Unsupervised Homography Estimation on Multimodal Image Pair via Alternating Optimization

*Corresponding author

Sanghyeob Song[1,3] **Jaihyun Lew**[1] **Hyemi Jang**[2] **Sungroh Yoon**[1,2*]

[1] Interdisciplinary Program in Artificial Intelligence, Seoul National University

[2] Department of Electrical and Computer Engineering, Seoul National University

[3] Samsung Electro-Mechanics

`{songsang7, fudojhl, wkdal9512, sryoon}@snu.ac.kr`

Online poster

https://neurips.cc/virtual/2024/poster/92937

Unsupervised Homography Estimation on Multimodal Image Pair via Alternating Optimization

*Corresponding author

Sanghyeob Song[1,3] **Jaihyun Lew**[1] **Hyemi Jang**[2] **Sungroh Yoon**[1,2*]

[1]Interdisciplinary Program in Artificial Intelligence, Seoul National University

[2]Department of Electrical and Computer Engineering, Seoul National University

[3]Samsung Electro-Mechanics

`{songsang7, fudojhl, wkdal9512, sryoon}@snu.ac.kr`

"Hello everyone. Today, I'll be presenting my paper, titled *"Unsupervised Homography Estimation on Multimodal Image Pair via Alternating Optimization."*
My name is Sanghyoeb Song, and I'm honored to share my work with you.
Before I begin, I'll be using a speech AI(F5 TTS) for clearer audio. Now, let's get started."

Online poster

https://neurips.cc/virtual/2024/poster/92937

Contents

- **Introduction**
- **Preliminaries**
- **Method**
- **Experiments**

Contents

- **Introduction**
- **Preliminaries**
- **Method**
- **Experiments**

Introduction

- Homography Estimation

Homography Estimation is the process of determining a transformation matrix that aligns two images captured from different perspectives.

🟩 : Input

🟧 : Output

Image 1
(Moving image)

Geometric transformation (Homography)

$$\begin{pmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & 1 \end{pmatrix}$$

Image 2
(Fixed image)

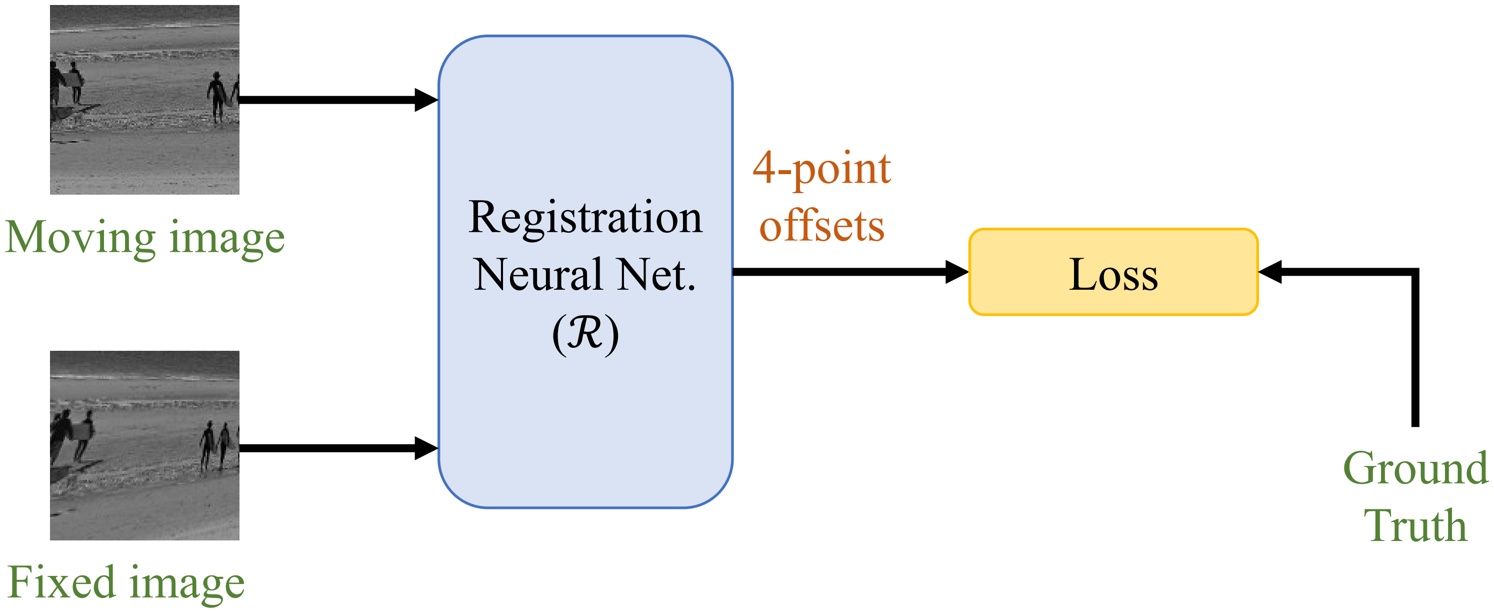
Introduction

- End-to-End Learning Approaches – Supervised Learning

  - Introduced by the paper *Deep image homography estimation* [1].

  - Finding 1-homography from 4-corresponding pairs is a determined problem.

  - Direct Linear Transformation (DLT) is used to convert 4-corresponding pairs to 1-homography.



[1] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Deep image homography estimation. CoRR, abs/1606.03798, 2016.

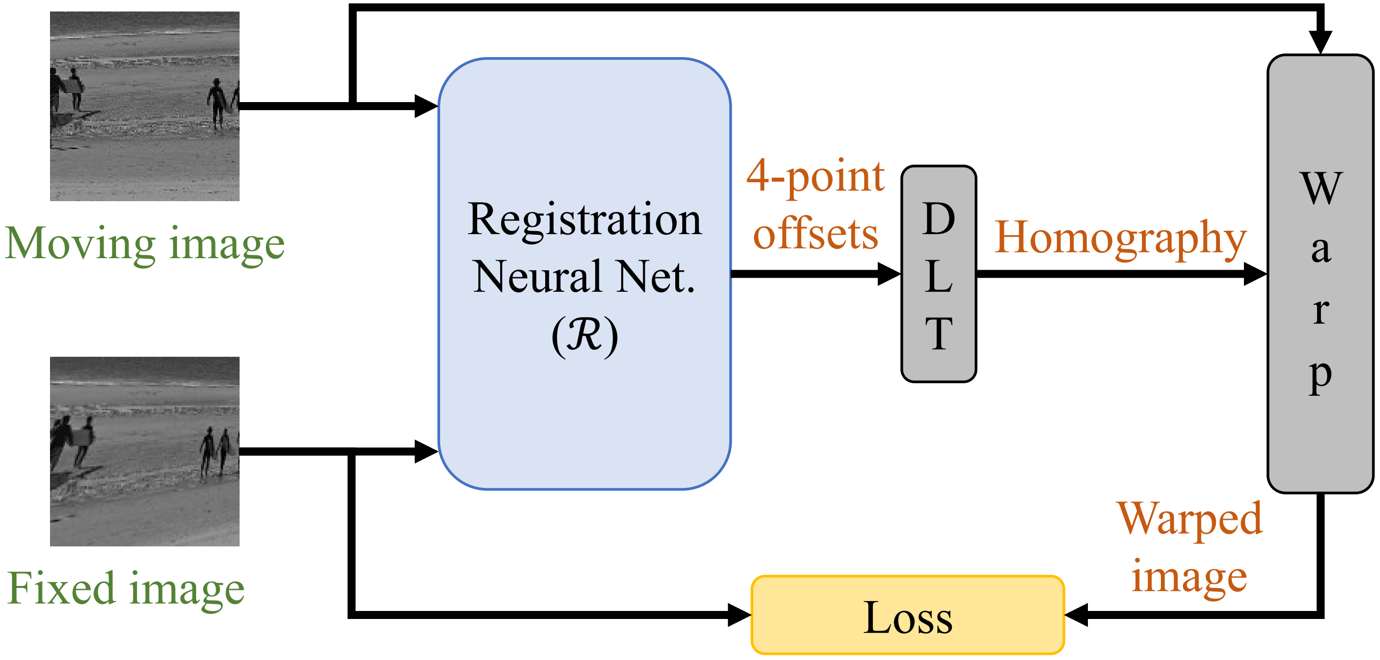
Introduction

- End-to-End Learning Approaches – Unsupervised Learning

  - Introduced by the paper *Unsupervised deep homography: A fast and robust homography estimation model* [2].

  - Finding 1-homography from 4-corresponding pairs is a determined problem.

  - Direct Linear Transformation (DLT) is used to convert 4-corresponding pairs to 1-homography.

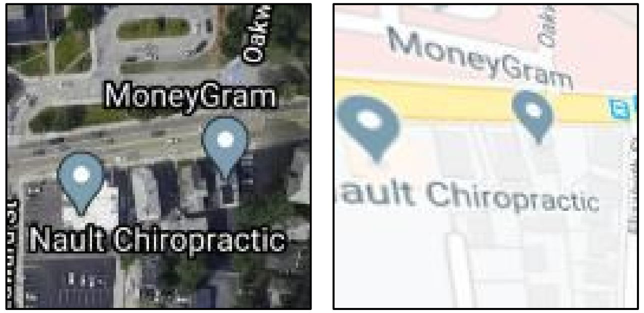


[2] Ty Nguyen, Steven W. Chen, Shreyas S. Shivakumar, Camillo Jose Taylor, and Vijay Kumar. Unsupervised deep homography: A fast and robust homography estimation model. IEEE Robotics Autom. Lett., 3(3):2346–2353, 2018.

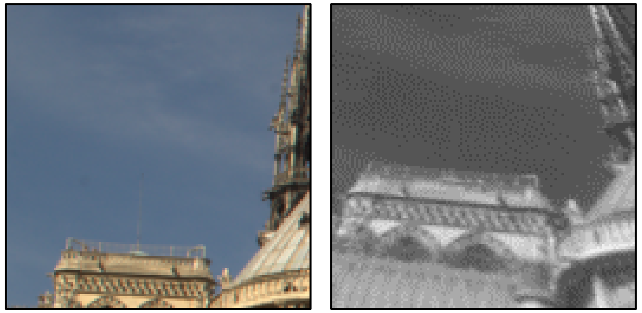
Introduction

- Multimodal Image Pair

Multimodal image pairs refer to pairs of images from different domains, offering complementary insights for a deeper analysis.



Optical – Map



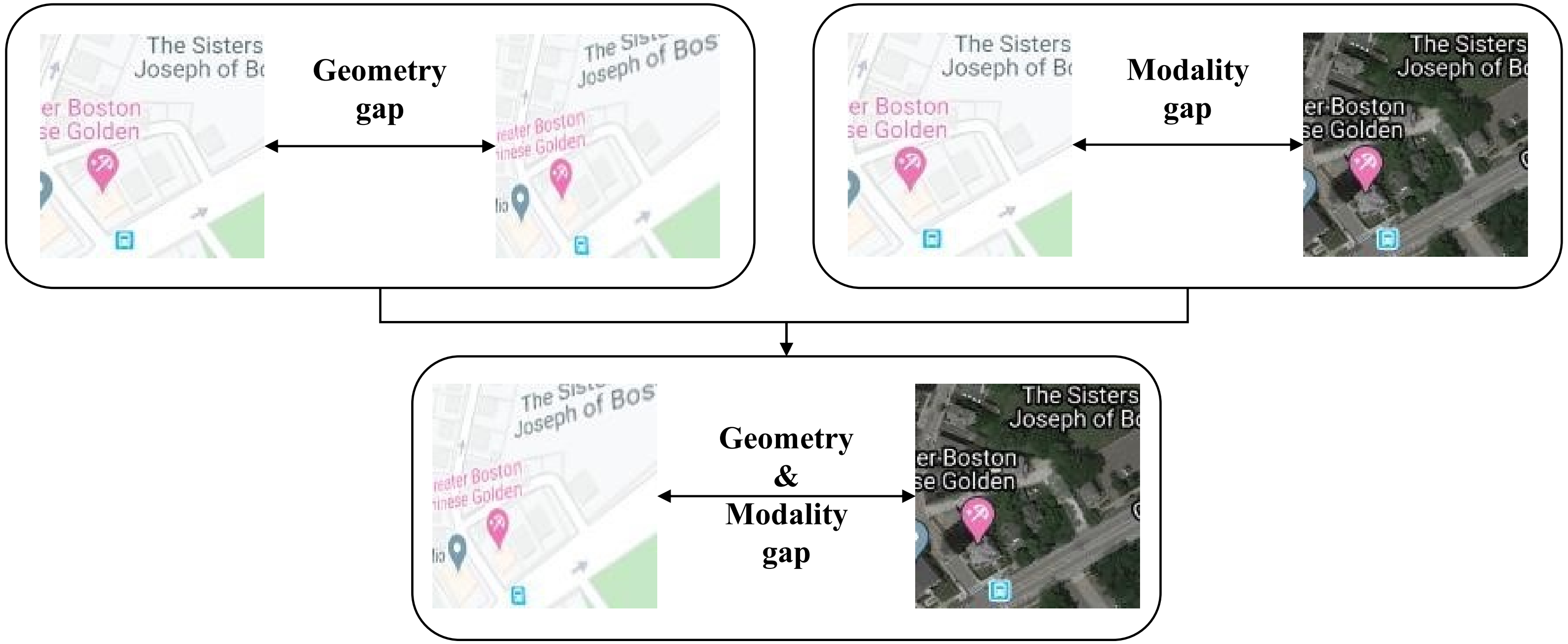
RGB – NIR

Introduction

- Unsupervised Homography Estimation on Multimodal Image Pair

  - Cases of misaligned multimodal image pairs.

  - No ground-truth data → Unsupervised learning



Contents

- Introduction
- **Preliminaries**
- Method
- Experiments

Preliminaries

- Trivial Solution Problem

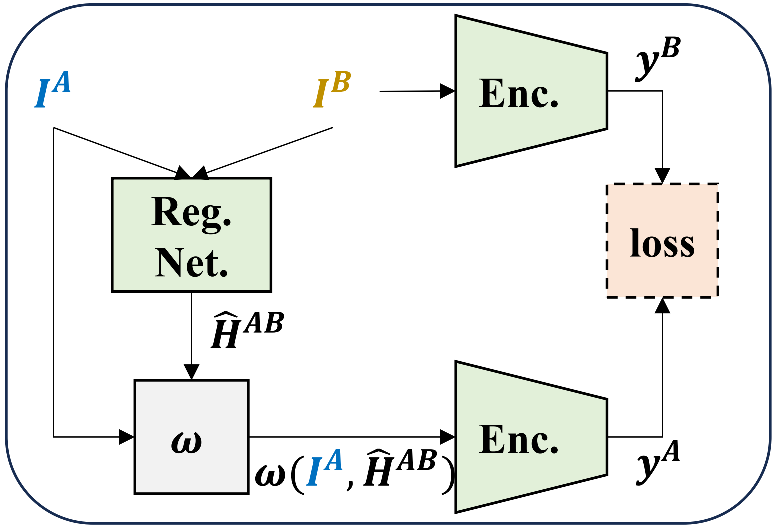
  - To solve our problem, there is a simple and straightforward approach.

    : Add encoder (or translator) to map images to a common space, enabling simple calculation of L1 or L2 loss.

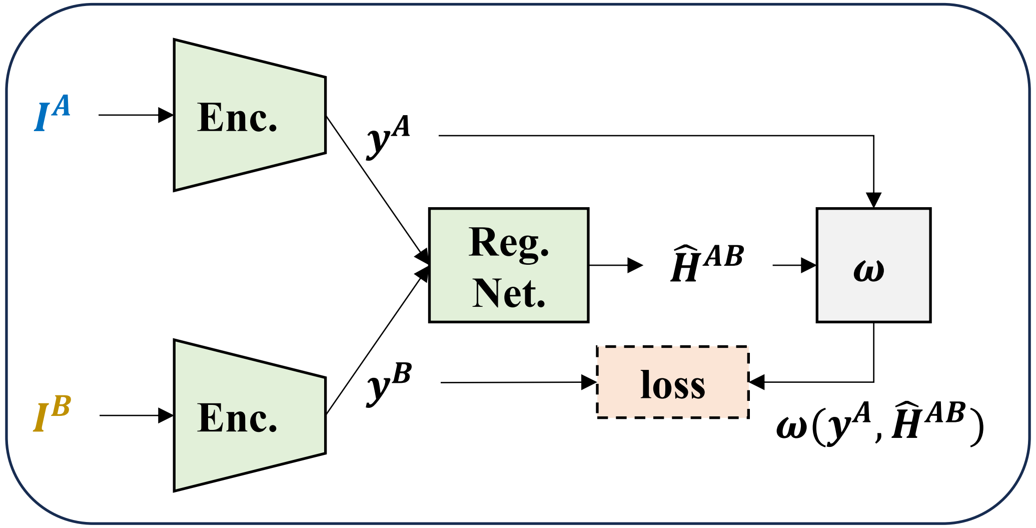
    : There are 2 options for placing encoder(s)

① Just before the loss function ② Between each image and registration network($\mathcal{R}$)

①



②



Preliminaries

- Trivial Solution Problem

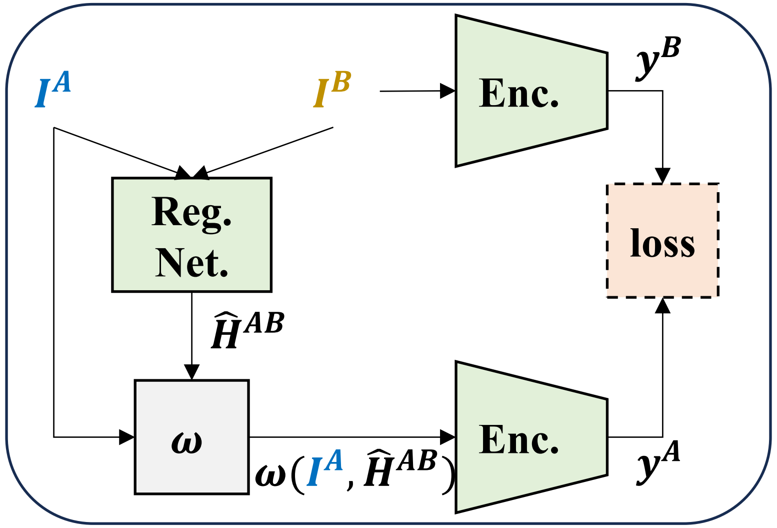
- However, both cases fail to solve the problem due to trivial solutions:

Encoders output a constant value, regardless of input. ($\boldsymbol{y}^{\boldsymbol{A}} = \boldsymbol{y}^{\boldsymbol{B}} = \boldsymbol{v}$)

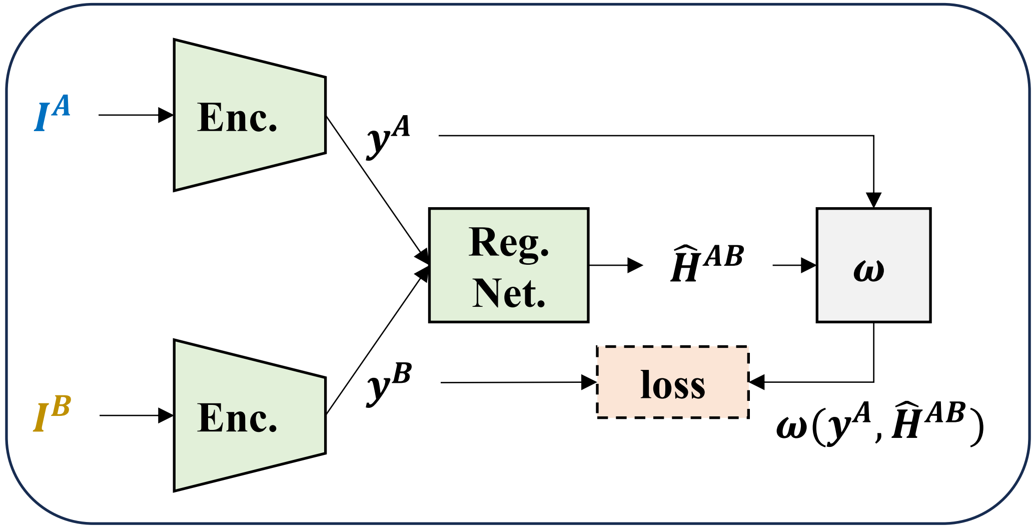
Registration network ($\mathcal{R}$) outputs an identity matrix as the homography. ($\widehat{\boldsymbol{H}}^{\boldsymbol{AB}} = \boldsymbol{I}$)

- A different approach is needed to avoid trivial solutions.

①



②

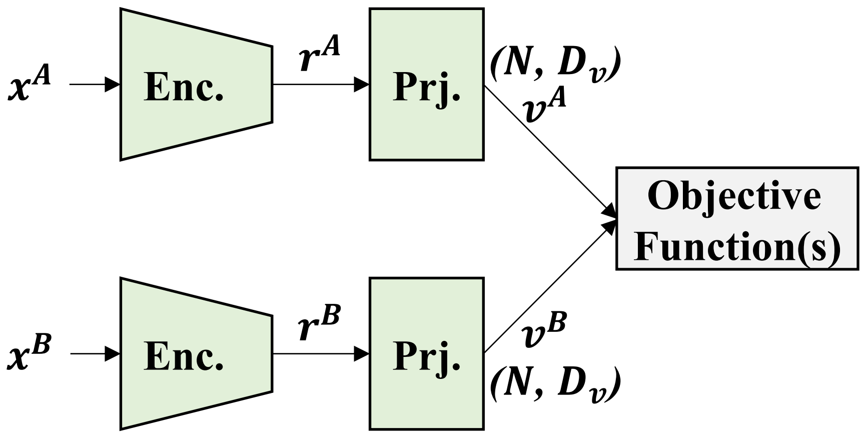


Preliminaries

- Siamese Self-Supervised Learning

  - Extracting the same features from a pair of partially different images.

  - SimCLR [3], Barlow Twins [4], VIC-Reg [5], ...



[3] TingChen, SimonKornblith, MohammadNorouzi, andGeoffreyE.Hinton. Asimpleframework for contrastive learning of visual representations. In Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event, volume 119 of Proceedings of Machine Learning Research, pages 1597–1607. PMLR, 2020.

[4] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self supervised learning via redundancy reduction. In Marina Meila and Tong Zhang, editors, Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event, volume 139 of Proceedings of Machine Learning Research, pages 12310–12320. PMLR, 2021.

[5] Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. In The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022. OpenReview.net, 2022.

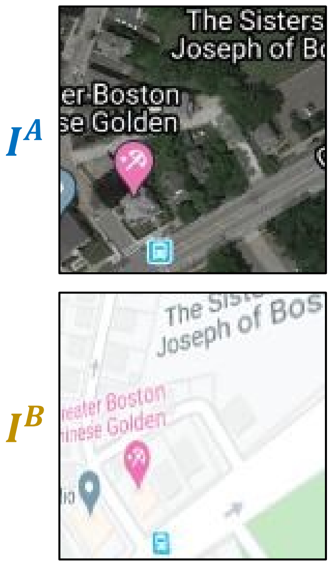
Contents

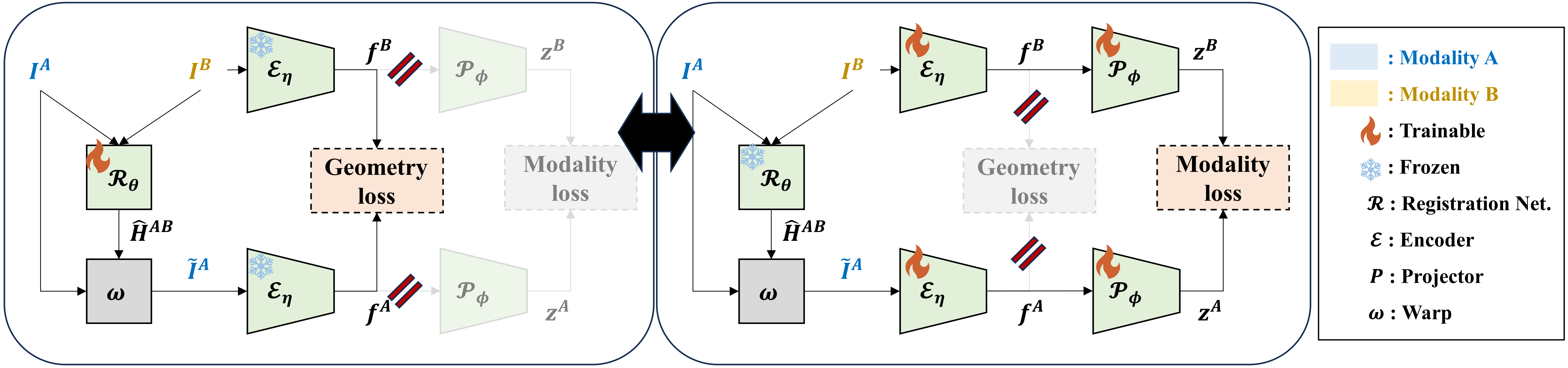
- Introduction
- Preliminaries
- **Method**
- Experiments

Method

- Framework - Alternating Optimization (AltO)

$$\begin{cases} \textbf{GL Phase}: & \theta_t \leftarrow \underset{\theta}{\mathrm{argmin}}[\mathrm{GeometryGap}(\theta_{t-1}, \eta_{t-1}, \phi_{t-1})] \\ \textbf{MARL Phase}: & \eta_t, \phi_t \leftarrow \underset{\eta,\phi}{\mathrm{argmin}}[\mathrm{ModalityGap}(\theta_t, \eta_{t-1}, \phi_{t-1})] \end{cases}$$





Method

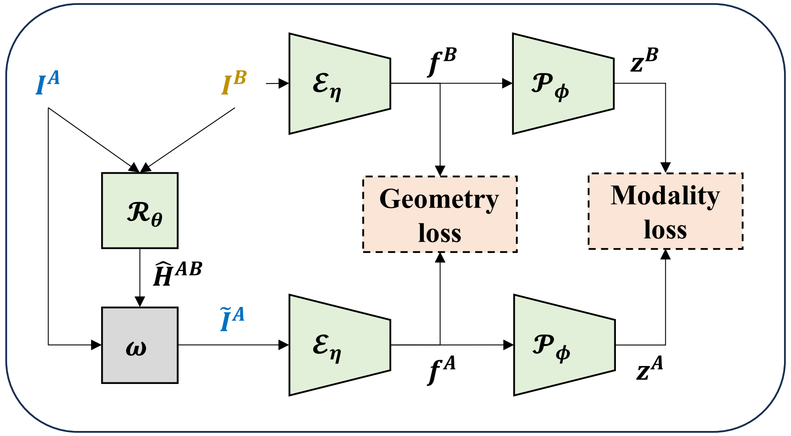
- Components – Losses

Geometry loss : Spatially extended version of the Barlow Twins loss

$$\mathcal{L}_g = \mathbb{E}_n\left[\sum_i (1 - C_{(n,ii)})^2 + \lambda \sum_i \sum_{j\neq i} C^2_{(n,ij)}\right], \text{ where } C_{(n,ij)} = \frac{\sum_{h,w}(\bar{f}^A_{(n,i,h,w)} \bar{f}^B_{(n,j,h,w)})}{\sqrt{\sum_{h,w}(\bar{f}^A_{(n,i,h,w)})^2}\sqrt{\sum_{h,w}(\bar{f}^B_{(n,j,h,w)})^2}}$$

Modality loss : The same as the Barlow Twins loss

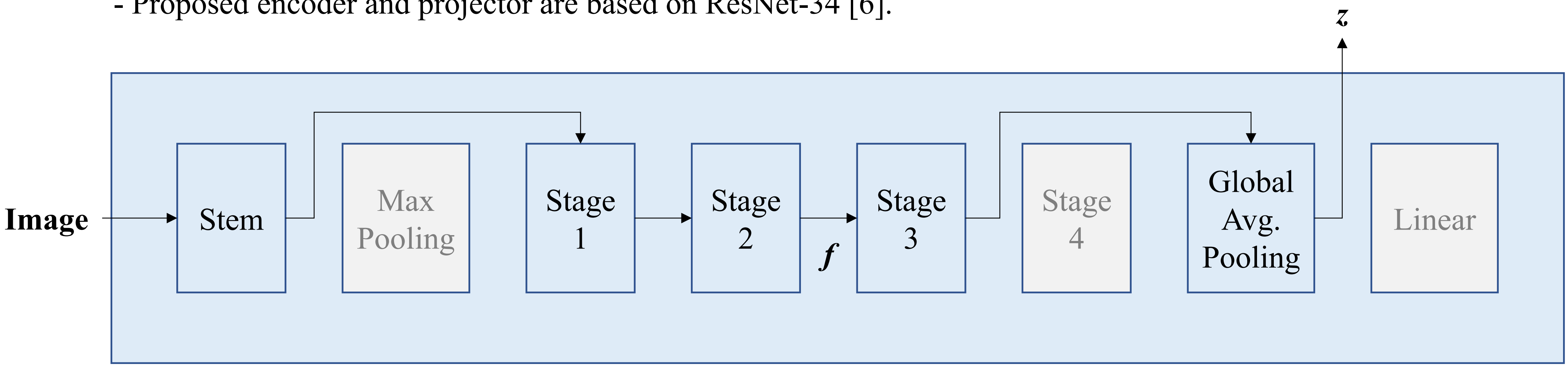
$$\mathcal{L}_m = \sum_i (1 - C_{ii})^2 + \lambda \sum_i \sum_{j\neq i} C^2_{ij}, \text{ where } C_{ij} = \frac{\sum_n(\bar{z}^A_{(n,i)} \bar{z}^B_{(n,j)})}{\sqrt{\sum_n(\bar{z}^A_{(n,i)})^2}\sqrt{\sum_n(\bar{z}^B_{(n,j)})^2}}$$



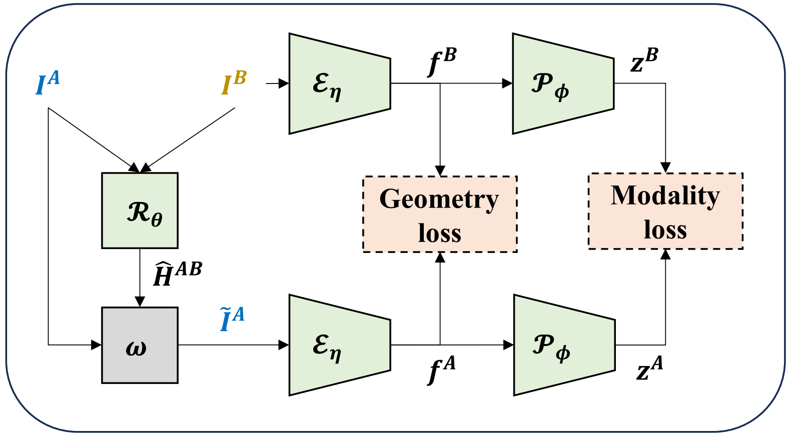
Method

- Components – Encoder and Projector

  - Proposed encoder and projector are based on ResNet-34 [6].



ResNet-34



[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016, pages 770–778. IEEE Computer Society, 2016.

Contents

- Introduction
- Preliminaries
- Method
- **Experiments**

Experiments

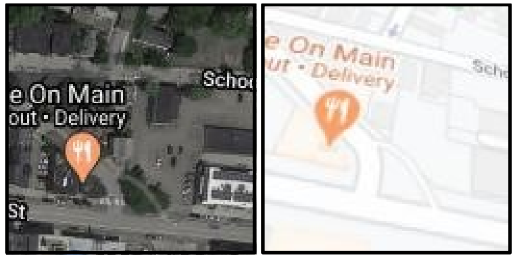
- Evaluation Metric

  - Mean Average Corner Error (MACE)

$$\mathrm{MACE}(H^{AB}, \hat{H}^{AB}) = \mathop{\mathbb{E}}_{n \in N}\left[\mathop{\mathbb{E}}_{c \in \mathcal{C}}\left[||\omega(c, H^{AB}) - \omega(c, \hat{H}^{AB})||_2\right]\right]$$

- Datasets

  - Google Map [7]
  - Google Earth [7]
  - Deep NIR [8]



(a) Google Map

(b) Google Earth

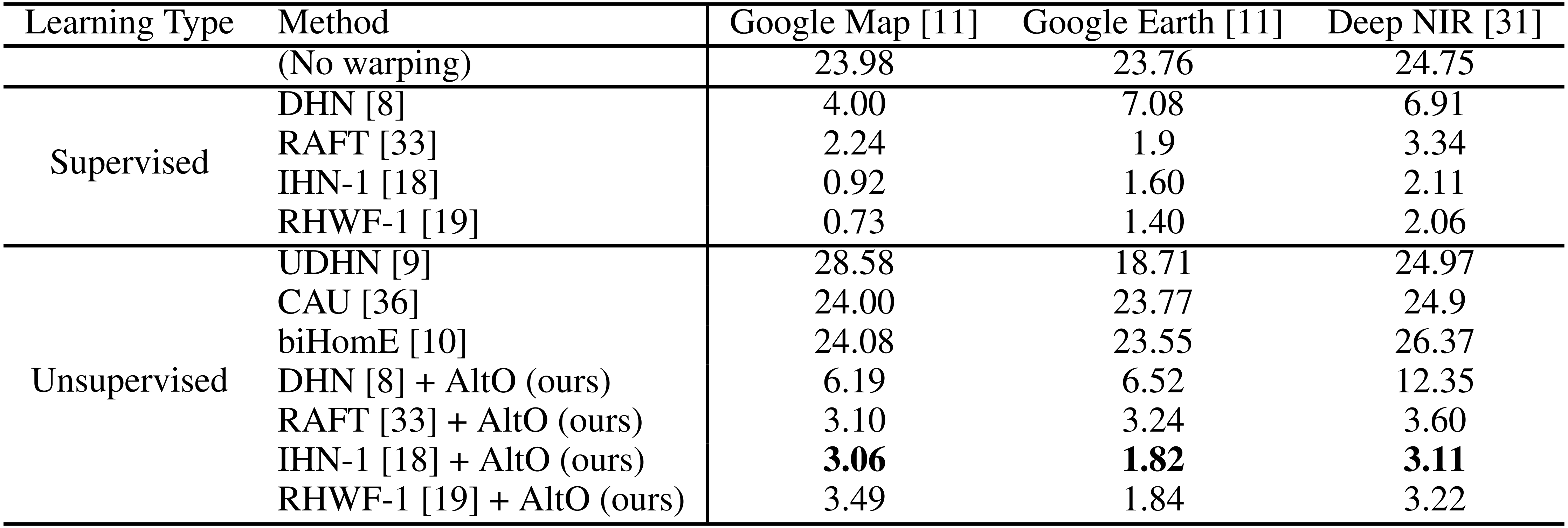
(c) Deep NIR

[7] Yiming Zhao, Xinming Huang, and Ziming Zhang. Deep lucas-kanade homography for multimodal image alignment. In IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021, pages 15950–15959. Computer Vision Foundation / IEEE, 2021.

[8] Inkyu Sa, Jong Yoon Lim, Ho Seok Ahn, and Bruce A. MacDonald. deepnir: Datasets for generating synthetic NIR images and improved fruit detection system using deep learning techniques. Sensors, 22(13):4721, 2022.

Experiments

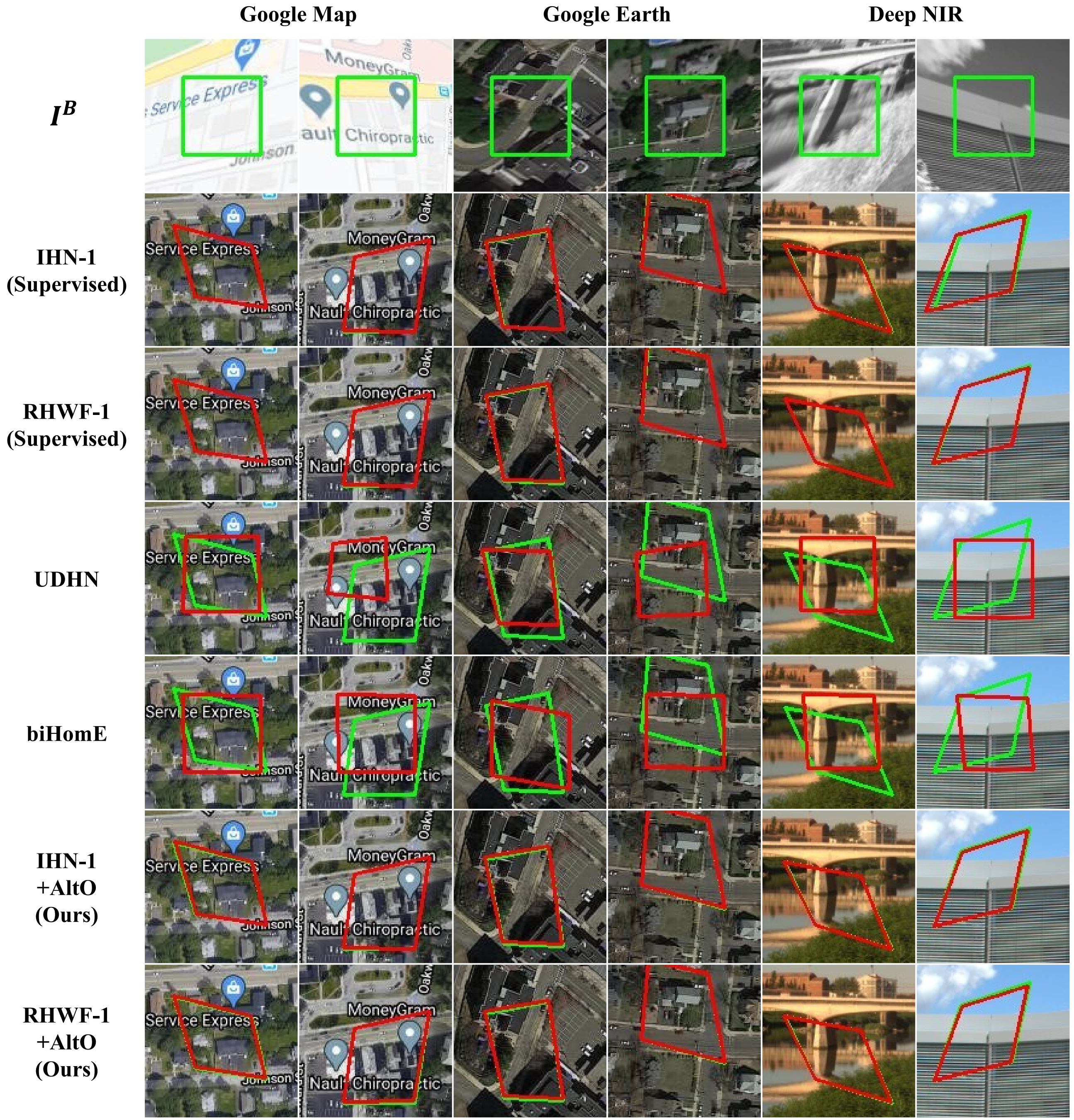
- Quantitative Result



Learning Type	Method	Google Map [11]	Google Earth [11]	Deep NIR [31]
	(No warping)	23.98	23.76	24.75
Supervised	DHN [8]	4.00	7.08	6.91
	RAFT [33]	2.24	1.9	3.34
	IHN-1 [18]	0.92	1.60	2.11
	RHWF-1 [19]	0.73	1.40	2.06
Unsupervised	UDHN [9]	28.58	18.71	24.97
	CAU [36]	24.00	23.77	24.9
	biHomE [10]	24.08	23.55	26.37
	DHN [8] + AltO (ours)	6.19	6.52	12.35
	RAFT [33] + AltO (ours)	3.10	3.24	3.60
	IHN-1 [18] + AltO (ours)	**3.06**	**1.82**	**3.11**
	RHWF-1 [19] + AltO (ours)	3.49	1.84	3.22

Experiments

- Qualitative Result



🟩 : Ground Truth

🟥 : Prediction

Thank you

Online poster

https://neurips.cc/virtual/2024/poster/92937