Shape-based Image Correspondence

Berk Sevilmis
berk_sevilmis@brown.edu
Benjamin B. Kimia
benjamin_kimia@brown.edu

LEMS
Brown University
Providence, RI 02912 USA

Abstract

Current state-of-the-art dense correspondence algorithms establish correspondences between pair of images by searching for a flow field that minimizes the distance between local signatures (*e.g.,* color histogram, SIFT descriptor) of aligned pixels while preserving smoothness. Agnostic to the global signatures (*e.g.,* object membership, category of object), these local signatures face difficulties in resolving alignment ambiguities when scene content undergoes type and configuration variation. In this paper, we investigate the effect of adding shape correspondence constraints either in the form of pair of corresponding contour fragments or pair of closed curves. We find the shape does not play a significant role in optical flow and stereo correspondence but it does play a significant role when scene content changes are large. We also explore using object proposals as a way of providing shape constraints with encouraging results.

1 Introduction

Many of the computer vision tasks such as stereo correspondence, optical flow, biometric user verification, and object recognition require the establishment of dense pixel correspondences between pair of images which can differ in image acquisition setting, *i.e.,* scene content and scene configuration. On the one end of the spectrum is the narrow-baseline stereo correspondence, where these variations are at a minimum since the same 3D scene is captured from slightly different viewpoints. On the other extreme is the *semantic image alignment*, where photometric and geometric variations are unbounded but it is still relevant to establish correspondences between image pairs. This typically involves images captured from different 3D scenes sharing similar characteristics such as containing same but different instances of objects. In this setting, dense correspondences are useful in semantic image segmentation [16], video depth estimation [12], image enhancement [9] *etc.*

Recent state-of-the-art approaches [3, 10, 13, 16, 18, 20, 22] attempt to compute correspondences between pair of images by matching image signatures, *e.g.,* color histograms, SIFT descriptor [19], CNN features [14], extracted locally from pixels and enforce smoothness on the correspondence field by enforcing spatial regularity. This type of an approach works well when the underlying scene that gives rise to the two images are the same but viewed under slightly different conditions, *i.e.,* optical flow with small displacement, and narrow-baseline stereo, or even wide-baseline multiview imagery. They perform reasonably well when the object and/or the scene change instance to a very similar type. However, given that the signature variation measure does not capture any semantic aspect of the scene beyond a local histogram over a neighborhood, it is challenged by semantically related images featuring large visual variations.

Pages 66.1-66.12
DOI: https://dx.doi.org/10.5244/C.30.66

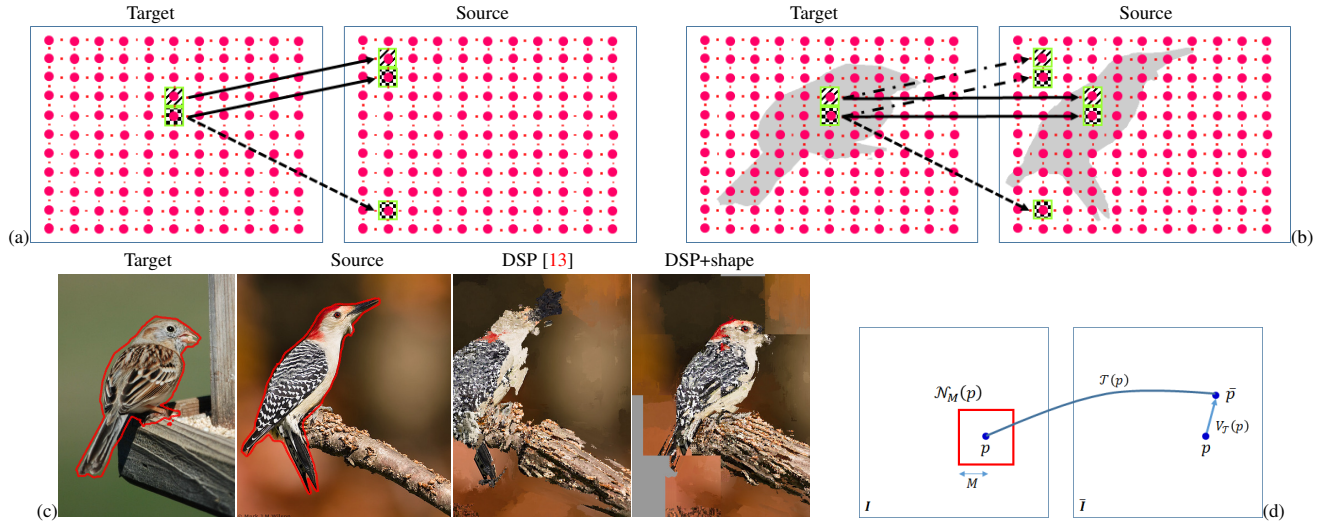

Figure 1: **Shape aligned dense correspondence.** (a) Spatial regularity in current state-of-the-art methods only disambiguate matches which are not locally consistent, *i.e.,* preferring the solid line correspondence to dashed one. (b) Shape alignment can reduce the ambiguity further by ruling out correspondences which violate inside-outside consistency. (c) A visual result. The warped source using shape alignment constraint is clearly superior. (d) A correspondence is a transformation $\mathcal{T}$ mapping a point p in I to a point $\bar{p}$ in $\bar{I}$. A local neighborhood $\mathcal{N}_M(p)$ restricts pixels which contribute to the descriptor at p.

Our approach is to introduce certain *semantic concepts* into the correspondence process. Specifically, in this paper, we explore the effect of shape as an additional guideline to the variational correspondence process. We ask whether specifying a pair of corresponding shapes can influence the correspondence process significantly and under what scenarios. We also ask whether shape should be specified in the form of a contour fragment or in the form of a closed curve bounding a region. Finally, when such corresponding shape constraints are not available, we ask whether object proposals can serve this purpose and under what conditions.

The rest of the paper is organized as follows. In Section 2, we formulate four related state-of-the-art approaches under an umbrella. In Section 3 we discuss how shape can be introduced as a constraint in the context of this formulation and how object proposals may be used. Section 4 describes experiments which explore the effectiveness of the role of shape in producing better correspondences. We will conclude that shape plays little role in optical flow, improves on correspondences for wide-baseline imagery, and is essential towards providing meaningful correspondences under semantic alignment.

2 A Unifying View of Current Approaches

The problem of computing dense correspondences between two images, I and $\bar{I}$, has generally relied on optimizing image similarity and smoothness of the correspondence. Formally, let $\mathcal{N}_M(p)$ denote a neighborhood of size M of point p, Fig. 1d. A dense correspondence is a transformation $\mathcal{T}$ which takes a point p in image I to a point $\bar{p}$ in image $\bar{I}$, $\bar{p} = \mathcal{T}(p)$. In analogy with optical flow, a vector field (flow field) $\mathrm{V}_{\mathcal{T}}(p)$ is defined as $\mathrm{V}_{\mathcal{T}}(p) = \bar{p} - p = \mathcal{T}(p) - p$, over which the regularity of $\mathcal{T}$ is represented. In some approaches the transformation $\mathcal{T}$ not only defines the dense correspondence, but also the scale of image at which correspondence makes most sense.

Previous approaches have followed a variational approach to find the best correspondence by defining an energy functional consisting of four terms,

$$E(\mathcal{T}) = E_{\text{data}}(\mathcal{T}) + E_{\text{limit flow}}(\mathcal{T}) + E_{\text{smooth}}(\mathcal{T}) + E_{\text{scale}}(\mathcal{T})$$
$$= \sum_p \left[f_{\text{data}}(\mathcal{N}_M(p), \mathcal{N}_{\bar{M}}(\mathcal{T}(p))) + f_{\text{limit flow}}(\mathcal{T}(p) - p) \right] + \sum_{(p,\hat{p}) \in \mathrm{E}} \left[f_{\text{smooth}}(\mathcal{T}(p), \mathcal{T}(\hat{p})) + f_{\text{scale}}(M(p), \hat{M}(\hat{p})) \right] \quad (1)$$

which in turn *(i)* optimize the similarity of the corresponding points p and $\bar{p}$, *(ii)* penalize or limit the extent of $\bar{p}$ deviating from p, *(iii)* enforce spatial smoothness on the correspondence, and *(iv)* enforce scale consistency. We now describe four of the state-of-the-art techniques in this framework.

First, the Patch Match approach [3] defines its data term as the sum of squared differences

$$f_{\text{data}}(\mathcal{N}_M(p),\mathcal{N}_{\bar{M}}(\mathcal{T}(p))) = \sum_{\hat{p}\in\mathcal{N}_M(p)} \|I(\hat{p}) - \bar{I}(\mathcal{T}(\hat{p}))\|^2, \tag{2}$$

or the difference in the SIFT descriptor, SD, over the neighborhood $\mathcal{N}_M$,

$$f_{\text{data}}(\mathcal{N}_M(p),\mathcal{N}_{\bar{M}}(\mathcal{T}(p))) = \|SD(\mathcal{N}_M(p)) - SD(\mathcal{N}_M(\mathcal{T}(p)))\|^2. \tag{3}$$

Patch Match [3] does not have an explicit $E_{\text{limit flow}}$ nor E_{smooth}. The energy is optimized by an iterative approach where an initial correspondence $\mathcal{T}^0(p)$ is defined by a random assignment. The transform is then iteratively updated to the transform of that neighbor that minimizes the data term, *i.e.,*

$$\mathcal{T}^{i+1}(p) = \mathcal{T}^i(\ \underset{\hat{p}|(p,\hat{p})\in \mathrm{E}}{\operatorname{argmin}}\ f_{\text{data}}(\mathcal{N}_M(p),\mathcal{N}_M(\mathcal{T}^i(\hat{p}) + (p-\hat{p})))\ ) + (p-\hat{p}). \tag{4}$$

In addition to the immediate neighbors, some random points in a larger neighborhood are also considered to help the process converge globally. The optimization stops after a fixed number of iterations. The Patch Match [3] algorithm approach finds an alignment of similar patches, but these patches may not enjoy spatial coherence, mainly because there is no smoothness term in the correspondence energy.

Second, the SIFT Flow approach [16] uses the L^1 norm variant of the SIFT descriptor difference of Eq. 3 as its data term. The limit flow energy term is defined as $\|\mathrm{V}_{\mathcal{T}}(p)\|_1$. The smoothness energy is a truncated L^1 norm of the flow field variation in $\mathcal{N}_M$,

$$f_{\text{smooth}}(\mathcal{T}(p),\mathcal{T}(\hat{p}))\big|_{(p,\hat{p})\in \mathrm{E}} = \|\mathrm{V}_{\mathcal{T}}(p) - \mathrm{V}_{\mathcal{T}}(\hat{p})\|_1, \tag{5}$$

with the modification that the variation in horizontal and vertical components are upper bounded by a predetermined threshold. The optimization is a variant of loopy belief propagation. The results are excellent when the scale and orientation of objects are similar and when the global arrangement is roughly similar. SIFT Flow [16] depicts better spatial coherence as compared to Patch Match [3], Tables 1 and 2.

Third, the Deformable Spatial Pyramid (DSP) approach [13] is a two stage process. In the first stage, the image is divided into blocks and a displacement is found for each block. In the second stage, the flow field for each pixel in each block refines the block displacement by optimizing the L^1 norm variant of the data term in Eq. 3 in a small window. The displacement in the first stage is found by representing the image by a three level pyramid consisting of the image divided into 1×1, 2×2 and 4×4 blocks,respectively. The pyramid is represented by an interconnected graph of 21 nodes with connections to neighbors within and across layers. The displacement at each node is constrained by parents and by neighbors and is found by a SIFT Flow [16] like computation. The pyramid structure affords a global perspective and improves correspondences as a result. When this global view fails, however, the failed effect is propagated to finer scales in an unrecoverable fashion. Also, the second stage is not always able to sufficiently fine tune neighbors across two adjacent blocks, leading to a blocky appearance.

Fourth, the Scale-Space SIFT Flow (SSF) approach [20] addresses the scale issue of SIFT Flow [16] by introducing the concept of an optimal scale at each point. Initially, SIFT Flow [16] correspondences for SIFT descriptors of varying scales, *i.e.,* 12×12, 24×24, 48×48, 72×72, 96×96 are found, to obtain a scale-dependent data term for each point. The initial scales of descriptors at each point are then computed by optimizing the scale-dependent data term as well as a novel term involving scale smoothness, *i.e.,*

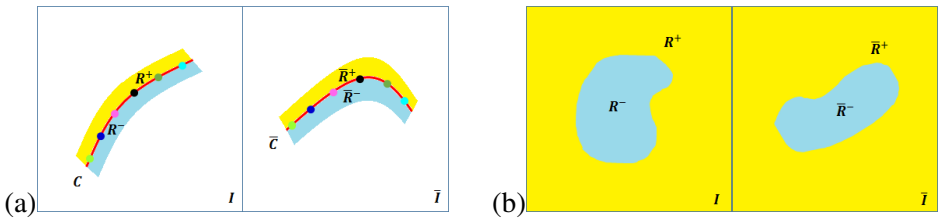

Figure 2: **Alignment constraints.** (a) Contour fragment alignment prevents cross talk between the right (R^-, in blue) and left (R^+, in yellow) regions. (b) Closed curve alignment prevents cross talk between inside (R^-, in blue), and outside (R^+, in yellow).

$$f_{\text{data}}(\mathcal{N}_M(p),\mathcal{N}_{\bar{M}}(\mathcal{T}(p))) = \|SD(\mathcal{N}_M(p)) - SD(\mathcal{N}_{\bar{M}}(\mathcal{T}(p)))\|_1,\ f_{\text{scale}}(M(p),\hat{M}(\hat{p}))\big|_{(p,\hat{p})\in \mathrm{E}} = \|M(p)-\hat{M}(\hat{p})\|_1 \tag{6}$$

where $\bar{M}$ is fixed at 12 while M varies as specified above. The algorithm alternates in optimizing $f_{\text{data}} + f_{\text{smooth}}$ and $f_{\text{data}} + f_{\text{scale}}$ until convergence. This algorithm improves SIFT Flow [16] results when there are significant scale changes, but unfortunately it inherits the remaining issues surrounding SIFT Flow [16].

3 The Role of Shape in Improving Image Correspondences

The role of shape as a constraint to improve correspondences can be explored by specifying either a contour fragment, representing a portion of the shape silhouette, or a closed curve bounding a full object or object part. First, assume that the contour fragment $C(s)$ in image I, where s is an arbitrary parameter, matches the contour $\bar{C}(\bar{s})$ in image $\bar{I}$ as shown in Fig. 2a. Each curve separates two regions, (R^-,R^+) for $C(s)$ and $(\bar{R}^-,\bar{R}^+)$ for $\bar{C}(\bar{s})$, each limited in spatial extent. The shape alignment requires that the correspondence respects such separation, and avoid mapping R^- to $\bar{R}^+$ and R^+ to $\bar{R}^-$. We extend the variational formulation in Eq. 1 by introducing an energy term that prevents "cross-talk", *i.e.,* $E(\mathcal{T}) = E_{\text{previous}}(\mathcal{T}) + E_{\text{crosstalk}}(\mathcal{T})$,

$$E_{\text{crosstalk}}(\mathcal{T}) = \begin{cases} 0 & L(p) = \bar{L}(\mathcal{T}(p)) \\ \infty & L(p) \neq \bar{L}(\mathcal{T}(p)), \end{cases} \tag{7}$$

where $E_{\text{previous}}(\mathcal{T})$ is the energy term corresponding to any of the four previous techniques, Eq. 1, and L and $\bar{L}$ are label images indicating contour or region membership.

Second, the shape constraint can take the form of correspondence between two closed curves, *i.e.,* closed curve $C(s)$ in image I mapping to closed curve $\bar{C}(\bar{s})$ in image $\bar{I}$. Here the entire region inside/outside of one curve maps to the entire region inside/outside of the corresponding curve respectively, as shown in Fig. 2b. Analogous to the case of contour fragment correspondence, we prevent inside-outside crosstalk by using the same energy functional as Eq. 7. The details of optimizing the energy functional with the shape-based energy term are given in the supplementary material.

Shape alignment, whether in the form of a contour fragment or closed contour, provides additional information which constrains the correspondence problem. Thus, it is naturally expected that the performance would be improved. However, it is not a priori clear whether it is the use of shape as a separator of regions in the image that is at work or whether just adding additional correspondences is improving the results. Thus, we compare two scenarios, one in which a set of N random point correspondences constrain the correspondence, and one in which these N point correspondences arrange to define a shape. Our experiments in Section 4 indicate that the addition of random correspondences has a marginal effect while for points arranged as a shape, the effect can be significant and meaningful. Section 4 also

shows that specifying corresponding shapes can significantly improve the correspondence with large visual variation in pose or in content.

While these corresponding shapes can be provided by user annotation, we pose the question of whether object proposals can provide such constraints automatically. For automatically proposing shape constraints in the form of closed curves bounding regions, our approach makes the assumption that the category of one or more objects the two images have in common is known, *e.g.,* bird or horse. An object detector such as R-CNN [8] is used to detect objects in images. Given a limited region of interest, category independent object proposal methods, *e.g.,* MCG [2] or CPMC [5] generate candidates for the shape of the detected object. Then, a structured prediction framework is used to train a ranker to sort the proposals generated within the region detected by R-CNN [8].

Specifically, let $\mathbf{x} = \{p_1, p_2, ..., p_n\}$ denote the list of n proposals generated and $\mathbf{y}$ be the ranking output where $y_{ij} = +1$ whenever the proposal p_i is ranked higher than p_j and $y_{ij} = -1$ otherwise. We use the Jaccard index metric with respect to ground truth segmentations to rank the proposals. The joint input/output feature map, $\psi(\boldsymbol{x},\boldsymbol{y})$, is given as follows:

$$\psi(\boldsymbol{x},\boldsymbol{y}) = \sum_i \sum_j y_{ij} \cdot (\phi(p_i) - \phi(p_j)) \quad i,j = 1...n \tag{8}$$

where $\phi(p_i)$ denotes the feature space representation of proposal p_i. Following [5] for extracting regional, shape and texture features from the proposals, the margin rescaled learning algorithm [11] is used to maximize the joint score $\hat{\boldsymbol{y}}(\boldsymbol{x};\boldsymbol{w})$:

$$\hat{\boldsymbol{y}}(\boldsymbol{x};\boldsymbol{w}) = \underset{\boldsymbol{y}}{\operatorname{argmax}} \langle \boldsymbol{w}, \psi(\boldsymbol{x},\boldsymbol{y}) \rangle \tag{9}$$

where $\mathbf{w}$ represents the learned weights. Further details on training the object proposal ranker are provided in the supplementary material. The top ranked object proposal from each image can then be used as a shape constraint for aligning two images. The experiments in Section 4 show a significant improvement in correspondence.

4 Experiments

In this section we investigate the effect of shape on establishing dense correspondences between pair of images. We consider three types of computer vision tasks: *(i)* same scene being imaged under slightly different image acquisition parameters, *e.g.,* optical flow; *(ii)* same scene being imaged under large changes in viewing pose, *e.g.,* wide-baseline stereo correspondence; *(iii)* scene content being categorically the same but instances of objects and object configurations not being identical. We refer to this case as semantic image alignment, which is clearly the most challenging of the three types.

Specifically, we use the MPI Sintel Flow dataset [4] for the task of optical flow, the DTU Robot Image datasets [1] for wide-baseline stereo correspondence, and finally the PASCAL-Part [6] and the CUB-200-2011 [21] datasets for semantic image alignment. In the task of semantic image alignment, a subset of most similar image pairs are selected using the protocol of [17] by maximizing pyramidic histogram intersection of HOG visual words [15]. In total, 196 image pairs from the MPI Sintel Flow dataset, 60 image pairs from the DTU Robot Image datasets, 10,054 image pairs from the PASCAL-Part dataset [6], and 5,794 image pairs from the CUB-200-2011 dataset [21] are used in our experiments.

Optical flow and multiview stereo datasets provide dense per pixel ground truth correspondences. The PASCAL-Part dataset [6] has fine object part annotations whereas the CUB-200-2011 dataset [21] provides 15 point-wise part locations. To evaluate quantitative performance, we measure the flow error magnitude $\|\mathrm{V}_{\mathcal{T}} - \mathrm{V}_{\mathcal{T}}^{GT}\|$ as well as the flow angular error $\angle(\mathrm{V}_{\mathcal{T}}, \mathrm{V}_{\mathcal{T}}^{GT})$, on the optical flow and multiview stereo datasets, where $\mathrm{V}_{\mathcal{T}}^{GT}$ is the

ground truth flow vector. The CUB-200-2011 dataset [21] only provides point-wise part locations. Intersection over union $IoU(i,j) = \frac{|A_i \cap A_j|}{|A_i \cup A_j|}$, also known as Jaccard index, is used on object part intersections to measure the quantitative performance on the PASCAL-Part dataset [6], where A_i denotes the set of pixels belonging to region i. We use the same set of parameters provided by the implementations in [3, 13, 16, 20].

Exploring the role of shape in image correspondence: We begin with the question of whether providing a pair of corresponding shapes improves dense correspondences, qualitatively and quantitatively. We use the segmentations provided by the MPI Sintel Segmentation Training Data for the task of optical flow, manually annotated object masks for wide-baseline stereo correspondence and the segmentations of objects provided by the PASCAL-Part dataset [6] and the CUB-200-2011 dataset [21] for semantic image alignment.

	Flow Error Magnitude				Flow Angular Error			
data,algorithm	Patch Match [3]	SIFT Flow [16]	DSP [13]	SSF [20]	Patch Match [3]	SIFT Flow [16]	DSP [13]	SSF [20]
optical flow, traditional	**36.8 ± 133.0**	4.7 ± 7.3	**6.6 ± 9.2**	4.7 ± 7.4	**36.2 ± 26.9**	13.4 ± 18.5	**22.4 ± 30.6**	13.1 ± 18.2
optical flow, traditional+shape	37.6 ± 134.0	**4.7 ± 7.1**	8.2 ± 9.5	**4.7 ± 7.2**	36.5 ± 27.3	**13.1 ± 18.2**	25.7 ± 32.3	**12.8 ± 17.7**
wide-baseline stereo,traditional	220.6 ± 195.3	95.5 ± 62.9	103.2 ± 62.5	73.4 ± 57.0	55.9 ± 56.8	45.8 ± 42.2	68.3 ± 49.1	32.0 ± 39.3
wide-baseline stereo,traditional+shape	**213.0 ± 195.9**	**92.7 ± 62.6**	**84.7 ± 67.2**	**69.0 ± 57.6**	**53.7 ± 55.1**	**40.8 ± 41.1**	**54.6 ± 48.0**	**28.6 ± 40.5**

Table 1: **Flow error magnitude and flow angular errors in optical flow and multiview stereo datasets.**

Experimental results for the three types of data are revealing. First, for dataset depicting slight visual variations, traditional methods are effective and do not benefit from the introduction of a shape correspondence constraint as seen in Table 1 top two rows. Second, for datasets depicting large visual variation with the same scene context, the shape correspondence constraints improve the correspondence in the range of 7% as seen in Table 1 bottom two rows. In both these cases, matching local image signatures with a smoothness constraint already achieves good performances.

parts	Patch Match [3]	Patch Match+shape	SIFT Flow [16]	SIFT Flow+shape	DSP [13]	DSP+shape	SSF [20]	SSF+shape
back	0.20 ± 0.13	**0.12 ± 0.08**	0.09 ± 0.07	**0.06 ± 0.05**	0.08 ± 0.05	**0.05 ± 0.05**	0.14 ± 0.09	**0.07 ± 0.06**
beak	0.25 ± 0.16	**0.17 ± 0.14**	0.15 ± 0.11	**0.11 ± 0.12**	0.12 ± 0.09	**0.08 ± 0.11**	0.18 ± 0.12	**0.12 ± 0.12**
belly	0.21 ± 0.13	**0.11 ± 0.08**	0.09 ± 0.06	**0.06 ± 0.05**	0.08 ± 0.05	**0.06 ± 0.05**	0.12 ± 0.08	**0.06 ± 0.05**
breast	0.23 ± 0.16	**0.14 ± 0.12**	0.11 ± 0.08	**0.08 ± 0.07**	0.09 ± 0.06	**0.07 ± 0.07**	0.16 ± 0.10	**0.08 ± 0.09**
crown	0.23 ± 0.16	**0.14 ± 0.12**	0.12 ± 0.09	**0.08 ± 0.09**	0.10 ± 0.07	**0.05 ± 0.07**	0.16 ± 0.10	**0.08 ± 0.09**
forehead	0.23 ± 0.16	**0.14 ± 0.12**	0.13 ± 0.10	**0.09 ± 0.10**	0.11 ± 0.08	**0.06 ± 0.09**	0.17 ± 0.11	**0.10 ± 0.11**
lefteye	0.20 ± 0.17	**0.11 ± 0.12**	0.12 ± 0.09	**0.08 ± 0.09**	0.10 ± 0.08	**0.06 ± 0.08**	0.17 ± 0.11	**0.09 ± 0.10**
leftleg	0.23 ± 0.14	**0.15 ± 0.10**	0.10 ± 0.07	**0.07 ± 0.06**	0.09 ± 0.06	**0.07 ± 0.06**	0.12 ± 0.08	**0.07 ± 0.06**
leftwing	0.19 ± 0.13	**0.11 ± 0.08**	0.11 ± 0.08	**0.08 ± 0.06**	0.10 ± 0.06	**0.08 ± 0.06**	0.14 ± 0.09	**0.09 ± 0.06**
nape	0.22 ± 0.14	**0.12 ± 0.09**	0.10 ± 0.07	**0.07 ± 0.06**	0.09 ± 0.06	**0.06 ± 0.06**	0.15 ± 0.09	**0.07 ± 0.07**
righteye	0.20 ± 0.17	**0.11 ± 0.12**	0.12 ± 0.09	**0.08 ± 0.09**	0.09 ± 0.07	**0.05 ± 0.08**	0.15 ± 0.10	**0.08 ± 0.09**
rightleg	0.23 ± 0.14	**0.15 ± 0.10**	0.10 ± 0.07	**0.07 ± 0.06**	0.09 ± 0.06	**0.07 ± 0.06**	0.13 ± 0.08	**0.08 ± 0.07**
rightwing	0.20 ± 0.13	**0.11 ± 0.08**	0.11 ± 0.08	**0.09 ± 0.07**	0.10 ± 0.07	**0.08 ± 0.07**	0.15 ± 0.09	**0.10 ± 0.08**
tail	0.26 ± 0.16	**0.19 ± 0.12**	0.18 ± 0.15	**0.15 ± 0.15**	0.14 ± 0.12	**0.12 ± 0.13**	0.19 ± 0.14	**0.16 ± 0.16**
throat	0.23 ± 0.15	**0.13 ± 0.11**	0.12 ± 0.09	**0.08 ± 0.09**	0.10 ± 0.07	**0.07 ± 0.08**	0.16 ± 0.10	**0.09 ± 0.09**

Table 2: **Normalized part location error in the CUB-200-2011 dataset [21].** The aim here is not to declare a winner but to emphasize the significant improvements obtained, irrespective of the method used, when shape constraints are used.

The third type of dataset depicting instance and configuration variation tells a different story, however. In the CUB-200-2011 dataset [21], Table 2, and the PASCAL-Part dataset [6], Fig. 3, there are significant improvements, 35% and 170% respectively. Shape seems to help bring pixels into proper registration, as evidenced by the examples in Figs. 4 and 5.

Is it really shape of object that is helpful?: The question arises whether the demonstrated improvement arising from providing the shape correspondence constraint is simply due to

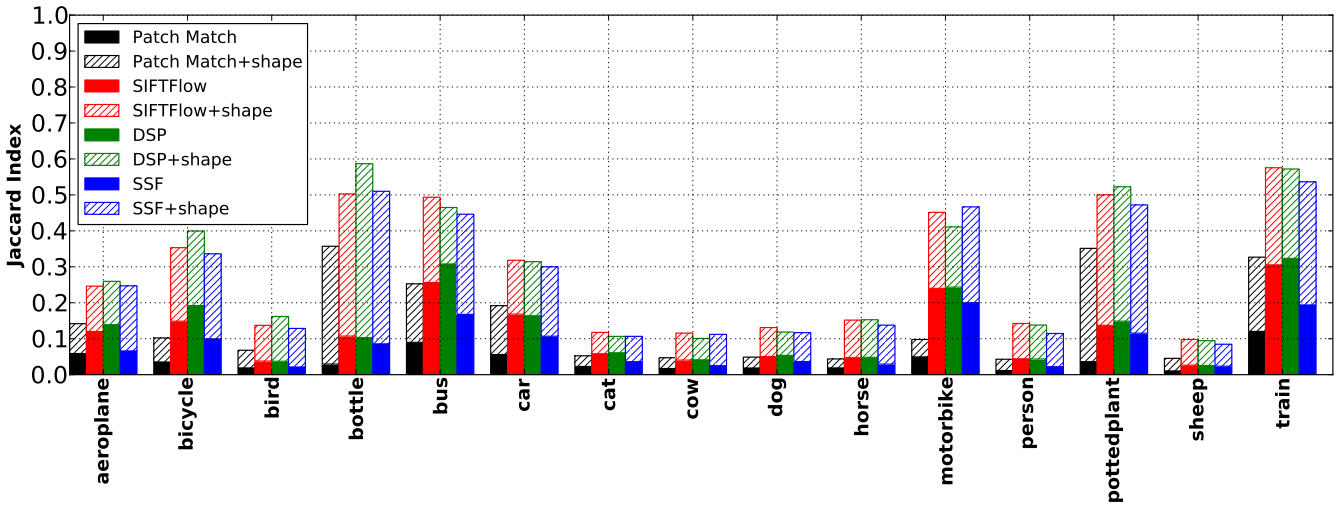

Figure 3: **Jaccard index performance of object classes in the PASCAL-Part dataset [6].** There is improvement in **every** case as indicated by the hashed portion of each bar. Per part performances of object classes are included in supplementary material.

	Flow Magnitude Error				Flow Angular Error			
Constraints	Patch Match [3]	SIFT Flow [16]	DSP [13]	SSF [20]	Patch Match [3]	SIFT Flow [16]	DSP [13]	SSF [20]
traditional	36.8 ± 133.1	4.7 ± 7.3	**6.6 ± 9.2**	4.7 ± 7.4	**36.2 ± 26.9**	13.4 ± 18.5	**22.4 ± 30.8**	13.1 ± 18.2
random points	**36.8 ± 133.0**	4.7 ± 7.2	6.6 ± 9.2	4.7 ± 7.3	**36.2 ± 26.9**	13.3 ± 18.5	**22.4 ± 30.8**	12.9 ± 17.9
contour fragment	37.8 ± 133.9	**4.6 ± 7.2**	8.5 ± 13.0	**4.5 ± 7.0**	36.4 ± 27.2	**13.0 ± 18.5**	23.5 ± 31.5	**12.6 ± 17.4**
closed curve	37.6 ± 134.0	4.7 ± 7.1	8.2 ± 9.5	4.7 ± 7.2	36.5 ± 27.3	13.1 ± 18.2	25.7 ± 32.3	12.8 ± 17.7

(a)

	Flow Magnitude Error				Flow Angular Error			
Constraints	Patch Match [3]	SIFT Flow [16]	DSP [13]	SSF [20]	Patch Match [3]	SIFT Flow [16]	DSP [13]	SSF [20]
traditional	220.6 ± 195.3	95.5 ± 62.9	103.2 ± 62.5	73.4 ± 57.0	55.9 ± 56.8	45.8 ± 42.2	68.3 ± 49.1	32.0 ± 39.3
random points	220.1 ± 195.1	94.6 ± 64.0	103.2 ± 62.4	68.5 ± 54.7	55.8 ± 56.7	45.5 ± 43.7	68.3 ± 49.1	**28.0 ± 39.1**
contour fragment	219.5 ± 195.0	93.6 ± 62.4	103.4 ± 64.7	**66.6 ± 53.8**	55.4 ± 56.5	43.5 ± 42.1	75.0 ± 53.0	28.3 ± 39.0
closed curve	**213.0 ± 195.9**	**92.7 ± 62.6**	**84.7 ± 67.2**	69.0 ± 57.6	**53.7 ± 55.1**	**40.8 ± 41.1**	**54.6 ± 48.0**	28.6 ± 40.5

(b)

Table 3: **Flow error magnitude and flow angular error using various alignment constraints:** in (a) optical flow and (b) multiview stereo datasets.

providing a set of additional point-pair correspondences or due to the fact that they provide geometrical constraints. In addition, we ask whether a contour correspondence constraint provides a significantly better constraint if the contour is closed. These questions can be answered by providing a set of N points which are in three configurations: *(i)* randomly placed points, *(ii)* points arranged as a contour, and *(iii)* points arranged as a closed contour.

Table 3 compares the results of adding each of these three constraints to the traditional methods experimented on optical flow and multiview stereo datasets. (We have not experimented with configurations *(i)*, and *(ii)* in the task of semantic image alignment, main reason being the unavailability of densely annotated ground truth flow fields. Also, the results shown in Table 4 and Fig. 5, though being valid for configuration *(iii)*, are not repeated.) In all the experiments here, N is empirically chosen to be 50. Table 3a, using optical flow dataset, shows no improvement based on shape as before, and Table 3b, using multiview stereo dataset, shows modest improvement due to shape. The experiment shows that closed curves generally perform better than contour fragments which in turn perform better than a random set of points as a constraint to finding image correspondence.

Object proposals and shape alignment: Since MCG [2] and CPMC [5] use the PASCAL dataset [7] in tuning hyperparameters of their object proposal pipelines, the CUB-200-2011 dataset [21] is used to perform an unbiased experiment. Correspondences are evaluated for the accuracy of part transfer compared to ground truth. Table 4 shows the quantitative performances achieved in object part transfer when object shapes are automatically proposed, whereas Fig. 6 demonstrates the qualitative results obtained for the same image pair shown

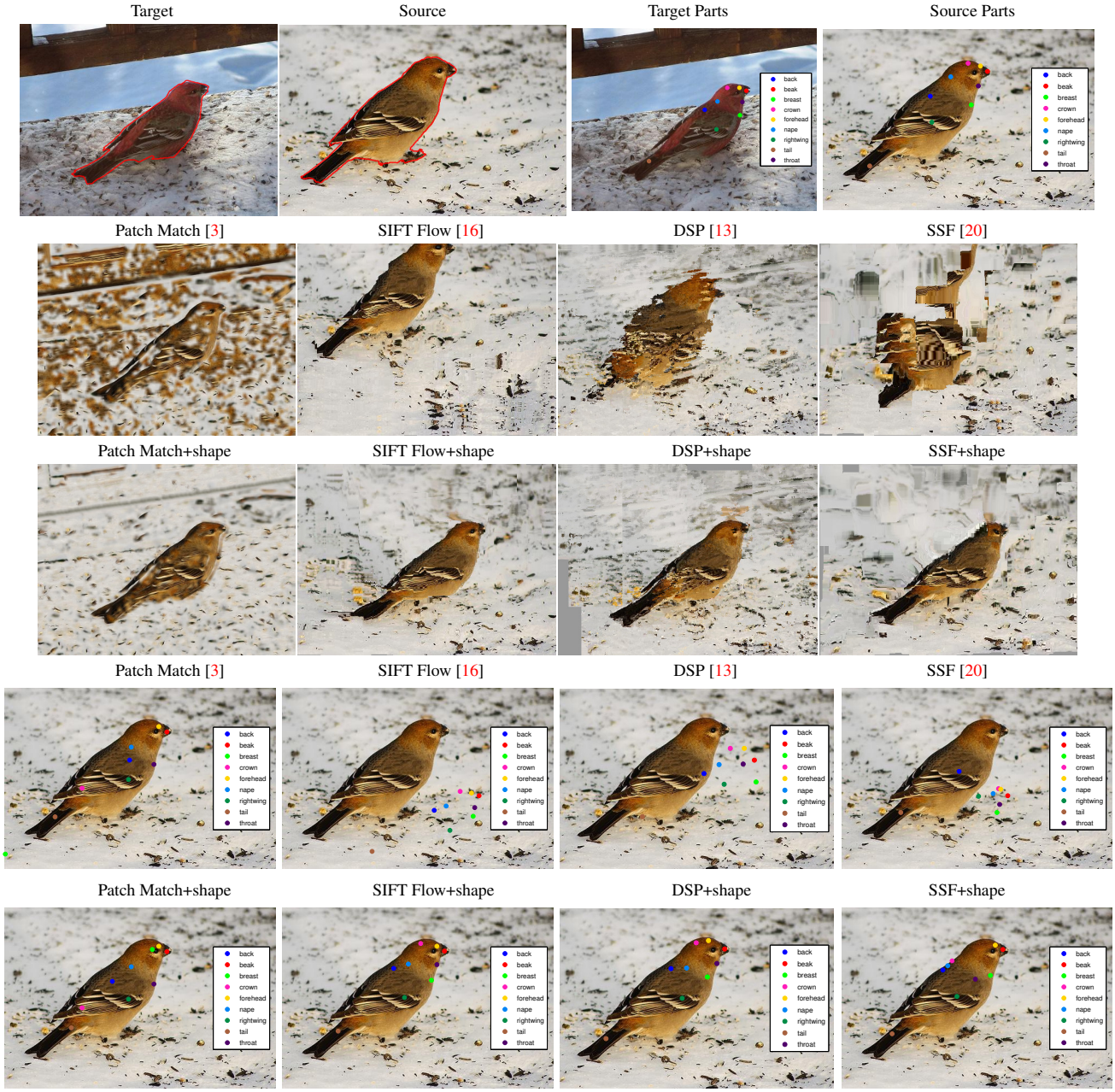

Figure 4: **A qualitative result from the CUB-200-2011 dataset [21].** First row shows the target and source images with ground truth segmentations marked in red along with the target and source image part locations. Second and third row show the warped source image using the default settings and with shape constraints respectively. The expectation is that the warped source image reconstructs the target image with the appearance of the source image. The fourth and fifth row show the matchings of the target image part locations on the source image using the default settings and with shape constraints respectively. Ideally, these should be equal to the source image part locations shown in first row fourth column. See also supplementary material.

in Fig. 4. The average Jaccard index of the segmentations provided by object proposals automatically is 73.8%. Compared to Table 2, which uses ground truth segmentations (hence having a Jaccard index of 100%), there is a 9.8% drop in performance. (Table 2 shows 35% improvement over default algorithms compared to 25.2% improvement shown in Table 4.) Note that due to discarding image pairs used to train our ranker, the default performances of algorithms shown in Table 4 are slightly different than the ones shown in Table 2.

5 Conclusion

In this paper we have explored the inclusion of shape constraints in improving dense correspondence computations. We have found that shape constraints *(i)* do not help with images

Figure 5: **A qualitative result from the PASCAL-Part dataset [6].** First row shows the target and source images with ground truth segmentations marked in red along with the target and source image part masks. Second and third row show the warped source image using the default settings and with shape constraints respectively. The expectation is that the warped source image reconstructs the target image with the appearance of the source image. The fourth and fifth row show the part mask transfer from the source image to the target image using the default settings and with shape constraints respectively. Ideally, the part masks transferred from the source image should coincide with the part masks of the target image. See also supplementary material.

of the same scene acquired under slightly different conditions; *(ii)* improve the results of pairs of images of the same scene but under largely different imaging conditions. Finally, we found the highest impact of shape constraints in *(iii)* semantic image alignment, *i.e.,* when the scene content of the two images are only categorically related and when scene component configurations vary. Such shape correspondences can be manually specified. However, we have begun exploring the use of object proposals and have shown very promising results in this direction.

parts	Patch Match [3]	Patch Match+shape	SIFT Flow [16]	SIFT Flow+shape	DSP [13]	DSP+shape	SSF [20]	SSF+shape
back	0.20 ± 0.13	**0.13 ± 0.09**	0.09 ± 0.07	**0.07 ± 0.06**	0.08 ± 0.05	**0.06 ± 0.06**	0.14 ± 0.09	**0.08 ± 0.07**
beak	0.25 ± 0.16	**0.21 ± 0.15**	0.14 ± 0.11	**0.12 ± 0.11**	0.12 ± 0.10	**0.09 ± 0.11**	0.18 ± 0.11	**0.13 ± 0.11**
belly	0.21 ± 0.13	**0.13 ± 0.10**	0.09 ± 0.06	**0.07 ± 0.05**	0.08 ± 0.05	**0.06 ± 0.05**	0.12 ± 0.08	**0.08 ± 0.06**
breast	0.21 ± 0.14	**0.12 ± 0.10**	0.11 ± 0.08	**0.09 ± 0.07**	0.09 ± 0.06	**0.07 ± 0.07**	0.14 ± 0.09	**0.09 ± 0.08**
crown	0.23 ± 0.16	**0.17 ± 0.14**	0.12 ± 0.09	**0.09 ± 0.09**	0.10 ± 0.07	**0.06 ± 0.08**	0.16 ± 0.10	**0.10 ± 0.09**
forehead	0.23 ± 0.16	**0.17 ± 0.14**	0.13 ± 0.10	**0.10 ± 0.10**	0.11 ± 0.08	**0.07 ± 0.09**	0.17 ± 0.11	**0.11 ± 0.10**
lefteye	0.19 ± 0.17	**0.13 ± 0.14**	0.12 ± 0.09	**0.09 ± 0.09**	0.10 ± 0.08	**0.06 ± 0.07**	0.17 ± 0.11	**0.10 ± 0.10**
leftleg	0.23 ± 0.14	**0.21 ± 0.13**	0.10 ± 0.07	**0.09 ± 0.06**	0.09 ± 0.06	**0.08 ± 0.06**	0.12 ± 0.08	**0.09 ± 0.07**
leftwing	0.19 ± 0.13	**0.13 ± 0.10**	0.11 ± 0.07	**0.09 ± 0.07**	0.10 ± 0.06	**0.09 ± 0.07**	0.14 ± 0.09	**0.10 ± 0.07**
nape	0.21 ± 0.14	**0.13 ± 0.11**	0.10 ± 0.07	**0.07 ± 0.06**	0.09 ± 0.06	**0.06 ± 0.06**	0.15 ± 0.09	**0.09 ± 0.07**
righteye	0.20 ± 0.17	**0.14 ± 0.14**	0.11 ± 0.09	**0.09 ± 0.08**	0.09 ± 0.07	**0.06 ± 0.09**	0.15 ± 0.10	**0.09 ± 0.08**
rightleg	0.23 ± 0.14	**0.21 ± 0.13**	0.10 ± 0.06	**0.09 ± 0.06**	0.09 ± 0.06	**0.08 ± 0.06**	0.13 ± 0.08	**0.10 ± 0.07**
rightwing	0.20 ± 0.13	**0.13 ± 0.09**	0.11 ± 0.08	**0.10 ± 0.07**	0.10 ± 0.07	**0.09 ± 0.08**	0.15 ± 0.09	**0.10 ± 0.08**
tail	0.26 ± 0.16	**0.24 ± 0.15**	0.18 ± 0.15	**0.17 ± 0.15**	0.14 ± 0.12	**0.13 ± 0.13**	0.19 ± 0.14	**0.17 ± 0.15**
throat	0.24 ± 0.15	**0.15 ± 0.12**	0.12 ± 0.09	**0.09 ± 0.09**	0.10 ± 0.08	**0.07 ± 0.08**	0.16 ± 0.10	**0.10 ± 0.09**

Table 4: **Normalized part location error in the CUB-200-2011 dataset [21] using R-CNN [8] and object proposals.** The aim here is not to declare a winner but to emphasize the significant improvements obtained, irrespective of the method used, when shape constraints are used.

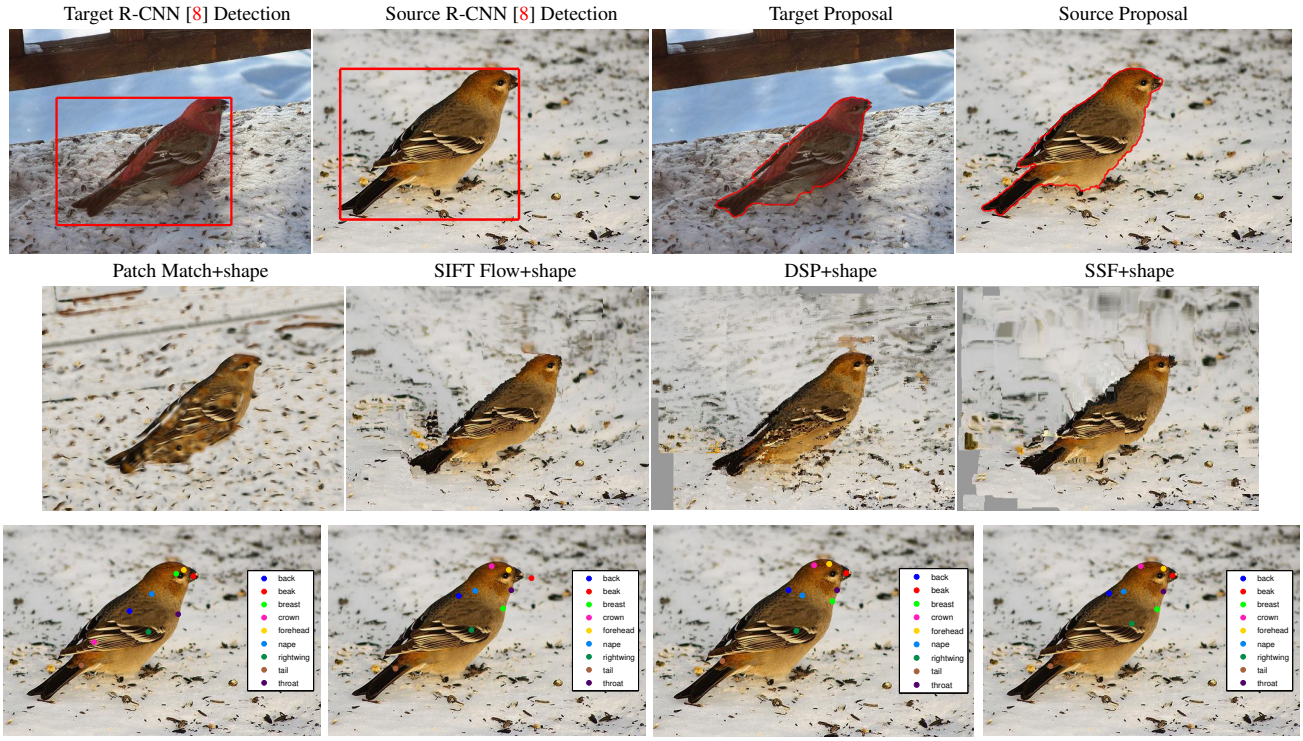

Figure 6: **The qualitative result using the same image pair as shown in Fig. 4 from the CUB-200-2011 dataset [21] using R-CNN [8] and object proposals.** First row shows the R-CNN [8] bounding box detection results along with the top ranked object proposals for source and target images. We do not repeat the results of the four methods in default settings, as shown in Fig. 4, and instead show the results of the four methods using shape constraints provided by object proposals and the results of part matchings in the second and third rows respectively.

References

[1] H. Aanæs, A.L. Dahl, and K. Steenstrup Pedersen. Interesting interest points. *International Journal of Computer Vision*, 97:18–35, 2012.

[2] Pablo Andrés Arbeláez, Jordi Pont-Tuset, Jonathan T. Barron, Ferran Marqués, and Jitendra Malik. Multiscale combinatorial grouping. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, pages 328–335, 2014.

[3] Connelly Barnes, Eli Shechtman, Dan B. Goldman, and Adam 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.

[4] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black. A naturalistic open source movie for optical flow evaluation. In *European Conf. on Computer Vision (ECCV)*, pages 611–625, 2012.

[5] João Carreira and Cristian Sminchisescu. CPMC: automatic object segmentation using constrained parametric min-cuts. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(7):1312–1328, 2012.

[6] Xianjie Chen, Roozbeh Mottaghi, Xiaobai Liu, Sanja Fidler, Raquel Urtasun, and Alan Yuille. Detect what you can: Detecting and representing objects using holistic models and body parts. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[7] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2): 303–338, June 2010.

[8] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[9] Yoav HaCohen, Eli Shechtman, Dan B. Goldman, and Dani Lischinski. Non-rigid dense correspondence with applications for image enhancement. *ACM Trans. Graph.*, 30(4):70, 2011.

[10] Junhwa Hur, Hwasup Lim, Changsoo Park, and Sang Chul 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.

[11] Thorsten Joachims. Training linear svms in linear time. In *Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia, PA, USA, August 20-23, 2006*, pages 217–226, 2006.

[12] Kevin Karsch, Ce Liu, and Sing Bing 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.

[13] Jaechul Kim, Ce Liu, Fei Sha, and Kristen 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.

[14] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1106–1114, 2012.

[15] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2006), 17-22 June 2006, New York, NY, USA*, pages 2169–2178, 2006.

[16] Ce Liu, Jenny Yuen, and Antonio Torralba. SIFT flow: Dense correspondence across scenes and its applications. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(5):978–994, 2011.

[17] Ce Liu, Jenny Yuen, and Antonio Torralba. Nonparametric scene parsing via label transfer. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(12):2368–2382, 2011.

[18] Jonathan Long, Ning Zhang, and Trevor 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.

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

[20] Weichao Qiu, Xinggang Wang, Xiang Bai, Alan L. Yuille, and Zhuowen 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.

[21] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

[22] Hongsheng Yang, Wen-Yan Lin, and Jiangbo 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.