2019 IEEE Winter Conference on Applications of Computer Vision

Alignment by Composition

Berk Sevilmis Benjamin B. Kimia
Brown University
{berk_sevilmis, benjamin_kimia}@brown.edu

Abstract

We propose an unsupervised method to establish dense semantic correspondences between images depicting different instances of the same object category. We posit that alignment is compositional in nature and requires the detection of a similar visual concept between images. We realize this in a top-down fashion using objectness, saliency, and visual similarity cues to co-localize the regions of holistic foreground objects. Jointly maximizing visual similarity and bounding the geometric distortion induced by their configuration, the target foreground object is then composed by the subregions of the source foreground object. The resultant composition is used to form a dense motion field enabling the alignment. Experimental results on several benchmark datasets support the efficacy of the proposed method.

1. Introduction

Image alignment is one of the fundamental problems in computer vision, finding many applications such as optical flow [8], stereo matching [24], scene parsing [18], video depth estimation [10], image enhancement [5], *etc.*

The difficulty in aligning images from the same 3D scene, as encountered in optical flow estimation or in wide-baseline stereo matching, lies in photometric changes and/or changes in the image acquisition process such as view geometry. The dense motion field is expected to encode these external factors and/or the displacement of objects in the scene. In the task of semantic alignment, however, the goal is to establish anatomical and/or functional correspondences between the images. The dense motion field is expected to encode not only the factors mentioned above, but also the intra-class variations between the instances of objects being imaged. The task becomes further challenging when object instances are viewed in disparate scenes forming background clutter.

Many of the earlier proposed works cast the semantic alignment problem as a smooth registration of densely and locally extracted features used to represent regions of a predefined spatial support. This type of an approach neglects to discover the extent of the existent semantic overlap between the pair of images. As a result, the alignment process is mistakenly dominated by the futile effort of matching outlier features arising from background clutter.

In this paper, we propose an unsupervised method to align image pairs which are semantically related. To this end, we advocate a top-down compositional process. The alignment is only meaningful between instances of objects and/or between scenes of similar appearance. This requires detecting akin visual concept demonstrated in the pair of images. We co-localize such regions using objectness, saliency and appearance similarity measures. The alignment is then guided by a region matching process targeted to compose one of the images from the other by discovering similar looking subregions under a low distortion piecewise affine geometric transformation model, see Fig. 1.

In summary, our key contributions are: (1) the use of co-localization to discover foreground regions of similar appearance by simultaneous consideration of image pairs, and (2) the introduction of a low geometric distortion inducing composition of regions in aligning the semantically related visual concept. Experiments show promising results supporting the effectiveness of our compositional process.

The rest of the paper is organized as follows. In Section 2, we review the relevant work on semantic image alignment. In Section 3, we introduce the algorithmic details of the proposed method. Section 4 presents the benchmark datasets and experimental results and Section 5 concludes the paper.

2. Related work

The prior work on semantic image alignment could be broadly categorized into unsupervised and supervised methods. We provide a brief, non-exhaustive overview of some of the techniques proposed so far.

Unsupervised methods. SIFT Flow [19] is one of the earliest papers to attempt registering images of different but similar looking scenes. SIFT features [21] at a predefined scale and orientation are densely extracted and matched hierarchically using loopy belief propagation. The method successfully aligns pairs of images which are already fairly

978-1-7281-1975-5/19/$31.00 ©2019 IEEE
DOI 10.1109/WACV.2019.00218

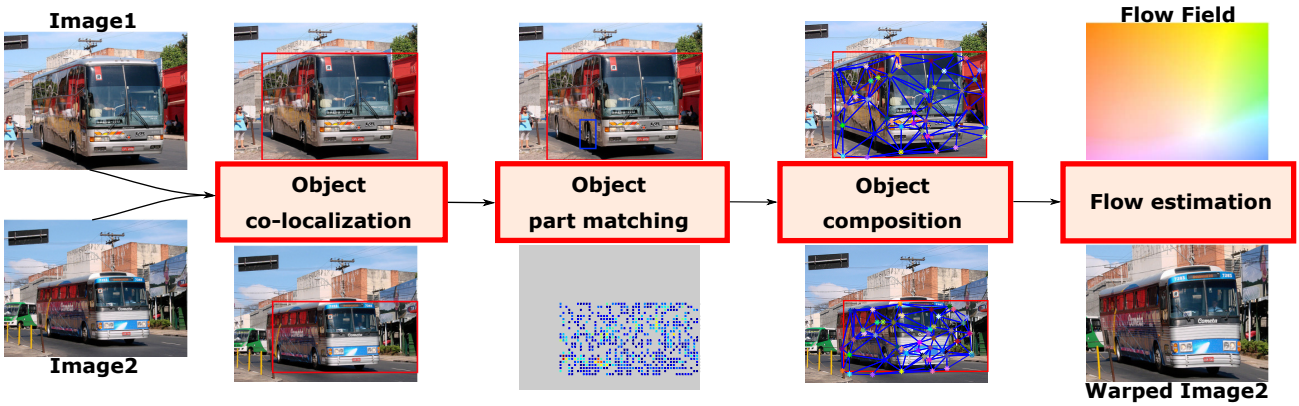

Figure 1: **Alignment by composition.** We align a pair of semantically related input images by means of a top-down compositional process. Holistic foreground objects are first co-localized and a set of object part proposals from Image1 is extracted. Each object part proposal seeks a region match support in Image2 in a coarse sliding window search fashion. Piece-wise affine transformations with bounded conformal distortion are used to model the geometric variation between the images. Simultaneously maximizing the region matching scores and bounding the distortions of the piece-wise affine transformations, a dense motion field is estimated to compose the foreground object in Image1 with regions from Image2.

aligned and which have similar view geometry. Deformable Spatial Pyramid (DSP) [11] provides modest improvement upon SIFT Flow by considering a pyramid graph consisting of cells whose spatial support covers a set of pixels providing context. These methods are highly challenged when objects and/or scenes are viewed under different scales and orientations. Scale-Space SIFT Flow [22] extends the search space by extracting a set of SIFT descriptors with different scales. The Generalized Deformable Spatial Pyramid [9] considers a set of different orientations as well. The Subpixel Semantic Flow [25] uses Geometric Blur [2] descriptors and continuous domain variational formulation to obtain subpixel resolution flow fields.

The Generalized PatchMatch [1] algorithm and the Daisy Filter Flow [31] are proposed to obtain fast dense correspondences. The former sacrifices geometric consistency and aims to improve a random correspondence field by propagation and random search, while the latter uses filter-based inference to obtain geometrically more coherent matches.

Proposal Flow (PF) [6] extracts object proposals and uses region matching to align images. Taniai *et al.* [26], related to our work, aims to jointly cosegment the foreground regions and establish a dense correspondence between them by a hierarchical Markov random field model. Our method does not rely on obtaining accurate segmentations and we use higher order potentials to constrain the distortion of the geometric transformation model relating the pair of images. Yang *et al.* [30] uses DSP-like graph representation of the estimated foreground region which is obtained using saliency cues. Unlike SIFT descriptors used in DSP, the grid cells of the graph are represented by HOG features [7], and the method trains an online discriminative classifier for each cell. Contrary to their approach, our foreground object co-localization utilizes information from both of the images as salient regions in one of the images might not arise from the similar visual concept pictured in the pair.

Supervised methods. Convolutional neural networks (CNNs) have also been used to train feature embeddings that can be robustly matched or to estimate the parameters of a geometric transformation model in order to semantically align a pair of images. The biggest difficulty faced in training a neural network is the lack of an abundant set of supervisory annotation.

Zhou *et al.* [33] uses 3D CAD models as auxiliary dense pixel-wise annotations to train a network that outputs a flow field. Geometric CNN [23] estimates the parameters of an affine transformation as well as a thin-plate spline (TPS) interpolant to align images in two steps. The affine transformation, first, globally aligns the images and the TPS interpolant accounts for the residual local deformations. Fully convolutional self similarity (FCSS) [12] network trains a descriptor based on local self-similarity and uses SIFT Flow and PF based optimization. Discrete continuous trnasformation matching (DCTM) [13] uses the FCSS descriptor, however, formulates an iterative discrete-continuous optimization. Ufer *et al.* [28] uses AlexNet [14] based deep features to generate a set of candidate point correspondences and uses a matching objective consisting of unary and pairwise match potentials.

3. Approach

Given two images I_1 and I_2, our goal is to find a flow vector $\mathbf{w}(\mathbf{p}) = (u(\mathbf{p}), v(\mathbf{p}))$ at each point $\mathbf{p} = (x, y)$ in I_1 to comprise a dense motion field enabling semantic registration. Fig. 1 shows the pipeline of our algorithm. We detail each component of our framework in the following subsections.

3.1. Object co-localization

Though the pair of input images are semantically related, the different instances of the same category objects that are meaningful to align need to be identified, as the objects could be viewed in disparate backgrounds. We first use the Selective Search (SS) [29] algorithm to generate proposals consisting of holistic objects and object parts. Let R_1 and R_2 denote the sets of extracted region proposals from I_1 and I_2 respectively. We propose the following optimization function to co-localize the foreground objects:

$$\{\overline{r_1}, \overline{r_2}\} = \underset{r_1 \in R_1, r_2 \in R_2}{\arg\max}\ O(r_1, r_2) + S(r_1, r_2) + A(r_1, r_2). \tag{1}$$

The definitions and the justifications of each term are as follows:

Objectness term $O(r_1, r_2)$. The object category similarity of a region pair, (r_1, r_2), could be quantified with an object detection framework. We run the Region based Fully Convolutional Network (R-FCN) [4] on the input images to obtain a set of regions of detected objects and their labels, *i.e.,* $\{D_i, L_i\}$ $i = 1, 2$. Using the detected object regions and their category labels, we then assign each proposal a category score which is the maximum Intersection-over-Union (IoU) score achieved with one of the detected objects, and a category label which is of that detected object, *i.e.,*

$$\text{score}(r) = \text{IoU}(r, d)\ r \in R_i,\ d \in D_i,\ i = 1, 2. \tag{2}$$

$$\ell_r = \ell_{d^*} \text{ where } d^* = \underset{d \in D_i}{\arg\max}\ \text{IoU}(r, d)\ r \in R_i,\ i = 1, 2. \tag{3}$$

Based on the scores and the labels of the proposals, the objectness term is defined as follows:

$$O(r_1, r_2) = \frac{1}{2} \cdot \mathbb{1}_{(\ell_{r_1} = \ell_{r_2})} \cdot [\text{score}(r_1) + \text{score}(r_2)] \tag{4}$$

which helps eliminate selecting regions bounding objects from different categories.

Saliency term $S(r_1, r_2)$. As being the subject of the image, foreground objects tend to be more salient than their surrounding background. We use deep learning based features [16] to compute saliency maps. Let $SM_i : I_i \rightarrow [0, 1]$, $i = 1, 2$ denote a real-valued saliency map computed on an input image. Every proposal is assigned a saliency value which is the average saliency computed on a proposal's spatial support. To help eliminate selecting probably background regions, a proposal pair having high saliency is favored to be picked:

$$S(r_1, r_2) = \tfrac{1}{2} \cdot \left[\frac{\sum_{\mathbf{p} \in r_1} SM_1(\mathbf{p})}{\text{area}(r_1)} + \frac{\sum_{\mathbf{p} \in r_2} SM_2(\mathbf{p})}{\text{area}(r_2)} \right]. \tag{5}$$

Appearance similarity term $A(r_1, r_2)$. While objectness and saliency terms are useful in filtering the regions arising from background clutter, the use of appearance similarity term is twofold: to assess the visual similarity of proposals and to select regions bounding holistic objects rather than object parts. We use a slight variant of the standout score proposed in [3] to compute the appearance similarity of a region pair (r_1, r_2):

$$A(r_1, r_2) = F(r_1, r_2) - \max_{r_{1b} \in B(r_1)} F\left(r_{1b}, \underset{q_2 \in R_2}{\arg\max} F(r_{1b}, q_2)\right)$$
$$B(r_1) = \{r_{1b} \,|\, r_1 \subsetneq r_{1b}, r_{1b} \in R_1\} \tag{6}$$

where $F(\cdot)$ measures the similarity of the features extracted from the regions and $B(\cdot)$ returns regions which enclose the region input as its argument. In an unsupervised setting, for the task of semantic correspondence, it has been shown that deep features extracted from mid-layers of CNNs trained for object detection achieved similar performance as handcrafted low level features [20]. Throughout our work, we use the whitened HOG features [7] to extract local visual information. We use the same inclusive relation used in [3] to construct the set $B(r_1)$: (1) the area of the region, r_1, is at most 50% of the areas of the regions surrounding it, $\{r_{1b}\}$, and (2) the surrounding regions contain at least 80% of the area of the region considered.

Eq. 6 measures the appearance similarity contrast by considering the difference between the appearance similarity of the candidate regions and the maximal similarity achieved by the regions containing them. As stated in [3], the standout score is high when the regions considered contain holistic objects of the same category, and is relatively low when the regions instead contain object parts. This is due to the fact that the appearance similarity contrast is maximized when the candidate region is the tight holistic object bounding box separating the foreground object from the background for which the appearance similarity is expected to be low.

3.2. Object part matching

The holistic object and the object part proposals generated by the SS may not be perfectly repeatable across different images. In other words, an object part such as the wheel of a car might exist in the pool of proposals in one image and might not be included in the other. To overcome the inherent imperfection in the region generation process, after the foreground regions are co-localized, we use a coarse sliding window search for every region contained in the foreground in I_1 satisfying the above-mentioned inclusive relation, *i.e.,* $\forall r \subsetneq \overline{r_1}$, $r \in R_1$ to generate matching candidates for the object parts. The notion of parts we use in this context do not need to be semantic or functional as certain portions of objects could be matched seamlessly with the embedded feature considered.

The foreground objects could be viewed under different scales. To be able to robustly match its parts with

varying scales in a sliding window fashion, we make use of the relative scale information captured by the co-localized foreground bounding boxes $\overline{r_1}$ and $\overline{r_2}$. Let $\{TL_x, TL_y, BR_x, BR_y\}$ denote the top-left and bottom-right coordinates of a bounding box. We use the following simple transformation, $\tau_{\{s_x,s_y,t_x,t_y\}}$, involving only scale and translation, to relate the coordinates of $\overline{r_1}$ and $\overline{r_2}$ and hence any region from R_1 to R_2, see Fig. 2:

$$\begin{bmatrix} TL_x|_{\overline{r_2}} \\ TL_y|_{\overline{r_2}} \end{bmatrix} = \begin{bmatrix} s_x & 0 \\ 0 & s_y \end{bmatrix} \cdot \begin{bmatrix} TL_x|_{\overline{r_1}} \\ TL_y|_{\overline{r_1}} \end{bmatrix} + \begin{bmatrix} t_x \\ t_y \end{bmatrix}$$
$$\begin{bmatrix} BR_x|_{\overline{r_2}} \\ BL_y|_{\overline{r_2}} \end{bmatrix} = \begin{bmatrix} s_x & 0 \\ 0 & s_y \end{bmatrix} \cdot \begin{bmatrix} BR_x|_{\overline{r_1}} \\ BR_y|_{\overline{r_1}} \end{bmatrix} + \begin{bmatrix} t_x \\ t_y \end{bmatrix} \quad (7)$$

where $\{s_x, s_y, t_x, t_y\}$ represent the relative scale and translation in x and y coordinates respectively. Using s_x and s_y, we transform the scales of the regions contained in $\overline{r_1}$ and do coarse sliding window search within $\overline{r_2}$ in I_2 with a stride of 8 pixels to obtain object part matching responses.

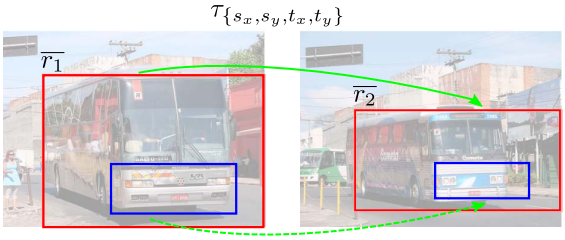

Figure 2: **The similarity transformation.** The bounding boxes of the foreground objects could be used to define a similarity transformation, *i.e.,* $\tau_{\{s_x,s_y,t_x,t_y\}}$ shown by the solid green curve. The same transformation could be used to infer scales and positions of object parts, one of which is shown by the dashed green curve.

3.3. Object composition

Object part matching responses computed using the coarse sliding window search provide candidate region matches between the input images. Such a process which only uses visual similarity cues lacks geometric consistency. The rear wheel of a car, in isolation, can be visually very similar to the front wheel and hence can attain a high region matching score. Such a match is invalidated only when a simultaneous organization of parts is considered. In order to establish dense correspondences between the input images, we aim to compose the foreground region $\overline{r_1}$ from the regions contained in $\overline{r_2}$ organized as a graph capturing the interactions between object parts. To this end, we first list the desired properties of an ideal compositional process.

The regions that compose $\overline{r_1}$ should: (1) have high object part matching scores indicating successful and distinctive visual alignment, (2) cover the domain of $\overline{r_1}$ as much as possible, and (3) have a configuration resulting in a low geometric distortion as objects are naturally coherent in form

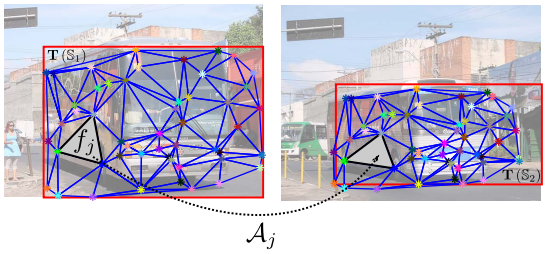

Figure 3: **The Delaunay triangulation.** We capture the geometry of object parts using Delaunay triangulations. The vertices are the center coordinates of object part regions and faces are triplets of noncollinear vertices. Each face f_j is only allowed to undergo an affine transformation $\mathcal{A}_j$ with a bounded conformal distortion.

and can only undergo a certain extent of deformation. We propose to optimize an energy functional which targets to select a maximal subset of the object part matching candidates satisfying these properties. Before elaborating the energy functional considered, we introduce some notation.

Let $\mathcal{S}_1 = \{r_i\}$, $i = 1, .., N$ represent a set of object part regions contained in $\overline{r_1}$. For each region in $\overline{r_1}$, we keep the K best matching object part regions found in $\overline{r_2}$ resulting in the set $\mathcal{S}_2 = \{\hat{r_{i,k}}\}$, $k = 1, .., K$. Let $\mathbb{S}_1 = \{r_{i_o}\}$, $o = 1, .., n \leq N$, $i_o \in [1..N]$ denote the maximal subset of object part regions selected from $\overline{r_1}$ maximizing the energy functional. Similarly, let $\mathbb{S}_2 = \{\hat{r_{i_o,k_o}}\}$, $k_o \in [1..K]$ denote the maximal subset of matched object part regions in $\overline{r_2}$.

We represent the organization of object parts using Delaunay triangulations. Let $\text{center}(r)$ denote an operator returning the center coordinate of a region r. The Delaunay triangulation of the centers of object part regions in $\mathbb{S}_1$ is denoted with $\mathbf{T}(\mathbb{S}_1) = (\mathbf{V}(\mathbb{S}_1), \mathbf{F}(\mathbb{S}_1))$ representing the resulting vertices and faces of the triangulation. $\mathbf{V}(\mathbb{S}_1) = \{\text{center}(r_{i_o})\}$ is the set of points and $\mathbf{F}(\mathbb{S}_1) = \{f_j\}$ is the set of faces indexed by triplets of noncollinear vertices. $\mathbf{T}(\mathbb{S}_2)$ is defined similarly. We consider piecewise affine transformations to relate the matched point sets $\{\text{center}(r_{i_o})\}$ and $\{\text{center}(\hat{r_{i_o,k_o}})\}$ so that the vertices of the face j undergoes the affine transformation $\mathcal{A}_j$. The conformal distortion of each $\mathcal{A}_j$ is measured using its linear part as defined in [17], *i.e.,* $\mathbf{D}(\mathcal{A}_j) = \frac{\sigma_{\max}(\mathcal{A}_j)}{\sigma_{\min}(\mathcal{A}_j)}$ where $\sigma_{\max}(\mathcal{A}_j)$ and $\sigma_{\min}(\mathcal{A}_j)$ denote the maximum and minimum singular values of the affine transformation $\mathcal{A}_j$, see Fig. 3.

The energy functional we are interested in maximizing is as follows:

$$E(\mathbb{S}_1, \mathbb{S}_2) = \max_{\substack{\mathbb{S}_1 \subseteq \mathcal{S}_1 \\ \mathbb{S}_2 \subseteq \mathcal{S}_2}} \psi_1(\mathbb{S}_1, \mathbb{S}_2) + \psi_2(\mathbb{S}_1) + \psi_3(\mathbb{S}_1)$$
$$s.t. \quad \max_j \mathbf{D}(\mathcal{A}_j) \leq C \quad (8)$$

Normalized Matching Score $\psi_1(\mathbb{S}_1, \mathbb{S}_2)$. We want to select regions with high matching scores signalling successful

visual alignment as stated in property (1). To achieve this, we use the following term in the objective functional:

$$\psi_1(\mathbb{S}_1, \mathbb{S}_2) = \frac{\sum_{o=1}^{n} F(r_{i_o}, r_{\hat{i_o,k_o}})}{\max\limits_{\substack{r_i \in \mathcal{S}_1 \\ r_{\hat{i,k}} \in \mathcal{S}_2}} F(r_i, r_{\hat{i,k}})} \tag{9}$$

which normalizes the object part matching scores.

Normalized Grid Score $\psi_2(\mathbb{S}_1)$. Representing regions with their center coordinates normalizes them across regional area. This representation helps define a maximal cover of the domain of $\overline{r_1}$ without biasing the process towards selecting regions with greater area. We first divide $\overline{r_1}$ into m tiles. Let Q_m denote the region of the mth tile. We count the number of tiles covered by $\mathbb{S}_1$ normalized by its maximum achievable value by $\mathcal{S}_1$ which defines the normalized grid score:

$$\psi_2(\mathbb{S}_1) = \frac{|\{m : \exists\, \text{center}(r_{i_o}) \in Q_m\}|}{|\{m : \exists\, \text{center}(r_i) \in Q_m\}|} \tag{10}$$

where $|\cdot|$ is the cardinality operator.

Normalized Convex Hull Score $\psi_3(\mathbb{S}_1)$. Though the average grid score, $\psi_2(\mathbb{S}_1)$, targets to select as many uniquely centered regions as possible, it has no control over their configuration. The regions corresponding to the center coordinates that are nearby are likely to cover $\overline{r_1}$ locally compared to the regions whose center coordinates are far apart. Hence, the convex hull score, on the other hand, can complement the normalized grid score by maximizing the spatial diversity of the selected regions in $\mathbb{S}_1$:

$$\psi_3(\mathbb{S}_1) = \frac{\text{area}(\text{Conv}(\{\text{center}(r_{i_o})\}))}{\text{area}(\text{Conv}(\{\text{center}(r_i)\}))} \tag{11}$$

where Conv denotes the convex hull. The normalized grid and convex hull scores together are used to achieve the (2)nd desirable property listed.

The faces, $\mathbf{F}(\mathbb{S}_1)$, are assumed to undergo piecewise affine transformations. By simultaneously composing $\overline{r_1}$ and bounding the maximal conformal distortion of the faces of the Delaunay triangulation, $\mathbf{T}(\mathbb{S}_1)$, we aim to realize a compositional process satisfying all the abovementioned properties.

3.3.1 Optimization

Beam Search. Maximizing the objective functional in Eq. 8 is NP-hard. Note that the objective functional without the constraints is unbounded, *i.e.,* the maximum is achieved when $\mathbb{S}_1 = \mathcal{S}_1$. However, the cardinality of $\mathbb{S}_1$ is not known beforehand when constraints are considered as well. To propose a solution, we opt for using a greedy algorithm based on beam search [27]. Beam search is a heuristic algorithm which, at each iteration, expands a solution set from the most promising limited set of solutions of the previous iteration. We start with an empty solution set and try to populate as many matching candidates as possible maximizing the objective while satisfying the distortion constraints, see Algorithm 1.

Algorithm 1 Beam Search

Input: $\mathcal{S}_1, \mathcal{S}_2$, and b
Output: $\mathbb{S}_1$ and $\mathbb{S}_2$
Let $result$ be a max priority queue of size 1 storing the solution sets $\mathbb{S}_1, \mathbb{S}_2$ and its priority being the energy defined in Eq. 8.

Let $beam$, and $beam^*$ be max priority queues of size b storing the candidate solution sets sorted with respect to their energies defined in Eq. 8.

procedure BEAMSEARCH($\mathcal{S}_1, \mathcal{S}_2, b$)
  $result \leftarrow \{\}$
  $beam \leftarrow \{\}$
  Pick the best b candidate solution sets of size 3 using brute force search maximizing Eq. 8 and update $beam$ and $result$.
  repeat
    $energy \leftarrow$ OBJECTIVEVALUE($result$)
    $beam^* \leftarrow \{\}$
    for $c \in beam$ **do**
      $result$.push(c)
      for $i = 1$ to N **do**
        if $(\nexists e \in c : \text{center}(r_i) = \text{center}(e))$ **then** ▷ There does not exist any element in c with center coordinate center (r_i)
          for $k = 1$ to K **do**
            add $c \cup \{r_i, \text{center}(r_i), r_{\hat{i,k}}, \text{center}(r_{\hat{i,k}})\}$ to $beam^*$
          end for
        end if
      end for
    end for
    $beam \leftarrow beam^*$
  until $energy =$ OBJECTIVEVALUE($result$)
  return $result.top()$
end procedure

function OBJECTIVEVALUE($result$)
  return $\begin{cases} -\infty & \text{if } result = \{\} \\ E(result.top()) & \text{otherwise} \end{cases}$
end function

Factor Graph. Once an approximate solution, $\mathbb{S}_1$ and $\mathbb{S}_2$, to Eq. 8 is obtained, we propose to fine-tune it as the region matches might not be perfectly localized due to the coarse sliding window search process. To this end, we do a dense but local sliding window search around the center coordinates of the regions in $\mathbb{S}_2$, and generate at most Z locally maximal fine match candidates for each region in $\mathbb{S}_1$. Let $\mathbb{S}_3 = \{r_{\tilde{i_o,z_o}}\}$, $z_o \in [1..Z]$ denote the set of fine match region candidates. We consider a factor graph model defined on the graph of the Delaunay triangulation introduced in the subsection 3.3 *i.e.,* $\mathbf{T}(\mathbb{S}_1) = (\mathbf{V}(\mathbb{S}_1), \mathbf{F}(\mathbb{S}_1))$. Our factor graph model contains two types of factors: (1) a factor function for each vertex quantifying the region matching and (2) a factor function for each face representing the conformal distortion induced by its respective affine transformation. Intuitively, these two types of factors serve for simultaneously finding a high scoring configuration of matches with low geometric distortion. The factor graph model is given as follows:

$$P_\Phi(\mathbb{S}_3) = \frac{1}{Z} \prod_{o=1}^{n} \phi_{\mathbf{V}}(r_{i_o}, r_{\tilde{i_o,z_o}}) \prod_{j=1}^{|\mathbf{F}(\mathbb{S}_1)|} \phi_{\mathbf{F}}\left(\tilde{\mathcal{A}}_j\right) \tag{12}$$

where $\phi_{\mathbf{V}}\left(r_{i_o}, r_{\tilde{i_o,z_o}}\right) = F\left(r_{i_o}, r_{\tilde{i_o,z_o}}\right)$ and $\phi_{\mathbf{F}}\left(\tilde{\mathcal{A}}_j\right) = \mathbf{D}\left(\tilde{\mathcal{A}}_j\right) = \frac{\sigma_{\max}\left(\tilde{\mathcal{A}}_j\right)}{\sigma_{\min}\left(\tilde{\mathcal{A}}_j\right)}$ are the vertex and face factor functions respectively. We use the sum-product algorithm to obtain the most likely configuration estimate of Eq. 12.

3.4. Flow estimation

Having estimated the most likely match configuration under the piecewise affine geometric transformation model, we compose the dense flow field relating I_1 and I_2 in two different ways which are detailed next.

Linear Warp. The bounding boxes of the matched object parts $\{r_{i_o}\}$, and $\{r_{\tilde{i_o,z_o}}\}$ could be used to define piecewise linear warps in order to compose $\overline{r_1}$. As the object part regions can have overlaps, it is likely that a point in the domain of $\overline{r_1}$ is covered by many different object part regions. This necessitates a linear warp transformation assignment for each point. We propose to assign each point, the linear warp transformation induced by the region matches having the highest matching score. The flow vectors for the points that are not composed and that are outside $\overline{r_1}$ are obtained by propagation as done in [30]. We use a bilateral filter as a final step to get rid of any possible boundary artifacts.

Thin Plate Spline Warp. The thin plate spline interpolant (TPS) could also be used to obtain a dense flow field. We use the sets of matched center coordinate pairs $\{\text{center}\left(r_{i_o}\right)\}$, and $\{\text{center}\left(r_{\tilde{i_o,z_o}}\right)\}$ to obtain the TPS warp.

4. Experiments

We use several semantic correspondence benchmark datasets to evaluate our proposed method quantitatively and qualitatively. In all the following experiments, we fix the parameters of our algorithm to the following values: K, which controls the number of coarse region match candidates, is set to 20, C, the distortion threshold, is set to 2, the beam size b of the Beam Search algorithm is set to 4, and Z, which denotes the number of fine region match candidates, is set to 10.

The JR dataset. The JR dataset [26] consists of three sets: FG3DCar, JODS, and PASCAL totalling 400 image pairs. The dataset contains dense flow field ground-truth of foreground objects. The matching accuracy is evaluated on a normalized scale of foreground object bounding box whose larger dimension is 100 pixels. The Euclidean distance between the ground-truth and the estimated flow vector is measured and matches with errors of under 5 pixels are deemed correct, following the work of [26].

Table 1 shows the comparison of the matching accuracies of several state-of-the-art algorithms. Among the others, our method achieves the best performance on the PASCAL subset. The performances obtained for the FG3DCar and the JODS subsets are also promising. Fig. 4 shows some qualitative dense correspondence results. All the competing state-of-the-art algorithms are challenged when the scales of the foreground objects are highly different. Thanks to the co-localization step, our method is able to correctly align the foreground objects.

Methods	supervision	FG3DCar	JODS	PASCAL	Avg.
SIFT Flow [19]	u	0.63	0.51	0.36	0.50
DSP [11]	u	0.49	0.47	0.38	0.45
PF [6]	u	0.79	0.65	0.53	0.66
Yang *et al.* [30]	u	**0.87**	**0.71**	0.73	**0.77**
Zhou *et al.* [33]	s	0.72	0.51	0.44	0.56
FCSS [12]	s	0.83	0.66	0.49	0.66
Ours (Linear Warp)	u	0.77	0.67	**0.80**	0.75
Ours (TPS Warp)	u	0.78	0.68	**0.80**	0.75

Table 1: Flow accuracy rate with an error threshold of 5 pixels on the JR dataset. (u:unsupervised, s:supervised)

The PF-WILLOW dataset. The PF-WILLOW dataset [6] contains, in total, 900 pairs of images from 5 object classes, which are face, car, motorbike, duck and bottle. The dataset contains ground-truth keypoint annotations. The matching accuracy is evaluated using the PCK metric [32] between the warped and the ground-truth keypoints. If the ground-truth keypoint lies within $\alpha \max\left(h, w\right)$ pixels of the warped keypoint, where h and w denote the height and the width of the foreground object bounding box, then the estimation is deemed correct. The standard benchmark on this dataset uses $\alpha = 0.1$, which is also used in this paper.

Table 2 shows the comparison of the matching accuracies of several state-of-the-art algorithms. On average, our proposed method achieves comparable performance and in several object categories such as car, motorbike and bottle obtains higher performances. Fig. 5 shows some qualitative dense correspondence results. Our method obtains high quality correspondences and keeps the integrity of objects under the warp transformation. The competing state-of-the-art algorithms are again challenged when there is significant amount of background clutter and/or when the viewpoint changes.

The PF-PASCAL dataset. The PF-PASCAL dataset [6] contains 1351 pairs of images of 20 PASCAL object categories. The matching accuracy is again evaluated using the PCK metric.

Table 3 shows the comparison of the matching accuracies of several state-of-the-art algorithms. Our method obtains very good performance on some of the object categories and outperforms many other competing algorithms. Fig. 6 shows some qualitative dense correspondence results.

The Caltech-101 dataset. The Caltech-101 dataset [15] contains images of 101 object categories. The benchmark protocol for semantic correspondence uses 15 random image pairs from each category making 1515 distinct image pairs in total. We use the same set of images used in [6]. Three different quantitative evaluation metrics are used to assess the quality of a dense correspondence field. These

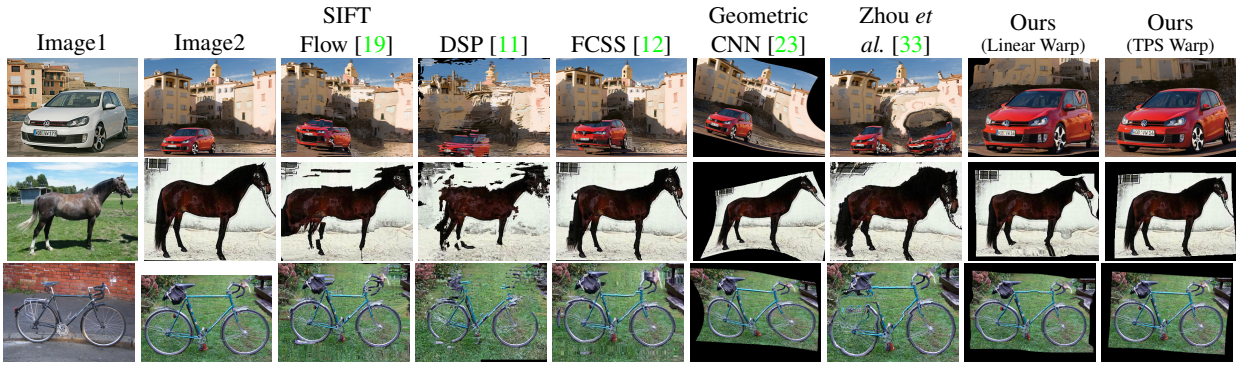

Figure 4: **Qualitative dense correspondence results on the JR dataset.** The first two columns show a pair of images and columns three to nine show warpings of the second image to the first image using dense pixel-wise correspondences. The last column shows the ground-truth warp. Our method achieves high quality warps. Note how competing algorithms are challenged when the foreground objects are viewed under different scales as demonstrated in the first row. Zoom in for better visibility.

Methods	supervision	car(S)	car(G)	car(M)	duc(S)	mot(S)	mot(G)	mot(M)	win(w/o C)	win(w/ C)	win(M)	Avg.
SIFT Flow [19]	u	0.54	0.37	0.36	0.32	0.41	0.20	0.23	0.83	0.16	0.33	0.38
DSP [11]	u	0.46	0.30	0.32	0.25	0.31	0.15	0.14	0.85	0.25	0.64	0.37
PF [6]	u	**0.86**	0.60	**0.53**	**0.64**	0.49	0.25	**0.29**	**0.91**	0.37	**0.65**	0.56
Zhou *et al.* [33]	s	0.77	0.34	0.52	0.42	0.34	0.19	0.20	0.78	0.19	0.38	0.41
FCSS [12]	s	-	-	-	-	-	-	-	-	-	-	0.53
Geometric CNN [23]	s	-	-	-	-	-	-	-	-	-	-	**0.57**
Ours (Linear Warp)	u	0.83	**0.62**	**0.53**	0.44	**0.52**	**0.27**	0.28	0.87	**0.46**	0.60	0.54
Ours (TPS Warp)	u	0.82	0.60	0.51	0.42	0.51	**0.27**	**0.29**	0.86	0.42	0.56	0.53

Table 2: PCK metric ($\alpha = 0.1$) comparison of dense flow field on the PF-WILLOW dataset. (u:unsupervised, s:supervised)

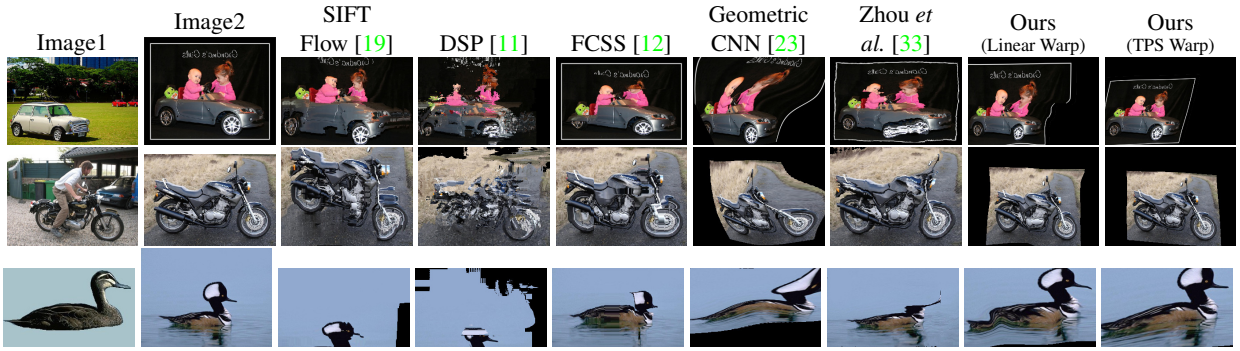

Figure 5: **Qualitative dense correspondence results on the PF-WILLOW dataset.** The first two columns show a pair of images and columns three to nine show warpings of the second image to the first image using dense pixel-wise correspondences. Our method achieves high quality warps. Large intra-class variations result in alignment failures for many of the state-of-the-art methods. Zoom in for better visibility.

are (1) the label transfer accuracy (LT-ACC), which transfers the annotated class labels of an exemplar image using the estimated dense correspondence field and counts the number of correctly labeled pixels in the test image, (2) the IoU metric, and (3) the localization error (LOC-ERR), which measures the localization error of pixels with respect to object bounding boxes.

Table 4 shows the comparison of the matching accuracies of several state-of-the-art algorithms. Our method, on this dataset, does not have any significant advantage over the other methods. The reason for this is that the dataset contains many different object categories for which the object detector we use is not trained, hindering the process of correctly co-localizing the foreground objects. Nevertheless, as demonstrated in Fig. 7, high quality dense correspondence results could sometimes be achieved.

Failure cases: Upon visual inspection of the failure cases, we have observed that many of them stem from an incorrect foreground object co-localization. An example is shown in Fig. 8.

5. Conclusion

We have proposed an unsupervised dense semantic correspondence algorithm based on a compositional process by first detecting and co-localizing the foreground objects. Owing to the bounded geometric distortion constraints on the piece-wise affine transformations adopted, high quality dense motion fields are obtained. The proposed method

Methods	supervision	aero	bike	bird	boat	bot	bus	car	cat	cha	cow	tab	dog	hor	mbik	pers	plnt	she	sofa	trai	tv	Avg.
SIFT Flow [19]	u	0.61	0.56	0.20	0.34	0.32	0.54	0.56	0.26	0.29	0.21	**0.33**	0.17	0.23	0.43	0.18	0.17	0.17	0.31	0.41	0.34	0.33
DSP [11]	u	0.64	0.56	0.17	0.27	0.38	0.51	0.55	0.20	0.23	0.24	0.19	0.15	0.23	0.41	0.15	0.11	0.18	0.27	0.35	0.28	0.30
PF [6]	u	**0.75**	**0.76**	**0.34**	**0.41**	**0.55**	0.71	0.73	0.32	**0.41**	0.41	0.21	**0.27**	0.38	0.57	**0.29**	0.17	0.33	**0.34**	**0.54**	**0.46**	0.45
Zhou *et al.* [33]	s	0.58	0.35	0.15	0.27	0.36	0.40	0.42	0.23	0.26	0.29	0.22	0.20	0.13	0.33	0.16	0.18	**0.48**	0.27	0.34	0.28	0.30
FCSS [12]	s	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	**0.46**
Ours (Linear Warp)	u	0.71	0.73	0.27	0.22	0.48	**0.72**	**0.76**	**0.38**	0.27	**0.50**	0.22	0.25	**0.39**	**0.58**	**0.29**	**0.30**	0.27	0.28	0.40	0.35	0.42
Ours (TPS Warp)	u	0.71	0.72	0.25	0.20	0.44	0.68	0.74	**0.38**	0.27	0.47	0.17	0.24	0.36	0.57	0.28	**0.30**	0.27	0.27	0.35	0.27	0.40

Table 3: PCK metric ($\alpha = 0.1$) comparison of dense flow field on the PF-PASCAL dataset. (u:unsupervised, s:supervised)

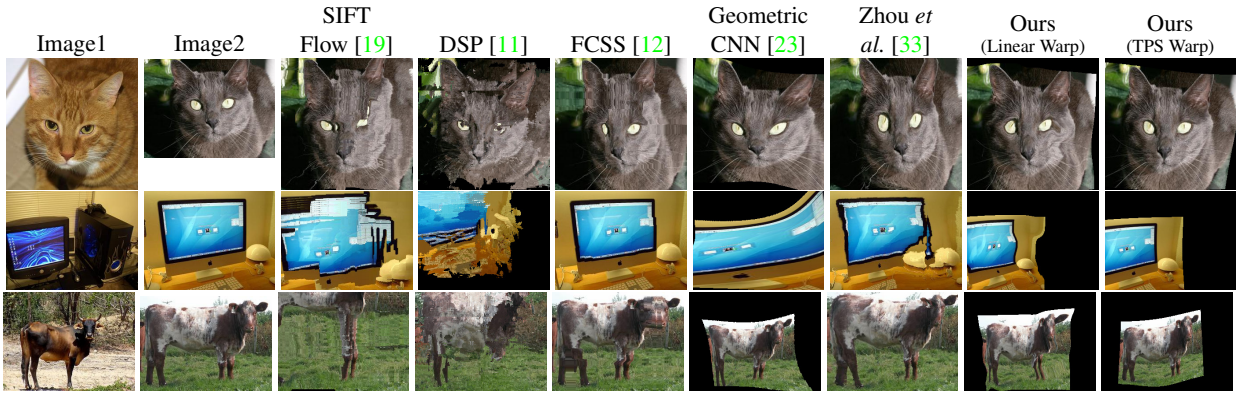

Figure 6: **Qualitative dense correspondence results on the PF-PASCAL dataset.** The first two columns show a pair of images and columns three to nine show warpings of the second image to the first image using dense pixel-wise correspondences. Our method achieves high quality warps. Zoom in for better visibility.

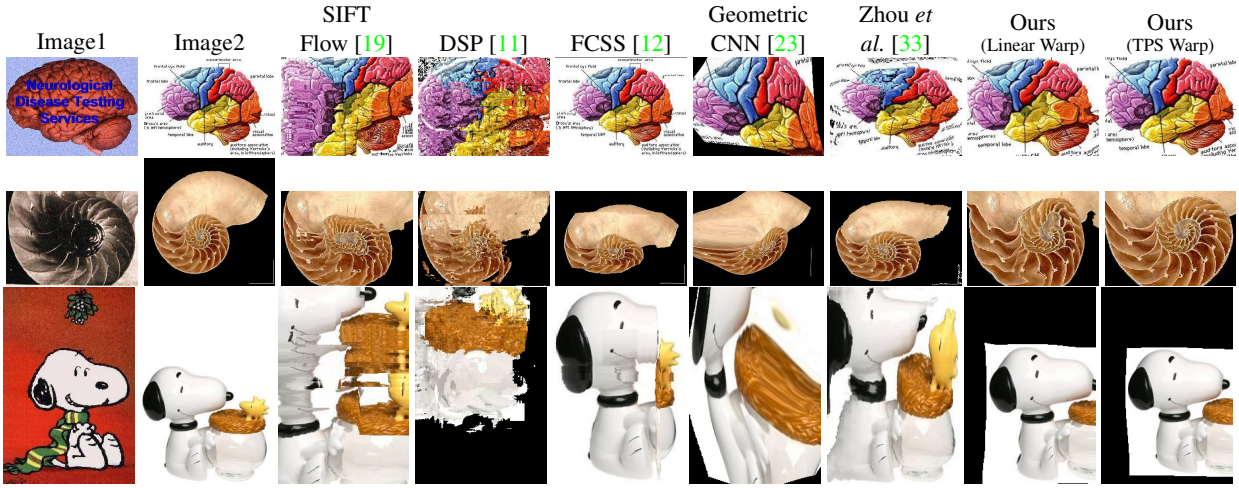

Figure 7: **Qualitative dense correspondence results on the Caltech-101 dataset.** The first two columns show a pair of images and columns three to nine show warpings of the second image to the first image using dense pixel-wise correspondences. Our method achieves high quality warps. Note how the corresponding spirals of a pair of nautili in the second row are correctly aligned. Zoom in for better visibility.

yields encouraging performances as demonstrated on several semantic dense correspondence benchmark datasets both qualitatively and quantitatively.

Figure 8: **A failure example from the Caltech-101 dataset.** Although it is able to establish geometrically coherent matches, our algorithm can fail if the foreground objects are incorrectly co-localized.

Acknowledgements: Part of this research was conducted using computational resources and services at the Center for Computation and Visualization (CCV), Brown University.

Methods	supervision	LT-ACC	IoU	LOC-ERR
SIFT Flow [19]	u	0.75	0.48	0.32
DSP [11]	u	0.77	0.47	0.35
PF [6]	u	0.78	0.50	0.25
Yang *et al.* [30]	u	0.81	0.55	**0.19**
FCSS [12]	s	0.80	0.50	0.21
Geometric CNN [23]	s	**0.82**	**0.56**	0.25
Ours (Linear Warp)	u	0.76	0.44	0.37
Ours (TPS Warp)	u	0.76	0.43	0.39

Table 4: Matching accuracy on the Caltech-101 dataset. (u:unsupervised, s:supervised)

References

[1] C. Barnes, E. Shechtman, D. B. Goldman, and A. Finkelstein. The generalized patchmatch correspondence algorithm. In *Computer Vision - ECCV 2010, 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part III*, pages 29–43, 2010. 2

[2] A. C. Berg and J. Malik. Geometric blur for template matching. In *2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2001), with CD-ROM, 8-14 December 2001, Kauai, HI, USA*, pages 607–614, 2001. 2

[3] M. Cho, S. Kwak, C. Schmid, and J. Ponce. Unsupervised object discovery and localization in the wild: Part-based matching with bottom-up region proposals. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 1201–1210, 2015. 3

[4] J. Dai, Y. Li, K. He, and J. Sun. R-FCN: object detection via region-based fully convolutional networks. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 379–387, 2016. 3

[5] Y. HaCohen, E. Shechtman, D. B. Goldman, and D. Lischinski. Non-rigid dense correspondence with applications for image enhancement. *ACM Trans. Graph.*, 30(4):70:1–70:10, 2011. 1

[6] B. Ham, M. Cho, C. Schmid, and J. Ponce. Proposal flow: Semantic correspondences from object proposals. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(7):1711–1725, 2018. 2, 6, 7, 8

[7] B. Hariharan, J. Malik, and D. Ramanan. Discriminative decorrelation for clustering and classification. In *Computer Vision - ECCV 2012 - 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part IV*, pages 459–472, 2012. 2, 3

[8] B. K. P. Horn and B. G. Schunck. Determining optical flow. *Artif. Intell.*, 17(1-3):185–203, 1981. 1

[9] J. Hur, H. Lim, C. Park, and S. C. Ahn. Generalized deformable spatial pyramid: Geometry-preserving dense correspondence estimation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 1392–1400, 2015. 2

[10] K. Karsch, C. Liu, and S. B. Kang. Depth extraction from video using non-parametric sampling. In *Computer Vision - ECCV 2012 - 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part V*, pages 775–788, 2012. 1

[11] J. Kim, C. Liu, F. Sha, and K. Grauman. Deformable spatial pyramid matching for fast dense correspondences. In *2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, June 23-28, 2013*, pages 2307–2314, 2013. 2, 6, 7, 8

[12] S. Kim, D. Min, B. Ham, S. Jeon, S. Lin, and K. Sohn. FCSS: fully convolutional self-similarity for dense semantic correspondence. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 616–625, 2017. 2, 6, 7, 8

[13] S. Kim, D. Min, S. Lin, and K. Sohn. DCTM: discrete-continuous transformation matching for semantic flow. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 4539–4548, 2017. 2

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States.*, pages 1106–1114, 2012. 2

[15] F. Li, R. Fergus, and P. Perona. One-shot learning of object categories. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(4):594–611, 2006. 6

[16] G. Li and Y. Yu. Visual saliency based on multiscale deep features. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 5455–5463, 2015. 3

[17] Y. Lipman, S. Yagev, R. Poranne, D. W. Jacobs, and R. Basri. Feature matching with bounded distortion. *ACM Trans. Graph.*, 33(3):26:1–26:14, 2014. 4

[18] C. Liu, J. Yuen, and A. Torralba. Nonparametric scene parsing via label transfer. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(12):2368–2382, 2011. 1

[19] C. Liu, J. Yuen, and A. Torralba. SIFT flow: Dense correspondence across scenes and its applications. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(5):978–994, 2011. 1, 6, 7, 8

[20] J. Long, N. Zhang, and T. Darrell. Do convnets learn correspondence? In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 1601–1609, 2014. 3

[21] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004. 1

[22] W. Qiu, X. Wang, X. Bai, A. L. Yuille, and Z. Tu. Scale-space SIFT flow. In *IEEE Winter Conference on Applications of Computer Vision, Steamboat Springs, CO, USA, March 24-26, 2014*, pages 1112–1119, 2014. 2

[23] I. Rocco, R. Arandjelovic, and J. Sivic. Convolutional neural network architecture for geometric matching. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 39–48, 2017. 2, 7, 8

[24] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision*, 47(1-3):7–42, 2002. 1

[25] B. Sevilmis and B. B. Kimia. Subpixel semantic flow. In *Proceedings of the British Machine Vision Conference 2017, BMVC 2017, London, UK, September 4-7, 2017*, 2017. 2

[26] T. Taniai, S. N. Sinha, and Y. Sato. Joint recovery of dense correspondence and cosegmentation in two images. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 4246–4255, 2016. 2, 6

[27] E. S. Tellez, G. Ruiz, E. Chávez, and M. Graff. Local search methods for fast near neighbor search. *CoRR*, abs/1705.10351, 2017. 5

[28] N. Ufer and B. Ommer. Deep semantic feature matching. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 5929–5938, 2017. 2

[29] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders. Selective search for object recognition. *International Journal of Computer Vision*, 104(2):154–171, 2013. 3

[30] F. Yang, X. Li, H. Cheng, J. Li, and L. Chen. Object-aware dense semantic correspondence. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 4151–4159, 2017. 2, 6, 8

[31] H. Yang, W. Lin, and J. Lu. DAISY filter flow: A generalized discrete approach to dense correspondences. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, pages 3406–3413, 2014. 2

[32] Y. Yang and D. Ramanan. Articulated human detection with flexible mixtures of parts. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(12):2878–2890, 2013. 6

[33] T. Zhou, P. Krähenbühl, M. Aubry, Q. Huang, and A. A. Efros. Learning dense correspondence via 3d-guided cycle consistency. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 117–126, 2016. 2, 6, 7, 8