Deep Semantic Feature Matching

Nikolai Ufer and Björn Ommer
Heidelberg Collaboratory for Image Processing
IWR, Heidelberg University, Germany
nikolai.ufer@iwr.uni-heidelberg.de ommer@uni-heidelberg.de

Abstract

Estimating dense visual correspondences between objects with intra-class variation, deformations and background clutter remains a challenging problem. Thanks to the breakthrough of CNNs there are new powerful features available. Despite their easy accessibility and great success, existing semantic flow methods could not significantly benefit from these without extensive additional training. We introduce a novel method for semantic matching with pre-trained CNN features which is based on convolutional feature pyramids and activation guided feature selection. For the final matching we propose a sparse graph matching framework where each salient feature selects among a small subset of nearest neighbors in the target image. To improve our method in the unconstrained setting without bounding box annotations we introduce novel object proposal based matching constraints. Furthermore, we show that the sparse matching can be transformed into a dense correspondence field. Extensive experimental evaluations on benchmark datasets show that our method significantly outperforms existing semantic matching methods.

1. Introduction

Finding correspondences between images is a fundamental problem of computer vision and key to many applications like 3D reconstruction, video analysis, image retrieval and object recognition. Classical correspondence methods like stereo matching [21] and optical flow [23, 35] consider input images showing same objects or scenes from different viewpoints. With the development of better features which are more robust against deformations and appearance changes, researchers started to estimate correspondences across different instances and scenes of the same semantic category. In the literature this problem is often denoted as *semantic matching* or *semantic flow* in the case of dense correspondences.

Despite of the success of deep features in many fields of computer vision, previous work on semantic matching [20, 33] reported that pre-trained CNN features perform similarly or even worse compared to hand-engineered features such as SIFT [34] or HOG [12, 22]. In this work we revisit deep semantic feature matching and propose an efficient algorithm specifically designed for this task, which addresses the following main issues of previous approaches:

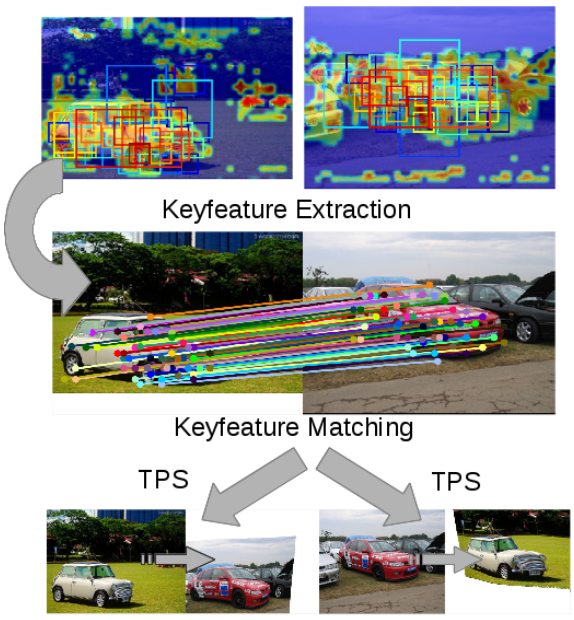


Figure 1: Overview of our approach for dense semantic matching. After salient feature extraction we utilize a MRF for finding sparse correspondences, which are used to estimate a dense flow field using thin plate splines (TPS).

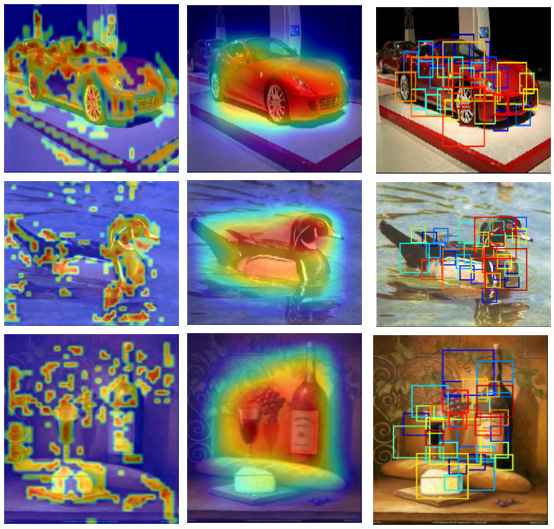
(1) Local structures are not robust against intra-class variation and for finding semantic correspondences more context is necessary. We extract particular context-sensitive features by utilizing a deep feature pyramid representation [19], where we encode image regions by aggregating respective cells of the feature pyramid over two levels. This provides more discriminative region descriptors compared to methods where regions are just cropped, rescaled and

passed through the network [20, 50].

(2) Recently, Ham et al. [20] introduced a more general task of semantic matching, namely to align two objects in real-world images without information about their class, scale and location. In this unconstrained setting with no bounding box annotations and severe background clutter, matching approaches using grid-based feature sampling [28, 31] or classical feature detectors like MSER [37] are prone to incorrect correspondences. To focus on object like structures, Ham et al. [20] utilize modern object proposal methods [4, 36, 45, 51, 24]. But this does not address the issue that convolutional filters learn in particular to respond to image regions which are discriminative for the original classification task. In contrast to approaches which learn latent parts using a large set of instances of the same object category [42, 3, 16, 40], we try to find well-encoded latent structures using the convolutional filter responses of a single image. Although, not the whole object is covered by these regions, they guide weaker encoded regions towards geometric consistent matches.

(3) Semantic matching of objects in real-world images requires spatial regularization which is on the one hand capable to overcome the matching ambiguity and on the other hand flexible enough to adjust to non-rigid deformations. Therefore, we introduce a sparse MRF framework which reduces matching ambiguities by enforcing geometric consistency between all feature pairs. This is infeasible for pixel-level approaches [28, 31] which estimate a continuous displacement field. Moreover, our framework is capable of adjusting to more complex object deformations compared to Hough Voting based approaches [20]. Feature detection is in particular not stable against intra-class variation, meaning that detected features in one image may not have a correspondence in the set of detected features in the other image. Therefore, standard graph matching approaches [43, 9, 10] extract a large set of features in both images and search for correspondences between these sets. This leads to a large number of outliers and matching ambiguities. In contrast, we use salient features as sliding-window detectors and extract a small number of nearest neighbors as matching candidates. In this way we simplify our optimization problem and alleviate the combinatorially difficult one-to-one matching constraint.

Contributions. The main contributions of this paper are threefold: (1) We present an efficient MRF framework utilizing deep feature pyramids [19] and convolutional activation guided feature selection for semantic matching. (2) In the unconstrained setting of unknown object location, we introducing new unary and pairwise terms for incorporating object proposals in our formulation. (3) We demonstrate that the proposed method significantly outperforms state-of-the-art semantic matching methods on challenging benchmark datasets.



(a) (b) (c)

Figure 2: Visualization of our key feature extraction approach. Column (a) shows the rescaled cell entropy of the first level of the feature pyramid, as defined in Equ. 1. Column (b) visualizes the pixel-wise object probability defined in Equ. 11. And column (c) shows the selected key features using non-maximum suppression on the combined entropy and pixel-wise object probability maps.

2. Related work

Image alignment is a key problem of computer vision and a large body of preliminary work exists. In the following we will focus on the most relevant research in the field of semantic matching and semantic flow.

First steps towards semantic matching was done by Liu et al. with the development of SIFT Flow [31]. Inspired by optical flow methods they densely sampled SIFT features and formulated a discrete optimization problem for solving a displacement field in a hierarchical scheme. Kim et al. [28] extended this approach by incorporating links between pyramid levels in the graph and defining matching costs of nodes using multiple descriptors. Inspired by deep convolutional neural networks Weinzaepfl et al. [39] estimated dense correspondences by using a multi-layer architecture of several layers interleaving convolutions and max pooling. More recently, Bristow et al. [6] used the graphical model of SIFT Flow and replaced the unary term with similarities of pixel-wise LDA classifier for improving the robustness against intra-class variations. For the task of object discovery and localization without any information about input images Cho et al. [8] introduced a region matching approach using off-the-shelf object proposals as candidate regions and a probabilistic Hough voting scheme as a spatial regularizer. Proposal Flow [20] extended the region-

matching idea by introducing a local Hough Voting based on neighboring regions and estimated a dense flow field using region correspondences.

In recent years more and more deep learning based methods have been proposed. There are approaches using CNNs as plain feature extractors without any taks specific design or training. For example Fischer et al. [18] reported that CNN features clearly outperform SIFT in the task of nearest neighbor matching. Long et al. [33] studied the capabilities of deep features for semantic alignment by investigating a SIFT Flow version with CNN features of a pre-trained classification network. But they achieved only slightly better performance results compared to the original SIFT Flow algorithm. Also Ham et al. [20] investigated the impact of deep features on their approach without any performance gain. Several deep learning based methods are utilizing specifically designed and trained architectures. Most of these methods need additional ground truth data for their learning procedure. For example, Kanazawa et al. [25] use a pre-trained classification network in combination with thin plate splines extracted from segmentation masks for learning a spatial correspondence prior. Moreover, they use additional ground truth segmentation masks for their final ratio-test based matching. Additional data in form of synthetic rendered 3D models are used by Zhou et al. [48] for formulating a cycle constraint between images. And Choy et al. [11] use ground truth correspondences for optimizing a correspondence objective for semantic matching. Our approach is in line of the first mentioned research direction and uses pre-trained CNN features extracted from a classification network without any additional data or training.

3. Proposed approach

In this chapter we present our semantic flow algorithm which is based on pre-trained CNN features and sparse MRF matching. It first builds a feature pyramid of each image and selects salient features for matching on different scales using informatic criterion on the cell activations of the pyramid. For each selected feature a set of matching candidates is extracted and the final assignment is obtained by solving an energy minimization problem with an unary appearance and a binary geometric term. Long-range contextual relationships are preserved by a fully connected graph. To improve results in real-world images without bounding box annotations, additional unary and binary objectness potentials are introduced. Finally, we estimate a dense flow field from the sparse correspondences using thin plate splines (TPS) [5, 14].

3.1. Spatial feature pyramid

Since objects may occur at different scales our algorithm is based on a multiple image resolution approach. In particular, we extract a spatial feature pyramid analogously to Girshick et al. [19]. For each input image a pyramid of 8 levels with scaling factor $2^{-1/2}$ is generated. For the first level the input image is padded and rescaled without change of aspect ratio and fixing the largest dimension. Based on the image pyramid a feature pyramid is generated, where each level consists of convolutional features extracted from images with decreasing resolution. In our experiments we use $\mathrm{Conv}\,4$ features of AlexNet [29] pre-trained on ILSVRC2012 [13]. Each cell of the feature pyramid has 384 channels and for each cell we associate a squared region in the input image at the center of its receptive field. We set the size of these regions to the receptive field stride, i.e. 16 pixels. To increase the descriptiveness we concatenate neighboring 5×5 cells over two pyramid levels and obtain an overall feature vector dimension of $19 \cdot 10^3$. The concatenated cells induce a grid of squared regions with sizes between 64 up to 724 pixels in the original image space. The receptive field sizes of these regions are between 195 up to the full image size.

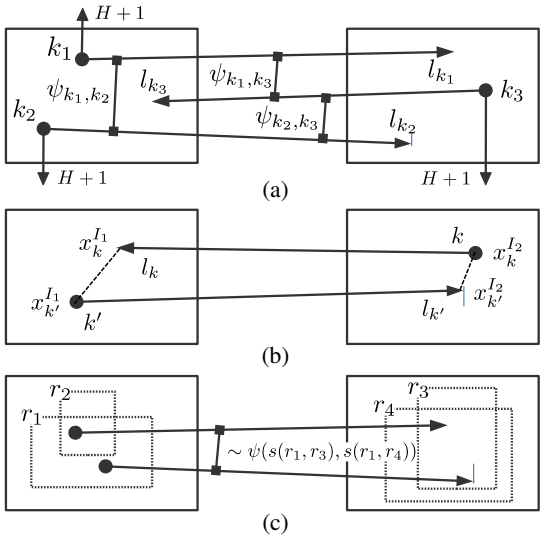


Figure 3: Figure (a) shows the binary geometric potentials of our basic formulation, (b) edges between two assignments and (c) the additional binary potential of our object-proposal guided matching.

3.2. Key feature selection

We select salient features, which we will denote *key features* in the following, based on the extracted pyramid feature maps. We introduce two criteria. Firstly, the overall signal should be strong. We quantify this by computing the cell-wise sum of activations over all feature map channels and set a threshold. Secondly, the information contained in the signal should be as high as possible. Therefore, we use the entropy according to Shannon's formula [41] as a

standard information theoretic-ranking criterion, i.e.

$$S_H(\boldsymbol{p}_s) = -\sum_{c=1}^{N_c} f_c(\boldsymbol{p}_s) log_2(f_c(\boldsymbol{p}_s)), \quad (1)$$

where $f(\boldsymbol{p}_s)$ are the normalized features at position $\boldsymbol{p}_s$ and feature pyramid level s. Based on the entropy maps over several levels we select a fixed number of features using non-maximum suppression, which leads to a good coverage and key features of different scales. Since too large image regions may lead to bad localization results we restrict the feature selection to some of the first pyramid levels. Fig. 2 illustrates our key feature extraction approach in the unconstraint setting, where we include the pixel-wise object probabilities as additional criteria, which will be explained in Sect. 3.4.

3.3. Key feature based matching

In general feature detection is not perfectly repeatable [44], which means detected features in one image may not have a correspondence in the set of detected features in the other image. This holds in particular for semantic matching and our feature extraction. Therefore, we directly search for nearest neighbors as matching candidates in the opposite image, where we use a sliding window approach over several pyramid levels and the cosine similarity metric to find them.

Matching energy function. The overall matching task is to assign each key feature $k \in \mathcal{K}$ to its most consistent matching candidate $l_k \in \mathcal{H}_k$, which can be formulated as an energy minimization problem,

$$E_s(\boldsymbol{l}) = \sum_{k \in \mathcal{K}} \psi_k(l_k) + \sum_{(k,k') \in \mathcal{K} \times \mathcal{K}} \psi_{k,k'}(l_k, l_{k'}), \quad (2)$$

where ψ_k models feature similarity and $\psi_{k,k'}$ geometric compatibility of pairwise assignments. For simplicity, we assume a fixed number H of matching candidates for all K key features. In the following we explain the energy potentials in more detail.

Unary appearance potential. The function ψ_k favors correspondences between features with similar appearance, which is defined as

$$\psi_k(l_k) = \begin{cases} \lambda_f \left(e^{(1-sim(f_k, f_{l_k}))^2/\sigma_f^2} - 1 \right), & \text{if } l_k \neq H+1, \\ \lambda_{occ} e_{occ}, & \text{otherwise,} \end{cases} \quad (3)$$

where f_k, f_{l_k} are respective feature descriptors as explained in Sect. 3.1 and sim the cosine similarity. At this point, we introduced an additional label $H+1$ which accounts for the possibility that features are not assigned to any candidate, which imposes a constant penalty e_{occ}. If the key feature is assigned to a candidate we call the assignment as being active.

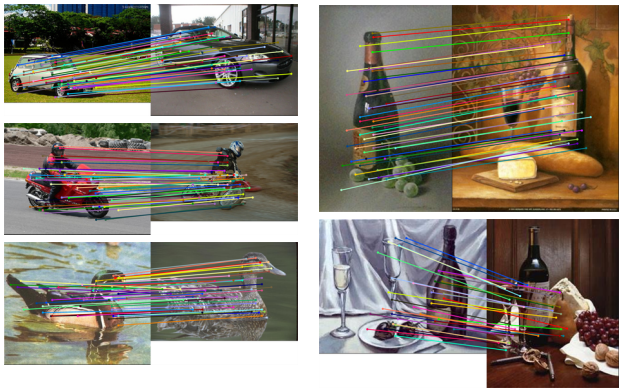


Figure 4: Matching examples on the Proposal Flow dataset [20] using our object proposal guided matching. The bottom right example shows a failure case.

Binary geometric potential. The function $\psi_{k,k'}$ enforces spatial consistency between assignments and consists of two terms,

$$\psi_{k,k'}(l_k, l_{k'}) = \delta(k, k', l_k, l_{k'}) \cdot \hat{\psi}_{k,k'}(l_k, l_{k'}), \quad (4)$$

where $\hat{\psi}_{k,k'}$ measures the geometric consistency and δ models the spatial range of this term, which will be described in the next sub-point. Inspired by graph matching approaches [43] we enforce geometric consistency between two active assignments by the relative length difference and absolute angle of corresponding edges, i.e.

$$\hat{\psi}_{k,k'}(l_k, l_{k'}) = \Big[\lambda_d \left(e^{(d^2_{k,k'}(l_k, l_{k'})/\sigma_d^2)} - 1 \right) + \lambda_\gamma \left(e^{(\gamma^2_{k,k'}(l_k, l_{k'})/\sigma_\gamma^2)} - 1 \right) \Big], \quad (5)$$

where λ_d and λ_γ are scalar weights. See Fig. 3b for an illustration of edges and the notation of feature locations. The function $d_{k,k'}$ measures the relative length difference of edges between two assignments, i.e.

$$d_{k,k'}(l_k, l_{k'}) = \left| \|x_k^{I_1} - x_{k'}^{I_1}\|/D_{I_1} - \|x_k^{I_2} - x_{k'}^{I_2}\|/D_{I_2} \right|, \quad (6)$$

where D_{I_1}, D_{I_2} are bounding box diagonals and $x_k^{I_1}$, $x_k^{I_2}$ feature locations of assignment $k \mapsto l_k$ in image I_1, I_2, respectively. If no bounding box annotations are given we set term (6) to zero. The function $\gamma_{k,k'}$ in (5) measures the absolute angle between two edges, i.e.

$$\gamma_{k,k'}(l_k, l_{k'}) = \arccos \left(\frac{x_k^{I_1} - x_{k'}^{I_1}}{\|x_k^{I_1} - x_{k'}^{I_1}\|} \cdot \frac{x_k^{I_2} - x_{k'}^{I_2}}{\|x_k^{I_2} - x_{k'}^{I_2}\|} \right), \quad (7)$$

Geometric interaction range. Stronger geometric constraints help to overcome matching ambiguities and to find

consistent matches. Therefore, we consider a fully connected graph which enforces geometric consistency between all feature pairs. But in the case of severe view-point changes and object deformations this may lead to geometric inflexibilities. To balance this effect gracefully, we include a damping function $\delta(\cdot)$ which reduces the influence of the binary term if both feature pairs are far away from each other. This is done by a sigmoid function, i.e.

$$\delta(k, k', l_k, l_{k'}) = \frac{1}{1 + e^{-(d_{min} - d_o)/\sigma_\delta}}, \tag{8}$$

$$d_{\min} := \min(\|x_k^{I_1} - x_{k'}^{I_1}\|/D_{I_1}, \|x_k^{I_2} - x_{k'}^{I_2}\|/D_{I_2}), \tag{9}$$

where d_o and σ_δ determines the offset and steepness. This term depends on the object scale and we set it to one if no bounding box information is available.

Model properties. Our model is invariant to translation and scale but due to the use of absolute angles it is rotation dependent. By setting pairwise costs of two assignments with one identical feature to infinity, for example if two key feature have the same matching candidate, our model produces valid one-to-one matchings. But this occurs very rarely and the overall influence of this constraint is very limited.

3.4. Object-proposal guided matching

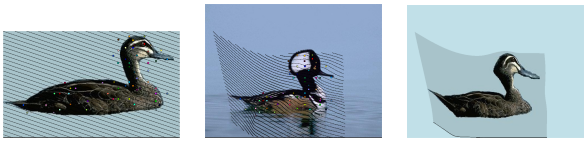
In this section, we consider the unconstrained setting of aligning objects without bounding box annotations and background clutter. To reduce the resulting matching ambiguities we utilize generic object proposal methods [45] inspired by Ham et al. [20]. In particular we modify the key feature extraction approach and introduce an additional unary and binary term.

Unary objectness potential. For each image we assume that one object proposal $\hat{r}$ in the set of all extracted proposals $\mathcal{R}$ exists, which covers the object to be matched perfectly. Then we estimate the probability of a pixel x being in $\hat{r}$ with the following marginal probability over all region proposals,

$$p(x|\mathcal{R}) = \sum_{r \in \mathcal{R}} p(x|r)p(r|\mathcal{R}) \approx \sum_{r \in \mathcal{R}(\{x\})} p(r|\mathcal{R}), \tag{10}$$

where $\mathcal{R}(\mathcal{P})$ denotes object proposals containing all pixels in the set $\mathcal{P}$. Notice, the probability $p(x|r)$ is zero if $x \notin r$ and in the other case we assume a uniform distribution, since we do not consider restrictions regarding the object position. We estimate the second probability with a Gibbs distribution over the object proposal scores and obtain our final pixel-wise object probability, which is given by

$$p(x|\mathcal{R}) = \frac{1}{Z} \sum_{r \in \mathcal{R}(x)} e^{\beta s_{obj}(r)}, \tag{11}$$



(a) (b) (c)

Figure 5: Thin plate spline transformation from source image (a) to target image (b) using the estimated point correspondences, where image (c) shows the warped result.

where Z is the partition function of the Gibbs distribution and $s_{obj}(\cdot)$ the region proposal score function. We utilize this objectness prior in two ways. Firstly, we include this as an additional unary term,

$$\psi_k^{uo}(l_k) = -\lambda_{uo} \sum_{j=1,2} \log p\big(x_k^{I_j}|\mathcal{R}\big), \tag{12}$$

in our energy function (2) with a weighting factor λ_{uo}. Secondly, we utilize the pixel-wise object probability as an additional criterion for our key feature selection, such that our selection approach focuses on regions where the probability is high that the object is located there. In Fig. 2 some examples of object probability maps are shown.

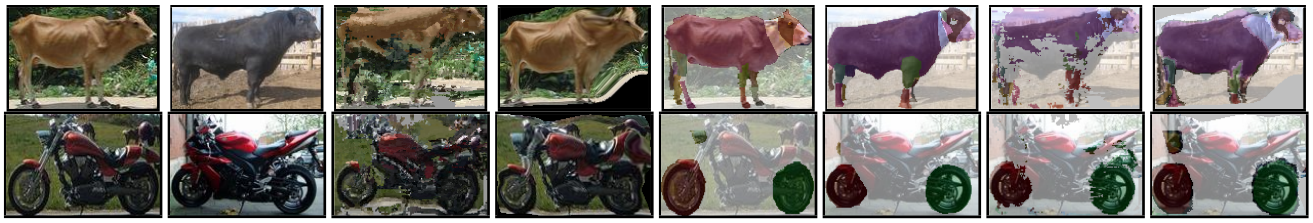
Binary objectness potential. Besides the unary term we include an additional binary term. Given object proposal sets $\mathcal{R}_1$, $\mathcal{R}_2$ in images I_1, I_2, respectively, we estimate the probability that key features k, k' are assigned to hypotheses l_k, $l_{k'}$ and both are lying inside the dominant object, with the marginal distribution

$$p\left(l_k, l_{k'}|\mathcal{R}, \mathcal{R}'\right) = \sum_{\substack{r_1 \in \mathcal{R}_1(k,k') \\ r_2 \in \mathcal{R}_2(k,k')}} p\left(l_k, l_{k'}|r_1, r_2\right) p(r_1, r_2|\mathcal{R}_1, \mathcal{R}_2)\,, \tag{13}$$

where $\mathcal{R}_j(k, k') := \mathcal{R}_j(\{x_k^{(j)}, x_{k'}^{(j)}\})$. Again, the second probability is zero if one of the assigned features is not located in one of the object proposals. Since the dominant objects in both images belong to the same category we model the second probability with a Gibbs distribution, which favors similar appearance and aspect ratio of proposals r_1 and r_2. Regarding the first probability, we use the given object proposals as reference frames and model it as a Gibbs distribution favoring similar edge lengths $\|x_k^{I_1} - x_{k'}^{I_1}\|$ and $\|x_k^{I_2} - x_{k'}^{I_2}\|$ relative to the diagonals of r_1 and r_2. Analogously to the unary prior, we include the probability (13) as an additional binary term,

$$\psi_{k,k'}^{bo}(l_k, l_{k'}) = -\lambda_{bo} \log p\big(l_k, l_{k'}|\mathcal{R}_1, \mathcal{R}_2\big), \tag{14}$$

in our matching energy with a weighting factor λ_{bo}.



(a) Source. (b) Target. (c) DSP. (d) Ours. (e) Source. (f) Target. (g) DSP. (h) Ours.

Figure 6: Qualitative examples on the PASCAL-Part dataset [7]. Column (a), (b): Source and target image. Column (c), (d): Warping results using DSP and our method. Column (e), (f): Annotated part segments for source and target image. Column (g), (h): Predicted part correspondences using DSP and our method. (Best viewed in pdf.)

Methods	aero	bike	boat	bottle	bus	car	chair	table	mbike	sofa	train	tv	Avg.
Ours	**0.23**	**0.36**	**0.05**	0.36	**0.45**	**0.42**	**0.14**	**0.08**	**0.23**	0.18	0.16	**0.33**	**0.25**
Ours (EB)	0.23	0.31	0.05	**0.37**	0.41	0.38	**0.14**	**0.08**	0.20	**0.19**	**0.17**	**0.33**	0.24
Ours (NN)	0.20	0.21	0.04	0.21	0.29	0.29	0.06	0.03	0.11	0.09	0.10	0.19	0.15
DSP [28]	0.17	0.3	0.05	0.19	0.33	0.34	0.09	0.03	0.17	0.12	0.12	0.18	0.17
Collection Flow [27]	0.16	0.17	0.04	0.31	0.25	0.16	0.09	0.02	0.08	0.07	0.06	0.09	0.12
RASL [38]	0.18	0.17	0.04	0.33	0.31	0.17	0.09	0.04	0.12	0.1	0.11	0.23	0.16
Congealing [30]	0.12	0.23	0.03	0.22	0.19	0.14	0.06	0.04	0.12	0.07	0.08	0.06	0.11
Flow Web [49]	0.29	0.41	0.04	0.34	0.54	0.5	0.14	0.04	0.21	0.15	0.15	0.33	0.26

Table 1: PCK on 12 rigid PASCAL-Part classes using FlowWeb [49] clusters ($\alpha = 0.05$).

Methods	IOU	PCK
Ours	**0.43**	**0.25**
Proposal Flow [20]	0.41	0.17
Congealing [30]	0.38	0.11
RASL [38]	0.39	0.16
Collection Flow [27]	0.38	0.12
DSP [28]	0.39	0.17
Flow Web [49]	0.43	0.26

Table 2: Evaluation of dense flow field on the PASCAL-Part dataset following the FlowWeb [49] evaluation protocol.

3.5. Inference

The discrete optimization problem in Equ. 2 is an Integer Quadratic Program (IQP) which is NP hard and optimization methods with polynomial complexity do not exist. Therefore, we have to use approximate inference methods. For solving the optimization problem we use the discrete graphical model library OpenGM [2] and use the fusion algorithm from Kappes et al. [26] for inference, where we choose Loopy Belief Propagation [17] as proposal generator and Lazy Flipping of search depth 2 [1] as fusion operator.

3.6. Semantic flow field

Depending on the input image pair, our sparse graph matching gives a set of 30-60 point correspondences, see Fig. 4. In most cases, these are quite uniformly distributed over the whole object and a standard TPS [5, 14] then generalizes this to a dense flow field, see Fig. 5.

4. Experimental evaluation

In this chapter we present comparative evaluations and diagnostic experiments using the publicly available benchmark datasets of PASCAL-Part [7] and Proposal Flow [20].

4.1. Key feature based matching

Firstly, we evaluate our key feature based matching in the setting of known object locations. We measure the accuracy of transferred keypoints and segmentation masks by following the evaluation protocol of FlowWeb [49]. The dataset consists of representative viewpoint clusters of the PASCAL-Part dataset [7]. In addition, body part masks and keypoint annotations are provided [46].

Experimental details. We pad images by 24 pixels on all sides and upscale them to 721 pixels maximum dimension. We set the number of key features K per image to 35 and the number of hypotheses H to 10. The key feature selection is restricted to the first three pyramid levels. Since object location and scale is given we utilize all MRF terms introduced in Sect. 3.3. We determined the parameters of our MRF using cross-validation on a small subset.

Part segment matching. We evaluate the quality of estimated flow fields based on the transformation of part segmentation masks. As quantitative measure, we use the weighted intersection over union (IOU), where the weights are determined by the area of each part. For classes without

Method	car(S)	car(G)	car(M)	duck(S)	mot(S)	mot(G)	mot(M)	win(w/o C)	win(w/ C)	win(M)	Avg.
Ours	0.91	**0.67**	0.66	**0.77**	**0.63**	**0.35**	**0.4**	0.88	**0.67**	**0.8**	**0.68**
Ours (w/o BOP)	0.9	0.66	**0.67**	0.74	0.62	**0.35**	**0.4**	**0.89**	0.63	0.78	0.66
Ours (w/o BOP,UOP)	0.86	0.66	0.62	0.61	0.55	0.33	0.34	0.83	0.63	0.78	0.62
LOM [20]	**1.0**	0.59	0.51	0.65	0.47	0.27	0.27	0.91	0.41	0.67	0.56

Table 3: Detailed per class PCK comparison ($\alpha = 0.1$) between Proposal flow [20] and our approach.

Method	PCK
Ours	**0.68**
LOM [20]	0.56
GMK [15]	0.27
SIFT Flow [32]	0.38
DSP [28]	0.37

Table 4: PCK evaluation ($\alpha = 0.1$) of dense flow field on the PF dataset.

part annotations, object silhouettes are used. In Table 2 the mean IOU value over all classes is provided. Our method outperforms all other pairwise correspondence methods and only Flow Web shows similar performance. Fig. 6 shows two qualitative examples.

Keypoint matching. For measuring keypoint transformation accuracy we use the percentage of correct keypoints (PCK) [47]. Each keypoint is transferred using the estimated flow field and we determine whether the keypoint is transferred correctly by measuring the Euclidean distance between predicted and annotated ground-truth correspondences. The predicted correspondence is correct if the Euclidean distance is lower than $\alpha \cdot \max(H, W)$, where H and W are the image height and width. The mean PCK values ($\alpha = 0.05$) over all classes are reported in Tab. 2 and a more detailed comparison per class in Tab. 1. Our method significantly outperforms all methods except for Flow Web, which is rather a post-processing method since it refines initial correspondences using cycle constraints between several images. Notice, we also showed significant improvement over DSP which is comparable to Long et al. [33].

Key feature selection. To demonstrate the influence of our proposed key feature selection method we perform the same experiment using the objectness score of EdgeBox [51]. To do so we compute the scores for each patch which belongs to a concatenation of 5×5 cells within a given feature pyramid and apply non-maximum suppression to get a good coverage of the image. The results are summarized in Tab. 1, where (EB) denotes the EdgeBox based selection procedure. Our specific feature detection improves the overall performance.

Nearest neighbour matching. We investigated the effect of nearest neighbor (NN) matching without any spatial constraint, see Tab. 1. The drastic performance drop shows the importance of our spatial regularization.

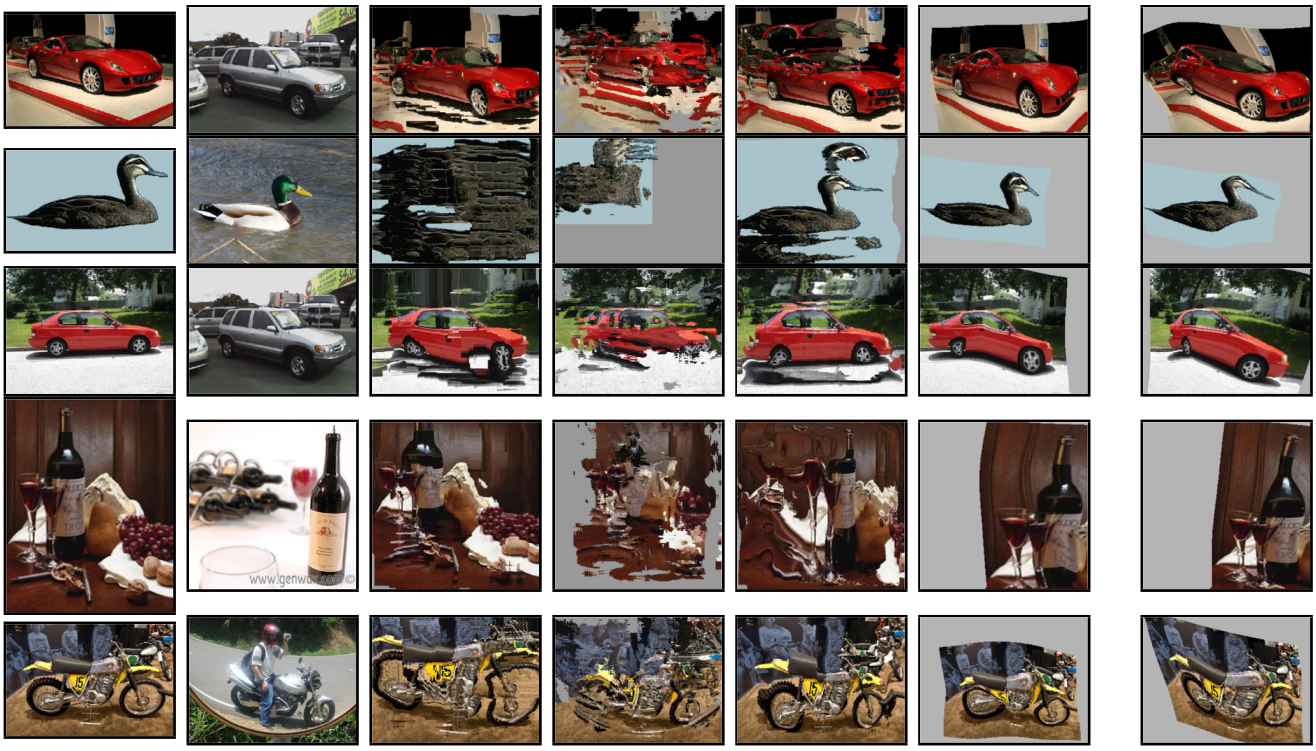
4.2. Object proposal guided matching

In this section, we evaluate our object proposal guided matching and follow the evaluation protocol of the Proposal Flow [20] benchmark. The dataset contains images with background clutter, intra-class variations, viewpoint changes and deformations. It consists of 4 main classes with several sub-classes according to background clutter and viewpoint changes [1].

Experimental details. We pad images by 64 pixels on all sides and upscale them to 931 pixels maximum dimension. Wet set the number of key features K to 35 and number of candidate matches to 5. The key feature selection is restricted to the first 4 pyramid levels. Since no bounding box annotations are available we neglect the MRF terms (6) and (8). Therefore, the only spatial regularization is enforced by the angle between edges. For the unary and binary objectness potentials we extract around 1000 object proposals using Selective Search [45]. The similarity between object proposals in Equ. 13 are determined using cosine similarity of the associated cells in the feature pyramid, where we use bilinear interpolation to get the same feature dimensionality. We determine the parameters of the MRF using cross-validation on a small subset. The time for inference during testing (including terms (3), (4), (12),(14)) is about 10 seconds.

Keypoint matching. Considering the keypoint matching accuracy, the variables H and W are now height and width of the rectangle drawn by annotated keypoints. In Tab. 4 we give a quantitative comparison with several semantic flow methods for $\alpha = 0.1$. Our approach significantly outperforms all methods. In Tab. 3 we give a detailed per class comparison with Proposal Flow [20]. Our method shows superior results for all classes, except for cars from the side and wine bottles without background clutter. For these classes a translation with unequal scaling is sufficient for getting a good alignment of keypoints. The table indicates that the classes mot(S) and mot(M) are much more

[1] The abbreviations (S) and (G) stand for side and general viewpoints and (C) for background clutter. The (M) indicates a mixture of images, i.e. mixed viewpoints for the class cars and a mixture of images with and without background clutter for the class wine bottles.



(a) Source (b) Target (c) SIFT Flow (d) DSP (e) PF (f) Ours (g) GT-TPS

Figure 7: Qualitative examples on the Proposal Flow dataset [20]. The source image (a) is warped to the target image (b) using various methods: SIFT Flow [32] (c), DSP [28] (d), Proposal Flow [20] (with SS and HOG) (e), our method (f) and TPS [5, 14] using annotated keypoints.

difficult compared to the rest. This is reasonable since a lot of the motorbikes are tilted or turned sideways and our model is not invariant against rotations. Fig. 7 gives some qualitative examples. Overall our method is more robust against view-point changes and background clutter.

Unary and binary objectness potentials. For evaluating the influence of the additional terms of our object proposal guided matching, we perform the following ablation studies. First we set the binary (w/o BOP) and then the binary and unary (w/o BOP,UOP) terms in addition with the modified key feature selection to zero and run our algorithm again, see Tab. 3. The unary term has overall more influence. This is reasonable since it guides assignments towards object like structures. We also measured the percentage of inlier correspondences between both objects. Therefore, we manually labeled ground-truth boxes covering the whole object, and measured the percentage of correspondences lying in both bounding boxes. By including the unary and binary objectness potentials the percentage of inliers increases from 60 to 88 percent.

5. Conclusion

We have presented a semantic matching algorithm using standard pre-trained CNN features without additional data or training. Our approach is based on a convolutional feature pyramid representation in combination with a salient feature selection method for extracting discriminative descriptors. Tailored to these descriptors we have proposed a candidate driven MRF matching formulation which circumvents the combinatorically difficult one-to-one matching constraint. Moreover, we have improved our method for the challenging task of matching unknown objects across different scenes by introducing new object-proposal based matching constraints, which leads to the majority of sparse correspondences are lying inside the unknown object bounding boxes. Experiments have shown competitive performance on standard semantic matching benchmark datasets.

6. Acknowledgement

This work has been supported in part by the German Research Foundation (DFG) within grant GRK 1653 and by a hardware donation from NVIDIA. The authors would also like to thank Jörg Hendrik Kappes for additional support.

References

[1] B. Andres, J. H. Kappes, U. Koethe, and F. A. Hamprecht. The Lazy Flipper: MAP Inference in Higher-Order Graphical Models by Depth-limited Exhaustive Search. In *Computational Complexity*, 9 2010. 6

[2] B. Andres, J. H. Kappes, U. Köthe, C. Schnörr, and F. A. Hamprecht. An empirical comparison of inference algorithms for graphical models with higher order factors using OpenGM. In *Lecture Notes in Computer Science*, volume 6376 LNCS, pages 353–362, 2010. 6

[3] B. Antic and B. Ommer. Learning latent constituents for recognition of group activities in video. *Proceedings of the European Conference on Computer Vision*, 2014. 2

[4] P. Arbelaez, J. Pont-Tuset, J. T. Barron, F. Marques, and J. Malik. Multiscale Combinatorial Grouping. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 500:328–335, 2014. 2

[5] F. L. Bookstein. Principal Warps: Thin-Plate Splines and the Decomposition of Deformations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(6):567–585, 6 1989. 3, 6, 8

[6] H. Bristow, J. Valmadre, and S. Lucey. Dense semantic correspondence where every pixel is a classifier. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4024–4031, 2016. 2

[7] X. Chen, R. Mottaghi, X. Liu, S. Fidler, R. Urtasun, and A. Yuille. Detect What You Can: Detecting and Representing Objects using Holistic Models and Body Parts. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1971–1978, 2014. 6

[8] 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. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1201–1210, 2015. 2

[9] M. Cho, J. Lee, and K. Lee. Reweighted random walks for graph matching. *Proceedings of the European Conference on Computer Vision*, pages 492–505, 2010. 2

[10] M. Cho, J. Sun, O. Duchenne, and J. Ponce. Finding matches in a haystack: A max-pooling strategy for graph matching in the presence of outliers. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2091–2098, 2014. 2

[11] C. B. Choy, J. Gawk, S. Savarese, and M. Chandraker. Universal Correspondence Network. *Advances in Neural Information Processing Systems*, pages 2406–2414, 2016. 3

[12] N. Dalal and W. Triggs. Histograms of Oriented Gradients for Human Detection. *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1(3):886–893, 2004. 1

[13] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 3

[14] G. Donato and S. Belongie. Approximate Thin Plate Spline Mapping. In *Computer Vision*, pages 21–31. Springer Berlin Heidelberg, 2002. 3, 6, 8

[15] O. Duchenne, A. Joulin, and J. Ponce. A Graph-Matching Kernel for Object Categorization. *IEEE International Conference on Computer Vision*, pages 1792–1799, 2011. 7

[16] A. Eigenstetter, M. Takami, and B. Ommer. Randomized Max-Margin Compositions for Visual Recognition. *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3590–3597, 6 2014. 2

[17] P. Felzenszwalb, P. Felzenszwalb, D. Huttenlocher, and D. Huttenlocher. Belief Propagation for Early Vision. *International journal of computer vision*, 70:41–54, 2006. 6

[18] P. Fischer, A. Dosovitskiy, and T. Brox. Descriptor Matching with Convolutional Neural Networks: a Comparison to SIFT. *arXiv:1405.5769*, 5 2014. 3

[19] R. Girshick, F. Iandola, T. Darrell, and J. Malik. Deformable Part Models are Convolutional Neural Networks. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 437–446, 9 2014. 1, 2, 3

[20] B. Ham, M. Cho, C. Schmid, and J. Ponce. Proposal Flow. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3475–3484, 2016. 1, 2, 3, 4, 5, 6, 7, 8

[21] M. J. Hannah. Computer Matching of Areas in Stereo Images. Technical report, 1974. 1

[22] B. Hariharan, J. Malik, and D. Ramanan. Discriminative Decorrelation for Clustering and Classification. 1:1–14. 1

[23] B. Horn and B. Schunck. Determining optical flow. *Artificial Intelligence*, 17(1-2):185–203, 1981. 1

[24] J. Hosang, R. Benenson, P. Dollar, and B. Schiele. What Makes for Effective Detection Proposals? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(4):814–830, 2 2016. 2

[25] A. Kanazawa, D. W. Jacobs, and M. Chandraker. WarpNet: Weakly Supervised Matching for Single-view Reconstruction. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3253–3261, 2016. 3

[26] J. H. Kappes, T. Beier, and C. Schnoerr. MAP-Inference on Large Scale Higher-Order Discrete Graphical Models by Fusion Moves. *European Conference on Computer Vision Workshop*, pages 469–484, 2014. 6

[27] I. Kemelmacher-Shlizerman and S. M. Seitz. Collection flow. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1792–1799, 2012. 6

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

[29] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, pages 1–9, 2012. 3

[30] E. G. Learned-Miller. Data driven image models through continuous joint alignment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(2):236–250, 2006. 6

[31] C. Liu, J. Yuen, and A. Torralba. Sift flow: Dense correspondence across scenes and its applications. *IEEE transactions on pattern analysis and machine intelligence*, 33(5):978–994, 2011. 2

[32] Z. Liu, P. Monasse, and R. Marlet. Match Selection and Refinement for Highly Accurate Two-View Structure from Motion. *European Conference on Computer Vision*, pages 818–833, 2014. 7, 8

[33] J. Long, N. Zhang, and T. Darrell. Do Convnets Learn Correspondence? *Advances in Neural Information Processing Systems*, pages 1601–1609, 11 2014. 1, 3, 7

[34] D. G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60(2):91–110, 11 2004. 1

[35] B. D. Lucas and T. Kanade. An Iterative Image Registration Technique with an Application to Stereo Vision. *Imaging*, 130:674–679, 1981. 1

[36] S. Manen, M. Guillaumin, and L. Van Gool. Object Proposals with Randomized Prim's Algorithm. *IEEE International Conference on Computer Vision*, pages 2536–2543, 2013. 2

[37] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust Wide Baseline Stereo from Extremal, Maximally Stable. *In British Machine Vision Conference*, pages 384–393, 2002. 2

[38] Y. Peng, A. Ganesh, J. Wright, and Y. Ma. RASL Robust Alignment by Sparse and Low-Rank Decomposition for Linearly Correlated Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11):2233–2246, 2012. 6

[39] J. Revaud, P. Weinzaepfel, Z. Harchaoui, and C. Schmid. DeepMatching: Hierarchical Deformable Dense Matching. *International Journal of Computer Vision*, 120(3):1–24, 2016. 2

[40] J. C. Rubio, A. Eigenstetter, and B. Ommer. Generative regularization with latent topics for discriminative object recognition. *Pattern Recognition*, 48(12):3871–3880, 2015. 2

[41] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423, 1948. 3

[42] S. Singh, A. Gupta, and A. A. Efros. Unsupervised discovery of mid-level discriminative patches. *Proceedings of the European Conference on Computer Vision*, pages 73–86, 2012. 2

[43] L. Torresani, V. Kolmogorov, and C. Rother. A dual decomposition approach to feature correspondence. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35:259–271, 2013. 2, 4

[44] T. Tuytelaars and K. Mikolajczyk. Local Invariant Feature Detectors: A Survey. *Computer Graphics and Vision*, 3(3):177–280, 2008. 4

[45] J. 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. 2, 5, 7

[46] Y. Xiang, R. Mottaghi, and S. Savarese. Beyond pascal: A benchmark for 3d object detection in the wild. *IEEE Winter Conference on Applications of Computer Vision*, pages 75–82, 2014. 6

[47] Y. Yang and D. Ramanan. Articulated Human Detection with Flexible Mixtures-of-Parts. *IEEE transactions on pattern analysis and machine intelligence*, pages 1–15, 2012. 7

[48] T. Zhou and A. A. Efros. Learning Dense Correspondences via 3D-guided Cycle Consistency. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 117–126, 2016. 3

[49] T. Zhou, Y. J. Lee, S. Yu, and A. Efros. Flowweb: Joint image set alignment by weaving consistent pixel-wise correspondences. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1191–1200, 2015. 6

[50] X. Zhou, M. Zhu, and K. Daniilidis. Multi-Image Matching via Fast Alternating Minimization. *Proceedings of the IEEE International Conference on Computer Vision*, pages 4032–4040, 2015. 2

[51] C. L. Zitnick and P. Dollár. Edge boxes: Locating object proposals from edges. *Proceedings of the IEEE International Conference on Computer Vision*, pages 391–405, 2014. 2, 7