Learning Knowledge-Rich Sequential Model for Planar Homography Estimation in Aerial Videos

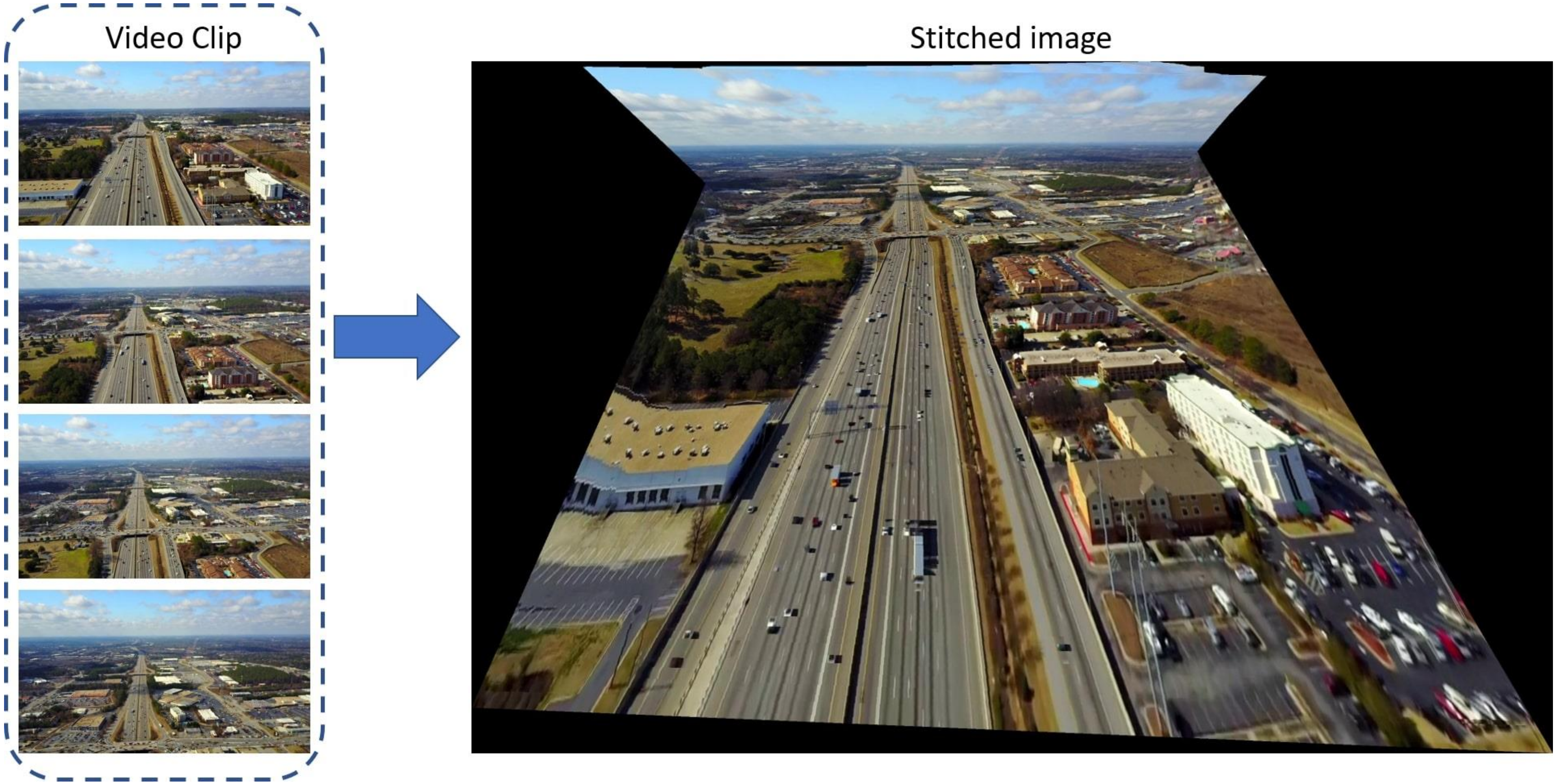
Pu Li and Xiaobai Liu

San Diego State University

Dataset and code available: https://github.com/Paul-LiPu/DeepVideoHomography

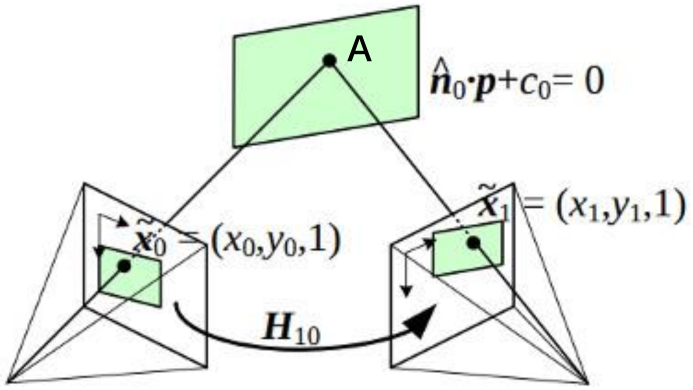
Our task

- Planar Homograph Estimation in Aerial videos



Background

- Homography estimation and Image stitching



Images for the same planar object by different camera position

$$s\begin{bmatrix} x_1 \\ y_1 \\ 1 \end{bmatrix} = H_{10}\begin{bmatrix} x_0 \\ y_0 \\ 1 \end{bmatrix} = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix}\begin{bmatrix} x_0 \\ y_0 \\ 1 \end{bmatrix}$$

transformation between two image

Illustration image from: https://docs.opencv.org/master/d9/dab/tutorial_homography.html

Related works

- Geometry-based homography estimators
  - Feature-based method, e.g., ORB[1] + RANSAC[2]
  - Direct pixel-based method, e.g., ECC[3]
- Learning-based homography estimators
  - Deep Homography[4]
  - Unsupervised deep homography[5]

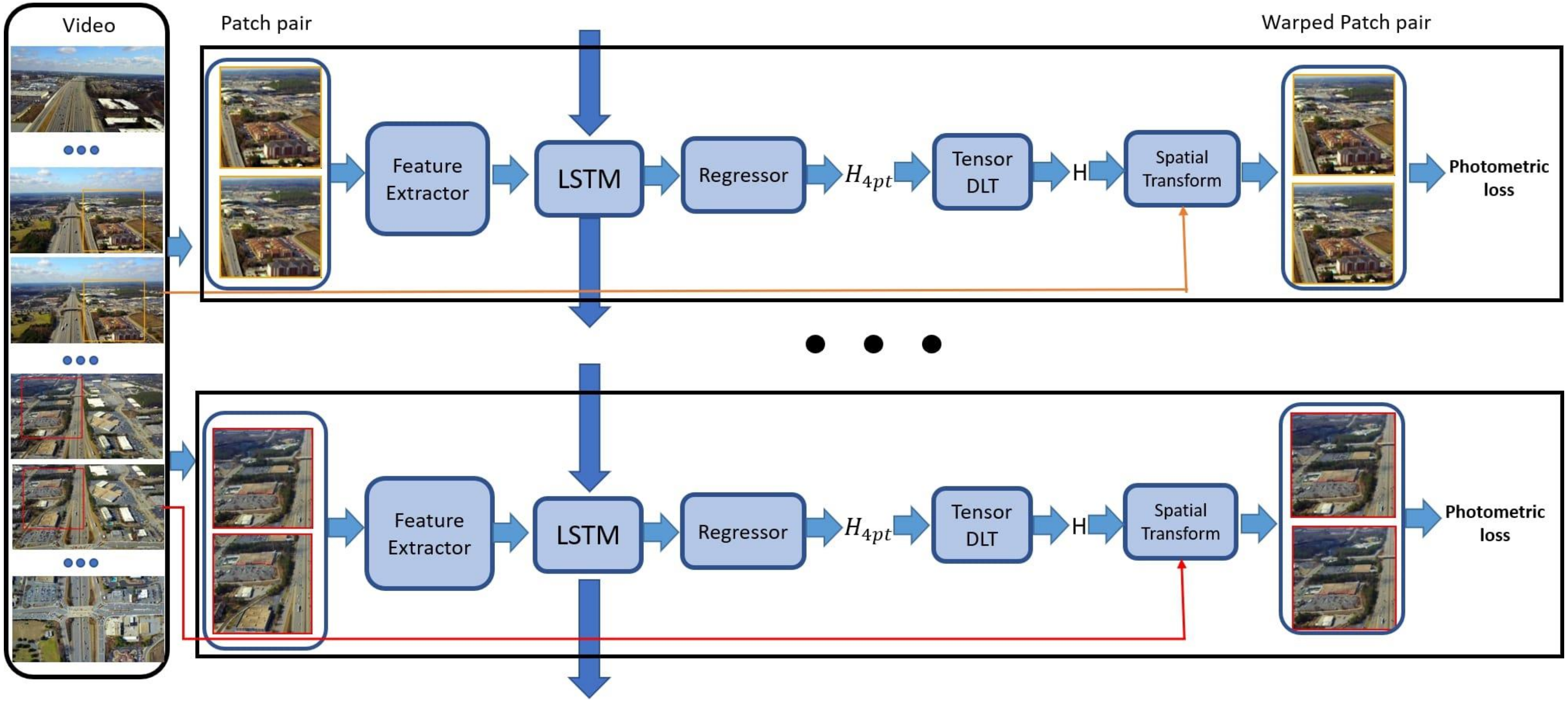
Limitation of previous works

- The previous works are designed for image pairs, and fail to model the temporal or sequential knowledge for homography estimation tasks.

- Existing deep network based homography estimators (supervised or unsupervised) suffer from overfitting issue.

Our Methods

Sequential Homograph Estimation for Aerial Videos



Our Methods

- Knowledge based regularization terms
  - Spatial regularization
    - $R_p(I) = \sum_t \sum_{a \neq b} ||H_{a,t,t+1} - H_{b,t,t+1}||_1$
  - Scale regularization
    - $R_s(I) = \sum_t \sum_{<m,n>} ||H_{m,t,t+1} - H_{n,t,t+1}||_1$
  - Temporal regularization
    - $R_{t1}(I) = \sum_t \sum_{<k,l>} ||H_{k,t,t+1} - H_{l,t+1,t+2}||_1$
    - $R_{t2}(I) = \sum_t \sum_{s=t+2}^{t+K-1} \sum_{x \in I_t} ||I_t(x) - I_s(H_{[t,s]} \cdot x)||_1$

Experiments: Quantitative Results

$$MACE = \frac{1}{\sum_{i=1}^{M}(N_i-1)} \sum_{m=1}^{M} \sum_{t=2}^{N_i} \left(\frac{1}{K_t} \sum_{j=1}^{K_t} \|\hat{x_j}^t - x_j^t\|_2\right)$$

Experiments	MACE
Identity	35.69
ORB2+RANSAC	12.02
BASE	13.66
REG-T	9.95
REG-P	11.57
REG-S	11.44
REG-ALL	9.16
LSTM	14.10
LSTM-REG-ALL	**8.77**

MACE performances of different experiments.

Dataset information

- 22 testing clips (1280 x720 pixels)
- Key points annotated every 30 frames.

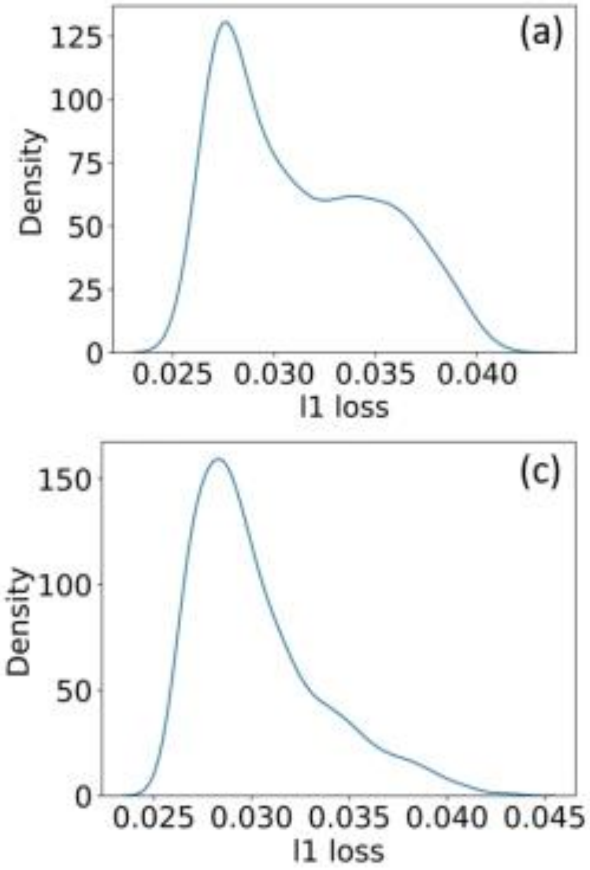
Experiments: Qualitative results

BASE

REG-T

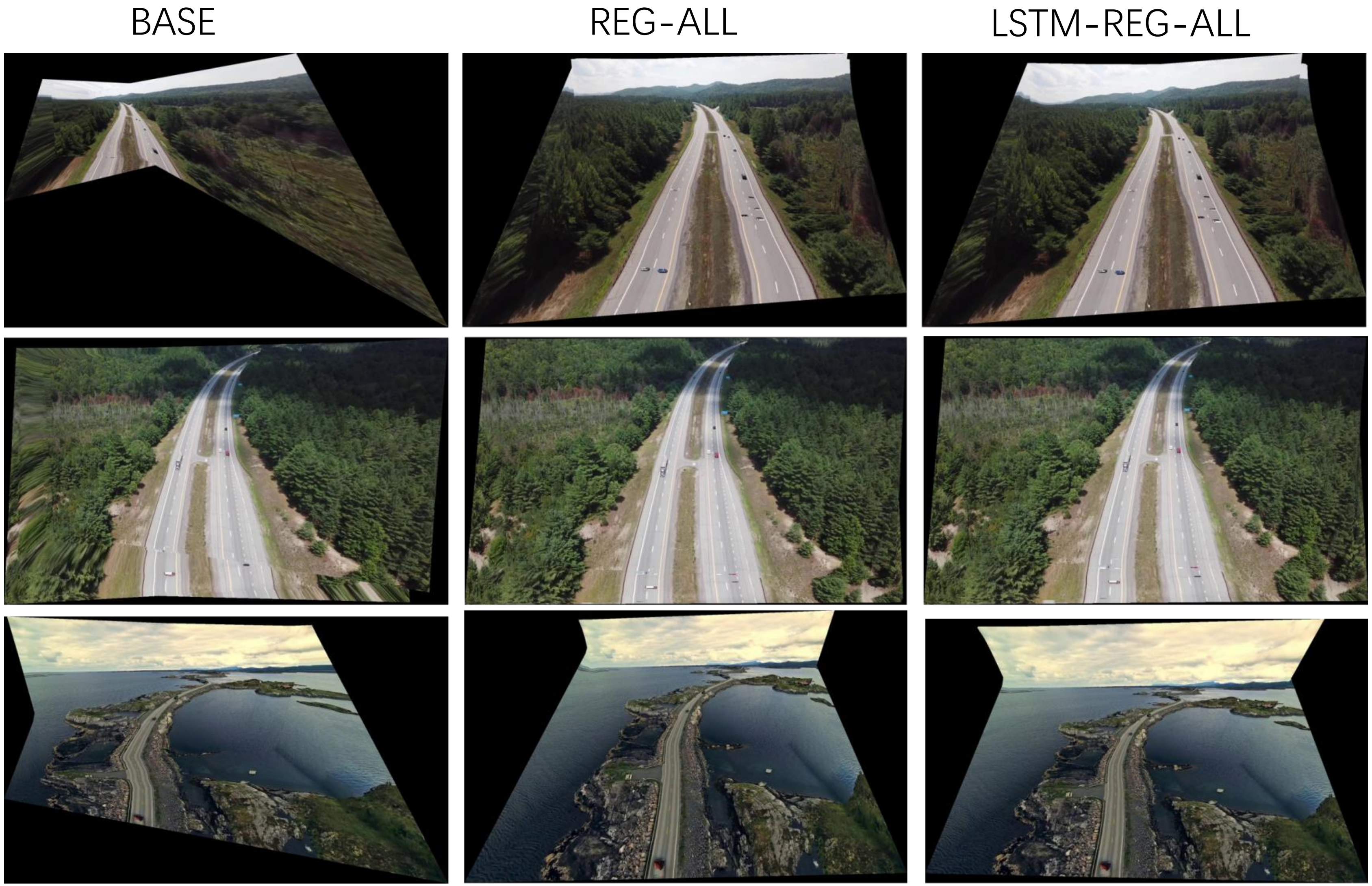
REG-P

REG-S



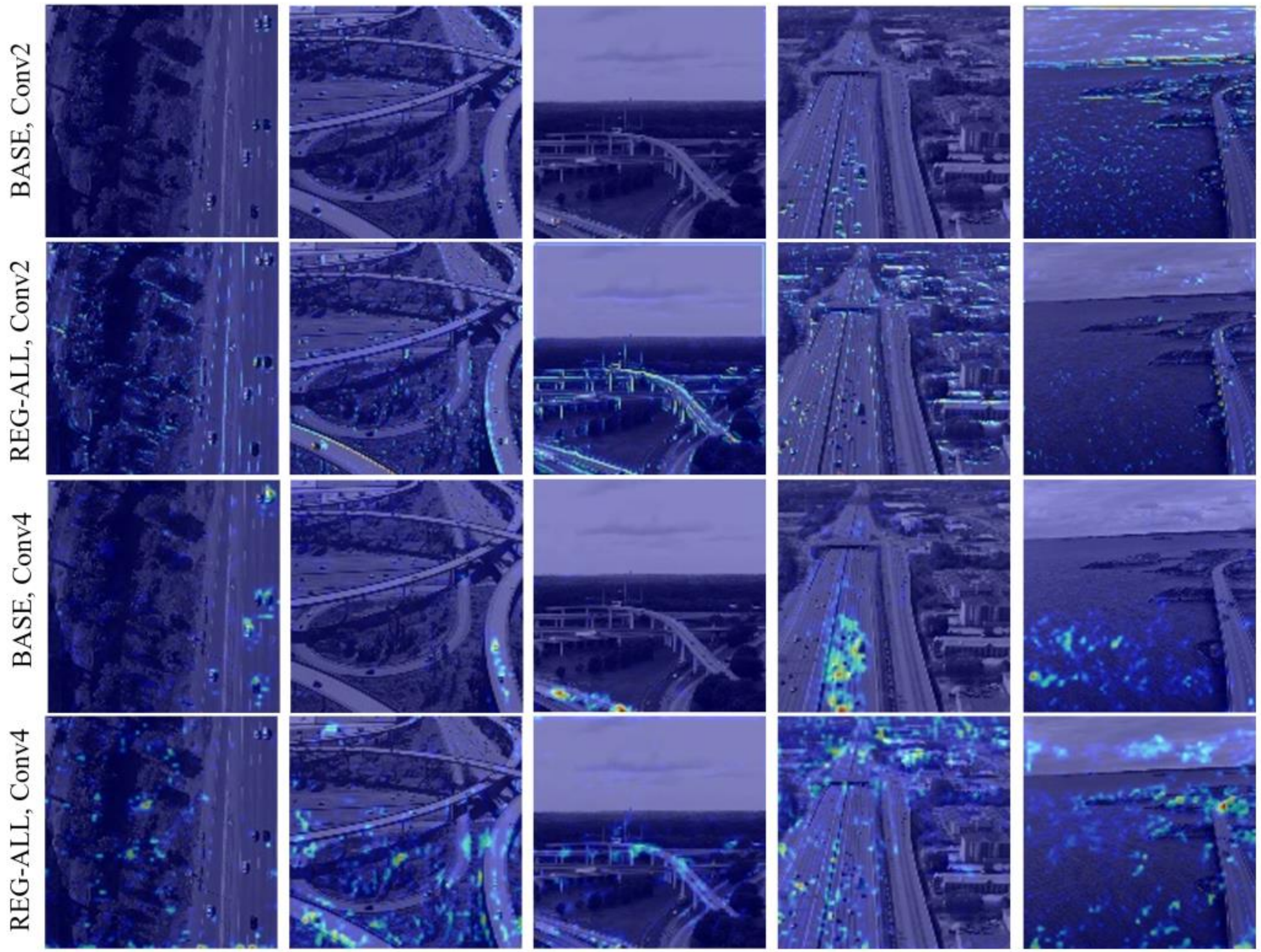
Photometric loss distribution from one thousand pairs of patches in one pair of image.

Experiments: Qualitative results



Examples Image stitching result.

Experiments: Qualitative results



Visualization of network activation by GradCam[6].

Contributions

- We reformulate the homography estimation of aerial videos to be a sequence-to-sequence task and develop a LSTM network to estimate the sequence of homography parameters.
- We employ a set of spatial-scale-temporal knowledge to regularize training of the LSTM model and empirically validate its superior performance over alternative methods on challenging aerial videos.

Reference

[1] E. Rublee, V. Rabaud, K. Konolige, and G. R. Bradski, "Orb: An efficient alternative to sift or surf." in ICCV, vol. 11, no. 1. Citeseer, 2011, p. 2.

[2] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[3] G. D. Evangelidis and E. Z. Psarakis, "Parametric image alignment using enhanced correlation coefficient maximization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 10, pp. 1858–1865, 2008

[4] D. DeTone, T. Malisiewicz, and A. Rabinovich, "Deep image homography estimation," arXiv preprint arXiv:1606.03798, 2016.

[5] T. Nguyen, S. W. Chen, S. S. Shivakumar, C. J. Taylor, and V. Kumar, "Unsupervised deep homography: A fast and robust homography estimation model," IEEE Robotics and Automation Letters, vol. 3, no. 3, pp. 2346–2353, 2018.

[6] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in Proceedings of the IEEE international conference on computer vision, 2017, pp. 618–626.