.

The few-shot learning ability is obtained through extensive pre-training (meta-learning) on a large corpus of talking head videos corresponding to different speakers with diverse appearance. In the course of meta-learning, our system simulates few-shot learning tasks and learns to transform landmark positions into realistically-looking personalized photographs, given a small training set of images with this person. After that, a handful of photographs of a new person sets up a new adversarial learning problem with a high-capacity generator and discriminator pre-trained via meta-learning. The new adversarial problem converges to the state that generates realistic and personalized images after a few training steps.

In the experiments, we provide comparisons of talking heads created by our system with alternative neural talking head models via quantitative measurements and a user study, where our approach generates images of sufficient realism and personalization fidelity to deceive the study participants. We demonstrate several uses of our talking head models, including video synthesis using landmark tracks extracted from video sequences of the same person, as well as puppeteering (video synthesis of a certain person based on the face landmark tracks of a different person).

CONCLUSION

We have presented a framework for meta-learning of adversarial generative models, which is able to train highlyrealistic virtual talking heads in the form of deep generator networks. Crucially, only a handful of photographs (as little as one) is needed to create a new model, whereas the model trained on 32 images achieves perfect realism and personalization score in our user study (for 224p static images).

Currently, the key limitations of our method are the mimics representation (in particular, the current set of landmarks does not represent the gaze in any way) and the lack of landmark adaptation. Using landmarks from a different person leads to a noticeable personality mismatch. So, if one wants to create “fake” puppeteering videos without such mismatch, some landmark adaptation is needed. We note, however, that many applications do not require puppeteering a different person and instead only need the ability to drive one’s own talking head. For such scenario, our approach already provides a high-realism solution.

APPENDIX

FIGURE 1

FIGURE 2

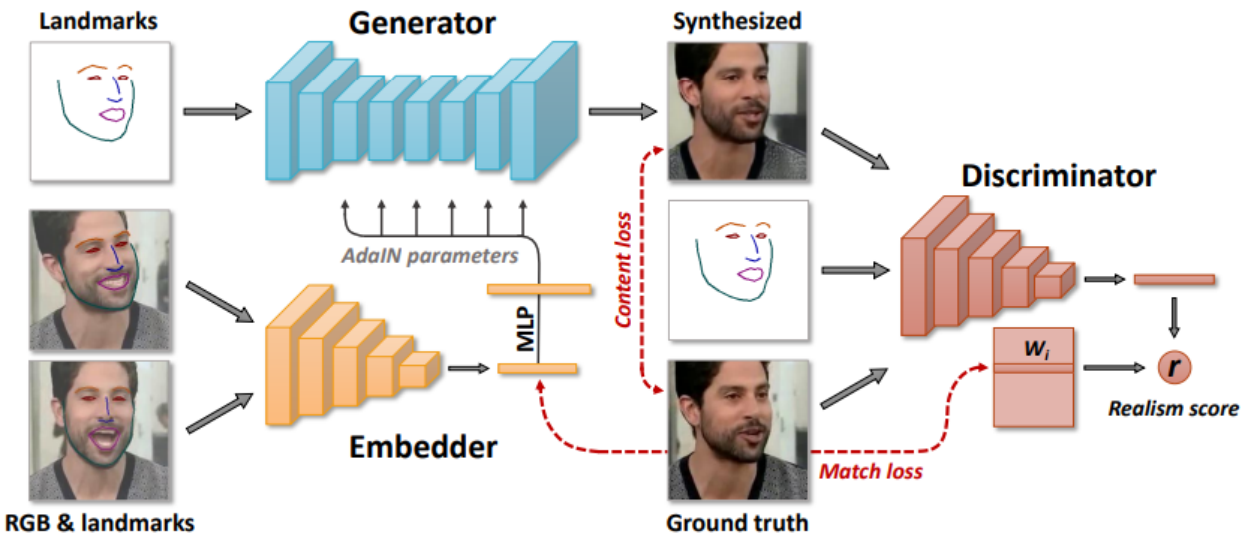


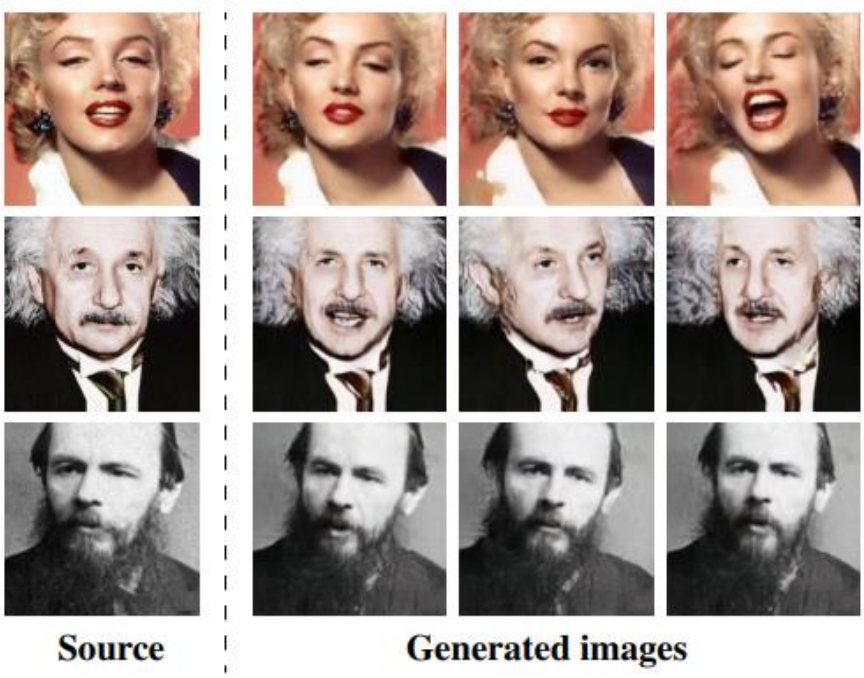
TABLE 1

Method (T)	FID↓	SSIM↑	CSIM↑	USER↓
VoxCeleb1				
X2Face (1)	45.8	**0.68**	**0.16**	0.82
Pix2pixHD (1)	**42.7**	0.56	0.09	0.82
Ours (1)	43.0	0.67	0.15	**0.62**
X2Face (8)	51.5	**0.73**	**0.17**	0.83
Pix2pixHD (8)	**35.1**	0.64	0.12	0.79
Ours (8)	38.0	0.71	**0.17**	**0.62**
X2Face (32)	56.5	**0.75**	0.18	0.85
Pix2pixHD (32)	**24.0**	0.70	0.16	0.71
Ours (32)	29.5	0.74	**0.19**	**0.61**
VoxCeleb2				
Ours-FF (1)	**46.1**	0.61	**0.42**	**0.43**
Ours-FT (1)	48.5	**0.64**	0.35	0.46
Ours-FF (8)	42.2	0.64	**0.47**	0.40
Ours-FT (8)	**42.2**	**0.68**	0.42	**0.39**
Ours-FF (32)	40.4	0.65	**0.48**	0.38
Ours-FT (32)	**30.6**	**0.72**	0.45	**0.33**

FIGURE 3

FIGURE 4

FIGURE 5



REFERENCES

A. Antoniou, A. J. Storkey, and H. Edwards. Augmenting image classifiers using data augmentation generative adversarial networks. In Artificial Neural Networks and Machine Learning - ICANN, pages 594–603, 2018. 2

A. Bulat and G. Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230, 000 3d facial landmarks). In IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22- 29, 2017, pages 1021–1030, 2017. 3

A. Nagrani, J. S. Chung, and A. Zisserman. Voxceleb: a large-scale speaker identification dataset. In INTERSPEECH, 2017. 5

C. Finn, P. Abbeel, and S. Levine. Model-agnostic metalearning for fast adaptation of deep networks. In Proc. ICML, pages 1126–1135, 2017. 2

C. Yin, J. Tang, Z. Xu, and Y. Wang. Adversarial metalearning. CoRR, abs/1806.03316, 2018. 2

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. CoRR, abs/1412.6980, 2014. 5

D. Ulyanov, A. Vedaldi, and V. S. Lempitsky. Instance normalization: The missing ingredient for fast stylization. CoRR, abs/1607.08022, 2016. 5

H. Averbuch-Elor, D. Cohen-Or, J. Kopf, and M. F. Cohen. Bringing portraits to life. ACM Transactions on Graphics (TOG), 36(6):196, 2017. 1

H. Kim, P. Garrido, A. Tewari, W. Xu, J. Thies, M. Nießner, P. Perez, C. Richardt, M. Zollh ´ ofer, and C. Theobalt. Deep ¨ video portraits. arXiv preprint arXiv:1805.11714, 2018. 2

H. Zhang, I. J. Goodfellow, D. N. Metaxas, and A. Odena. Self-attention generative adversarial networks. arXiv:1805.08318, 2018. 5

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In Advances in neural information processing systems, pages 2672–2680, 2014. 2

J. Deng, J. Guo, X. Niannan, and S. Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In CVPR, 2019. 6

J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In Proc. ECCV, pages 694–711, 2016. 4, 5

J. S. Chung, A. Nagrani, and A. Zisserman. Voxceleb2: Deep speaker recognition. In INTERSPEECH, 2018. 5
J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Nießner. Face2face: Real-time face capture and reenactment of RGB videos. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2387–2395, 2016. 2
K. Nagano, J. Seo, J. Xing, L. Wei, Z. Li, S. Saito, A. Agarwal, J. Fursund, H. Li, R. Roberts, et al. paGAN: real-time avatars using dynamic textures. In SIGGRAPH Asia 2018 Technical Papers, page 258. ACM, 2018. 2
K. S. Andrew Brock, Jeff Donahue. Large scale gan training for high fidelity natural image synthesis. arXiv:1809.11096, 2018. 2, 5
K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In Proc. ICLR, 2015. 4
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. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, Advances in Neural Information Processing Systems 30, pages 6626–6637. Curran Associates, Inc., 2017. 6
M. K. Takeru Miyato. cgans with projection discriminator. arXiv:1802.05637, 2018. 2, 4 [33] M. K. Y. Y. Takeru Miyato, Toshiki Kataoka. Spectral normalization for generative adversarial networks. arXiv:1802.05957, 2018. 5
M. Mori. The uncanny valley. Energy, 7(4):33–35, 1970. 1
O. Alexander, M. Rogers, W. Lambeth, J.-Y. Chiang, W.- C. Ma, C.-C. Wang, and P. Debevec. The Digital Emily project: Achieving a photorealistic digital actor. IEEE Computer Graphics and Applications, 30(4):20–31, 2010. 2
O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In Proc. BMVC, 2015. 4
O. Wiles, A. Sophia Koepke, and A. Zisserman. X2face: A network for controlling face generation using images, audio, and pose codes. In The European Conference on Computer Vision (ECCV), September 2018. 1, 2, 6
P. Isola, J. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In Proc. CVPR, pages 5967–5976, 2017. 2
R. Zhang, T. Che, Z. Ghahramani, Y. Bengio, and Y. Song. Metagan: An adversarial approach to few-shot learning. In NeurIPS, pages 2371–2380, 2018. 2
S. Arik, J. Chen, K. Peng, W. Ping, and Y. Zhou. Neural voice cloning with a few samples. In Proc. NIPS, pages 10040–10050, 2018. 2
S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37, ICML'15, pages 448–456. JMLR.org, 2015. 5
S. Lombardi, J. Saragih, T. Simon, and Y. Sheikh. Deep appearance models for face rendering. ACM Transactions on Graphics (TOG), 37(4):68, 2018. 2
S. M. Seitz and C. R. Dyer. View morphing. In Proceedings of the 23rd annual conference on Computer graphics and interactive techniques, pages 21–30. ACM, 1996. 1
S. O. Mehdi Mirza. Conditional generative adversarial nets. arXiv:1411.1784. 2
S. Suwajanakorn, S. M. Seitz, and I. KemelmacherShlizerman. Synthesizing Obama: learning lip sync from audio. ACM Transactions on Graphics (TOG), 36(4):95, 2017. 2
T. A. Tero Karras, Samuli Laine. A style-based generator architecture for generative adversarial networks. arXiv:1812.04948. 2
T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018. 4, 6
T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, G. Liu, A. Tao, J. Kautz, and B. Catanzaro. Video-to-video synthesis. arXiv preprint arXiv:1808.06601, 2018. 2

V. Blanz, T. Vetter, et al. A morphable model for the synthesis of 3d faces. In Proc. SIGGRAPH, volume 99, pages 187–194, 1999. 2
X. Huang and S. Belongie. Arbitrary style transfer in realtime with adaptive instance normalization. In Proc. ICCV, 2017. 2, 5
Y. Ganin, D. Kononenko, D. Sungatullina, and V. Lempitsky. Deepwarp: Photorealistic image resynthesis for gaze manipulation. In European Conference on Computer Vision, pages 311–326. Springer, 2016. 1
Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. arXiv preprint arXiv:1408.5093, 2014. 5
Y. Jia, Y. Zhang, R. Weiss, Q. Wang, J. Shen, F. Ren, P. Nguyen, R. Pang, I. L. Moreno, Y. Wu, et al. Transfer learning from speaker verification to multispeaker text-tospeech synthesis. In Proc. NIPS, pages 4485–4495, 2018. 2
Z. Shu, M. Sahasrabudhe, R. Alp Guler, D. Samaras, N. Paragios, and I. Kokkinos. Deforming autoencoders: Unsupervised disentangling of shape and appearance. In The European Conference on Computer Vision (ECCV), September 2018. 1
Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: From error visibility to structural similarity. Trans. Img. Proc., 13(4):600–612, Apr. 2004. 6

TRANSLATED VERSION: SPANISH

Below is a rough translation of the insights presented above. This was done to give a general understanding of the ideas presented in the paper. Please excuse any grammatical mistakes and do not hold the original authors responsible for these mistakes.

VERSION TRADUCIDA: ESPAÑOL

A continuación se muestra una traducción aproximada de las ideas presentadas anteriormente. Esto se hizo para dar una comprensión general de las ideas presentadas en el documento. Por favor, disculpe cualquier error gramatical y no responsabilite a los autores originales de estos errores.

INTRODUCCIÓN

En este trabajo, consideramos la tarea de crear modelos de cabezas parlantes fotorrealistas personalizados, es decir, sistemas que pueden sintetizar secuencias de video plausibles de expresiones de habla e imitaciones de un individuo en particular. Más específicamente, consideramos el problema de sintetizar imágenes de cabeza personalizadas fotorrealistas dadas un conjunto de puntos de referencia faciales, que impulsan la animación del modelo. Esta capacidad tiene aplicaciones prácticas para la telepresencia, incluidas las videoconferencias y los juegos multijugador, así como la industria de efectos especiales. Se sabe que sintetizar secuencias de cabezas parlantes realistas es difícil por dos razones. Primero, las cabezas humanas tienen una alta complejidad fotométrica, geométrica y cinemática. Esta complejidad se debe no solo al modelado de rostros (para el cual existe una gran cantidad de enfoques de modelado) sino también al modelado de la cavidad bucal, el cabello y las prendas. El segundo factor de complicación es la agudeza del sistema visual humano hacia errores incluso menores en el modelado de apariencia de cabezas humanas (el llamado efecto valle inquietante). Esta baja tolerancia a los errores de modelado explica la prevalencia actual de avatares similares a dibujos animados no fotorrealistas en muchos sistemas de teleconferencia prácticamente implementados. Para superar los desafíos, varios trabajos han propuesto sintetizar secuencias de cabezas articuladas deformando un marco estático único o múltiple. Tanto los algoritmos de deformación clásicos como los campos de deformación sintetizados mediante el aprendizaje automático (incluido el aprendizaje profundo) se pueden utilizar para tales fines. Si bien los sistemas basados en deformaciones pueden crear secuencias de cabezas parlantes a

partir de una sola imagen, la cantidad de movimiento, la rotación de la cabeza y la disoclusión que pueden manejar sin artefactos perceptibles es limitado.

La síntesis directa (libre de deformaciones) de fotogramas de vídeo utilizando redes convolucionales profundas entrenadas de forma adversaria (ConvNets) presenta la nueva esperanza para cabezas parlantes fotorrealistas. Muy recientemente, estos sistemas han demostrado algunos resultados notablemente realistas. Sin embargo, para tener éxito, estos métodos tienen que entrenar grandes redes, donde tanto el generador como el discriminador tienen decenas de millones de parámetros para cada cabeza parlante. Por lo tanto, estos sistemas requieren un video de varios minutos de duración o un gran conjunto de datos de fotografías, así como horas de capacitación en la GPU para crear un nuevo modelo de cabeza parlante personalizado. Si bien este esfuerzo es menor que el requerido por los sistemas que construyen modelos de cabeza fotorrealistas utilizando sofisticados modelos físicos y ópticos, sigue siendo excesivo para la mayoría de los escenarios prácticos de telepresencia, donde queremos permitir a los usuarios crear sus modelos de cabeza personalizados con tan poco esfuerzo como sea posible.

En este trabajo, presentamos un sistema para crear modelos de cabezas parlantes a partir de un puñado de fotografías (el llamado aprendizaje de pocas instantáneas) y con un tiempo de entrenamiento limitado. De hecho, nuestro sistema puede generar un resultado razonable basado en una sola fotografía (aprendizaje de una sola toma), mientras que agregar algunas fotografías más aumenta la fidelidad de la personalización. De manera similar, las cabezas parlantes creadas por nuestro modelo son ConvNets profundas que sintetizan cuadros de video de manera directa mediante una secuencia de operaciones convolucionales en lugar de deformaciones. Las cabezas parlantes creadas por nuestro sistema pueden, por lo tanto, manejar una gran variedad de poses que van más allá de las capacidades de los sistemas basados en deformaciones.

La capacidad de aprendizaje de pocas tomas se obtiene a través de un extenso pre-entrenamiento (meta-aprendizaje) en un gran corpus de videos de cabezas parlantes correspondientes a diferentes hablantes con apariencia diversa. En el curso del metaaprendizaje, nuestro sistema simula tareas de aprendizaje de pocas tomas y aprende a transformar posiciones históricas en fotografías personalizadas de aspecto realista, con un pequeño conjunto de imágenes de entrenamiento con esta persona. Después de eso, un puñado de fotografías de una nueva persona establece un nuevo problema de aprendizaje adversario con un generador de alta capacidad y un discriminador preentrenado mediante metaaprendizaje. El nuevo problema adversarial converge hacia el estado que genera imágenes realistas y personalizadas luego de unos pocos pasos de entrenamiento.

En los experimentos, proporcionamos comparaciones de cabezas parlantes creadas por nuestro sistema con modelos alternativos de cabezas parlantes neuronales a través de mediciones cuantitativas y un estudio de usuario, donde nuestro enfoque genera imágenes de suficiente realismo y fidelidad de personalización para engañar a los participantes del estudio. Demostramos varios usos de nuestros modelos de cabezas parlantes, incluida la síntesis de video utilizando pistas históricas extraídas de secuencias de video de la misma persona, así como el titiriterio (síntesis de video de una determinada persona basada en las huellas de la cara de otra persona).

CONCLUSIÓN

Hemos presentado un marco para el metaaprendizaje de modelos generativos adversarios, que es capaz de entrenar cabezas parlantes virtuales altamente realistas en forma de redes generadoras profundas. Fundamentalmente, solo se necesita un puñado de fotografías (tan poco como una) para crear un nuevo modelo, mientras que el modelo entrenado en 32 imágenes logra un realismo perfecto y una puntuación de personalización en nuestro estudio de usuario (para imágenes estáticas de 224p).

Actualmente, las limitaciones clave de nuestro método son la representación mímica (en particular, el conjunto actual de hitos no representa la mirada de ninguna manera) y la falta de adaptación de hitos. El uso de puntos de referencia de una persona diferente conduce a un desajuste notable de personalidad. Entonces, si uno quiere crear videos de titiriteros "falsos" sin tal desajuste, se necesita alguna adaptación histórica. Sin embargo, observamos que muchas aplicaciones no requieren el manejo de títeres de una

persona diferente y, en cambio, solo necesitan la capacidad de manejar la propia cabeza parlante. Para tal escenario , nuestro enfoque ya proporciona una solución de alto realismo.

TRANSLATED VERSION: FRENCH

Below is a rough translation of the insights presented above. This was done to give a general understanding of the ideas presented in the paper. Please excuse any grammatical mistakes and do not hold the original authors responsible for these mistakes.

VERSION TRADUITE: FRANÇAIS

Voici une traduction approximative des idées présentées ci-dessus. Cela a été fait pour donner une compréhension générale des idées présentées dans le document. Veuillez excuser toutes les erreurs grammaticales et ne pas tenir les auteurs originaux responsables de ces erreurs.

INTRODUCTION

Dans ce travail, nous considérons la tâche de créer des modèles de têtes parlantes photoréalistes personnalisés, c'est-à-dire des systèmes capables de synthétiser des séquences vidéo plausibles d'expressions vocales et des imitations d'un individu particulier. Plus précisément, nous considérons le problème de la synthèse d'images de tête personnalisées photoréalistes à partir d'un ensemble de repères de visage, qui animent l'animation du modèle. Une telle capacité a des applications pratiques pour la téléprésence, y compris la vidéoconférence et les jeux multi-joueurs, ainsi que l'industrie des effets spéciaux. La synthèse de séquences de têtes parlantes réalistes est connue pour être difficile pour deux raisons. Premièrement, les têtes humaines ont une complexité photométrique, géométrique et cinématique élevée. Cette complexité provient non seulement de la modélisation des visages (pour lesquels il existe un grand nombre d'approches de modélisation) mais également de la modélisation de la cavité buccale, des cheveux et des vêtements. Le deuxième facteur de complication est l'acuité du système visuel humain vis-à-vis d'erreurs même mineures dans la modélisation de l'apparence des têtes humaines (le soi-disant effet de vallée étrange). Une telle faible tolérance aux erreurs de modélisation explique la prévalence actuelle d'avatars de type bande dessinée non photoréalistes dans de nombreux systèmes de téléconférence pratiquement déployés. Pour surmonter les défis, plusieurs travaux ont proposé de synthétiser des séquences de têtes articulées en déformant une ou plusieurs trames statiques. Les algorithmes de déformation classiques et les champs de déformation synthétisés à l'aide de l'apprentissage automatique (y compris l'apprentissage en profondeur) peuvent être utilisés à ces fins. Alors que les systèmes basés sur la déformation peuvent créer des séquences de têtes parlantes à partir d'une seule image, la quantité de mouvement, la rotation de la tête et la désocclusion qu'ils peuvent gérer sans artefacts visibles est limité.

La synthèse directe (sans déformation) d'images vidéo à l'aide de réseaux convolutionnels profonds entraînés de manière adverse (ConvNets) présente le nouvel espoir de têtes parlantes photoréalistes. Très récemment, des résultats remarquablement réalistes ont été démontrés par de tels systèmes. Cependant, pour réussir, de telles méthodes doivent former de grands réseaux, où le générateur et le discriminateur ont des dizaines de millions de paramètres pour chaque tête parlante. Ces systèmes nécessitent donc une vidéo de plusieurs minutes ou un grand ensemble de données de photographies ainsi que des heures de formation GPU afin de créer un nouveau modèle de tête parlante personnalisé. Bien que cet effort soit inférieur à celui requis par les systèmes qui construisent des modèles de tête photo-réalistes utilisant une modélisation physique et optique sophistiquée, il reste excessif pour la plupart des scénarios de téléprésence pratiques, où nous voulons permettre aux utilisateurs de créer leurs modèles de tête personnalisés avec aussi peu effort que possible.

Dans ce travail, nous présentons un système de création de modèles de têtes parlantes à partir d'une poignée de photographies (apprentissage en quelques clichés) et avec un temps de formation limité. En fait, notre système peut générer un résultat raisonnable basé sur une seule photo (apprentissage en une seule

prise), tandis que l'ajout de quelques photos supplémentaires augmente la fidélité de la personnalisation. De même, les têtes parlantes créées par notre modèle sont des ConvNets profonds qui synthétisent les images vidéo de manière directe par une séquence d'opérations convolutives plutôt que par déformation. Les têtes parlantes créées par notre système peuvent donc gérer une grande variété de poses qui vont au-delà des capacités des systèmes basés sur la déformation.

La capacité d'apprentissage en quelques plans est obtenue grâce à une pré-formation approfondie (méta-apprentissage) sur un large corpus de vidéos de tête parlante correspondant à différents locuteurs d'apparence variée. Au cours du méta-apprentissage, notre système simule des tâches d'apprentissage en quelques plans et apprend à transformer des positions de repère en photographies personnalisées d'aspect réaliste, grâce à un petit ensemble d'images de formation avec cette personne. Après cela, une poignée de photographies d'une nouvelle personne pose un nouveau problème d'apprentissage contradictoire avec un générateur de grande capacité et un discriminateur pré-formés via le méta-apprentissage. Le nouveau problème contradictoire converge vers l'état qui génère des images réalistes et personnalisées après quelques étapes de formation.

Dans les expériences, nous fournissons des comparaisons de têtes parlantes créées par notre système avec des modèles alternatifs de têtes parlantes neuronales via des mesures quantitatives et une étude utilisateur, où notre approche génère des images d'un réalisme suffisant et d'une fidélité de personnalisation pour tromper les participants à l'étude. Nous démontrons plusieurs utilisations de nos modèles de têtes parlantes, y compris la synthèse vidéo à l'aide de pistes de repère extraites de séquences vidéo de la même personne, ainsi que le marionnettisme (synthèse vidéo d'une certaine personne basée sur les traces de repère du visage d'une personne différente).

CONCLUSION

Nous avons présenté un cadre pour le méta-apprentissage de modèles génératifs contradictoires, qui est capable de former des têtes parlantes virtuelles hautement réalistes sous la forme de réseaux de générateurs profonds. Surtout, seule une poignée de photographies (aussi peu qu'une seule) est nécessaire pour créer un nouveau modèle, alors que le modèle formé sur 32 images atteint un réalisme parfait et un score de personnalisation dans notre étude utilisateur (pour des images statiques de 224p).

Actuellement, les principales limites de notre méthode sont la représentation mimique (en particulier, l'ensemble actuel de points de repère ne représente en aucune façon le regard) et le manque d'adaptation des points de repère. L'utilisation de points de repère d'une personne différente conduit à une discordance de personnalité notable. Donc, si l'on veut créer de «fausses» vidéos de marionnettistes sans un tel décalage, une adaptation historique est nécessaire. Nous notons, cependant, que de nombreuses applications ne nécessitent pas de marionnettiste d'une personne différente et ont uniquement besoin de la capacité de conduire sa propre tête parlante. Pour un tel scénario , notre approche fournit déjà une solution très réaliste.

TRANSLATED VERSION: GERMAN

Below is a rough translation of the insights presented above. This was done to give a general understanding of the ideas presented in the paper. Please excuse any grammatical mistakes and do not hold the original authors responsible for these mistakes.

ÜBERSETZTE VERSION: DEUTSCH

Hier ist eine ungefähre Übersetzung der oben vorgestellten Ideen. Dies wurde getan, um ein allgemeines Verständnis der in dem Dokument vorgestellten Ideen zu vermitteln. Bitte entschuldigen Sie

alle grammatikalischen Fehler und machen Sie die ursprünglichen Autoren nicht für diese Fehler verantwortlich.

EINFÜHRUNG

In dieser Arbeit betrachten wir die Aufgabe, personalisierte fotorealistische Sprechkopfmodelle zu erstellen, dh Systeme, die plausible Videosequenzen von Sprachausdrücken und Nachahmungen eines bestimmten Individuums synthetisieren können. Insbesondere betrachten wir das Problem der Synthese fotorealistischer personalisierter Kopfbilder anhand einer Reihe von Gesichtsmarkierungen, die die Animation des Modells steuern. Diese Fähigkeit hat praktische Anwendungen für die Telepräsenz, einschließlich Videokonferenzen und Multiplayer-Spiele sowie für die Spezialeffektindustrie. Es ist bekannt, dass die Synthese realistischer Sprechkopfsequenzen aus zwei Gründen schwierig ist. Erstens weisen menschliche Köpfe eine hohe photometrische, geometrische und kinematische Komplexität auf. Diese Komplexität ergibt sich nicht nur aus der Modellierung von Gesichtern (für die es eine Vielzahl von Modellierungsansätzen gibt), sondern auch aus der Modellierung von Mundhöhlen, Haaren und Kleidungsstücken. Der zweite komplizierende Faktor ist die Schärfe des menschlichen visuellen Systems gegenüber selbst geringfügigen Fehlern bei der Modellierung des Erscheinungsbilds menschlicher Köpfe (der sogenannte unheimliche Taleffekt). Diese geringe Toleranz gegenüber Modellierungsfehlern erklärt die derzeitige Verbreitung nicht fotorealistischer Cartoon-ähnlicher Avatare in vielen praktisch eingesetzten Telekonferenzsystemen. Um die Herausforderungen zu bewältigen, haben mehrere Arbeiten vorgeschlagen, artikulierte Kopfsequenzen durch Verziehen eines einzelnen oder mehrerer statischer Rahmen zu synthetisieren. Für solche Zwecke können sowohl klassische Warping-Algorithmen als auch Warping-Felder verwendet werden, die mithilfe von maschinellem Lernen (einschließlich Deep Learning) synthetisiert wurden. Während Warping-basierte Systeme aus nur einem Bild sprechende Kopfsequenzen erstellen können, können Bewegung, Kopfdrehung und Disokklusion berücksichtigt werden dass sie ohne erkennbare Artefakte umgehen können, ist begrenzt.

Die direkte (verzerrungsfreie) Synthese von Videobildern mit kontradiktorisch trainierten Deep Convolutional Networks (ConvNets) bietet die neue Hoffnung für fotorealistische Sprechköpfe. In jüngster Zeit wurden mit solchen Systemen einige bemerkenswert realistische Ergebnisse gezeigt. Um jedoch erfolgreich zu sein, müssen solche Verfahren große Netzwerke trainieren, in denen sowohl Generator als auch Diskriminator zig Millionen Parameter für jeden sprechenden Kopf haben. Diese Systeme erfordern daher ein mehrere Minuten langes Video oder einen großen Datensatz mit Fotos sowie stundenlanges GPU-Training, um ein neues personalisiertes Modell für sprechende Köpfe zu erstellen. Während dieser Aufwand geringer ist als der, der von Systemen benötigt wird, die fotorealistische Kopfmodelle unter Verwendung ausgefeilter physikalischer und optischer Modelle erstellen, ist er für die meisten praktischen Telepräsenzszenarien, in denen Benutzer ihre personalisierten Kopfmodelle mit möglichst wenig erstellen können, immer noch zu groß Anstrengung wie möglich.

In dieser Arbeit stellen wir ein System zur Erstellung von Sprechkopfmodellen aus einer Handvoll Fotos (sogenanntes Wenigschuss-Lernen) und mit begrenzter Einarbeitungszeit vor. Tatsächlich kann unser System auf der Grundlage eines einzelnen Fotos ein vernünftiges Ergebnis erzielen (einmaliges Lernen), während das Hinzufügen einiger weiterer Fotos die Genauigkeit der Personalisierung erhöht. Ähnlich wie bei den von unserem Modell erstellten sprechenden Köpfen handelt es sich um tiefe ConvNets, die Videobilder auf direkte Weise durch eine Folge von Faltungsoperationen und nicht durch Verziehen synthetisieren. Die von unserem System erzeugten Sprechköpfe können daher eine Vielzahl von Posen verarbeiten, die über die Fähigkeiten von Warping-basierten Systemen hinausgehen.

Die Lernfähigkeit mit wenigen Schüssen wird durch ein umfangreiches Pre-Training (Meta-Learning) an einem großen Korpus von Videos mit sprechendem Kopf erreicht, die verschiedenen Sprechern mit unterschiedlichem Erscheinungsbild entsprechen. Im Verlauf des Meta-Lernens simuliert unser System Lernaufgaben mit wenigen Schüssen und lernt, Orientierungspunkte in realistisch aussehende personalisierte Fotos umzuwandeln, wenn eine kleine Anzahl von Bildern mit dieser Person trainiert wird. Danach stellen eine Handvoll Fotos einer neuen Person ein neues kontroverses Lernproblem mit

einem Generator und Diskriminator mit hoher Kapazität auf, die über Meta-Learning vorab trainiert wurden. Das neue kontroverse Problem konvergiert zu dem Zustand, der nach einigen Trainingsschritten realistische und personalisierte Bilder erzeugt .

In den Experimenten bieten wir Vergleiche von Sprechköpfen, die von unserem System erstellt wurden, mit alternativen neuronalen Sprechkopfmodellen über quantitative Messungen und eine Benutzerstudie, wobei unser Ansatz Bilder mit ausreichendem Realismus und Personalisierungstreue erzeugt, um die Studienteilnehmer zu täuschen. Wir demonstrieren verschiedene Verwendungszwecke unserer sprechenden Kopfmodelle, einschließlich der Videosynthese unter Verwendung von Orientierungspunktspuren, die aus Videosequenzen derselben Person extrahiert wurden, sowie des Puppenspiels (Videosynthese einer bestimmten Person basierend auf den Gesichtsspuren einer anderen Person).

FAZIT

Wir haben einen Rahmen für das Meta-Lernen von kontradiktorischen generativen Modellen vorgestellt, mit dem hochrealistische virtuelle Sprechköpfe in Form von tiefen Generatornetzwerken trainiert werden können. Entscheidend ist, dass nur eine Handvoll Fotos (nur eines) benötigt werden, um ein neues Modell zu erstellen, während das auf 32 Bildern trainierte Modell in unserer Benutzerstudie (für statische 224p-Bilder) einen perfekten Realismus- und Personalisierungswert erzielt.

Derzeit sind die Haupteinschränkungen unserer Methode die Nachahmung der Darstellung (insbesondere repräsentiert der aktuelle Satz von Orientierungspunkten den Blick in keiner Weise) und das Fehlen einer Anpassung der Orientierungspunkte. Die Verwendung von Orientierungspunkten einer anderen Person führt zu einer spürbaren Nichtübereinstimmung der Persönlichkeit. Wenn man also "gefälschte" Puppenspielvideos ohne solche Nichtübereinstimmung erstellen möchte, ist eine wegweisende Anpassung erforderlich. Wir stellen jedoch fest, dass viele Anwendungen nicht das Puppenspiel einer anderen Person erfordern, sondern nur die Fähigkeit, den eigenen sprechenden Kopf zu steuern. Für ein solches Szenario bietet unser Ansatz bereits eine realistische Lösung.

TRANSLATED VERSION: PORTUGUESE

Below is a rough translation of the insights presented above. This was done to give a general understanding of the ideas presented in the paper. Please excuse any grammatical mistakes and do not hold the original authors responsible for these mistakes.

VERSÃO TRADUZIDA: PORTUGUÊS

Aqui está uma tradução aproximada das ideias acima apresentadas. Isto foi feito para dar uma compreensão geral das ideias apresentadas no documento. Por favor, desculpe todos os erros gramaticais e não responsacule os autores originais responsáveis por estes erros.

INTRODUÇÃO

Neste trabalho, consideramos a tarefa de criar modelos fotorrealistas personalizados de head talkers, ou seja, sistemas que podem sintetizar sequências de vídeo plausíveis de expressões de fala e mímica de um determinado indivíduo. Mais especificamente, consideramos o problema de sintetizar imagens fotorrealistas personalizadas da cabeça a partir de um conjunto de pontos de referência do rosto, que conduzem a animação do modelo. Essa capacidade tem aplicações práticas para telepresença, incluindo videoconferência e jogos para vários jogadores, bem como na indústria de efeitos especiais. Sintetizar sequências realistas de falantes é conhecido por ser difícil por dois motivos. Em primeiro lugar, as cabeças humanas têm alta complexidade fotométrica, geométrica e cinemática. Essa complexidade decorre não apenas da modelagem de faces (para a qual existe um grande número de abordagens de modelagem), mas

também da modelagem da cavidade bucal, cabelo e roupas. O segundo fator complicador é a agudeza do sistema visual humano com relação a erros ainda menores na modelagem da aparência de cabeças humanas (o chamado efeito de vale sobrenatural). Essa baixa tolerância a erros de modelagem explica a prevalência atual de avatares não fotorrealistas semelhantes a desenhos animados em muitos sistemas de teleconferência implantados de forma prática. Para superar os desafios, diversos trabalhos propuseram sintetizar sequências de cabeças articuladas por meio do empenamento de um único ou múltiplos quadros estáticos. Tanto algoritmos de warping clássicos quanto campos de warping sintetizados usando aprendizado de máquina (incluindo aprendizado profundo) podem ser usados para tais propósitos. Embora os sistemas baseados em warping possam criar sequências de cabeças falantes a partir de uma única imagem, a quantidade de movimento, rotação da cabeça e desoclusão que eles podem manipular sem artefatos perceptíveis é limitado.

A síntese direta (livre de distorção) de quadros de vídeo usando redes convolucionais profundas treinadas adversarialmente (ConvNets) apresenta a nova esperança para cabeças falantes fotorrealistas. Muito recentemente, alguns resultados notavelmente realistas foram demonstrados por tais sistemas. No entanto, para ter sucesso, tais métodos precisam treinar grandes redes, onde tanto o gerador quanto o discriminador têm dezenas de milhões de parâmetros para cada cabeça falante. Esses sistemas, portanto, requerem um vídeo de vários minutos ou um grande conjunto de dados de fotografias, bem como horas de treinamento em GPU para criar um novo modelo de cabeça falante personalizado. Embora esse esforço seja menor do que o exigido por sistemas que constroem modelos de cabeça foto-realistas usando modelagem física e óptica sofisticada, ainda é excessivo para a maioria dos cenários de telepresença práticos, onde queremos permitir que os usuários criem seus modelos de cabeça personalizados com o mínimo esforço possível.

Neste trabalho, apresentamos um sistema para a criação de modelos de cabeças falantes a partir de um punhado de fotografias (o chamado fewshot learning) e com tempo de treinamento limitado. Na verdade, nosso sistema pode gerar um resultado razoável com base em uma única fotografia (aprendizagem única), enquanto adicionar mais algumas fotos aumenta a fidelidade da personalização. Da mesma forma, para, as cabeças falantes criadas por nosso modelo são ConvNets profundas que sintetizam quadros de vídeo de maneira direta por uma sequência de operações convolucionais em vez de distorção. As cabeças falantes criadas por nosso sistema podem, portanto, lidar com uma grande variedade de poses que vão além das habilidades dos sistemas baseados em warping.

A habilidade de aprendizado de poucos instantes é obtida por meio de extenso pré-treinamento (meta-aprendizado) em um grande corpus de vídeos falantes correspondentes a diferentes falantes com aparência diversa. No curso de meta-aprendizagem, nosso sistema simula tarefas de aprendizagem de poucos instantes e aprende a transformar posições de referência em fotografias personalizadas de aparência realista, dado um pequeno conjunto de imagens de treinamento com essa pessoa. Depois disso, um punhado de fotos de uma nova pessoa cria um novo problema de aprendizado adversário com gerador de alta capacidade e discriminador pré-treinado via meta-aprendizado. O novo problema adversário converge para o estado que gera imagens realísticas e personalizadas após algumas etapas de treinamento.

Nos experimentos, fornecemos comparações de cabeças falantes criadas por nosso sistema com modelos alternativos de cabeças falantes neurais por meio de medições quantitativas e um estudo de usuário, onde nossa abordagem gera imagens de realismo suficiente e fidelidade de personalização para enganar os participantes do estudo. Demonstramos vários usos de nossos modelos de cabeças falantes, incluindo a síntese de vídeo usando trilhas de referência extraídas de sequências de vídeo da mesma pessoa, bem como a manipulação de marionetes (síntese de vídeo de uma determinada pessoa com base nas trilhas de referência de rosto de uma pessoa diferente).

CONCLUSÃO

Apresentamos uma estrutura para meta-aprendizagem de modelos gerativos adversários, que é capaz de treinar falantes virtuais altamente realistas na forma de redes geradoras profundas. Crucialmente, apenas

um punhado de fotografias (tão pequeno quanto um) é necessário para criar um novo modelo, enquanto o modelo treinado em 32 imagens atinge realismo perfeito e pontuação de personalização em nosso estudo de usuário (para imagens estáticas de 224p).

Atualmente, as principais limitações do nosso método são a representação da mímica (em particular, o conjunto atual de marcos não representa o olhar de forma alguma) e a falta de adaptação dos marcos. Usar pontos de referência de uma pessoa diferente leva a uma incompatibilidade perceptível de personalidade. Portanto, se alguém quiser criar vídeos “falsos” de marionetes sem essa incompatibilidade, é necessária alguma adaptação histórica. Notamos, no entanto, que muitas aplicações não requerem manipular uma pessoa diferente e, em vez disso, precisam apenas da capacidade de dirigir a própria cabeça falante. Para tal cenário , nossa abordagem já fornece uma solução de alto realismo.