Looking Beyond Corners: Contrastive Learning of Visual Representations for Keypoint Detection and Description Extraction

Henrique Siqueira\textsuperscript{1}  Patrick Ruhkamp\textsuperscript{1,2}  Ibrahim Halfaoui\textsuperscript{1,2}  Markus Karmann\textsuperscript{1}  Onay Urfalioglu\textsuperscript{1}

\textsuperscript{1} Technical University of Munich  \textsuperscript{2} HUAWEI, Munich Research Center

\texttt{first.second@huawei.com / tum.de}

Abstract

Learnable keypoint detectors and descriptors are beginning to outperform classical hand-crafted feature extraction methods. Recent studies on self-supervised learning of visual representations have driven the increasing performance of learnable models based on deep networks. By leveraging traditional data augmentations and homography transformations, these networks learn to detect corners under adverse conditions such as extreme illumination changes. However, their generalization capabilities are limited to corner-like features detected a priori by classical methods or synthetically generated data.

In this paper, we propose the Correspondence Network (CorrNet) that learns to detect repeatable keypoints and to extract discriminative descriptions via unsupervised contrastive learning under spatial constraints. Our experiments show that CorrNet is not only able to detect low-level features such as corners, but also high-level features that represent similar objects present in a pair of input images through our proposed joint guided backpropagation of their latent space. Our approach obtains competitive results under viewpoint changes and achieves state-of-the-art performance under illumination changes.

1. Introduction

“The AI revolution will not be supervised”.
- A. Efros \textsuperscript{40}.

Keypoint detection and description extraction in images are important tasks for many real-world applications including pose estimation \textsuperscript{11} \textsuperscript{30} \textsuperscript{15} and object tracking \textsuperscript{13}. Traditionally, those tasks are performed by extracting hand-crafted features based on prior information such as gradient-based methods \textsuperscript{19} \textsuperscript{5}. With the success of deep learning in computer vision, recent approaches that rely on learnable features by training deep neural networks have rapidly gained popularity \textsuperscript{10} \textsuperscript{24}. Most learnable methods, however, are trained under a (semi-)supervised regime and require a huge amount of labelled data from some source \textsuperscript{10} \textsuperscript{9}.

In fact, supervised learning approaches depend heavily on the amount and quality of labelled training data available to perform well under real-world conditions \textsuperscript{10} \textsuperscript{9}. This limiting factor has encouraged the scientific community to explore alternatives. Self-supervised methods, in particular

![The CorrNet framework](image-url)
Motivated by the latest achievements on contrastive learning, we present a novel paradigm for detecting repeatable and discriminative keypoints - beyond extracting sole corner and edge detections - by learning visual feature representations from weakly augmented image pairs in a fully unsupervised manner. By training the proposed Correspondence Network (CorrNet) under a contrastive regime with spatial constraints from weakly augmented image pairs, CorrNet neither relies on additional knowledge of the scene and camera poses [11, 30], nor strong priors such as homographic viewpoint changes [10, 4], supervised pre-training [10], or hand-crafted detection anchors [4] which are typical methods adopted to train modern feature extractors. In our approach, keypoints are detected based on novel guided grad-CAM [38, 36] of CorrNet’s latent space. Feature descriptors are extracted by computing representations of local patches surrounding detected keypoints with the identical network.

2. Related Work

Feature extraction (i.e., keypoint detection and description extraction) is a prerequisite for several computer vision applications. Early approaches [13, 19, 5, 25] proposed the use of hand-crafted feature extraction methods for corner and edge detection, followed by descriptor extraction of local information. However, their generalization capabilities are limited and their performance often deteriorates under adverse conditions such as strong illumination changes [10, 34, 8, 25].

In order to improve generalization performance for real-world applications, classical supervised learning approaches [31, 32, 1, 33] were applied for feature extraction, enhancing generalization by becoming invariant to various input variations including scale, rotation, and illumination in a sequential pipeline from keypoint detection to description extraction. Sequential approaches for feature extraction are still vastly adopted in computer vision [37, 39, 3, 24, 21, 20], but at present, the disadvantage of the performance of the latter (description extraction) being tightly dependent on the performance of the former (keypoint detection), limits capabilities and applicability.

Another paradigm for learnable feature extraction is to jointly train a model to detect keypoints and extract descriptions [41, 26, 10, 27, 11, 30]. For instance, in the work of DeTone et al. [10], a multi-head deep neural network (i.e., SuperPoint) is trained under a multitask learning setting where one head minimizes the location of keypoints and the other minimize a similarity measure description vectors with respect to the detected keypoints with weak supervision. The labels used to train SuperPoint are synthetically generated through homography transformations in the training images. Their approach achieved state-of-the-art results on the well-established HPatches benchmark [2], and hence, is defined as a baseline in our experiments.

3. The Correspondence Network Framework

The Correspondence Network (CorrNet) is an unsupervised visual representation learning approach for joint keypoint detection and description extraction. Unlike the majority of contrastive learning-based methods which use self-supervision as a precursory mechanism to improve performance on a subsequent task (e.g. classification), CorrNet leverages contrastive learning as the main training objective. We claim that a deep neural network trained for constraining similar pairs of images is able to learn repeatable and discriminative visual features. These features should be able to describe the similarity between two images, as well as recurrent patterns inherent in the training set. Consequently, they can be potentially considered as keypoints together with their latent representations as descriptors.

To detect repeatable and discriminative features in the input space, we propose an alternative approach to the guided grad-CAM algorithm [36] that is best tailored to approaches for metric learning such as CorrNet. In this section, we describe the CorrNet training procedure using contrastive learning under spatial constraints. Then, we conclude the section by describing the full pipeline in Figure 3 using first a CorrNet model for keypoint detection and global description, and then a fine-tuned one for discriminative local descriptions of the previously detected keypoints.

3.1. Contrastive Learning with Spatial Constraints

Motivated by the latest results on self-supervised learning [8], the core of the CorrNet framework consists of a siamese convolutional neural network trained via unsupervised contrastive learning. The final objective is to automatically learn visual similarities and differences between image pairs by exploiting ideas from SimCLR [7]. Unlike previous methods [35], SimCLR does not require accurate sample selection. Despite its flexibility, the data augmentation scheme plays an important role in the approach. Without proper data augmentation, the network tends to fail to learn useful visual representations as demonstrated by Chen et al. [7]. As a result, the performance of the network in subsequent tasks would be negatively affected. For SimCLR, colour distortions, Gaussian blur and random crops were used to transform the input images. The spatial information, however, was not explicitly taken into consideration by randomly cropping input regions and considering two crops without overlap as a positive pair. In our approach, we prove the potential of high repeatability of keypoints and discriminative local descriptions under contrastive learning, whereas spatial constraints should be modelled as soft con-
Fostering highly repeatable keypoints. With the aim to preserve spatial relationships between images, a positive image pair is defined as two random crops with overlapping regions in the image, as shown by the green crops in Figure 2 (top illustration). To increase the difficulty of the optimization objective and force CorrNet to learn efficient representations [42, 7], another pair can be randomly extracted from the same image but without overlapping regions with respect to the first pair (red crops). According to the object classification goal in SimCLR, the green crops should be considered similar to the red crops since both represent features of the same object, Big Ben. However, in our approach, they should be considered different, since there are no repeatable keypoints between them as they do not overlap. In addition to this sampling method, photometric and weak geometric transformations are applied to each image to make CorrNet invariant to input changes caused by illumination or perspective variations.

Fostering highly discriminative descriptions. As the contrastive training progresses, the network learns gradually to become invariant to strong input changes given the set of data augmentations. Assuming that two keypoints are close to each other in a reference image with common neighbouring pixels, a network trained to become invariant to this sort of situation would generate ambiguous description vectors, making the task of finding corresponding keypoints in the target image harder for the matching algorithm. Thus, we propose to train an additional set of weights for CorrNet under the output constraints as illustrated in Figure 2 (bottom) to generate better descriptors. With the goal to learn fine-grained visual differences between keypoint regions, a positive pair is then defined as crops of local regions centred at detected keypoints (the green crop) and the neighbouring local crops in red are considered to be the negative samples. This definition serves to illustrate the intuition behind our approach, but in practice, each image crop in the neighbourhood should get closer to its transformed version and repel away from all other local crops.

Both, CorrNet for keypoint detection and global-level description (referred to as CorrNet) and CorrNet for local discriminative descriptor extraction (named as l-CorrNet), are trained to minimize the normalized temperature-scaled cross entropy loss function (NT-Xent) [7], defined for a single positive pair as follows:

$$l_{i,j} = - \log \frac{\exp(\text{sim}(z_i, z_j)/\tau)}{\sum_{k=1}^{N} \mathbb{1}_{k \neq i} \exp(\text{sim}(z_i, z_k)/\tau)}$$

A positive pair of similar transformed input images \((x_i, x_j)\) are presented to the network to compute their respective description vectors \(z_i\) and \(z_j\). The exponential cosine similarity between the description vectors of the positive pair \((z_i, z_j)\) is divided by the summation of the exponential cosine similarities from negative pairs \((z_i, z_k)\) where \(x_i \neq x_k\). The temperature parameter \(\tau\) is utilized for normalization.

### 3.2. Joint Guided Gradient Back-Propagation

Figure 3 depicts the complete pipeline for joint keypoint detection and description extraction with the CorrNet framework. The reference and target images \((x, x')\) are introduced to the model to compute their description vectors \((z, z')\). In the case of joint input including a source and a target image, keypoints are detected by applying a modified version of the guided grad-CAM algorithm [36] on the CorrNet’s latent units with the goal to focus on salient shared regions between both images. Gradients are exclusively backpropagated for these regions with high weighting after being multiplied with the resulting embedding of the opposite image generated by the modified Grad-CAM approach. A qualitative impression is illustrated in Fig. 3. Our novel approach can robustly filter out many noisy keypoints in the non-overlapping regions in the background with non-similar image content in comparison with the standard guided backpropagation.

\[\text{\textbf{Figure 2: The CorrNet contrastive learning paradigm. Top:}}\]
\[\text{Large crops are extracted from an image to build positive and negative examples to learn visual feature representations with CorrNet for keypoint detection. Bottom: A multitude of small negative crops are sampled around a positive small crop to train a feature descriptor with CorrNet.} \]
CorrNet architecture: CorrNet encodes visual feature representations which are utilized to extract the highest activated neurons. Our novel guided backpropagation scheme identifies image regions with similar visual content, such that we can extract keypoints with NMS. Local patches around extracted keypoints are encoded with the same architecture as descriptors for further matching and subsequent vision tasks. Please note the overlap of keypoint activation in the GradCam stage: activated areas in the bottom correlate with the perspective scene change to correspond with top image.

Therefore, keypoints should be exclusively detected in these shared regions. In the case of a single image input to CorrNet, the standard guided backpropagation is performed [38]. In the literature, the guided grad-CAM algorithm [36] has been utilized to indicate units in the input space responsible for the high activation of a target neuron in the output layer, usually the highest activated unit. In our approach, we first project $z$ into $z'$ by calculating the vector multiplication between the description vectors. We define the unit with the highest value as the target for visualization. We claim that the high activation of this unit is the result of the presence of common visual features in the input space due to similarity between images $x$ and $x'$. The application of the guided backpropagation [38] on the target output unit with respect to the input image results in a tensor with the same size as the input. We consider the maximum gradient among the colour channels and normalize the values to the range [0, 1]. The resulting images can be seen in Figure 5 (the guided backpropagation step). The larger the gradient, the larger the influence of the input element on the target output element. Keypoints are detected by applying non-maximum suppression on the guided grad-CAM image and then selecting the top-k units with the largest gradients. Finally, the last part of the correspondence network framework extracts local crops centred at the detected keypoints. Each crop is an input to l-CorrNet to compute its descriptions. The cosine similarity between the local description vectors from l-CorrNet can be used to determine matches.

4. Experiments

In this section, we systematically evaluate CorrNet as a joint detector and descriptor under different experimental conditions. Firstly, we describe the experimental datasets, followed by the evaluation metrics and the implementation details. Afterwards, we present and discuss quantitative and qualitative results and analyze extensive ablation studies on CorrNet for keypoint detection.

4.1. Datasets

In our work, we follow the experimental protocol adopted by DeTone et al. [10] - thus adopting the resolution of 240x320 - which includes the use of the MS-COCO [18] as a training set and HPatches [2] as a test set.

Ideally, keypoints should be repeatable under various scene and image conditions, such as image noise, illumination and perspective changes. A widely used criterion is repeatability as introduced by Mikolajczyk and Schmid [23]. It represents the ratio between the number of repeatable keypoints over the total number of detections defined as follows:

$$\text{REP} = \frac{1}{N_1 + N_2} \left( \sum_i \gamma(x_i) + \sum_j \gamma(x_j) \right).$$ (2)
As a second metric, the localization error is used to evaluate the accuracy of the position of detected keypoints as follows:

\[
LE = \frac{1}{N} \sum_{i=1}^{N} \min_{j \in \{1, \ldots, N_2\}} \|x_i - \hat{x}_i\|,
\]

where correctness \(\gamma(x_i)\) is defined with respect to the threshold \(\epsilon\) depicting the maximum correct distance between two points, having \(N_1\) points in the first image and \(N_2\) points in the second image. We define correctness \([10]\) as:

\[
\gamma(x_i) = \left( \min_{j \in \{1, \ldots, N_2\}} \|x_i - \hat{x}_i\| \right) \leq \epsilon. \tag{4}
\]

The third metric is based on the estimated homography \(\hat{H}\) against the ground truth homography \(H\). We evaluate how accurately the homography transforms the four corners of one image onto the other. If the source image has corners \(c_1, c_2, c_3, c_4\) respectively, the ground-truth \(H\) homography is applied to recover ground truth corners in the target image \(c'_1, c'_2, c'_3, c'_4\) and the estimated homography \(\hat{H}\) to get \(\tilde{c}_1, \tilde{c}_2, \tilde{c}_3, \tilde{c}_4\). The threshold \(\epsilon\) is used to denote a correct homography and resulting scores should range between 0 and 1 where higher is always better.

\[
CorrH = \frac{1}{N} \sum_{i=1}^{N} \left( \sum_{j=1}^{4} \|c_{i,j} - \tilde{c}_{i,j}\| \right) \leq \epsilon \tag{5}
\]

Finally, we rely on the publicly available evaluation scripts from Jau \([15]\) and Sarlin \([34]\). In these scripts, keypoints detected in a reference image that would correspond to a location outside the target image (and vice-versa) are filtered out using the ground-truth homography provided in the HPatches dataset. Although this practice has been previously used in the literature \([10, 3]\), the use of the ground-truth homography would be unrealistic since it is not available in real-world applications. Therefore, this filtering method is omitted in our experiments in order to present results that better reflect the performance of the approaches in real-world applications.

### 4.3. Implementation Details

Our architecture consists of a siamese convolutional neural network. The CorrNet encoder, as illustrated in Figure 3, is a ResNet50 with no max-pooling layers and pre-trained weights of SimCLR\([17]\). The projection head, on the other hand, was not made available. In our approach, it comprises a multilayer perceptron of three layers with 2048, 512, and \(d\) neurons, respectively. The parameter \(d\) denotes the size of the descriptor vector. The Adam algorithm \([17]\) was used to optimize CorrNet. After extensive exploratory experiments, we fixed the learning rate to \(1e^{-3}\), weight decay to \(1e^{-6}\), batch size of 200 and the temperature parameter of the NT-Xent loss to 0.5.

### 4.4. Evaluation of the Keypoints

| Approach            | Nature | Illumination | Viewpoint |
|---------------------|--------|--------------|-----------|
| CorrNet (ours)      | us     | 0.86         | 1.95      |
| CorrNet (ours, single) | us     | 0.87         | 1.39      |
| Harris \([13]\)     | hc     | 74.3%        | 66.8%     |
| Shi-Tomasi \([16]\) | hc     | 73.9%        | 66.8%     |
| FAST \([32]\)       | hc     | 72.9%        | 65.7%     |
| SuperPoint (Gauss) \([15]\) | se    | 72.4%        | 64.1%     |
| SuperPoint \([10]\) | se     | 71.5%        | 55.8%     |
| ELF \([6]\)         | lo     | 74.3%        | 60.6%     |
| KeyNet \([4]\)      | se     | 70.6%        | 59.9%     |
| ASLFeat \([24]\)    | se     | 67.9%        | 54.0%     |
| R2D2 \([30]\)       | us     | 67.2%        | 46.3%     |
| SimCLR \([17]\)     | us     | 58.2%        | 38.4%     |
| Random              | hc     | 52.5%        | 26.2%     |

Table 1: Comparison with state of the art. Repeatability (rep.) and localization error (loc. error) of keypoints on HPatches for illumination and viewpoint changes. \(se\): semi-supervised, \(us\): unsupervised, and \(hc\): hand-crafted.

Table 1 presents the performance of our approach, CorrNet, in comparison with state-of-the-art methods for keypoint detection. To conduct a fair evaluation for all of the methods, they are strictly evaluated under the same condition. In summary, the input image size is 240x320, non-maximum suppression (NMS) of 3x3 is applied to increase the distribution of the detected key points in the image, no filtering method using ground truth is utilized, and the top 1000 keypoints are detected at each image. Moreover, only original implementations from their official repositories are used such as SuperPoint\([17]\), SuperPoint (Gauss)\([5]\) and KeyNet\([4]\). For the classical methods, we use the OpenCV implementations.

CorrNet achieves state-of-the-art results for illumination changes on both repeatability and localization error, demonstrating its robustness against adverse environmental conditions, and competitive results on viewpoint changes. Surprisingly, consistent with recent findings from Balntas and Lenc et al. \([2]\), classical hand-crafted methods can still play an important role in modern applications. Moreover, CorrNet’s superior results when compared against SimCLR (pre-trained network from Chen et al. \([7]\) using our guided grad-CAM that correlates two input images by exchanging the grad-CAM flow between images) and ELF \([6]\) (a
Figure 5: Qualitative results of keypoint detection using CorrNet, Harris [13] and SuperPoint [10]. Repeatability scores are shown on top of each pair of images. On the left results under illumination changes, and on the right results under viewpoint changes.

Figure 6: Qualitative results of keypoint matching using keypoints and descriptions from CorrNet, SIFT [19] and SuperPoint [10]. On the left results under illumination changes, and on the right results under viewpoint changes.

gradient-based method for keypoint detection using pre-trained networks), show that the proposed training method via contrastive learning under spatial constraints provides a significant performance improvement and plays an important role in detecting highly repeatable keypoints.

Figure 5 presents qualitative results on repeatability from CorrNet, Harris, and SuperPoint detections. While CorrNet is able to detect repeatable and stable keypoints under different illumination changes, Harris and SuperPoint suffer from the challenging illumination condition in the target image where keypoints are detected in dark regions due to the false-positive corner and edge detection caused by shades. Under different viewpoints, CorrNet achieves higher repeatability scores on these examples by successfully focusing the detection on similar objects present in the reference and target images, whereas the other methods blindly detect keypoints on the entire images.

Finally, it is important to notice that although CorrNet is trained in a fully unsupervised manner with weakly augmented image pairs, our generic training scheme can yield comparable results on viewpoint changes. In fact, our approach even outperforms semi- and self-supervised methods such as SuperPoint [10, 15], ELF [6], ASLFeat [22], and R2D2 [30]. We believe further research including perspective viewpoint changes, and exploiting the temporal domain of sequences would lead to further improvements.

4.5. Evaluation of the Descriptions

In this experiment, we evaluate l-CorrNet as a descriptor extraction method and compare our results with the state-of-the-art. Table 2 summarizes the success rate of homography estimation under different thresholds: 1, 3, and 5. Under illumination changes, l-CorrNet (single) achieves the most accurate homography estimation with a threshold of 1 by a large margin, and similar results compared to HardNet under a threshold of 3. With a threshold of 5 l-CorrNet
achieves comparable results to HardNet. For homography estimation from different viewpoints, SIFT shows to be the most reliable and accurate method. We observe l-CorrNet (single) achieves better results compared to l-CorrNet for homography estimation. Our hypothesis is, that keypoints that are essential for the homography computation may be located within regions that are less relevant in terms of similar visual features and are therefore not entirely recovered. Additionally, large homographic changes may lead to only a small overlap between image pairs, which would foster keypoint clusters with our novel approach, making it more challenging for homography computation.

From a qualitative analysis of CorrNet’s matching presented in Figure 6, we show CorrNet’s robustness against extreme illumination changes. In this second column, SIFT fails to match enough keypoints for accurate homography estimation. SuperPoint achieves similar qualitative results as CorrNet. Nonetheless, keypoints detected by CorrNet are more distributed over the images which often leads to more accurate homography estimation. Under viewpoint changes, the distribution of keypoints over the images may have affected homography estimation using CorrNet when compared against SIFT and SuperPoint (Figure 6 right).

### 4.6. Ablation Studies

Aiming to identify the major factors with a direct impact on CorrNet’s performance as a joint keypoint detector and descriptor extractor, we carefully investigate different parameters in the CorrNet framework. The research questions to be studied in this section are the following: (1) what is the impact of the description size on keypoint repeatability? (2) which latent vectors, $h$ or $z$, better encodes the visual similarities between two input images, reference and target? (3) What is the impact of different levels of spatial constraints on the repeatability of keypoints? (4) How can the likelihood of the detection of repeatable keypoints be increased? And finally, (5) what is the best set of data augmentation methods for keypoint detection?

**Description size and the most informative latent space.** Figure 4 depicts the training dynamics of CorrNet with respect to different description sizes (128, 256, and 512 units) by showing their repeatability over epochs. Solid lines show the repeatability of keypoints by applying the guided backpropagation with respect to the latent vector $h$, whereas dashed lines with respect to the latent vector $z$. Our results are consistent with the work of Chen et al. [7] which argues that the latent vector $h$ provides more useful visual representations of the data. Indeed, the repeatability scores are consistently higher for all of the CorrNets at different epochs when the guided backpropagation algorithm is applied from $h$. Note CorrNet’s low repeatability scores for both illumination (58.2%) and viewpoint (38.4%) in the beginning changes drastically during the first few epochs. These results suggest, that the pre-trained SimCLR, while performing highly accurate on ImageNet trained without spatial constraint, is not able to detect a high number of repeatable keypoints when compared to a mature CorrNet trained under the proposed spatial constraints. Regarding the description size, all of the CorrNets show similar behaviours. They achieve high repeatability scores after a few epochs and the difference in performance decreases over time.

**Contrastive learning under spatial constraints.** Figure 7 also shows repeatability over training epochs for different CorrNets trained under different spatial constraints on the illumination and viewpoint sets of the HPatches.
In this paper, we introduce CorrNet, a fully unsupervised approach for feature extraction. Our approach provides a flexible training strategy based on contrastive learning under spatial constraints and a novel keypoint detection algorithm based on an alternative to guided grad-CAM for metric learning. Our experiments demonstrate CorrNet’s robustness and advantages against previous methods under challenging illumination changes, achieving state-of-the-art results on HPatches [2], and competitive results under viewpoint changes.

We believe that fully unsupervised approaches based on contrastive learning will play a crucial role in many machine learning applications. The flexibility of setting up a
training strategy that guides the learned representations will potentially yield increasingly superior results on feature extraction in the wild. As a future direction, we will extend CorrNet to be able to perform lifelong learning \cite{28} of visual features for continually improving its performance on feature extraction under a pose estimation task.

References

[1] Pablo F Alcantarilla and T Solutions. Fast explicit diffusion for accelerated features in nonlinear scale spaces. *IEEE Trans. Patt. Anal. Mach. Intell.*, 34(7):1281–1298, 2011.

[2] Vassileios Balntas, Karel Lenc, Andrea Vedaldi, and Krystian Mikolajczyk. Ipatches: 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.

[3] Vassileios Balntas, Edgar Riba, Daniel Ponsa, and Krystian Mikolajczyk. Learning local feature descriptors with triplets and shallow convolutional neural networks. In *Bmvc*, volume 1, page 3, 2016.

[4] Axel Barroso-Laguna, Edgar Riba, Daniel Ponsa, and Krystian Mikolajczyk. Key.Net: Keypoint Detection by Handcrafted and Learned CNN Filters. In *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision*, 2019.

[5] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In *European conference on computer vision*, pages 404–417. Springer, 2006.

[6] Assia Benbihi, Matthieu Geist, and Cedric Pradalier. Elf: Embedded localisation of features in pre-trained cnn. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7940–7949, 2019.

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[8] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big self-supervised models are strong semi-supervised learners. In *Pre-proceedings of Advances in Neural Information Processing Systems*, volume 34, pages --, 2021.

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[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.

[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 *CVPR 2019-IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[12] Marco Gori. *Machine Learning: A constraint-based approach*. Morgan Kaufmann, 2017.

[13] Christopher G Harris and Mike Stephens. A combined corner and edge detector. Citeseer, 1988.

[14] Hou-Ning Hu, Yung-Hsu Yang, Tobias Fischer, Trevor Darrell, Fisher Yu, and Min Sun. Monocular quasi-dense 3d object tracking. *arXiv preprint arXiv:2103.07351*, 2021.

[15] You-Yi Jau, Rui Zhu, Hao Su, and Mannohman Chandraker. Deep keypoint-based camera pose estimation with geometric constraints. In *IROS*, 2020.

[16] Jianbo Shi and Tomasi. Good features to track. In 1994 *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 593–600, 1994.

[17] Diedrik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[18] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[19] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[20] Zixin Luo, Tianwei Shen, Lei Zhou, Jiahui Zhang, Yao Yao, Shiwei Li, Tian Fang, and Long Quan. Contextdesc: Local descriptor augmentation with cross-modality context. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2527–2536, 2019.

[21] Zixin Luo, Tianwei Shen, Lei Zhou, Siyu Zhu, Runze Zhang, Yao Yao, Tian Fang, and Long Quan. Geodesc: Learning local descriptors by integrating geometry constraints. In *Proceedings of the European conference on computer vision (ECCV)*, pages 168–183, 2018.

[22] 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/CVF conference on computer vision and pattern recognition*, pages 6589–6598, 2020.

[23] Krystian Mikolajczyk and Cordelia Schmid. Scale & affine invariant interest point detectors. *International journal of computer vision*, 60(1):63–86, 2004.

[24] Anastasiya Mishchuk, Dmytro Mishkin, Filip Radenovic, and Jiri Matas. Working hard to know your neighbor’s margins: Local descriptor learning loss. *arXiv preprint arXiv:1705.10872*, 2017.

[25] Mark Nixon and Alberto Aguado. *Feature extraction and image processing for computer vision*. Academic press, 2019.

[26] Hyeonwoo Noh, Andre Araujo, Jack Sim, Tobias Weyand, and Bohyung Han. Large-scale image retrieval with attentive deep local features. In *Proceedings of the IEEE international conference on computer vision*, pages 3456–3465, 2017.

[27] Yuki Ono, Eduard Trulls, Pascal Fua, and Kwang Moo Yi. Lt-net: learning local features from images. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 6237–6247, 2018.
[28] German I. Parisi, Ronald Kemker, Jose L. Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *Neural Networks*, 113:54–71, 2019.

[29] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, ICML 2021, 18-24 July 2021, Virtual Event, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 2021.

[30] Jerome Revaud, Cesar De Souza, Martin Humenberger, and Philippe Weinzaepfel. R2d2: Reliable and repeatable detector and descriptor. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[31] Edward Rosten and Tom Drummond. Machine learning for high-speed corner detection. In *European conference on computer vision*, pages 430–443. Springer, 2006.

[32] Edward Rosten, Reid Porter, and Tom Drummond. Faster and better: A machine learning approach to corner detection. *IEEE transactions on pattern analysis and machine intelligence*, 32(1):105–119, 2008.

[33] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. Orb: An efficient alternative to sift or surf. In *2011 International conference on computer vision*, pages 2564–2571. Ieee, 2011.

[34] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperGlue: Learning feature matching with graph neural networks. In *CVPR*, 2020.

[35] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[36] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[37] Edgar Simo-Serra, Eduard Trulls, Luis Ferraz, Iasonas Kokkinos, Pascal Fua, and Francesc Moreno-Noguer. Discriminative learning of deep convolutional feature point descriptors. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 118–126, 2015.

[38] J Springenberg, Alexey Dosovitskiy, Thomas Brox, and M Riedmiller. Striving for simplicity: The all convolutional net. In *ICLR (workshop track)*, 2015.

[39] Yurun Tian, Bin Fan, and Fuchao Wu. L2-net: Deep learning of discriminative patch descriptor in euclidean space. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 661–669, 2017.

[40] Yann LeCun. Yann LeCun - Self Supervised Learning — ICLR 2020. [https://www.youtube.com/watch?v=8ITK-Dd0H9U](https://www.youtube.com/watch?v=8ITK-Dd0H9U) 2020. Online; accessed 17 March 2021.

[41] 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.

[42] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European conference on computer vision*, pages 649–666. Springer, 2016.