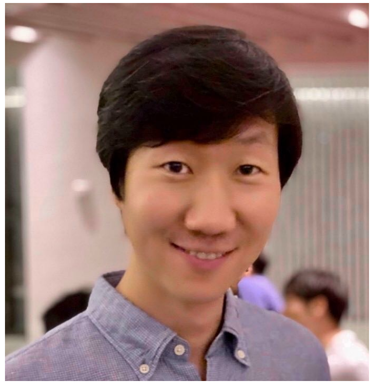
POSTECH
POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY
ComputerVision Lab.

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

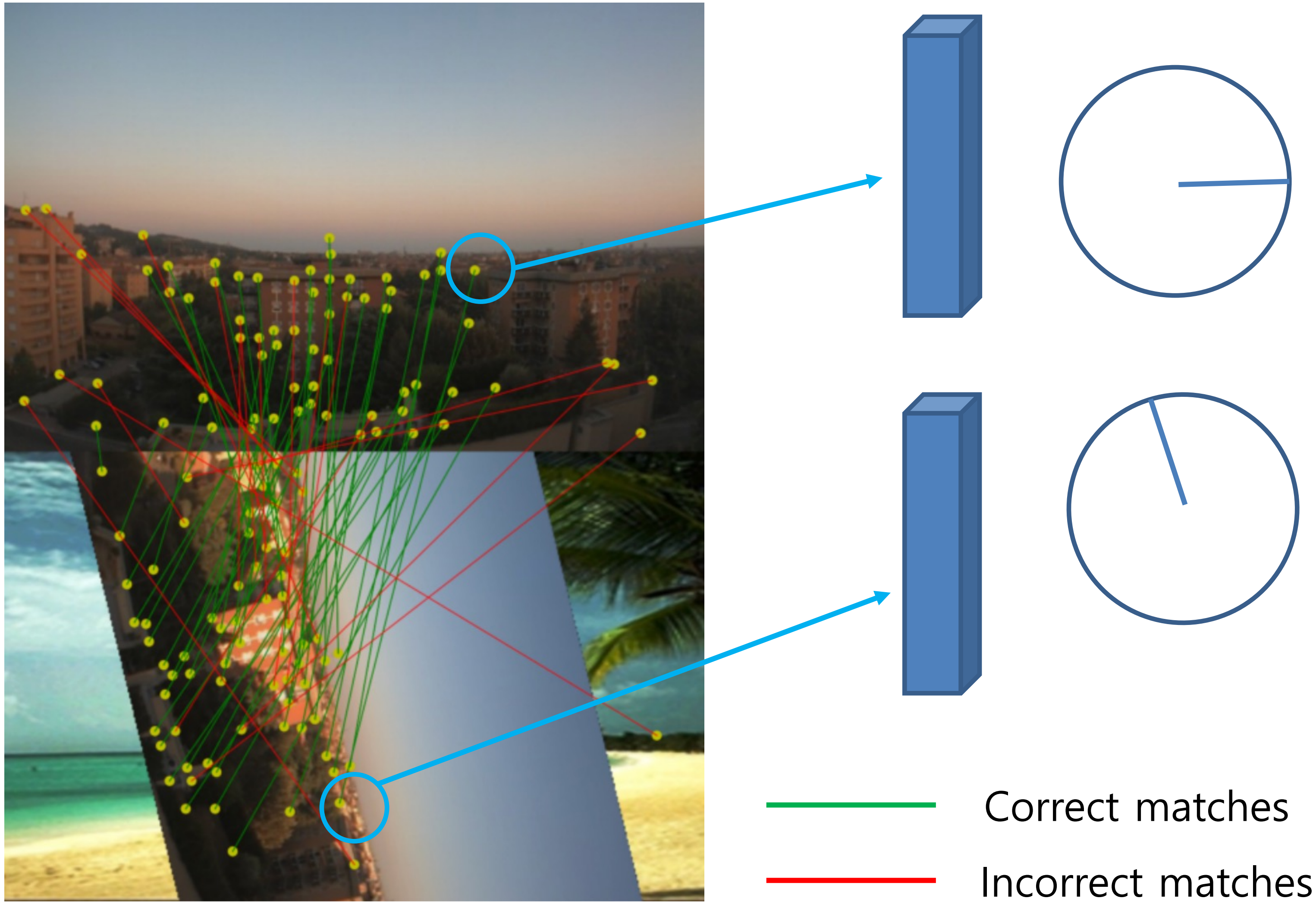
Poster Session THU-PM-112

Learning Rotation-Equivariant Features for Visual Correspondence



Jongmin Lee Byungjin Kim Seungwook Kim Minsu Cho

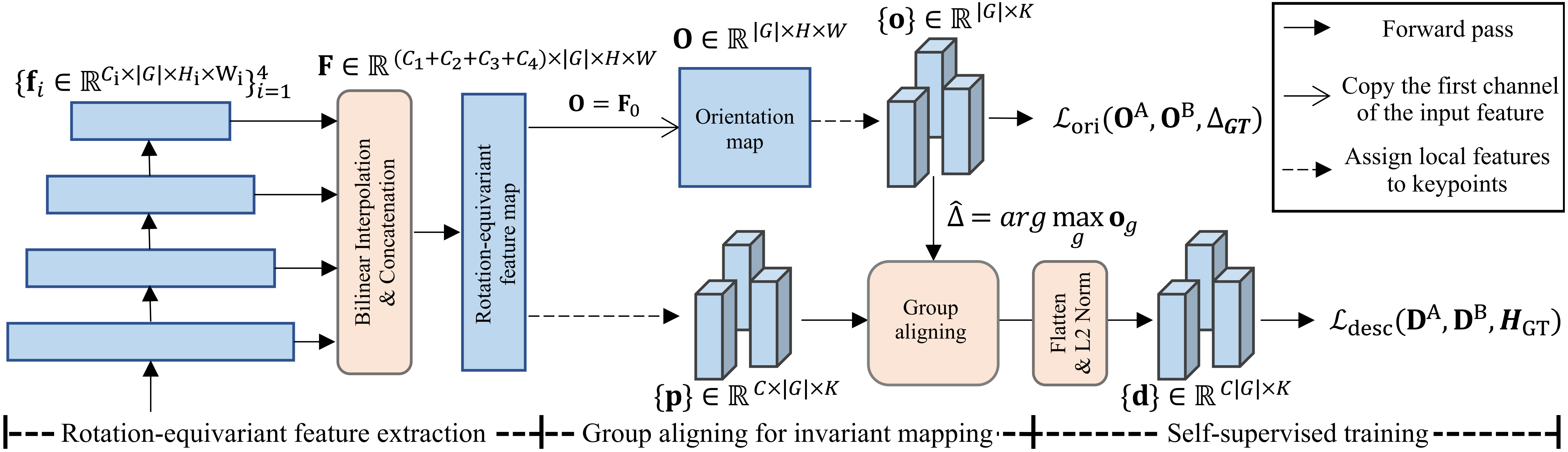
Local Features for Visual Correspondence



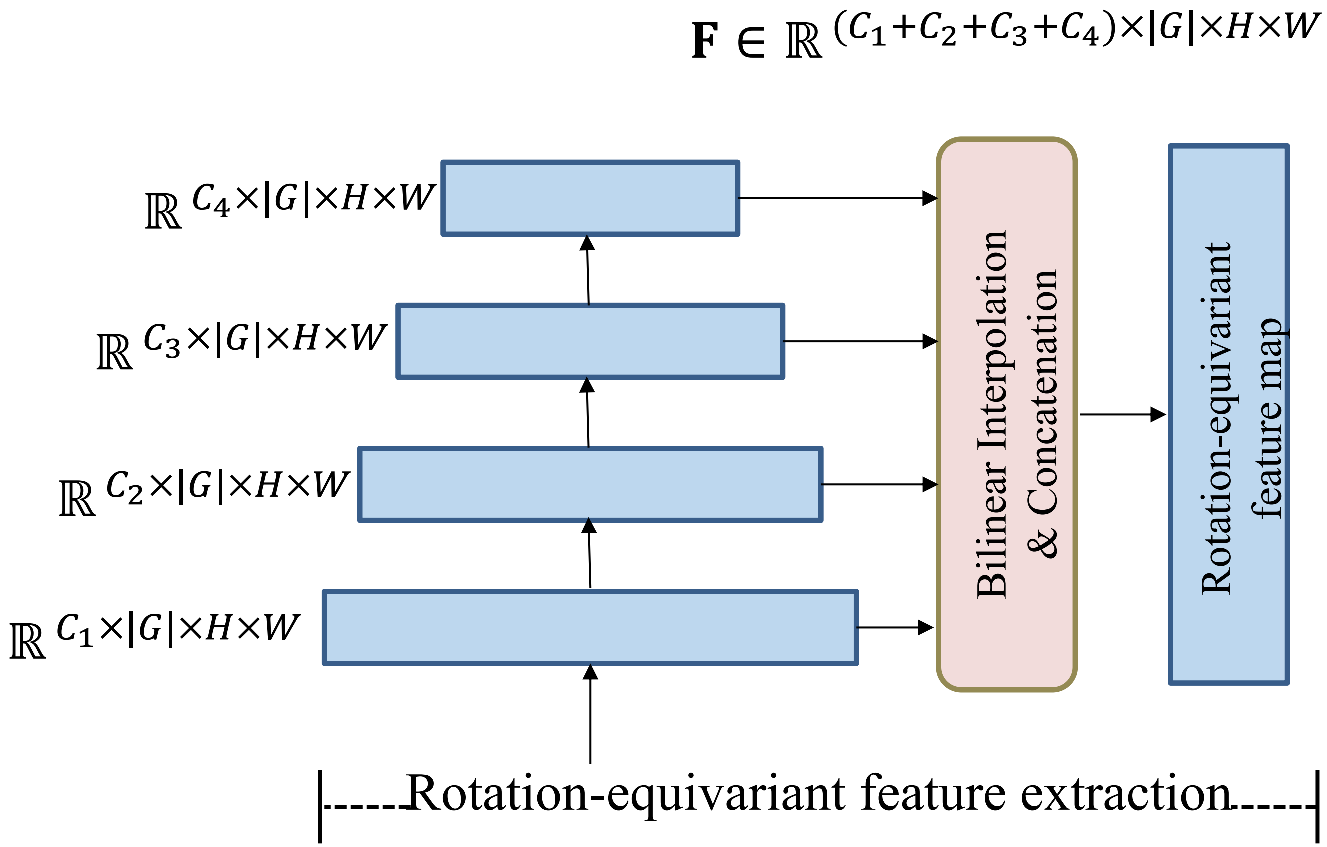
Local Features
(Descriptors, orientations …)

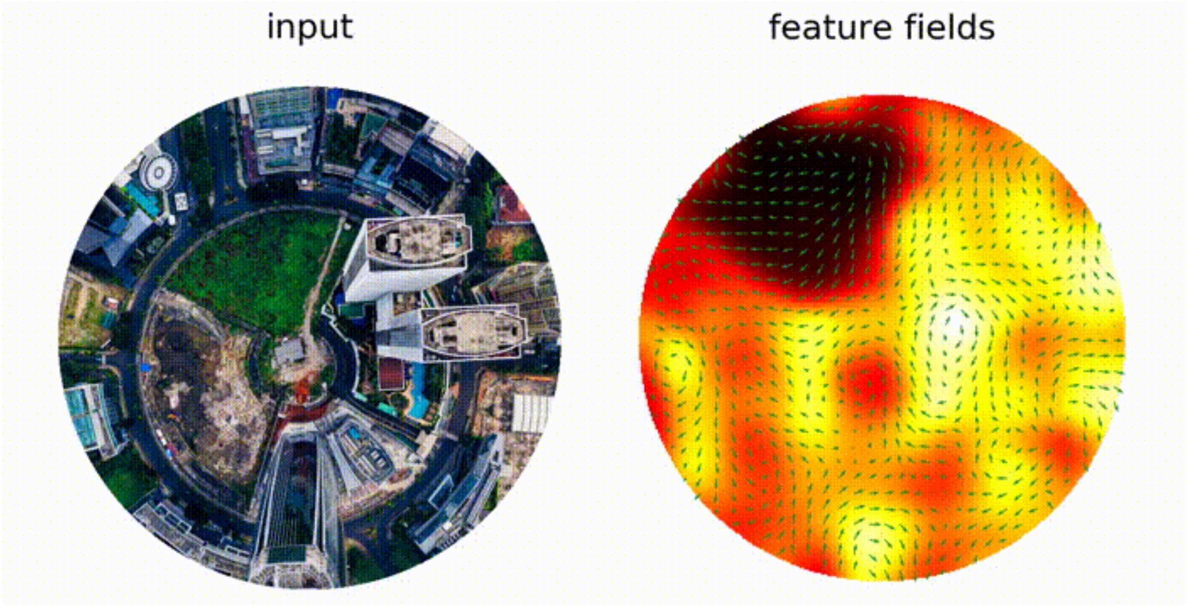
Discriminative and Invariant
to imaging variations

Rotation-Equivariant Local Features (RELF)



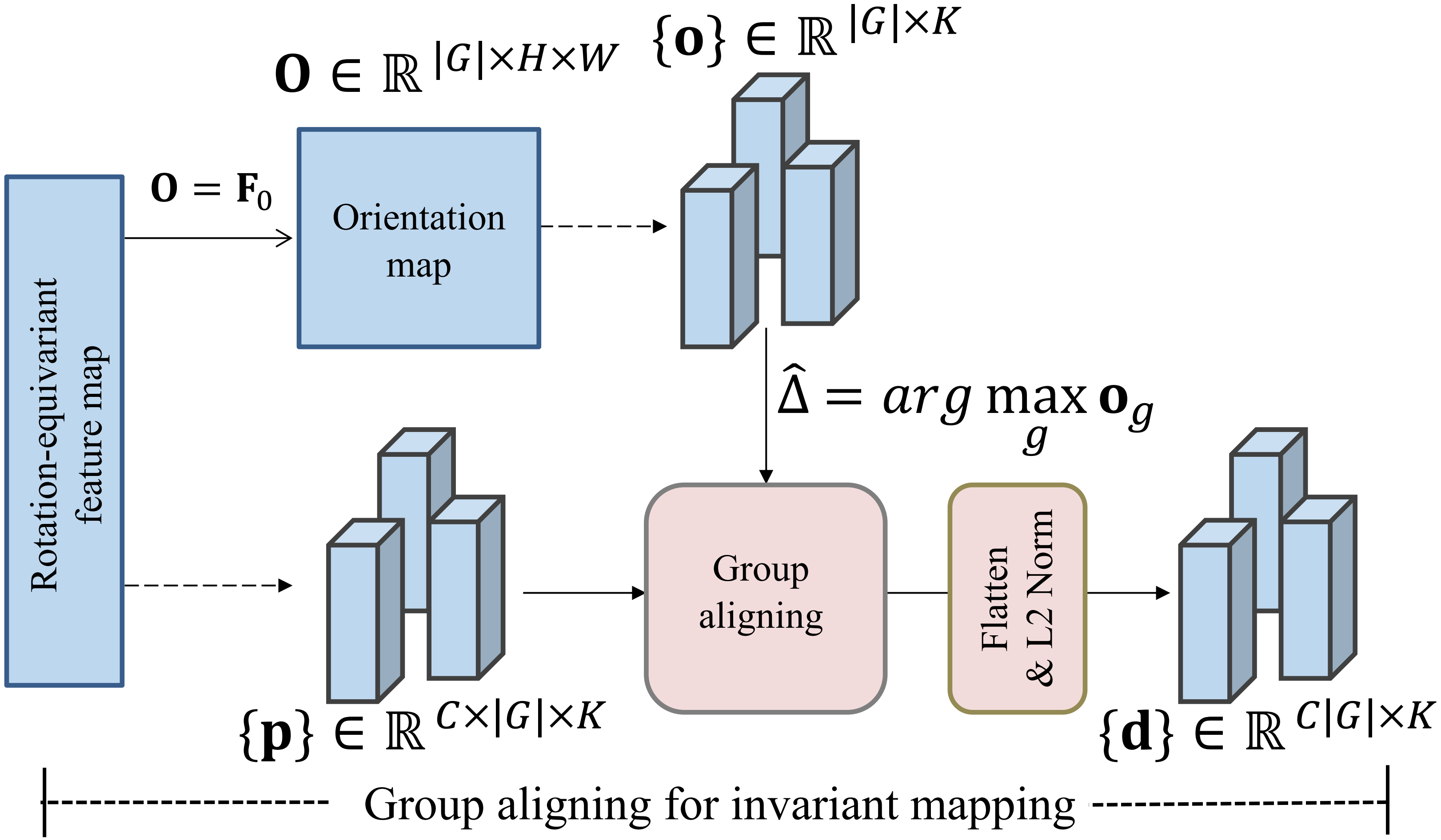
Rotation-Equivariant Local Features (RELF)



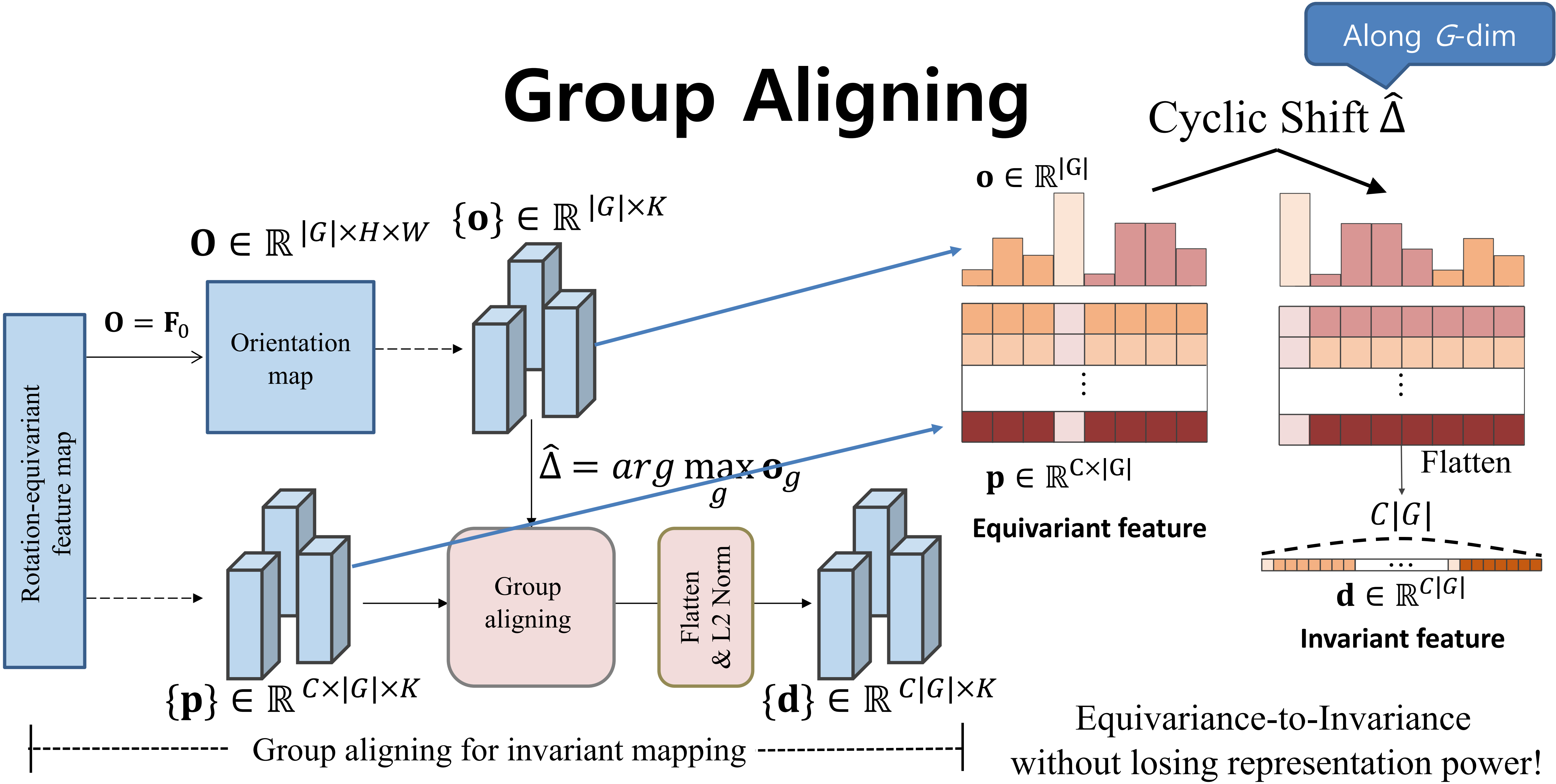


General E(2)-Equivariant Steerable CNNs (Weiler and Cesa, NeurIPS 2019)

Rotation-Equivariant Local Features (RELF)

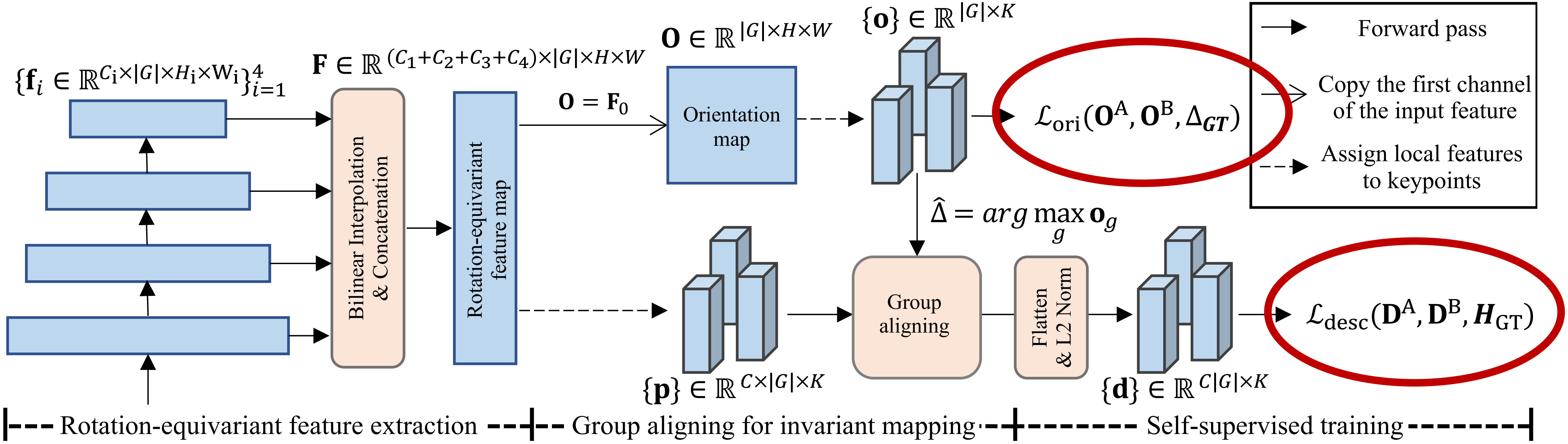


Group Aligning



Equivariance-to-Invariance
without losing representation power!

Rotation-Equivariant Local Features (RELF)



POSTECH
POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY
ComputerVision Lab.

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

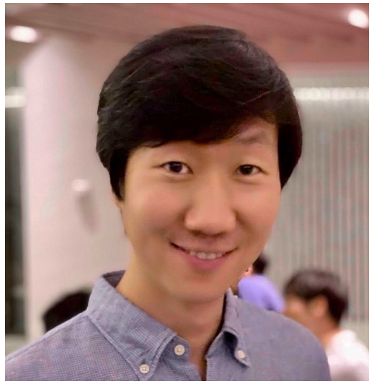
Poster Session THU-PM-112

Learning Rotation-Equivariant Features for Visual Correspondence


Jongmin Lee


Byungjin Kim


Seungwook Kim


Minsu Cho

Main Contributions

- ✓ Group-aligning for invariant mapping
  - ✓ Extracting rotation-invariant and discriminative local descriptors without collapsing the group dimension
- ✓ Self-supervised equivariant learning
  - ✓ Self-supervised losses to extract reliable orientations and descriptors robust to illumination/viewpoint changes
  - ✓ Using E(2)-CNN[1] for rotational equivariance with structural guarantees

[1] General E(2)-Equivariant Steerable CNNs (Weiler and Cesa, NeurIPS 2019)

Motivation

Exploit group-equivariant features to extract invariant descriptors

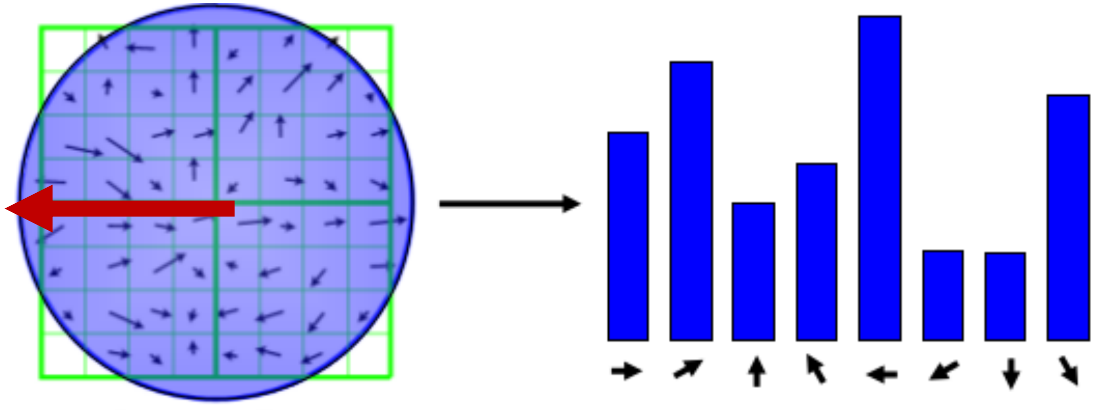
Invariance
&
Discriminativeness

Classical descriptor

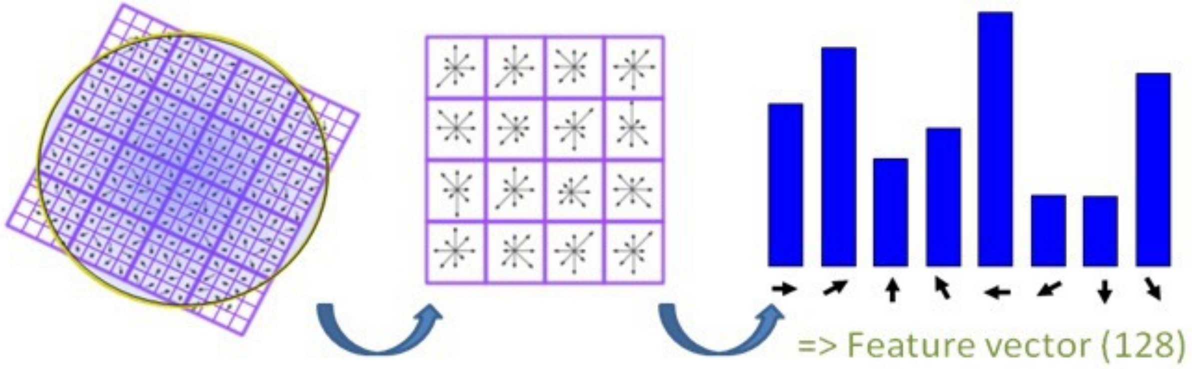
Standard neural net.

Related Work

- Handcrafted[1], [2], [3]



Orientation histogram by aggregating local image gradients



Orientation-normalized descriptor

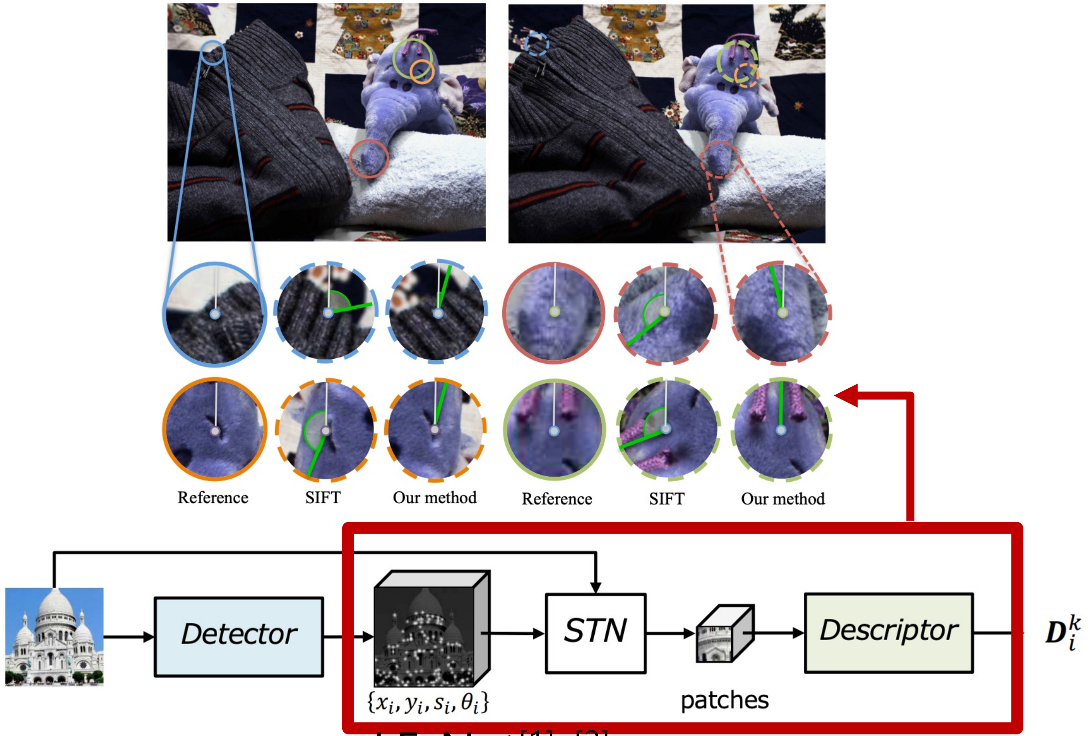
[1] Distinctive image features from scale-invariant keypoints. (Lowe, IJCV 2004)
[2] ORB: An efficient alternative to SIFT or SURF (Rublee et al., ICCV 2011)
[3] Rotationally Invariant Descriptors Using Intensity Order Pooling (Fan et al., TPAMI 2011)

Related Work

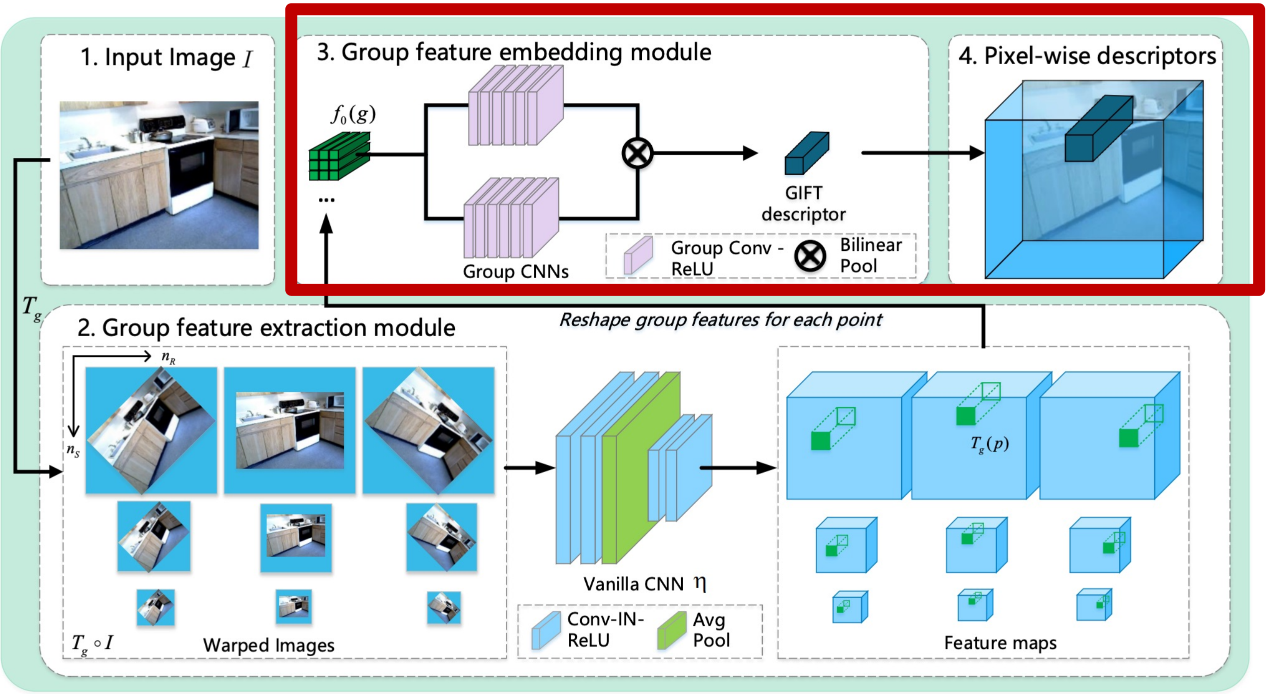
Not explicitly equivariant /invariant to rotation

Lose representation power by collapsing *G*-dim using group pooling

- Learning-based method [1], [2], [3]



LF-Net[1], [2]

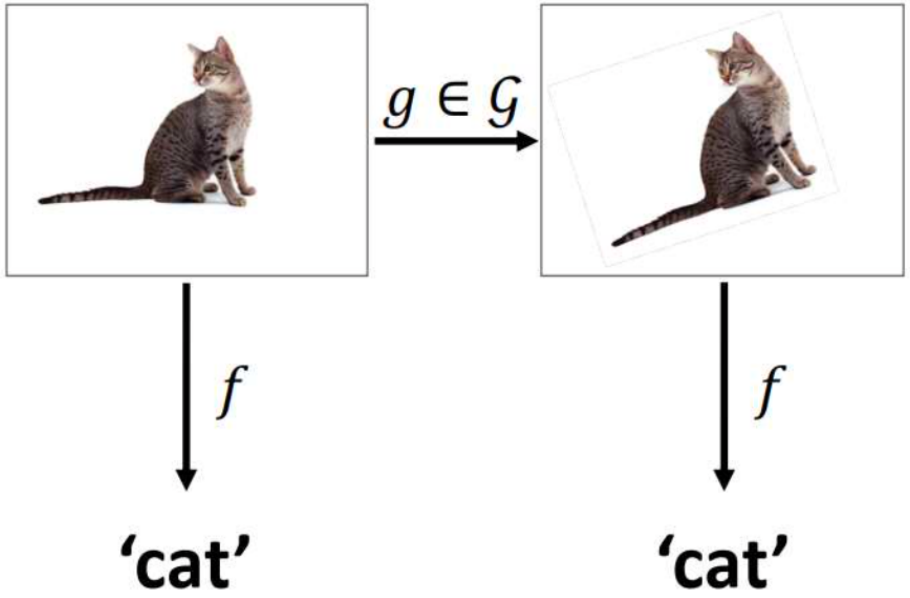


GIFT[3]

[1] Learning to Assign Orientations to Feature Points (Yi et al., CVPR 2016)
[2] LF-Net: Learning Local Features from Images (Ono et al., NIPS 2018)
[3] GIFT: Learning Transformation-Invariant Dense Visual Descriptors via Group CNNs (Liu et al., NIPS 2019)

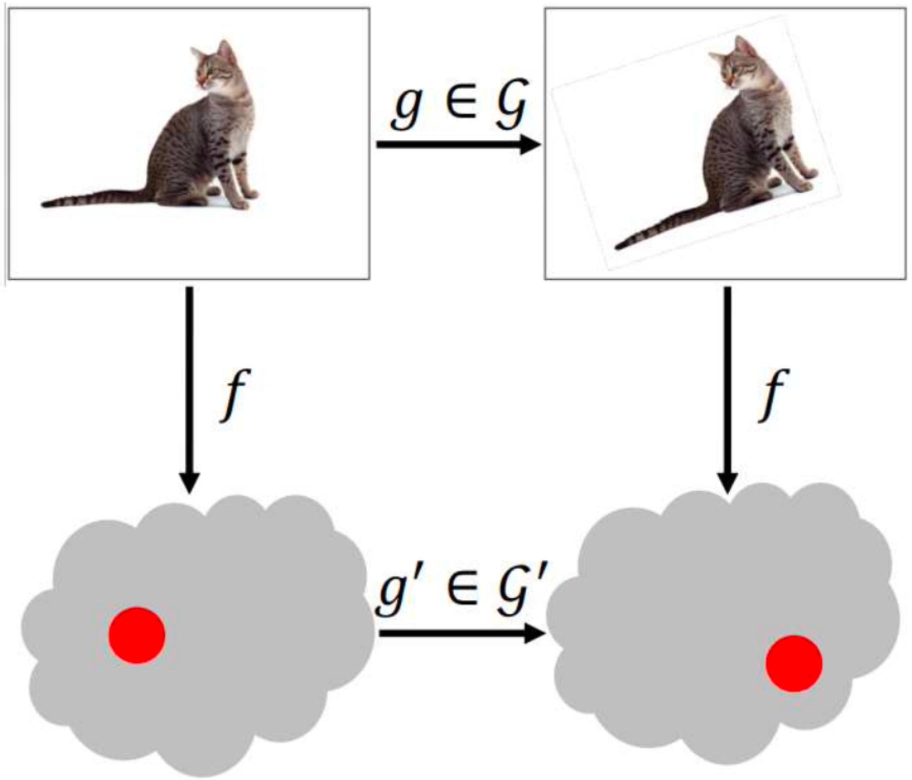
Invariance and Equivariance

Invariance



$$f(g(x)) = f(x)$$

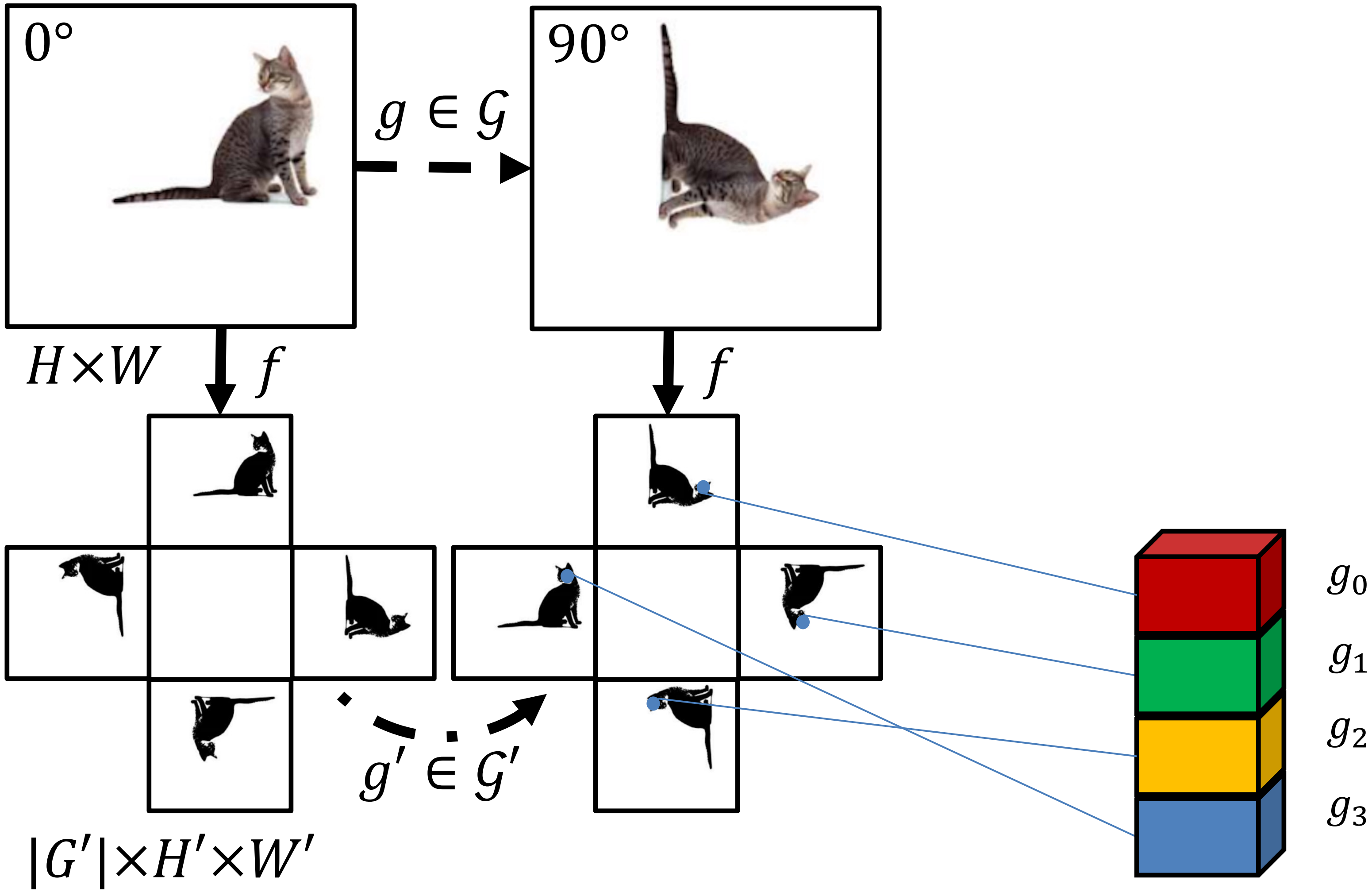
Equivariance



$$f(g(x)) = g'(f(x))$$

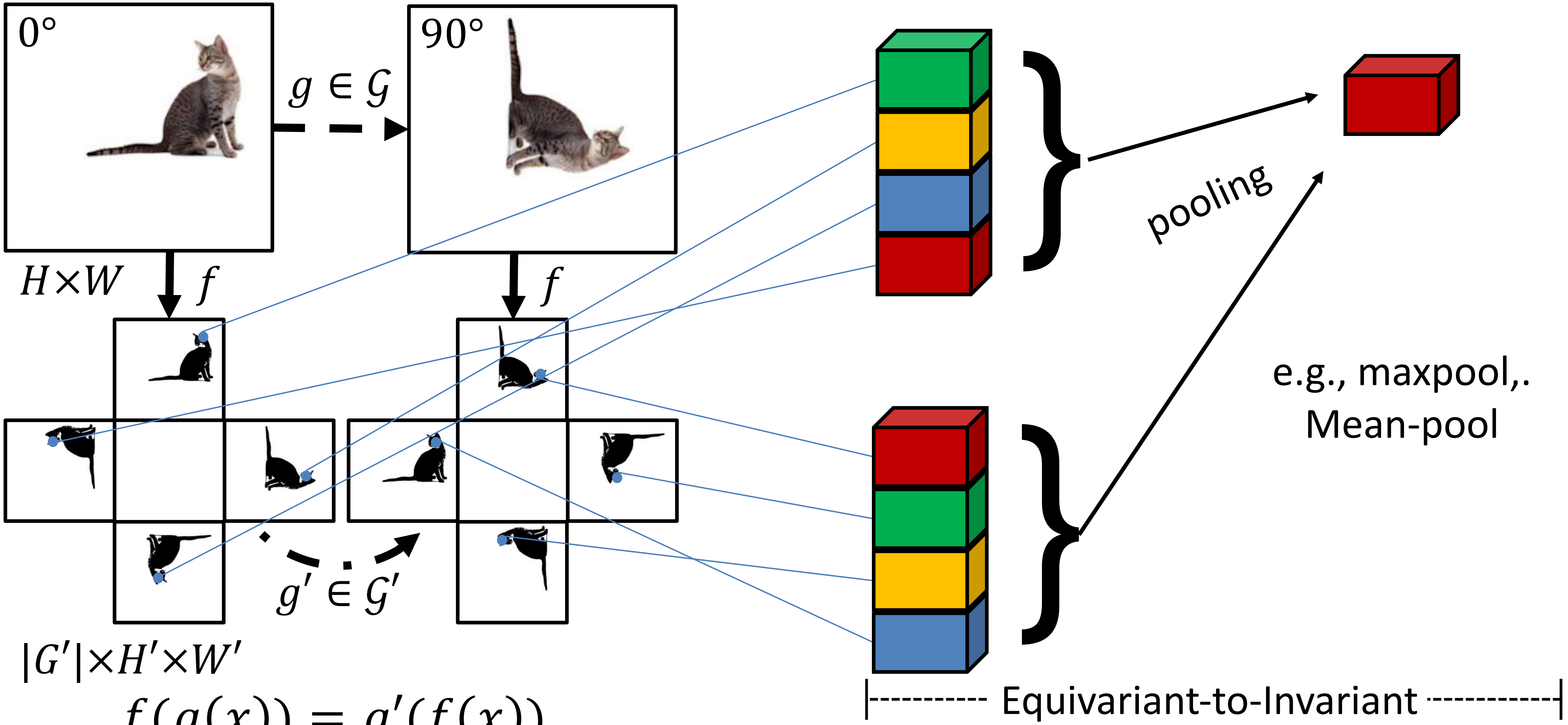
Figure from: Deep Learning – Bernhard Kainz

Group-Equivariance



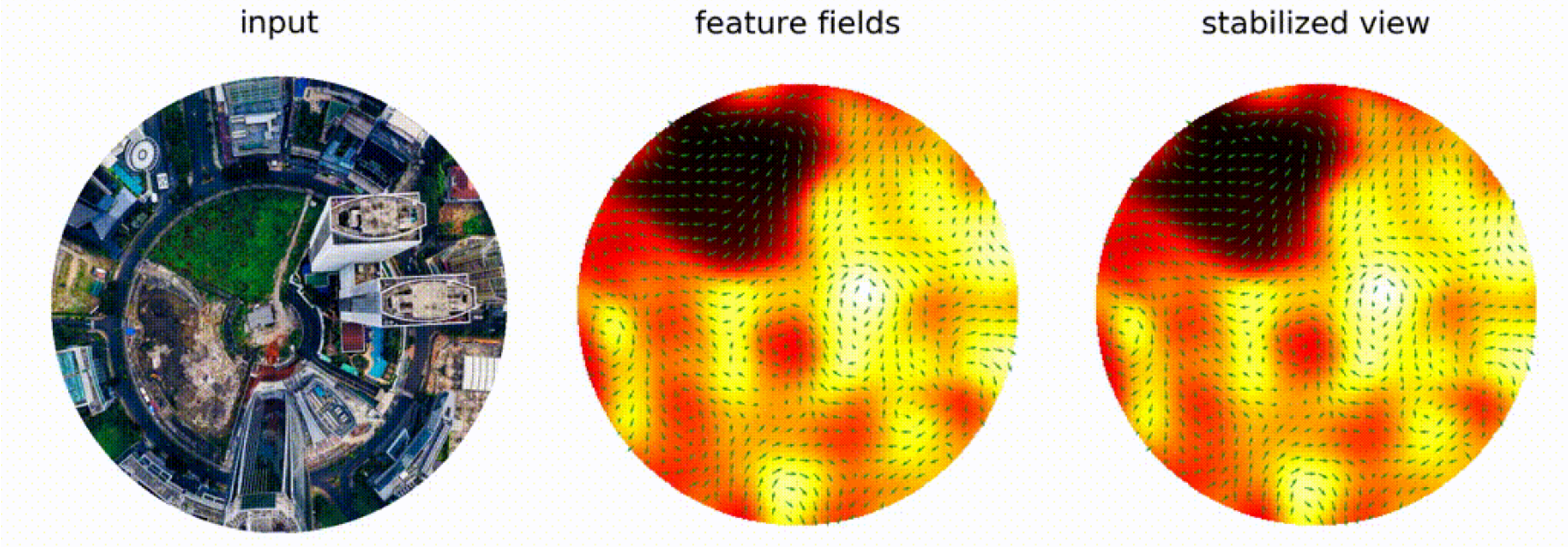
$$f(g(x)) = g'(f(x))$$

Group-Equivariance



$$f(g(x)) = g'(f(x))$$

Group-Equivariant Features



Rotation-Equivariant Feature

- Rotation-equivariant features contributes to generate **rotation-invariant descriptors** and **rotation-equivariant orientation**.

Animation: General E(2)-Equivariant Steerable CNNs (Weiler and Cesa, NeurIPS 2019)

Equivariant features, invariant descriptors

Local Features: keypoint, descriptor, scale, orientation, affine shape …

Equivariant features, invariant descriptors

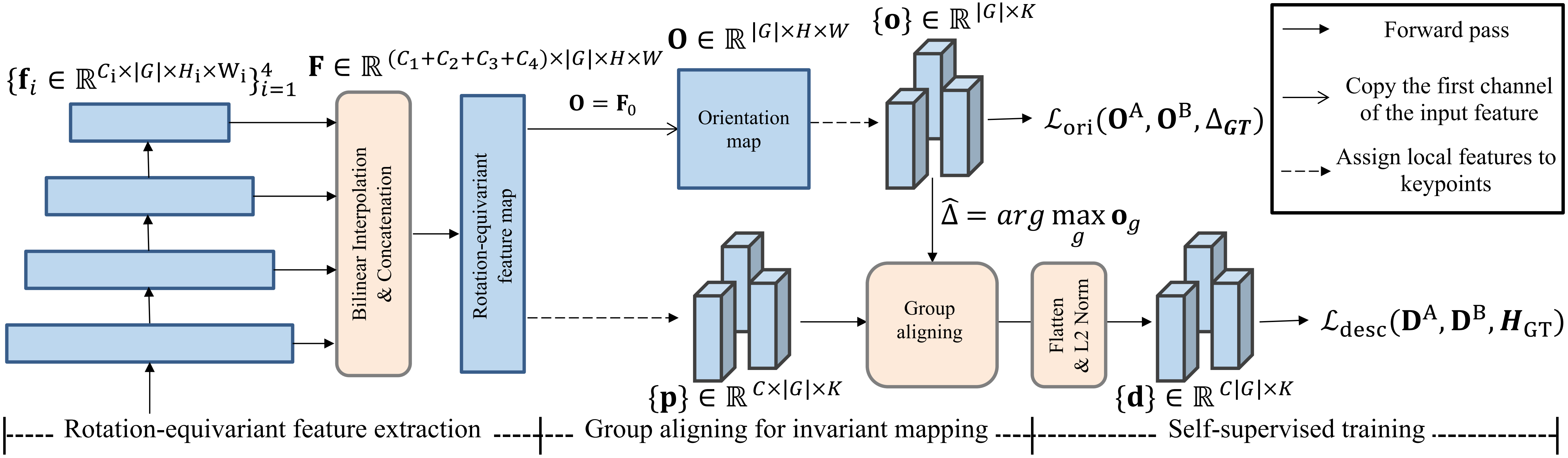
Local Features: keypoint, **descriptor**, scale, **orientation**, affine shape …

Should be invariant to rotation

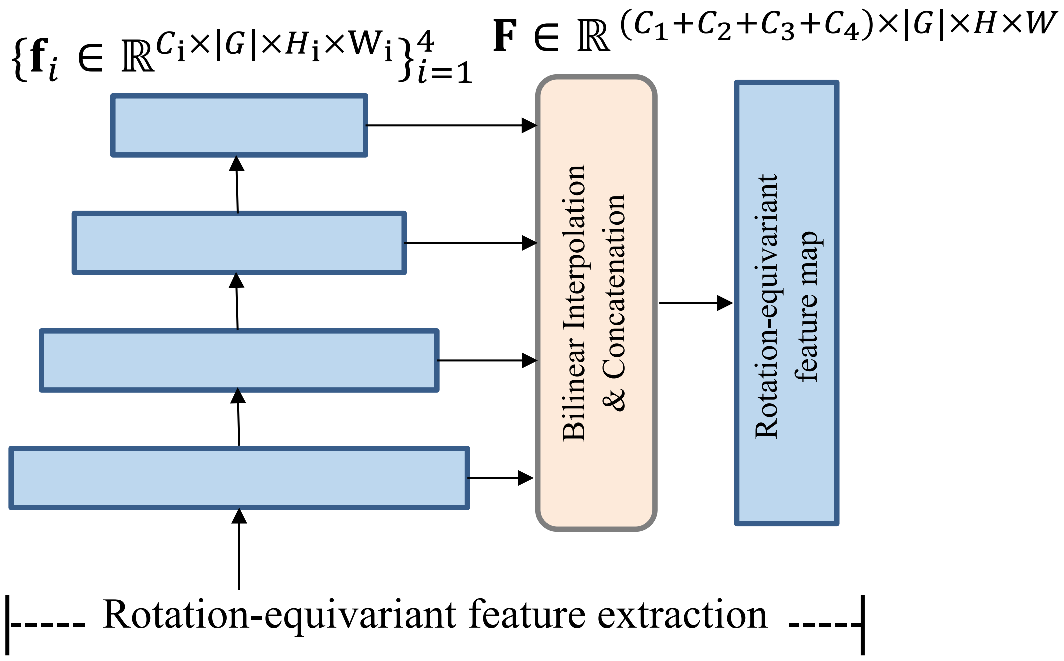
Should be equivariant to rotation

- Technical contributions
  1) Multi-scale feature extraction with rotation-equivariant backbone[1]
  2) Group aligning to obtain rotation-invariant local descriptors
  3) Self-supervised training by synthetic geometric transformation

Overall Architecture



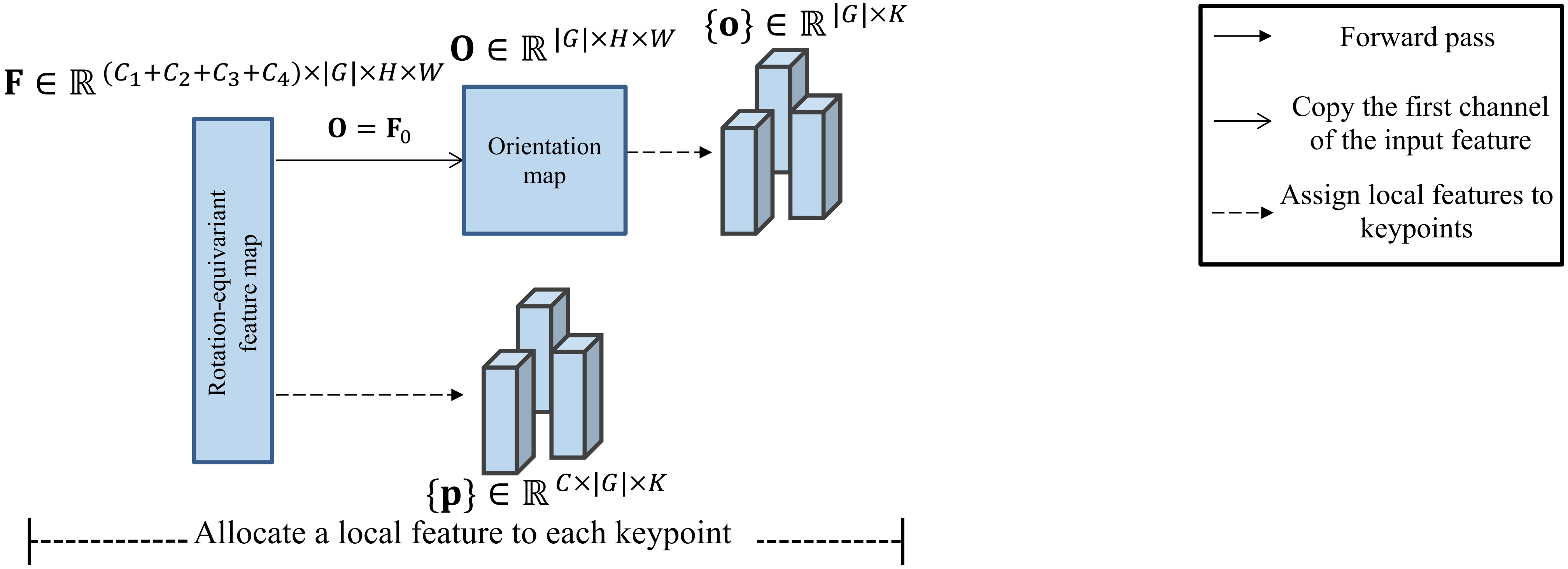
Rotation-Equivariant Feature Extraction



- Rotation-Equivariant ResNet-18 (ReResNet18)[1] constructed by equivariant convolutional layers[2]
- Multi-layer features to exploit the **low-level geometry information** and **high-level semantics** in the local features.

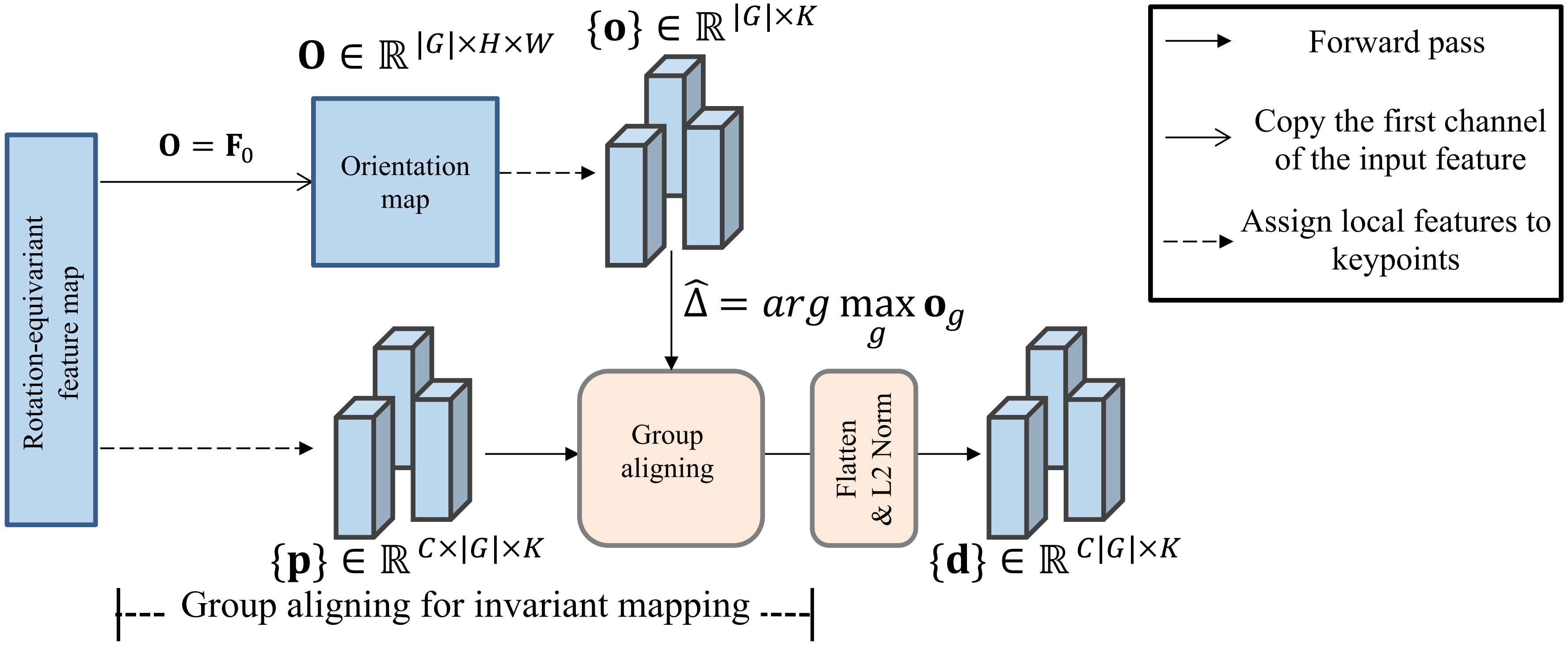
[1] ReDet: A Rotation-Equivariant Detector for Aerial Object Detection. (Han et al., CVPR 2021)
[2] General E(2)-equivariant Steerable CNNs. (Weiler et al., NIPS 2019)

Assigning Local Features to Keypoints



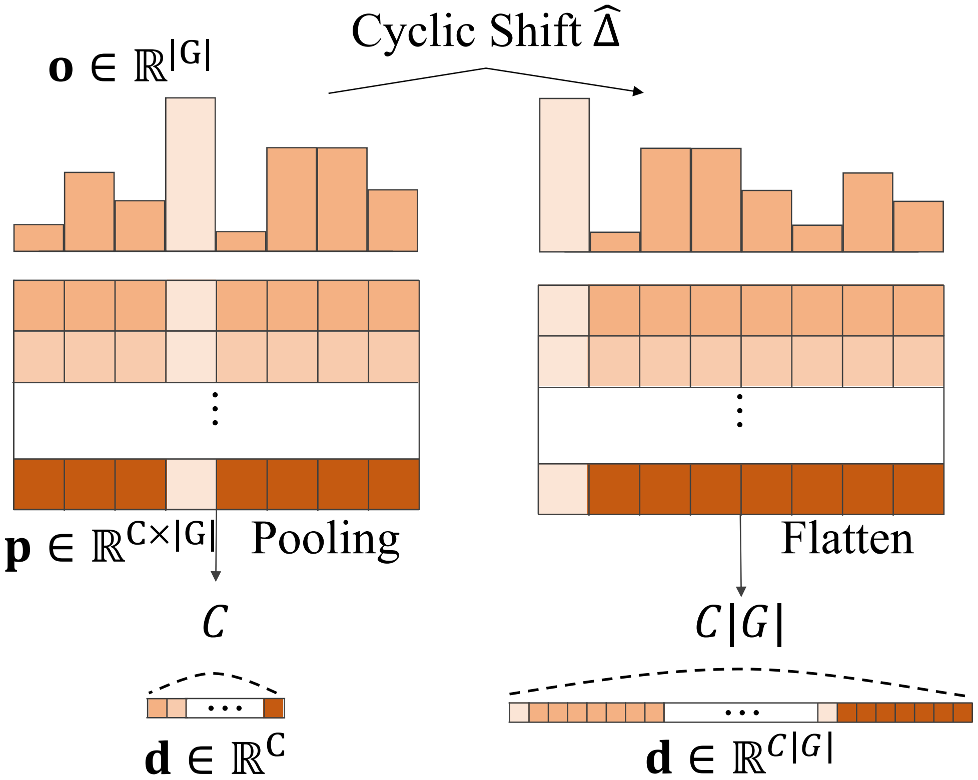
- Extract K number of keypoints ($K = 512$ in training time, using Harris corner detection)
- Allocate a local feature $\mathbf{p} \in \mathbb{R}^{C\times|G|}$ to each keypoint
- Obtain an orientation map $\mathbf{O}$ by selecting the first channel of rotation-equivariant $\mathbf{F}$
    - Allocate an orientation vector $\mathbf{o} \in \mathbb{R}^{|G|}$ to a keypoint

Group Aligning for Invariant Mapping



- Estimating the dominant orientation and the shifting value
- Group aligning
- Descriptor vector normalization (L2 Norm)

Group Aligning vs. Group Pooling



(a) Group pooling (b) Group aligning

$\mathbf{o}$: orientation histogram
$\mathbf{p}$: equivariant feature
$\mathbf{d}$: invariant descriptor
$|G|$: the order of group

Group aligning process

1. Shift a group-equivariant feature $\mathbf{p}$ along the group dimension
2. By its dominant orientation $\hat{\Delta}$

$$\hat{\Delta} = arg\max \mathbf{o}$$

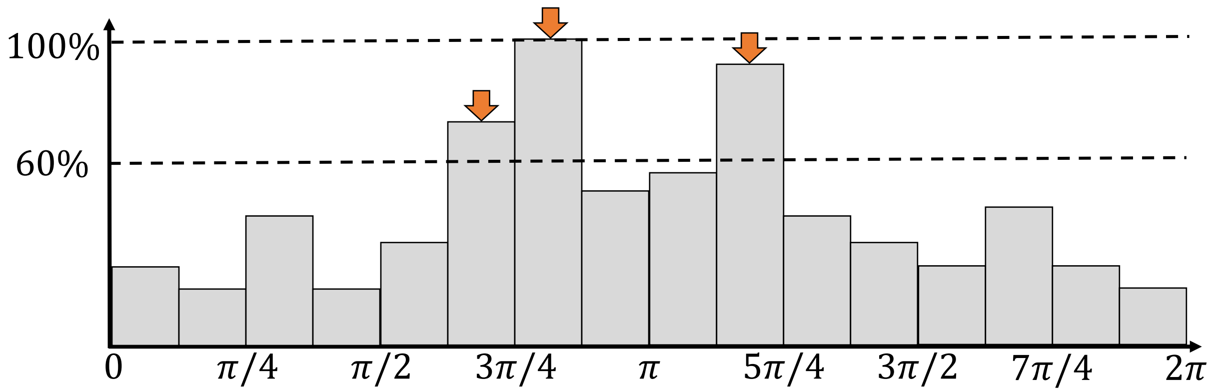
3. Obtain a rotation-invariant descriptor $\mathbf{d}$

- Advantages
  - Without **having to collapse the group information** unlike group pooling
  - Preserving **feature discriminability**.

The final output descriptor size is 1,024 with $C = 64, |G| = 16$.

Additional Functionality of Group Aligning

- **Multiple descriptor extraction** using orientation candidates



An example of multiple descriptor extraction.

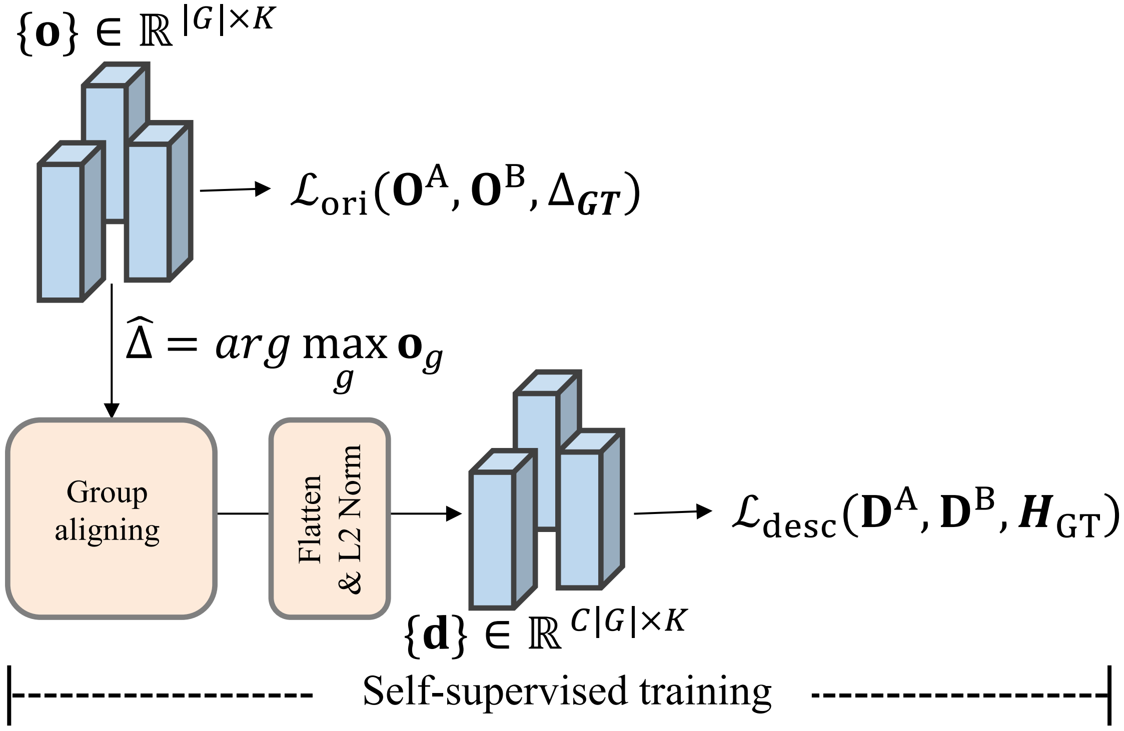
In an orientation histogram $o \in \mathbb{R}^{16}$,
we select **multiple candidates of dominant orientations.**

Different alignments in *Group*-dim

- **Compensating for incorrect orientation predictions**
- Improving matching accuracy **by correcting false matches** using all the hypotheses

Self-Supervised Equivariant Training

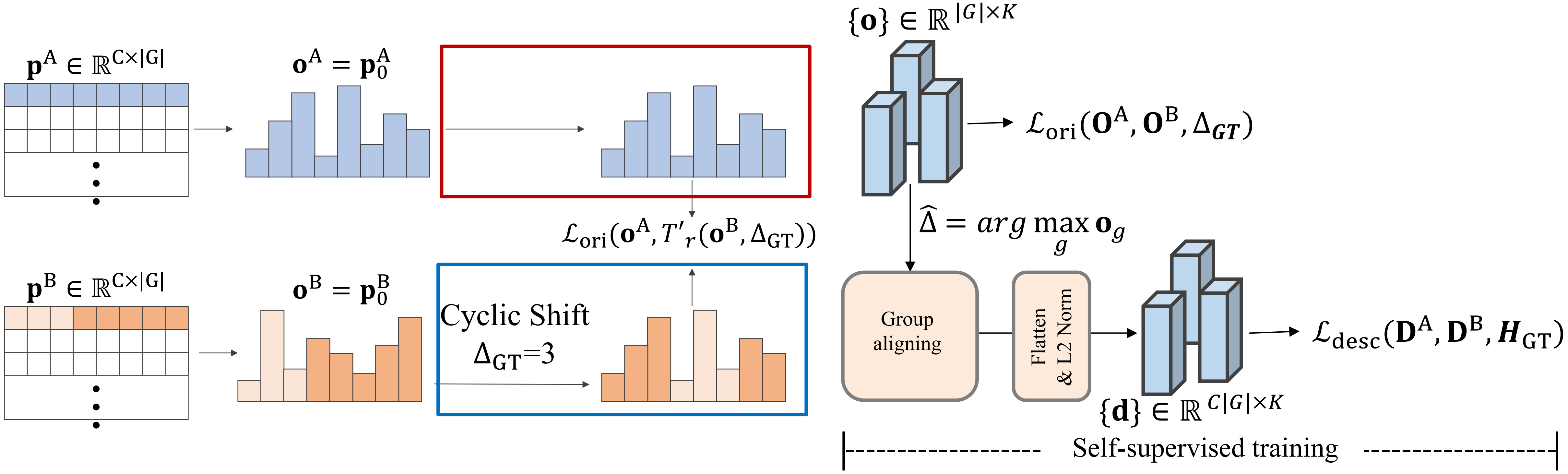
$$\mathcal{L}^{\text{desc}}(\mathbf{D}^{\mathrm{A}}, \mathbf{D}^{\mathrm{B}}) = \sum_{(\mathbf{d}_i^{\mathrm{A}}, \mathbf{d}_i^{\mathrm{B}}) \in (\mathbf{D}^{\mathrm{A}}, \mathbf{D}^{\mathrm{B}})} -\log \frac{\exp(\mathrm{sim}(\mathbf{d}_i^{\mathrm{A}}, \mathbf{d}_i^{\mathrm{B}})/\tau)}{\sum_{k \in K \setminus i} \exp(\mathrm{sim}(\mathbf{d}_i^{\mathrm{A}}, \mathbf{d}_k^{\mathrm{B}}))/\tau)},$$



- Two Loss functions: output robust to the other imaging variations (e.g., illumination, affine ...)
  - Orientation alignment loss
  - Contrastive descriptor loss[1]

[1] A Simple Framework for Contrastive Learning of Visual Representations (Chen et al., ICML 2020)

Orientation Alignment Loss[1], [2]

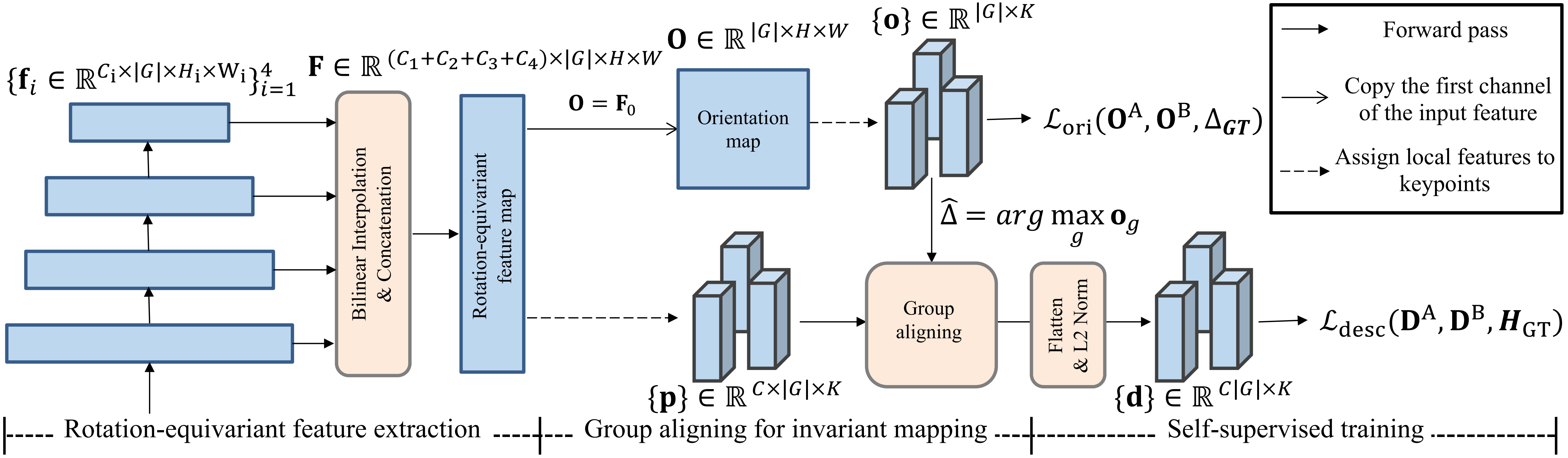


$$\mathcal{L}^{\mathrm{ori}}(\mathbf{O}^{\mathrm{A}}, \mathbf{O}^{\mathrm{B}}, \Delta_{\mathrm{GT}}) = -\sum_{k\in K}\sum_{g\in G} \sigma\left(\boxed{\mathbf{O}^{\mathrm{A}}_{g,k}}\right)\log\left(\sigma\left(\boxed{T'_r(\mathbf{O}^{\mathrm{B}}_{g,k}, \Delta_{\mathrm{GT}})}\right)\right),$$

$$\underline{T'_r(\mathbf{O}_i, \Delta_{\mathrm{GT}}) = \mathbf{O}_{(i+\Delta_{\mathrm{GT}}) \bmod |G|}}$$

[1] Self-Supervised Learning of Image Scale and Orientation (Lee et al., BMVC 2021)
[2] Self-Supervised Equivariant Learning for Oriented Keypoint Detection (Lee et al., CVPR 2022)

Overall architecture



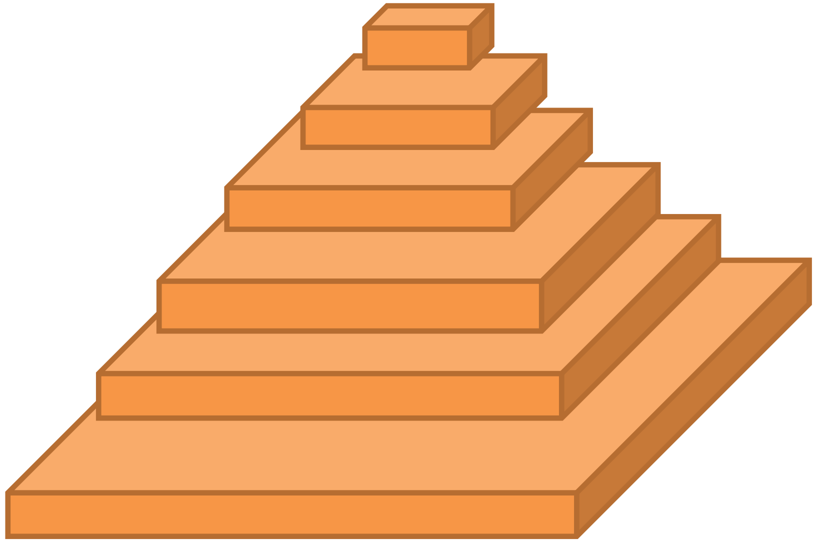
Experimental Settings

Joint training

$$\mathcal{L} = \alpha\mathcal{L}^{\mathrm{ori}} + \mathcal{L}^{\mathrm{desc}},$$

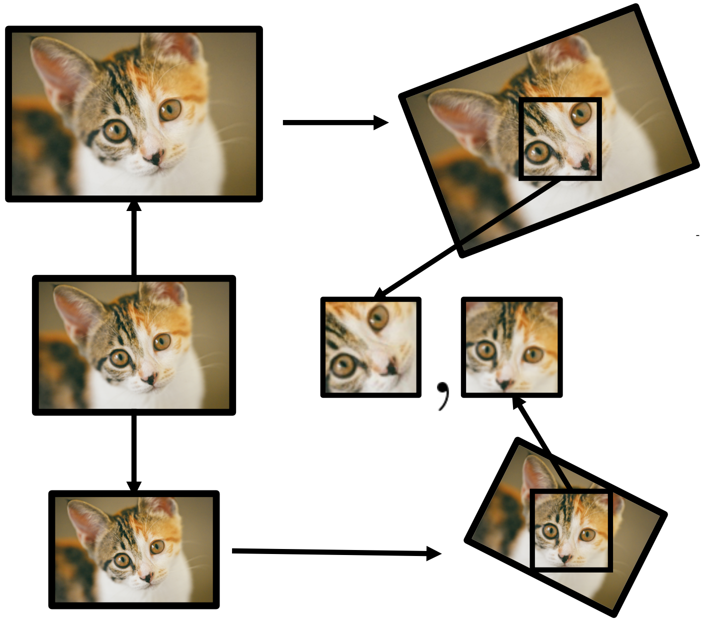
α is 10.

Image pyramid robust to scale changes



Synthetic training dataset

$$D_i = (I_i, I_i', \mathcal{H}_i^{\mathrm{GT}}, \theta_i^{\mathrm{GT}})$$



$$\theta^{\mathrm{GT}} = \arctan\left(\frac{\mathcal{H}_{21}^{\mathrm{GT}}}{\mathcal{H}_{11}^{\mathrm{GT}}}\right)$$

$\mathcal{H}^{\mathrm{GT}}$ is 3×3 martrix.

Evaluation Datasets and Metrics

- Roto-360
  - 360 image pairs
  - **In-plane rotation** from $0°$ to $350°$ at $10°$ intervals

- HPatches
  - 57 scenes, illumination / 59 scenes, viewpoint
  - Each scene contains five image pairs with ground-truth **planar homography**

- MVS dataset
  - Six image sequences of outdoor scenes
  - Ground-truth **camera pose**

To evaluate the rotational invariance

- Roto-360, HPatches
  - **Mean matching accuracy (MMA)** of 3/5/10 pixel thresholds (precision)
  - **The number of predicted matches** (recall)

- MVS dataset
  - **Relative pose estimation accuracy** of $5°/10°/20°$ angular difference thresholds.

To evaluate the general transformations (homography, viewpoint change)

Comparison to Other Invariant Mappings

- Evaluation on Roto-360

	MMA @1px	pred.
Group aligning	**97.54**	**84.9**
Average pooling	57.92	60.8
Max pooling	33.72	51.5
Bilinear pooling[1]	26.42	43.6
w/o invariant map	23.97	32.6

with GT keypoint pairs without training

	MMA @10px	MMA @5px	MMA @3px	pred.
Align	**93.08**	**91.35**	**90.18**	688.3
Avg	85.84	82.12	81.05	**705.9**
Max	82.61	78.00	77.79	686.0
Bilinear	42.69	41.03	40.51	332.5
w/o	19.68	18.81	18.57	349.1

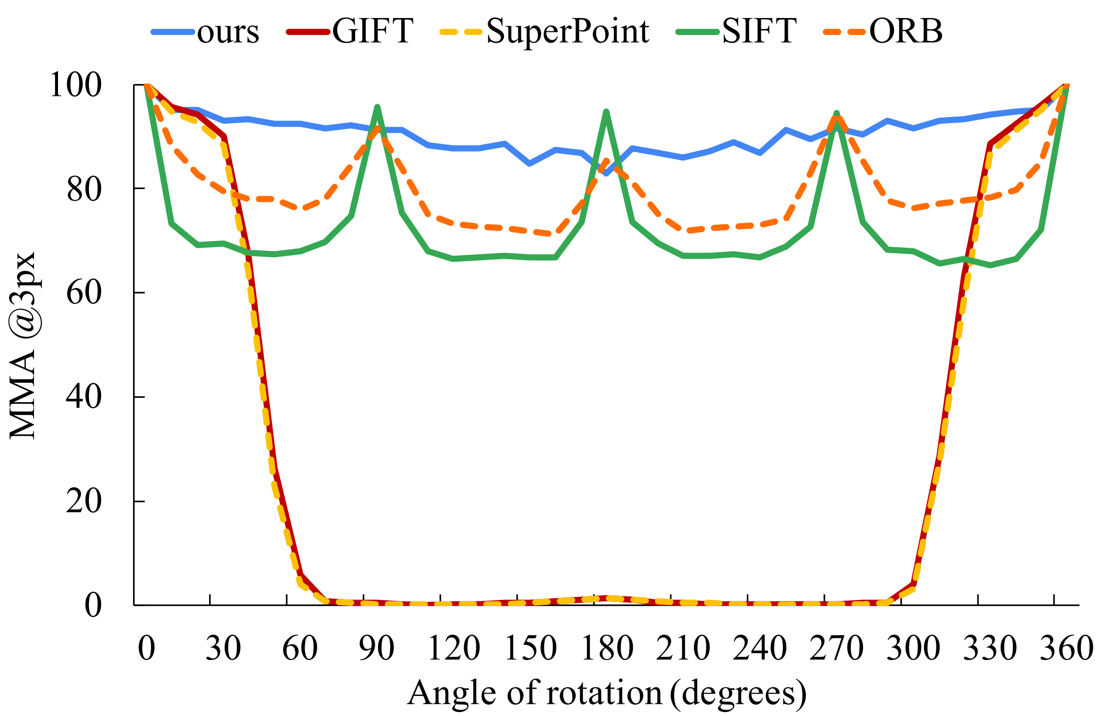
with predicted keypoint pairs with training

[1] GIFT: Learning Transformation-Invariant Dense Visual Descriptors via Group CNNs (Liu et al., NIPS 2019)

Comparison to Existing Local Descriptors

Detector	Descriptor	MMA @10px	MMA @5px	pred.	total.
SIFT	SIFT	78.86	78.59	774.1	1500
	GIFT	37.97	36.82	531.2	1500
	ours	84.67	79.85	558.3	1500
	ours*	**84.91**	**80.09**	759.8	2219
LF-Net	LF-Net	75.05	**74.30**	386.7	1024
	GIFT	35.56	33.82	426.3	1024
	ours	79.90	71.63	431.8	1024
	ours*	**80.32**	71.99	591.4	1503
SuperPoint	SuperPoint	22.85	22.10	462.6	1161
	GIFT	42.35	42.05	589.2	1161
	ours	93.08	91.35	688.3	1161
	ours*	**94.35**	**92.82**	1333.0	2340

bold: best result. Underlined: second best. *: multiple descriptor extraction.



Matching accuracies according to varying degree of rotations on Roto-360.

Distinctive image features from scale-invariant keypoints. (Lowe, IJCV 2004)
LF-Net: Learning Local Features from Images (Ono et al., NIPS 2018)
SuperPoint: Self-Supervised Interest Point Detection and Description (DeTone et al., CVPRW 2018)
GIFT: Learning Transformation-Invariant Dense Visual Descriptors via Group CNNs (Liu et al., NIPS 2019)

Results on Homography & Viewpoint Changes

Method	HP-illu		HP-view		MVS-Pose		
	@5px	@3px	@5px	@3px	20°	10°	5°
SIFT	49.08	44.62	53.57	47.96	0.02	0.00	0.00
SuperPoint	74.63	67.53	64.96	56.17	0.20	0.07	0.01
LF-Net	62.21	57.63	50.88	47.00	0.06	0.03	0.01
RF-Net	61.63	57.46	56.62	51.49	0.10	0.04	0.01
GIFT	**79.71**	**71.89**	72.48	62.88	**0.60**	0.28	0.09
$\text{ours}_{\text{avgpool}}$	62.28	56.27	65.85	59.55	0.27	0.10	0.05
$\text{ours}_{\text{maxpool}}$	59.66	53.91	63.42	57.64	0.27	0.11	0.03
$\text{ours}_{\text{bilinearpool}}$	45.13	41.57	46.03	42.22	0.35	0.17	0.09
$\text{ours}_{\text{groupalign}}$	70.39	62.88	70.97	63.95	0.58	0.26	0.12
$\text{ours}_{\text{groupalign}}$*	73.13	65.33	74.69	67.38	0.56	0.30	0.12
$\text{ours}_{\text{bilinearpool}}$†	57.32	52.67	60.06	54.83	0.24	0.11	0.03
$\text{ours}_{\text{groupalign}}$†	77.94	69.35	**78.06**	**70.03**	0.56	**0.33**	**0.14**

Our invariant pooling still performs best.

* denotes multiple descriptor extraction. † is larger backbone.

Results on Homography & Viewpoint Changes

Method	HP-illu		HP-view		MVS-Pose		
	@5px	@3px	@5px	@3px	20°	10°	5°
SIFT	49.08	44.62	53.57	47.96	0.02	0.00	0.00
SuperPoint	74.63	67.53	64.96	56.17	0.20	0.07	0.01
LF-Net	62.21	57.63	50.88	47.00	0.06	0.03	0.01
RF-Net	61.63	57.46	56.62	51.49	0.10	0.04	0.01
GIFT	**79.71**	**71.89**	72.48	62.88	**0.60**	0.28	0.09
$\text{ours}_{\text{avgpool}}$	62.28	56.27	65.85	59.55	0.27	0.10	0.05
$\text{ours}_{\text{maxpool}}$	59.66	53.91	63.42	57.64	0.27	0.11	0.03
$\text{ours}_{\text{bilinearpool}}$	45.13	41.57	46.03	42.22	0.35	0.17	0.09
$\text{ours}_{\text{groupalign}}$	70.39	62.88	70.97	63.95	0.58	0.26	0.12
$\text{ours}_{\text{groupalign}}$*	73.13	65.33	74.69	67.38	0.56	0.30	0.12
$\text{ours}_{\text{bilinearpool}}$†	57.32	52.67	60.06	54.83	0.24	0.11	0.03
$\text{ours}_{\text{groupalign}}$†	77.94	69.35	**78.06**	**70.03**	0.56	**0.33**	**0.14**

Multiple descriptor extraction increases the performance.

* denotes multiple descriptor extraction. † is larger backbone.

Results on Homography & Viewpoint Changes

Method	HP-illu		HP-view		MVS-Pose		
	@5px	@3px	@5px	@3px	20°	10°	5°
SIFT	49.08	44.62	53.57	47.96	0.02	0.00	0.00
SuperPoint	74.63	67.53	64.96	56.17	0.20	0.07	0.01
LF-Net	62.21	57.63	50.88	47.00	0.06	0.03	0.01
RF-Net	61.63	57.46	56.62	51.49	0.10	0.04	0.01
GIFT	**79.71**	**71.89**	72.48	62.88	**0.60**	0.28	0.09
ours$_{avgpool}$	62.28	56.27	65.85	59.55	0.27	0.10	0.05
ours$_{maxpool}$	59.66	53.91	63.42	57.64	0.27	0.11	0.03
ours$_{bilinearpool}$	45.13	41.57	46.03	42.22	0.35	0.17	0.09
ours$_{groupalign}$	70.39	62.88	70.97	63.95	0.58	0.26	0.12
ours$_{groupalign}$*	73.13	65.33	74.69	67.38	0.56	0.30	0.12
ours$_{bilinearpool}$†	57.32	52.67	60.06	54.83	0.24	0.11	0.03
ours$_{groupalign}$†	77.94	69.35	**78.06**	**70.03**	0.56	**0.33**	**0.14**

Our group aligning performs better than bilinear pooling proposed in GIFT

* denotes multiple descriptor extraction. † is larger backbone.

Ablation Study & Design Choice

	HP-all		Roto-360		params.
	@5px	@3px	@5px	@3px	(millions)
ours (proposed $\|G\| = 16$)	**70.69**	**63.42**	91.35	90.18	0.62M
w/o orientation loss	66.41	58.61	85.29	83.26	0.62M
w/o descriptor loss	27.49	24.83	25.64	24.98	0.62M
w/o image scale pyramid	68.77	62.25	**91.47**	**90.43**	0.62M
w/o equivariant backbone	47.25	42.52	8.65	8.51	11.18M
$\|G\| = 64$	63.96	57.35	85.12	83.32	**0.16M**
$\|G\| = 36$	68.17	60.95	87.78	85.89	0.26M
$\|G\| = 32$	69.44	62.08	89.10	87.31	0.31M
$\|G\| = 24$	69.72	62.21	90.27	88.34	0.39M
$\|G\| = 8$	65.74	58.92	87.16	85.57	1.24M

Ablation Study & Design Choice

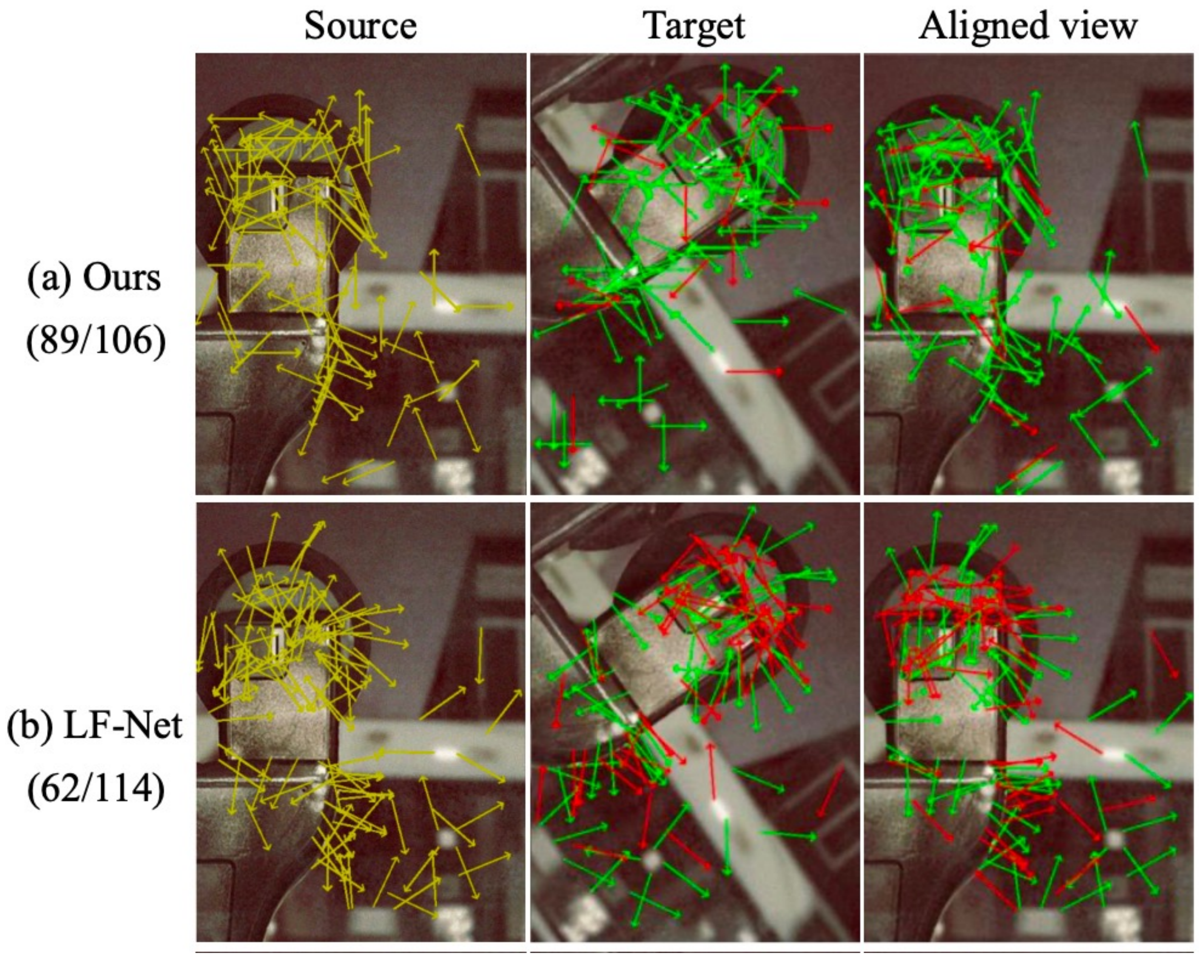
	HP-all		Roto-360		params. (millions)
	@5px	@3px	@5px	@3px	
ours (proposed $\|G\| = 16$)	**70.69**	**63.42**	<u>91.35</u>	<u>90.18</u>	0.62M
w/o orientation loss	66.41	58.61	85.29	83.26	0.62M
w/o descriptor loss	27.49	24.83	25.64	24.98	0.62M
w/o image scale pyramid	<u>68.77</u>	<u>62.25</u>	**91.47**	**90.43**	0.62M
w/o equivariant backbone	47.25	42.52	8.65	8.51	11.18M
$\|G\| = 64$	63.96	57.35	85.12	83.32	**0.16M**
$\|G\| = 36$	68.17	60.95	87.78	85.89	0.26M
$\|G\| = 32$	69.44	62.08	89.10	87.31	0.31M
$\|G\| = 24$	69.72	62.21	90.27	88.34	0.39M
$\|G\| = 8$	65.74	58.92	87.16	85.57	1.24M

Ablation Study & Design Choice

We take our best model $|G| = 16$

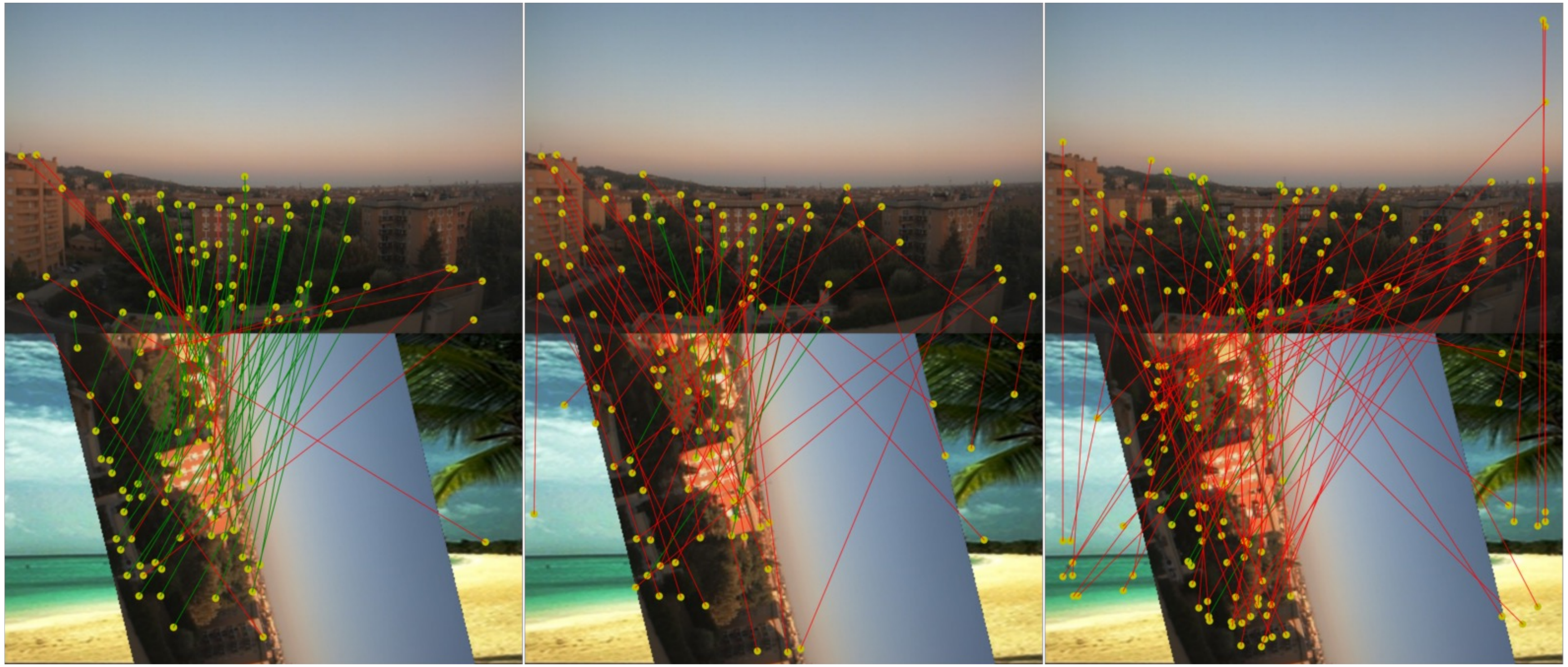
	HP-all		Roto-360		params.
	@5px	@3px	@5px	@3px	(millions)
ours (proposed $\|G\| = 16$)	**70.69**	**63.42**	<u>91.35</u>	<u>90.18</u>	0.62M
w/o orientation loss	66.41	58.61	85.29	83.26	0.62M
w/o descriptor loss	27.49	24.83	25.64	24.98	0.62M
w/o image scale pyramid	<u>68.77</u>	<u>62.25</u>	**91.47**	**90.43**	0.62M
w/o equivariant backbone	47.25	42.52	8.65	8.51	11.18M
$\|G\| = 64$	63.96	57.35	85.12	83.32	**0.16M**
$\|G\| = 36$	68.17	60.95	87.78	85.89	0.26M
$\|G\| = 32$	69.44	62.08	89.10	87.31	0.31M
$\|G\| = 24$	69.72	62.21	90.27	88.34	0.39M
$\|G\| = 8$	65.74	58.92	87.16	85.57	1.24M

Qualitative Results



▲ Consistency of estimated orientations

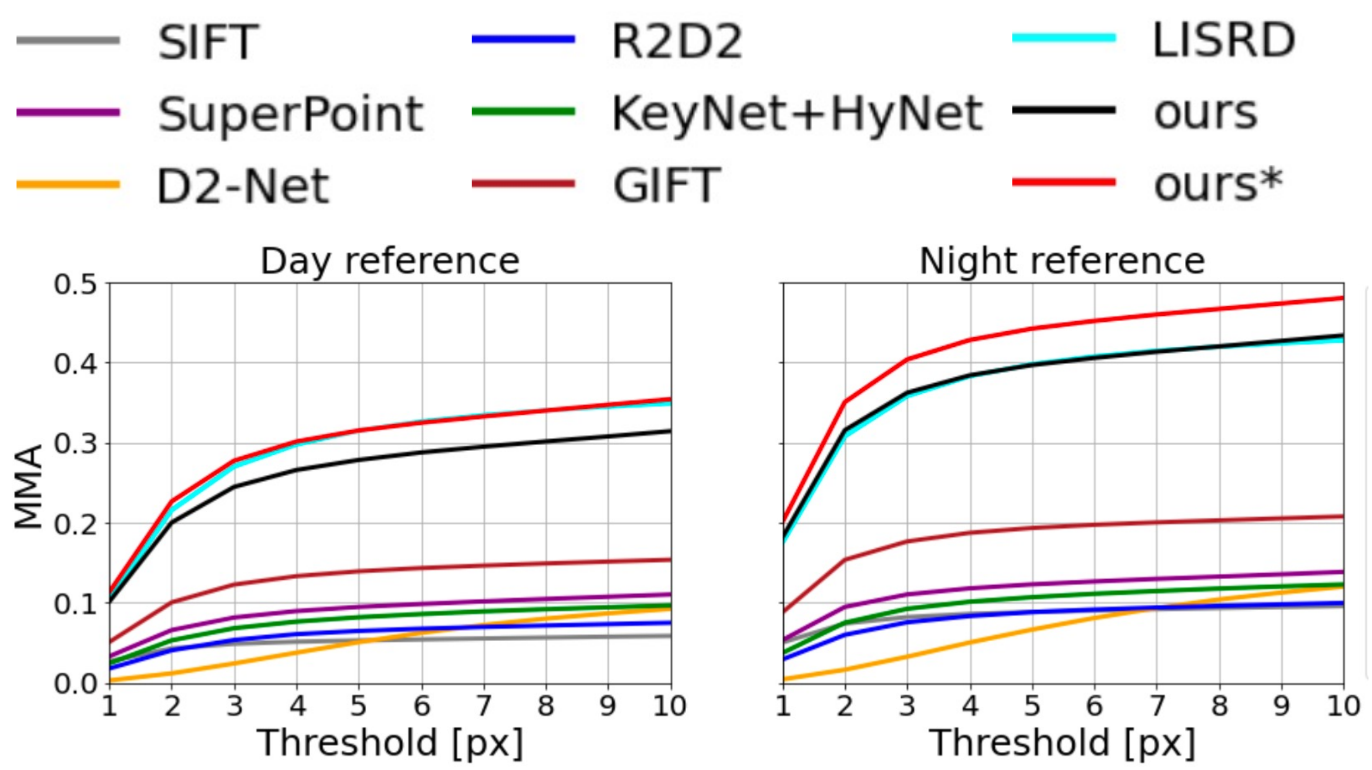
Qualitative Results



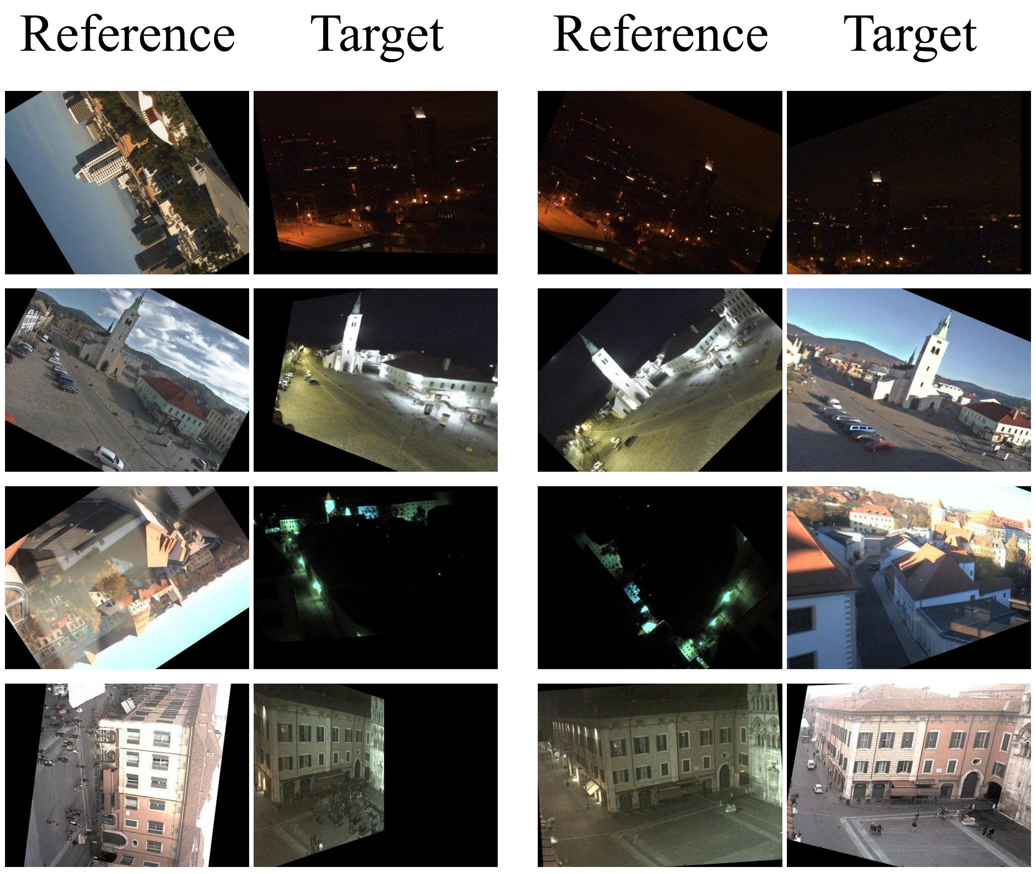
(a) ours (b) GIFT [3] (c) LF-Net [2]

▲ Predicted Matches

Results on *extreme* Rotated Day-Night Matching



- Evaluation to compare the rotational robustness
- Under both geometric/illumination changes



Examples of *e*RDNIM

Conclusion

- Self-supervised rotation-equivariant network
  - for visual correspondence
  - to improve the discriminability of local descriptors
- New invariant mapping operation
  - group-aligning shifts the rotation-equivariant features along the group dimension
  - based on the orientation value to produce rotation-invariant descriptors
  - while preserving the feature discriminability,
  - without collapsing the group dimension.
- Experiments
  - best performance in obtaining rotation-invariant descriptors on Roto-360
  - transferable to tasks such as keypoint matching and camera pose estimation.

Poster Session THU-PM-112

Thursday (22nd, Jun), 4:00pm - 6:00pm

See you soon!

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Thank you!

Project Page **Paper** **Code**

POSTECH
POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Computer**Vision** Lab.