Vol.2022-CG-185 No.13
2022/3/12

顔画像のスワイプ操作による動画生成

西川 隆盛[1,a)]　栗山 繁[1,2,b)]

概要：近年，深層学習を用いて自然画像から動画を end-to-end に生成するのに、フレームを予測的に補間する手法などが提案されている．これらは動画の生成に複数の静止画が必要であるが，二枚以上の画像を入手できない場合には、一枚の画像から生成できることが望ましい．また，end-to-end な生成では動画中のオブジェクトの動きを直接制御できないので，ユーザの意図した動画の生成が困難である．この解決策として，ユーザが画像中で局所的に変化する部分の大きさと向きをスワイプ操作によって対話的に指定して、一枚の画像から確率的に動画を生成する手法（iPOKE）が提案されている．本報告では，この既存手法を直観的な顔アニメーションの生成に応用する場合の表現力と有効性を，ユーザ評価実験を通じて検証する．

キーワード：確率的動画生成，スワイプ操作，顔アニメーション

Video generation by swiping a face image

Abstract:
Recently, frame interpolations by prediction have been proposed for the end-to-end generation of videos from natural images using deep learning. These require multiple still images to generate a video, but it is desirable to generate a video from a single image when more than two images are not available. In addition, end-to-end generation does not allow direct control over the motion of objects in the video, making it difficult to generate a video intended by the user. A method for generating videos probabilistically from a single image (iPOKE) has been proposed to solve this problem. The user interactively specifies the size and orientation of locally changing parts in the image by swiping. In this study, we examine this existing method's expressive power and effectiveness when applied to the generation of intuitive facial animations through user evaluations.

Keywords: Probabilistic video generation, Swipe operation, Facial animations

1. はじめに

近年，深層学習を用いて自然画像から動画を end-to-end に生成する手法が多く提案されている．自然画像から動画を生成する手法は，フレーム予測をする Deep Voxel Flow[1] やフレーム補間をする hierarchial VRNN[2] などがある．これらは動画の生成に 2 フレーム以上での静止画が必要であるが，これらを必ずしも与える事ができるとは限らない．したがって，フレーム予測のタスクとしては 1 フレーム分の，すなわち 1 枚の静止画像から生成できることが望ましい．また，end-to-end な生成では動画中のオブジェクトの動きを制御できず，ユーザの意図した動画の生成が困難である．そこで，ユーザのスワイプ操作（2 次元平面での移動ベクトルを指定する操作）によって対話的に動画を生成する手法である iPOKE [3] が提案されている．本報告ではこの手法を顔画像に対して適用し，スワイプ操作という単純で直観的な操作で，1 枚の肖像画像から表現豊かな顔アニメーションを自動生成する技術の可能性を調査する．

iPOKE では，双方向での学習が可能な conditional Invertible Neural Network (cINN) を使用することで，1 フレームの入力からの確率的動画生成を達成しており，スワイプ操作によってオブジェクトを任意の動きに編集できる．しかし，植物の動画では葉の構造が崩れてしまう等の問題が指摘されているので，本研究では構造の崩れに対して同様な問題が発生するであろう顔画像での性能を検証する．

[1] 豊橋技術科学大学
TUT,Toyohashi,Aichi 441-8580,Japan
[2] 株式会社サイバーエージェント
CyberAgent inc., Shibuya, Tokyo 150–0042, Japan
[a)] nishikawa.ryusei.ag@tut.jp
[b)] sk@tut.jp

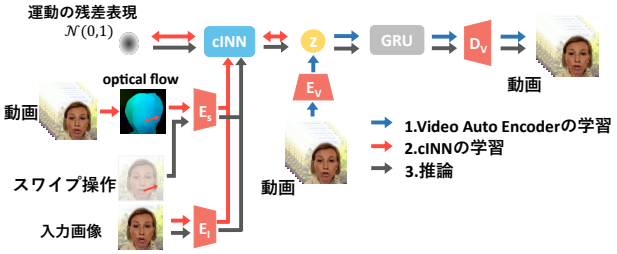

図 1 iPOKE のモデル図.
Fig. 1 Model diagram of iPOKE.

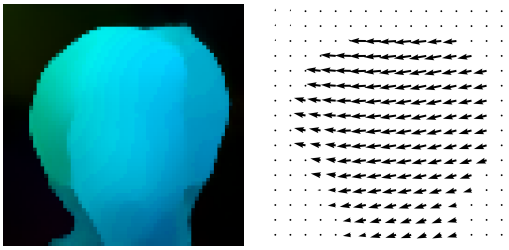

図 2 Flownet で抽出したオプティカルフローとシフトベクトル.
Fig. 2 Optical flow and shift vectors extracted by Flownet.

2. 学習モデル

図 1 に示す iPOKE は，以下の手順で学習と推論を行う.

(1) 動画を計算可能な次元数へと変換するために，Video Auto Encoder によって潜在変数 z を学習する.
(2) 潜在変数，入力画像，スワイプ操作を cINN へ入力し，運動の残差表現を学習する.
(3) 運動の残差表現をサンプリングし，入力画像とスワイプ操作を cINN に入力して動画を生成する.

ただし，ここでの運動の残差表現とは，画像には含まれないが動画には存在するオブジェクトの運動の情報を指す．この運動の残差表現をサンプリングし，初期フレームとスワイプ操作を条件として cINN に入力することで，確率的に動画を生成する.

2.1 スワイプ操作

スワイプ操作の学習データには，学習済みの Flownet 2.0 [4] によって，動画の最初と最後のフレームからオプティカルフローを抽出し，そのシフトベクトルを模擬的なスワイプ操作として用いている．抽出されたオプティカルフロー とシフトベクトルを**図 2** に示す．ただし，推論時にはオプティカルフローを必要とせず，マウスのドラッグ操作をスワイプ操作として入力する点に注意されたい.

2.2 Video Auto Encoder (Video-AE)

先述した通り，cINN の計算量を減らすために Video-AE によって，動画の潜在変数 z を学習する．エンコーダ E_v では，動画の時間方向を考慮するために 3D-ResNet-encoder[5] を用い，潜在変数 z は後続の GRU の初期隠れ状態として利用される．デコーダ D_v は 2D-ResNet ブロックの連結で構成され，GRU の出力である特徴マップをアップサンプリングして動画を生成する．動画の各フレームに対する損失関数は L1 loss と perceptual loss ℓ^ϕ[6] を用いた以下の式で与える.

$$\mathcal{L}_{rec} = \frac{1}{T}\sum_{i=1}^{T}\left[\|x_i - \hat{x_i}\|_1 + \ell^\phi(x_i, \hat{x_i})\right] , \qquad (1)$$

ただし，x_i と $\hat{x_i}$ は元動画と生成動画の i 番目のフレームを表す.

また，空間的な識別器 $\mathcal{D}_S$ と時間的な識別器 $\mathcal{D}_T$ を用いた損失関数を，各々以下の式 (2) と式 (3) で与える.

$$\mathcal{L}_{\mathcal{D}_S} = \frac{1}{T}\sum_{i=1}^{T}\left[\mathcal{D}_S(\hat{x_i}) - \ell_{F_S}(x_i, \hat{x_i})\right] , \qquad (2)$$

$$\mathcal{L}_{\mathcal{D}_T} = \frac{1}{T}\sum_{i=1}^{T}\left[\mathcal{D}_T(\hat{x_i}) + \ell_{F_T}(x_i, \hat{x_i})\right] . \qquad (3)$$

ここで，ℓ_{F_S} と ℓ_{F_T} は feature matching loss [7] である．Video-AE は，これらの損失関数の和

$$\mathcal{L} = \mathcal{L}_{rec} + \mathcal{L}_{\mathcal{D}_S} + \mathcal{L}_{\mathcal{D}_T} , \qquad (4)$$

を用いて学習される.

2.3 スワイプ操作と入力画像に対するエンコーダ

スワイプ操作と入力画像の潜在変数を cINN の入力とするために，オートエンコーダとして事前学習し，それらのエンコーダ部を各々 E_s, E_i とする．デコーダは学習時にのみ使用され，推論時には使用しないので，図 1 では省略している.

画像に対するオートエンコーダは入力画像を再構築するように学習し，エンコーダの出力が入力画像の潜在変数表現となって cINN へ入力される．一方，スワイプ操作の学習にはオプティカルフローのシフトベクトルを使用する．シフトベクトルは 2 チャネルのマップで表現されるため，スワイプ操作は入力画像と同様な方法で学習される点に注意されたい．ただし，双方のオートエンコーダの損失関数には 2.2 節と同様に L1 loss と perceptual loss を用いる.

2.4 conditional Invertible Neural Network(cINN)

cINN は convolutional normalizing flow model[8] として，二層の act norm，四層の masked convolution，および一層の coupling layer で構成される複数個の sub block を連結し，glow block は各 sub block の後に連結される（**図 3** 参照）.

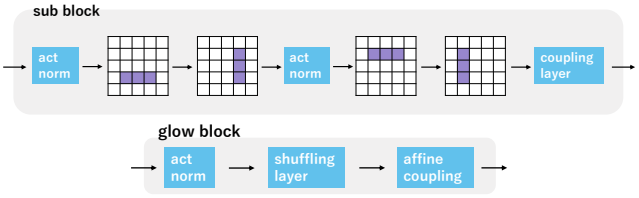

図 3 cINN の構成図.
Fig. 3 Architecture of cINN.

表 1 YouTube の動画データセット
Table 1 YouTube Video dataset

動画	時間長	フレーム数	fps
May	4 分 3 秒	6000	25
Obama	5 分 24 秒	8000	30
Nadella	4 分 27 秒	9700	30

Flow は逆変換可能な関数で構成されるため，入力データから潜在変数への学習のみで，潜在変数から入力データへの変換は逆関数の演算で算出できるという特徴がある．よって，Variational Autoencoder や Generative Adversarial Network では潜在変数から入力データへと逆写像するエンコーダが必要であったが，flow では必要ない点に留意されたい．

Glow block は ActNorm，affine coupling[9]，shuffling layer で構成され，shuffling layer は [8] で用いられる invertible 1×1 convolutions に相当する．ここで，学習時間を抑えるために，sub block は masked convolution[10] によって flow を構成する関数のヤコビ行列を効率的に計算できる．

3. 実験

3.1 データセット

データセットには，VideoForensicsHQ[11] と YouTube の 2 種類の動画を使用し，訓練データと検証データの割合は 8:2 に設定した．また，学習時には動画の全フレームを 64×64 画素にリサイズした．モデル全体の学習には，Tesla A100 40GB を 1 台用いて約 1 週間を要した．

VideoForensicsHQ は正面を向いて話している動画で構成されており，そこに含まれる 1 人の女性動画を使用した．使用した動画の再生時間さは約 7 分であり，総フレーム数は約 9,000 フレーム，フレームレートは 25 fps である．

一方，YouTube 動作では人が正面を向いて話している動画をダウンロードし，顔が中心に来るようにトリミングした．各動画の詳細を**表 1** に示す．

3.2 評価指標

動画の評価指標には Fréchet Video Distance(FVD)[12] を用いる．これは 2 つの分布間の距離を表す指標である Fréchet Inception Distance(FID)[13] を，動画を評価するための指標として拡張したものである．FVD は動画の個々のフレームの品質と動画の時間的な一貫性を評価する以下の式で定義される．

$$d(P_R, P_G) = \min_{X,Y} E|X - Y|^2 \,. \tag{5}$$

式 (5) は，分布 P_R,P_G の確率変数 X,Y について最小化をする．分布を多変量ガウス分布と仮定すると，式 (5) の右辺は式 (6) に変形できる．

$$|\mu_R + \mu_G|^2 = \mathrm{Tr}(\textstyle\sum_R + \sum_G - 2(\sum_R \sum_G)^{\frac{1}{2}}) \,. \tag{6}$$

ここで，μ_R，μ_G はそれぞれ分布 P_R,P_G の平均，$\sum_R$，$\sum_G$ は共分散行列である．

動画の特徴表現として多変量ガウス分布を得るために，事前学習済みの Inflated 3D Convnet [14] を使用する．このネットワークは，画像識別機を動画へと適応させるために時間方向へと拡張し，YouTube 上の動画データセット [15] を用いて，人間の行動認識をするために学習される．

3.3 1 名の顔データによる学習

iPOKE のモデルを VideoForensicsHQ で学習した結果を**図 4** に，YouTube の中から一人の女性の動画のみを用いて学習した結果を**図 5** に示す．これらは，口や目などを動かす操作を与えて 10 フレーム長の動画を生成した．図 4 では，口や顔が動くことが確認できたが，同時に背景も動く結果となった．これは，学習に使用した動画の背景にも変化があり，モデルが前景と背景を区別できず，背景の動きも同時に学習したことが原因と考えられる．一方，図 5 では背景の変化はなかったが，頭部を動かす操作を与えた場合，顔の構造がぼやけた結果となった．また，目を動かす操作を入力した場合に，瞬きをする動画が口を動かす動画に比べて生成され難いという問題があった．オプティカルフローでは，変化する時間間隔が短い表情よりも，それが長い頭部の変化を捉えてしまうので，表情の変化を捉えられていないと推察される．ゆえに，オプティカルフローの推定間隔を短くすることで表情の変化が捉えられやすくなると考えられる．

3.4 オプティカルフローの推定間隔の影響

オプティカルフローの推定間隔を 10 フレームから 5 フレームに変えて，モデルを再学習した結果を**図 6** に示す．ただし，初期フレーム，スワイプ操作，運動の残差表現をサンプリングした点は統一している．その結果，異なる表情の動きを生成できたが，顔の構造がぼやけたりするような問題は解決されなかった．これは，大きな値を持つシフトベクトルが優先的に学習されたことが原因として考えられる．そのため，オプティカルフローの推定間隔が変わって

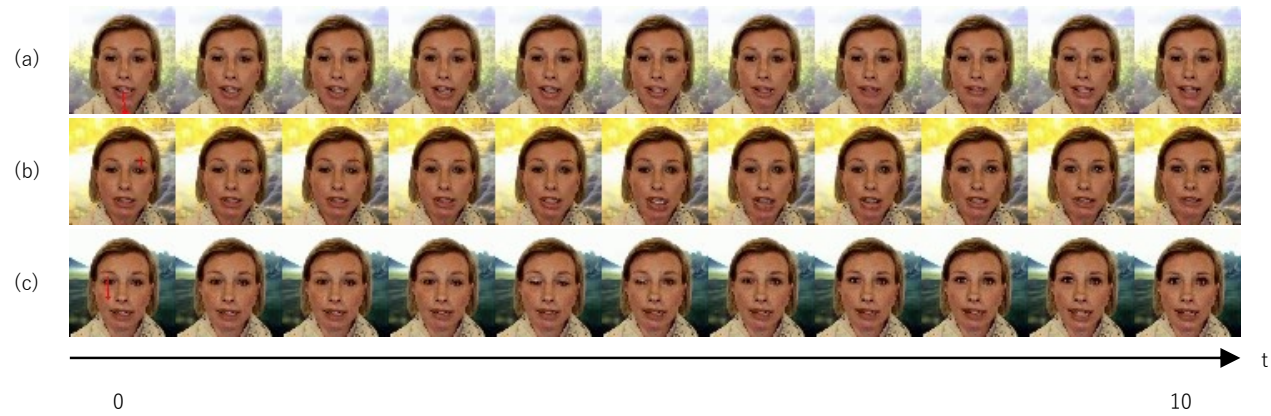

図 4 VideoForensicsHQ のデータセットを使用して (a) 口を開く操作，(b) 目を開く操作，(c) 瞬きする操作を与えて生成した動画.

Fig. 4 Videos generated with VideoForensicsHQ for the operations of (a) opening mouth, (b) opening eye, and (c) blinking.

図 5 YouTube を使用して (a) 顔を右に振る操作，(b) 口を開く操作，(c) 目を開く操作を与えて生成した動画.

Fig. 5 Videos generated with YouTube for the operations of (a) turning face to the right operation, (b) opening mouth, and (c) opening eye.

も，表情の小さい動きは学習にはあまり使用されず，頭部の大きな動きが学習に大きく影響していると考えられる．

3.5 3 名の顔データによる学習

YouTube の全てのデータセットで学習した結果を**図 7** に示す．3.3 節の結果と比べると，顔の構造が大きく崩れてしまう場合があり，動画としては不自然な結果となった．図 7 の（c），（d），（e）を見ると，目や口の構造が崩れていることが確認できる．これは人の顔の違いが学習ノイズとなったことが原因であると考えられる．また 3.3 節の結果と比べて，スワイプ操作を与えても，生成された動画が初期フレームからほぼ変化しない（b），（f）のような結果が多く確認された．これは，複数人の動画での iPOKE の学習の難しさを示唆している．

3.6 定量評価

動画の評価指標である FVD によって，生成した動画を比較した結果を**表 2** に示す．結果から 3 人の動画で学習した場合は動画の品質が大きく低下することが確認された．この原因として，人の顔の特徴の違いで Video-AE による学習が不十分であったことが推察される．

表 2 FVD による定量評価
Table 2 Quantitative evaluation by FVD

学習データ	FVD ↓
1 人（背景変化なし）	57.53
3 人（背景変化なし）	121.78
iPOKE（植物）	63.06
iPOKE（人の全身）	81.49

学習の過程における Video-AE での FVD の変化を示したグラフを**図 8** に示す．Video-AE の学習時点で，FVD の値が低くならないため，その結果が cINN の学習にも影響を及ぼしていると考えられる．

図 6 オプティカルフローの異なるフレーム推定間隔 (a) 10 フレーム，(b) 5 フレームにおける，YouTube データセットを使用して生成した動画.

Fig. 6 Videos generated using YouTube at different frame estimation intervals of (a) 10 and (b) 5 for optical flow.

図 7 YouTube のデータセットを使用して生成した動画．各々，目を開く (a,e)，口を開く (b,d,f)，および頭部を上に動かす (c) 操作を与えた場合.

Fig. 7 Videos generated using YouTube for the operation of opening eyes (a,e), opening mouth (b, d, f), and moving head upward (c).

3.7 定性評価

人間が見て自然な動画であるかを定性的に判断するため，ユーザ評価としてアンケートを実施した．元動画，および１人または３人の顔データで学習して生成した動画を各４枚用意し，無作為な順でユーザに動画を提示して自然さを５段階で評価してもらった．４２人のアンケート回答で得られた各評価の平均値を**図 9** に示す．その結果，３人同時に学習した場合が最もユーザ評価が低くなったが，これは 3.5 節でも述べた通り，顔の構造が大きく崩れているのが主たる原因と考えられる．

4. おわりに

本報告では顔画像を用いた顔アニメーション生成に対する iPOKE の有効性を検証をした．頭部を振る動きと表情の動きが混在している場合，オプティカルフローは頭部を振る動きを主に捉えてしまい，表情の動きは十分に捉えられないことが判明した．また，スワイプ操作には大きな値を持つシフトベクトルを使用するため，瞬きのようなフレーム間で変化が小さい動きを学習することが困難である．さらに，iPOKE では複数人の動画を学習データとし

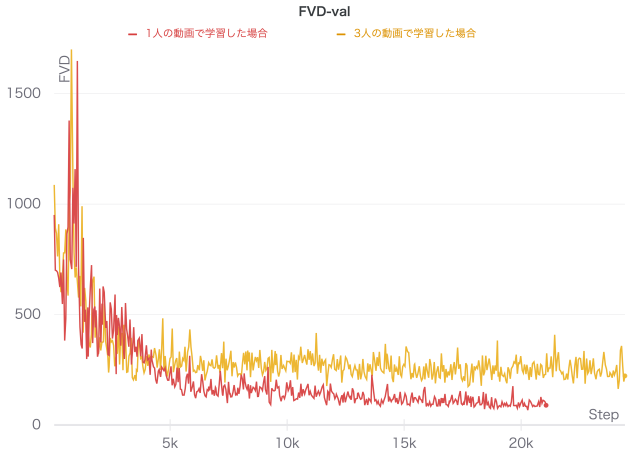

図 8 1 人と 3 人の動画で学習した場合の Video-AE の FVD の比較.

Fig. 8 Comparison of FVD of Video-AE trained with the video for one or three persons.

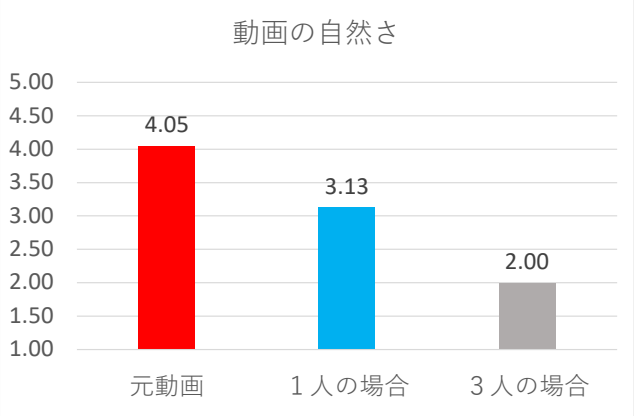

図 9 元動画と，1 人ずつおよび 3 人同時に学習した動画の自然さを 1〜5 で評価した平均値の比較.

Fig. 9 Comparison of the average rating from 1 to 5 for the naturalness, with respect to ground truth data, and videos trained by one or three persons.

て用いた場合，顔の構造が崩れるために自然な動画を生成できない．したがって，iPOKE を顔画像の編集に利用する際には，各個人に対して個別に学習する必要がある．

iPOKE の顔画像に対する表現力を向上させる方法として，入力に顔画像の特徴点を導入することが挙げられる．特徴点を使用することで，顔のパーツを個別に捉えることができるため，動きを分離した学習が可能になると考えられる．また，顔の個人差を軽減できるため，学習に用いる表情の多様性が増えることで生じる，生成動画の品質低下の問題を解決できる．

今回の実験では単一のスワイプ操作で動画を生成していたが，複数のスワイプ操作入力を同時に学習できれば，より多様な動画生成が可能になると期待できる．

参考文献

[1] Lluis Castrejon, Nicolas Ballas, and Aaron Courville. Improved conditional vrnns for video prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7608–7617, 2019.

[2] Ziwei Liu, Raymond A Yeh, Xiaoou Tang, Yiming Liu, and Aseem Agarwala. Video frame synthesis using deep voxel flow. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4463–4471, 2017.

[3] Andreas Blattmann, Timo Milbich, Michael Dorkenwald, and Björn Ommer. ipoke: Poking a still image for controlled stochastic video synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14707–14717, 2021.

[4] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul 2017.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[6] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pp. 694–711. Springer, 2016.

[7] 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*, pp. 8798–8807, 2018.

[8] Diederik P. Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In *NeurIPS*, 2018.

[9] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. *ArXiv*, Vol. abs/1605.08803, , 2017.

[10] Xuezhe Ma and Eduard H. Hovy. Macow: Masked convolutional generative flow. In *NeurIPS*, 2019.

[11] Gereon Fox, Wentao Liu, Hyeongwoo Kim, Hans-Peter Seidel, Mohamed Elgharib, and Christian Theobalt. VideoForensicsHQ: Detecting high-quality manipulated face videos. In *IEEE International Conference on Multimedia and Expo (ICME 2021)*, Shenzhen, China (Virtual), 2021. IEEE.

[12] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *ArXiv*, Vol. abs/1812.01717, , 2018.

[13] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NIPS*, 2017.

[14] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[15] Will Kay, João Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Apostol Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset. *ArXiv*, Vol. abs/1705.06950, , 2017.