動き情報を加えたPredNetによる未来画像生成の高精度化

西片智広† 山内 悠嗣†
†中部大学
E-mail: tr21010-0146@sti.chubu.ac.jp yuu@isc.chubu.ac.jp

1 はじめに

モビリティ分野では自動運転技術の実用化を目指して盛んに研究開発が行われているが，その1つに自車周辺の環境を認識するセンシング技術が挙げられる．安全安心な自動運転を実現するためには，自動運転車に搭載された多数のセンサにより周囲を観測し，獲得した情報から周囲の環境を認識する必要がある．さらに，危険な状態を早期に把握するためには，この先に発生する事象の予測が重要であることから自車及びその周辺の状況を予測する研究が進められている．

センサから獲得される情報の1つに車載カメラからの映像がある．カメラから観測できる画像から，この先に観測できる未来の画像を生成することができれば，より高度な予測の実現が期待できる．未来の画像を生成する研究 [1, 2, 3, 4] は幾つか提案されているが，その中でも PredNet[3] は，畳み込み Long Short Term Memory(LSTM)[5] をベースとした深層学習ネットワークであり，高精度な画像の予測を実現している．PredNet は連続した数フレームの画像群を入力し，1フレーム先の未来画像を生成する．PredNet が対象としている車載カメラの映像は，走行する自車の動きと，カメラで観測している自動車や歩行者等の移動体の動きの2つの動きが含まれる．これら2つの動きをより把握することができれば，この先の映像がどのように変化するか予想できる．

そこで，本研究では高精度な未来画像を生成するために，未来画像の生成時に動きの情報を加味する．明示的に動きの情報を加味することで，画像内で移動する物体や動的な背景を考慮することが可能となり，高精度な未来画像の生成が期待できる．動きの情報を加味するために，本研究では深層学習に基づく DDFlow[6] により推定したオプティカルフローを用いる．

2 関連研究

2.1 予測に関する研究

自車及びその周辺の状況を予測する研究の1つとして危険予測に関する研究 [7, 8] が提案されている．これらの研究では，交通事故を記録した車載カメラの映像を用いて，物体検出器と Recurrent Neural Network により事故の発生する可能性を予測する．また，自動車の速度や加速度などの数値情報と周囲の環境情報から車両の軌跡を予測する手法 [9] が提案されている．これらの研究は，過去に獲得したセンサ情報に基づいて先の事象を予測する．この先に発生する事象や危険リスクの予測ではなく，センサから獲得されるであろう未来の画像を予測できれば，より高度な予測の実現が期待できる．

2.2 未来の画像を生成する深層学習ネットワーク

深層学習により未来でカメラから観測されるであろう画像を予測する手法が提案されている．Xingjian らは，降水量を記録した天気図を入力とし，畳み込み LSTM によりこの先の天気図を生成する手法 [5] を提案している．Xiaodan らは，車載カメラの映像を入力とし，Generator で生成した一時刻先の未来画像と，実際の画像を Discriminator に入力し敵対的学習をさせる General Adversrial Network[10] を用いて高精度な未来画像を生成する手法 [11] を提案している．Junhyuk らは，Autoencoder[12] を用いて，入力されたビデオゲームの画像を圧縮し，その画像の一時刻先の未来画像を復元する手法 [13] を提案している．

また，多くの手法ではネットワークの入力に動画像だけではなく，他の情報を追加することで高精度な未来画像の生成を実現している．Fragkiadak らは，ビリヤードのシミュレーションデータを用いて，俯瞰撮影したビリヤード台の動画像とビリヤードの球に加わる力を入力とし，LSTM によりビリヤードの球の軌跡を画像として生成する手法 [1] を提案している．Finn らは，ロボットハンドが物体を押すまたは引き寄せる動作をしている動画像，ロボットハンドの姿勢と動作の種類の3つを入力とし，畳み込み LSTM によりロボットハンドの未来画像を生成する手法 [2] を提案している．Villegas らは，テニスのサーブを打つ構えをしている画像とサーブ中の骨格情報を入力とし，LSTM によりサーブをしている未来画像を生成する手法 [14] を提案している．

3　提案手法

提案手法は，動きの情報を明示的に捉えることで高精度な未来画像の生成を実現する．

3.1　DDFlow によるオプティカルフローの推定

オプティカルフローは，時間軸上で隣接したフレーム間の物体の動きの変化量を表したベクトルである．オプティカルフローを推定する手法 [15, 16, 17] は幾つか提案されているが，本研究では深層学習に基づきオプティカルフローを推定する DDFlow を採用する．DDFlow は高性能かつ高速にオプティカルフローを推定可能な PWC-Net[18] に隠れに対する頑健性を加えた手法である．

図 1 に DDFlow のネットワークの概要を示す．DDFlow は 2 段階の処理によりネットワークを学習する．1 段階目は全てのピクセルのうち，隠れが発生しないピクセルのみを対象としてフローを推定する．時間軸上で連続した画像 I_1, I_2 のそれぞれから，隠れの発生しないピクセル群を推定し，推定したピクセル群に対するフローを畳み込みニューラルネットワークにより推定する．

2 段階目は画像 I_1 から画像 I_2 に変化する際，画像の外に出ることで隠れが発生するピクセルのフローを推定する．入力画像 I_1, I_2 をランダムな位置と大きさで切り取ることにより小さくした画像 $\tilde{I}_1, \tilde{I}_2$ を作成する．これにより画像 I_1, I_2 では隠れの発生しないピクセルが，画像 $\tilde{I}_1, \tilde{I}_2$ では画像の外に出ることにより隠れが発生するピクセルとなる．$\tilde{I}_1, \tilde{I}_2$ にて故意に隠れを発生させたピクセルのフローを推定し，1 段階目で求めたフローを正解ラベルとして扱い損失を求める．これらの 2 段階の処理により，隠れに対して頑健なオプティカルフローを推定することが可能である．

DDFlow により推定したオプティカルフローの可視化した画像を図 2 に示す．可視化したオプティカルフロー画像から高精度かつ密なオプティカルフローであることが確認できる．

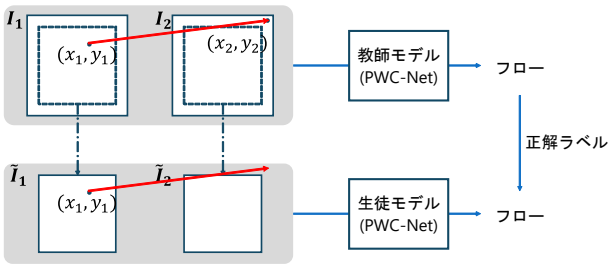

図 1　DDFlow の概要．

3.2　PredNet

PredNet は畳み込み LSTM をベースとした深層学習ネットワークである．PredNet は連続した数フレームの画像群を受け取り，1 フレーム先の未来画像を生成する．PredNet は LSTM をベースにしたネットワークであるため，入力された時系列データが多くなるほどネットワークに情報が蓄積され，より高精度な未来画像の生成が可能となる．図 3 に提案手法のネットワークの構成を示す．時刻 t，レイヤー数 l のとき，畳み込み LSTM 層 R_l^t，予測表現層 $\hat{A}_l^t$，入力表現層 A_l^t，誤差表現層 E_l^t の 4 つのユニットから構成され，複数のレイヤーを 1 つのネットワークとする．ネットワークの入力 x_t は画像と可視化したオプティカルフローである．各ユニットの更新式を式 (1)~(4) に示す．

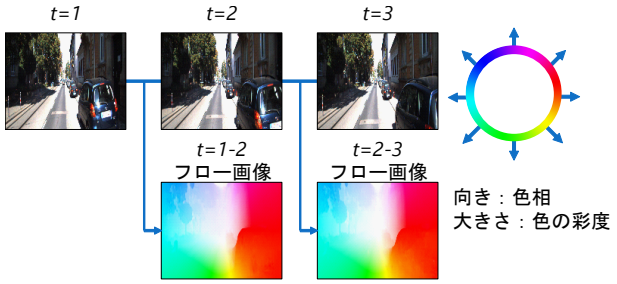

図 2　可視化したオプティカルフロー画像の例．

$$R_l^t = \mathrm{CoNVLSTM}\left(E_l^{t-1}, R_l^{t-1}, \mathrm{UPSAMPLE}\left(R_{l+1}^t\right)\right) \tag{1}$$

$$A_l^t = \begin{cases} x_t & l = 0 \\ \mathrm{MAXPOOL}\left(\mathrm{RELU}\left(\mathrm{CoNV}\left(E_{l-1}^t\right)\right)\right) & l > 0 \end{cases} \tag{2}$$

$$\hat{A}_l^t = \mathrm{RELU}\left(\mathrm{CONV}\left(R_l^t\right)\right) \tag{3}$$

$$E_l^t = \left(\mathrm{RELU}\left(A_l^t - \hat{A}_l^t\right) + \mathrm{RELU}\left(\hat{A}_l^t - A_l^t\right)\right) \tag{4}$$

畳み込み LSTM 層 R_l^t は一時刻前の誤差表現層 E_l^{t-1}，一時刻前の畳み込み LSTM 層 R_l^{t-1}，1 レイヤー上の畳み込み LSTM 層 R_{l+1}^t を受け取り，予測表現層 $\hat{A}_l^t$ で未来画像を生成する．

3.3　外挿

PredNet は 1 フレーム先の未来画像の生成のみではなく，外挿によって数フレーム先の未来画像の生成も可能である．図 4 に外挿の概要を示す．外挿では，PredNet に画像を入力し，1 フレーム先の未来画像を生成することで情報を蓄積する．その後，生成した未来画像を入力として受け取り，1 フレーム先の未来画像を生成する．生成した画像を次は入力として受け取り，1 フレーム先の未来画像を生成する．これを繰り返すことで数フレーム先の未来画像を生成する．外挿による数フレーム先の未来画像生成は，PredNet の生成画像を入力とするため，生成する時刻が外挿開始の時刻から離れるほど画像生成の精度は低下する．

4　評価実験

提案手法の有効性を確認するために 2 つの実験を行う．1 つ目は 1 フレーム先の未来画像の生成に関する評価実験，2 つ目は外挿における未来画像の生成に関する

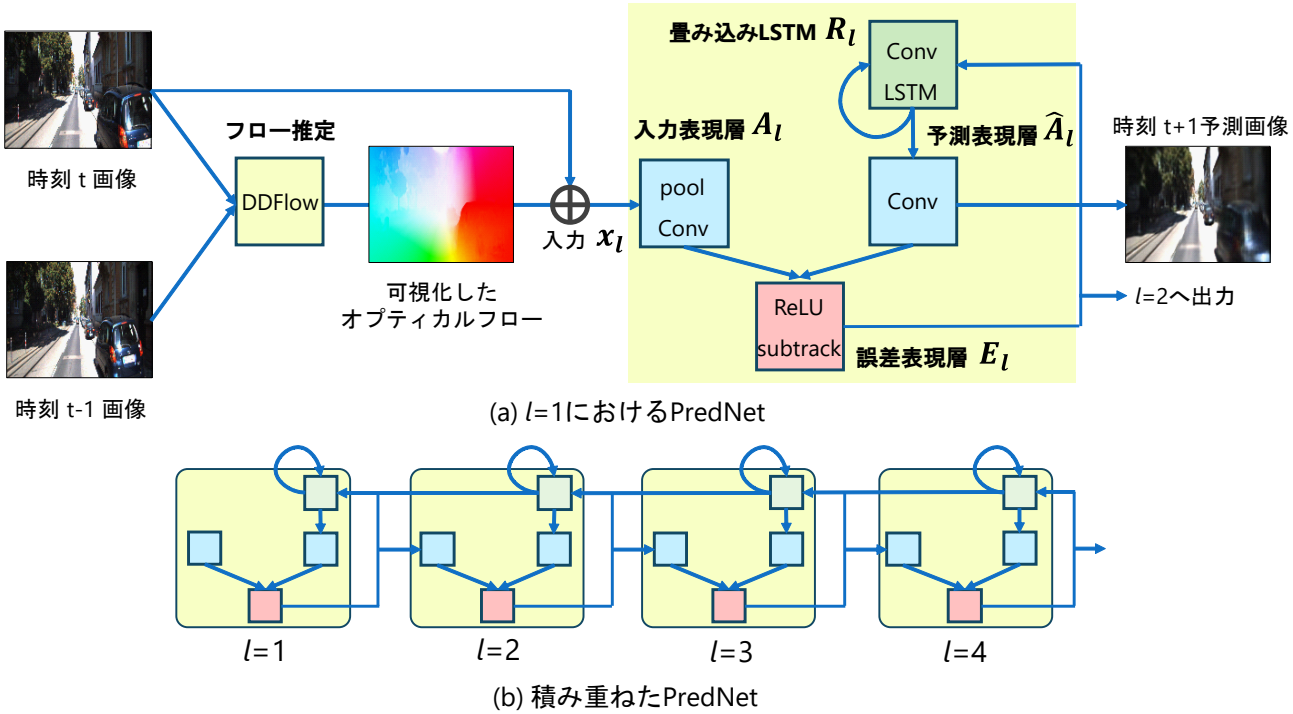

図 3　提案手法におけるネットワークの構成.

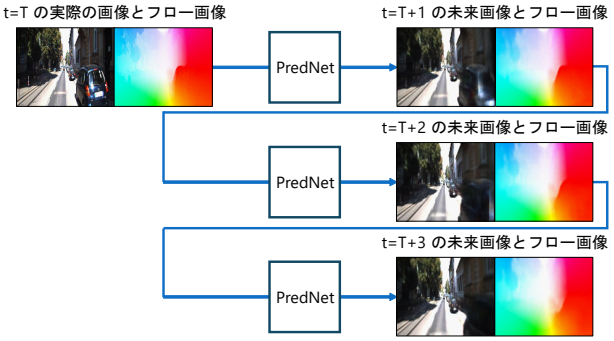

図 4　外挿時の概要.

実験である．比較手法は PredNet に画像のみを用いて学習したモデルと，画像と可視化したオプティカルフローの 2 つを用いて学習したモデルである．学習，評価データは，車載カメラで撮影した画像のデータセットである KITTI Dataset[19] を使用する．10Hz 間隔の RGB 画像を学習用として約 37,900 枚，評価用として約 3,800 枚使用した．評価用データは KITTI Dataset の中から加速減速，直進，カーブ，停止の 4 種類のアクションを選定し，どのようなアクションに提案手法が有効であるか検証する．評価指標には，$t+1$ 時刻に撮影した画像と生成した未来画像との平均二乗誤差 (MSE) を用いる．

4.1　1 フレーム先の未来画像生成

1 フレーム先の未来画像生成の実験結果を表 1 に示す．どのアクションにおいても提案手法が従来法を上回る結果となった．最も効果のあった停止のアクションでは，MSE を約 30% 減少させることができた．カーブのアクションでは MSE が約 15% 減少し，1 フレームの間で左右方向に変化の大きい場面でも有効であることが確認できた．

図 5 に生成した未来画像とその差分画像の例を示す．$t=2$ の差分画像より，通過する自動車の位置と背景の建物の差分が小さいことがわかる．

表 1　MSE の比較 (1×10^{-3}).

	加速減速	直進	カーブ	停止	平均
従来法	7.65	11.36	10.63	4.11	8.44
提案手法	6.99	10.49	8.95	2.86	7.32

4.2　数フレーム先の未来画像生成

外挿時の未来画像の生成における各時刻の MSE を表 2 に示す．t が増加するにつれて予測が困難になり MSE が増加するが，提案手法では $t=7$ のとき MSE を約 7% 減少させることが可能である．図 6 に，停止のアクションを用いて生成した数フレーム先の未来画像とその差分画像を示す．提案手法の差分画像より，背景の差分が小さいことがわかる．このことから動きの情報を追加することで動きのある物体と動きのない物体を認識し，高精度な未来画像生成が可能であることがわかる．

表 2　外挿時の MSE の比較 (1×10^{-3}).

	$t=2$	$t=3$	$t=4$	$t=5$	$t=6$	$t=7$
従来法	19.43	10.72	15.38	22.07	25.09	28.10
提案手法	11.81	10.31	14.49	17.60	20.46	23.15

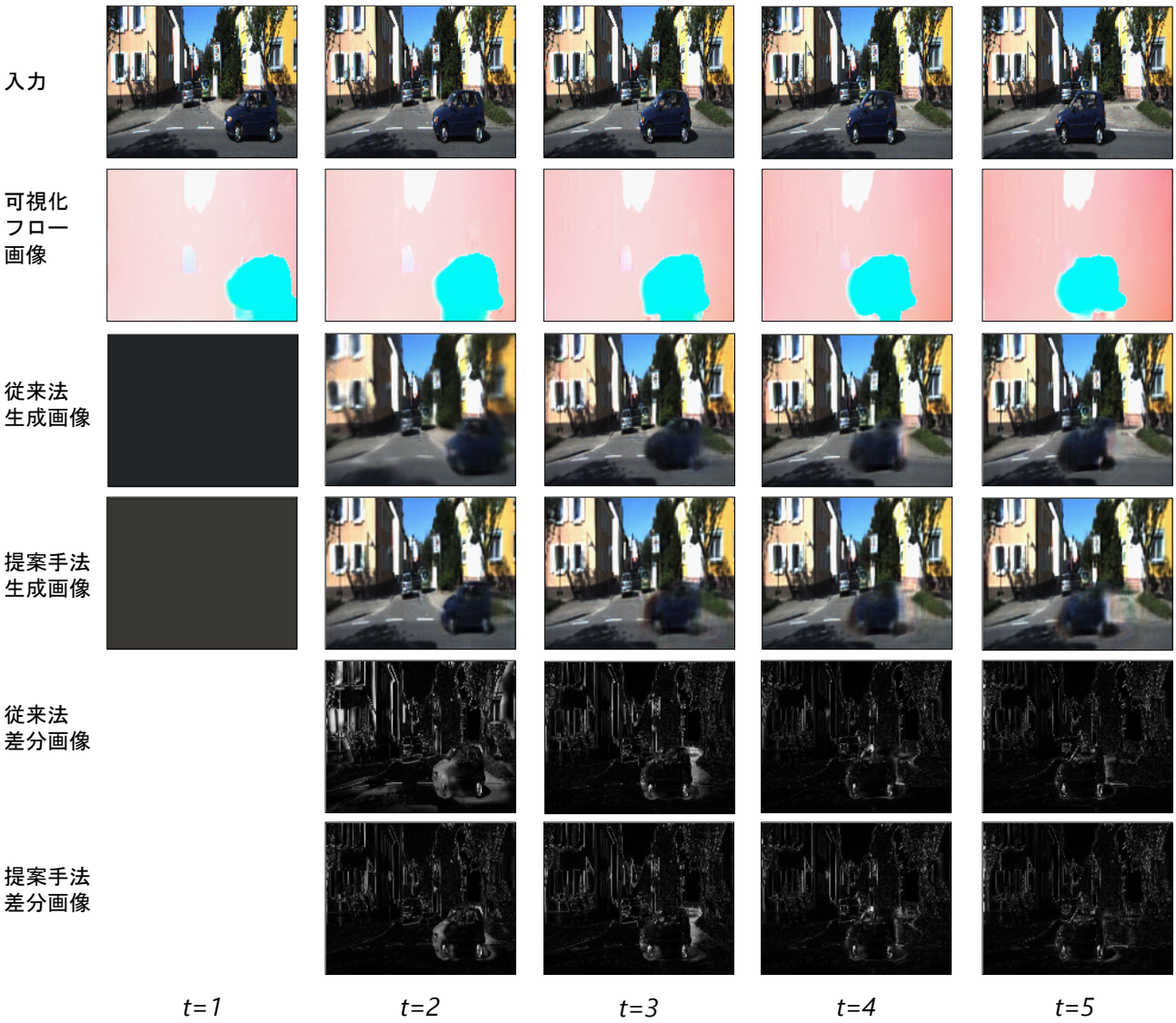

図 5　各手法による生成画像と入力との差分画像の例.

5　おわりに

本研究では動き情報を加えた PredNet による未来画像生成の高精度化を提案した．評価実験より，オプティカルフローを追加することで生成画像の精度が向上することを確認した．今後は，オプティカルフローの生成と未来画像生成を 1 つのネットワークで構成することを検討している．

参考文献

[1] K. Fragkiadaki, P. Agrawal, S. Levine, and J. Malik, “Learning visual predictive models of physics for playing billiards”, *ICLR*, 2016.

[2] C. Finn, I. Goodfellow, and S. Levine, “Unsupervised learning for physical interaction through video prediction”, *NIPS*, vol. 29, pp. 64–72, 2016.

[3] W. Lotter, G. Kreiman, and D. Cox, “Deep predictive coding networks for video prediction and unsupervised learning”, *ICLR*, 2017.

[4] V. Ruben, Y. Jimei, Z. Yuliang, S. Sungryull, X. Lin, and L. Honglak, “Learning to generate long-term future via hierarchical prediction”, *ICML*, 2017.

[5] S. Xingjian, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo, “Convolutional lstm network: A machine learning approach for precipitation nowcasting”, *NIPS*, pp. 802–810, 2015.

[6] P. Liu, I. King, M. R. Lyu, and J. Xu, “Ddflow: Learning optical flow with unlabeled data distillation”, *AAAI*, 2019.

[7] Y. Yao, M. Xu, Y. Wang, D. J. Crandall, and E. M. Atkins, “Unsupervised traffic accident detection in first-person videos”, *IROS*, 2019.

[8] F.-H. Chan, Y.-T. Chen, Y. Xiang, and M. Sun, “Anticipating accidents in dashcam videos”, *ACCV*, 2016.

[9] G. Xie, H. Gao, L. Qian, B. Huang, K. Li, and J. Wang, “Vehicle trajectory prediction by integrating physics-and maneuver-based approaches using interactive multiple models”, *IEEE Trans. Ind. Electron*, vol. 65, no. 7, pp. 5999–6008, 2017.

[10] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets”, *NIPS*, 2014.

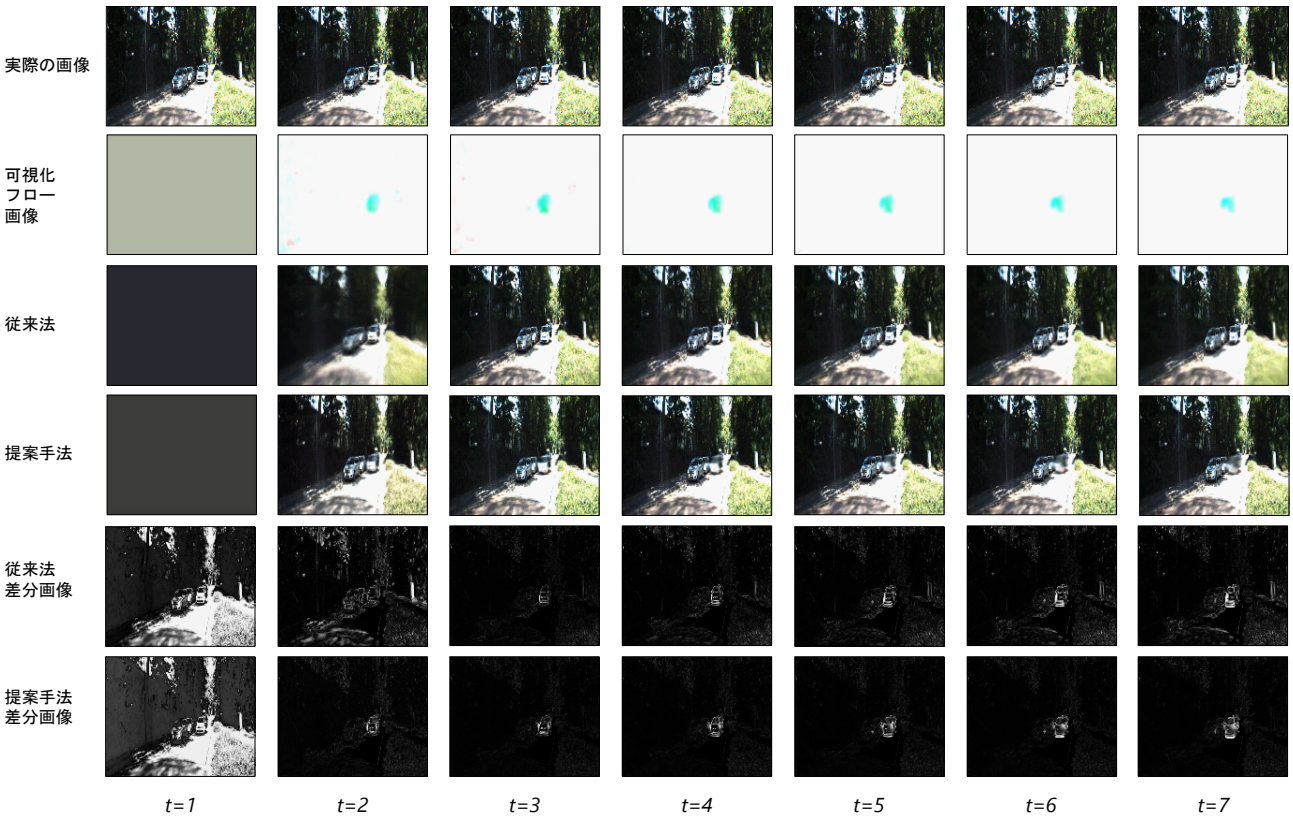

図 6　外挿時の生成画像の例.

[11] X. Liang, L. Lee, W. Dai, and E. P. Xing, “Dual motion gan for future-flow embedded video prediction”, *ICCV*, 2017.

[12] K. D. P., and W. Max, “Auto-Encoding Variational Bayes.” *ICLR*, 2014.

[13] J. Oh, X. Guo, H. Lee, R. L. Lewis, and S. Singh, “Action-Conditional Videos Prediction using Deep Networks in Atari Games”, *NIPS*, 2015.

[14] R. Villegas, J. Yang, Y. Zou, S. Sohn, X. Lin, and H. Lee, “Learning to generate long-term future via hierarchical prediction”, *Proc. of PMLR*, vol. 70, pp. 3560–3569, 2017.

[15] B. D. Lucas, and T. Kanade, “An iterative image registration technique with an application to stereo vision”, *IJCAI*, 1981.

[16] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, “Flownet 2.0: Evolution of optical flow estimation with deep networks”, *CVPR*, 2017.

[17] Z. Ren, J. Yan, B. Ni, B. Liu, X. Yang, and H. Zha, “Unsupervised deep learning for optical flow estimation”, *Proc. of AAAI Conf. on Artif Intell*, vol. 31, no. 1, 2017.

[18] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz, “PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume”, *CVPR*, 2018.

[19] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, “Vision meets robotics: The kitti dataset”, *IJRR*, 2013.