ETH zürich Computer Vision and Geometry Lab

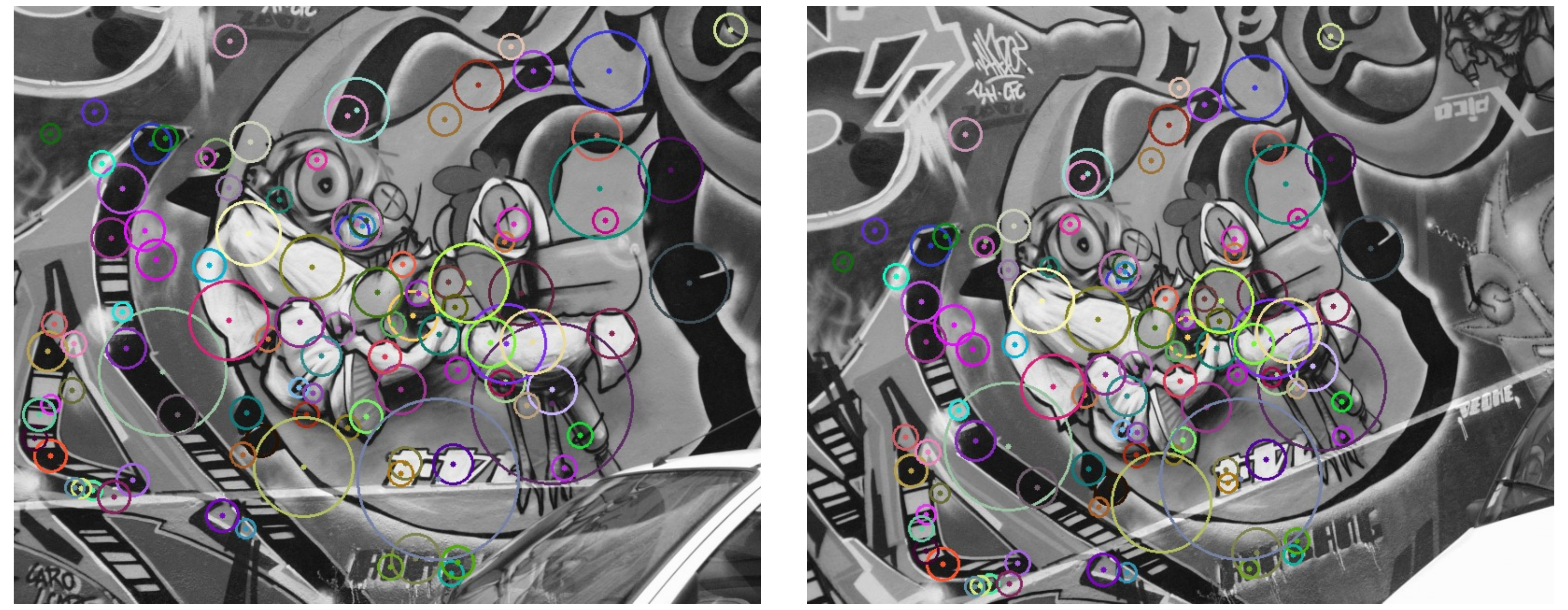
# QUAD-NETWORKS: UNSUPERVISED LEARNING TO RANK FOR INTEREST POINT DETECTION

NIKOLAY.SAVINOV@INF.ETHZ.CH, AKIHITO.SEKI@TOSHIBA.CO.JP, {LUBOR.LADICKY, TORSTEN.SATTLER, MARC.POLLEFEYS}@INF.ETHZ.CH

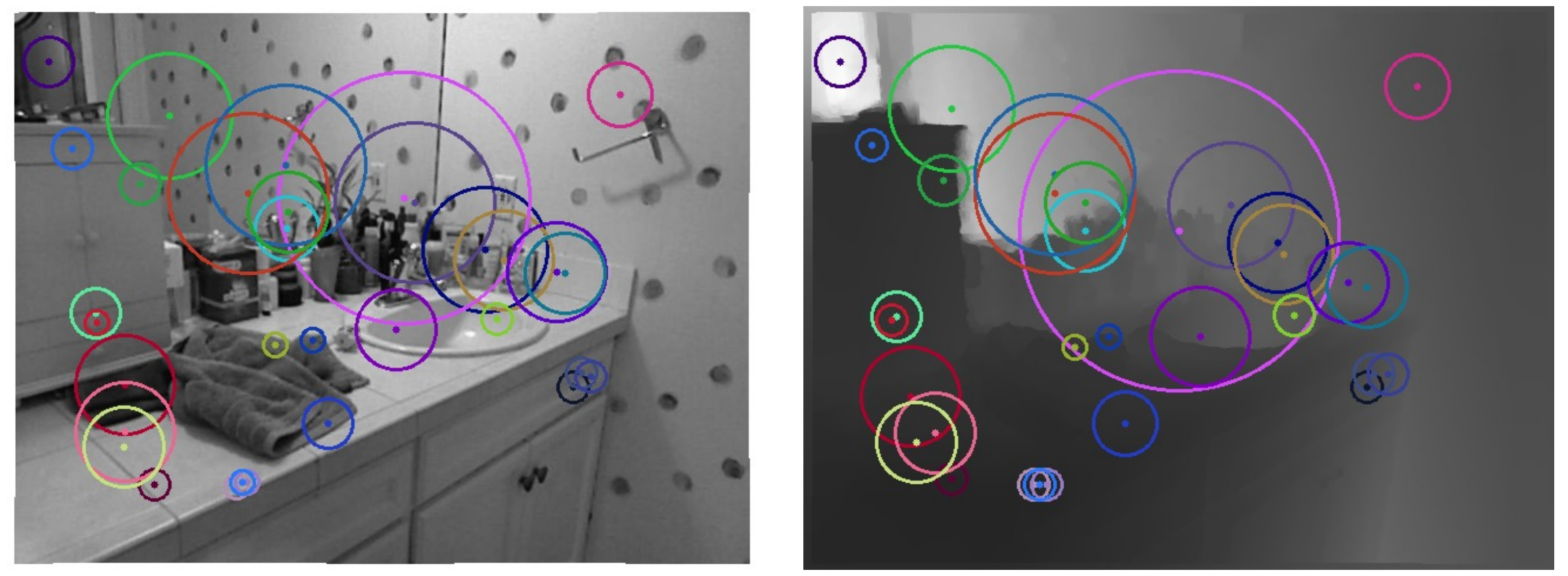
CVPR July 21-26 HONOLULU 2017

## TASK: INTEREST POINT DETECTION

Traditional setting (RGB/RGB):



Cross-modal setting (e.g., RGB/depth):

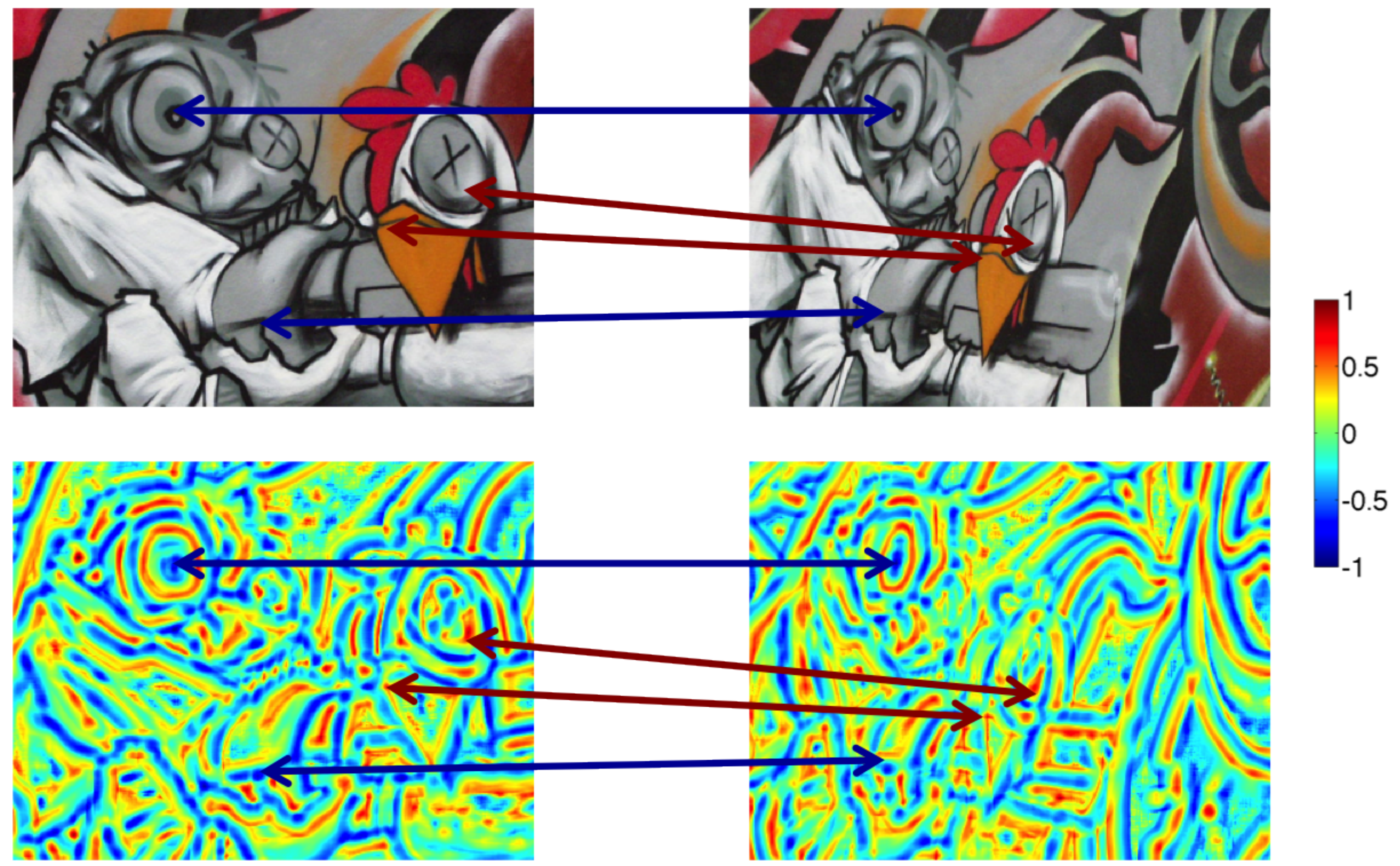


- Goal: detect a sparse subset of points, re-detect the same points after transformations.
- Transformations could be arbitrary: viewpoint/modality/illumination change.
- If further matched, those points allow to estimate the transformation.

## WHY UNSUPERVISED?

- Traditionally, detectors were hand-designed: corners, blobs.
- Unfortunately, in some cases humans have no intuition what points could be "interesting".
- For example: interest point detection between two different modalities, RGB and depth map.
- Simple heuristics will fail: the strongest corners/blobs in RGB might come from texture which is missing in depth maps.
- If we cannot say what interesting and what not, let's avoid such labeling at all — thus unsupervised formulation needed.
- Such unsupervised formulations have not been explored in the previous work: most works learn how to filter points detected by DoG.

## OUR APPROACH: DETECTION BY RANKING



- Top: an image undergoes a perspective change transformation.
- Bottom: our learned response function, visualized as a heat map, produces a ranking of image locations that is reasonably invariant under the transformation.
- Since the resulting ranking is largely repeatable, the top/bottom quantiles of the response function are also repeatable (examples of interest points are shown by arrows).

## FORMULATION

- We want to rank object points and represent this ranking with a single real-valued response function $H(p|w)$.
- $H$ — deep net, $p$ — image patch, $w$ — learned parameters.

Our goal is to have a ranking satisfying

$$\begin{cases} H(p_d^i|w) > H(p_d^j|w) \quad \& \quad H(p_{t(d)}^i|w) > H(p_{t(d)}^j|w) \\ \qquad\qquad\qquad\text{or} \\ H(p_d^i|w) < H(p_d^j|w) \quad \& \quad H(p_{t(d)}^i|w) < H(p_{t(d)}^j|w) \ . \end{cases} \tag{1}$$

We introduce the ranking agreement function

$$R(p_d^i, p_d^j, p_{t(d)}^i, p_{t(d)}^j|w) = \\ (H(p_d^i|w) - H(p_d^j|w))(H(p_{t(d)}^i|w) - H(p_{t(d)}^j|w)) \tag{2}$$

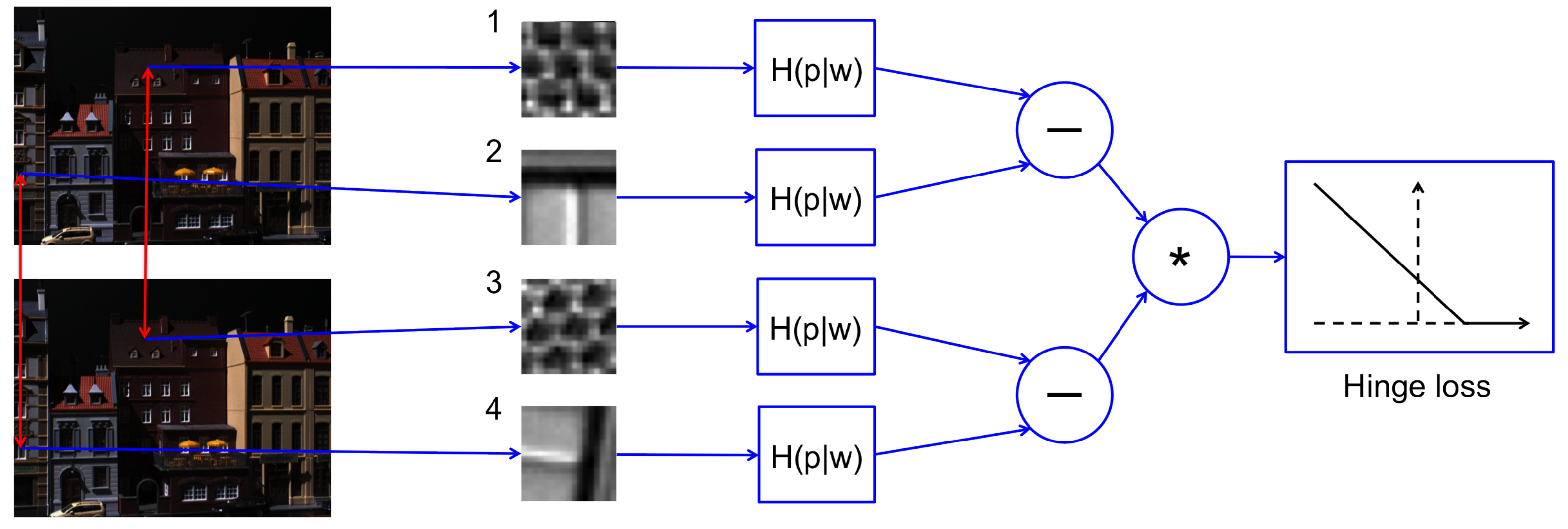
and want to achieve

$$R(p_d^i, p_d^j, p_{t(d)}^i, p_{t(d)}^j|w) > 0 \tag{3}$$

by minimizing the loss

$$L(w) = \sum_{d\in D}\sum_{t\in T}\sum_{i,j\in C_{dt}} \max(0, 1 - R(p_d^i, p_d^j, p_{t(d)}^i, p_{t(d)}^j|w)) \ . \tag{4}$$

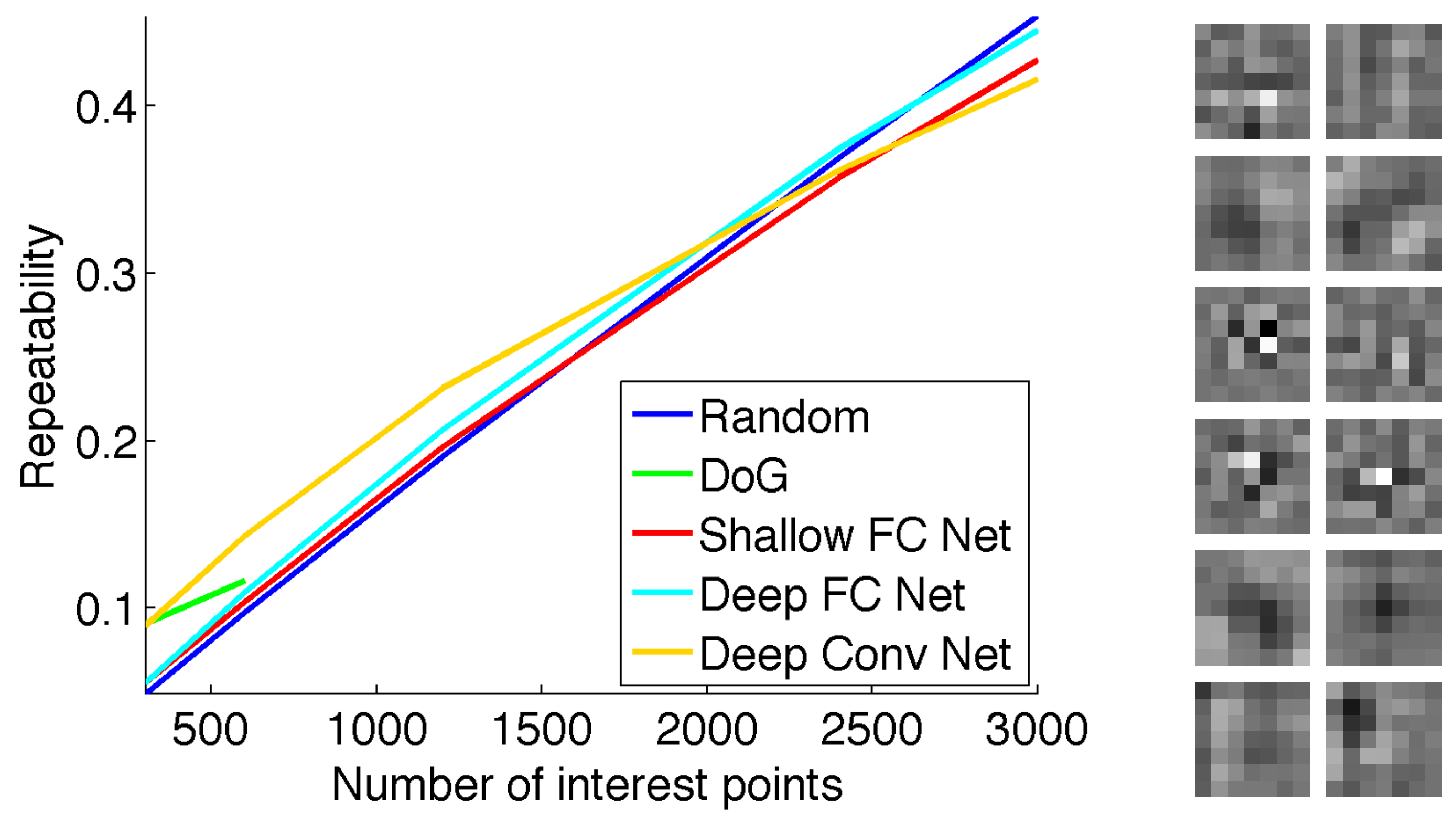
## TRAINING



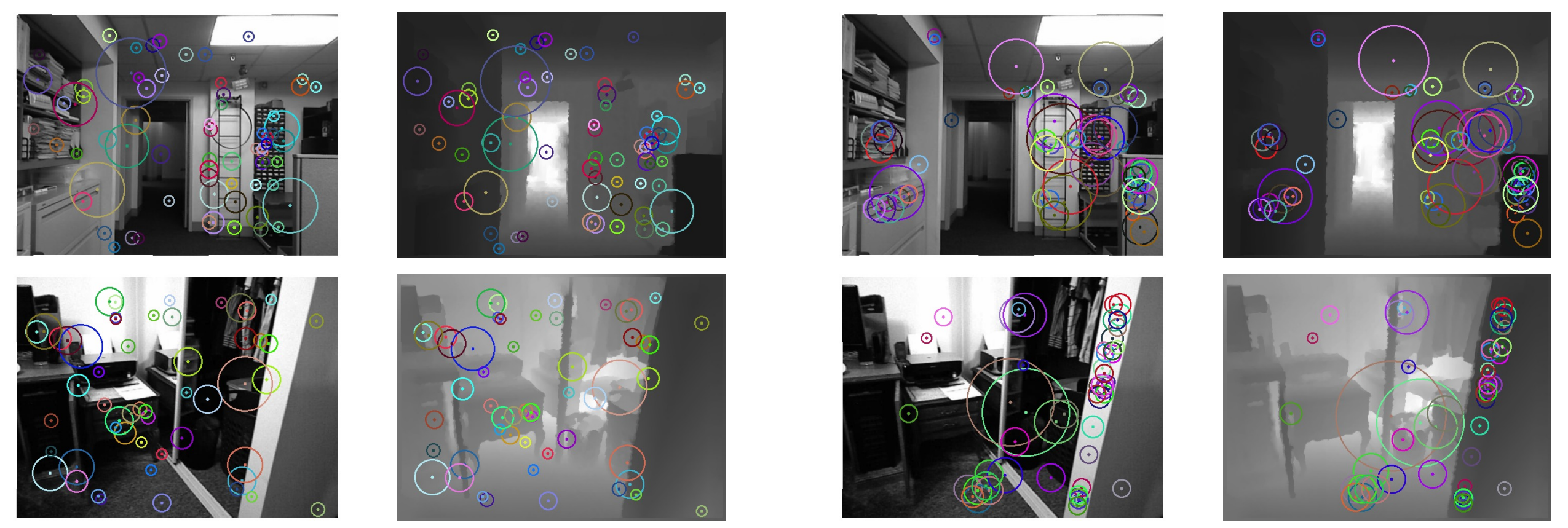
- Quad-network forward pass on a training quadruple is shown.
- Patches $(1, 3)$ and $(2, 4)$ are correspondence pairs between two different images, so $1, 2$ come from the first image and $3, 4$ come from the second image.
- All of the patches are extracted with a random rotation.

## EXPERIMENTS: CROSS-MODAL DETECTION

Repeatability (the higher the curve, the better) and filters from our "Deep Conv Net" model:



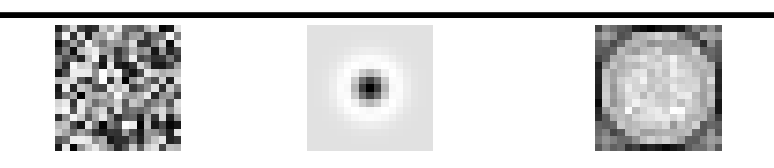
Detections from DoG and our "Deep Conv Net" model:



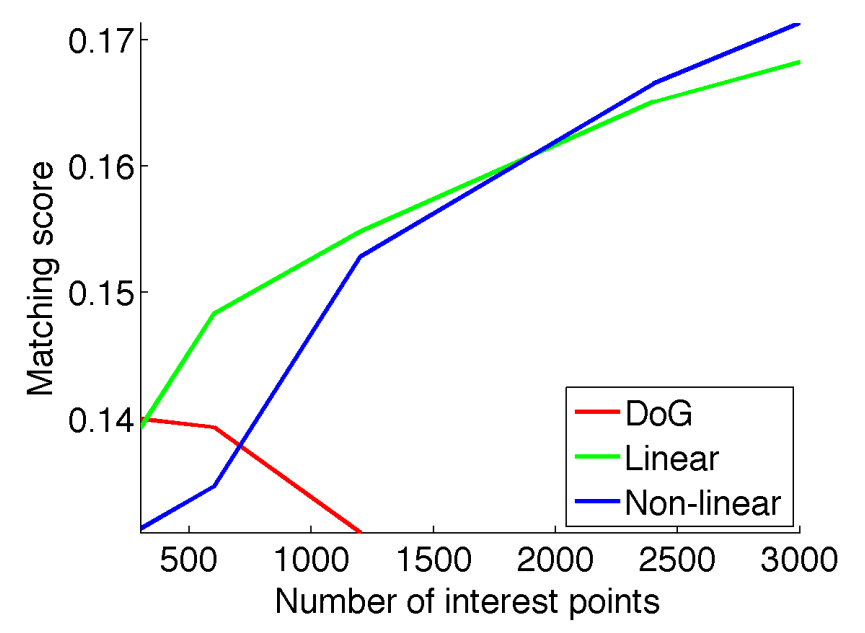
DoG image DoG depth ours image ours depth

## EXPERIMENTS: TRADITIONAL DETECTION

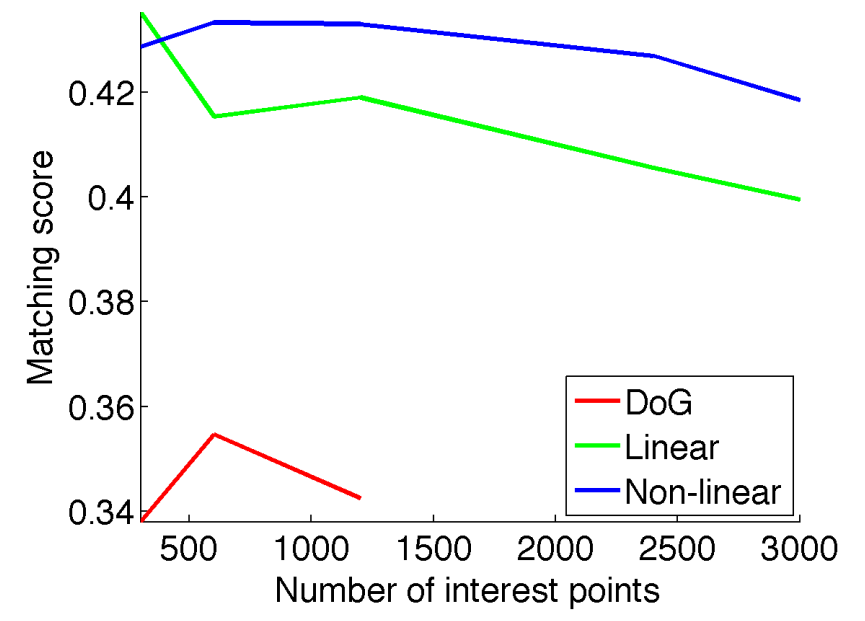
|  |  |  | Number of interest points |  |  |  |  |
|---|---|---|---|---|---|---|---|
| T | Data | Method | 300 | 600 | 1200 | 2400 | 3000 |
| VP | graf | Random | 0.06 | 0.08 | 0.12 | 0.17 | 0.19 |
|  |  | DoG | **0.21** | **0.2** | 0.18 | - | - |
|  |  | Linear | 0.17 | 0.18 | 0.19 | 0.21 | 0.22 |
|  |  | Non-lin | 0.17 | 0.19 | **0.21** | **0.24** | **0.25** |
|  | wall | Random | 0.18 | 0.22 | 0.27 | 0.33 | 0.36 |
|  |  | DoG | 0.27 | 0.28 | 0.28 | - | - |
|  |  | Linear | **0.33** | **0.36** | **0.39** | 0.43 | 0.44 |
|  |  | Non-lin | 0.3 | 0.35 | **0.39** | **0.44** | **0.46** |
| Z+R | bark | Random | 0.02 | 0.03 | 0.05 | 0.08 | 0.1 |
|  |  | DoG | 0.13 | 0.13 | - | - | - |
|  |  | Linear | **0.14** | **0.15** | **0.15** | 0.15 | - |
|  |  | Non-lin | 0.12 | 0.13 | 0.14 | **0.16** | **0.16** |
|  | boat | Random | 0.03 | 0.05 | 0.08 | 0.11 | 0.12 |
|  |  | DoG | 0.26 | 0.25 | 0.2 | - | - |
|  |  | Linear | **0.27** | **0.27** | 0.27 | 0.26 | 0.25 |
|  |  | Non-lin | 0.21 | 0.24 | **0.28** | **0.28** | **0.29** |
| L | leuven | Random | 0.51 | 0.57 | 0.63 | 0.69 | 0.71 |
|  |  | DoG | 0.51 | 0.51 | 0.5 | - | - |
|  |  | Linear | 0.69 | 0.69 | 0.73 | 0.73 | 0.72 |
|  |  | Non-lin | **0.7** | **0.72** | **0.75** | **0.76** | **0.77** |
| Blur | bikes | Random | 0.36 | 0.42 | 0.48 | 0.53 | 0.54 |
|  |  | DoG | 0.41 | 0.41 | 0.39 | - | - |
|  |  | Linear | **0.53** | **0.53** | 0.49 | **0.55** | **0.57** |
|  |  | Non-lin | 0.52 | 0.51 | **0.51** | 0.49 | 0.49 |
|  | trees | Random | 0.21 | 0.26 | 0.32 | 0.4 | 0.43 |
|  |  | DoG | 0.29 | 0.3 | 0.31 | - | - |
|  |  | Linear | 0.34 | 0.37 | 0.42 | 0.45 | **0.5** |
|  |  | Non-lin | **0.36** | **0.39** | **0.44** | **0.49** | **0.5** |
| JPEG | ubc | Random | 0.42 | 0.47 | 0.53 | 0.59 | 0.61 |
|  |  | DoG | **0.68** | 0.6 | - | - | - |
|  |  | Linear | 0.55 | **0.62** | **0.66** | 0.67 | 0.68 |
|  |  | Non-lin | 0.58 | **0.62** | 0.64 | **0.69** | **0.7** |



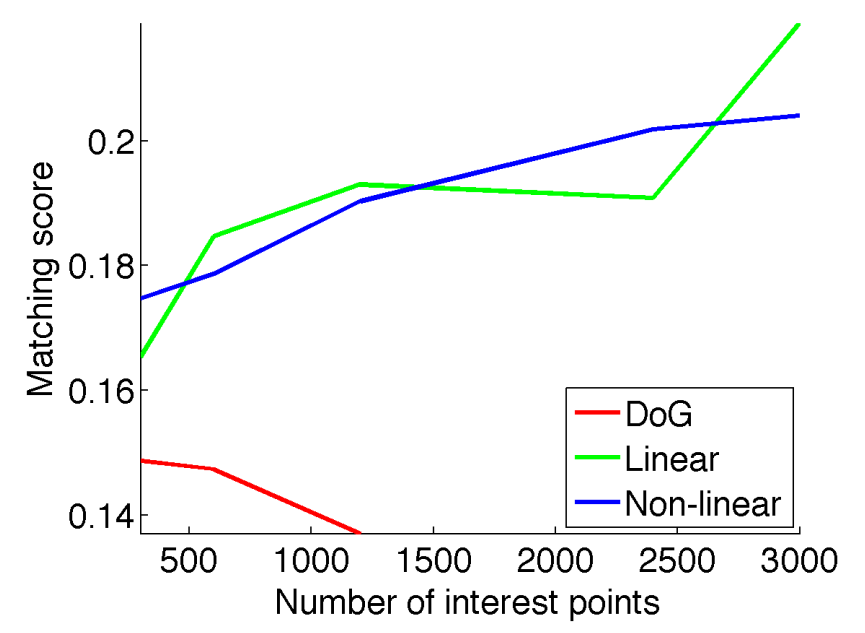
random DoG ours linear



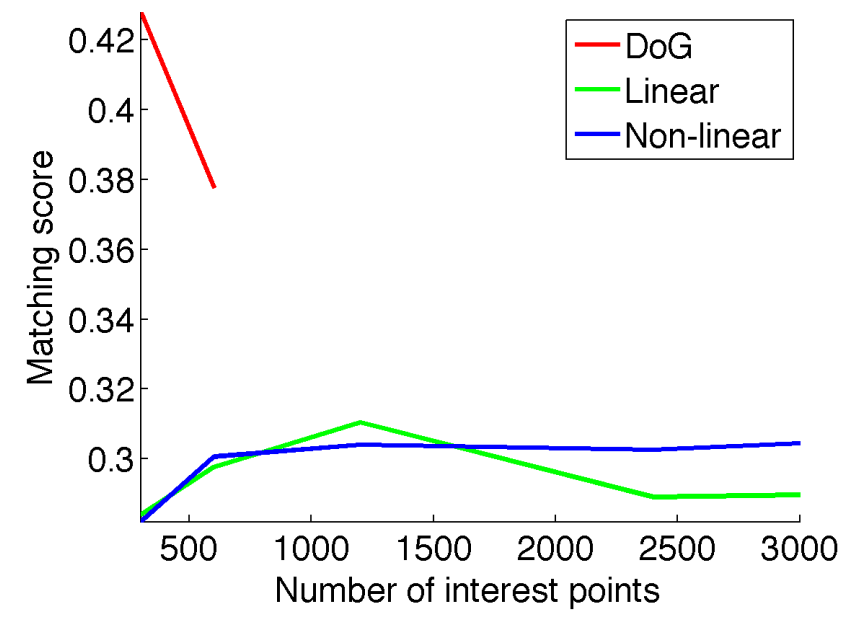
Wall (viewpoint)



Leuven (illumination)



Trees (blur)



UBC (jpeg)

- Left: repeatability (the higher, the better) and filters.
- Right: matching score (the higher the curve, the better).
- The detections from our methods ("Linear", "Non-linear") have similar or better repeatability/matching score compared to DoG.

## WHAT'S NEXT?

- Learning the descriptor jointly with our detector.
- Trying our method for other modality pairs (e.g., infrared and RGB).
- Applying our method to detection beyond images (e.g., to interest frame detection in videos).

## ACKNOWLEDGEMENTS

We would like to thank Andrea Cohen, Dmitry Laptev and Victor Lempitsky for their valuable feedback on this work. This work is partially funded by the Swiss NSF project 163910, the Max Planck CLS Fellowship and the Swiss CTI project 17136.1 PFES-ES.