RANSAC-Flow: generic two-stage image alignment

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

[1] LIGM (UMR 8049) - Ecole des Ponts, UPE
[2] 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/.

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 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.

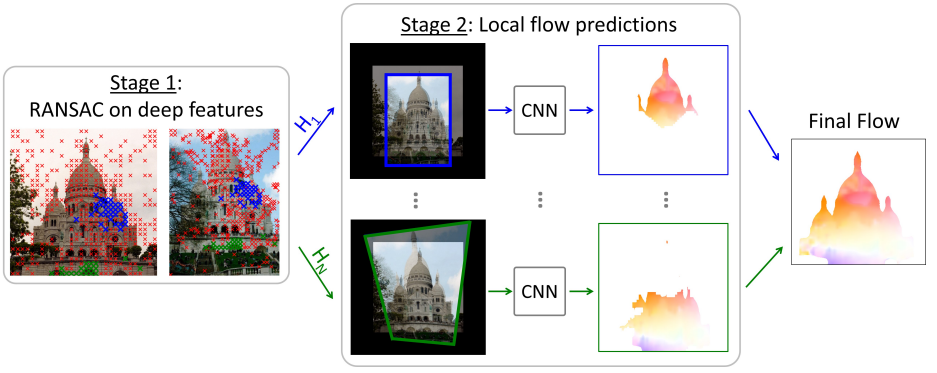


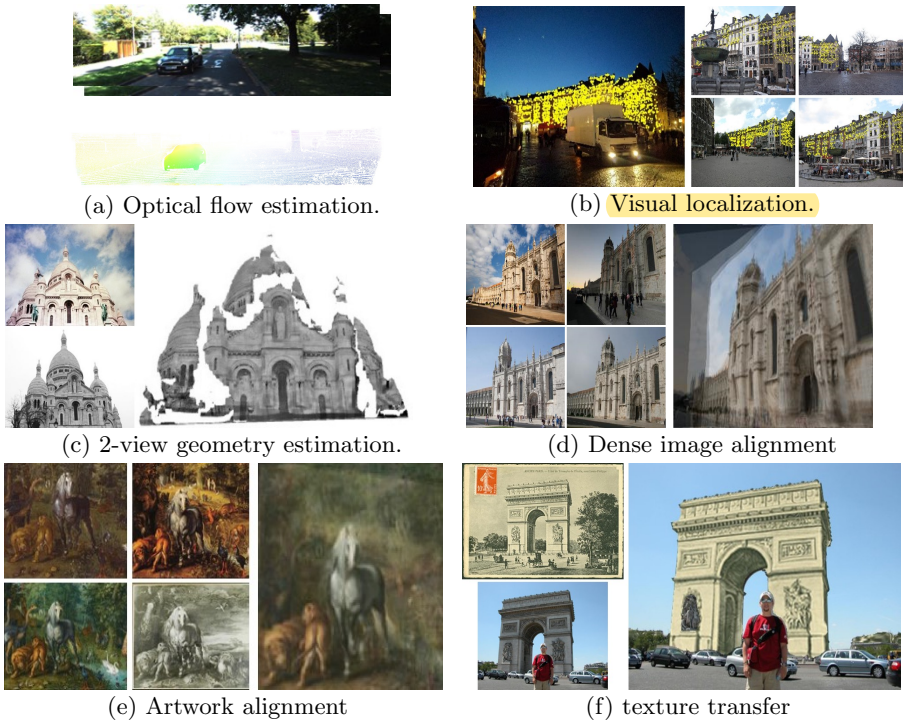
Fig. 1: **Overview of RANSAC-Flow.** Stage 1: given a pair of images, we compute sparse correspondences (using off-the-shelf deep features), use RANSAC to estimate a homography, and warp second image using it. Stage 2: given two coarsely aligned images, our self-supervised fine flow network generates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

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) [76] between the pixels of the warped and the original images, plus a cycle-consistency loss [84].

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 [42] and HPatches [5], (b) visual localization on Aachen Day-Night [60], (c) 2-view geometry estimation on YFCC100M [70], (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 [35],



(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 enable new challenging applications.

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 [16] to fit a simple geometric transformation (e.g. affine or homography) [69]. Recently, many works proposed to learn better local features [37,12,71,43,38,54]. Differentiable and trainable version of RANSAC have also been developed [79,48,47,50].

Using mid-level features [67] instead of local keypoints, proved to be beneficial for matching visual content across modalities, e.g. 3D models and paintings [3]. Recently, [64] 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. [57] 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 [24] 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 [36], 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 [69], as well as using local features in addition to pixels to increase robustness [6,55,4,20]. However, all such methods are still limited to aligning very similar images, where the brightness constancy assumption mostly holds. SIFT-Flow [33] 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 [83]). In the deep era, [34] showed that ConvNet activation features can be used for correspondence, achieving similar performance to SIFT-Flow. [9] proposed to learn matches with a Correspondence Contrastive loss, producing semi-dense matches. [58] 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 [13,21], but transfer to real data remains a difficult problem. Unsupervised training through reconstruction has been proposed in several works, targeting brightness consistency [2,75], gradient consistency [52] or high SSIM [25,78]. This idea of learning correspondences through reconstruction has been applied to video, reconstructing colors [73], predicting weights for frame reconstruction [27,29], or directly optimizing feature consistency in the warped images [74]. Several papers have introduced cycle consistency as an additional supervisory signal for image alignment [84,74]. Recently, feature correlation became a key part of several architectures [21,68] aiming at predicting dense flows. Particularly relevant to us is the approach of [41] which includes a feature correlation layer in a U-Net [59] architecture to improve flow resolution. A similar approach has been used in [31] which predicts dense correspondences for image retrieval.

Hybrid parametric/non-parametric image alignment. Classic "plane + parallax" approaches [62,28,23] 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 [10]. Recently, [78,8] proposed to learn optical flow by jointly optimizing with depth and ego-motion for stereo videos.

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 [18], and obtained similar results. To obtain good results, we found it was crucial to perform feature matching at different scales. We kept the images aspect ratio and resized them with 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 [19]+H*" and "*MoCo [18]+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 deformation flow $F_{I_s \to I_t}$ between them. We write $\mathcal{F}_{I_s \to I_t}$ as the mapping function associated to the flow $F_{I_s \to I_t}$. Since we only expect the fine alignment to work in image regions where the homoraphy is a good approximation of the deformation, we also want to predict a matchability mask $M_{I_s \to I_t}$, 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 each of these three components in the following paragraphs. Each loss is defined pixel-wise and $\cdot$ denotes the element-wise multiplication.

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 $\mathcal{F}_{I_s \to I_t}(I_s)$ should be aligned to the target image I_t. We use the structural similarity (SSIM) [76] as a robust similarity measure:

$$\mathcal{L}_{rec}^{SSIM}(I_s, I_t) = (1 - SSIM\left(\mathcal{F}_{I_s \to I_t}(I_s), I_t\right)) \cdot M_{I_t} \tag{2}$$

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

$$\mathcal{L}_c(I_s, I_t) = \|\mathcal{F}_{I_t \to I_s}(F_{I_t \to I_s}) + F_{I_s \to I_t}\| \cdot M_{I_t} \tag{3}$$

Note we also experimented with cycle consistency between triplets of images, but this overly constrains the training data since triplets of matching images are required and the results didn't lead to significant improvements.

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 in I_s as:

$$M^{cycle}_{I_s \to I_t} = \mathcal{F}_{I_t \to I_s}(M_{I_t \to I_s}) \cdot M_{I_s \to I_t} \tag{4}$$

where $M_{I_s \to I_t}$ is the matchability predicted from source to target, $M_{I_t \to I_s}$ the one predicted from target to source and $\mathcal{F}_{I_t \to I_s}$ the map associated to the flow predicted from target to source. $M^{cycle}_{I_s \to I_t}$ 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:

$$\mathcal{L}_m(I_s, I_t) = \|M^{cycle}_{I_s \to I_t} - 1\| \tag{5}$$

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.

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 [65,72] 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 lead 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 [32] 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 [70] dataset is about five. While more complex combinations could be considered, this simple approach provides surprisingly robust results. In our experiments, we validate quantitatively the benefits of using these multiple homographies ("*multi-H* ").

3.4 Architecture and Implementation Details

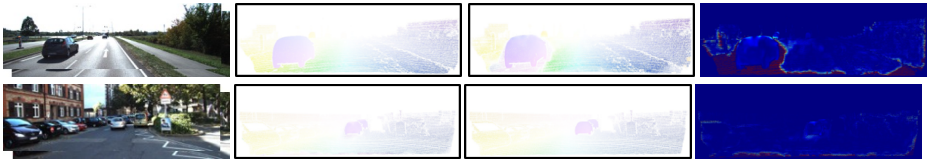
Architecture In our fine-alignment network, the input source and target images (I_s, I_t) are first processed separately by a fully-convolutional *feature extractor* which outputs two feature maps (f_s, f_t). Each feature from the source image is then compared to features in a $(2K+1) \times (2K+1)$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

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

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 $F_{I_s \to I_t}$ and matchability $M_{I_s \to I_t}$.

Our *feature extractor* is similar to the *Conv3* feature extractor in ResNet-18 [19] 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-aliazing version proposed in [80]. 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 convolutionnal networks composed of three Conv+Relu+BN blocks (Convolution, Relu activation and Batch Normalization [22]) 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.

For image pair selection and training, all images were resized so that their minimum dimension is 480. The hyper-parameters of our objective are set to $\lambda = 0.01$, $\mu = 1$. The entire fine alignment model is learned from random initialization using the Adam optimizer [26] 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)



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

Fig. 3: Visual results on KITTI [42]. We show the predicted flow, ground-truth flow and the error map in (b), (c) and (d) respectively.

Table 1: Dense correspondences evaluation on Hpatches and KITTI 2015. We report the AEE (Average Endpoint Error, lower is better) and Fl-all (Ratio of pixels where flow estimate is wrong by both 3 pixels and $\geq 5\%$, lower is better). Note that the computational time for EpicFlow and FlowField is 16s and 23s respectively, while our approach takes approximately 4s. Supervised optical flow methods are trained on the FlyingChairs [13], FlyingThings3D [40] and Sintel [7] for Hpatches and KITTI train. Cao et al.[8] uses 2D object bounding boxes annotations.

Method	Viewpoint (AEE) 1	2	3	4	5
Supervised Approaches					
SPyNet [51,41]	36.94	50.92	54.29	62.60	72.57
FlowNet2 [21,41]	5.99	15.55	17.09	22.13	30.68
PWC-Net [68,41]	4.43	11.44	15.47	20.17	28.30
Rocco [57,41]	9.59	18.55	21.15	27.83	35.19
DGC-Net [41]	1.55	5.53	8.98	11.66	16.70
DGC-Nc-Net [31]	1.24	4.25	8.21	9.71	13.35
Weakly Supervised Approaches					
ImageNet [19] + H	1.33	3.34	3.71	6.04	10.07
Unsupervised Approaches					
Moco [18] + H	1.47	2.96	3.43	7.73	10.53
DeepMatching [56,41]	5.84	4.63	12.43	12.17	22.55
Moco Feature					
Ours	0.52	2.13	4.83	5.13	6.36
w/o fine-tuning	0.53	2.04	2.32	6.54	6.79
ImageNet Feature					
Ours	0.51	2.36	2.91	4.41	5.12
w/o fine-tuning	0.51	2.37	2.64	4.49	5.16

(a) Hpatches [5]

Method	Train (AEE) noc	all	Test (Fl-all) noc	all
Supervised Approaches				
SPyNet [51]	-	-	26.71	35.07
FlowNet2 [21,78]	4.93	10.06	6.94	10.41
FlowNetS [13,78]	8.12	14.19	-	-
PWC-Net [68]	-	10.35	6.12	9.60
Weakly Supervised Approaches				
ImageNet[19] + H	13.49	17.26	-	-
Cao et al. [8]	4.19	5.13	-	-
Unsupervised Approaches				
Moco[18] + H	13.86	17.60	-	-
DSTFlow [53]	6.96	16.79	-	39
OAFlow [75]	-	8.8	-	31.2
GeoNet [78]	6.77	10.81	-	-
DeepFlow [77]	-	-	18.35	28.48
EpicFlow [55,78]	4.45	9.57	16.69	26.29
CPMFlow [20]	-	-	13.85	22.40
FlowField [4]	-	-	10.98	19.80
Moco Feature				
Ours	4.15	12.63	14.60	26.16
w/o fine-tuning	4.67	13.51	-	-
w/o Multi-H	7.04	14.02	-	-
ImageNet Feature				
Ours	3.87	12.48	14.12	25.76
w/o fine-tuning	4.55	13.51	-	-
w/o Multi-H	6.74	13.77	-	-

(b)KITTI 2015 [42]

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 dataset. 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).

4.1 Direct correspondences evaluation

Optical flow. We evaluate the quality of our dense flow on the Hpatches [5] and KITTI 2015 flow [42] datasets and report the results in Table 1.

(a) Souce (b) Target (c) Coarse alignment (d) Fine alignment (e) Flows

Fig. 4: Visual results on RobotCar [39] (first line), Megadepth [32] (second line) and Hpatches [5] (third line) using a single homography. We show the source, target in (a) and (b). The results of (c) coarse and (d) fine alignment are represented as overlapped images with zoomed details and flow maps. The coarse (top) and fine (bottom) flows are visualized in (e)

Hpatches is especially adapted to our method since it consists of planar scenes and our hypothesis of using homography approximation of the alignment is thus valid, which explains that we obtain very strong results outperforming any baseline method by a clear margin. Note however that adding the fine flow network significantly boosts the results compared to using only our coarse approach (" *ImageNet [19] +H*" and " *MoCo [18] +H*"). This can be understood by looking at the qualitative results in Figure 4 (last raw): the fine flow predicts a flow very similar to an homography which really improves alignment. Note that since the scenes are planar we always obtain a single homography for coarse alignment. Also note that fine-tunning on the HPatches images makes little difference in these results.

On KITTI, we evaluated both on the 200 training pairs and on the test set since other approaches report results on one or the other. Note we could not perform 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 of [8]. Unsurprisingly, our method is much weaker on occluded regions since our algorithm is not desgined specifically for optical flow performances and has no reason to handle occluded regions in a good way. From the qualitative results in Figure 3, we can see the largest errors are actually in occluded regions and near the occlusion and image boundaries. Interestingly, our ablations show that the use of multiple homographies is critical to the quality of our results even if the input images appear quite similar.

Table 2: Evaluation of sparse correspondences on the RobotCar [39,30] and MegaDepth [32] dataset. We report the accuracy over all annotated alignments (∼340M correspondences for RobotCar and ∼367K for MegaDepth) for pixel error smaller than d pixels with d = 1, 3, 5. Note that, all the test images are with the original aspect ratio and resized to have minimum dimension 480 pixels.

Method	Test Acc(%, ≤ d pixels)					
	RobotCar			MegaDepth		
	d = 1	d = 3	d = 5	d = 1	d = 3	d = 5
ImageNet [19]+H	1.03	8.12	19.21	3.49	23.48	43.94
Moco [18]+H	1.08	8.77	20.05	3.70	25.12	**45.45**
SIFT-Flow[3] [33]	1.12	8.13	16.45	**8.70**	12.19	13.30
NcNet [58]+Homography	0.81	7.13	16.93	1.98	14.47	32.80
NcNet [58]+Bilinear	0.61	5.32	14.04	1.09	9.42	24.04
DGC-Net [41]	**1.53**	**11.80**	**24.92**	8.29	**31.94**	43.35
DGC+M-Net [41]	1.19	9.35	20.17	3.55	20.33	34.28
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

While this results demonstrate that our approach is reasonable, these datasets only contain very similar and almost aligned pairs while the main goal of our approach 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 [39,30] and MegaDepth [32] datasets. In Robotcar, we evaluated on all the 6 511 pairs images the correspondences provided by [30], 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 [32] 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 [33], variants of NcNet [58] and DGC-Net [41]. In the results provided in Table 2, we can see that our approach consistently improves performances by a large margin on both datasets. Using multiple homographies provides a much clearer boost on MegaDepth [32] than on RobotCar [39], which can be explained by the large viewpoint variations on this dataset. This qualitative difference between the datasets can be seen in the visual results using a single homography

(a) Query (b) Matched images (c) Localization

Fig. 5: Visual results of Aachen day-night dataset. We show the 3D points in yellow in (a) and (b), the localisation map with all the camera in red in (c).

Table 3: (a) Two-view geometric estimation for 4 scenes from YFCC100M [70] dataset. (b) Results on Visual Localization on Aachen nighttime [60,61]. The benchmark is referred from R2D2 [54].

Method	mAP@5°	mAP@10°	mAP@20°
SIFT [35]	46.83	68.03	80.58
Contextdesc [37]	47.68	**69.55**	**84.30**
Superpoint [12]	30.50	50.83	67.85
PointCN [45,79]	47.98	-	-
PointNet++ [49,79]	46.23	-	-
$\mathcal{N}^3$Net [47,79]	49.13	-	-
DFE [50,79]	49.45	-	-
OANet [79]	**52.18**	-	-
Moco Feature			
Ours	**64.88**	81.75	91.33
w/o multi-H	61.10	79.90	89.58
w/o fine-tuning	63.48	**82.40**	**91.63**
ImageNet Feature			
Ous	**62.45**	78.38	88.58
w/o multi-H	59.90	77.70	87.23
w/o fine-tuning	62.10	**79.45**	**88.75**

(a) Two-view geometry, YFCC100M [70]

Method	0.5m,2°	1m,5°	5m,10°
Upright RootSIFT [35]	36.7	54.1	72.5
DenseSfM [60]	39.8	60.2	84.7
HAN + HN++ [43,44]	39.8	61.2	77.6
Superpoint [12]	42.8	57.1	75.5
DELF [46]	39.8	61.2	85.7
D2-net [14]	44.9	66.3	**88.8**
R2D2 [54]	**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**

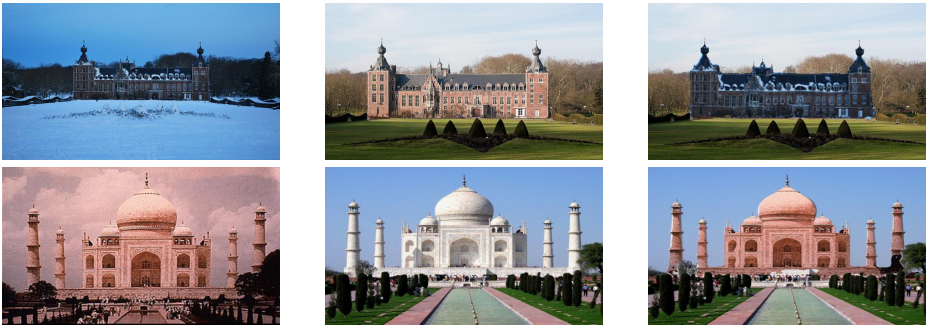
(b) Localization, Aachen night-time [60,61]

provided in Figure 4 (first and second rows). Note that we can clearly see the effect of fine flows on the zoomed overlapped images.

4.2 Evaluation for downstream tasks.

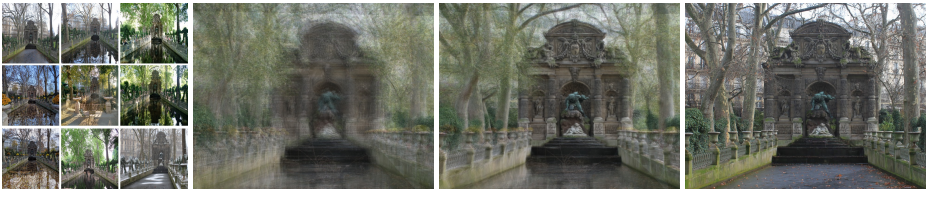
Given the limitations of the correspondence benchmarks discussed in the previous paragraph, and to demonstrate the practical interest of our results, we now evaluate our correspondences on two standard geometry estimation benchmarks where many results from competing approaches exist. Note that competing 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 our approach, we follow the standard setup of [79] evaluating on 4×1000 image pairs for 4 scenes from YFCC100M [70] 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 [17]. To avoid



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

Fig. 6: Texture transfer. We show the (a) source, (b) target and (c) texture transferred result from source to target.



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

Fig. 7: Aligning a group of Internet images from the Medici Fountain, similar to [66]. (a) source images; (b) average after coarse alignment (homography); (c) average after fine alignment, please zoom to see the high quality details; (d) animation (view with Acrobat Reader).

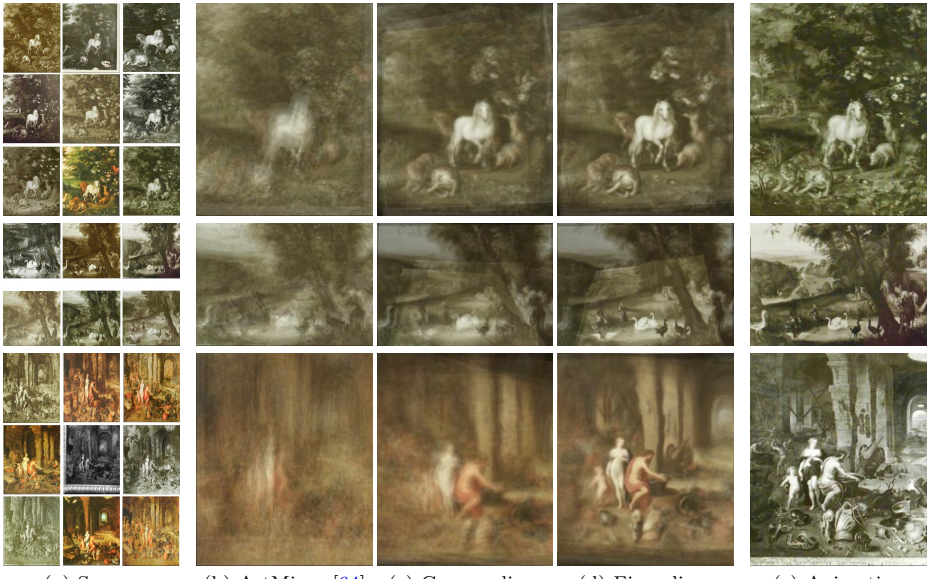
correspondences in the sky, we used the pre-trained the segmentation network provided in [81] to remove them. While this require some supervision, this is reasonable since most of the baselines we compare to have been trained in a supervised way. As can be seen in Table 3, our method outperforms all the baselines by a large margin including the recent OANet [79] method which is trained with ground truth calibration of cameras: our mAP at 5° is 64.88%, improving the performance on this benchmark by more than 12%. Interestingly, we found that the self-supervised MoCo [18] feature allows us to achieve better performance than features trained on ImageNet [11] classification. 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 [60,61]. 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 [60] and first compute image matching for a

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

Fig. 8: Aligning pairs of similar artworks from the Brueghel collection [1]: (a) source and target; (b) average w/o alignment; (c) average after coarse alignment (homography); (d) average after fine alignment; (e) flows for coarse (top) and fine (bottom) alignments.



(a) Source (b) ArtMiner [64] (c) Coarse align. (d) Fine align. (e) Animation

Fig. 9: Aligning groups of patterns discovered by ArtMiner [64]: (a) source images; (b) average from [64]; (c) average after coarse alignment (homography); (d) average after fine alignment; (e) animation (view with Acrobat Reader).

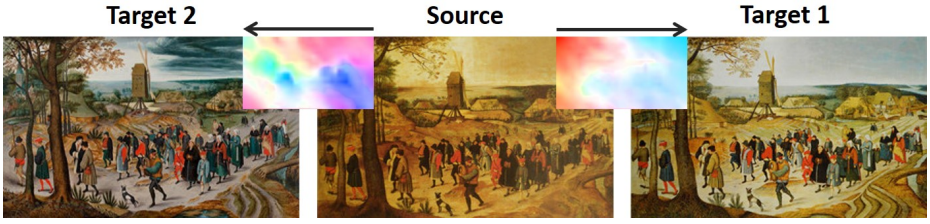


Fig. 10: Analyzing copy process from flow. From the middle painting to the right one, the flow is smooth while from the middle painting to the left one, the flow is more irregular, which suggests two different copy processes.

list of image pairs and then give them as input to COLMAP [63] that provides a localisation estimation for the query nighttime images. To limit the number of correspondences we input to COLMAP we use only correspondences on a sparse set of keypoints, selected using the Superpoint [12] keypoints. Our results are reported in Table 3(b) and are on par with state of the art results. As can be seen in the examples of Figure 5, our approach discovers large number of 3D points under significant appearance and viewpoint variations.

4.3 Applications

We believe that 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, that showcase the robustness of our approach.

Texture transfer. Because it provides dense correspondences, our approach can be used to transfer texture between images. In Figure 6 and Figure 2(f) we show results using historical and modern images from the LTLL dataset [15]. We use the pre-trained segmentation network of [82], and transfer the texture from the source image to the target building regions.

Internet images alignment. As visualized in Figures 2(d) and 7, we can align sets of Internet images, aligning them and computing their average image, similar to [66]. Even if our image set is not precisely the same, we note that much more details can be seen in the average of our fine-aligned images.

Artwork analysis. Finding and matching near-duplicate patterns in a historical body of artwork is an important problem for art historians. Computationally, it is difficult because the duplicate appearance can be very different [64]. In Figure 8, we show visual results of aligning different versions of artworks from the Brueghel dataset [64] 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 for each individual artwork. Indeed, we found that some artwork were copied in a spatially consistent way, while in others, different parts of the picture were not aligned with each other. This can be immediately and clearly seen in the flows, which are either very regular or very discontinuous, as illustrated in Figure 10. The same process can, of course, be applied to more than a single pair of images, as illustrated in Figure 2(e)

and 9 where we align together many examples of similar details identified by [64]. Visualizing the succession of the finely aligned images allows to easily identify their differences.

5 Conclusion

We have introduced a new unsupervised approach for generic dense image alignment which is able to perform 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. Our method can predict accurate dense correspondences for challenging image pairs, exhibiting large appearance and viewpoint variations, and performs on par with state of the art methods both on classical optical flow benchmarks and on difficult two-view geometry and localization benchmarks. We also demonstrated it allows completely new applications of computer vision for art4work analysis.

Acknowledgements: This work was supported in part by ANR project EnHerit ANR-17-CE23-0008, project Rapid Tabasco, NSF IIS-1633310, grant from SAP, the France-Berkeley Fund, 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: 2016 IEEE international conference on image processing (ICIP). pp. 1629–1633. IEEE (2016)
3. Aubry, M., Russell, B.C., Sivic, J.: Painting-to-3d model alignment via discriminative visual elements. ACM Transactions on Graphics (ToG) **33**(2), 14 (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. pp. 4015–4023 (2015)
5. Balntas, V., Lenc, K., Vedaldi, A., Mikolajczyk, K.: Hpatches: A benchmark and evaluation of handcrafted and learned local descriptors. In: CVPR (2017)
6. Brox, T., Bregler, C., Malik, J.: Large displacement optical flow. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 41–48. IEEE (2009)
7. Butler, D.J., Wulff, J., Stanley, G.B., Black, M.J.: A naturalistic open source movie for optical flow evaluation. In: A. Fitzgibbon et al. (Eds.) (ed.) European Conf. on Computer Vision (ECCV). pp. 611–625. Part IV, LNCS 7577, Springer-Verlag (Oct 2012)
8. 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. pp. 5594–5603 (2019)
9. Choy, C.B., Gwak, J., Savarese, S., Chandraker, M.: Universal correspondence network. In: Advances in Neural Information Processing Systems. pp. 2414–2422 (2016)
10. 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. pp. 11–20 (1996)
11. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
12. DeTone, D., Malisiewicz, T., Rabinovich, A.: Superpoint: Self-supervised interest point detection and description. In: CVPR Deep Learning for Visual SLAM Workshop (2018), http://arxiv.org/abs/1712.07629
13. 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. pp. 2758–2766 (2015)
14. 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. pp. 8092–8101 (2019)
15. Fernando, B., Tommasi, T., Tuytelaars, T.: Location recognition over large time lags. Computer Vision and Image Understanding **139**, 21–28 (2015)
16. 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 **24**(6), 381–395 (1981)

17. Hartley, R., Zisserman, A.: Multiple view geometry in computer vision. Cambridge university press (2003)
18. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. arXiv preprint arXiv:1911.05722 (2019)
19. 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. pp. 770–778 (2016)
20. 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. pp. 5704–5712 (2016)
21. 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. pp. 2462–2470 (2017)
22. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167 (2015)
23. Irani, M., Anandan, P., Cohen, M.: Direct recovery of planar-parallax from multiple frames. IEEE Transactions on Pattern Analysis and Machine Intelligence **24**(11), 1528–1534 (2002)
24. Jaderberg, M., Simonyan, K., Zisserman, A., et al.: Spatial transformer networks. In: Advances in neural information processing systems. pp. 2017–2025 (2015)
25. Jason, J.Y., Harley, A.W., Derpanis, K.G.: Back to basics: Unsupervised learning of optical flow via brightness constancy and motion smoothness. In: European Conference on Computer Vision. pp. 3–10. Springer (2016)
26. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
27. Kong, S., Fowlkes, C.: Multigrid predictive filter flow for unsupervised learning on videos. arXiv preprint arXiv:1904.01693 (2019)
28. 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. vol. 1, pp. 685–688. IEEE (1994)
29. Lai, Z., Xie, W.: Self-supervised learning for video correspondence flow. arXiv preprint arXiv:1905.00875 (2019)
30. 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. pp. 9532–9542 (2019)
31. Laskar, Z., Melekhov, I., Tavakoli, H.R., Ylioinas, J., Kannala, J.: Geometric image correspondence verification by dense pixel matching. arXiv preprint arXiv:1904.06882 (2019)
32. 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. pp. 2041–2050 (2018)
33. Liu, C., Yuen, J., Torralba, A.: Sift flow: Dense correspondence across scenes and its applications. IEEE transactions on pattern analysis and machine intelligence **33**(5), 978–994 (2010)
34. Long, J.L., Zhang, N., Darrell, T.: Do convnets learn correspondence? In: Advances in neural information processing systems. pp. 1601–1609 (2014)
35. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International journal of computer vision **60**(2), 91–110 (2004)
36. Lucas, B.D., Kanade, T., et al.: An iterative image registration technique with an application to stereo vision (1981)

37. Luo, Z., Shen, T., Zhou, L., Zhang, J., Yao, Y., Li, S., Fang, T., Quan, L.: Contextdesc: Local descriptor augmentation with cross-modality context. Computer Vision and Pattern Recognition (CVPR) (2019)
38. 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 (ECCV). pp. 168–183 (2018)
39. Maddern, W., Pascoe, G., Linegar, C., Newman, P.: 1 year, 1000 km: The oxford robotcar dataset. The International Journal of Robotics Research **36**(1), 3–15 (2017)
40. Mayer, N., Ilg, E., Häusser, P., Fischer, P., Cremers, D., Dosovitskiy, A., Brox, T.: A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In: IEEE International Conference on Computer Vision and Pattern Recognition (CVPR) (2016), http://lmb.informatik.uni-freiburg.de/Publications/2016/MIFDB16, arXiv:1512.02134
41. Melekhov, I., Tiulpin, A., Sattler, T., Pollefeys, M., Rahtu, E., Kannala, J.: Dgc-net: Dense geometric correspondence network. In: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 1034–1042. IEEE (2019)
42. Menze, M., Geiger, A.: Object scene flow for autonomous vehicles. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
43. 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. pp. 4826–4837 (2017)
44. Mishkin, D., Radenovic, F., Matas, J.: Repeatability is not enough: Learning affine regions via discriminability. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 284–300 (2018)
45. 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. pp. 2666–2674 (2018)
46. 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. pp. 3456–3465 (2017)
47. Plötz, T., Roth, S.: Neural nearest neighbors networks. In: Advances in Neural Information Processing Systems. pp. 1087–1098 (2018)
48. 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. pp. 652–660 (2017)
49. 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. pp. 5099–5108 (2017)
50. Ranftl, R., Koltun, V.: Deep fundamental matrix estimation. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 284–299 (2018)
51. Ranjan, A., Black, M.J.: Optical flow estimation using a spatial pyramid network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4161–4170 (2017)
52. 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)
53. 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)

54. Revaud, J., De Souza, C., Humenberger, M., Weinzaepfel, P.: R2d2: Reliable and repeatable detector and descriptor. In: Advances in Neural Information Processing Systems. pp. 12405–12415 (2019)
55. 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. pp. 1164–1172 (2015)
56. Revaud, J., Weinzaepfel, P., Harchaoui, Z., Schmid, C.: Deepmatching: Hierarchical deformable dense matching. International Journal of Computer Vision **120**(3), 300–323 (2016)
57. 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. pp. 6148–6157 (2017)
58. Rocco, I., Cimpoi, M., Arandjelović, R., Torii, A., Pajdla, T., Sivic, J.: Neighbourhood consensus networks. In: Advances in Neural Information Processing Systems. pp. 1651–1662 (2018)
59. 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. pp. 234–241. Springer (2015)
60. 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. pp. 8601–8610 (2018)
61. Sattler, T., Weyand, T., Leibe, B., Kobbelt, L.: Image retrieval for image-based localization revisited. In: BMVC. vol. 1, p. 4 (2012)
62. Sawhney, H.S.: 3d geometry from planar parallax. In: CVPR. vol. 94, pp. 929–934 (1994)
63. Schonberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4104–4113 (2016)
64. 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 (CVPR) (2019)
65. 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. pp. 3118–3126 (2018)
66. 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. pp. 1–10 (2011)
67. Singh, S., Gupta, A., Efros, A.A.: Unsupervised discovery of mid-level discriminative patches. In: European Conference on Computer Vision. pp. 73–86. Springer (2012)
68. 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. pp. 8934–8943 (2018)
69. Szeliski, R.: Image alignment and stitching: A tutorial. Found. Trends. Comput. Graph. Vis. **2**(1), 1–104 (Jan 2006). https://doi.org/10.1561/0600000009, http://dx.doi.org/10.1561/0600000009
70. 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. arXiv preprint arXiv:1503.01817 (2015)

71. 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. pp. 661–669 (2017)
72. Ulyanov, D., Vedaldi, A., Lempitsky, V.: Deep image prior. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 9446–9454 (2018)
73. Vondrick, C., Shrivastava, A., Fathi, A., Guadarrama, S., Murphy, K.: Tracking emerges by colorizing videos. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 391–408 (2018)
74. 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 (CVPR) (June 2019)
75. Wang, Y., Yang, Y., Yang, Z., Zhao, L., Wang, P., Xu, W.: Occlusion aware unsupervised learning of optical flow. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)
76. 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 **13**(4), 600–612 (2004)
77. Weinzaepfel, P., Revaud, J., Harchaoui, Z., Schmid, C.: Deepflow: Large displacement optical flow with deep matching. In: Proceedings of the IEEE international conference on computer vision. pp. 1385–1392 (2013)
78. 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. pp. 1983–1992 (2018)
79. 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. International Conference on Computer Vision (ICCV) (2019)
80. Zhang, R.: Making convolutional networks shift-invariant again. arXiv preprint arXiv:1904.11486 (2019)
81. 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)
82. 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)
83. 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. pp. 1191–1200 (2015)
84. 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. pp. 117–126 (2016)