Unsupervised Learning of Consensus Maximization for 3D Vision Problems

Thomas Probst, Danda Pani Paudel, Ajad Chhatkuli, Luc Van Gool

Computer Vision Laboratory, ETH Zurich, Switzerland

Abstract

Consensus maximization is a key strategy in 3D vision for robust geometric model estimation from measurements with outliers. Generic methods for consensus maximization, such as Random Sampling and Consensus (RANSAC), have played a tremendous role in the success of 3D vision, in spite of the ubiquity of outliers. However, replicating the same generic behaviour in a deeply learned architecture, using supervised approaches, has proven to be difficult. In that context, unsupervised methods have a huge potential to adapt to any unseen data distribution, and therefore are highly desirable. In this paper, we propose for the first time an unsupervised learning framework for consensus maximization, in the context of solving 3D vision problems. For that purpose, we establish a relationship between inlier measurements, represented by an ideal of inlier set, and the subspace of polynomials representing the space of target transformations. Using this relationship, we derive a constraint that must be satisfied by the sought inlier set. This constraint can be tested without knowing the transformation parameters, therefore allows us to efficiently define the geometric model fitting cost. This model fitting cost is used as a supervisory signal for learning consensus maximization, where the learning process seeks for the largest measurement set that minimizes the proposed model fitting cost. Using our method, we solve a diverse set of 3D vision problems, including 3D-3D matching, non-rigid 3D shape matching with piece-wise rigidity and image-to-image matching. Despite being unsupervised, our method outperforms RANSAC in all three tasks for several datasets.

1. Introduction

In 3D vision, problems such as Structure-from-Motion (SfM) [26, 16, 12] and image registration [17, 30] are geometrically solved from noisy measurements with outliers. In that context, consensus maximization among inlier measurements, is very often a crucial step. Typically, the maximum consensus is searched using the Random Sampling and Consensus (RANSAC) algorithm [13] or its derivatives [35, 44, 37]. During this process, almost all geometric models are represented by a system of polynomials whose common root specifies the desired transformation parameters. Polynomial-based geometric model representations, when used with RANSAC, offer accurate transformation parameters, therefore are used for a diverse set of problems [12, 17, 30, 20]. Globally optimal methods [10, 33, 42, 2, 15, 3, 23, 49], which overcome limitations of RANSAC, further bolster the importance of maximizing consensus. However, most methods that seek consensus maximization solely depend upon the geometric models. They do not exploit knowledge about the scene or their measurement distributions.

Learning for consensus maximization is an alternative approach which has the potential of providing a higher inlier/outlier classification accuracy, by leveraging the distribution of the given data. Additionally, the supervisory signal for consensus maximization may help to learn other related tasks, within the framework of multi-task learning. Owing to the success of deep learning, recent methods have tackled the consensus maximization problem for image matching using the Fundamental matrix [36], the Essential matrix [48] and for the absolute pose [7]. These methods use supervision through ground-truth labels to train a neural network for classifying correspondences as inliers or outliers. Other methods, [36, 48] respectively use Fundamental or Essential matrix models, which require supervision by their associated matrices to train the network parameters. Unfortunately, such networks, trained in a supervised manner, are not on par with RANSAC in terms of their generality to unseen data distribution. Moreover, ground-truth geometric models are sometimes difficult to obtain, or even non-existent [16]. In this scenario, unsupervised methods have a huge potential, as they can adapt to any unseen data distribution, and therefore are highly desirable.

In this paper, we propose an unsupervised framework to learn consensus maximization in the context of geometric 3D vision problems. We model the geometric transformations using polynomials, as commonly done in the literature [12, 16, 30]. We then develop a framework of fitting polynomials in a deep architecture while maximizing the consensus. To develop such method, we first establish a relationship between inlier measurements, represented by an ideal of the inlier measurements, and the subspace of polynomials representing the space of target transformations. This relationship is then used to formulate a loss function for fitting polynomials, as the singular values minimization problem on the so-called Vandermonde matrix [4, 8].While minimizing the proposed loss function, our training process also seeks for the largest consensus among the measurements. Thereby our formulation evaluates the consistency to the geometric model without regressing the model parameters, which is known to be sensitive for robust estimation tasks within supervised setups [36, 48]. Regardless, one may still think of using a robust regression-based formulation, e.g. based on m-estimators. However, minimizing such robust loss function may provide satisfactory results only when outliers are relatively few [31, 2].

Our loss function on the other hand is designed to train a correspondence classification network, and it naturally extends to many transformation models that can be expressed by one or more polynomials. Notably, we neither require classification labels, nor the ground-truth transformations. To the best of our knowledge, our work is the first to learn a deep architecture in an unsupervised manner for consensus maximization in 3D vision problems. Our method is adapted to a diverse set of 3D vision problems: 3D-3D rigid body transformation, non-rigid shape matching with piecewise rigidity and uncalibrated 2D transformations (Fundamental matrix and homography). We experimentally show that our method is able to outperform RANSAC in all of the mentioned tasks, while being unsupervised. We further empirically show how the accuracy of supervised methods worsens when tested on different data statistics. This deterioration in accuracy could be recovered to a large extend, using the proposed unsupervised training framework.

2. Related Work

We briefly summarize the related work to put our paper into context. Consensus maximization is a well studied topic [12, 17, 30, 20], and is usually solved with RANSAC [13, 35, 44, 37]. In contrast to heuristic approaches, global methods provide optimality guaranties [10, 33, 42, 2, 15, 3, 23, 49]. Recently, supervised machine learning has been leveraged to solve consensus maximization and robust estimation. With regards to two-view geometry, [48, 36] learn from keypoint correspondences, whereas [32, 29] regress transformation parameters directly from the input images. [7] integrate a differentiable version of RANSAC into their network to robustify camera localization. Although they may show beneficial accuracy (w.r.t. to classical RANSAC) and speed (compared to global methods), the generalizability of supervised methods is limited to the domain of the training data and to the level of abstraction of the input data. In contrast, our method takes inspiration from algebraic varieties [4, 9] to train a deep network for inlier/outlier classification in an unsupervised fashion. We build on permutation-invariant networks which recently gained attention for learning on unordered point sets [34, 24, 46], by adapting the PointNet [34] architecture.

3. Background and Theory

3.1. Consensus Maximization

Given a set $\mathcal{X} = \{(\mathsf{u}_i, \mathsf{v}_i), i = 1, \dots, m\}$, of corresponding measurements, the consensus maximization problem involves finding the largest subset $\Omega \subseteq \mathcal{X}$ that can be explained by a single parametric transformation Φ. For every pair $(\mathsf{u}, \mathsf{v}) \in \Omega$, the distance between v and the transformed measurement $\Phi(\mathsf{u})$ is smaller than a threshold ϵ. Mathematically, the problem of consensus maximization is,

$$\begin{aligned} &\max_{\Phi, \Omega \subseteq \mathcal{X}} \quad |\Omega| \\ &\text{s.t.} \quad \mathbf{d}(\Phi(\mathsf{u}_i), \mathsf{v}_i) \leq \epsilon, \quad \forall (\mathsf{u}_i, \mathsf{v}_i) \in \Omega. \end{aligned} \tag{1}$$

We represent Φ using polynomials as commonly done in the literature. In this regard, (1) is an algebraic problem of finding a variety $\mathcal{V}$ of known dimension – representing Φ – such that the distance from every inlier member of $\mathcal{X}$ to $\mathcal{V}$ is bounded by a given threshold ϵ. After a general formulation of our task, we first discuss the problem of consensus maximization in the absence of noise. The case of noisy data is considered later.

3.2. Problem Formulation

Consider the ring $\mathcal{R}[\mathsf{x}] := \mathbb{R}[x_1, ..., x_n]_d$ of multivariate polynomials of degree $\leq d$ and an algebraic variety $\mathcal{V} \subseteq \mathbb{R}^n$ defined such that $\mathcal{V} := \{\mathsf{x} \in \mathbb{R}^n : p_j(\mathsf{x}) = 0\}$. Let $\mathsf{x}_i = (\mathsf{u}_i^\intercal, \mathsf{v}_i^\intercal)^\intercal \in \mathbb{R}^n$ be a measurement vector representing a pair of correspondences in $\mathcal{X}$. Primarily, we are interested in finding $\Omega \subseteq \mathcal{X}$ which vanishes on some variety $\mathcal{V}$ constrained by Φ, where $\mathcal{X}$ is corrupted by outliers. Recovering $\mathcal{V}$ exactly is an NP-hard problem because every polynomial in the ideal $\mathcal{I}(\mathcal{V})$:= $\{\sum_j g_j(\mathsf{x}) p_j(\mathsf{x}) : g_j(\mathsf{x}) \in \mathcal{R}[\mathsf{x}]\}$ vanishes on $\mathcal{V}$ as well. Nevertheless, the existence of the ideal $\mathcal{I}(\mathcal{V})$ implies the existence of $\mathcal{V}$. In this context, an example problem of fitting a 3D line, represented by Φ, involves an ideal $\mathcal{I}(\mathcal{V})$ of some one-dimensional variety $\mathcal{V}$, which is parameterized by two intersecting planes $p_1(\mathsf{x})$ and $p_2(\mathsf{x})$. As shown in Fig. 1, the ideal $\mathcal{I}(\mathcal{V})$ is represented by a pencil of planes passing through the line. The desired inlier set is the largest $\Omega \subseteq \mathcal{X}$ for which there exists a line – represented by $\mathcal{V}$ of dimension one – passing through all 3D points $\mathsf{x} \in \Omega$. While seeking for Ω, we ensure the existence of $\mathcal{V}(\Omega) := \{\mathcal{V} : \Omega \subseteq \mathcal{V}\}$ by ensuring the existence of $\mathcal{I}(\Omega)$.

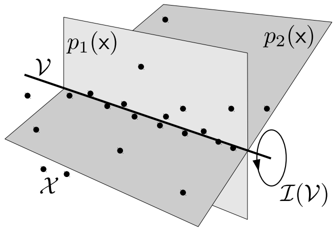

Figure 1. **3D Line Fitting.** Finding a one-dimensional variety $\mathcal{V}$ from a sample set $\mathcal{X}$. $\mathcal{V}$ is the intersection of planes $p_1(\mathsf{x})$ and $p_2(\mathsf{x})$. The ideal $\mathcal{I}(\mathcal{V})$ is a pencil of planes.

Definition 3.1 *The ideal $\mathcal{I}(\Omega)$ is a set of polynomials that vanishes on the samples of Ω. i.e. $\mathcal{I}(\Omega) := \mathcal{I}(\mathcal{V}(\Omega))$.*

One can argue that there always exists some $\mathcal{I}(\Omega)$, even when $\Omega = \mathcal{X}$. However, we are interested in only those $\mathcal{I}(\Omega)$ which lie in the space of valid transformations Φ, represented by the polynomial basis $\mathcal{B} \subseteq \mathcal{R}[\mathsf{x}]$. We assume that Φ resides within the vector space spanned by $\mathcal{B}$.

Definition 3.2 *The basis $\mathcal{B}$ is a set of monomial bases of $\mathcal{R}[\mathsf{x}]$ and $\mathcal{R}_{\mathcal{B}} = \mathbb{R}[\mathcal{B}]$ is the subspace of polynomials spanned by $\mathcal{B}$. Any polynomial in $\mathcal{R}_{\mathcal{B}}$ can therefore be represented by a coefficient vector in $\mathbb{R}^s$ for $s = |\mathcal{B}|$.*

The polynomial representation of Φ involves r linearly independent equations in $\mathcal{R}_{\mathcal{B}}$ that vanish for all $\mathsf{x} \in \Omega$. For example, while fitting lines and spheres in 3D, $(r, \mathcal{R}_{\mathcal{B}})$ are $(2, \mathbb{R}[x_1, x_2, x_3]_1)$ and $(1, \mathbb{R}[x_1, x_2, x_3]_2)$, respectively.

The consensus maximization problem can be reformulated as the search for the largest inlier set Ω, for which there exists some ideal $\mathcal{I}(\Omega)$ whose intersection with $\mathcal{R}_{\mathcal{B}}$ spans exactly r dimensions:

$$\max_{\Omega \subseteq \mathcal{X}} |\Omega| : \dim(\mathcal{I}(\Omega) \cap \mathcal{R}_{\mathcal{B}}) = r. \tag{2}$$

Intuitively, the problem of (2) demands that $\mathcal{I}(\Omega)$ must vanish on all polynomials represented by r equations in $\mathcal{R}_{\mathcal{B}}$, as required by Φ. Recall the example of Fig. 1. The line $\mathcal{V}$ lives in the space of linear polynomials $\mathcal{R}_{\mathcal{B}} = \mathbb{R}[x_1, x_2, x_3]_1$ of dimension 2. In other words, one needs 2 independent linear equations to represent a line in 3D. Among all linear polynomials (in $\mathcal{R}_{\mathcal{B}}$) that vanish on Ω, we require exactly two to be independent. These equations indeed represent two intersecting planes, thus the line $\mathcal{V}$.

3.3. Ideals and Sample Sets

The relationship between an ideal $\mathcal{I}(\Omega)$ and the sample set $\Omega \subset \mathbb{R}^n$ can be established with the help of the so-called Vandermonde matrix. This also allows us to reason about the existence of $\mathcal{I}(\Omega) \cap \mathcal{R}_{\mathcal{B}}$ for the chosen bases $\mathcal{B}$.

Definition 3.3 *The Vandermonde matrix $\mathsf{M}_d(\Omega) \in \mathbb{R}^{m \times s}$ is a matrix with the terms of a geometric progression monomials in each row, such that the entries m_{ij} are the monomials $\mathsf{x}^{\mathsf{e}} = x_1^{e_1} x_2^{e_2} \dots x_n^{e_n}$ of degree at most d.*

For example, if $n = 1, d = 3$, and $\Omega = \{x_1, x_2, x_3\}$ then $\mathsf{M}_3(\Omega)$ is the Vandermonde matrix of the form,

$$\mathsf{M}_3(\Omega) = \begin{bmatrix} x_1^3 & x_1^2 & x_1 & 1 \\ x_2^3 & x_2^2 & x_2 & 1 \\ x_3^3 & x_3^2 & x_3 & 1 \end{bmatrix}.$$

Note that $\mathsf{M}_d(\Omega)$ grows linearly in the number of samples m as well as in the number of monomials s, and therefore is a compact representation. One of the key properties of the Vandermonde matrix, which allows us to analyze the existence of $\mathcal{I}(\Omega) \cap \mathcal{R}_{\mathcal{B}}$, is stated below.

Theorem 3.4 *The kernel ker$(\mathsf{M}_d(\Omega))$ of the Vandermonde matrix $\mathsf{M}_d(\Omega)$ equals to the vector space $\mathcal{I}(\Omega) \cap \mathcal{R}_{\mathcal{B}}$. i.e. all polynomials that are linear combinations of $\mathcal{B}$ and vanish on Ω are represented by $\mathcal{I}(\Omega) \cap \mathcal{R}_{\mathcal{B}} =$ ker$(\mathsf{M}_d(\Omega))$.*

Using Theorem 3.4, the problem of (1) is expressed as the following constrained cardinality maximization problem,

$$\max_{\Omega \subseteq \mathcal{X}} |\Omega|, \quad \text{s.t.} \quad \dim(\ker(\mathsf{M}_d(\Omega))) = r. \tag{3}$$

In the presence of noise, however, the constraint on the kernel dimension of the Vandermonde matrix $\mathsf{M}_d(\Omega)$, given in (3), is difficult to satisfy. Therefore, we enforce the dimensionality constraint by minimizing the singular values of $\mathsf{M}_d(\Omega)$ [18]. For descending singular values $\sigma_1, \sigma_2, \dots \sigma_s$ of $\mathsf{M}_d(\Omega)$, the trailing r singular values must be zero for the constraint of (3) to be true. Hence, we relax the problem of (3), for a given scalar λ, as follows,

$$\max_{\Omega \subseteq \mathcal{X}} |\Omega| - \lambda \sum_{k=0}^{r-1} \sigma_{s-k}(\mathsf{M}_d(\Omega)). \tag{4}$$

Maximizing the cardinality can be thought of a subset selection problem, expressed using a set of binary variables $w_i \in \{0, 1\}$ (de-)activating the corresponding rows in $\mathsf{M}_d(\Omega)$.

$$\max_{\mathsf{w} \in \{0,1\}^m} \sum_{i=1}^{m} w_i - \lambda \sum_{k=0}^{r-1} \sigma_{s-k}(\text{diag}(\mathsf{w}) \mathsf{M}_d(\mathcal{X})). \tag{5}$$

Exact solving of (5) involves combinatorial optimization and is not tractable in practice. Therefore, we relax the problem by introducing a soft selection of inliers using continuous sample weights w_i as follows.

$$\max_{\mathsf{w} \in \mathbb{R}^m} \sum_{i=1}^{m} w_i - \lambda \sum_{k=0}^{r-1} \sigma_{s-k}(\text{diag}(\mathsf{w}) \mathsf{M}_d(\mathcal{X})), \quad \text{s.t. } 0 \le w_i \le 1. \tag{6}$$

Equation (6) is still not convex, however, it is differentiable and can be optimized using gradient-based methods.

3.4. Recovering the Ideal

In some cases, we are interested in actually recovering the ideal $\mathcal{P} = \mathcal{I}(\Omega) \cap \mathcal{R}_{\mathcal{B}}$, for example to compute the parameters of the transformation Φ. Consider the SVD decomposition of $\mathsf{M}(\mathcal{X}) = \mathsf{U}\Sigma\mathsf{V}^T$. Recall that the ideal of our interest $\mathcal{P}$ lies on the kernel of $\mathsf{M}(\mathcal{X})$. Therefore we can extract the corresponding polynomials from the nullspace of $\mathsf{M}(\mathcal{X})$, represented by the trailing r right singular vectors,

$$\mathsf{B} = \left[\, \mathsf{v}_s \;\; \mathsf{v}_{s-1} \;\; \dots \;\; \mathsf{v}_{s-r+1} \,\right] \in \mathbb{R}^{s \times r}. \tag{7}$$

Let $\mathsf{p}(\mathsf{x})$ be the vector of monomials in $\mathcal{B}$. The recovered ideal $\mathcal{P}$ is then defined by $\mathcal{P} = \{\mathsf{B}^T \mathsf{p}(\mathsf{x}) = 0\}$.

3.5. Deep Learning for Consensus Maximization

Besides direct optimization of Eq. (6) for a given set of correspondences, it can be thought of a supervisory signal for learning consensus maximization from data. Given a neural network $\mathbf{w}_\theta(\mathcal{X}) : \mathbb{R}^{m \times n} \to [0,1]^m$ parametrized by θ, we wish to learn prediction score w_i for each sample $\mathsf{x}_i \in \mathcal{X}$, that maximizes the number of inliers ($w_i \to 1$), while rejecting outliers ($w_i \to 0$). To this end, we define a differentiable supervisory signal that requires neither point-wise labels, nor knowledge about the ground truth transformation between correspondences. Given sample set $\mathcal{X}$, we aim to learn the optimal parameters θ by minimizing the following empirical loss $\ell(\theta, \mathcal{X})$ based on Eq. (6).

$$\ell(\theta, \mathcal{X}) = -\left\|\mathbf{w}_\theta(\mathcal{X})\right\|_1 + \lambda \sum_{k=0}^{r-1} \sigma_{s-k}(\operatorname{diag}(\mathbf{w}_\theta(\mathcal{X}))\mathsf{M}_d(\mathcal{X})). \tag{8}$$

The input to our network is a set of correspondences $\mathcal{X}$. Consequently, we require a architecture that is invariant to the permutation of the input, whereas most neural networks were designed for ordered input data, e.g. 2D images. However, recent advances in deep learning on unordered point sets [34, 24, 46] allow a suitable choice of architecture for our problem. We employ the PointNet [34] segmentation architecture to encourage a global reasoning about the underlying transformation. The key component of the architecture is a max-pool operation across correspondences before computing a global feature vector (GFV). The GFV is then concatenated to the pointwise features to exploit the global context for the point-wise predictions. Since our goal is binary inlier classification, we add an element-wise sigmoid layer that outputs inlier prediction scores w_i in the range $[0,1]$. We define our prediction function $\mathbf{w}_\theta(\mathcal{X})$ as

$$\mathbf{w}_\theta(\mathcal{X})_i = \mathbf{s}(\mathcal{C}_\theta(\mathcal{X})_i), \quad \mathbf{s}(x) = (1 + e^{-x})^{-1}, \tag{9}$$

with the PointNet-seg output denoted as $\mathcal{C}_\theta(\mathcal{X}) \in \mathbb{R}^m$. Together with the loss function (8), our architecture can the be modeled using standard building blocks: after construction of the Vandermonde Matrix $\mathsf{M}_d(\mathcal{X}) \in \mathbb{R}^{m \times s}$, every row i gets weighted with the corresponding inlier probability w_i. Then we compute the last r singular values of the weighted Vandermonde matrix using the differentiable SVD operation. The architecture is illustrated in Fig. 2. We implement the network in tensorflow [1] and use the ADAM [19] optimizer to learn the parameters θ.

By design, our approach generalizes over transformation functions that can be represented by polynomial equations. In the following sections we explain how the Vandermonde loss can be adapted to geometric transformation problems.

4. 3D Vision Problems

In this section, we present four examples of 3D vision problems for consensus maximization problems and introduce different problem specific $\mathcal{R}_{\mathcal{B}}$ subspace constraints.

Unfortunately, non-linear constraints on transformation parameters can not be directly applied in this framework. However, since we can compute the ideal $\mathcal{P}$ and extract the parameters of the model, we are able to introduce a regularization term on the solutions.

4.1. Rigid Body Transformation

We consider correspondences between two point clouds that differ by a 3D rigid body transformation. Let $\{\mathsf{u}, \mathsf{v}\}$ be euclidean coordinates of a pair of points such that

$$\mathsf{v} = \mathsf{R}\mathsf{u} + \mathsf{t}, \quad \mathsf{R} \in SO(3), \, \mathsf{t} \in \mathbb{R}^3. \tag{10}$$

We can see that Eq. (10) involves $r = 3$ linear equations in the coordinates of corresponding points. Therefore we can restrict the polynomial subspace $\mathcal{R}_{\mathcal{B}}$ to linear terms $\mathcal{B} = \{u_x, u_y, u_z, v_x, v_y, v_z, 1\}$. This leads to a Vandermonde matrix $\mathsf{M}(\mathcal{X}) \in \mathbb{R}^{m \times 7}$, with kernel dimension 3.

Note that this representation holds for any 3D affine transformation, since it does not enforce rotation manifold and scale constraints. We therefore introduce an additional regularization term on the recovered ideal $\mathcal{P}$. Recall that we extract the basis B of $\mathcal{P}$ according to Eq. (7). The recovered polynomials in $\mathcal{P}$ are some linear combination of Eq. (10). To recover the components of R and t, we need a change of basis to separate v_x, v_y and v_z in each equation. The desired basis B' of the form in Eq. (10) can be obtained by

$$\mathsf{B}' = -\left[\, \mathsf{b}_4^T \;\; \mathsf{b}_5^T \;\; \mathsf{b}_6^T \,\right]^{-T} \mathsf{B} = \left[\, \hat{\mathsf{R}} \;\; -\mathsf{I}_{3\times3} \;\; \hat{\mathsf{t}} \,\right], \tag{11}$$

where b_i denotes the ith row of the matrix B. This form offers direct access to the estimated rotation and translation parameters. To avoid numerical issues, we add a small identity matrix before the matrix inversion in our implementation. Based on this observation, we define a regularizer ℓ_r,

$$\ell_r(\theta, \mathcal{X}) = \log\left(1 + \left\|\hat{\mathsf{R}}\hat{\mathsf{R}}^T - \mathsf{I}_{3\times3}\right\|_2\right). \tag{12}$$

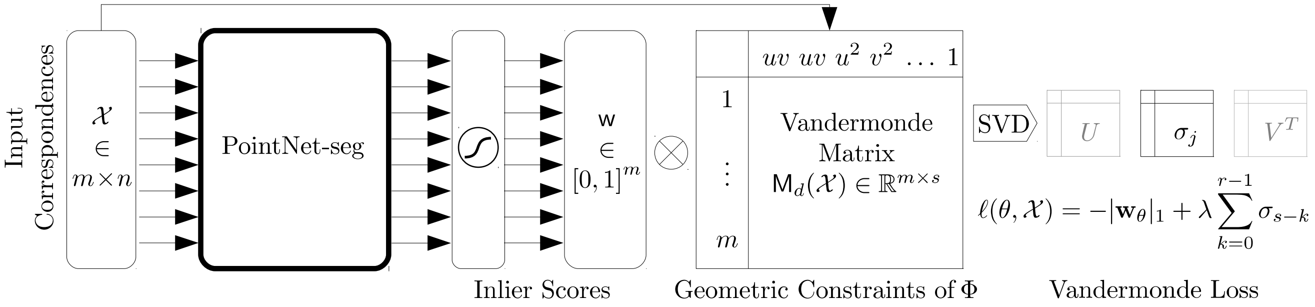
Input Correspondences
$\mathcal{X}$
$\in$
$m \times n$
PointNet-seg
w
$\in$
$[0,1]^m$
uv uv u^2 v^2 $\dots$ 1
1
$\vdots$
m
Vandermonde Matrix
$\mathsf{M}_d(\mathcal{X}) \in \mathbb{R}^{m \times s}$
SVD
U
σ_j
V^T
$\ell(\theta, \mathcal{X}) = -|\mathbf{w}_\theta|_1 + \lambda \sum_{k=0}^{r-1} \sigma_{s-k}$
Inlier Scores
Geometric Constraints of Φ
Vandermonde Loss

Figure 2. **Method Overview.** A set of correspondences $\mathcal{X}$ is fed to a network that outputs inlier scores $\mathbf{w}$. Every score weights its corresponding row in the Vandermonde matrix $\mathsf{M}_d(\mathcal{X})$. We minimize the last singular values of $\mathsf{M}_d(\mathcal{X})$, while maximizing the number of inliers. Training uses only the knowledge about the polynomial structure of transformation Φ.

The log serves for extenuation of spike-like behaviour. We add $\lambda_r \ell_r(\theta, \mathcal{X})$ to the Vandermonde loss (8). Note that (12) can be also adapted to similarity transform by encouraging orthogonality after normalizing by scale.

Non-rigid Extension. Theoretically, there is no reason to assume that our network can learn the inlier statistics of only one rigid transformation in the data. As long as a consistent transformation pattern is present, correlations between inliers can be extracted by an accordingly trained network [45]. In this work, we investigate this idea in the context of unsupervised learning. We assume that the transformation can be approximated by piece-wise rigidity, a well studied approximation for non-rigid surfaces [43, 28, 38, 21]. Consequently, we model the global non-rigid deformation by rigid transformations on local neighborhoods. A straight forward approach is to compute our loss defined for 3D rigid transformation on local neighborhoods of the input point set. Given a K-neighborhood $\mathcal{N}_a$ of an anchor point $\mathsf{u}_a \in \mathsf{U}$ on the first point cloud, we assemble a local Vandermonde matrix $\mathsf{M}(\mathcal{N}_a)$. The loss is then computed according to Alg. 1.

Algorithm 1 Piecewise-rigid loss $\ell_{\text{pr}}(\theta, \mathcal{X}, K)$

```
0. Define a reference point set $\mathsf{U} = \{\mathsf{u}_i : (\mathsf{u}_i, \mathsf{v}_i) \in \mathcal{X}\}$.
1. Randomly sample an anchor point $\mathsf{u}_a \in \mathsf{U}$.
2. Compute the (geodesic) neighborhood of $\mathsf{u}_a$
   $\mathcal{N}_a = \{(\mathsf{u}_i, \mathsf{v}_i) \in \mathcal{X} : \mathbf{d}(\mathsf{u}_a, \mathsf{u}_i) \le \delta\}$, where $|\mathcal{N}_a| = K$.
3. Assemble the local Vandermonde matrix $\mathsf{M}(\mathcal{N}_a)$.
4. Extract $\mathsf{B}_a$ (7) of $\mathsf{M}(\mathcal{N}_a)$ to compute $\ell_r(\theta, \mathcal{N}_a)$ (12).
5. Compute $\ell(\theta, \mathcal{N}_a)$ (8).
6. Return $\ell(\theta, \mathcal{N}_a) + \lambda_r \ell_r(\theta, \mathcal{N}_a)$.
```

Random sampling of anchor points u_a may result in gradients of high variance and unstable training. We observed that larger batch size and low learning rate gives stable gradients and improves learning significantly.

4.2. Uncalibrated 2-View Geometry

We now consider correspondences $\{\mathsf{u}_i, \mathsf{v}_i\}_{i=1}^m$ between 2D image points of uncalibrated perspective cameras. Depending on the camera motion, the relationship between two views can be described by either a Fundamental matrix or homography, as expressed by

$$\mathsf{u}^T \mathsf{F} \mathsf{v} \sim 0, \qquad \text{for Fundamental matrix } \mathsf{F} \in \mathbb{R}^{3\times3}. \tag{13}$$

$$\mathsf{u} - \mathsf{H}\mathsf{v} \sim 0, \qquad \text{for Homography } \mathsf{H} \in \mathbb{R}^{3\times3}. \tag{14}$$

An interesting property is that both share the same polynomial subspace $\mathcal{R}_\mathcal{B}$ of second degree polynomials in 2 variables: $\mathcal{B} = \{u_x, u_y, v_x, v_y, u_x v_x, u_x v_y, u_y v_x, u_y, v_y 1\}$. However, the Fundamental matrix (13) is represented by $r = 1$ basis in $\mathcal{R}_\mathcal{B}$, whereas homography (13) is constrained by $r = 3$ bases. The similarity of the polynomials of (13) and (14) allows us to train one network that can handle both cases simultaneously, by simply minimizing one or three singular values. This gives a significant practical advantage over other approaches including RANSAC, that results in an incorrect Fundamental matrix under degenerate motions.

Note that in case of the Fundamental matrix we cannot directly enforce F to be rank-deficient. Given the estimated Fundamental matrix $\hat{\mathsf{F}} \in \mathbb{R}^{3\times3}$, recovered from the basis $\mathsf{B} \in \mathbb{R}^9$ by reshaping, we thus define a regularization term

$$\ell_f(\theta, \mathcal{X}) = \sigma_3(\hat{\mathsf{F}}), \tag{15}$$

that minimizes the last singular value. Again, we add $\lambda_f \ell_f(\theta, \mathcal{X})$ to form the complete loss.

5. Experimental Results

We conduct a variety of experiments to validate the developed theoretical framework and to demonstrate the performance of unsupervised learning for consensus maximization. When comparing with RANSAC, we assume 50% outlier rate and tune the parameters accordingly. In experiments with real data, we compute the ROC curves and select the optimal operating point for each method.

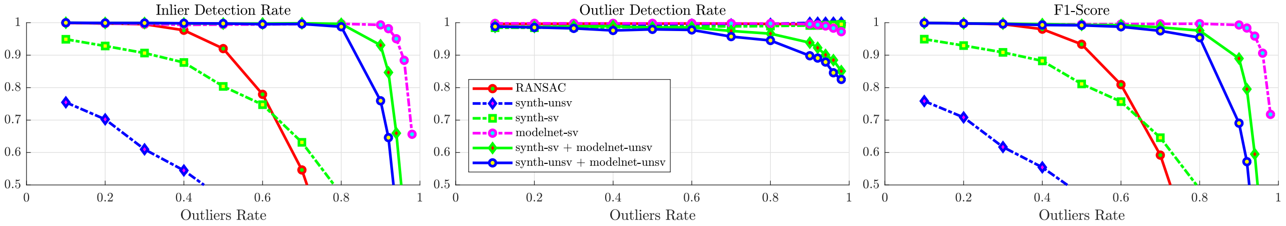

Figure 3. **3D-3D Rigid body transformation estimation with increasing outlier rate.** We evaluate inlier detection rate, outlier detection rate, and F1-score on the ModelNet-40 test set by varying the ratio of synthetically introduced outliers. We compare training on synthetic data and ModelNet-40 training set, for both supervised and unsupervised training setups.

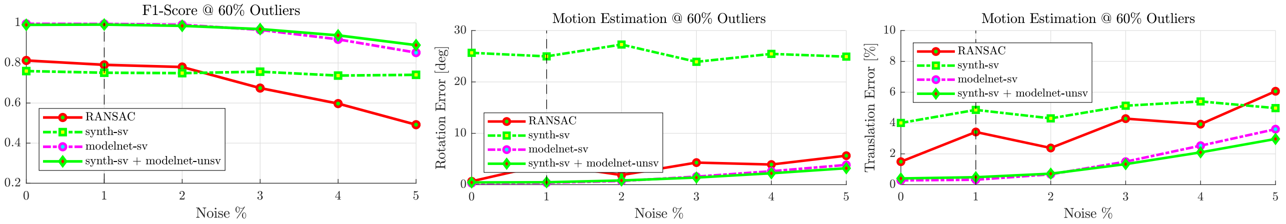

Figure 4. **3D-3D Rigid body transformation estimation with increasing noise level.** From left to right: Evaluation of the impact on F1-score, rotation error, and translation error on ModelNet-40 test set using 60% synthetic outliers.

Synthetic data supported training. Although we are able to train our network from scratch, with competitive results in the presence of up to 50% outliers, the process beyond that becomes delicate. Nevertheless, when supported with initialization using synthetic data, the training behaviour turns out to be as expected, with very competitive results for outliers up to 98%, on real data with very different distributions. Note that we are concerned to maximize the consensus for 3D vision problems. In this context, generating such synthetic data from noise is straightforward. However, as expected, training on synthetic data alone is not enough to get good overall performance on real data, given the immutable difference in statistics.

Implementation details. We trained using a batch size of 64 samples, each containing 512 correspondences. Training was performed for a fixed number of 100 epochs with learning rate decay of 0.9 every 10 epochs, starting from 10^{-3}. The hyperparameters were set as $\lambda = 0.15$, $\lambda_r = \lambda_f = 0.01$. We generated synthetic data for pretraining by uniform sampling of 3D points, and applying a random 6-DoF pose. For the two-view data, we uniformly sampled three euler angles from $[-60, 60]$ degrees and a random translation.

5.1. 3D-3D Rigid Body Transformation

We begin by investigating the behaviour of our method in scenarios of varying outlier rates and noise levels in semi-synthetic experiments on 3D-3D rigid body transformation. The following experiments were conducted on the ModelNet-40 [47] dataset, with the default train/test split. We sampled 512 points from each model, and added 1% noise on the points. We then applied an unconstrained rigid transformation and randomly mixed correspondences to generate the desired number of outliers. In Fig. 3 we plot the inlier and outlier detection rates and the F1-measure in order to compare to several baselines. Here, we evaluate the performance of two methods for pretraining on the synthetic data: supervised (**synth-sv**) and unsupervised (**synth-unsv**). For **synth-unsv**, we trained using the Vandermonde Loss (8), starting with only 10% outlier rate and gradually increasing to 95% outlier rate, whereas **synth-sv** is directly trained using cross entropy loss. Naturally, supervised pretraining performs better. More interestingly, with unsupervised finetuning we observe large improvement over both pretrained methods. Among the finetuned models, **synth-sv+modelnet-unsv** slightly outperforms **synth-unsv+modelnet-unsv** in the domain of high outlier rates, while approaching the performance of the end-to-end supervised method **modelnet-sv**. We can therefore conclude that 1) supervised pretraining is more effective, 2) data specific finetuning is necessary and 3) unsupervised finetuning is able to adapt to the different statistics of ModelNet-40. Moreover, the experiments demonstrate that RANSAC is not useful as a supervisory signal at the outlier rates beyond 60%.

For the second experiment, we varied the noise level from 0% to 5% at a fixed outlier rate of 60%. As depicted in Fig. 4 the F1-score for all trained methods is fairly stable, whereas RANSAC starts deteriorating beyond 2% noise. More importantly, evaluating the rotation and translation errors of the models estimated from the respective inliers, we find that **synth-sv** performs exceptionally bad. We attribute this to the fact that the domain gap results in a fairly good accuracy (similar to RANSAC), but it is unable to reject some very bad outliers, thus leading to an inadequate

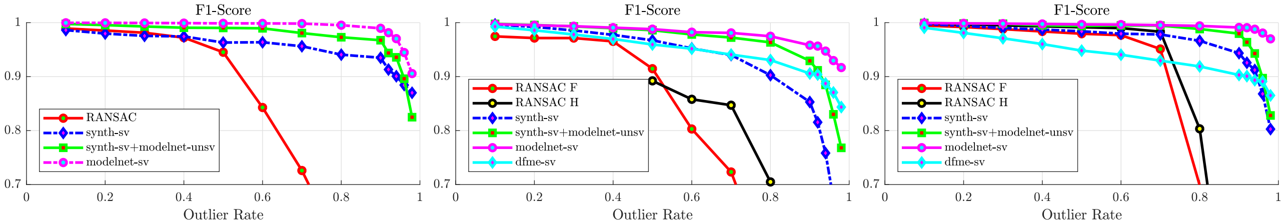

Figure 5. **2-View geometry estimation with increasing outliers.** From left to right: Evaluation on ModelNet-40 test data with Fundamental matrix, on a 50-50 mixture of Homography and Fundamental matrices, and on pure Homography.

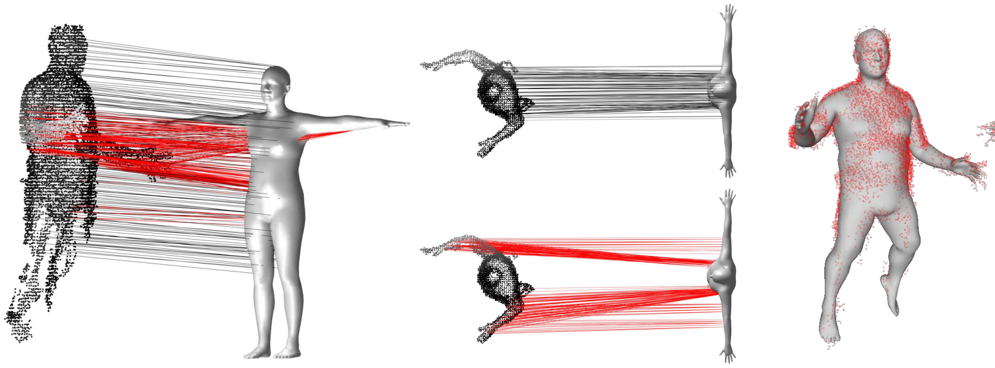

Figure 6. **Non-rigid shape matching.** From left to right: Sparse SfM source point cloud; the reference SMPL [27] model; correspondences classified by our network (inliers in black and outliers in red); top views of inliers (on top) and outliers (at bottom) shown separately; and front and side views of the mesh model obtained after performing articulated ICP initialized using our detected inliers. Red points in the last two images are the SfM point cloud superimposed on the fitted mesh model.

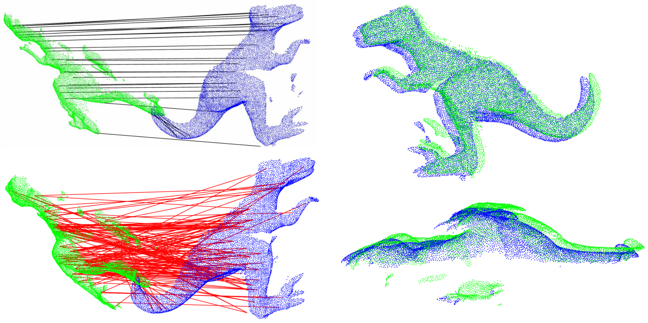

Figure 7. **Rigid body transformation.** Correspondence classification result of our approach on the T-Rex dataset [10] (85% outliers) using PFH [39] matches. Inliers and outliers are visualized on the left side, the aligned models in two views on the right side.

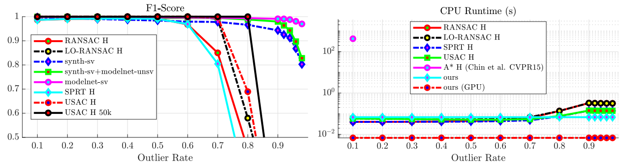

Figure 8. **2D-2D Homography estimation with increasing outliers.** Left: Comparing F1-score with variants of RANSAC on ModelNet-40 test set. Right: Run-time of various methods.

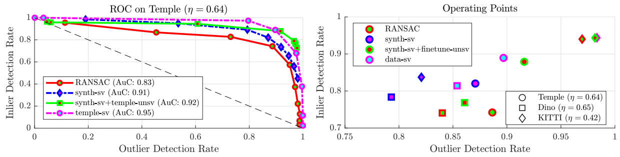

Figure 9. **2D-2D Fundamental matrix estimation.** ROC curve on Middlebury Temple (top). The right figure shows operating points of four methods on three real datasets. η is the outlier rate.

fit. Our method **synth-sv+modelnet-unsv** performs competitive to the supervised model, consistently outperforming other methods in all metrics. We conclude that minimizing our loss gives better classification accuracy and leads to better models, thus validating our theoretical considerations.

In Fig. 7 we show a qualitative result on the T-Rex dataset [10] (85% outliers) of our unsupervisedly trained outlier removal network for PFH [39] feature matches.

5.2. 2D-2D Homography & Fundamental Matrix

We now analyze the performance on 2D-2D Fundamental matrix and homography estimation tasks. First, we investigate the behaviour on semi-synthetically generated data. We projected 3D points from ModelNet-40 data on two views with varying motions and focal lengths, and added 3px noise. The left plot in Fig. 5 compares pretraining, our method, and supervised learning on Fundamental matrix estimation. We observe slight improvement over pretraining, with finetuning getting close to supervised performance. This can be attributed to the similar motion statistics used in pretraining. Here, we omitted comparisons with other supervised methods [36] that showed performance on par with **modelnet-sv**. In the middle and right plot of Fig. 5, we tested on 50% and 100% homographies, respectively. Here, we trained our model on both homography and Fundamental matrix (see Eq. (14) and (13)), only knowing the type of the motion (not the parameters) at training time (not at test time). We compare to RANSAC with Fundamental matrix and homography model, and to a state-of-the-art supervised method for Fundamental matrix estimation [36] **dfme-sv**. We can clearly see that our unsupervised method can handle both cases, as opposed to the single model approaches that fail to to reject outliers by fitting Fundamental matrix to homography data. For up to 80% outliers, the unsupervised approach is very competitive with the fully supervised **modelnet-sv**.

We further compare the F1-score and runtime with various variants of RANSAC [35] and a globally optimal method [11] on homography estimation in Fig. 8. The best baseline USAC works reliably until 70% outlier rate. Extending to 80% involves increasing the max. iterations 100-fold to 50k, severely impacting runtime. The global optimal method [11] is orders of magnitude slower and fails above 10% outliers, making it not practical in general. Note that our implementation is by no means optimized for efficiency.

The next set of experiments was conducted with real matches on three different datasets where the groundtruth motion was known: Temple and Dino from the Middlebury-MultiView dataset [41], and one sequence of KITTI [14]. We disjointly sampled frames for training and testing, computed SIFT keypoints and matched across distance of up to 5 neighboring frames. In the left plot of Fig. 9 we visualize the classification performance by plotting the ROC curve over varying classification thresholds. We observe that unsupervised training gives AuC and operating points very competitive to supervised training. In the middle of Fig. 9 we plot operating points for all methods on all three datasets. Note that the pretrained model **synth-sv** yields suboptimal results and is consistently improved by unsupervised finetuning to a level similar to the supervised model.

5.3. Non-rigid 3D Shape Matching

To gauge the capabilities of our method to learn non-rigid transformations, we experiment on the FAUST [6] dataset. The dataset offers 10 subjects of different body shape in 10 different poses. We take one subject as a reference model, and precompute the geodesic neighborhood. For every training sample, we compute the local loss according to Alg. 1. Training is conducted in a curriculum learning fashion: starting with one global transformation ($K = 512$), we linearly reduce the number of neighbors in every epoch down to $K = 100$. Results on two datasets are reported in Table 1: intra-subject (mostly isometric, pose variation) and inter-subject (non-isometric, similar poses). Initial matches are computed using DFM [25] and KM [22]. Again, we observe that the unsupervised method adapts to the method-specific outlier statistics, whereas the method pretrained on synthetic outliers fails to generalize. Compared to the isometric consensus maximization method MFCM [33], we loose more inliers, which can be attributed to the fact that piecewise rigidity is not the entirely correct deformation model. In Fig. 6 we give a qualitative result on matching a sparse SfM point cloud, reconstructed from 72 images using Colmap [40] to the SMPL [27] reference model. We train our network to filter matches from KM [22]. The resulting set of matches enables ICP-based refinement [5] on body pose and shape.

	Inlier / Outl.	Time [s]	Inl. / Outlier	Time [s]
DFM [25]	4211 / 772	1.0	3756 / 1227	1.0
MFCM [33]	3918 / **31**	24	3437 / 93	19
synth-sv	3812 / 74	**0.8**	2601 / 82	**0.8**
synth-sv+unsv	3814 / 58	**0.8**	3122 / **11**	**0.8**
KM [22]	4736 / 181	89	4051 / 860	92
MFCM [33]	4556 / **17**	110	3634 / 161	115
synth-sv	3811 / 23	**0.8**	2371 / 171	**0.8**
synth-sv+unsv	3957 / 19	**0.8**	3303 / **40**	**0.8**
	Intra-subject		Inter-subject	

Table 1. **Non-rigid 3D shape matching.** Our results on FAUST [6] intra- and inter-subjects vs. [22] and [33]. We report the number of true positive (inliers) and false positive (remaining outliers) matches, as well as run time on CPU.

6. Conclusions

In this paper we introduced a method for unsupervised learning of consensus maximization. Based on the relationship between inlier measurements, represented by an ideal of the inlier set, and the subspace of polynomials representing the space of target transformations, our formulation allows to train a neural network for inlier/outlier classification, without knowing ground truth correspondences or the model parameters. Our experiments confirm that there is a huge potential in adapting learning-based methods to unseen data domains. We demonstrate on a diverse set of 3D vision problems that our method can successfully finetune to new data without external supervision, thus replicating the generic behaviour of RANSAC. For future work, we are interested in investigating the case where the type of transformation is also unknown, by jointly finding the polynomial basis $\mathcal{R}_{\mathcal{B}}$, on which suitable polynomials reside.

Acknowledgements. This research was funded by the EU Horizon 2020 programme, grant No. 687757 – REPLICATE and by the Swiss Commission for Technology and Innovation (CTI), Grant No. 26253.1 PFES-ES – EXASOLVED.

References

[1] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D. G. Murray, B. Steiner, P. A. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu, and X. Zhang. Tensorflow: A system for large-scale machine learning. In *OSDI*, 2016. 4

[2] J.-C. Bazin, H. Li, I. S. Kweon, C. Demonceaux, P. Vasseur, and K. Ikeuchi. A branch-and-bound approach to correspondence and grouping problems. *IEEE transactions on pattern analysis and machine intelligence*, 35(7):1565–1576, 2013. 1, 2

[3] J. C. Bazin, Y. Seo, R. I. Hartley, and M. Pollefeys. Globally optimal inlier set maximization with unknown rotation and focal length. In *ECCV*, 2014. 1, 2

[4] A. Björck and V. Pereyra. Solution of vandermonde systems of equations. *Mathematics of Computation*, 24(112):893–903, 1970. 2

[5] F. Bogo, A. Kanazawa, C. Lassner, P. V. Gehler, J. Romero, and M. J. Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *ECCV*, 2016. 8

[6] F. Bogo, J. Romero, M. Loper, and M. J. Black. FAUST: Dataset and evaluation for 3D mesh registration. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Piscataway, NJ, USA, 2014. IEEE. 8

[7] E. Brachmann, A. Krull, S. Nowozin, J. Shotton, F. Michel, S. Gumhold, and C. Rother. Dsac-differentiable ransac for camera localization. In *CVPR*, volume 3, 2017. 1, 2

[8] P. Breiding, S. K. Verovsek, B. Sturmfels, and M. Weinstein. Learning algebraic varieties from samples. *arXiv preprint arXiv:1802.09436*, 2018. 2

[9] P. Breiding, S. K. Verovsek, B. Sturmfels, and M. Weinstein. Learning algebraic varieties from samples. *arXiv preprint arXiv:1802.09436*, 2018. 2

[10] T. J. Chin, Y. H. Kee, A. Eriksson, and F. Neumann. Guaranteed outlier removal with mixed integer linear programs. In *CVPR*, 2016. 1, 2, 7

[11] T.-J. Chin, P. Purkait, A. P. Eriksson, and D. Suter. Efficient globally optimal consensus maximisation with tree search. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2413–2421, 2015. 8

[12] O. D. Faugeras. What can be seen in three dimensions with an uncalibrated stereo rig. In *ECCV*, pages 563–578, 1992. 1, 2

[13] M. A. Fischler and R. C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6):381–395, 1981. 1, 2

[14] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 8

[15] R. I. Hartley and F. Kahl. Global optimization through rotation space search. *IJCV*, 82(1):64–79, 2009. 1, 2

[16] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521540518, second edition, 2004. 1, 2

[17] J. A. Hesch and S. I. Roumeliotis. A direct least-squares (DLS) method for PnP. In *ICCV*, pages 383–390, 2011. 1, 2

[18] Y. Hu, D. Zhang, J. Ye, X. Li, and X. He. Fast and accurate matrix completion via truncated nuclear norm regularization. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(9):2117–2130, 2013. 3

[19] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. 4

[20] L. Kneip, H. Li, and Y. Seo. Upnp: An optimal o(n) solution to the absolute pose problem with universal applicability. In D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, editors, *Computer Vision – ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I*, 2014. 1, 2

[21] S. Kumar, Y. Dai, and H. Li. Monocular dense 3d reconstruction of a complex dynamic scene from two perspective frames. In *ICCV*, 2017. 5

[22] Z. Lähner, M. Vestner, A. Boyarski, O. Litany, R. Slossberg, T. Remez, E. Rodolà, A. M. Bronstein, M. M. Bronstein, R. Kimmel, and D. Cremers. Efficient deformable shape correspondence via kernel matching. In *3DV*, 2017. 8

[23] H. Li. Consensus set maximization with guaranteed global optimality for robust geometry estimation. In *ICCV*, 2009. 1, 2

[24] J. Li, B. M. Chen, and G. H. Lee. So-net: Self-organizing network for point cloud analysis. *CoRR*, abs/1803.04249, 2018. 2, 4

[25] O. Litany, T. Remez, E. Rodola, A. M. Bronstein, and M. M. Bronstein. Deep functional maps: Structured prediction for dense shape correspondence. In *ICCV*, 2017. 8

[26] H. Longuet-Higgins. A computer algorithm for reconstructing a scene from two projections. *Nature*, 293:133–135, 1981. 1

[27] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, 2015. 7, 8

[28] R. A. Newcombe, D. Fox, and S. M. Seitz. Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. In *CVPR*, 2015. 5

[29] T. Nguyen, S. W. Chen, S. S. Shivakumar, C. J. Taylor, and V. Kumar. Unsupervised deep homography: A fast and robust homography estimation model. *IEEE Robotics and Automation Letters*, 3:2346–2353, 2018. 2

[30] D. Nistér. An efficient solution to the five-point relative pose problem. *IEEE Trans. Pattern Anal. Mach. Intell.*, 26(6):756–777, 2004. 1, 2

[31] D. P. Paudel, A. Habed, C. Demonceaux, and P. Vasseur. Robust and optimal registration of image sets and structured scenes via sum-of-squares polynomials. *International Journal of Computer Vision*, 127:415–436, 2018. 2

[32] O. Poursaeed, G. Yang, A. Prakash, Q. Z. Fang, H. Jiang, B. Hariharan, and S. Belongie. Deep fundamental matrix estimation without correspondences. 2018. 2

[33] T. Probst, A. Chhatkuli, D. P. Paudel, and L. V. Gool. Model-free consensus maximization for non-rigid shapes. *CoRR*, abs/1807.01963, 2018. 1, 2, 8

[34] C. R. Qi, H. Su, K. Mo, and L. J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 77–85, 2017. 2, 4

[35] R. Raguram, O. Chum, M. Pollefeys, J. Matas, and J.-M. Frahm. Usac: a universal framework for random sample consensus. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(8):2022–2038, 2013. 1, 2, 8

[36] R. Ranftl and V. Koltun. Deep fundamental matrix estimation. In *ECCV*, pages 284–299, 2018. 1, 2, 8

[37] P. J. Rousseeuw. Least median of squares regression. *Journal of the American statistical association*, 79(388):871–880, 1984. 1, 2

[38] C. Russell, R. Yu, and L. Agapito. Video pop-up: Monocular 3d reconstruction of dynamic scenes. In *ECCV*, 2014. 5

[39] R. B. Rusu, N. Blodow, and M. Beetz. Fast point feature histograms (fpfh) for 3d registration. *2009 IEEE International Conference on Robotics and Automation*, pages 3212–3217, 2009. 7

[40] J. L. Schönberger and J.-M. Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 8

[41] S. M. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, 1:519–528, 2006. 8

[42] P. Speciale, D. P. Paudel, M. R. Oswald, T. Kroeger, L. V. Gool, and M. Pollefeys. Consensus maximization with linear matrix inequality constraints. In *CVPR*, 2017. 1, 2

[43] J. Taylor, A. D. Jepson, and K. N. Kutulakos. Non-rigid structure from locally-rigid motion. In *CVPR*, 2010. 5

[44] P. H. Torr and A. Zisserman. Mlesac: A new robust estimator with application to estimating image geometry. *Computer vision and image understanding*, 78(1):138–156, 2000. 1, 2

[45] N. Verma, E. Boyer, and J. Verbeek. Feastnet: Feature-steered graph convolutions for 3d shape analysis. 2017. 5

[46] Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. M. Bronstein, and J. M. Solomon. Dynamic graph cnn for learning on point clouds. *CoRR*, abs/1801.07829, 2018. 2, 4

[47] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao. 3d shapenets: A deep representation for volumetric shapes. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1912–1920, 2015. 6

[48] K. M. Yi, E. Trulls, Y. Ono, V. Lepetit, M. Salzmann, and P. Fua. Learning to find good correspondences. In *CVPR*, 2018. 1, 2

[49] Y. Zheng, S. Sugimoto, and M. Okutomi. Deterministically maximizing feasible subsystem for robust model fitting with unit norm constraint. In *CVPR*, 2011. 1, 2