

ScaleNet: A Shallow Architecture for Scale Estimation

Axel Barroso-Laguna

Yurun Tian

Krystian Mikolajczyk

Imperial College London

{axel.barroso17, yurun.tian, k.mikolajczyk}@imperial.ac.uk

Abstract

In this paper, we address the problem of estimating scale factors between images. We formulate the scale estimation problem as a prediction of a probability distribution over scale factors. We design a new architecture, ScaleNet, that exploits dilated convolutions as well as self- and cross-correlation layers to predict the scale between images. We demonstrate that rectifying images with estimated scales leads to significant performance improvements for various tasks and methods. Specifically, we show how ScaleNet can be combined with sparse local features and dense correspondence networks to improve camera pose estimation, 3D reconstruction, or dense geometric matching in different benchmarks and datasets. We provide an extensive evaluation on several tasks, and analyze the computational overhead of ScaleNet. The code, evaluation protocols, and trained models are publicly available at <https://github.com/axelBarroso/ScaleNet>.

1. Introduction

Establishing correspondences is the very first step in many different 3D pipelines. Advancing on this task will have a direct impact on the performance of downstream applications such as camera pose estimation [39], autonomous driving [6], or 3D reconstructions [41]. However, methods that search for correspondences between images face significant challenges, and although some solutions are more mature than others, the task still is far from being solved.

As the field advances, even though the intermediate tasks in the correspondence search remain the same, their methods are being revisited and redesigned, *e.g.*, keypoint detectors/descriptors [10, 11, 37], dense geometric matchers [25, 53], or geometric verification techniques [31, 40, 46]. These new approaches have shown that the downstream tasks can be pushed to new performance levels through robust correspondences. The key objective of these new methods is to handle more and more extreme cases where previous pipelines failed, and although some methods are arguably application-specific [60], their robustness to extreme

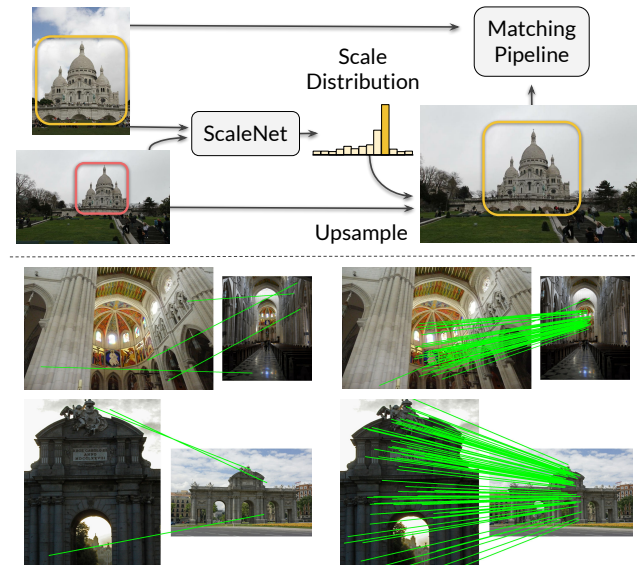


Figure 1. We propose a scale-aware system that helps establish correspondences under strong scale changes, allowing to correctly match pairs where previous pipelines were not successful. We introduce ScaleNet, a network that estimates the scale distribution between two images so that regions of interest can be corrected and exhibit the same scale factor (top). We display (bottom) an indoor and outdoor example where matches from R2D2 with multi-scale pyramid (left) increase if using ScaleNet rectification (right).

conditions is the main reason for success.

Inspired by the previous methods targeting visual robustness [29, 30, 58], we address the problem of handling the scale change between images, which is a long-standing challenge in computer vision [23, 26]. Scale robustness and estimation have been the focus of much research in the area of handcrafted feature extraction [23, 54, 60] as a reliable solution that can significantly boost the performance of existing methods. Moreover, the scale change is arguably the most challenging, and the most important parameter to estimate compared to rotation, translation, or even local affine deformations [54]. There are several strategies to deal with scale changes, with the multi-scale pyramid being one of the most popular solutions [3, 4, 11, 23, 37, 52, 54]. Although the multi-scale pyramid mitigates the problem of different

scales, it increases complexity and ambiguity as the matcher needs to establish correspondences among multiple scale levels. Figure 1 (bottom) shows an example of extreme pairs for which R2D2 [37] with multi-scale pyramid can only get a high number of correct matches once we rectify the scales. Besides the added complexity, a multi-scale pyramid is not always a straightforward solution to incorporate in some methods, such as dense correspondence networks [18,45,53]. In contrast to multi-scale pyramids, some works aim at being invariant to different scales through their learning process [10,34], however, as a side effect, they become progressively less discriminative [55]. Another possible direction, and popular strategy, is to estimate the local or global transformations and rectify the images prior to establishing the correspondences [30,36,50,58].

The scale factor characterizes the relationship between pairs of images and, in general, an accurate estimate can only be achieved when considering both images at the same time. Using pairs of images as input may increase the complexity beyond acceptable in some applications such as large-scale retrieval and localization unless used in their final verification stage. Nonetheless, solving the scale before the main analysis improves the discriminative power and allows less robust but more efficient methods to be used in challenging scenarios [34]. Hence, we propose a new approach that estimates and corrects the scale factor between a given pair of images before the correspondence search, which is illustrated at the top of figure 1. Our scale predictor network, termed ScaleNet, is inspired by dense geometric methods [25,51,53] and conditioned local features extractors [14,34,55]. ScaleNet extracts features from two low-resolution images and exploits CNN correlation layers to predict the scale factor. Due to the non-linear nature of scale changes, we formulate the scale regression problem as the estimation of a probability distribution in logarithmic space. We show how ScaleNet can be combined with different methods and demonstrate the improvements on different tasks and datasets.

Our contributions include: 1) a scale-aware matching system based on a novel scale predictor architecture, 2) a strategy to measure and label the scale factor between two images, and 3) a learning scheme that tackles the non-linear nature of scale changes.

2. Related work

Recent works have allowed significant progress in establishing good correspondences between images. Although many works have focused on solving entire tasks in an end-to-end manner [5,39], there are lots of efforts focused on identifying limitations and improving the robustness of individual steps in modern pipelines [3,28,40].

Image rectification consists of predicting or applying a

set of transformations to the images so that the search for correspondences is done in an optimum setting. Pioneering work on this area is ASIFT [58], which applies multiple affine transformations to find less challenging image pairs for matching with SIFT [23]. MODS [29] investigated this line of work by introducing an iterative scheme to generate intermediate synthetic views between images. MODS also proposed an adaptive system to avoid applying synthetic transformations to easy-to-match pairs, being faster and more versatile than previous ASIFT. Closer to our work is [60], which computes the scale factor between a pair of images by detecting and matching exhaustively SIFT features on multiple scale levels. Although it shows that they can deal with strong scale changes, their method is tied to the need of visiting all possible scaled images before the actual local feature matching stage. One negative aspect of these methods is that they still require a blind exploration of synthetic views to find the optimal image transformations. A few works have tried to overcome the previous limitation and proposed to learn such parameterizations directly from the images. One of the first attempts is [57], where authors introduced a neural network to assign a canonical orientation to every input image patch before the descriptor architecture. In AffNet [30], authors follow this trend and learn a full affine shape estimator to geometrically align input patches before the descriptor. Moreover, [36] proposed a scene-specific overlapping estimator and showed that scale rectification based on image overlapping can improve feature matching. Unlike previously learned methods, our labeling strategy is based on keypoint distance ratios, resulting in a scene-agnostic scale predictor that uses pairs of images as input. Even though our ScaleNet tackles only the scale factor out of several possible transformation parameters, we show in section 5 that the scale factor is crucial for boosting the performance of current methods.

Visual robustness has been the focus of numerous works in the field of correspondence search [4,13,24]. The rotation has been addressed by correcting input patches [13,23,30] before extracting local descriptors [28,48,49], or by designing robust architectures [4,22,24]. However, in the context of scale changes, the standard strategy is the well-known multi-scale (M-S) pyramid approach, which applies methods at different re-scaled versions of the image [11,37,47]. Even though M-S pyramids offer a versatile solution for many applications, it does not provide a suitable approach for extreme scale changes (cf. section 5), and thus, both, single and multi-scale feature extractors, benefit from correcting the scale factor before extracting features. Moreover, recent works show that there is growing interest in pair-wise methods, *i.e.*, that use two input images at the same time to establish the local or dense correspondences [14,18,25,35,45,53,55], but, in that scenario, there is no

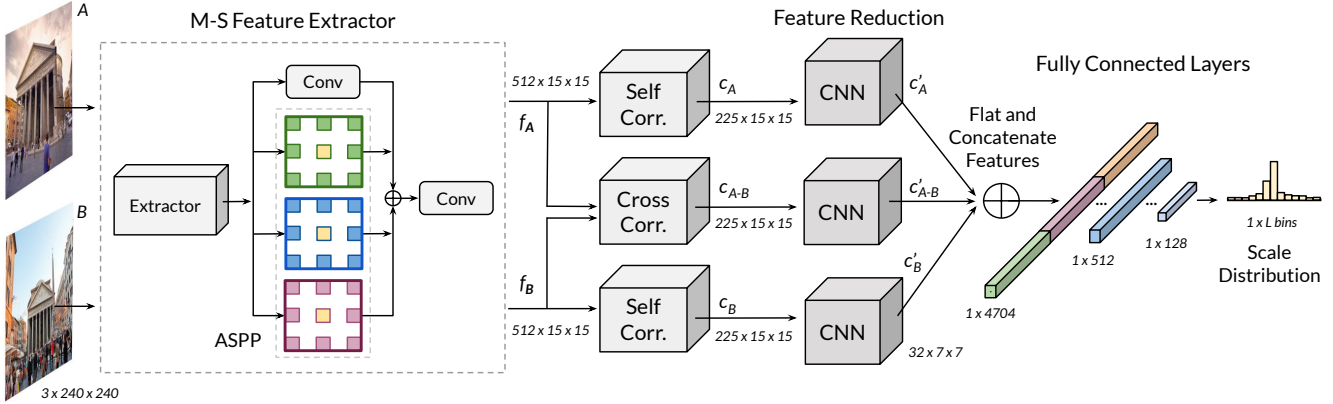


Figure 2. ScaleNet uses a multi-scale (M-S) feature extractor and a combination of self- and cross-correlation layers to calculate the relationship within each image and between them. The M-S extractor is composed of a common CNN and an ASPP block [7, 8]. The ASPP module consists of three 3×3 convolutional layers with different dilation rates. *Conv* blocks perform 1×1 convolutions to learn the cross-channel interactions that go into the self- and cross-correlation layers. Correlation volumes’ dimensionality is reduced through a CNN, and its results are concatenated into a flat feature vector. Finally, consecutive fully connected layers calculate the scale distribution.

clear or effective strategy for dealing with scale changes. Hence, given that such methods already take two images as inputs, ScaleNet rectification offers a more natural and intuitive process towards visual robustness than M-S pyramids.

3. Method

ScaleNet embraces several key concepts to deliver good performance in practical settings. The first key aspect is that it is effective for low-resolution images, which makes it more efficient. Another important idea is the formulation of the scale estimation as a distribution prediction in logarithmic space rather than a regression problem [17, 30, 56]. This allows using a simple and shallow, yet effective architecture. Figure 2 presents our ScaleNet architecture, and the following sections detail each of the aspects of ScaleNet and its learning scheme.

3.1. ScaleNet architecture

Consider A and B as input images to ScaleNet. Low-resolution images A and B are processed by a multi-scale feature extractor block, which is composed of a generic network, *e.g.*, VGG, or ResNet, followed by the atrous spatial pyramid pooling (ASPP) [7, 8]. The ASPP block achieves multi-scale robustness by applying to the feature map 3×3 dilated convolutions, each with a different dilation rate, *e.g.*, 2, 3, and 4. Thus, the ASPP block allows the network to compute and fuse features from different receptive fields. M-S features are concatenated and fed into a final 1×1 convolution, which combines the features from local and global areas at a minimum cost. We then apply self- and cross-correlation layers to multi-scale features f_A and f_B , and obtain the correlation maps, c_A , c_B , and c_{A-B} , which contain the self- and cross-pairwise similarities. As in [38], ReLU

and L2-normalization are applied to the correlation maps before the feature reduction blocks, which are composed of four Conv-Batch-ReLU layers each. Finally, c'_A , c'_B , and c'_{A-B} maps are flattened, concatenated, and fed into a set of fully connected layers to predict the scale distribution P , with a final softmax activation layer.

3.2. Predicting scale distributions

ScaleNet outputs a scale distribution rather than a regressed single scale factor, which helps the network to converge to a reliable model. In contrast, when tackling the problem as a regression task, the same network cannot predict an accurate scale (*cf.* appendix C.1). We attribute this to the fact that the network can learn and interpret the relationships between the quantized scale ranges and solve an easier classification task, which requires it to assign the weights to the predefined scale factors instead of predicting its actual value.

We formulate the scale estimation as a problem of predicting the probability distribution in a scale-space quantized into L bins. Given images A and B , our objective function measures the distance between our computed scale distribution, $P_{A \rightarrow B}$, and the ground-truth distribution $P_{A \rightarrow B}^{gt}$:

$$Loss(A, B) = KL(P_{A \rightarrow B}, P_{A \rightarrow B}^{gt}), \quad (1)$$

where $KL(\cdot, \cdot)$ is the Kullback-Leibler divergence loss. To obtain the scale factor from the probability distribution $P_{A \rightarrow B}$, we combine all scale levels using a soft-scale computation. It enables the network to output scale factors that interpolate between the quantized scale values, thus covering all possible scales between images A and B . Soft-assignment gives the architecture further flexibility and

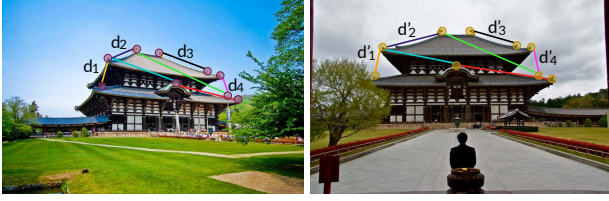


Figure 3. In the dataset generation, we randomly pick pairs of corresponding points in both views and compute their distances in the image planes. The final scale ground-truth is the average of the distance ratios of all picked keypoints.

robustness when inferring scales as shown in section 5.4.

The scale factor is a relative ratio operator, hence, it is non-linear. To avoid a bias by high scale values when computing the soft-scale, we transform the quantized scale classes, s_i , to logarithmic space. The logarithmic transformation allows to calculate the soft-scale, $\bar{S}_{A \rightarrow B}$, as a linear combination of the logarithmic scales, \bar{s}_i , weighted by the predicted scale probabilities from softmax output p_i . The global scale factor $\bar{S}_{A \rightarrow B}$ in log-scale is given as:

$$\bar{S}_{A \rightarrow B} = \sum_{i=0}^{L-1} p_i \cdot \ln(s_i), \quad (2)$$

where $s_i = \sigma^t$ corresponds to the quantized scale for bin i , σ is our predefined base scale factor, integer $t \in [-L/2, \dots, 0, \dots, +L/2]$, and $L = 2t + 1$ is the total number of scale bins. Moreover, we improve the robustness of our scale estimator by a simple yet effective consistency check trick, where we compute the scale factor, $\bar{S}_{A \rightarrow B}$, and its inverse, $\bar{S}_{B \rightarrow A}$, and combine them as:

$$\hat{S}_{A \rightarrow B} = \frac{\bar{S}_{A \rightarrow B} - \bar{S}_{B \rightarrow A}}{2} \quad \text{and} \quad S_{A \rightarrow B} = e^{\hat{S}_{A \rightarrow B}}, \quad (3)$$

with $S_{A \rightarrow B}$ as the final scale factor between the images.

3.3. Dataset generation

ScaleNet is trained with synthetically generated image pairs as well as images from real scenes.

Synthetic pairs. We define a set of planar affine transformations to map one image into another. Such image pairs are easy to generate on-demand for any ground-truth scale S_{gt} , however, they do not include the real noise from different imaging conditions.

Real pairs present more challenging conditions than synthetic pairs, *i.e.*, non-planar viewpoint changes, weather/illumination conditions, or occlusions, among others. In addition, in contrast to the previous synthetic global transformations, the scale between two real images may

spatially vary, and areas of different depth or strong perspective changes may include various scale factors. Thus, we introduce a new approach for obtaining training data by estimating the scale factors between real images. We use 3D reconstruction datasets, where 3D point clouds and their corresponding projected 2D positions on the images are available. First of all, given the 3D model, we find pairs of images with an overlap higher than 10% computed as in [11, 32]. For each pair of images A and B , we identify the 3D points of the model that are visible in both images. Using the registered 3D points from the model as opposed to sampling random positions ensures that the regions used for computing the scale are discriminative. Thus, given the covisible 3D points, we query its projected 2D positions, k_A , and k_B , on image A and B . We randomly sample pairs of points, k_A and k_B , and compute their distances as shown in figure 3.

The scale factor between two images with keypoints i and j is calculated as the ratio of their distances in image A and B :

$$S_{i-j} = \frac{\|k_{Bi} - k_{Bj}\|}{\|k_{Ai} - k_{Aj}\|} \quad \text{and} \quad i, j \in [1, \dots, K], \quad (4)$$

with K as the total number of covisible 3D points between images A and B . As different regions may exhibit different scale factors, we compute the global scale factor as the average of ratios in logarithmic scale after sampling R different pairs of points:

$$S_{A \rightarrow B}^{gt} = e^{\bar{S}_{A \rightarrow B}} \quad \text{where} \quad \bar{S}_{A \rightarrow B} = \frac{1}{R} \sum_{i \neq j} \ln(S_{i-j}). \quad (5)$$

As detailed in section 3.2, the ScaleNet learning scheme minimizes the K-L divergence between the predicted scales and ground-truth distributions. We, therefore, build the ground-truth scale distribution, $P_{A \rightarrow B}^{gt}$, such as it satisfies:

$$\ln(S_{A \rightarrow B}^{gt}) = \bar{S}_{A \rightarrow B}^{gt} = \sum_{i=0}^{L-1} p_i^{gt} * \ln(s_i), \quad (6)$$

where s_i are the quantized scale factors, and $P_{A \rightarrow B}^{gt} = [p_0^{gt}, \dots, p_{(L-1)}^{gt}]$ is the ground-truth distribution of the scale estimated as a normalized histogram of point pairs.

4. Implementation notes

ScaleNet details. ScaleNet predicts a scale distribution with $L = 13$ bins and $\sigma = \sqrt{2}$, giving possible scale factors in range $S \in [0.16, \dots, 6]$. ASPP module has 3 levels with dilation rates 2, 3, and 4. During training, ScaleNet uses Adam Optimizer with a learning rate of 10^{-4} and a decay factor of 0.1 every ten epochs. The training takes on average 40 epochs, 20 hours on a machine

	Pose estimation (AUC)			Time (ms)
	at 5°	at 10°	at 20°	
Baseline	4.8	7.4	10.4	-
VGG-16	5.4	8.3	11.8	8.3
ResNet-50	5.6	8.8	12.7	15.8
VGG w/ self-corr.	6.2	9.3	12.5	11.3
VGG w/ ASPP	7.3	11.1	15.7	16.6
Ours (VGG+self-corr+ASPP)	8.4	12.3	17.6	19.4
Ours + Consistency check	8.7	13.4	19.5	19.8

Table 1. Ablation study of the different ScaleNet’s design choices. Baseline refers to SuperPoint [10] without scale correction.

with an i7-7700 CPU running at 3.60GHz, and an NVIDIA GeForce GTX 1080-Ti. ScaleNet model and training scripts are implemented in PyTorch [33].

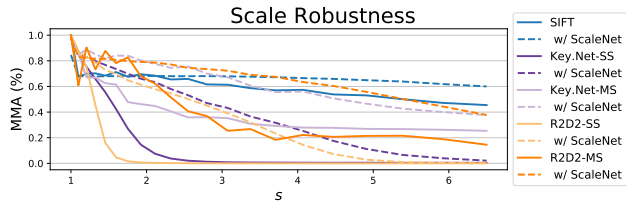
Dataset details. We use Megadepth dataset [20] for generating our custom training and testing dataset. We discard scenes that are in the PhotoTourism test from our training set as in [40] to avoid overlap. We keep 10% of the training scenes as our validation set. During data generation, we sample $R = 200$ random pairs of 3D points (cf. equation 5) for a robust scale estimation between the two images. We sample pairs of images with scale factors $S \in [0.16, \dots, 6]$ and create a collection of 250,000 training and 25,000 validation pairs where all scale factors are well represented. Synthetic pairs are generated on the fly from the Megadepth training images during training. We include more details and examples of our training set in appendix A.

5. Experiments

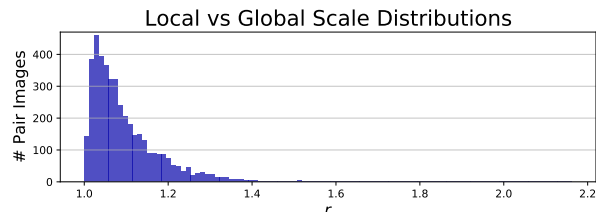
This section presents results for ScaleNet integrated with state-of-the-art methods on several datasets and tasks. Refer to appendix for more experiments and qualitative examples.

5.1. Preliminaries

Multi-scale pyramid & ScaleNet. Multi-scale (M-S) pyramids and ScaleNet aim at making methods more robust against arbitrary scale changes. Although both approaches tackle the same problem, each has its strengths, *e.g.*, M-S pyramids can compute a higher number of features by visiting multiple resized images, and ScaleNet offers a more natural integration into tasks where two images are given as input [14, 25, 53, 55]. Besides, we claim that ScaleNet improves not only single-scale feature extractors but also multi-scale ones. We analyze in figure 4a the robustness of methods against synthetic scale transformations and show how the combination with ScaleNet benefits them. In this experiment, we use 2,000 random images from the Megadepth dataset and scale them to create the pairs. We measure the mean matching accuracy (MMA) computed as in [37]. As expected, results indicate that the performance



(a) Mean matching accuracy under synthetic scale transformations, s .



(b) Histogram of the scale ratios, r , between the globally and locally estimated scales in image pairs used for training, with $r = s_a/s_b$, and where $s_a = \max(s_{local}, s_{global})$, and $s_b = \min(s_{local}, s_{global})$.

Figure 4. Analysis on the visual robustness against pure scale changes, and comparison of local vs global scales.

of single-scale methods, Key.Net [3] and R2D2 [37], drop notably even when images present small scale perturbations ($s > 1.5$). Meanwhile, M-S pyramid or ScaleNet strategies mitigate the effect and lead to a better approach. Moreover, we observe that the combination of M-S pyramids and ScaleNet achieves the top performance, and proves that both strategies contribute and work well together.

Local vs global scale estimation. ScaleNet can predict a scale for each point within the image, however, it is not straightforward to correct the scale factor locally for networks that process the whole image, *e.g.*, dense correspondences networks [25, 51, 53], or dense local feature extractors [10, 11, 37]. Figure 4b shows the histogram of the average scale ratios between the globally and the locally estimated scales per image. To compute the local scale values, we restrict the random sampling to spatially neighboring keypoints in equation 4 rather than points sampled across the whole image. Figure 4b shows that in the majority of images the differences between global and local estimations are small and within 1.0 and 1.2. A ratio r of 1.0 indicates that the local and the global scales between the two images are the same, and therefore, the global scale is valid across the whole image. Based on results in figure 4b, we argue that although local scales could bring an extra benefit in scenes with strong viewpoint or perspective changes, a single global scale will significantly contribute towards correcting images.

Ablation study in table 1 displays the contributions of each ScaleNet’s block towards robust scale estimation. We com-

	Pose estimation (AUC)		
	at 5°	at 10°	at 20°
SIFT	9.8	15.4	22.0
w/ ScaleNet	12.9 (+32%)	19.5 (+27%)	27.4 (+25%)
R2D2-SS	3.1	4.4	5.6
w/ ScaleNet	6.5 (+110%)	9.7 (+121%)	12.8 (+129%)
R2D2-MS	10.5	13.1	18.9
w/ ScaleNet	13.2 (+26%)	17.5 (+34%)	21.0 (+11%)
Key.Net-SS	4.7	6.9	9.3
w/ ScaleNet	8.9 (+89%)	14.0 (+103%)	19.5 (+110%)
Key.Net-MS	14.0	22.1	31.6
w/ ScaleNet	17.2 (+23%)	26.4 (+20%)	38.4 (+22%)
SuperPoint	5.4	8.2	11.0
w/ ScaleNet	8.1 (+50%)	12.4 (+51%)	17.3 (+57%)
SP+SuperGlue	16.5	25.3	35.0
w/ ScaleNet	22.5 (+36%)	33.9 (+34%)	45.5 (+30%)

Table 2. Relative camera pose results on custom Megadepth split with and without ScaleNet correction.

bine different architecture’s designs with SuperPoint [10] and test them in the task of relative camera pose estimation. We use Lowe’s ratio test [23] and MAGSAC [2] to compute camera poses, and, as in [40, 45], report the AUC of the pose errors at 5°, 10°, and 20°, where the error is calculated as the maximum of the rotation and translation angular errors. We sample 2,000 image pairs with scale factors between 0.16 and 6 from an independent validation set. Besides the AUC, we report the overhead inference time of each design.

We first compare the effect of different pre-trained feature extractors on ImageNet [9] (cf. figure 2), VGG-16 [44], and ResNet-50 [15], and see that the more complex ResNet representation contributes towards better poses with the downside of a higher computational cost. Thus, to keep our method light and fast, and given the similarity of the AUC scores, we use the VGG feature extractor for following experiments. In addition to the extractors, we analyze the effect of the self-correlation layers and the multi-scale ASPP component and observe that both boost the performance. Self-correlations capture the intra-image relationships and, therefore, give a better awareness of the global content. Additionally, ASPP offers a mechanism to extract more global features, thus, address larger scale changes. Furthermore, we show that the consistency check offers a higher AUC at a low computation cost. Note that M-S features and correlation layers only need to be computed once in inference and, hence, the extra cost for consistency check is small and proportional to running the dense layers twice.

5.2. Relative camera pose

Protocol. We first evaluate ScaleNet on the camera pose estimation task due to its natural integration into the existing pipelines. Similar to the previous ablation experiment, given a collection of image pairs, we calculate

	Geometric matching (PCK-5 %)		
	All	Easy	Hard
DGC-Net	40.2	34.4	4.5
w/ ScaleNet	41.4 (+3%)	36.8 (+7%)	20.1 (+347%)
GLU-Net	55.5	55.4	10.8
w/ ScaleNet	57.8 (+4%)	56.3 (+2%)	26.8 (+148%)

Table 3. Results on sparse correspondences for full, easy, and hard splits in Megadepth, consisting of 1,600, 627, and 440 pairs, respectively. Improvements from ScaleNet are across all splits, in particular, geometric matchers benefit largely from scale correction when there are large scale changes ($s > 1.8$).

the AUC of the camera pose error at 5°, 10°, and 20° as in [40, 45]. We use the test scenes from Megadepth and mine 4,000 images pairs with small and strong scale changes such as $s \in [0.16, 6]$. Appendix contains more details and examples of our dataset. We study the effect of ScaleNet on popular and publicly available local feature methods [3, 10, 23, 28, 37], and refer to Key.Net/HardNet as Key.Net in the following tables and figures.

Results in table 2 show that ScaleNet corrections boost the performance of all methods. As discussed previously, even though ScaleNet excels when combined with a single-scale method, *e.g.*, SuperPoint, ScaleNet is also able to improve the pose estimation of multi-scale extractors. We observe that the average improvements are 77% and 24% for single and multi-scale methods, respectively. In addition, we report results with SuperPoint and SuperGlue [40] and see that even this state-of-the-art matcher benefits (+33% on average) from a scale correction prior to the feature extraction.

5.3. Geometric matching

Protocol. We also evaluate ScaleNet on the geometric correspondence task by integrating it into the popular DGC-Net [25] and GLU-Net [53]. Note that ScaleNet can also be combined with other recent methods [45, 51, 52]. ScaleNet rescales one of the images before the dense correspondence network estimates the dense flow fields between the two images. Due to the lack of dense annotations on real image pairs with large viewpoint and illumination changes, we evaluate ScaleNet using sparse correspondences available in the Megadepth [20] dataset. Specifically, we follow the protocol of 1,600 image pairs introduced in [43] to compute the percentage of correct keypoints (PCK) under a 5 pixel acceptance threshold. The experiment is extended with multiple acceptance thresholds in appendix C.4.

Results in table 3 show the PCK scores obtained with and without scale correction before the dense architectures, DGC-Net and GLU-Net. Moreover, to highlight the benefit of ScaleNet for images at different scale factors, besides

		Reg. Images	Num. Obs.	Track Length	Rep. Error		Reg. Images	Num. Obs.	Track Length	Rep. Error
IMC Dataset (10 scenes)	SIFT	98.5	93k	7.26	0.80	SuperPoint	99.0	78k	10.47	1.17
	w/ ScaleNet	98.6	126k	10.20	0.87	w/ ScaleNet	99.5	135k	12.71	1.26
	w/ D-ScaleNet	98.5	118k	9.55	0.85	w/ D-ScaleNet	99.3	98k	11.96	1.22
	R2D2-SS	96.0	52k	8.35	0.93	R2D2-MS	97.9	86k	14.19	1.05
	w/ ScaleNet	97.2	89k	10.71	1.02	w/ ScaleNet	99.0	114k	15.35	1.13
	w/ D-ScaleNet	96.4	72k	9.43	1.01	w/ D-ScaleNet	97.7	95k	14.37	1.08
	Key.Net-SS	99.1	82k	11.05	1.06	Key.Net-MS	99.5	114k	16.23	0.96
	w/ ScaleNet	99.6	148k	14.22	1.16	w/ ScaleNet	99.6	148k	19.23	1.09
w/ D-ScaleNet	99.2	126k	12.71	1.10	w/ D-ScaleNet	99.5	147k	17.62	1.02	

Table 4. 3D reconstruction results on the 10 test scenes (100 images each) from the IMC dataset [59] with and without ScaleNet correction.

reporting the results for the full dataset (*All*), we create the *Easy* ($s > 1.2$) and *Hard* ($s > 1.8$) splits, where s indicates scale distortions factor between images. Each split has 1,600, 627, and 440 image pairs, respectively. We observe that ScaleNet integration does not bring significant improvements in the *All* data, where the majority of images have small scale variations, but it is important to note that it does not hurt the performance either. Meanwhile, in extreme cases (*Hard* split), current dense correspondence methods fail under severe scale changes, and their integration with ScaleNet improves the results by 347% and 148% for DGC-Net and GLU-Net.

5.4. 3D reconstruction

Protocol. ScaleNet can be easily integrated into geometric correspondence or relative camera pose pipelines, which are often a part of a more general 3D reconstruction system. To evaluate ScaleNet in this scenario, we follow the protocol proposed in the Local Feature Evaluation Benchmark [42] for building 3D reconstruction models. As ScaleNet can upsample one of the images, and that could result in a higher number of candidate keypoints, we limit the number of features to the top 2,048 keypoints based on the protocol proposed in [59]. We present results for the test split from the IMC dataset [59], which includes ten different scenes with 100 images each. IMC images pose significant challenges, *e.g.*, weather/illumination, perspective, scale, as well as strong occlusions.

As ScaleNet is applied to image pairs before the feature extraction, the detectors and descriptors are recomputed every time the scale is corrected. Hence, to reduce the computation time, we propose a discrete variant of ScaleNet (D-ScaleNet) that makes the extraction process more efficient. D-ScaleNet implements a hard-assignment by selecting the maximum scale instead of the soft-scale and consistency check from equation 2 and 3. Analogous to multi-scale pyramid approaches, we run the detectors/descriptors at multiple resized images but then select the optimal set of pre-computed features for matching based on D-ScaleNet estimation.

	Time (s)			
	Extraction	Matching	Reconst.	Total
SuperPoint	10.3	16.7	195.3	222.3
w/ ScaleNet	980.5	46.3	208.9	1235.7
w/ D-ScaleNet	141.8	161.8	205.6	509.2

Table 5. 3D reconstruction times on the *British Museum* scene (100 images) from IMC dataset [59].

Results in table 4 show the 3D reconstruction metrics of state-of-the-art methods with and without ScaleNet. We notice that image rectification by both, D-ScaleNet and ScaleNet, increases the number of registered images and the total number of observations in the 3D models. Track length is especially boosted by scale correction, meaning that the model was able to match the same keypoint simultaneously in more images. This increase of track length is particularly important since it proves that ScaleNet helps current methods distinguish and link points that were not possible without it, due to extreme view differences. On average, improvements added by ScaleNet are greater than those produced by D-ScaleNet, however, D-ScaleNet still brings a notable boost over baselines. On the opposite side, ScaleNet increases the reprojection error (Rep. Error) of the reconstructions by 0.09 points on average. We attribute this to their longer track lengths since more points are triangulated throughout the images, and thus, the reprojection error increases, which has also been reported in [42]. Longer tracks will benefit works that rely on complete and long tracks to refine the point positions and reduce their reprojection errors [12, 21]. In table 5, we show a comparison of the times taken to generate a 3D model when using ScaleNet and D-ScaleNet, and display the benefits in terms of computational time that D-ScaleNet provides.

5.5. Image matching

Protocol. We compute the Mean Matching Accuracy (MMA) [27] as the ratio of correctly matched features within a threshold (5 pixels) and the total number of features following the benchmark proposed in [11]. We report

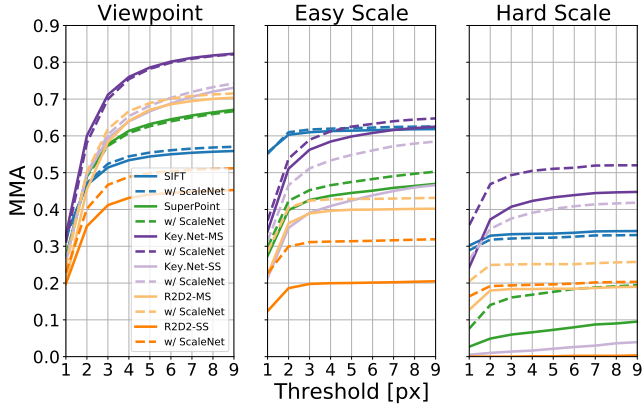


Figure 5. MMA on the full Viewpoint HPatches (left), scenes with easy (middle) and hard (right) scale transformations. Dashed lines (- -) show ScaleNet combined with other methods.

results for the 59 viewpoint scenes from the HPatches dataset [1]. Similar to the 3D task, we fix the number of features to top 1,000 as in [19] to eliminate the effect of increased matching scores from a high number of keypoints.

Results in figure 5 (left) show that ScaleNet improves the robustness of current methods overall, and excels for single-scale methods, *e.g.*, Key.Net-SS [3], or R2D2-SS [37]. However, in contrast to images used for statistics in figure 4b, most of the scenes in the HPatches Viewpoint contain strong perspective changes, *i.e.*, display different scale factors within the scene. Hence, to highlight the effect of ScaleNet, we show results for the subset of scenes with affine transformations, *i.e.*, scenes with stretch and skew transformations in addition to global scale and rotation. We select the splits such that the *Easy* has $s \in [1.2, \dots, 1.8]$ and the *Hard* has $s > 1.8$. When the scale factor is global across the image, ScaleNet correction can deal with planar scenes and improve the matching accuracy of single and multi-scale methods. As expected, results show that ScaleNet scaling is more critical for stronger scale changes in figure 5 (right). Only SIFT [23], which is specially designed to be robust against scale changes on planar scenes, does not benefit from ScaleNet correction in the *Hard* split. Moreover, to deal with scenes with possible strong perspective scale changes, we introduce in appendix Local-ScaleNet, which infers local scale factors and offers a more robust and functional alternative for scenes with perspective changes.

6. Discussion

Limitations. ScaleNet deals with arbitrary scale changes and, hence, it only brings improvements if such changes are present in the images. This makes ScaleNet useful in applications where those viewpoint changes prevent a successful matching of images, *e.g.*, extreme and sparse

collection of images, or ground-aerial applications. Nevertheless, even though ScaleNet does not boost performance when there is no scale change, it does not hurt either (cf. appendix C.3). Another limitation comes when ScaleNet needs to be applied to a large collection of images, *e.g.*, 3D tasks. ScaleNet works with pairs of images, hence, features can not be stored but need to be computed every time a new image is presented, increasing the feature extraction time as seen in table 5. Although we propose D-ScaleNet, which mitigates the complexity time for such tasks, ScaleNet can be further optimized for faster processing by replacing VGG with more compact models such as MobileNet [16], or by only using ScaleNet for computing camera pose after a restrictive retrieval search.

Societal impact. Image matching is a pivotal but small component within large systems that facilitate technologies like AR, 3D reconstruction, navigation, modeling, SLAM, among others. Hence, as we contribute towards more robust matching pipelines, ScaleNet’s societal impact is tied to the applications that rely on such technologies. Some applications may include smartphone apps, AR headsets, or autonomous cars. However, as our method cannot work independently of a larger system, the negative or ethical issues are not directly associated with our approach but rather with the specific business and final application where image matching may be used.

Reproducibility. The experiments are computed on standard and public datasets and tasks, and hence, they can be reproduced. Moreover, we made public the evaluation and training scripts, as well as our custom training dataset. In addition, to encourage the research on scale estimation, we published the test set of section 5.2 and splits of section 5.3 for easier comparison and support of future works.

7. Conclusions

We introduced ScaleNet, an approach that estimates the scale change between images and improves the performance of methods that search for correspondences throughout different views of the same scene. We proposed a novel learning scheme that formulates the problem of scale estimation as a prediction of a probability distribution of scales. We demonstrated how to make use of images from non-planar scenes to generate the training data. In addition to ScaleNet, we also introduced D-ScaleNet, a discrete variant of the proposed approach, and demonstrated its effectiveness in 3D-related tasks as well as computational time. We proved that ScaleNet can improve the results of popular pipelines in image matching for relative camera pose or 3D reconstruction while not being limited only to these tasks.

Acknowledgements. This project was supported by Chist-Era EPSRC IPALM EP/S032398/1 grant.

References

- [1] Vassileios Balntas, Karel Lenc, Andrea Vedaldi, and Krystian Mikolajczyk. Hpatches: A benchmark and evaluation of handcrafted and learned local descriptors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5173–5182, 2017. 8
- [2] Daniel Barath, Jiri Matas, and Jana Noskova. Magsac: marginalizing sample consensus. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10197–10205, 2019. 6
- [3] Axel Barroso-Laguna, Edgar Riba, Daniel Ponsa, and Krystian Mikolajczyk. Key.net: Keypoint detection by handcrafted and learned cnn filters. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019. 1, 2, 5, 6, 8
- [4] Axel Barroso-Laguna, Yannick Verdie, Benjamin Busam, and Krystian Mikolajczyk. Hdd-net: Hybrid detector descriptor with mutual interactive learning. In *Proceedings of the Asian Conference on Computer Vision*, 2020. 1, 2
- [5] Eric Brachmann and Carsten Rother. Neural-guided ransac: Learning where to sample model hypotheses. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4322–4331, 2019. 2
- [6] Mathias Bürki, Lukas Schaupp, Marcin Dymczyk, Renaud Dubé, Cesar Cadena, Roland Siegwart, and Juan Nieto. Vizard: Reliable visual localization for autonomous vehicles in urban outdoor environments. In *2019 IEEE Intelligent Vehicles Symposium (IV)*, pages 1124–1130. IEEE, 2019. 1
- [7] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017. 3
- [8] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 3
- [9] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009. 6
- [10] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 224–236, 2018. 1, 2, 5, 6
- [11] Mihai Dusmanu, Ignacio Rocco, Tomas Pajdla, Marc Pollefeys, Josef Sivic, Akihiko Torii, and Torsten Sattler. D2-net: A trainable cnn for joint detection and description of local features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 1, 2, 4, 5, 7
- [12] Mihai Dusmanu, Johannes L. Schönberger, and Marc Pollefeys. Multi-View Optimization of Local Feature Geometry. In *Proceedings of the European Conference on Computer Vision*, 2020. 7
- [13] Patrick Ebel, Anastasiia Mishchuk, Kwang Moo Yi, Pascal Fua, and Eduard Trulls. Beyond cartesian representations for local descriptors. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 253–262, 2019. 2
- [14] Hugo Germain, Guillaume Bourmaud, and Vincent Lepetit. S2dnet: Learning accurate correspondences for sparse-to-dense feature matching. In *European Conference on Computer Vision*, 2020. 2, 5
- [15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6
- [16] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *CoRR*, abs/1704.04861, 2017. 8
- [17] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *Advances in neural information processing systems*, pages 2017–2025, 2015. 3
- [18] Wei Jiang, Eduard Trulls, Jan Hosang, Andrea Tagliasacchi, and Kwang Moo Yi. Cotr: Correspondence transformer for matching across images. *arXiv preprint arXiv:2103.14167*, 2021. 2
- [19] Karel Lenc and Andrea Vedaldi. Large scale evaluation of local image feature detectors on homography datasets. In *The British Machine Vision Conference (BMVC)*, 2018. 8
- [20] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3302–3312, 2018. 5, 6
- [21] Philipp Lindenberger, Paul-Edouard Sarlin, Viktor Larsson, and Marc Pollefeys. Pixel-perfect structure-from-motion with featuremetric refinement. *arXiv preprint arXiv:2108.08291*, 2021. 7
- [22] Yuan Liu, Zehong Shen, Zhixuan Lin, Sida Peng, Hujun Bao, and Xiaowei Zhou. Gift: Learning transformation-invariant dense visual descriptors via group cnns. *arXiv preprint arXiv:1911.05932*, 2019. 2
- [23] David G. Lowe. Distinctive image features from scale-invariant keypoints. In *International booktitle of computer vision*, 2004. 1, 2, 6, 8
- [24] Zixin Luo, Lei Zhou, Xuyang Bai, Hongkai Chen, Jiahui Zhang, Yao Yao, Shiwei Li, Tian Fang, , and Long Quan. Aslfeat: Learning local features of accurate shape and localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 2
- [25] Iaroslav Melekhov, Aleksei Tiulpin, Torsten Sattler, Marc Pollefeys, Esa Rahtu, and Juho Kannala. DGC-Net: Dense geometric correspondence network. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2019. 1, 2, 5, 6
- [26] Krystian Mikolajczyk and Cordelia Schmid. Scale & affine invariant interest point detectors. *International journal of computer vision*, 60(1):63–86, 2004. 1
- [27] Krystian Mikolajczyk and Cordelia Schmid. A performance evaluation of local descriptors. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, number 10, pages 1615–1630. IEEE, 2005. 7

- [28] Anastasiia Mishchuk, Dmytro Mishkin, Filip Radenovic, and Jiri Matas. Working hard to know your neighbor’s margins: Local descriptor learning loss. In *Advances in Neural Information Processing Systems*, pages 4826–4837, 2017. [2](#), [6](#)
- [29] Dmytro Mishkin, Jiri Matas, and Michal Perdoch. Mods: Fast and robust method for two-view matching. In *Computer Vision and Image Understanding*, volume 141, pages 81–93. Elsevier, 2015. [1](#), [2](#)
- [30] Dmytro Mishkin, Filip Radenovic, and Jiri Matas. Repeatability is not enough: Learning affine regions via discriminability. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 284–300, 2018. [1](#), [2](#), [3](#)
- [31] Kwang Moo Yi, Eduard Trulls, Yuki Ono, Vincent Lepetit, Mathieu Salzmann, and Pascal Fua. Learning to find good correspondences. In *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2666–2674, 2018. [1](#)
- [32] Yuki Ono, Eduard Trulls, Pascal Fua, and Kwang Moo Yi. Lf-net: learning local features from images. In *Advances in Neural Information Processing Systems*, pages 6234–6244, 2018. [4](#)
- [33] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017. [5](#)
- [34] Rémi Pautrat, Viktor Larsson, Martin R. Oswald, and Marc Pollefeys. Online invariance selection for local feature descriptors. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. [2](#)
- [35] Rémi Pautrat, Viktor Larsson, Martin R. Oswald, and Marc Pollefeys. Online invariance selection for local feature descriptors. In *European Conference on Computer Vision*, 2020. [2](#)
- [36] Anita Rau, Guillermo Garcia-Hernando, Danail Stoyanov, Gabriel J Brostow, and Daniyar Turmukhambetov. Predicting visual overlap of images through interpretable non-metric box embeddings. In *European Conference on Computer Vision*, pages 629–646. Springer, 2020. [2](#)
- [37] Jerome Revaud, Philippe Weinzaepfel, César De Souza, and Martin Humenberger. R2d2: Repeatable and reliable detector and descriptor. In *Advances in Neural Information Processing Systems*, 2019. [1](#), [2](#), [5](#), [6](#), [8](#)
- [38] I. Rocco, R. Arandjelović, and J. Sivic. Convolutional neural network architecture for geometric matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. [3](#)
- [39] Paul-Edouard Sarlin, Cesar Cadena, Roland Siegwart, and Marcin Dymczyk. From coarse to fine: Robust hierarchical localization at large scale. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. [1](#), [2](#)
- [40] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020. [1](#), [2](#), [5](#), [6](#)
- [41] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4104–4113, 2016. [1](#)
- [42] Johannes L Schonberger, Hans Hardmeier, Torsten Sattler, and Marc Pollefeys. Comparative evaluation of hand-crafted and learned local features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1482–1491, 2017. [7](#)
- [43] Xi Shen, François Darmon, Alexei A Efros, and Mathieu Aubry. Ransac-flow: generic two-stage image alignment. *arXiv preprint arXiv:2004.01526*, 2020. [6](#)
- [44] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015. [6](#)
- [45] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8922–8931, 2021. [2](#), [6](#)
- [46] Weiwei Sun, Wei Jiang, Eduard Trulls, Andrea Tagliasacchi, and Kwang Moo Yi. Acne: Attentive context normalization for robust permutation-equivariant learning. In *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11286–11295, 2020. [1](#)
- [47] Yurun Tian, Vassileios Balntas, Tony Ng, Axel Barroso-Laguna, Yiannis Demiris, and Krystian Mikołajczyk. D2d: Keypoint extraction with describe to detect approach. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, 2020. [2](#)
- [48] Yurun Tian, Axel Barroso-Laguna, Tony Ng, Vassileios Balntas, and Krystian Mikołajczyk. Hynet: Local descriptor with hybrid similarity measure and triplet loss. In *Advances in neural information processing systems*, 2020. [2](#)
- [49] Yurun Tian, Xin Yu, Bin Fan, Fuchao Wu, Huub Heijnen, and Vassileios Balntas. Sosnet: Second order similarity regularization for local descriptor learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11016–11025, 2019. [2](#)
- [50] Carl Toft, Daniyar Turmukhambetov, Torsten Sattler, Fredrik Kahl, and Gabriel Brostow. Single-image depth prediction makes feature matching easier, 2020. [2](#)
- [51] Prune Truong, Martin Danelljan, Luc Van Gool, and Radu Timofte. Gocor: Bringing globally optimized correspondence volumes into your neural network, 2021. [2](#), [5](#), [6](#)
- [52] Prune Truong, Martin Danelljan, Luc Van Gool, and Radu Timofte. Learning accurate dense correspondences and when to trust them, 2021. [1](#), [6](#)
- [53] Prune Truong, Martin Danelljan, and Radu Timofte. GLU-Net: Global-local universal network for dense flow and correspondences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. [1](#), [2](#), [5](#), [6](#)
- [54] Tinne Tuytelaars and Krystian Mikołajczyk. Local invariant feature detectors: a survey. In *Foundations and Trends in Computer Graphics and Vision*, 2008. [1](#)

- [55] Olivia Wiles, Sebastien Ehrhardt, and Andrew Zisserman. Co-attention for conditioned image matching. In *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021. [2](#), [5](#)
- [56] Kwang Moo Yi, Eduard Trulls, Vincent Lepetit, and Pascal Fua. Lift: Learned invariant feature transform. In *European Conference on Computer Vision*, pages 467–483. Springer, 2016. [3](#)
- [57] Kwang Moo Yi, Yannick Verdie, Pascal Fua, and Vincent Lepetit. Learning to assign orientations to feature points. *CoRR*, abs/1511.04273, 2015. [2](#)
- [58] Guoshen Yu and Jean-Michel Morel. Asift: An algorithm for fully affine invariant comparison. *Image Processing On Line*, 1:11–38, 2011. [1](#), [2](#)
- [59] Jin Yuhe, Dmytro Mishkin, Anastasiia Mishchuk, Jiri Matas, Pascal Fua, Kwang Moo Yi, and Eduard Trulls. Image matching across wide baselines: From paper to practice. In *International Journal of Computer Vision*, 2020. [7](#)
- [60] Lei Zhou, Siyu Zhu, Tianwei Shen, Jinglu Wang, Tian Fang, and Long Quan. Progressive large scale-invariant image matching in scale space. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2362–2371, 2017. [1](#), [2](#)