HAL
open science

RANSAC-Flow: Generic Two-Stage Image Alignment

Xi Shen, François Darmon, Alexei Efros, Mathieu Aubry

To cite this version:

Xi Shen, François Darmon, Alexei Efros, Mathieu Aubry. RANSAC-Flow: Generic Two-Stage Image Alignment. European Conference on Computer Vision 2020, pp.618-637, 2020, 10.1007/978-3-030-58548-8_36 . hal-03131124

HAL Id: hal-03131124

https://hal.science/hal-03131124

Submitted on 4 Feb 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

RANSAC-Flow: generic two-stage image alignment

Xi Shen[1], François Darmon[1,2], Alexei A. Efros[3], and Mathieu Aubry[1]

[1] LIGM (UMR 8049) - Ecole des Ponts, UPE
[2] Thales Land and Air Systems
[3] UC Berkeley

Abstract. This paper considers the generic problem of dense alignment between two images, whether they be two frames of a video, two widely different views of a scene, two paintings depicting similar content, etc. Whereas each such task is typically addressed with a domain-specific solution, we show that a simple unsupervised approach performs surprisingly well across a range of tasks. Our main insight is that parametric and non-parametric alignment methods have complementary strengths. We propose a two-stage process: first, a feature-based parametric coarse alignment using one or more homographies, followed by non-parametric fine pixel-wise alignment. Coarse alignment is performed using RANSAC on off-the-shelf deep features. Fine alignment is learned in an unsupervised way by a deep network which optimizes a standard structural similarity metric (SSIM) between the two images, plus cycle-consistency. Despite its simplicity, our method shows competitive results on a range of tasks and datasets, including unsupervised optical flow on KITTI, dense correspondences on Hpatches, two-view geometry estimation on YFCC100M, localization on Aachen Day-Night, and, for the first time, fine alignment of artworks on the Brughel dataset. Our code and data are available at http://imagine.enpc.fr/~shenx/RANSAC-Flow/.

Keywords: unsupervised dense image alignment, applications to art

1 Introduction

Dense image alignment (also known as image registration) is one of the fundamental vision problems underlying many standard tasks from panorama stitching to optical flow. Classic work on image alignment can be broadly placed into two camps: parametric and non-parametric. Parametric methods assume that the two images are related by a global parametric transformation (e.g. affine, homography, etc), and use robust approaches, like RANSAC, to estimate this transformation. Non-parametric methods do not make any assumptions on the type of transformation, and attempt to directly optimize some pixel agreement metric (e.g. brightness constancy constraint in optical flow and stereo). However, both approaches have flaws: parametric methods fail (albeit gracefully) if the parametric model is only an approximation for the true transform, while

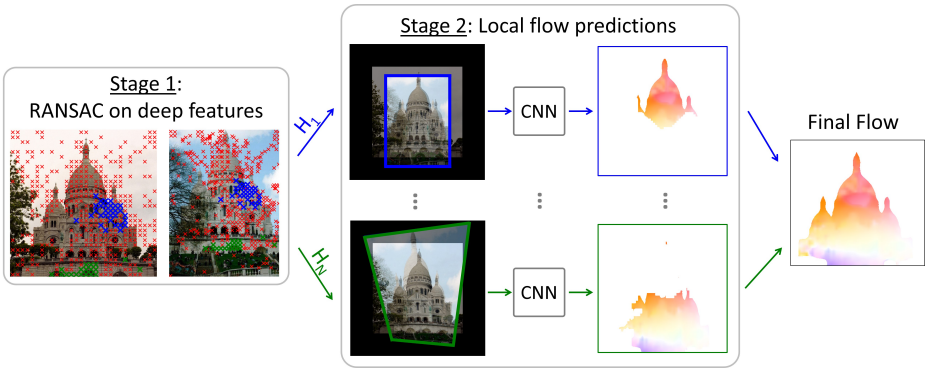
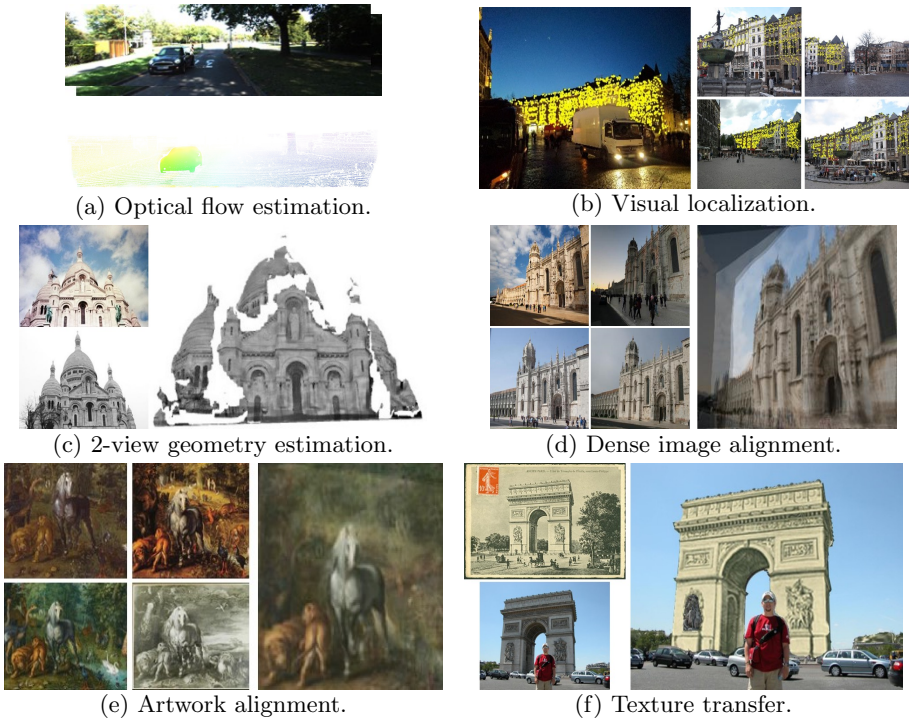


Fig. 1: **Overview of RANSAC-Flow.** Stage 1: given a pair of images, we compute sparse correspondences (using off-the-shelf deep features), use RANSAC to estimate a homography, and warp second image using it. Stage 2: given two coarsely aligned images, our self-supervised fine flow network generates flow predictions in the matchable region. To compute further homographies, we can remove matched correspondences, and iterate the process.

non-parametric methods have trouble dealing with large displacements and large appearance changes (e.g. two photos taken at different times from different views). It is natural, therefore, to consider a hybrid approach, combining the benefits of parametric and non-parametric methods together.

In this paper, we propose RANSAC-flow, a two-stage approach integrating parametric and non-parametric methods for generic dense image alignment. Figure 1 shows an overview. In the first stage, a classic geometry-verification method (RANSAC) is applied to a set of feature correspondences to obtain one or more candidate coarse alignments. Our method is agnostic to the particular choice of transformation(s) and features, but we've found that using multiple homographies and off-the-shelf self-supervised deep features works quite well. In the second non-parametric stage, we refine the alignment by predicting a dense flow field for each of the candidate coarse transformations. This is achieved by self-supervised training of a deep network to optimize a standard structural similarity metric (SSIM) [85] between the pixels of the warped and the original images, plus a cycle-consistency loss [93].

Despite its simplicity, the proposed approach turns out to be surprisingly effective. The coarse alignment stage takes care of large-scale viewpoint and appearance variations and, thanks to multiple homographies, is able to capture a piecewise-planar approximation of the scene structure. The learned local flow estimation stage is able to refine the alignment to the pixel level without relying on the brightness constancy assumption. As a result, our method produces competitive results across a wide range of different image alignment tasks, as shown in Figure 2: (a) unsupervised optical flow estimation on KITTI [48]



(a) Optical flow estimation. (b) Visual localization.

(c) 2-view geometry estimation. (d) Dense image alignment.

(e) Artwork alignment. (f) Texture transfer.

Fig. 2: RANSAC-Flow provides competitive results on a wide variety of tasks and enables new challenging applications.

and Hpatches [5], (b) visual localization on Aachen Day-Night [69], (c) 2-view geometry estimation on YFCC100M [79], (d) dense image alignment, and applications to (e) detail alignment in artwork and (f) texture transfer. Our code and data are available at http://imagine.enpc.fr/~shenx/RANSAC-Flow/.

2 Related Work

Feature-based image alignment. The classic approach to align images with very different appearances is to use sparse local image features, such as SIFT [40], which are designed to deal with large viewpoint and illumination differences as well as clutter and occlusion. These features have to be used together with a geometric regularization step to discard false matches. This is typically done using RANSAC [18,58,6,8] to fit a simple geometric transformation (e.g. affine or homography) [78]. Recently, many works proposed to learn better local features [42,14,80,49,43,64]. Differentiable and trainable version of RANSAC have also been developed [88,56,54,59].

Using mid-level features [76,29,28,30] instead of local keypoints, proved to be beneficial for matching visual content across modalities, e.g. 3D models and

paintings [3]. Recently, [73] learned deep mid-level features for matching across different visual media (drawings, oil paintings, frescoes, sketches, etc), and used them together with spatial verification to discover copied details in a dataset of thousands of artworks. [66] used deep feature map correlations as input to a regression network on synthetic image deformations to predict the parameters of an affine or thin-plate spline deformation. Finally, transformer networks [26] can also learn parametric alignment typically as a by-product of optimizing a classification task.

Direct image alignment. Direct, or pixel-based, alignment has its roots in classic optical flow methods, such as Lucas-Kanade [41], who solve for a dense flow field between a pair of images under a brightness constancy assumption. The main drawback is these methods tend to work only for very small displacements. This has been partially addressed with hierarchical flow estimation [78], as well as using local features in addition to pixels to increase robustness [9,62,4,22]. However, all such methods are still limited to aligning very similar images, where the brightness constancy assumption mostly holds. SIFT-Flow [38] was an early method that aimed at expanding optical flow-style approaches for matching pairs of images across physically distinct, and visually different scenes (and later generalized to joint image set alignment using cycle consistency [92]). Some approaches such as SCV [11] and MODS [50], were proposed to grow matches around initial warping. In the deep era, [39] showed that ConvNet activation features can be used for correspondence, achieving similar performance to SIFT-Flow. [12] proposed to learn matches with a Correspondence Contrastive loss, producing semi-dense matches. [67] introduced the idea of using 4D convolutions on the feature correlations to learn to filter neighbour consensus. Note that these latter works target semantic correspondences, whereas we focus on the case when all images depict the same physical scene.

Deep Flow methods. Deep networks can be trained to predict optical flow and to be robust to drastic appearance changes, but require adapted loss and architectures. Flows can be learned in a completely supervised way using synthetic data, e.g. in [15,23], but transfer to real data remains a difficult problem. Unsupervised training through reconstruction has been proposed in several works, targeting brightness consistency [2,84], gradient consistency [60] or high SSIM [27,87]. This idea of learning correspondences through reconstruction has been applied to video, reconstructing colors [82], predicting weights for frame reconstruction [32,34], or directly optimizing feature consistency in the warped images [83]. Several papers have introduced cycle consistency as an additional supervisory signal for image alignment [93,83]. Recently, feature correlation became a key part of several architectures [23,77] aiming at predicting dense flows. Particularly relevant to us is the approach of [47] which includes a feature correlation layer in a U-Net [68] architecture to improve flow resolution. A similar approach has been used in [36] which predicts dense correspondences. Recently, Glu-Net [55] learns dense correspondences by investigating the combined use of global and local correlation layers.

Hybrid parametric/non-parametric image alignment. Classic "plane + parallax" approaches [71,33,25,86] aimed to combine parametric and non-parametric alignment by first estimating a homography (plane) and then considering the violations from that homography (parallax). Similar ideas also appeared in stereo, e.g. model-based stereo [13]. Recently, [87,10] proposed to learn optical flow by jointly optimizing with depth and ego-motion for stereo videos. Our RANSAC-Flow is also related to the methods designed for geometric multi-model fitting, such as RPA [45], T-linkage [46] and Progressive-X [7].

3 Method

Our two-stage RANSAC-Flow method is illustrated in Figure 1. In this section, we describe the coarse alignment stage, the fine alignment stage, and how they can be iterated to use multiple homographies.

3.1 Coarse alignment by feature-based RANSAC

Our coarse parametric alignment is performed using RANSAC to fit a homography on a set of candidate sparse correspondences between the source and target images. We use off-the-shelf deep features (conv4 layer of a ResNet-50 network) to obtain these correspondences. We experimented with both pre-trained ImageNet features as well as features learned via MoCo self-supervision [20], and obtained similar results. We found it was crucial to perform feature matching at different scales. We fixed the aspect ratio of each image and extracted features at seven scales: 0.5, 0.6, 0.88, 1, 1.33, 1.66 and 2. Matches that were not symmetrically consistent were discarded. The estimated homography is applied to the source image and the result is given together with the target image as input to our fine alignment. We report coarse-only baselines in Experiments section for both features as "*ImageNet [21]+H*" and "*MoCo [20]+H*".

3.2 Fine alignment by local flow prediction

Given a source image I_s and a target image I_t which have already been coarsely aligned, we want to predict a fine flow $F_{s\to t}$ between them. We write $\mathbf{F}_{s\to t}$ as the mapping function associated to the flow $F_{s\to t}$. Since we only expect the fine alignment to work in image regions where the homography is a good approximation of the deformation, we predict a matchability mask $M_{s\to t}$, indicating which correspondences are valid. In the following, we first present our objective function, then how and why we optimize it using a self-supervised deep network.

Objective function. Our goal is to find a flow that warps the source into an image similar to the target. We formalize this by writing an objective function composed of three parts: a reconstruction loss $\mathcal{L}_{rec}$, a matchability loss $\mathcal{L}_m$ and a cycle-consistency loss $\mathcal{L}_c$. Given the pair of images(I_s, I_t) the total loss is:

$$\mathcal{L}(I_s, I_t) = \mathcal{L}_{rec}(I_s, I_t) + \lambda \mathcal{L}_m(I_s, I_t) + \mu \mathcal{L}_c(I_s, I_t) \tag{1}$$

with λ and μ hyper-parameters weighting the contribution of the matchability and cycle loss. We detail these three components in the following paragraphs. Each loss is defined pixel-wise.

Matchability loss. Our matchability mask can be seen as pixel-wise weights for the reconstruction and cycle-consistency losses. These losses will thus encourage the matchability to be zero. To counteract this effect, the matchability loss encourages the matchability mask to be close to one. Since the matchabiliy should be consistent between images, we define the cycle-consistent matchability at position (x,y) in I_t, (x',y') in I_s with $(x, y) = \mathbf{F}_{s\to t}(x', y')$ as:

$$M_t^{cycle}(x, y) = M_{t\to s}(x, y) M_{s\to t}(x', y') \tag{2}$$

where $M_{s\to t}$ is the matchability predicted from source to target and $M_{t\to s}$ the one predicted from target to source. M_t^{cycle} will be high only if both the matchability of the corresponding pixels in the source and target are high. The matchability loss encourages this cycle-consistent matchability to be close to 1:

$$\mathcal{L}_m(I_s, I_t) = \sum_{(x,y)\in I_t} |M_t^{cycle}(x, y) - 1| \tag{3}$$

Note that directly encouraging the matchability to be 1 leads to similar quantitative results, but using the cycle consistent matchability helps to identify regions that are not matchable in the qualitative results.

Reconstruction loss. Reconstruction is the main term of our objective and is based on the idea that the source image warped with the predicted flow $F_{s\to t}$ should be aligned to the target image I_t. We use the structural similarity (SSIM) [85] as a robust similarity measure:

$$\mathcal{L}_{rec}^{SSIM}(I_s, I_t) = \sum_{(x,y)\in I_t} M_t^{cycle}(x, y) \left(1 - SSIM\left(I_s(x', y'), I_t(x, y)\right)\right) \tag{4}$$

Cycle consistency loss. We enforce cycle consistency of the flow for 2-cycles:

$$\mathcal{L}_c(I_s, I_t) = \sum_{(x,y)\in I_t} M_t^{cycle}(x, y) \|(x', y'), \mathbf{F}_{t\to s}(x, y)\|_2 \tag{5}$$

Optimization with self-supervised network. Optimizing objective functions similar to the one described above is common to most optical flow approaches. However, this is known to be an extremely difficult task because of the highly non-convex nature of the objective which typically has many bad local minima. Recent works on the priors implicit within deep neural network architectures [74,81] suggest that optimizing the flow as the output of a neural network might overcome

these problems. Unfortunately, our objective is still too complex to obtain good result from optimization on just a single image pair. We thus built a larger database of image pairs on which we optimize the neural network parameters in a self-supervised way (i.e. without need for any annotations). The network could then be fine-tuned on the test image pair itself, but we have found that this single-pair optimization leads to unstable results. However, if several pairs similar to the test pair are available (i.e. we have access to the entire test set), the network can be fine-tuned on this test set which leads to some improvement, as can be seen in our experiments where we systematically report our results with and without fine-tuning.

To collect image pairs for the network training, we simply sample pairs of images representing the same scene and applied our coarse matching procedure. If it led to enough inliers, we added the pair to our training image set, if not we discarded it. For all the experiments, we sampled image pairs from the MegaDepth [37] scenes, using $20,000$ image pairs from 100 scenes for training and 500 pairs from 30 different scenes for validation.

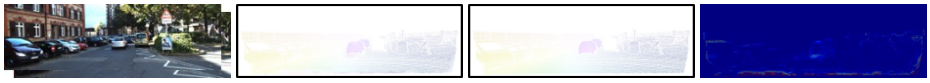
3.3 Multiple Homographies

The overall procedure described so far provides good results on image pairs where a single homography serves as a good (if not perfect) approximation of the overall transformation (e.g. planar scenes). This is, however, not the case for many image pairs with strong 3D effects or large objects displacements. To address this, we iterate our alignment algorithm to let it discover more homography candidates. At each iteration, we remove feature correspondences that were inliers for the previous homographies as well as from locations inside the previously predicted matchability masks, and recompute RANSAC again. We stop the procedure when not enough candidate correspondences remain. The full resulting flow is obtained by simply aggregating the estimated flows from each iteration together. The number of homographies considered depends on the input image pairs. For example, the average number of homographies we obtain from pairs for two-view geometry estimation in the YFCC100M [79] dataset is about five. While more complex combinations could be considered, this simple approach provides surprisingly robust results. In our experiments, we quantitatively validate the benefits of using these multiple homographies ("*multi-H*").

3.4 Architecture and Implementation Details

In our fine-alignment network, the input source and target images (I_s, I_t) are first processed separately by a fully-convolutional *feature extractor* which outputs two feature maps (f_s, f_t). Each feature from the source image is then compared to features in a $(2K+1) \times (2K+1)$ square neighbourhood in the target image using cosine similarity, similar to [15,23]. This results in a $W \times H \times (2K+1)^2$ similarity tensor s defined by:

$$s(i,j,(m+K+1)(n+K+1)) = \frac{f_s(i,j).f_t(i-m,j-n)}{\|f_s(i,j)\|\|f_t(i-m,j-n)\|} \tag{6}$$



(a) Input (b) Predicted (c) Ground truth (d) Error map

Fig. 3: Visual results on KITTI [48]. We show the predicted flow, ground-truth flow and the error map in (b), (c) and (d) respectively.

where $m, n \in [-K, ..., K]$ and "." denotes dot product. In all our experiments, we used $K = 3$. This similarity tensor is taken as input by two fully-convolutional *prediction networks* which predict flow and matchability.

Our *feature extractor* is similar to the *Conv3* feature extractor in ResNet-18 [21] but with minor modifications: the first 7×7 convolutional kernel of the network is replaced by a 3×3 kernel without stride and all the max-poolings and strided-convolution are replaced by their anti-aliasing versions proposed in [89]. These changes aim at reducing the loss of spatial resolution in the network, the output feature map being 1/8th of the resolution of the input images. The flow and matchability *prediction networks* are fully convolutional networks composed of three Conv+Relu+BN blocks (Convolution, Relu activation and Batch Normalization [24]) with 512, 256, 128 filters respectively and a final convolutional layer. The output flows and matchability are bilinearly upsampled to the resolution of the input images. Note we tried using up-convolutions, but this slightly decreased the performance while increasing the memory footprint.

We use Kornia [65] for homography warping. All images were resized so that their minimum dimension is 480 pixels. The hyper-parameters of our objective are set to $\lambda = 0.01$, $\mu = 1$. We provide a study of λ and μ in the supplementary material. The entire fine alignment model is learned from random initialization using the Adam optimizer [31] with a learning rate of 2e-4 and momentum terms β_1, β_2 set to 0.5, 0.999. We trained only with $\mathcal{L}_{rec}$ for the first 150 epochs then added $\mathcal{L}_c$ for another 50 epochs and finally trained with all the losses (Equation 1) for the final 50 epochs. We use a mini-batch size of 16 for all the experiments. The whole training converged in approximately 30 hours using a single GPU Geforce GTX 1080 Ti for the 20k image pairs from the MegaDepth. For fine-tuning on the target dataset, we used a learning rate of 2e-4 for another 10K iterations.

4 Experiments

In this section, we evaluate our approach in terms of resulting correspondences (Sec 4.1), downstream tasks (Sec 4.2), as well as applications to texture transfer and artwork analysis (Sec 4.3). We provide more visual results at http://imagine.enpc.fr/~shenx/RANSAC-Flow/.

4.1 Direct correspondences evaluation

Optical flow. We evaluate the quality of our dense flow on the KITTI 2015 flow [48] and Hpatches [5] datasets and report the results in Table 1.

Table 1: (a) Dense correspondences evaluation on KITTI 2015 [48] and Hpatches [5]. We report the AEE (Average Endpoint Error) and Fl-all (Ratio of pixels where flow estimate is wrong by both 3 pixels and $\geq 5\%$). The computational time for EpicFlow and FlowField is 16s and 23s respectively, while our approach takes 4s. (b) Sparse correspondences evaluation on RobotCar [44,35] and MegaDepth [37]. We report the accuracy over all annotated alignments for pixel error smaller than d pixels. All the images are resized to have minimum dimension 480 pixels.

Method	KITTI 2015 [48] Train (AEE ↓)		KITTI 2015 [48] Test (Fl-all ↓)		Hpatches [5] Viewpoint (AEE ↓)				
	noc	all	noc	all	1	2	3	4	5
Supervised Approaches									
FlowNet2 [23,47,87]	4.93	10.06	6.94	10.41	5.99	15.55	17.09	22.13	30.68
PWC-Net [77,47]	-	10.35	**6.12**	**9.60**	4.43	11.44	15.47	20.17	28.30
Rocco [66,47]	-	-	-	-	9.59	18.55	21.15	27.83	35.19
DGC-Net [47]	-	-	-	-	1.55	5.53	8.98	11.66	16.70
DGC-Nc-Net [36]	-	-	-	-	1.24	4.25	8.21	9.71	13.35
Glu-Net [55]	6.86	9.79	-	-	0.59	4.05	7.64	9.82	14.89
Weakly Supervised Approaches									
ImageNet [21] + H	13.49	17.26	-	-	1.33	3.34	3.71	6.04	10.07
Cao et al. [10]	4.19	5.13	-	-	-	-	-	-	-
Unsupervised Approaches									
Moco [20] + H	13.86	17.60	-	-	1.47	2.96	3.43	7.73	10.53
DeepMatching [63,47]	-	-	-	-	5.84	4.63	12.43	12.17	22.55
DSTFlow [61]	6.96	16.79	-	39	-	-	-	-	-
GeoNet [87]	6.77	10.81	-	-	-	-	-	-	-
EpicFlow [62,87]	4.45	**9.57**	16.69	26.29	-	-	-	-	-
FlowField [4]	-	-	10.98	19.80	-	-	-	-	-
Moco Feature									
Ours	4.15	12.63	14.60	26.16	0.52	2.13	4.83	5.13	6.36
w/o fine-tuning	4.67	13.51	-	-	0.53	**2.04**	**2.32**	6.54	6.79
w/o Multi-H	7.04	14.02	-	-	-	-	-	-	-
ImageNet Feature									
Ours	**3.87**	12.48	14.12	25.76	**0.51**	2.36	2.91	**4.41**	**5.12**
w/o fine-tuning	4.55	13.51	-	-	**0.51**	2.37	2.64	4.49	5.16
w/o Multi-H	6.74	13.77	-	-	-	-	-	-	-

(a) Dense correspondences evaluation on KITTI 2015 [48] and Hpatches [5].

Method	RobotCar [44,35] Acc($\leq$ d pixels ↑)			MegaDepth [37] Acc($\leq$ d pixels ↑)		
	1	3	5	1	3	5
ImageNet [21]+H	1.03	8.12	19.21	3.49	23.48	43.94
Moco [20]+H	1.08	8.77	20.05	3.70	25.12	45.45
SIFT-Flow [38]	1.12	8.13	16.45	8.70	12.19	13.30
NcNet [67]+H	0.81	7.13	16.93	1.98	14.47	32.80
DGC-Net [47]	1.19	9.35	20.17	3.55	20.33	34.28
Glu-Net [55]	**2.16**	**16.77**	**33.38**	25.2	51.0	56.8
Moco Feature						
Ours	2.10	16.07	31.66	**53.47**	83.45	**86.81**
w/o Multi-H	2.06	15.77	31.05	50.65	78.34	81.59
w/o Fine-tuning	2.09	15.94	31.61	52.60	**83.46**	86.80
ImageNet Feature						
Ours	2.10	16.09	31.80	53.15	83.34	86.74
w/o Multi-H	2.06	15.84	31.30	50.08	77.84	81.08
w/o Fine-tuning	2.09	16.00	31.90	52.80	83.31	86.64

(b) Sparse correspondences evaluation on the RobotCar [44,35] and MegaDepth [37].

On KITTI [48], we evaluated both on the training and the test set since other approaches report results on one or the other. Note we could not perform an ablation study on the test set since the number of submissions to the online server is strictly limited. We report results both on non-occluded (noc) and all regions. Our results are on par with state of the art unsupervised and weakly supervised results on non-occluded regions, outperforming for example the recent approach [10,55]. Unsurprisingly, our method is much weaker on occluded regions since our algorithm is not designed specifically for optical flow performances and has no reason to handle occluded regions in a good way. We find that the largest errors are actually in occluded regions and image boundaries (Figure 3). Interestingly, our ablations show that the multiple homographies is critical to our results even if the input images appear quite similar.

For completeness, we also present results on the Hpatches [5]. Note that Hpatches dataset is synthetically created by applying homographies to a set of real images, which would suggest that our coarse alignment alone should be enough. However, in practice, we have found that, due to the lack of feature correspondences, adding the fine flow network significantly boosts the results compared to using only our coarse approach.

While these results show that our approach is reasonable, these datasets only contain very similar and almost aligned pairs while the main goal of our approach

(a) Source (b) Target (c) Coarse align. (d) Fine align. (e) Flows

Fig. 4: Visual results on RobotCar [44] (1st row), Megadepth [37] (2nd row) and Hpatches [5] (3rd row) using one homography. We show the source and target in (a), (b). The overlapped images after coarse and fine alignment are in (c) and (d) with zoomed details. The coarse (top) and fine (bottom) flows are in (e).

is to be able to handle challenging cases with strong viewpoint and appearance variations.

Sparse correspondences. Dense correspondence annotations are typically not available for extreme viewpoint and imaging condition variations. We thus evaluated our results on sparse correspondences available on the RobotCar [44,35] and MegaDepth [37] datasets. In Robotcar, we evaluated on the correspondences provided by [35], which leads to approximately 340M correspondences. The task is especially challenging since the images correspond to different and challenging conditions (dawn, dusk, night, etc.) and most of the correspondences are on texture-less region such as roads where the reconstruction objective provides very little information. However, viewpoints in RobotCar are still very similar. To test our method on pairs of images with very different viewpoints, we used pairs of images from scenes of the MegaDepth [37] dataset that we didn't use for training and validation. Note that no real ground truth is available and we use as reference the result of SfM reconstructions. More precisely, we take 3D points as correspondences and randomly sample 1 600 pairs of images that shared more than 30 points, which results in approximately 367K correspondences.

On both datasets, we evaluated several baselines which provide dense correspondences and were designed to handle large viewpoint changes, inluding SIFT-Flow [38], variants of NcNet [67], DGC-Net [47] and the very recent, concurrently developed Glu-Net [55]. In the results provided in Table 1, we can see that our approach is comparable to Glu-Net on RobotCar [44,35] but largely improves performances on MegeDepth [37]. We believe this is because by the large viewpoint variations on MegeDepth is better handled by our method. This

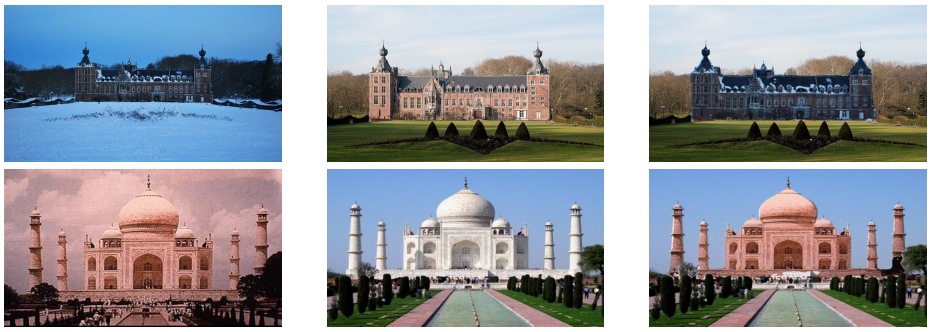
Table 2: (a) Two-view geometric estimation on YFCC100M [79,88]. (b) Visual Localization on Aachen night-time [69,70].

Method	mAP @5°	mAP@10°	mAP@20°
SIFT [40]	46.83	68.03	80.58
Contextdesc [42]	47.68	69.55	**84.30**
Superpoint [14]	30.50	50.83	67.85
PointCN [52,88]	47.98	-	-
PointNet++ [57,88]	46.23	-	-
N^3Net [54,88]	49.13	-	-
DFE [59,88]	49.45	-	-
OANet [88]	52.18	-	-
Moco Feature			
Ours	**64.88**	**73.31**	81.56
w/o multi-H	61.10	70.50	79.24
w/o fine-tuning	63.48	72.93	81.59
ImageNet Feature			
Ous	62.45	70.84	78.99
w/o multi-H	59.90	68.8	77.31
w/o fine-tuning	62.10	70.78	79.07

(a) Two-view geometry, YFCC100M [79]

Method	0.5m,2°	1m,5°	5m,10°
Upright RootSIFT [40]	36.7	54.1	72.5
DenseSfM [69]	39.8	60.2	84.7
HAN + HN++ [49,51]	39.8	61.2	77.6
Superpoint [14]	42.8	57.1	75.5
DELF [53]	39.8	61.2	85.7
D2-net [16]	44.9	66.3	**88.8**
R2D2 [64]	**45.9**	66.3	**88.8**
Moco Feature			
Ours	44.9	**68.4**	**88.8**
w/o Multi-H	42.9	**68.4**	**88.8**
w/o Fine-tuning	41.8	**68.4**	**88.8**
ImageNet Feature			
Ous	44.9	**68.4**	**88.8**
w/o Multi-H	43.9	66.3	**88.8**
w/o Fine-tuning	44.9	**68.4**	**88.8**

(b) Localization, Aachen night-time [69,70]



(a) Source (b) Target (c) Texture transfer

Fig. 5: Texture transfer : (a) source, (b) target and (c) texture transferred result.

qualitative difference between the datasets can be seen in the visual results in Figure 4. Note that we can clearly see the effect of fine flows on the zoomed details.

4.2 Evaluation for downstream tasks.

Given the limitations of the correspondence benchmarks discussed in the previous paragraph, and to demonstrate the practical interest of our results, we now evaluate our correspondences on two standard geometry estimation benchmarks where many results from competing approaches exist. Note that competing approaches typically use only sparse matches for these tasks, and being able to perform them using dense correspondences is a demonstration of the strength and originality of our method.

Two-view geometry estimation. Given a pair of views of the same scene, two-view geometry estimation aims at recovering their relative pose. To validate

(a) Source (b) Coarse align. (c) Fine align. (d) Animation

Fig. 6: Aligning a group of Internet images from the Medici Fountain, similar to [75]. We show the source images (a), the average image after coarse (b) and fine alignment (c). The animation (view with Acrobat Reader) is in (d).

our approach, we follow the standard setup of [88] evaluating on 4×1000 image pairs for 4 scenes from YFCC100M [79] dataset and reporting mAP for different thresholds on the angular differences between ground truth and predicted vectors for both rotation and translation as the error metric. For each image pair, we use the flow we predict in regions with high matchability (> 0.95) to estimate an essential matrix with RANSAC and the 5-point algorithm [19]. To avoid correspondences in the sky, we used the pre-trained the segmentation network provided in [90] to remove them. While this require some supervision, this is reasonable since most of the baselines we compare to have been trained in a supervised way. As can be seen in Table 2, our method outperforms all the baselines by a large margin including the recent OANet [88] method which is trained with ground truth calibration of cameras. Also note that using multiple homographies consistently boosts the performance of our method.

Once the relative pose of the cameras has been estimated, our correspondences can be used to perform stereo reconstruction from the image pair as illustrated in Figure 2(c) and in the project webpage. Note that contrary to many stereo reconstruction methods, we can use two very different input images.

Day-Night Visual Localization. Another task we performed is visual localization. We evaluate on the local feature challenge of the Visual Localization benchmark [69,70]. For each of the 98 night-time images contained in the dataset, up to 20 relevant day-time images with known camera poses are given. We followed evaluation protocol from [69] and first compute image matching for a list of image pairs and then give them as input to COLMAP [72] that provides a localisation estimation for the queries. To limit the number of correspondences we use only correspondences on a sparse set of keypoints using the Superpoint [14]. Our results are reported in Table 2(b) and are on par with state of the art results.

4.3 Applications

One of the most exciting aspect of our approach is that it enables new applications based on the fine alignment of historical, internet or artistic images.

(a) Inputs (b) W/o align. (c) Coarse align. (d) Fine align. (e) Flows

Fig. 7: Aligning pairs of similar artworks from the Brueghel [1]: We show the pairs in (a). The average images without alignment, after coarse and fine alignment are in (b), (c) and (d). The coarse (top) and fine (bottom) flows are in (e).

Texture transfer. Our approach can be used to transfer texture between images. In Figure 5 and 2(f) we show results using historical and modern images from the LTLL dataset [17]. We use the pre-trained segmentation network of [91], and transfer the texture from the source to the target building regions.

Internet images alignment. As visualized in Figures 2(d) and 6, we can align sets of internet images, similar to [75]. Even if our image set is not precisely the same, much more details can be seen in the average of our fine-aligned images.

Artwork analysis. Finding and matching near-duplicate patterns is an important problem for art historians. Computationally, it is difficult because the duplicate appearance can be very different [73]. In Figure 7, we show visual results of aligning different versions of artworks from the Brueghel dataset [73] with our coarse and fine alignment. We can clearly see that a simple homography is not sufficient and that the fine alignment improves results by identifying complex displacements. The fine flow can thus be used to provide insights on Brueghel's copy process. Indeed, we found that some artworks were copied in a spatially consistent way, while in others, different parts of the picture were not aligned with each other. This can be clearly seen in the flows in Figure 9, which are either very regular or very discontinuous. The same process can be applied to more than a single pair of images, as illustrated in Figure 2(e) and 8 where we align together many similar details identified by [73]. Visualizing the succession of the finely aligned images allows to identify their differences.

(a) Source (b) [73] (c) Coarse align. (d) Fine align. (e) Animation

Fig. 8: Aligning details discovered by [73]: (a) sources; average from [73] (b), with coarse (c) and fine (d) alignment; (e) animation (view with Acrobat Reader).

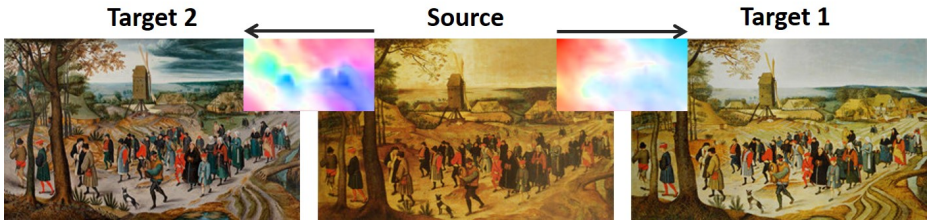


Fig. 9: Analyzing copy process from flow. The flow is smooth from the middle to the right one, while it is irregular from the middle to the left one.

5 Conclusion

We have introduced a new unsupervised method for generic dense image alignment which performs well on a wide range of tasks. Our main insight is to combine the advantages of parametric and non-parametric methods in a two-stage approach and to use multiple homography estimations as initializations for fine flow prediction. We also demonstrated it allows new applications for artwork analysis.

Acknowledgements: This work was supported by ANR project EnHerit ANR-17-CE23-0008, project Rapid Tabasco, NSF IIS-1633310, grants from SAP and Berkeley CLTC, and gifts from Adobe. We thank Shiry Ginosar, Thibault Groueix and Michal Irani for helpful discussions, and Elizabeth Alice Honig for her help in building the Brueghel dataset.

References

1. Brueghel family: Jan brueghel the elder." the brueghel family database. university of california, berkeley. http://www.janbrueghel.net/, accessed: 2018-10-16
2. Ahmadi, A., Patras, I.: Unsupervised convolutional neural networks for motion estimation. In: International Conference on Image Processing (2016)
3. Aubry, M., Russell, B.C., Sivic, J.: Painting-to-3d model alignment via discriminative visual elements. ACM Transactions on Graphics (ToG) (2014)
4. Bailer, C., Taetz, B., Stricker, D.: Flow fields: Dense correspondence fields for highly accurate large displacement optical flow estimation. In: Proceedings of the IEEE International Conference on Computer Vision (2015)
5. Balntas, V., Lenc, K., Vedaldi, A., Mikolajczyk, K.: Hpatches: A benchmark and evaluation of handcrafted and learned local descriptors. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2017)
6. Barath, D., Matas, J.: Graph-cut ransac. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2018)
7. Barath, D., Matas, J.: Progressive-x: Efficient, anytime, multi-model fitting algorithm. In: Proceedings of the IEEE International Conference on Computer Vision (2019)
8. Barath, D., Matas, J., Noskova, J.: Magsac: marginalizing sample consensus. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2019)
9. Brox, T., Bregler, C., Malik, J.: Large displacement optical flow. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2009)
10. Cao, Z., Kar, A., Hane, C., Malik, J.: Learning independent object motion from unlabelled stereoscopic videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2019)
11. Cech, J., Matas, J., Perdoch, M.: Efficient sequential correspondence selection by cosegmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence (2010)
12. Choy, C.B., Gwak, J., Savarese, S., Chandraker, M.: Universal correspondence network. In: Advances in Neural Information Processing Systems (2016)
13. Debevec, P.E., Taylor, C.J., Malik, J.: Modeling and rendering architecture from photographs: A hybrid geometry-and image-based approach. In: Proceedings of the 23rd annual conference on Computer graphics and interactive techniques (1996)
14. DeTone, D., Malisiewicz, T., Rabinovich, A.: Superpoint: Self-supervised interest point detection and description. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (2018)
15. Dosovitskiy, A., Fischer, P., Ilg, E., Hausser, P., Hazirbas, C., Golkov, V., Van Der Smagt, P., Cremers, D., Brox, T.: Flownet: Learning optical flow with convolutional networks. In: Proceedings of the IEEE international conference on computer vision (2015)
16. Dusmanu, M., Rocco, I., Pajdla, T., Pollefeys, M., Sivic, J., Torii, A., Sattler, T.: D2-net: A trainable cnn for joint description and detection of local features. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2019)
17. Fernando, B., Tommasi, T., Tuytelaars, T.: Location recognition over large time lags. Computer Vision and Image Understanding (2015)
18. Fischler, M.A., Bolles, R.C.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM (1981)

19. Hartley, R., Zisserman, A.: Multiple view geometry in computer vision. Cambridge university press (2003)
20. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2020)
21. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (2016)
22. Hu, Y., Song, R., Li, Y.: Efficient coarse-to-fine patchmatch for large displacement optical flow. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2016)
23. Ilg, E., Mayer, N., Saikia, T., Keuper, M., Dosovitskiy, A., Brox, T.: Flownet 2.0: Evolution of optical flow estimation with deep networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2017)
24. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: International Conference on Machine Learning (2015)
25. Irani, M., Anandan, P., Cohen, M.: Direct recovery of planar-parallax from multiple frames. IEEE Transactions on Pattern Analysis and Machine Intelligence (2002)
26. Jaderberg, M., Simonyan, K., Zisserman, A., et al.: Spatial transformer networks. In: Advances in Neural Information Processing Systems (2015)
27. Jason, J.Y., Harley, A.W., Derpanis, K.G.: Back to basics: Unsupervised learning of optical flow via brightness constancy and motion smoothness. In: Proceedings of the European Conference on Computer Vision (2016)
28. Kim, S., Lin, S., JEON, S.R., Min, D., Sohn, K.: Recurrent transformer networks for semantic correspondence. In: Advances in Neural Information Processing Systems (2018)
29. Kim, S., Min, D., Ham, B., Jeon, S., Lin, S., Sohn, K.: Fcss: Fully convolutional self-similarity for dense semantic correspondence. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2017)
30. Kim, S., Min, D., Jeong, S., Kim, S., Jeon, S., Sohn, K.: Semantic attribute matching networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2019)
31. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: International Conference for Learning Representations (2014)
32. Kong, S., Fowlkes, C.: Multigrid predictive filter flow for unsupervised learning on videos. arXiv preprint arXiv:1904.01693 (2019)
33. Kumar, R., Anandan, P., Hanna, K.: Direct recovery of shape from multiple views: A parallax based approach. In: Proceedings of 12th International Conference on Pattern Recognition (1994)
34. Lai, Z., Xie, W.: Self-supervised learning for video correspondence flow. In: BMVC (2019)
35. Larsson, M., Stenborg, E., Hammarstrand, L., Pollefeys, M., Sattler, T., Kahl, F.: A cross-season correspondence dataset for robust semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2019)
36. Laskar, Z., Melekhov, I., Tavakoli, H.R., Ylioinas, J., Kannala, J.: Geometric image correspondence verification by dense pixel matching. In: Winter Conference on Applications of Computer Vision (2020)

37. Li, Z., Snavely, N.: Megadepth: Learning single-view depth prediction from internet photos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2018)
38. Liu, C., Yuen, J., Torralba, A.: Sift flow: Dense correspondence across scenes and its applications. IEEE Transactions on Pattern Analysis and Machine Intelligence (2010)
39. Long, J.L., Zhang, N., Darrell, T.: Do convnets learn correspondence? In: Advances in neural information processing systems (2014)
40. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision (2004)
41. Lucas, B.D., Kanade, T., et al.: An iterative image registration technique with an application to stereo vision (1981)
42. Luo, Z., Shen, T., Zhou, L., Zhang, J., Yao, Y., Li, S., Fang, T., Quan, L.: Contextdesc: Local descriptor augmentation with cross-modality context. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2019)
43. Luo, Z., Shen, T., Zhou, L., Zhu, S., Zhang, R., Yao, Y., Fang, T., Quan, L.: Geodesc: Learning local descriptors by integrating geometry constraints. In: Proceedings of the European Conference on Computer Vision (2018)
44. Maddern, W., Pascoe, G., Linegar, C., Newman, P.: 1 year, 1000 km: The oxford robotcar dataset. The International Journal of Robotics Research (2017)
45. Magri, L., Fusiello, A.: T-linkage: A continuous relaxation of j-linkage for multi-model fitting. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2014)
46. Magri, L., Fusiello, A.: Multiple structure recovery via robust preference analysis. Image and Vision Computing (2017)
47. Melekhov, I., Tiulpin, A., Sattler, T., Pollefeys, M., Rahtu, E., Kannala, J.: Dgc-net: Dense geometric correspondence network. In: Winter Conference on Applications of Computer Vision (2019)
48. Menze, M., Geiger, A.: Object scene flow for autonomous vehicles. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2015)
49. Mishchuk, A., Mishkin, D., Radenovic, F., Matas, J.: Working hard to know your neighbor's margins: Local descriptor learning loss. In: Advances in Neural Information Processing Systems (2017)
50. Mishkin, D., Matas, J., Perdoch, M.: Mods: Fast and robust method for two-view matching. Computer Vision and Image Understanding (2015)
51. Mishkin, D., Radenovic, F., Matas, J.: Repeatability is not enough: Learning affine regions via discriminability. In: Proceedings of the European Conference on Computer Vision (2018)
52. Moo Yi, K., Trulls, E., Ono, Y., Lepetit, V., Salzmann, M., Fua, P.: Learning to find good correspondences. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2018)
53. Noh, H., Araujo, A., Sim, J., Weyand, T., Han, B.: Large-scale image retrieval with attentive deep local features. In: Proceedings of the IEEE International Conference on Computer Vision (2017)
54. Plötz, T., Roth, S.: Neural nearest neighbors networks. In: Advances in Neural Information Processing Systems (2018)
55. Prune, T., Martin, D., Radu, T.: GLU-Net: Global-local universal network for dense flow and correspondences. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2020)

56. Qi, C.R., Su, H., Mo, K., Guibas, L.J.: Pointnet: Deep learning on point sets for 3d classification and segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2017)
57. Qi, C.R., Yi, L., Su, H., Guibas, L.J.: Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In: Advances in Neural Information Processing Systems (2017)
58. Raguram, R., Chum, O., Pollefeys, M., Matas, J., Frahm, J.M.: Usac: a universal framework for random sample consensus. IEEE Transactions on Pattern Analysis and Machine Intelligence (2012)
59. Ranftl, R., Koltun, V.: Deep fundamental matrix estimation. In: Proceedings of the European Conference on Computer Vision (2018)
60. Ren, Z., Yan, J., Ni, B., Liu, B., Yang, X., Zha, H.: Unsupervised deep learning for optical flow estimation. In: Thirty-First AAAI Conference on Artificial Intelligence (2017)
61. Ren, Z., Yan, J., Ni, B., Liu, B., Yang, X., Zha, H.: Unsupervised deep learning for optical flow estimation. In: Thirty-First AAAI Conference on Artificial Intelligence (2017)
62. Revaud, J., Weinzaepfel, P., Harchaoui, Z., Schmid, C.: Epicflow: Edge-preserving interpolation of correspondences for optical flow. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2015)
63. Revaud, J., Weinzaepfel, P., Harchaoui, Z., Schmid, C.: Deepmatching: Hierarchical deformable dense matching. International Journal of Computer Vision (2016)
64. Revaud, J., Weinzaepfel, P., de Souza, C.R., Humenberger, M.: R2D2: repeatable and reliable detector and descriptor. In: Advances in Neural Information Processing Systems (2019)
65. Riba, E., Mishkin, D., Ponsa, D., Rublee, E., Bradski, G.: Kornia: an open source differentiable computer vision library for pytorch. In: Winter Conference on Applications of Computer Vision (2020)
66. Rocco, I., Arandjelovic, R., Sivic, J.: Convolutional neural network architecture for geometric matching. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2017)
67. Rocco, I., Cimpoi, M., Arandjelović, R., Torii, A., Pajdla, T., Sivic, J.: Neighbourhood consensus networks. In: Advances in Neural Information Processing Systems (2018)
68. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention (2015)
69. Sattler, T., Maddern, W., Toft, C., Torii, A., Hammarstrand, L., Stenborg, E., Safari, D., Okutomi, M., Pollefeys, M., Sivic, J., et al.: Benchmarking 6dof outdoor visual localization in changing conditions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2018)
70. Sattler, T., Weyand, T., Leibe, B., Kobbelt, L.: Image retrieval for image-based localization revisited. In: BMVC (2012)
71. Sawhney, H.S.: 3d geometry from planar parallax. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (1994)
72. Schonberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2016)
73. Shen, X., Efros, A.A., Aubry, M.: Discovering visual patterns in art collections with spatially-consistent feature learning. In: Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (2019)

74. Shocher, A., Cohen, N., Irani, M.: "zero-shot" super-resolution using deep internal learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2018)
75. Shrivastava, A., Malisiewicz, T., Gupta, A., Efros, A.A.: Data-driven visual similarity for cross-domain image matching. In: Proceedings of the 2011 SIGGRAPH Asia Conference (2011)
76. Singh, S., Gupta, A., Efros, A.A.: Unsupervised discovery of mid-level discriminative patches. In: Proceedings of the European Conference on Computer Vision (2012)
77. Sun, D., Yang, X., Liu, M.Y., Kautz, J.: Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2018)
78. Szeliski, R.: Image alignment and stitching: A tutorial. Found. Trends. Comput. Graph. Vis. (2006)
79. Thomee, B., Shamma, D.A., Friedland, G., Elizalde, B., Ni, K., Poland, D., Borth, D., Li, L.J.: Yfcc100m: The new data in multimedia research. Communications of the ACM (2016)
80. Tian, Y., Fan, B., Wu, F.: L2-net: Deep learning of discriminative patch descriptor in euclidean space. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2017)
81. Ulyanov, D., Vedaldi, A., Lempitsky, V.: Deep image prior. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2018)
82. Vondrick, C., Shrivastava, A., Fathi, A., Guadarrama, S., Murphy, K.: Tracking emerges by colorizing videos. In: Proceedings of the European Conference on Computer Vision (2018)
83. Wang, X., Jabri, A., Efros, A.A.: Learning correspondence from the cycle-consistency of time. In: The IEEE Conference on Computer Vision and Pattern Recognition (2019)
84. Wang, Y., Yang, Y., Yang, Z., Zhao, L., Wang, P., Xu, W.: Occlusion aware unsupervised learning of optical flow. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2018)
85. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing (2004)
86. Wulff, J., Sevilla-Lara, L., Black, M.J.: Optical flow in mostly rigid scenes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2017)
87. Yin, Z., Shi, J.: Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2018)
88. Zhang, J., Sun, D., Luo, Z., Yao, A., Zhou, L., Shen, T., Chen, Y., Quan, L., Liao, H.: Learning two-view correspondences and geometry using order-aware network. Proceedings of the IEEE International Conference on Computer Vision (2019)
89. Zhang, R.: Making convolutional networks shift-invariant again. In: International Conference on Machine Learning (2019)
90. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ade20k dataset. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2017)
91. Zhou, B., Zhao, H., Puig, X., Xiao, T., Fidler, S., Barriuso, A., Torralba, A.: Semantic understanding of scenes through the ade20k dataset. International Journal on Computer Vision (2018)

92. Zhou, T., Jae Lee, Y., Yu, S.X., Efros, A.A.: Flowweb: Joint image set alignment by weaving consistent, pixel-wise correspondences. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2015)
93. Zhou, T., Krahenbuhl, P., Aubry, M., Huang, Q., Efros, A.A.: Learning dense correspondence via 3d-guided cycle consistency. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2016)