浙江大學
ZHEJIANG UNIVERSITY

EUROPEAN CONFERENCE ON COMPUTER VISION

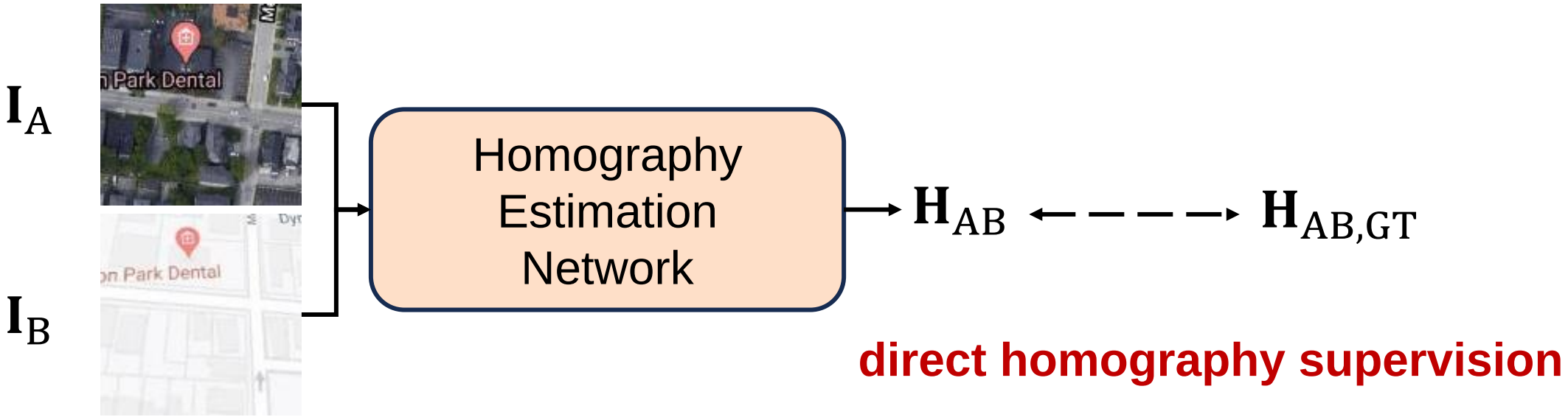
SCPNet: Unsupervised Cross-modal Homography Estimation via Intra-modal Self-supervised Learning

Runmin Zhang*, Jun Ma*, Si-Yuan Cao*†, Lun Luo,
Beinan Yu, Shu-Jie Chen, Junwei Li, and Hui-Liang Shen

https://github.com/RM-Zhang/SCPNet

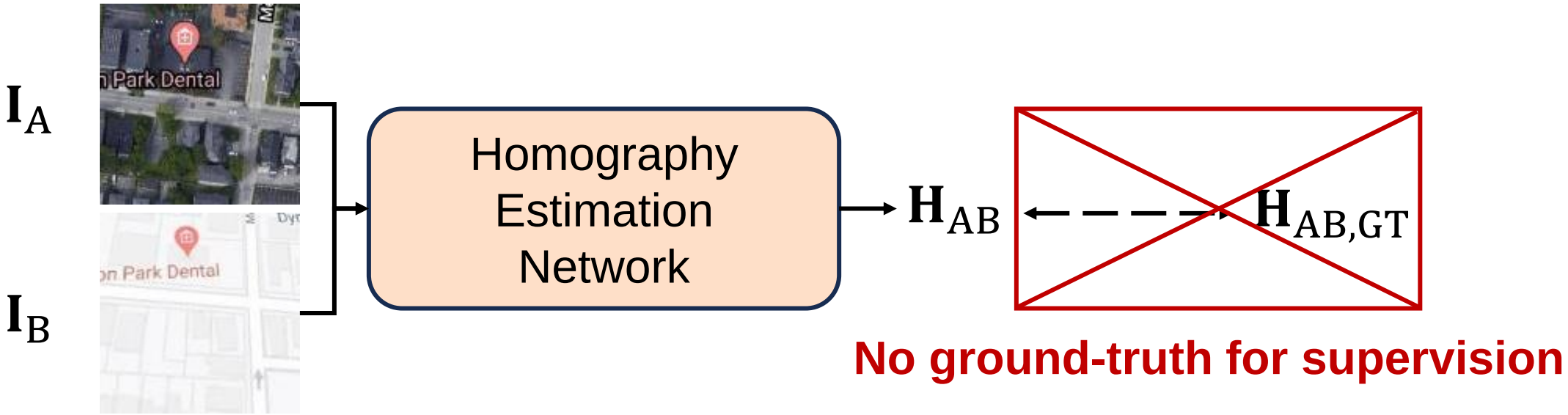
Background

Previous supervised approaches



Background

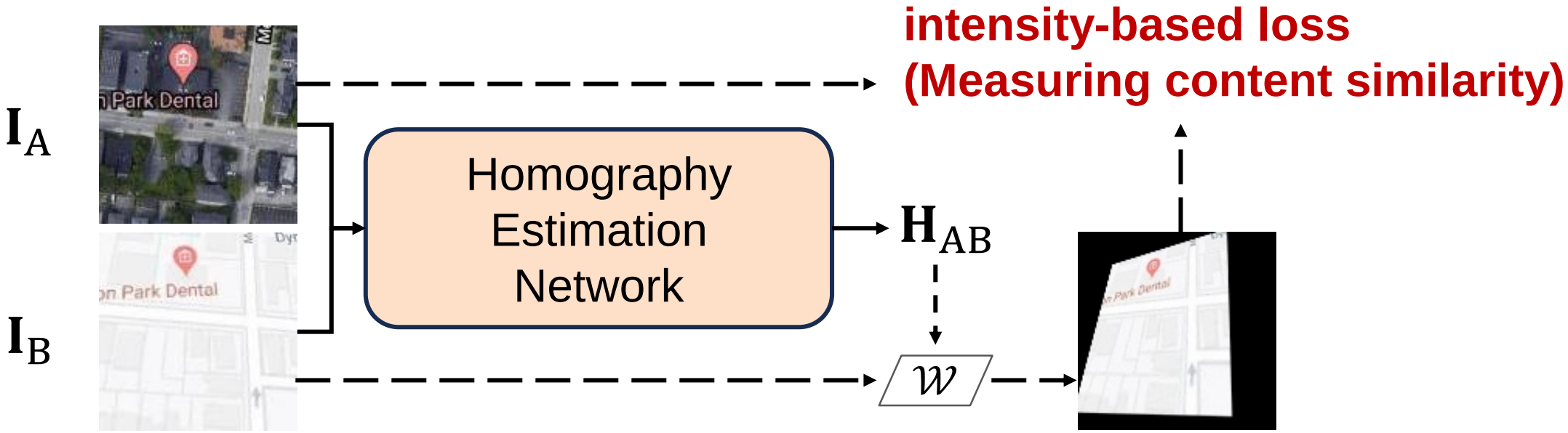
Unsupervised situation



No ground-truth for supervision

Background

Recent unsupervised approaches

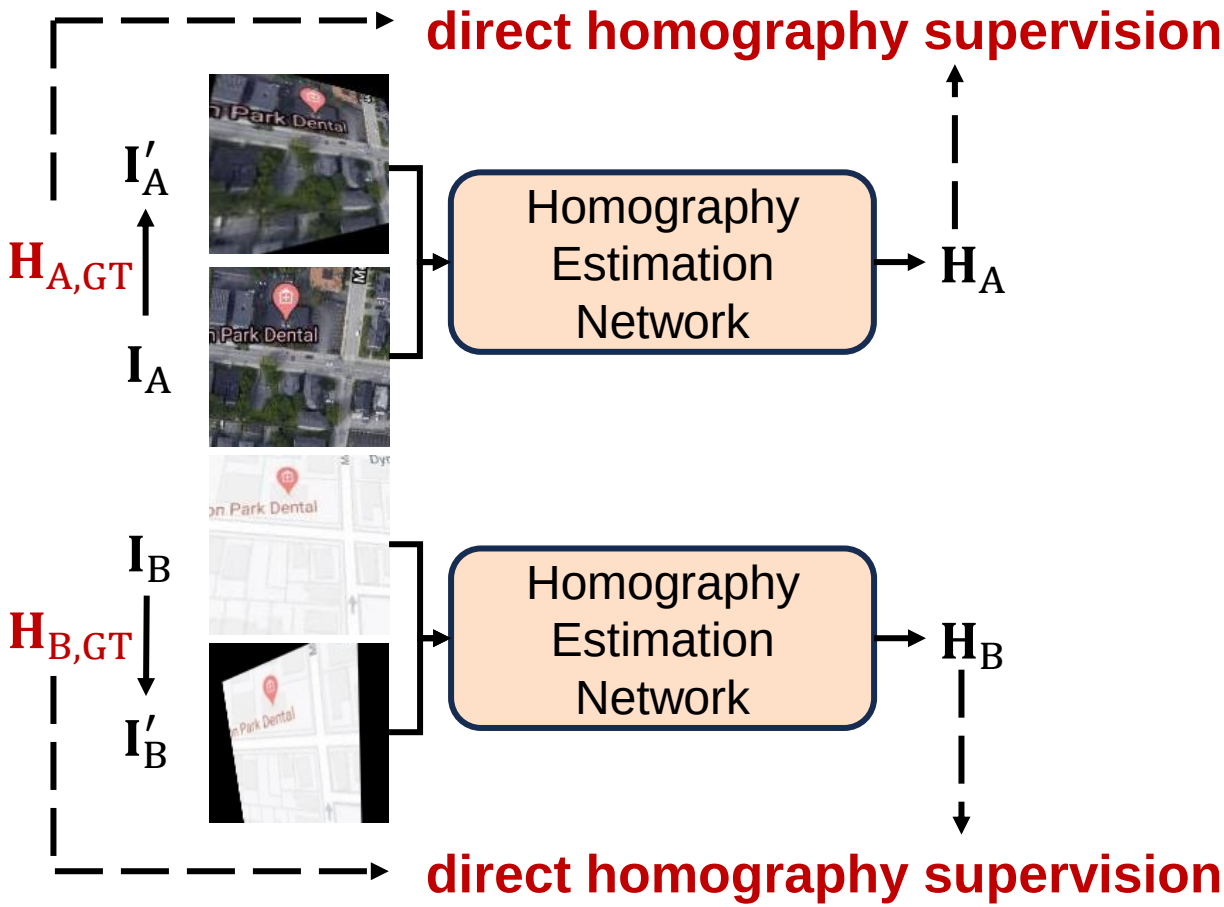


Fail under large homography deformation and modality variance.

Motivation

Our intra-modal self-supervised learning

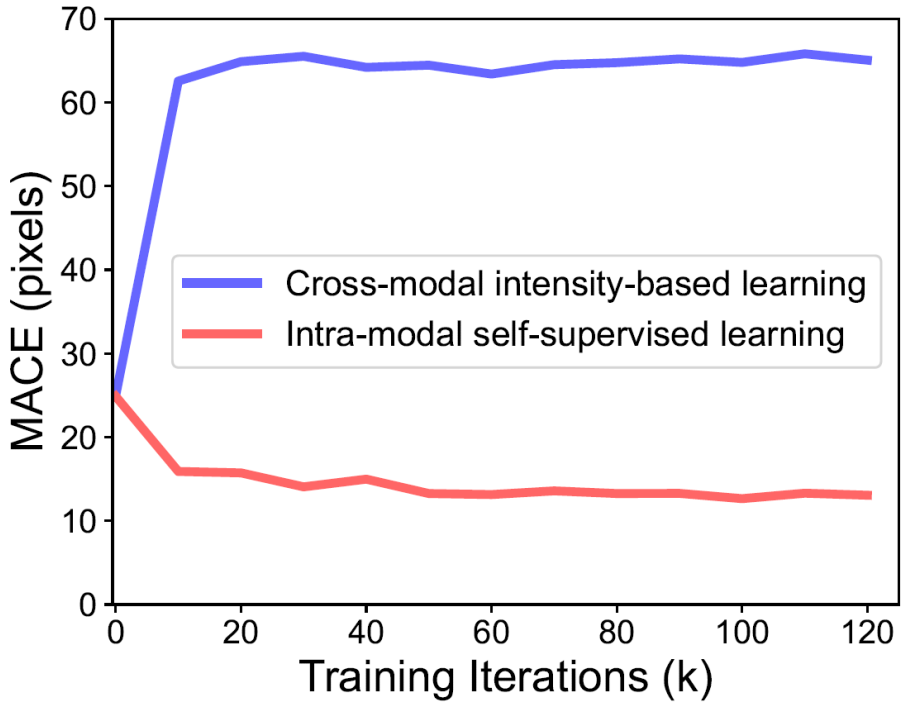
- Inspired by multitask learning, we introduce highly related tasks that provide direct supervision through intra-modal homography simulation.
- Training network to learn two-branch intra-modal homography.



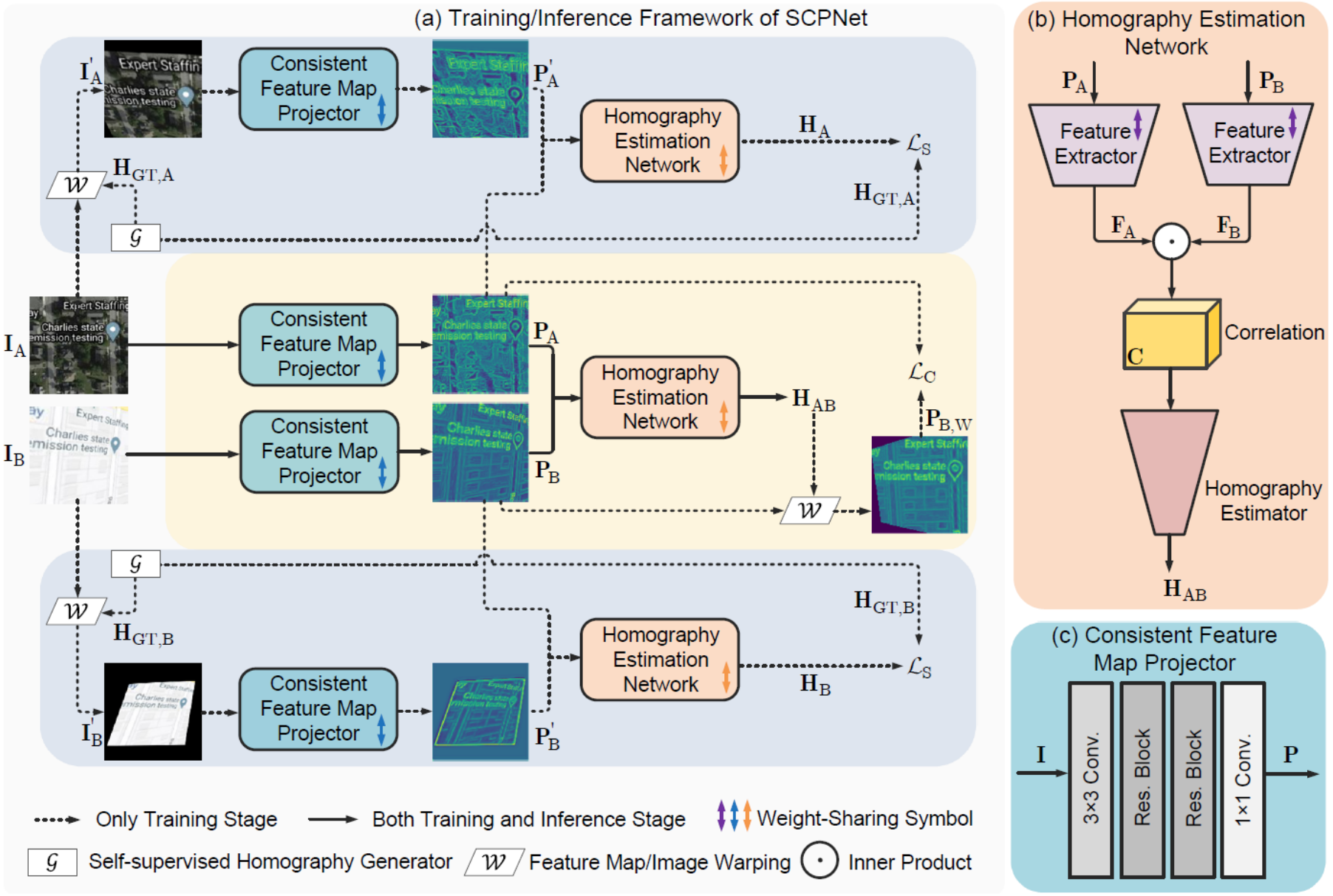
Motivation

Pilot experiments and finding

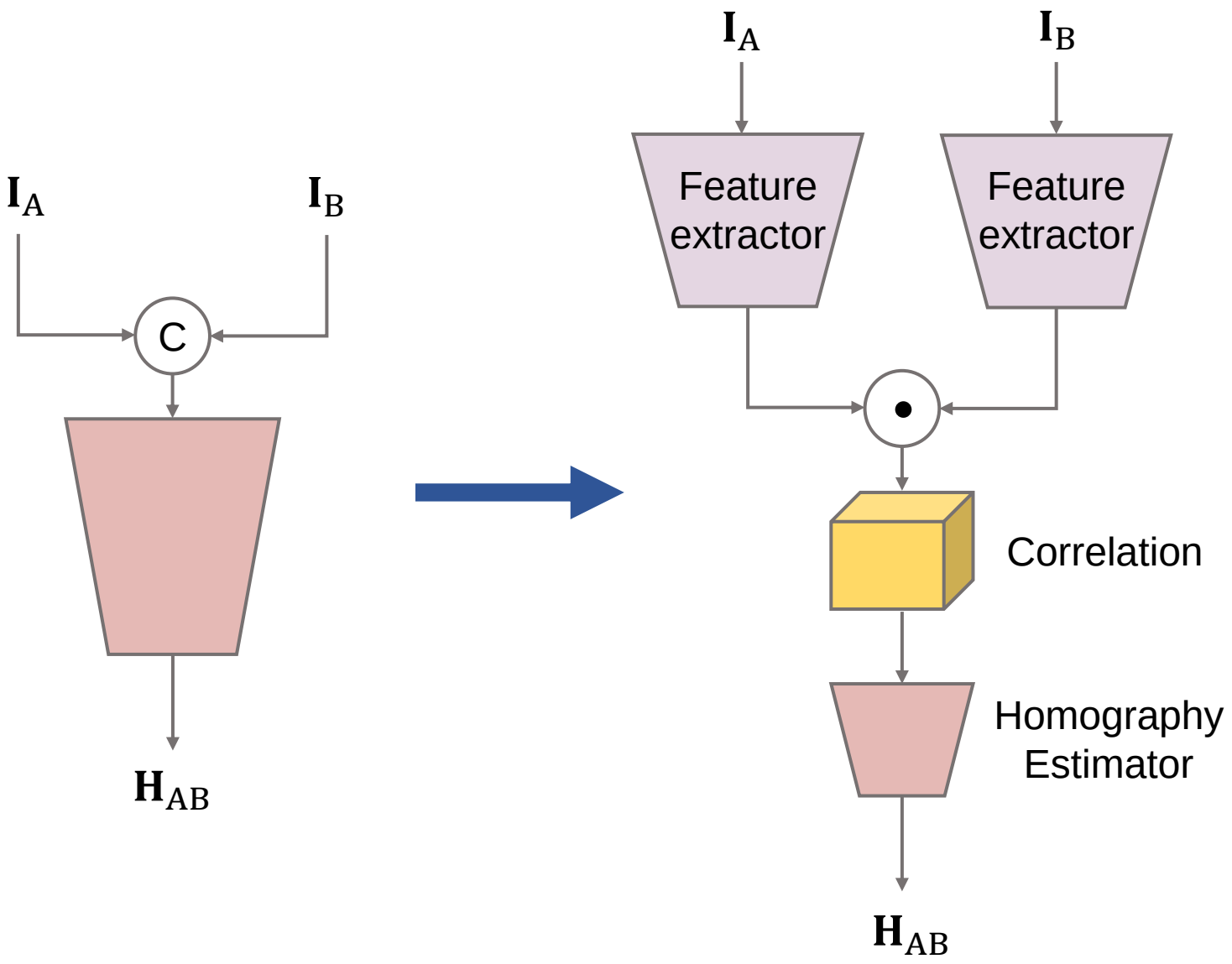
The cross-modal homography estimation can be indirectly facilitated by training the weight-shared network using the simulated transform within the two modalities.



SCPNet: Overall Framework

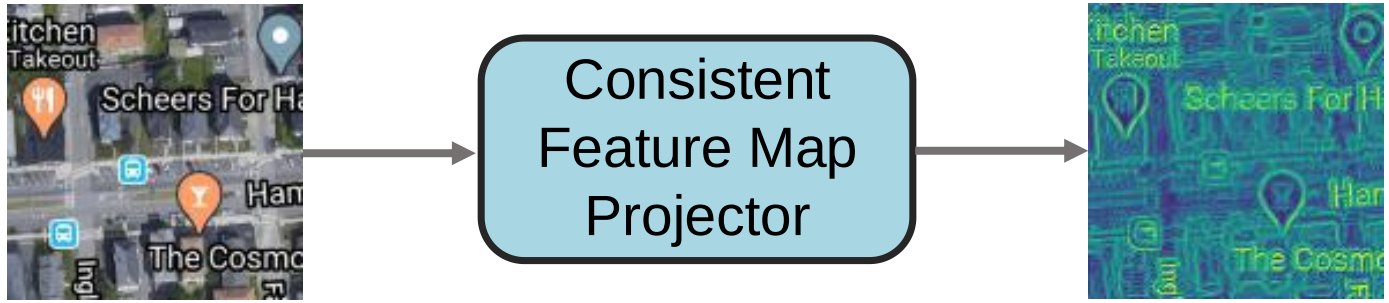


Correlation-based Homography Estimation Network

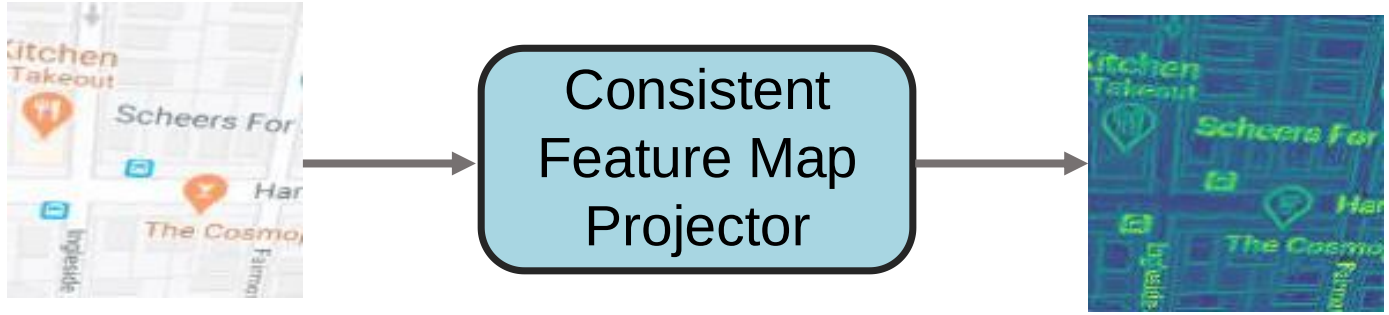


- More clear network constraint
- Stronger knowledge generalization ability

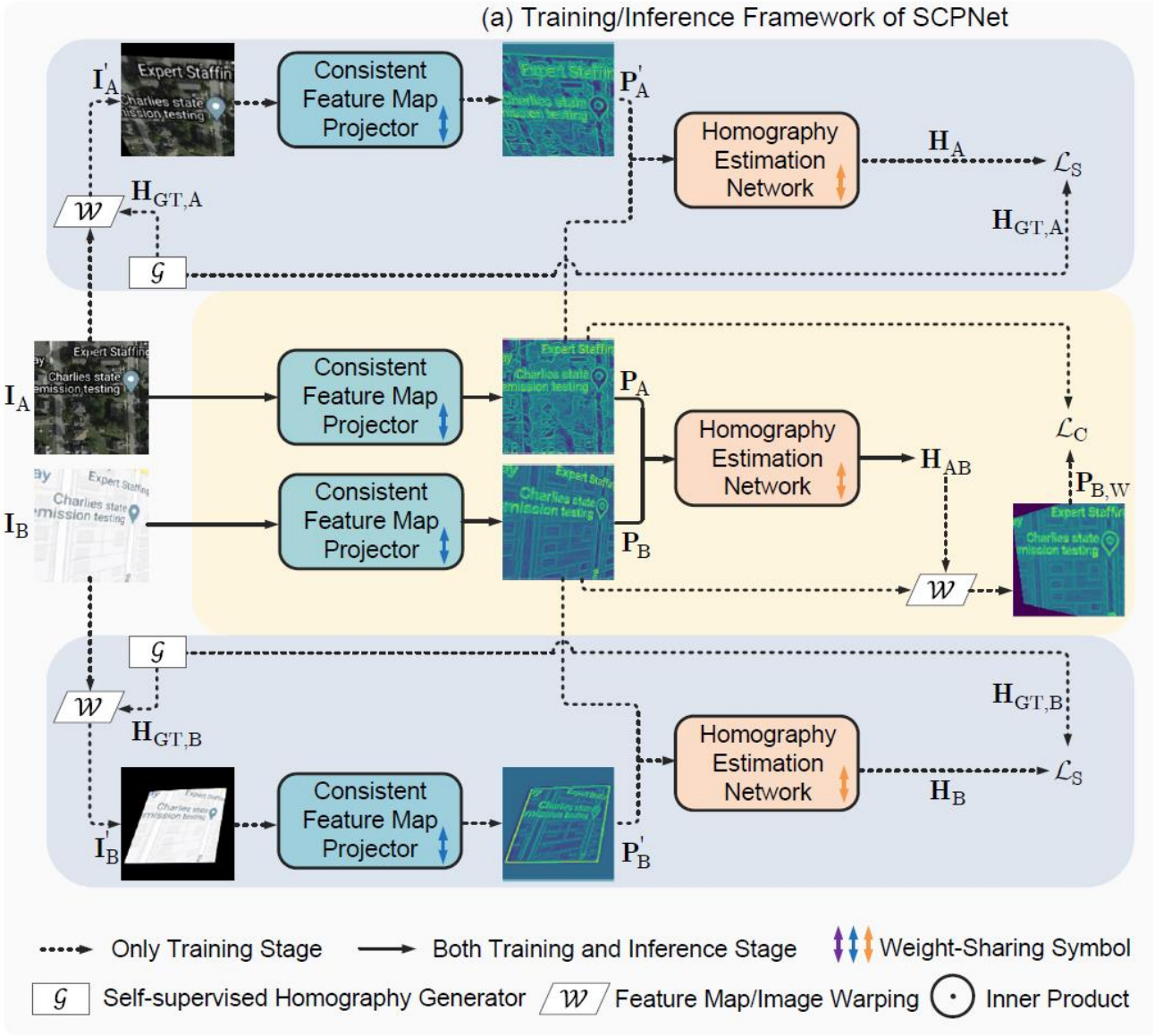
Consistent Feature Map Projector



- Consistent latent space projection
- Valid cross-modal supervision



Training/Inference Framework



Cross-modal intensity-based loss: supervise the content similarity of consistent feature maps.

$$\mathcal{L}_{\mathrm{C}} = \frac{\|\mathbf{P}_{\mathrm{A}} - \mathbf{P}_{\mathrm{B,W}}\|_1}{\|\mathbf{P}_{\mathrm{A}} - \mathbf{P}_{\mathrm{B}}\|_1}$$

$$\arg\min_{\xi,\zeta} \mathcal{L}_{\mathrm{C}}\left(\delta_\zeta(\mathbf{I}_{\mathrm{A}}), \mathcal{W}(\delta_\zeta(\mathbf{I}_{\mathrm{B}}), \psi_\xi(\delta_\zeta(\mathbf{I}_{\mathrm{A}}), \delta_\zeta(\mathbf{I}_{\mathrm{B}})))\right)$$
$$+\lambda\,\mathcal{L}_{\mathrm{S}}\left(\psi_\xi(\delta_\zeta(\mathbf{I}_{\mathrm{A}}), \delta_\zeta(\mathbf{I}'_{\mathrm{A}})), \mathbf{H}_{\mathrm{GT,A}}\right)$$
$$+\lambda\,\mathcal{L}_{\mathrm{S}}\left(\psi_\xi(\delta_\zeta(\mathbf{I}_{\mathrm{B}}), \delta_\zeta(\mathbf{I}'_{\mathrm{B}})), \mathbf{H}_{\mathrm{GT,B}}\right).$$

Intra-modal self-supervised loss: supervise the homography matrix by the offsets of four corner points.

Experiments

Ablation on GoogleMap dataset

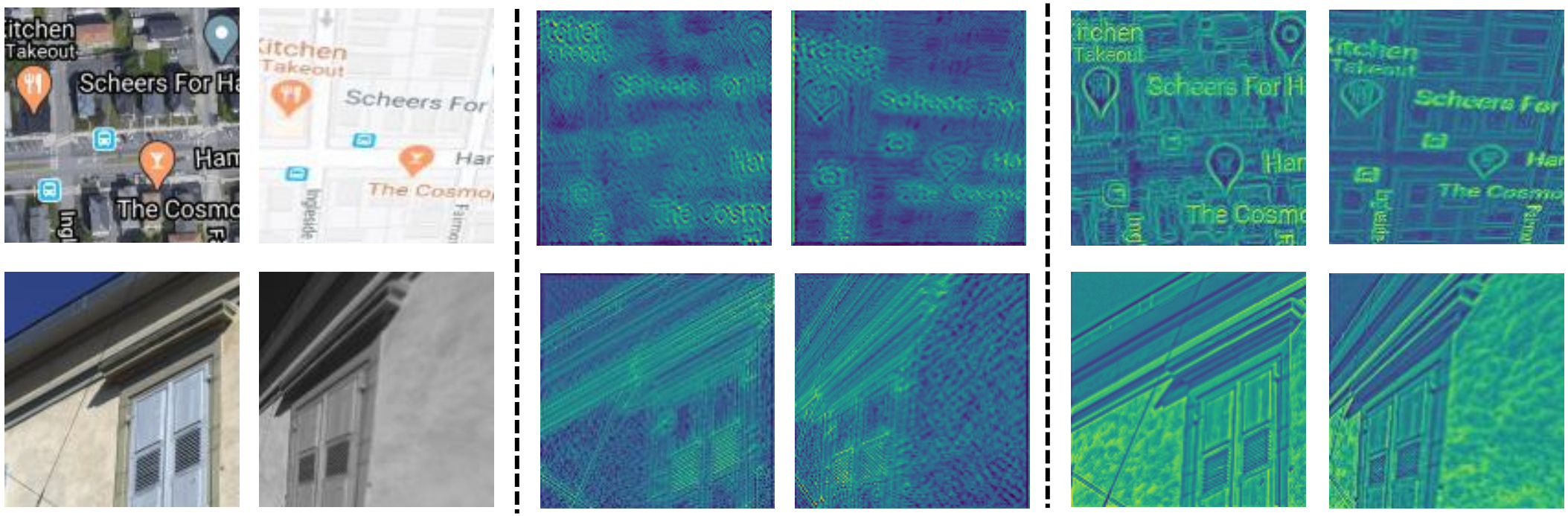
Table 1: Ablation study of SCPNet. NC denotes the training is not converged. Self denotes intra-modal self-supervised learning, projection denotes consistent feature map projection, and cross denotes cross-modal intensity-based learning.

Setting	Self	Correlation	Projection	Cross	MACE↓
1	✗	✗	✗	✓	NC
2	✗	✓	✗	✓	NC
3	✗	✗	✓	✓	24.64
4	✗	✓	✓	✓	24.80
5	✓	✗	✗	✗	13.06
6	✓	✓	✗	✗	9.68
7	✓	✗	✓	✗	10.01
8	✓	✓	✓	✗	7.70
9	✓	✓	✓	✓	4.35

Experiments

Feature map visualization

The correlation visibly facilitates the consistent feature map generation by the direct constraint of inner product.



Experiments

Quantitative results on cross-modal datasets

Table 2: Quantitative results of our SCPNet and other approaches on cross-modal datasets. NC denotes the training is not converged. **Bold** indicates the best result among unsupervised methods.

Dataset		GoogleMap				Flash/no-flash			
Offset		Easy	Moderate	Hard	Mean	Easy	Moderate	Hard	Mean
Handcrafted	SIFT [32]	19.17	23.87	29.04	24.53	14.61	18.69	23.85	19.53
	ORB [36]	19.11	23.9	29.02	24.52	16.91	22.44	27.01	22.63
	DASC [26]	14.29	20.73	28.12	21.76	11.64	19.50	28.11	20.59
	RIFT [29]	10.43	15.46	21.93	16.55	11.22	13.95	21.66	16.21
Unsupervised	UDHN [35]	18.63	21.55	26.89	22.84	16.27	21.27	24.85	21.20
	CA-UDHN [44]	19.31	23.92	29.10	24.61	16.01	21.54	25.14	21.32
	biHomE [27]	NC	NC	NC	NC	8.24	12.56	14.04	11.86
	BasesHomo [41]	19.43	23.97	28.66	24.49	19.45	24.73	29.66	25.12
	UMF-CMGR [14]	19.22	24.01	29.02	24.60	17.99	22.43	28.40	23.49
	SCPNet (Ours)	**3.60**	**4.44**	**4.85**	**4.35**	**1.80**	**2.33**	**3.59**	**2.67**
Supervised	DHN [12]	7.06	6.82	7.00	6.93	5.28	6.13	7.51	6.42
	MHN [28]	4.75	5.00	5.34	5.06	3.18	6.55	5.81	5.24
	LocalTrans [37]	0.91	1.43	6.30	3.22	0.49	0.67	4.05	1.96
	IHN [6]	0.70	0.96	1.06	0.92	0.76	0.65	0.94	0.80
	RHWF [8]	0.62	0.68	0.93	0.76	0.79	0.68	0.53	0.65

Experiments

Quantitative results on cross-spectral datasets

Table 3: Quantitative results of our SCPNet and other approaches on cross-spectral datasets. NC denotes the training is not converged. **Bold** indicates the best result among unsupervised methods.

Dataset		Harvard				RGB/NIR			
Offset		Easy	Moderate	Hard	Mean	Easy	Moderate	Hard	Mean
Handcrafted	SIFT [32]	17.27	21.49	26.70	22.30	15.54	23.90	28.81	24.40
	ORB [36]	18.61	23.06	28.29	23.82	17.75	22.84	27.01	23.00
	DASC [26]	11.85	18.29	25.03	19.05	13.50	17.91	25.73	19.78
	RIFT [29]	10.41	15.69	21.98	16.62	11.22	13.80	23.34	16.84
Unsupervised	UDHN [35]	18.03	22.20	26.55	22.69	18.54	23.27	27.16	23.43
	CA-UDHN [44]	18.77	23.64	28.55	24.14	18.31	23.88	28.66	24.12
	biHomE [27]	NC	NC	NC	NC	18.61	23.05	28.18	23.77
	BasesHomo [41]	19.77	24.20	28.46	24.57	19.23	23.44	28.89	24.41
	UMF-CMGR [14]	16.61	21.08	26.25	21.81	17.04	22.16	26.53	22.38
	SCPNet (Ours)	**2.34**	**3.70**	**5.48**	**4.00**	**1.65**	**4.69**	**7.13**	**4.78**
Supervised	DHN [12]	5.30	6.34	8.09	6.72	9.55	10.08	14.87	11.88
	MHN [28]	4.37	5.09	6.27	5.35	6.88	7.10	8.26	7.51
	LocalTrans [37]	0.27	0.43	4.58	2.04	0.53	0.77	5.15	2.47
	IHN [6]	1.40	1.72	2.03	1.75	1.25	2.14	2.21	1.90
	RHWF [8]	1.37	1.76	1.85	1.68	0.68	1.44	1.08	1.07

Experiments

Qualitative results

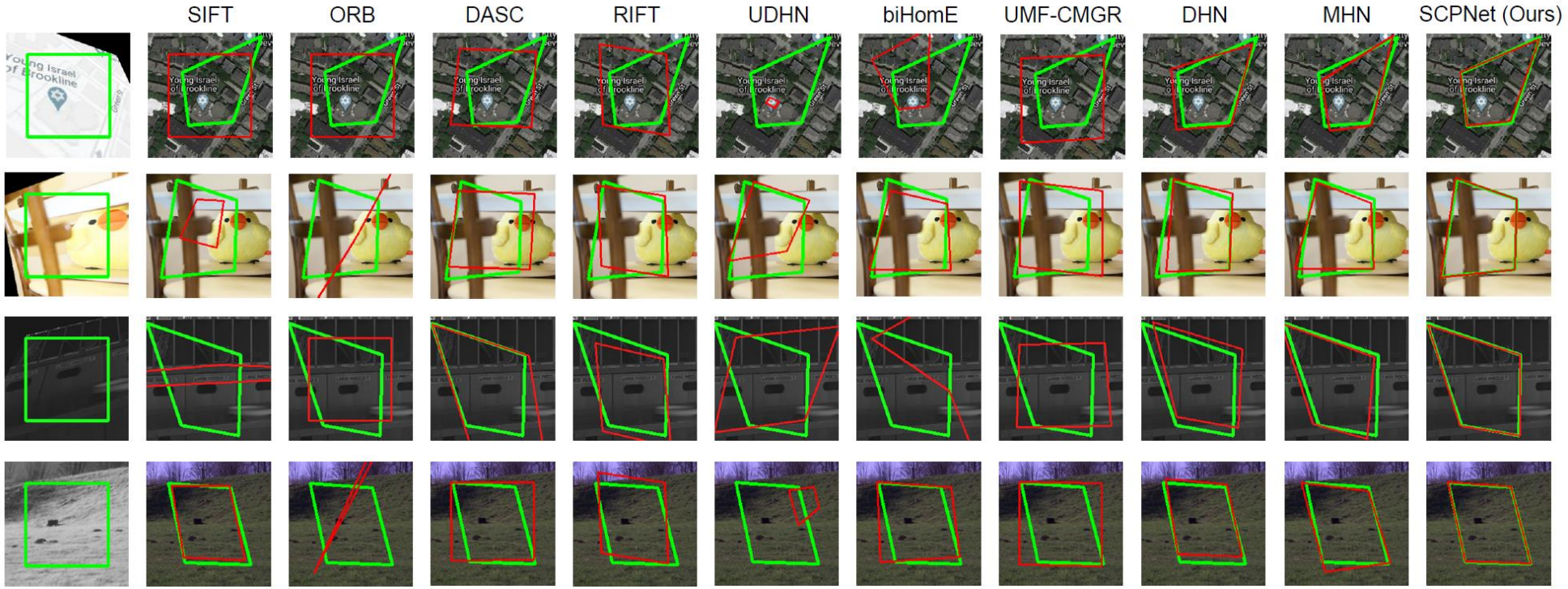


Fig. 5: Qualitative homography estimation results on GoogleMap, Flash/no-flash, Harvard, and RGB/NIR datasets respectively. Green polygons denote the ground-truth homography deformation from $\mathbf{I}_B$ (source, the deformed image) to $\mathbf{I}_A$ (target). Red polygons denote the estimated homography deformation using different algorithms on the target images.

Conclusions

- We propose **SCPNet**, a novel unsupervised cross-modal homography estimation framework, which combines three key components.
- We devise the intra-modal **S**elf-supervised learning to support the unsupervised learning framework, which mines the two-branch self-supervised information via applying simulated homography within the two modalities.
- We combine the **C**orrelation and consistent feature map **P**rojection to form a powerful unsupervised learning network architecture of SCPNet.
- SCPNet **ranks top** in the unsupervised homography estimation on cross-modal/spectral and manually-made inconsistent data under large offsets.

浙江大學
ZHEJIANG UNIVERSITY

EUROPEAN CONFERENCE ON COMPUTER VISION

Thanks for watching!