# Robust Learning-based Image Matching

## IMW CVPR 2019

Dawei Sun Zixin Luo Jiahui Zhang

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Altizure

intel

清華大學
Tsinghua University

# 00 Outline

An image matching pipeline: 1) local keypoint detection, 2) local keypoint description, 3) sparse matching.

# 00 Outline

An image matching pipeline: 1) local keypoint detection, 2) local keypoint description, 3) sparse matching.

Part 1

## ContextDesc: a learning-based local descriptor

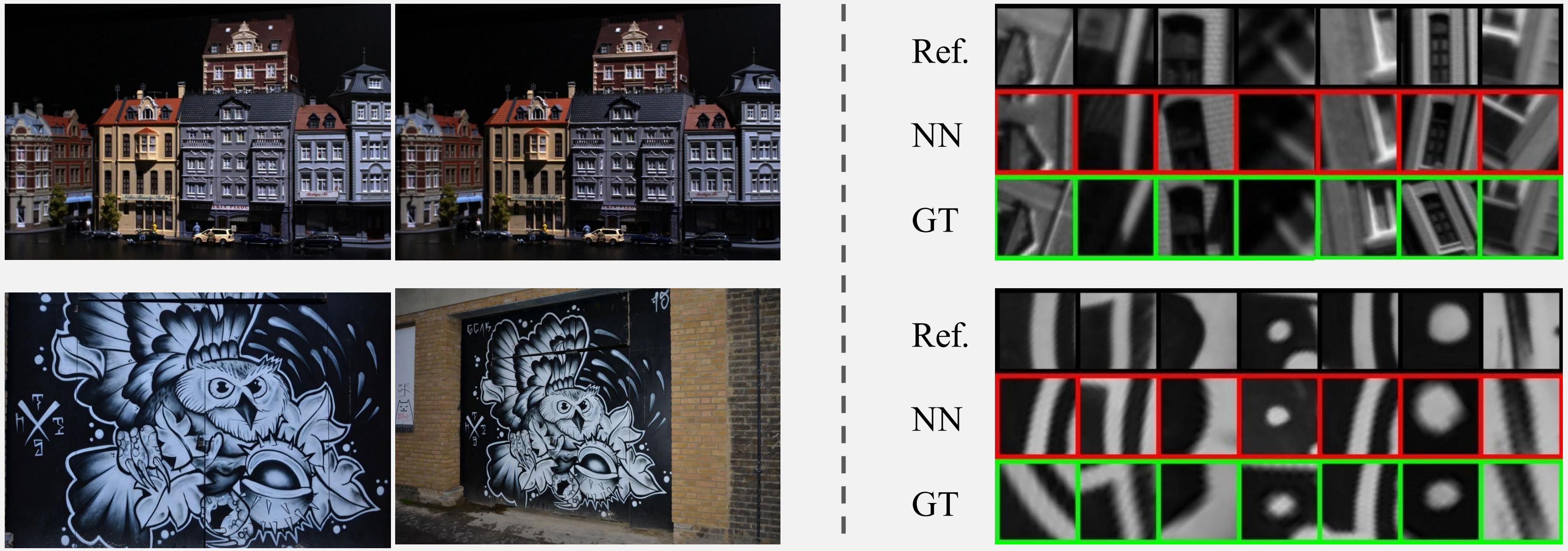
ContextDesc: Local Descriptor Augmentation with Cross-Modality Context, CVPR'19

Part 2

## A learning-based inlier classification and fundamental matrix estimation method

In submission

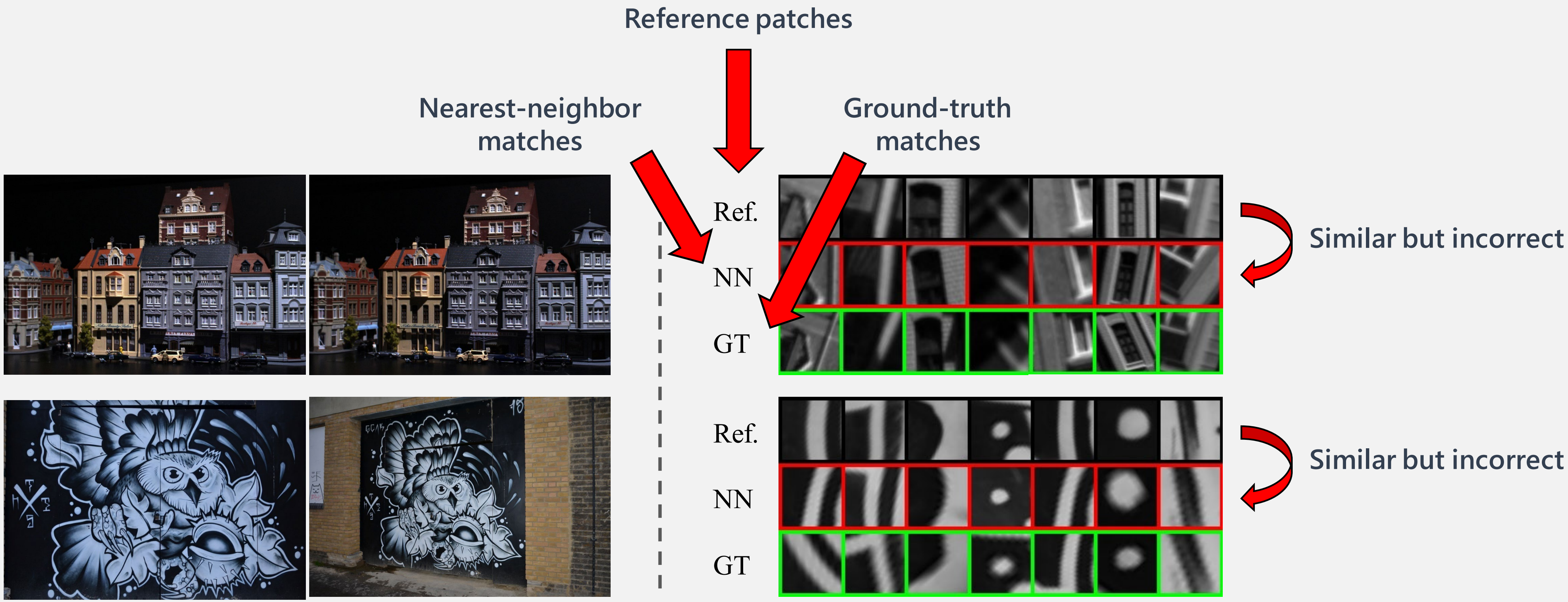
# 01 Motivation




The results are becoming saturated on standard benchmark

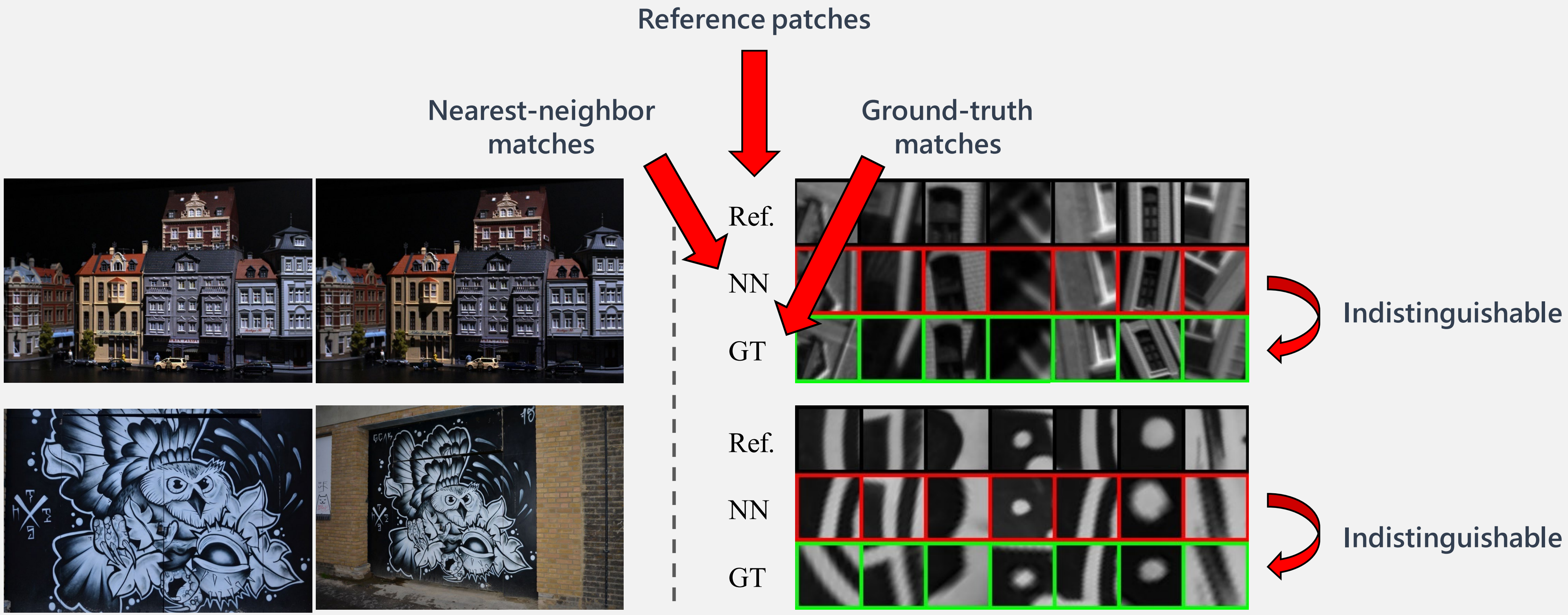
*Samples are from HPatches dataset.

# 01 Motivation




*GeoDesc is used for feature description.

# 01 Motivation




Locally distinguishable by visual appearance for,

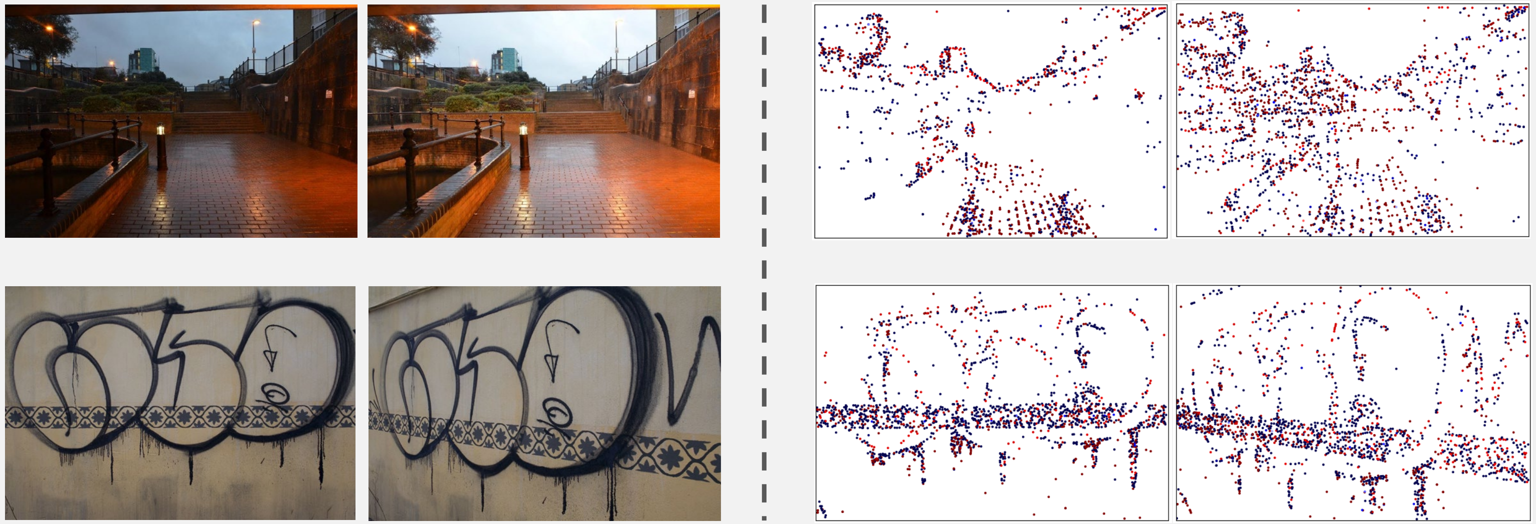
e.g., repetitive patterns?

# 01 Motivation

More *visual context!*

- The representation of both local details and richer context – *how to construct the network?*
- Construct feature pyramids – *too costly for this low-level task?*

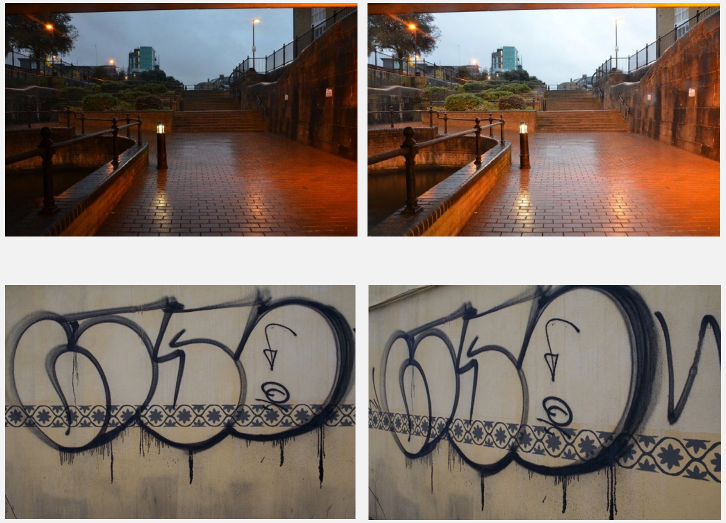
# 01 Motivation

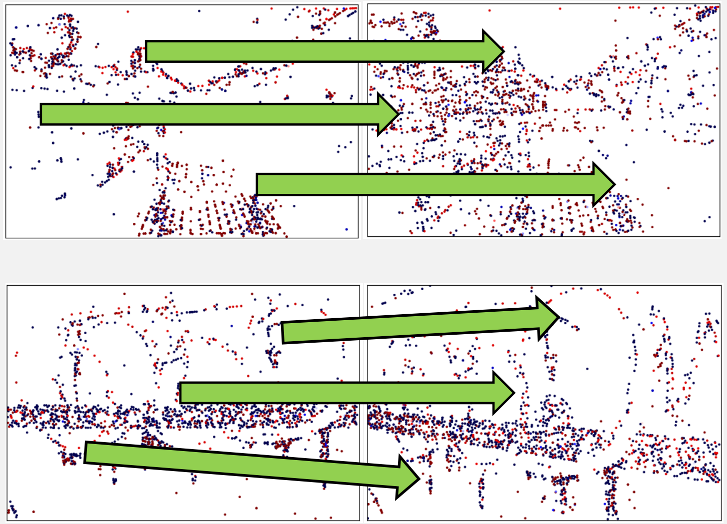


Keypoint distribution reveals meaningful scene structure

*Keypoints are derived from SIFT.

# 01 Motivation



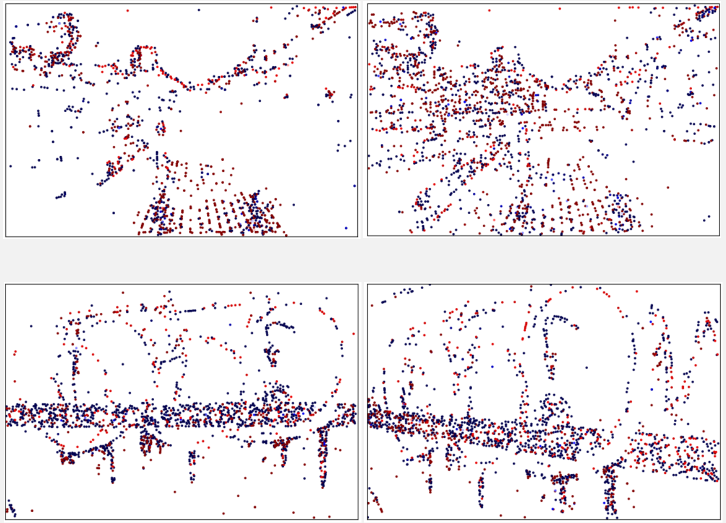


*Coarse* matches can be established, even *without* color information

**Keypoint distribution reveals meaningful scene structure**

*Keypoints are designed to be repeatable in the same underlying scene*

# 01 Motivation



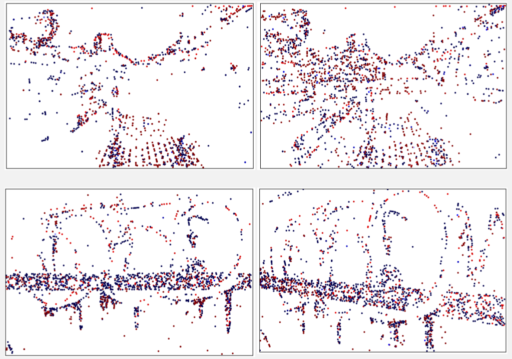
## Encoding *geometric context* from keypoint distribution of individual image

- **Keypoints are irregular and unordered** – *how to construct a proper encoder?*
- **Keypoints are not perfectly repeatable** – *how to acquire strong invariance property to different image variations?*

# 01 Motivation

## Visual context

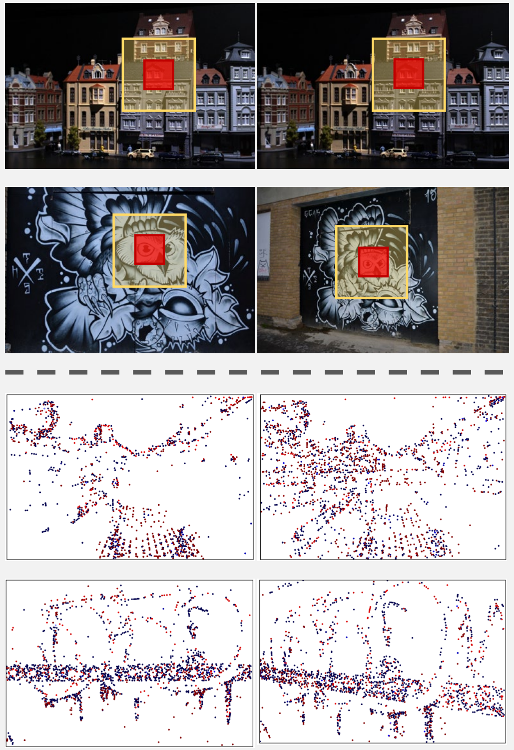
- Incorporate high-level visual information
- Resort to *regional representation* often used in image retrieval (one forward pass of the entire image)



## Geometric context

- Geometric cues from keypoint distribution.
- Resort to *PointNet-like architecture* to process 2D point sets

# 01 Motivation



*Based on off-the-shelf descriptors...*

A unified framework: Cross-modality local descriptor augmentation

# 02 Methods

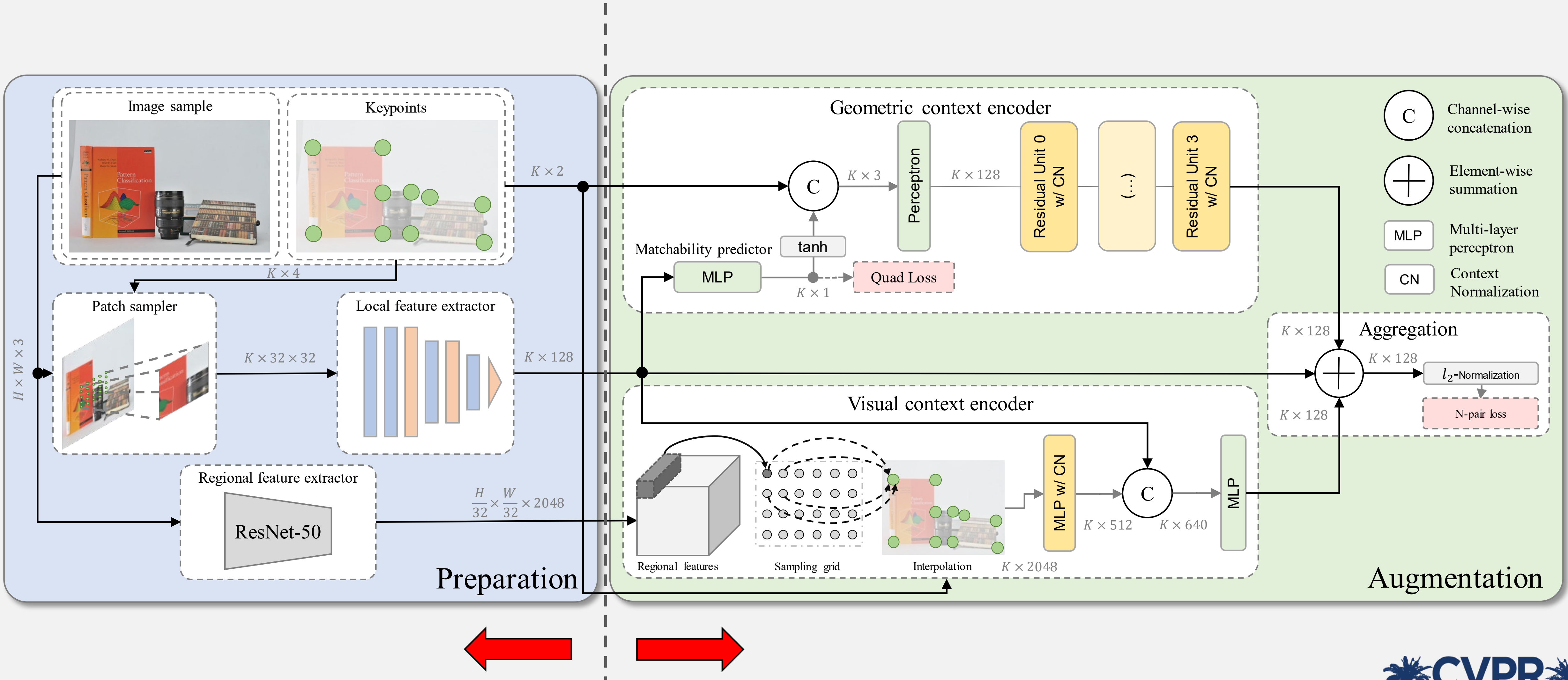
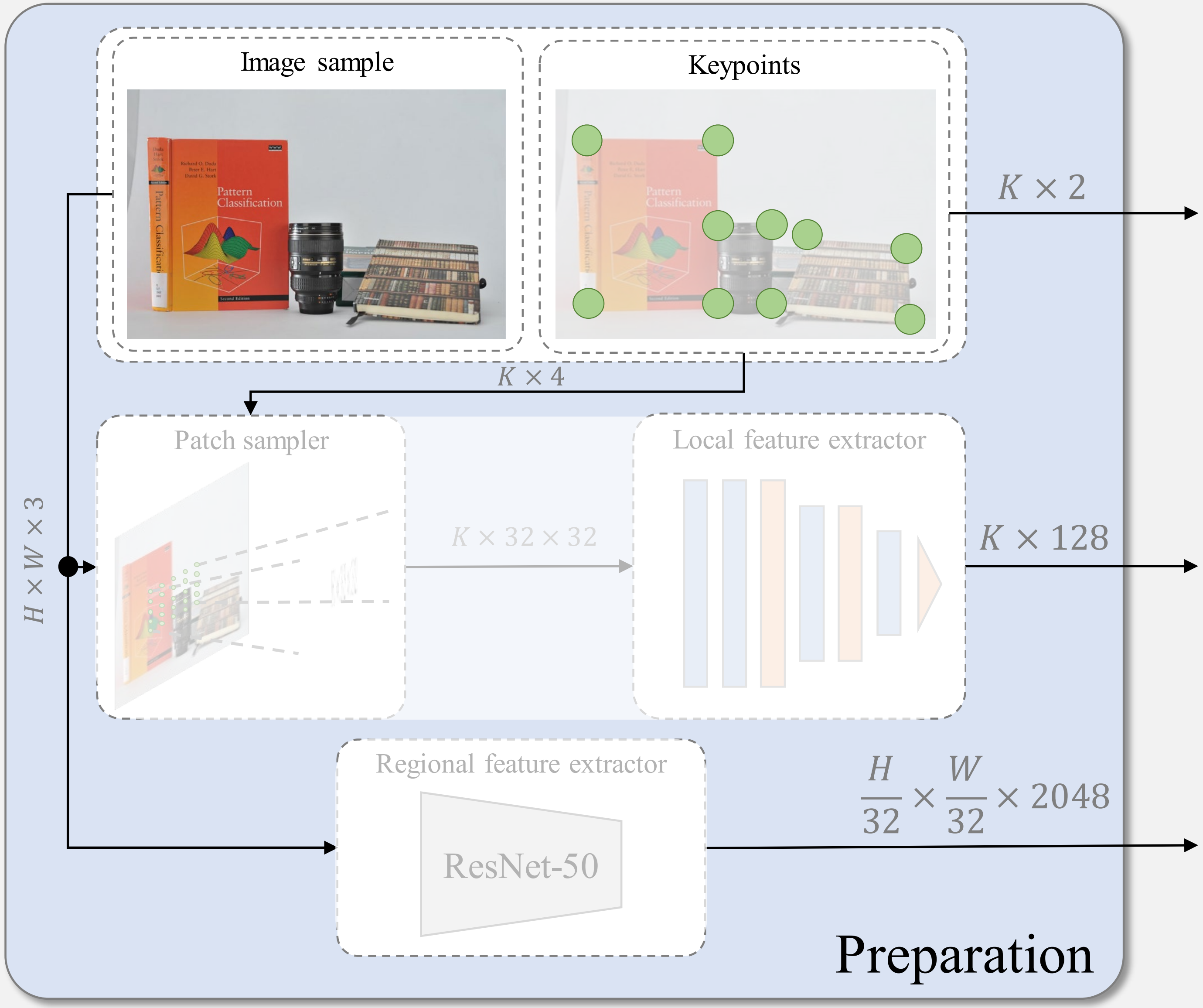


Image sample
Keypoints
K × 2
K × 4
Patch sampler
K × 32 × 32
Local feature extractor
K × 128
H × W × 3
Regional feature extractor
ResNet-50
H/32 × W/32 × 2048
Preparation
Geometric context encoder
C
K × 3
Perceptron
K × 128
Residual Unit 0 w/ CN
(...)
Residual Unit 3 w/ CN
tanh
Matchability predictor
MLP
K × 1
Quad Loss
Visual context encoder
Regional features
Sampling grid
Interpolation
K × 2048
MLP w/ CN
K × 512
K × 640
MLP
Aggregation
K × 128
l2-Normalization
N-pair loss
Channel-wise concatenation
Element-wise summation
Multi-layer perceptron
CN
Context Normalization
Augmentation

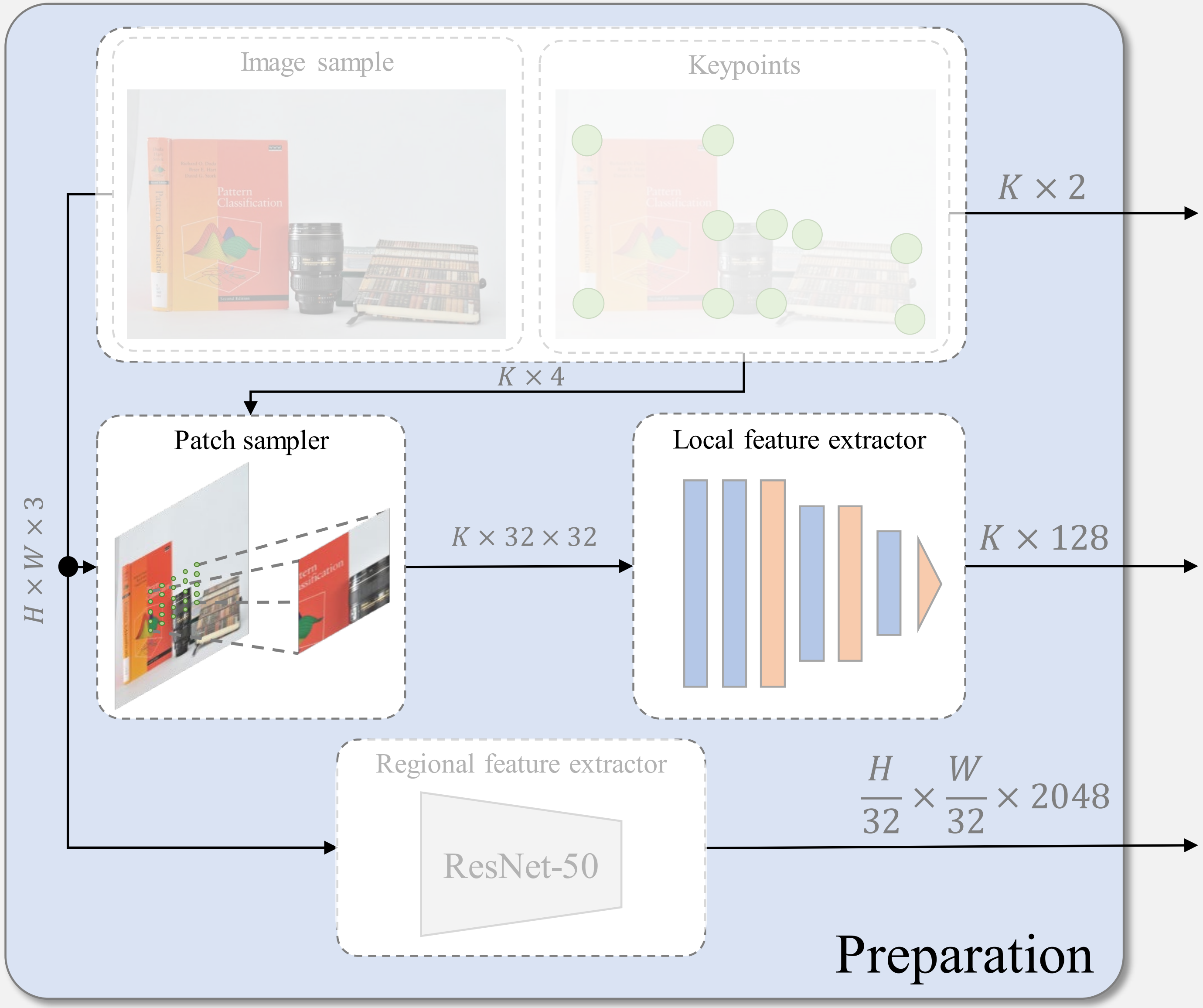
# 02 Methods




**Keypoint locations**

*E.g., SIFT*

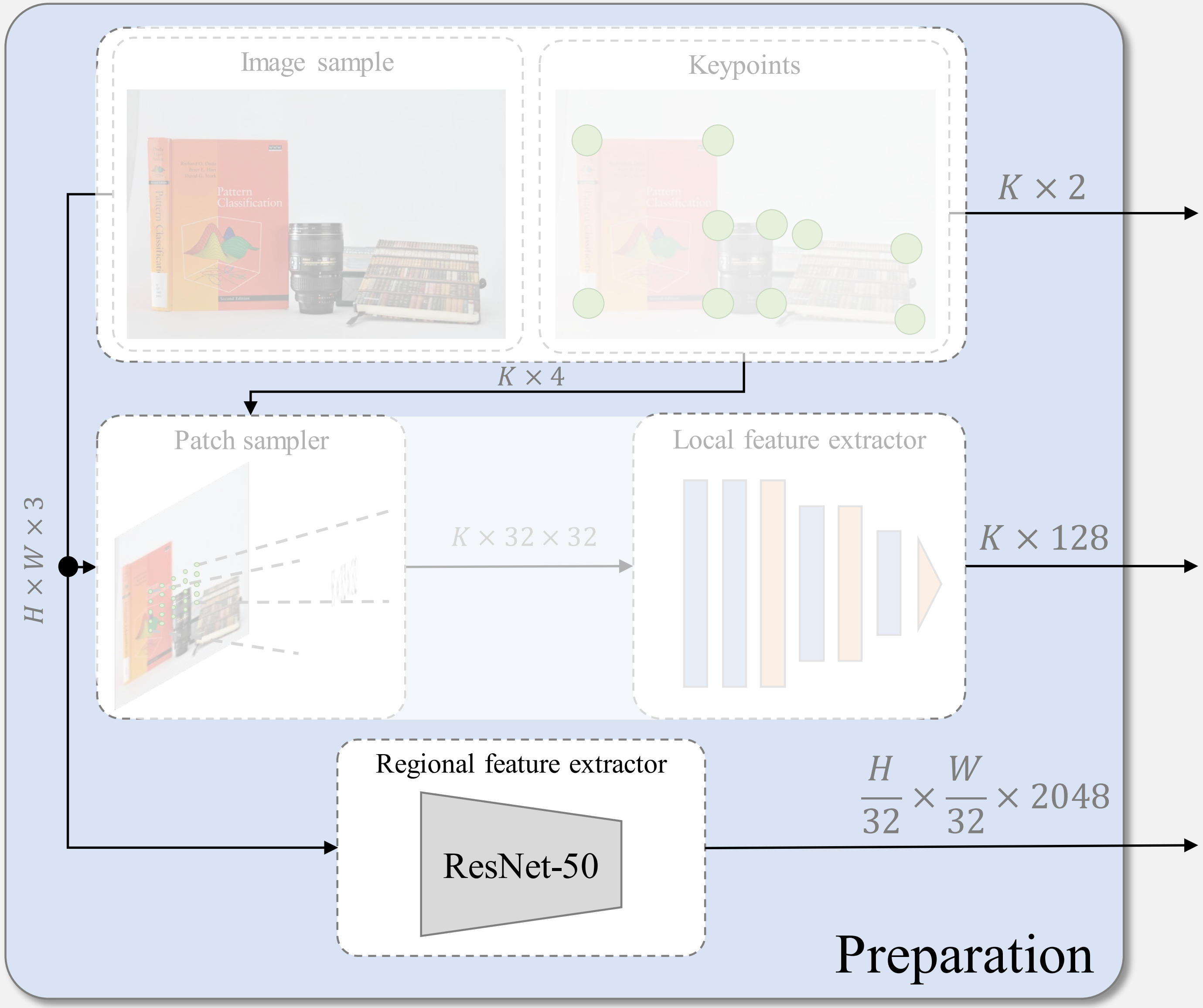
# 02 Methods




Raw local features

*E.g., GeoDesc*

# 02 Methods

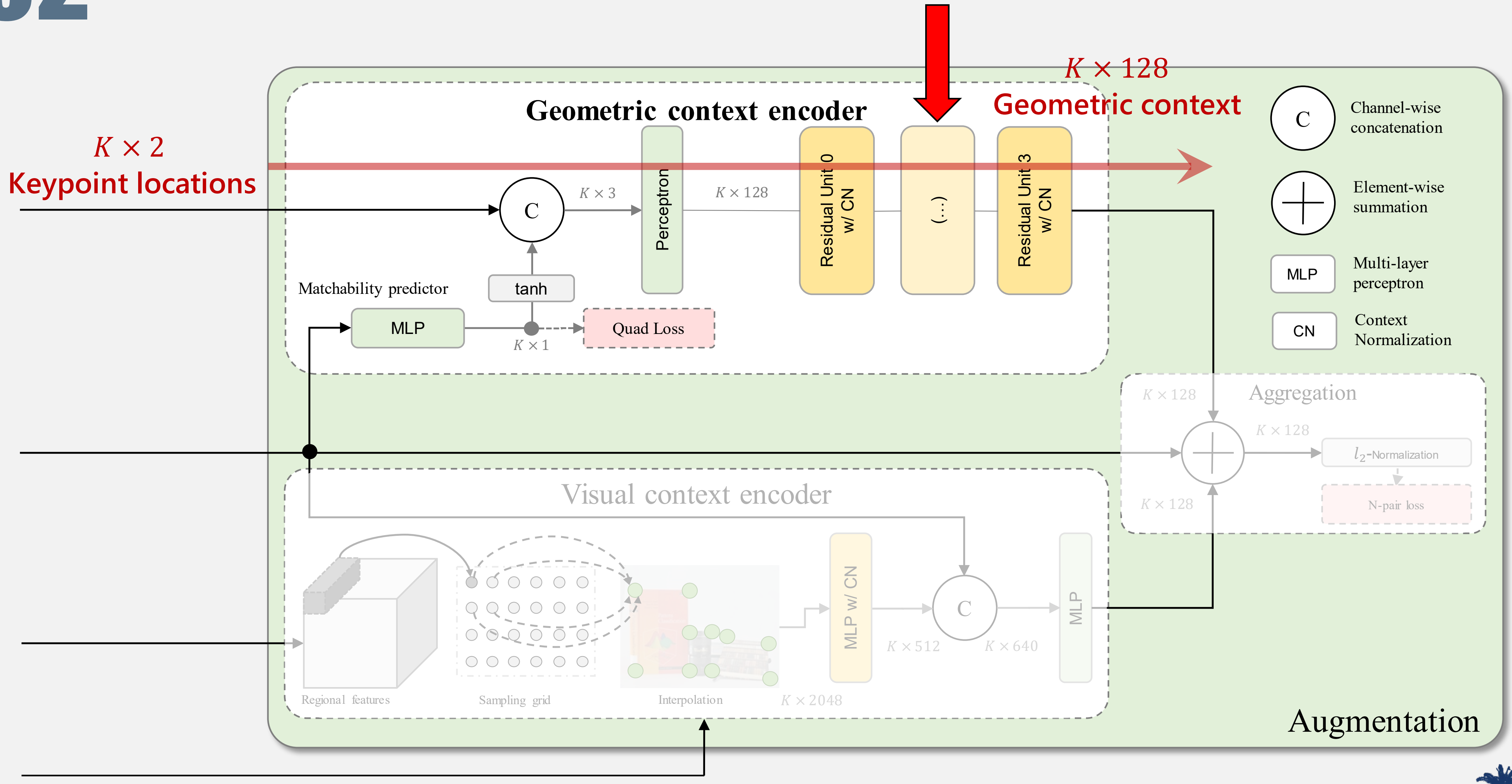



## Regional features

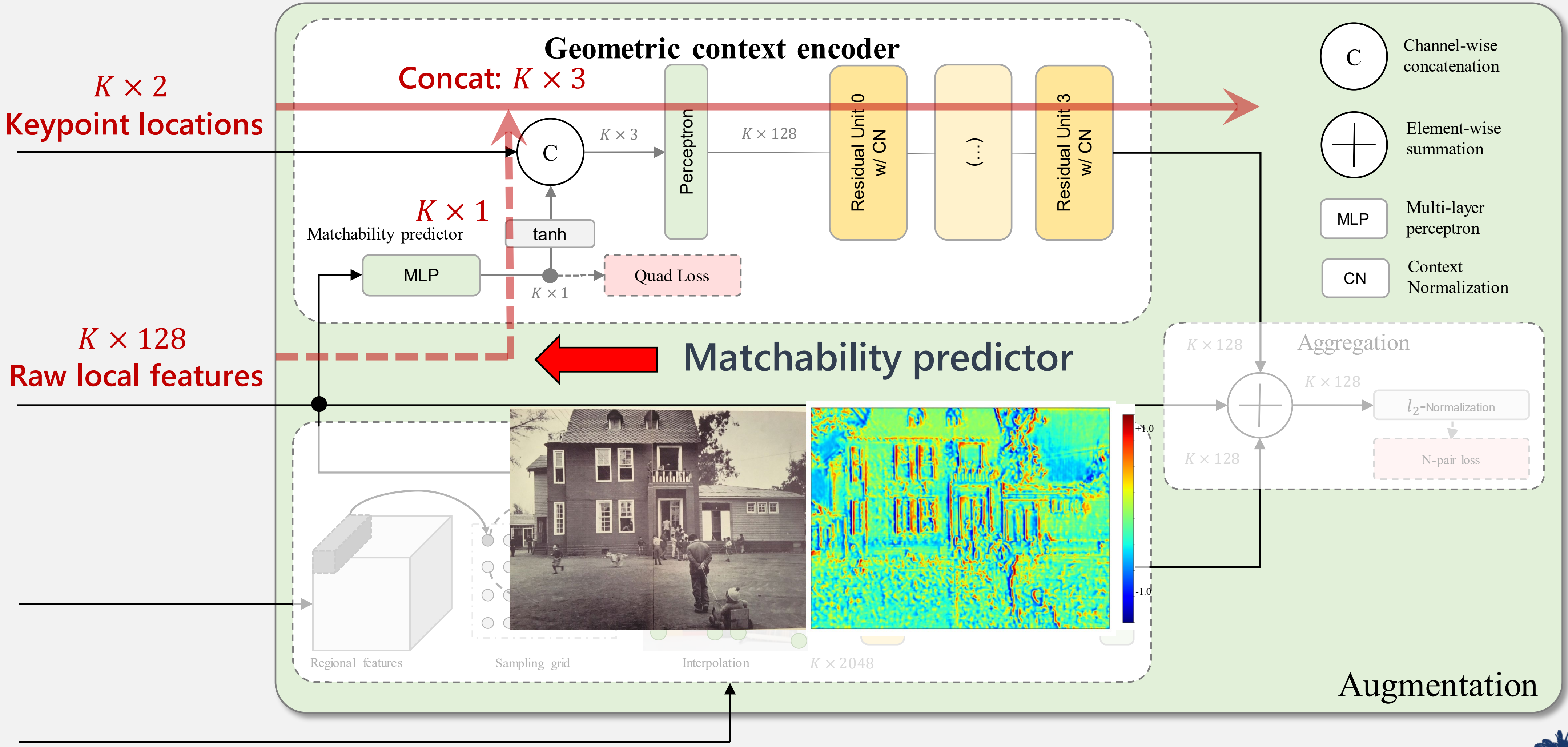
*An off-the-shelf image retrieval model*

# 02 Methods

A variant of PointNet

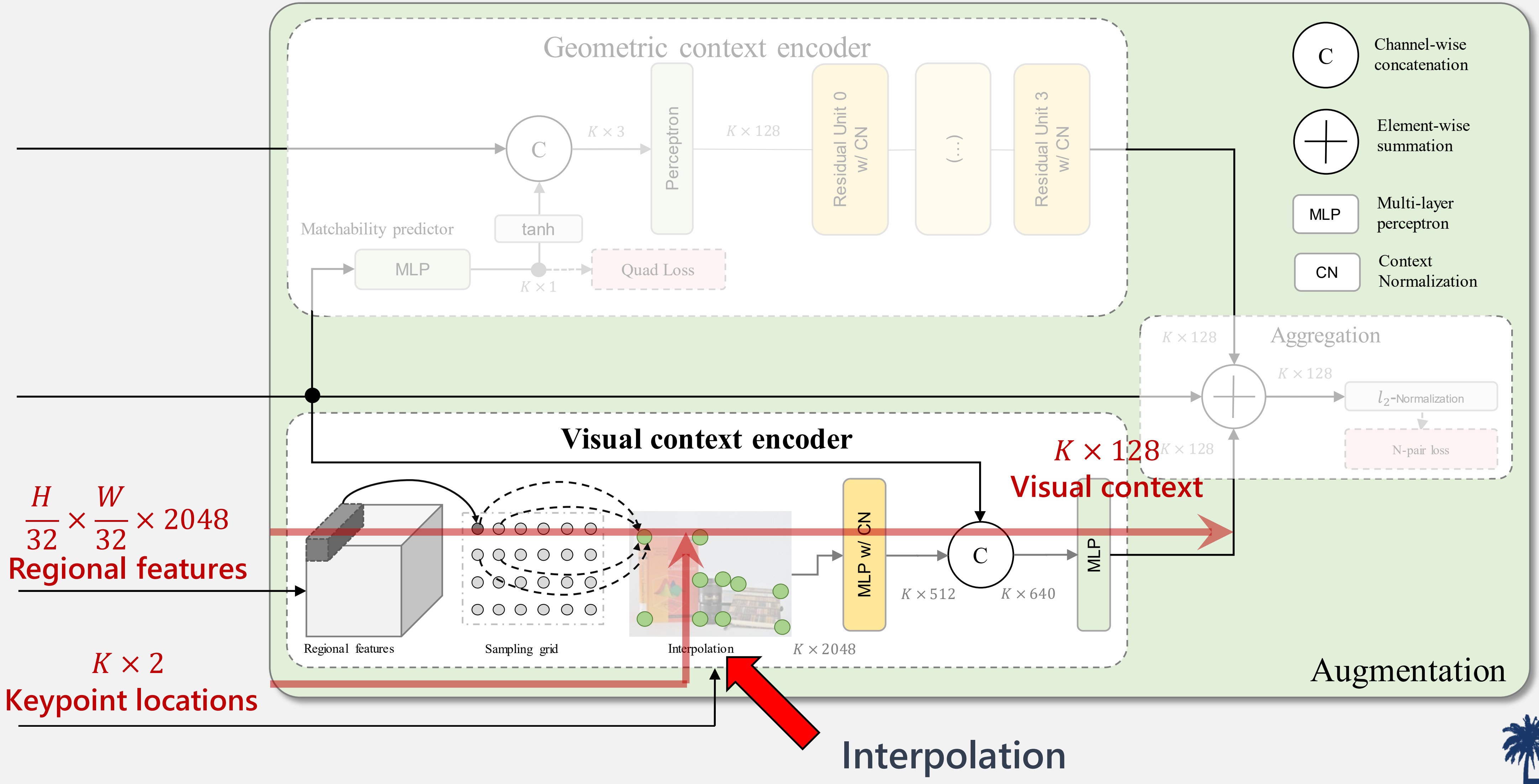


# 02 Methods



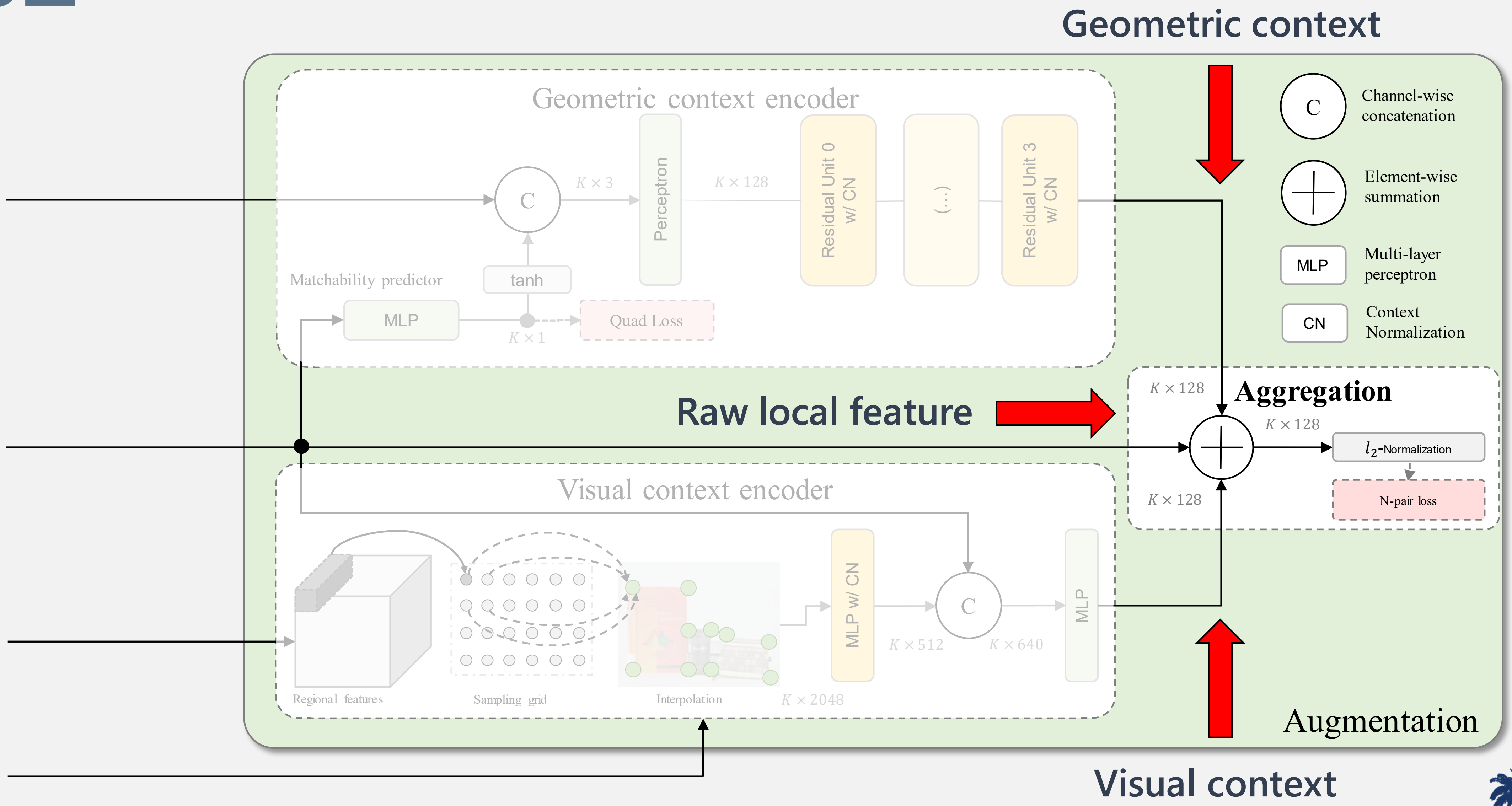
$K \times 2$
Keypoint locations
$K \times 128$
Raw local features
Geometric context encoder
Concat: $K \times 3$
C
$K \times 3$
Perceptron
$K \times 128$
Residual Unit 0 w/ CN
(...)
Residual Unit 3 w/ CN
$K \times 1$
Matchability predictor
tanh
MLP
Quad Loss
$K \times 1$
Matchability predictor
Channel-wise concatenation
Element-wise summation
MLP
Multi-layer perceptron
CN
Context Normalization
Aggregation
$K \times 128$
$K \times 128$
$l_2$-Normalization
N-pair loss
$K \times 128$
+1.0
-1.0
Regional features
Sampling grid
Interpolation
$K \times 2048$
Augmentation

# 02 Methods



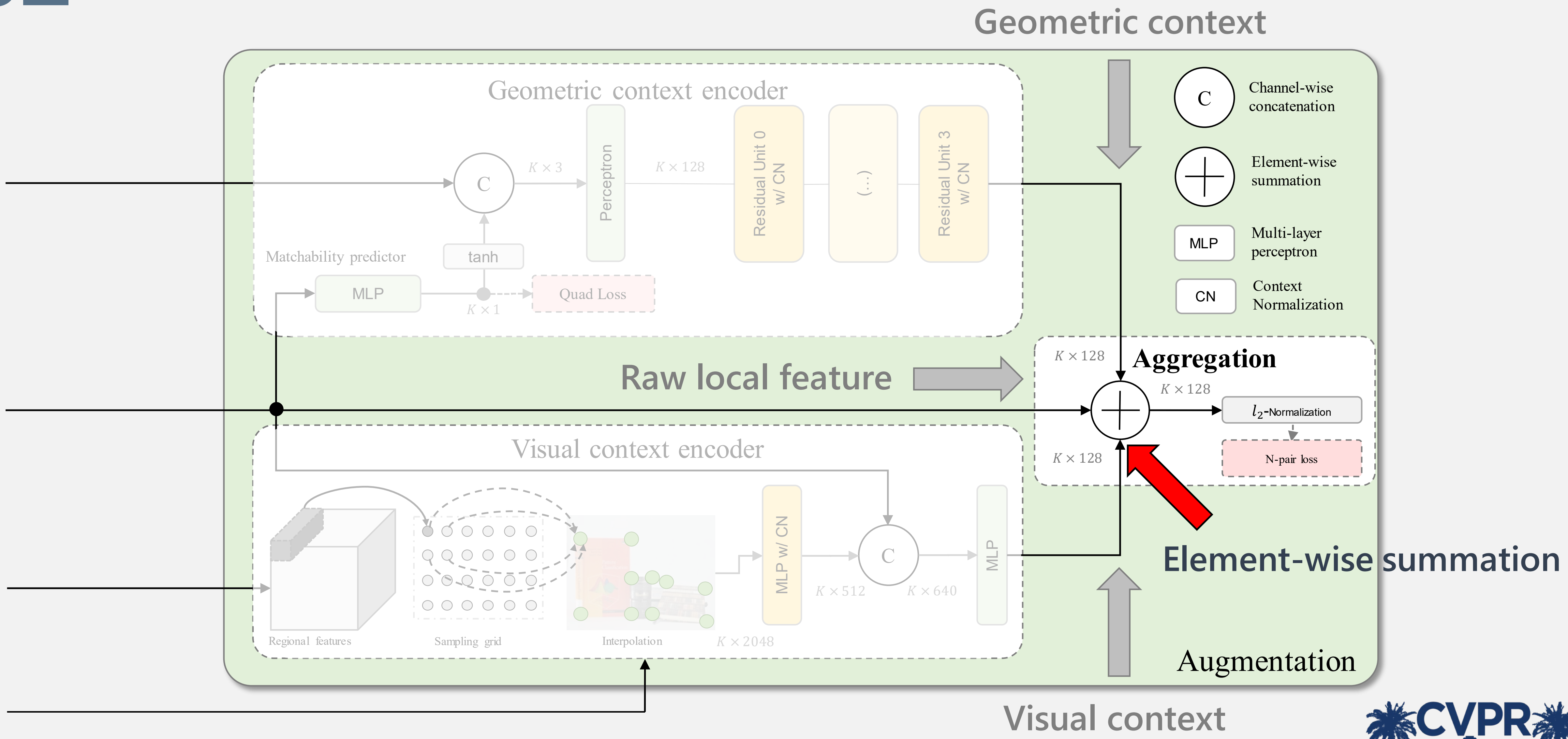
Geometric context encoder
C
$K \times 3$
Perceptron
$K \times 128$
Residual Unit 0 w/ CN
(...)
Residual Unit 3 w/ CN
Matchability predictor
tanh
MLP
$K \times 1$
Quad Loss
Channel-wise concatenation
Element-wise summation
Multi-layer perceptron
Context Normalization
Aggregation
$K \times 128$
$l_2$-Normalization
N-pair loss
Visual context encoder
$K \times 128$
Visual context
$\frac{H}{32} \times \frac{W}{32} \times 2048$
Regional features
Sampling grid
Interpolation
$K \times 2048$
MLP w/ CN
$K \times 512$
$K \times 640$
$K \times 2$
Keypoint locations
Augmentation
Interpolation

# 02 Methods



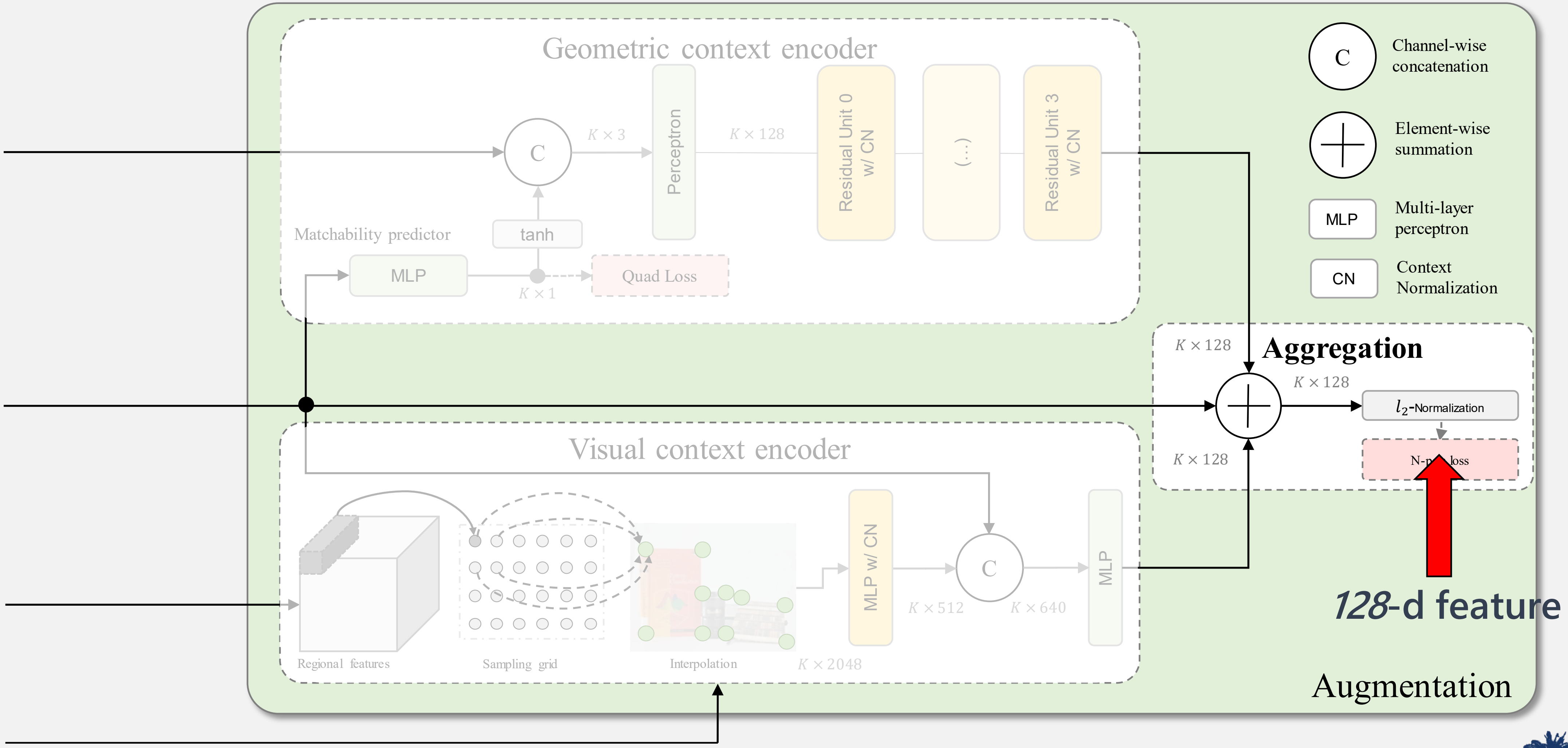
Geometric context
Geometric context encoder
C
$K \times 3$
Perceptron
$K \times 128$
Residual Unit 0 w/ CN
(...)
Residual Unit 3 w/ CN
Matchability predictor
tanh
MLP
$K \times 1$
Quad Loss
Channel-wise concatenation
Element-wise summation
MLP
Multi-layer perceptron
CN
Context Normalization
Raw local feature
Aggregation
$K \times 128$
$l_2$-Normalization
N-pair loss
Visual context encoder
Regional features
Sampling grid
Interpolation
$K \times 2048$
MLP w/ CN
$K \times 512$
$K \times 640$
Augmentation
Visual context

# 02 Methods



Geometric context
Geometric context encoder
C
$K \times 3$
Perceptron
$K \times 128$
Residual Unit 0 w/ CN
(...)
Residual Unit 3 w/ CN
Matchability predictor
tanh
MLP
$K \times 1$
Quad Loss
Channel-wise concatenation
Element-wise summation
Multi-layer perceptron
Context Normalization
CN
$K \times 128$
Aggregation
Raw local feature
$K \times 128$
$l_2$-Normalization
$K \times 128$
N-pair loss
Visual context encoder
MLP w/ CN
C
MLP
$K \times 512$
$K \times 640$
Element-wise summation
Regional features
Sampling grid
Interpolation
$K \times 2048$
Augmentation
Visual context

# 02 Methods



Geometric context encoder
C
$K \times 3$
Perceptron
$K \times 128$
Residual Unit 0 w/ CN
(...)
Residual Unit 3 w/ CN
Matchability predictor
tanh
MLP
$K \times 1$
Quad Loss
Channel-wise concatenation
Element-wise summation
MLP
Multi-layer perceptron
CN
Context Normalization
$K \times 128$
Aggregation
$K \times 128$
$l_2$-Normalization
$K \times 128$
N-pair loss
Visual context encoder
Regional features
Sampling grid
Interpolation
$K \times 2048$
MLP w/ CN
$K \times 512$
C
$K \times 640$
MLP
128-d feature
Augmentation

# 03 Ablations

Improvements from
*visual context*

| Visual context encoder | | |
|---|---|---|
| *Strategy* | *Recall i/v* | |
| baseline (GeoDesc [23]) | 59.46 | 71.24 |
| CS (256-d) [50, 19, 43] | 59.83 | 71.27 |
| w/ global feature [5] | 59.11 | 71.02 |
| w/ regional feature | 63.64 | 73.37 |
| **w/ regional feature + CN** | **63.98** | **73.63** |

| Geometric context encoder | | |
|---|---|---|
| *Network architecture* | *Recall i/v* | |
| baseline (GeoDesc [23]) | 59.46 | 71.24 |
| PointNet [31] | 59.61 | 70.96 |
| w/ CN (pre.) + xy | 61.67 | 72.63 |
| w/ CN (pre.) + xy + raw local feature | 60.91 | 72.99 |
| w/ CN (orig.) + xy + matchability | 59.94 | 71.25 |
| **w/ CN (pre.) + xy + matchability** | **62.82** | **73.40** |

| Comparison with other methods | | |
|---|---|---|
| *Method* | *Recall i/v* | |
| SIFT [22] | 47.36 | 53.06 |
| L2-Net [43] | 47.58 | 53.96 |
| HardNet [25] | 57.63 | 63.36 |
| GeoDesc [23] | 59.46 | 71.24 |
| **ContextDesc** | **66.55** | **75.52** |
| **ContextDesc+** | **67.14** | **76.42** |

# 03 Ablations

Improvements from *geometric context*

| Visual context encoder | | |
|---|---|---|
| *Strategy* | *Recall i/v* | |
| baseline (GeoDesc [23]) | 59.46 | 71.24 |
| CS (256-d) [50, 19, 43] | 59.83 | 71.27 |
| w/ global feature [5] | 59.11 | 71.02 |
| w/ regional feature | 63.64 | 73.37 |
| **w/ regional feature + CN** | **63.98** | **73.63** |

| Geometric context encoder | | |
|---|---|---|
| *Network architecture* | *Recall i/v* | |
| baseline (GeoDesc [23]) | 59.46 | 71.24 |
| PointNet [31] | 59.61 | 70.96 |
| w/ CN (pre.) + xy | 61.67 | 72.63 |
| w/ CN (pre.) + xy + raw local feature | 60.91 | 72.99 |
| w/ CN (orig.) + xy + matchability | 59.94 | 71.25 |
| **w/ CN (pre.) + xy + matchability** | **62.82** | **73.40** |

| Comparison with other methods | | |
|---|---|---|
| *Method* | *Recall i/v* | |
| SIFT [22] | 47.36 | 53.06 |
| L2-Net [43] | 47.58 | 53.96 |
| HardNet [25] | 57.63 | 63.36 |
| GeoDesc [23] | 59.46 | 71.24 |
| **ContextDesc** | **66.55** | **75.52** |
| **ContextDesc+** | **67.14** | **76.42** |

# 03 Ablations

| Visual context encoder | | |
|---|---|---|
| *Strategy* | *Recall i/v* | |
| baseline (GeoDesc [23]) | 59.46 | 71.24 |
| CS (256-d) [50, 19, 43] | 59.83 | 71.27 |
| w/ global feature [5] | 59.11 | 71.02 |
| w/ regional feature | 63.64 | 73.37 |
| **w/ regional feature + CN** | **63.98** | **73.63** |

| Geometric context encoder | | |
|---|---|---|
| *Network architecture* | *Recall i/v* | |
| baseline (GeoDesc [23]) | 59.46 | 71.24 |
| PointNet [31] | 59.61 | 70.96 |
| w/ CN (pre.) + xy | 61.67 | 72.63 |
| w/ CN (pre.) + xy + raw local feature | 60.91 | 72.99 |
| w/ CN (orig.) + xy + matchability | 59.94 | 71.25 |
| **w/ CN (pre.) + xy + matchability** | **62.82** | **73.40** |

Improvements from *cross-modality context*

| Comparison with other methods | | |
|---|---|---|
| *Method* | *Recall i/v* | |
| SIFT [22] | 47.36 | 53.06 |
| L2-Net [43] | 47.58 | 53.96 |
| HardNet [25] | 57.63 | 63.36 |
| GeoDesc [23] | 59.46 | 71.24 |
| **ContextDesc** | **66.55** | **75.52** |
| **ContextDesc+** | **67.14** | **76.42** |

# 04 Evaluations

SUN3D: indoor scenes

| | SIFT [22] | L2-Net [43] | HardNet [25] | GeoDesc [23] | Ours |
|---|---|---|---|---|---|
| | | | *median number of inlier matches* | | |
| *indoor* | 138 | 153 | 239 | 271 | **365** |
| *outdoor* | 168 | 173 | 219 | 214 | **482** |

| | SIFT [22] | GeoDesc [23] | Ours |
|---|---|---|---|
| | | *Recall* | |
| *JPEG* | 60.7 | 66.1 | **78.6** |
| *Blur* | 41.0 | 47.7 | **57.8** |
| *Exposure* | 78.2 | 86.4 | **88.2** |
| *Day-Night* | 29.2 | *39.6* | **43.3** |
| *Scale* | 81.2 | 85.8 | **88.1** |
| *Rotation* | 82.4 | **87.6** | 86.3 |
| *Scale-Rotation* | 29.6 | 33.7 | **38.0** |
| *Planar* | 48.2 | 59.1 | **61.7** |

| | | # Images | # Registered | # Sparse Points | # Observations |
|---|---|---|---|---|---|
| **Fountain** | *SIFT [22]* | 11 | 11 | 10,004 | 44K |
| | *GeoDesc [23]* | | 11 | 16,687 | 83K |
| | *Ours* | | 11 | **16,965** | **84K** |
| **Herzjesu** | *SIFT* | 8 | 8 | 4,916 | 19K |
| | *GeoDesc* | | 8 | 8,720 | 38K |
| | *Ours* | | 8 | **9,429** | **40K** |
| **South Building** | *SIFT* | 128 | 128 | 62,780 | 353K |
| | *GeoDesc* | | 128 | 170,306 | 887K |
| | *Ours* | | 128 | **174,359** | **893K** |
| **Roman Forum** | *SIFT* | 2,364 | 1,407 | 242,192 | 1,805K |
| | *GeoDesc* | | 1,566 | 770,363 | 5,051K |
| | *Ours* | | **1,571** | **848,319** | **5,484K** |
| **Alamo** | *SIFT* | 2,915 | 743 | 120,713 | 1,384K |
| | *GeoDesc* | | 893 | 353,329 | 3,159K |
| | *Ours* | | **921** | **424,348** | **3,488K** |

CVPR

# 04 Evaluations

## YFCC: outdoor scenes

| | SIFT [22] | L2-Net [43] | HardNet [[illegible]] | GeoDesc [23] | Ours |
|---|---|---|---|---|---|
| | | | *median number of inlier matches* | | |
| *indoor* | 138 | 153 | 239 | 271 | **365** |
| *outdoor* | 168 | 173 | 219 | 214 | **482** |

| | SIFT [22] | GeoDesc [23] | Ours |
|---|---|---|---|
| | | *Recall* | |
| *JPEG* | 60.7 | 66.1 | **78.6** |
| *Blur* | 41.0 | 47.7 | **57.8** |
| *Exposure* | 78.2 | 86.4 | **88.2** |
| *Day-Night* | 29.2 | *39.6* | **43.3** |
| *Scale* | 81.2 | 85.8 | **88.1** |
| *Rotation* | 82.4 | **87.6** | 86.3 |
| *Scale-Rotation* | 29.6 | 33.7 | **38.0** |
| *Planar* | 48.2 | 59.1 | **61.7** |

| | | # Images | # Registered | # Sparse Points | # Observations |
|---|---|---|---|---|---|
| **Fountain** | *SIFT [22]* | 11 | 11 | 10,004 | 44K |
| | *GeoDesc [23]* | | 11 | 16,687 | 83K |
| | *Ours* | | 11 | **16,965** | **84K** |
| **Herzjesu** | *SIFT* | 8 | 8 | 4,916 | 19K |
| | *GeoDesc* | | 8 | 8,720 | 38K |
| | *Ours* | | 8 | **9,429** | **40K** |
| **South Building** | *SIFT* | 128 | 128 | 62,780 | 353K |
| | *GeoDesc* | | 128 | 170,306 | 887K |
| | *Ours* | | 128 | **174,359** | **893K** |
| **Roman Forum** | *SIFT* | 2,364 | 1,407 | 242,192 | 1,805K |
| | *GeoDesc* | | 1,566 | 770,363 | 5,051K |
| | *Ours* | | **1,571** | **848,319** | **5,484K** |
| **Alamo** | *SIFT* | 2,915 | 743 | 120,713 | 1,384K |
| | *GeoDesc* | | 893 | 353,329 | 3,159K |
| | *Ours* | | **921** | **424,348** | **3,488K** |

# 04 Evaluations

| | SIFT [22] | L2-Net [43] | HardNet [25] | GeoDesc [23] | Ours |
|---|---|---|---|---|---|
| *median number of inlier matches* | | | | | |
| *indoor* | 138 | 153 | 239 | 271 | **365** |
| *outdoor* | 168 | 173 | 219 | 214 | **482** |

| | SIFT [22] | GeoDesc [23] | Ours |
|---|---|---|---|
| *Recall* | | | |
| *JPEG* | 60.7 | 66.1 | **78.6** |
| *Blur* | 41.0 | 47.7 | **57.8** |
| *Exposure* | 78.2 | 86.4 | **88.2** |
| *Day-Night* | 29.2 | *39.6* | **43.3** |
| *Scale* | 81.2 | 85.8 | **88.1** |
| *Rotation* | 82.4 | **87.6** | 86.3 |
| *Scale-Rotation* | 29.6 | 33.7 | **38.0** |
| *Planar* | 48.2 | 59.1 | **61.7** |

Oxford dataset: Different image variations

| | | # Images | # Registered | # Sparse Points | # Observations |
|---|---|---|---|---|---|
| **Fountain** | *SIFT [22]* | 11 | 11 | 10,004 | 44K |
| | *GeoDesc [23]* | | 11 | 16,687 | 83K |
| | *Ours* | | 11 | **16,965** | **84K** |
| **Herzjesu** | *SIFT* | 8 | 8 | 4,916 | 19K |
| | *GeoDesc* | | 8 | 8,720 | 38K |
| | *Ours* | | 8 | **9,429** | **40K** |
| **South Building** | *SIFT* | 128 | 128 | 62,780 | 353K |
| | *GeoDesc* | | 128 | 170,306 | 887K |
| | *Ours* | | 128 | **174,359** | **893K** |
| **Roman Forum** | *SIFT* | 2,364 | 1,407 | 242,192 | 1,805K |
| | *GeoDesc* | | 1,566 | 770,363 | 5,051K |
| | *Ours* | | **1,571** | **848,319** | **5,484K** |
| **Alamo** | *SIFT* | 2,915 | 743 | 120,713 | 1,384K |
| | *GeoDesc* | | 893 | 353,329 | 3,159K |
| | *Ours* | | **921** | **424,348** | **3,488K** |

# 04 Evaluations

| | SIFT [22] | L2-Net [43] | HardNet [25] | GeoDesc [23] | Ours |
|---|---|---|---|---|---|
| | *median number of inlier matches* | | | | |
| *indoor* | 138 | 153 | 239 | 271 | **365** |
| *outdoor* | 168 | 173 | 219 | 214 | **482** |

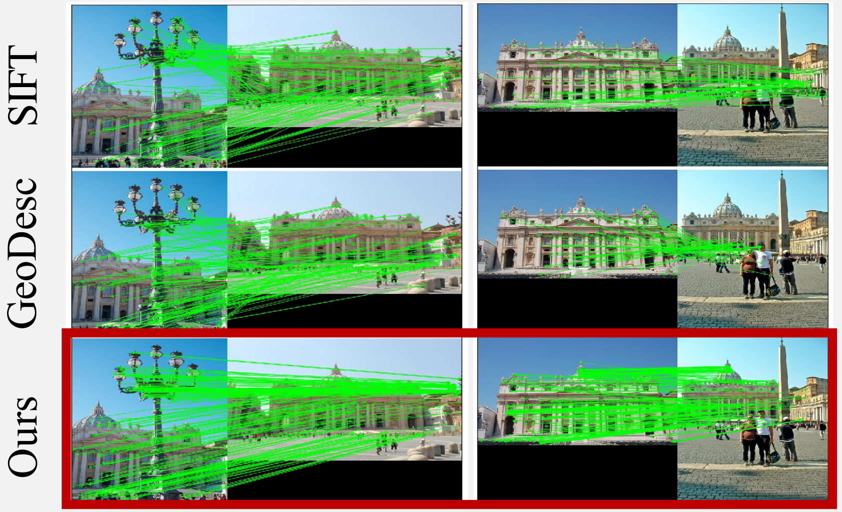
| | SIFT [22] | GeoDesc [23] | Ours |
|---|---|---|---|
| | *Recall* | | |
| *JPEG* | 60.7 | 66.1 | **78.6** |
| *Blur* | 41.0 | 47.7 | **57.8** |
| *Exposure* | 78.2 | 86.4 | **88.2** |
| *Day-Night* | 29.2 | *39.6* | **43.3** |
| *Scale* | 81.2 | 85.8 | **88.1** |
| *Rotation* | 82.4 | **87.6** | 86.3 |
| *Scale-Rotation* | 29.6 | 33.7 | **38.0** |
| *Planar* | 48.2 | 59.1 | **61.7** |

| | | # Images | # Registered | # Sparse Points | # Observations |
|---|---|---|---|---|---|
| **Fountain** | *SIFT [22]* | 11 | 11 | 10,004 | 44K |
| | *GeoDesc [23]* | | 11 | 16,687 | 83K |
| | *Ours* | | 11 | **16,965** | **84K** |
| **Herzjesu** | *SIFT* | 8 | 8 | 4,916 | 19K |
| | *GeoDesc* | | 8 | 8,720 | 38K |
| | *Ours* | | 8 | **9,429** | **40K** |
| **South Building** | *SIFT* | 128 | 128 | 62,780 | 353K |
| | *GeoDesc* | | 128 | 170,306 | 887K |
| | *Ours* | | 128 | **174,359** | **893K** |
| **Roman Forum** | *SIFT* | 2,364 | 1,407 | 242,192 | 1,805K |
| | *GeoDesc* | | 1,566 | 770,363 | 5,051K |
| | *Ours* | | **1,571** | **848,319** | **5,484K** |
| **Alamo** | *SIFT* | 2,915 | 743 | 120,713 | 1,384K |
| | *GeoDesc* | | 893 | 353,329 | 3,159K |
| | *Ours* | | **921** | **424,348** | **3,488K** |

3D reconstruction benchmark

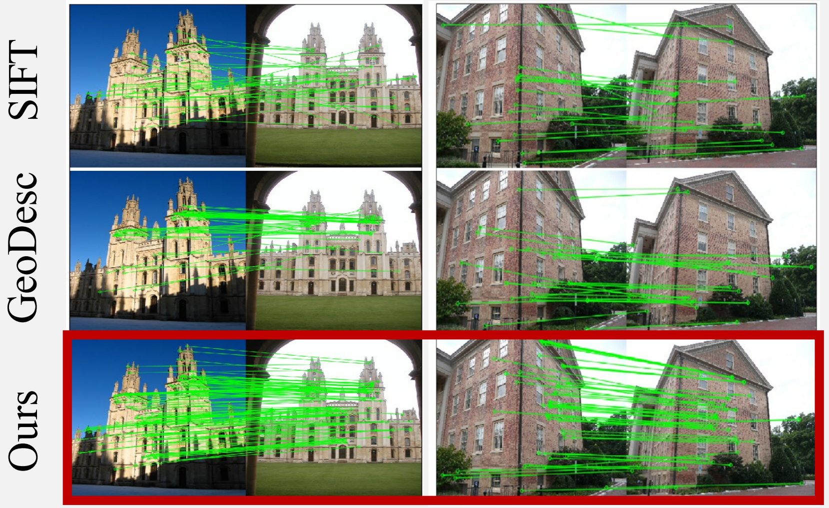
CVPR

# 04 Evaluations

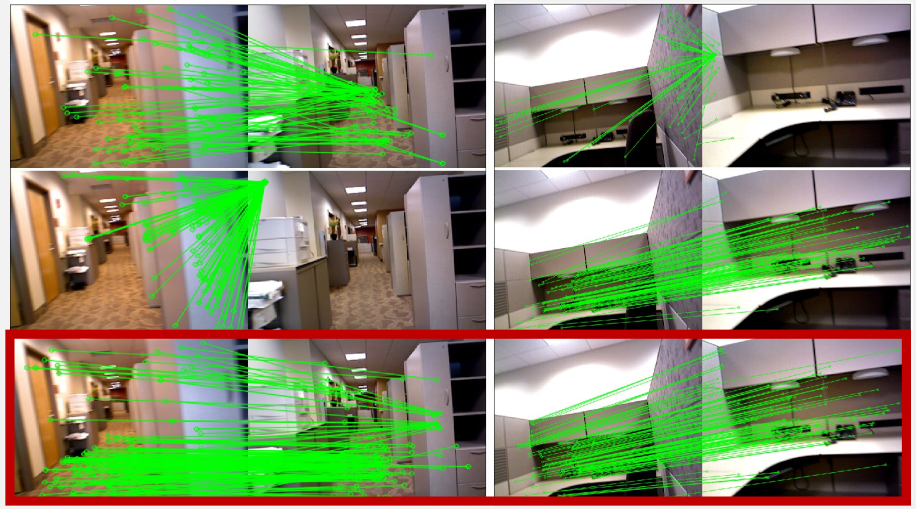



Scale or rotation change

Illumination change




Perspective change



Indoor scene

# 04 Evaluations

When pose accuracy as evaluation metric: consistent improvement, but less significant

| Method | Date | Type | Ims (%) | #Pts | SR | TL | $mAP^{5^\circ}$ | $mAP^{10^\circ}$ | $mAP^{15^\circ}$ | $mAP^{20^\circ}$ | $mAP^{25^\circ}$ | ATE |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| SIFT + ContextDesc kp:8000, match:nn | 19-05-09 | F | 97.9 | 6020.4 | 97.2 | 3.26 | 0.3828 | 0.4821 | 0.5399 | 0.5853 | 0.6226 | — |
| SIFT + GeoDesc kp:8000, match:nn | 19-04-24 | F | 97.3 | 5583.8 | 95.8 | 3.39 | 0.3858 | 0.4778 | 0.5317 | 0.5790 | 0.6139 | — |

After obtaining sufficient matches, what is the next bottleneck in order to improve the image matching?

# 05 Sparse matching

1) Establish putative matches (nearest-neighbor search/FLANN)
2) Outlier rejection (ratio test/mutual check/GMS)
3) Geometry computation (5-point/8-point algorithm with RANSAC)
4) Non-linear optimization for refinement

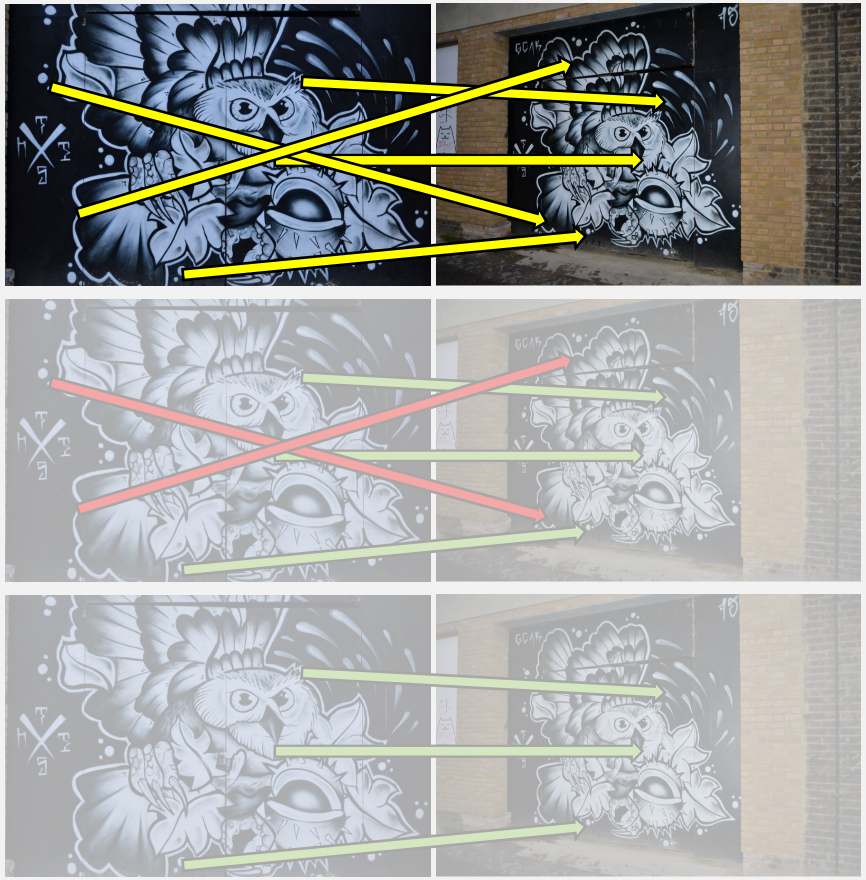
# 05 Sparse matching

1) Establish putative matches (nearest-neighbor search/FLANN)
2) Outlier rejection (ratio test/mutual check/GMS)
3) Geometry computation (5-point/8-point algorithm with RANSAC)
4) Non-linear optimization for refinement

Learning-based

*Yi et al.: Learning to find good correspondences, CVPR'18.

# 05 Sparse matching

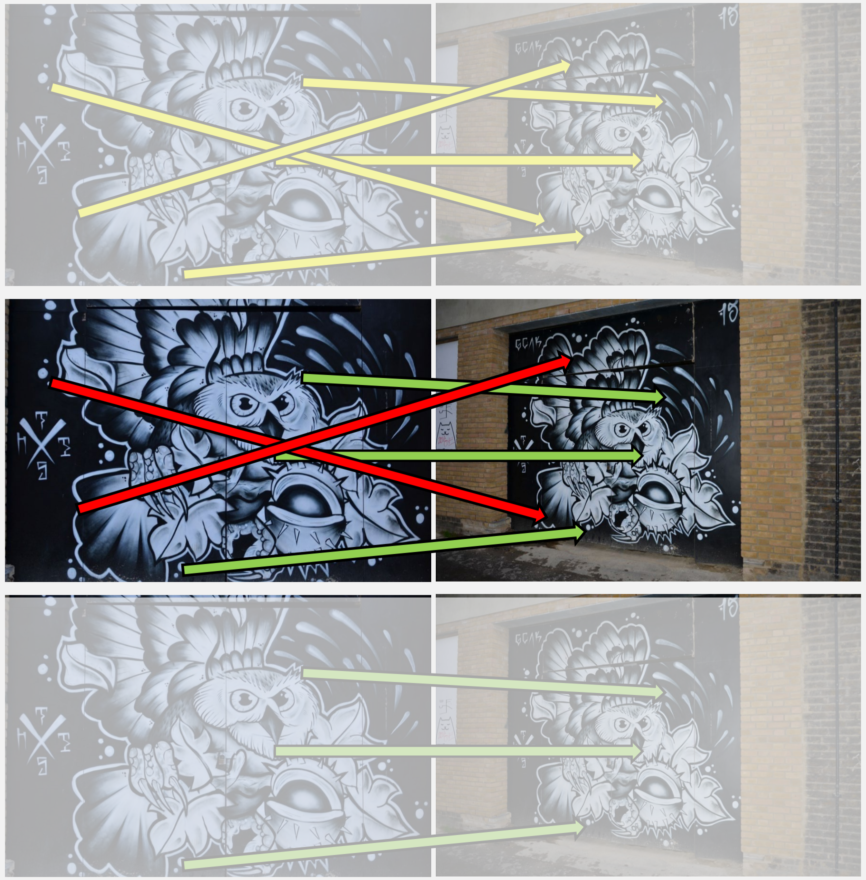


Given putative matches $N \times 4$, where each row vector denotes a correspondence $(x, y, x', y')$ of an image pair.

The network predicts the probability vector $N \times 1$ that indicates whether a correspondence is an inlier.

Only inlier matches (and its confidence) are used for computing the geometry.

# 05 Sparse matching

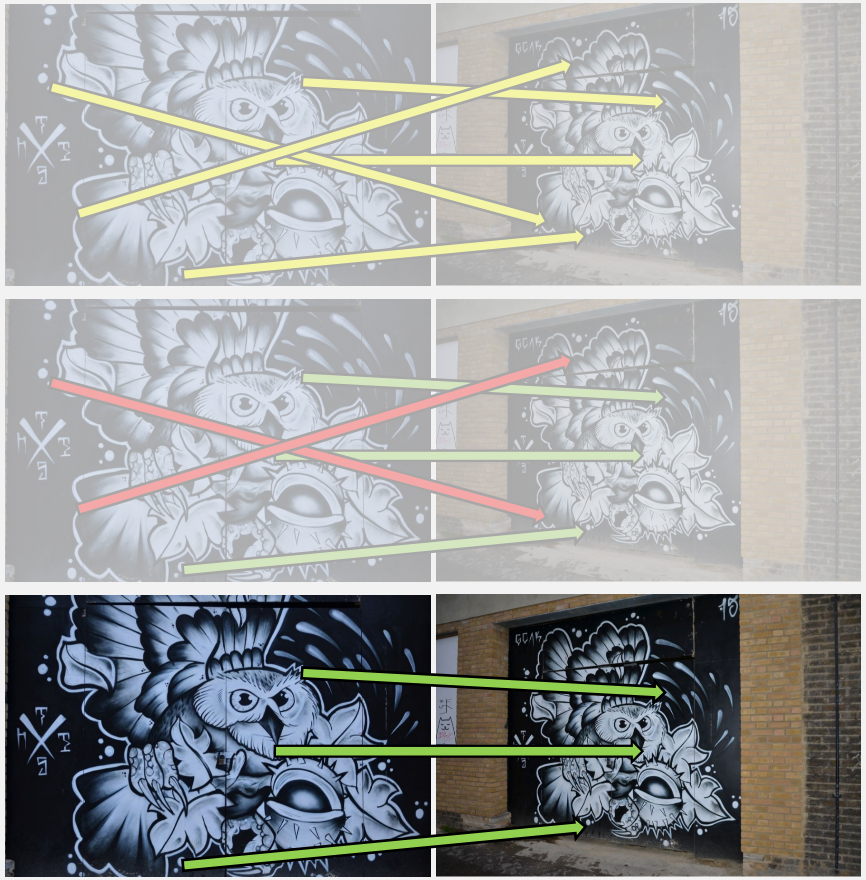


Given putative matches $N \times 4$, where each row vector denotes a correspondence $(x, y, x', y')$ of an image pair.

The network predicts the probability vector $N \times 1$ that indicates whether a correspondence is an inlier.

Only inlier matches (and its confidence) are used for computing the geometry.

# 05 Sparse matching



Given putative matches $N \times 4$, where each row vector denotes a correspondence $(x, y, x', y')$ of an image pair.

The network predicts the probability vector $N \times 1$ that indicates whether a correspondence is an inlier.

Only inlier matches (and their confidence) are used for solving the two-view geometry.

# 05 Sparse matching

## Why is it important?

| Method | Date | Type | Ims (%) | #Pts | SR | TL | mAP$^{5^o}$ | mAP$^{10^o}$ | mAP$^{15^o}$ | mAP$^{20^o}$ | mAP$^{25^o}$ | ATE |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **SIFT + ContextDesc + Inlier Classification V2** kp:8000, match:custom | 19-[illegible]-28 | [illegible] | [illegible] | 6126.0 | 97.5 | 3.44 | 0.5755 | 0.6830 | 0.7389 | 0.7750 | 0.8006 | — |
| **SIFT + ContextDesc** kp:8000, match:nn1to1 | [illegible]-06-07 | F | 98.1 | 6472.1 | 98.0 | 3.34 | 0.4287 | 0.5371 | 0.6017 | 0.6464 | 0.6826 | — |
| **SIFT + ContextDesc** kp:8000, match:nn | 19-05-09 | F | 97.9 | 6020.4 | 97.2 | 3.26 | 0.3828 | 0.4821 | 0.5399 | 0.5853 | 0.6226 | — |

Proposed learning-based

Mutual check

No outlier rejection

# 05 Sparse matching

### Previous method

- Adopt a PointNet-like architecture.
- Apply context normalization (instance normalization) on the entire point set to capture global context.

Local context, e.g., piece-wise smoothness (GMS matcher).

*Yi et al.: Learning to find good correspondences, CVPR'18.

# 05 Sparse matching

### Previous method

- Adopt a PointNet-like architecture.
- Apply context normalization (instance normalization) on the entire point set to capture global context.

### Local context, e.g., piece-wise smoothness (GMS matcher).

*Bian et al.: GMS: Grid-based Motion Statistics for Fast, Ultra-robust Feature Correspondence, CVPR'17.

# 05 Sparse matching

### Previous method

- Adopt a PointNet-like architecture.
- Apply context normalization (instance normalization) on the entire point set to capture global context.

### Proposed

- Learn to establish neighboring relations on unordered, non-Euclidean correspondence sets.
- Build a hierarchical architecture to capture both global and local context.

# 05 Sparse matching

| Method | Date | Type | Ims (%) | #Pts | SR | TL | mAP$^{5^o}$ | mAP$^{10^o}$ | mAP$^{15^o}$ | mAP$^{20^o}$ | mAP$^{25^o}$ | ATE |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **SIFT + ContextDesc + Inlier Classification V2** kp:8000, match:custom | 19-05-28 | F/M | 98.6 | 6126.0 | 97.5 | 3.44 | 0.5755 | 0.6830 | 0.7389 | 0.7750 | 0.8006 | — |
| **SIFT + ContextDesc + Inlier Classification V1** kp:8000, match:custom | 19-05-29 | F/M | 98.4 | 6045.8 | 97.8 | 3.43 | 0.5553 | 0.6633 | 0.7169 | 0.7545 | 0.7849 | — |

Proposed

Previous method

# 06 Future work

## Evaluation metric

- HPatches: patch verification/matching/retrieval - *Reflect the performance in real applications? [1]*
- Two-view image matching on pose recovery accuracy - *Involve other variables such as RANSAC-based algorithms? [2]*
- 3D reconstruction metrics – *Involve more variables such as image retrieval, SfM or bundle adjustment? [3]*
- The ground truth is often obtained from SfM with a traditional matching pipeline.

[1] Balntas et al.: HPatches: A benchmark and evaluation of handcrafted and learned local descriptors, CVPR'17
[2] Yi et al.: Learning to find good correspondences, CVPR'18.
[3] Schönberger et al.: Comparative Evaluation of Hand-Crafted and Learned Local Features, CVPR'17.

## Keypoint detection

- Challenging for learning-based methods – clear definition as supervision?
- Pixel-wise, even sub-pixel wise accuracy – preserve low-level details after multiple convolutions?
- Combine the advantage of both human priors and learned priors.

# 06 Future work

## Evaluation metric

- HPatches: patch verification/matching/retrieval - *Reflect the performance in real applications? [1]*
- Two-view image matching on pose recovery accuracy - *Involve other variables such as RANSAC-based algorithms? [2]*
- 3D reconstruction metrics – *Involve more variables such as image retrieval, SfM or bundle adjustment? [3]*
- The ground truth is often obtained from SfM with a traditional matching pipeline.

[1] Balntas et al.: HPatches: A benchmark and evaluation of handcrafted and learned local descriptors, CVPR'17
[2] Yi et al.: Learning to find good correspondences, CVPR'18.
[3] Schönberger et al.: Comparative Evaluation of Hand-Crafted and Learned Local Features, CVPR'17.

## Keypoint detection

- Challenging for learning-based methods – clear definition as supervision?
- Pixel-wise, even sub-pixel wise accuracy – preserve low-level details after multiple convolutions?
- Combine the advantage of both human priors and learned priors.

# 06 Future work

## Evaluation metric

- HPatches: patch verification/matching/retrieval - *Reflect the performance in real applications? [1]*
- Two-view image matching on pose recovery accuracy - *Involve other variables such as RANSAC-based algorithms? [2]*
- 3D reconstruction metrics – *Involve more variables such as image retrieval, SfM or bundle adjustment? [3]*
- The ground truth is often obtained from SfM with a traditional matching pipeline.

[1] Balntas et al.: HPatches: A benchmark and evaluation of handcrafted and learned local descriptors, CVPR'17
[2] Yi et al.: Learning to find good correspondences, CVPR'18.
[3] Schönberger et al.: Comparative Evaluation of Hand-Crafted and Learned Local Features, CVPR'17.

## Keypoint detection

- Challenging for learning-based methods – clear definition as supervision?
- Pixel-wise, even sub-pixel wise accuracy – preserve low-level details after multiple convolutions?
- Combine the advantage of both human priors and learned priors.

CVPR

# 06 Future work

## Evaluation metric

- HPatches: patch verification/matching/retrieval - *Reflect the performance in real applications? [1]*
- Two-view image matching on pose recovery accuracy - *Involve other variables such as RANSAC-based algorithms? [2]*
- 3D reconstruction metrics – *Involve more variables such as image retrieval, SfM or bundle adjustment? [3]*
- The ground truth is often obtained from SfM with a traditional matching pipeline.

[1] Balntas et al.: HPatches: A benchmark and evaluation of handcrafted and learned local descriptors, CVPR'17
[2] Yi et al.: Learning to find good correspondences, CVPR'18.
[3] Schönberger et al.: Comparative Evaluation of Hand-Crafted and Learned Local Features, CVPR'17.

## Keypoint detection

- Challenging for learning-based methods – clear definition as supervision?
- Pixel-wise, even sub-pixel wise accuracy – preserve low-level details after multiple convolutions?
- Combine the advantage of both human priors and learned priors.

# 06 Future work

## Evaluation metric

- HPatches: patch verification/matching/retrieval - *Reflect the performance in real applications? [1]*
- Two-view image matching on pose recovery accuracy - *Involve other variables such as RANSAC-based algorithms? [2]*
- 3D reconstruction metrics – *Involve more variables such as image retrieval, SfM or bundle adjustment? [3]*
- The ground truth is often obtained from SfM with a traditional matching pipeline.

[1] Balntas et al.: HPatches: A benchmark and evaluation of handcrafted and learned local descriptors, CVPR'17
[2] Yi et al.: Learning to find good correspondences, CVPR'18.
[3] Schönberger et al.: Comparative Evaluation of Hand-Crafted and Learned Local Features, CVPR'17.

## Keypoint detection

- Challenging for learning-based methods – clear definition as supervision?
- Pixel-wise, even sub-pixel wise accuracy – preserve low-level details after multiple convolutions?
- Combine the advantage of both human priors and learned priors.

# Thanks!

Code available at: https://github.com/lzx551402/contextdesc

LONG BEACH
CALIFORNIA
June 16-20, 2019