IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received November 30, 2019, accepted January 9, 2020, date of publication January 14, 2020, date of current version January 22, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2966655

Deep Semantic Feature Matching Using Confidential Correspondence Consistency

WEI LYU, LANG CHEN, ZHONG ZHOU, (Member, IEEE), AND WEI WU
State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing 100191, China

Corresponding author: Zhong Zhou (zz@buaa.edu.cn)

This work was supported in part by the National Key Research and Development Program of China under Grant 2018YFB2100601, and in part by the National Natural Science Foundation of China under Grant 61872024 and Grant 61872023.

ABSTRACT This work aims to establish visual correspondences between a pair of images depicting objects of the same semantic category. It encounters many challenges such as non-overlapping of scenes or objects, background clutter, and large intra-class variation. Existing methods handle this task with handcrafted features, which cannot effectively fit the correlations between non-overlapping images. Besides, additional training or information may be implemented into the learned features. In this paper, we propose a novel approach for semantic correspondence, which is based on deep feature representation, geometric and semantic associations between intra-class objects, and hierarchical matching selection according to the convolutional feature pyramid. Firstly, we construct the initial correspondence by developing a sparse feature matching model on the coarsest feature level, which enforces the nearest-neighbor searching under semantic and geometric consistency constraints. Further, a narrowing strategy is proposed and employed from the coarsest to the finest feature level, which hierarchically refine and optimize the correspondence. The results illustrate that this approach achieves competitive performance on the public datasets for semantic correspondence.

INDEX TERMS Feature matching, consistency constraints, nearest-neighbor searching, hierarchical optimization, convolution feature pyramid.

I. INTRODUCTION

Establishing correspondences between images is one of the fundamental problems in computer vision and graphics. Early works are concerned with calculating the correlations among multiple overlapping images (instance-level correspondence), such as image stitching [1], 3D reconstruction [2], [3], and stereo matching [4]. They assume that the input image pair shows a proportion of the same scenes or objects from different viewpoints, and correspondences are obtained by using the handcrafted feature descriptors, e.g., Scale-Invariant Feature Transform (SIFT) [10] and Speeded Up Robust Features (SURF) [11]. In the latest years, the semantic correspondence exploration [5]–[7], i.e., intra-class semantic object matching, has been developed (category-level correspondence), which is widely applied in various fields such as object recognition [8] and re-identification [9]. It establishes the correlations between different objects of the same semantic category. This is a challenging task due to non-overlapping, background clutter, intra-class variation, and difference in viewpoint.

Let us consider that both intra-class objects share similar semantic structure. In other words, extrinsic similar geometry and intrinsic semantic association between objects are available for semantic correspondence. The key issue is how to utilize the finite associated information to select enough salient features for matching. Traditional handcrafted features, such as SIFT [10] and SURF [11], work well on matching the overlapping images. But they are not suitable for the category-level correspondence since they are mainly designed for the same objects or scenes. Besides, the correlations between images can also be obtained by searching for the keypoint pair whose neighbors have similar displacement vector [12]. Some methods use the matching constraint to minimize the appearance matching cost and to preserve the geometric consistency [21], [22]. Meanwhile, Graph cut [15], random search [16], [17], and hierarchical optimization [13] are also used to further improve the correspondence. With the

The associate editor coordinating the review of this manuscript and approving it for publication was Zhenhua Guo.

development of strong representation ability of the Convolutional Neural Networks (CNNs), a number of works employ the CNN feature representations and show that the CNN features are more flexible than handcrafted features, and more effective for nearest-neighbor matching [18], [19]. Remarkably, since the shallower layer of a CNNs model focuses on low-level image features like color and texture whereas the deeper layer on high-level semantic information, how to effectively select and utilize CNN features is still an open question.

In this work, we build a novel coarse-to-fine framework for deep semantic feature matching, which fully utilizes the deep feature representations, semantic and geometric information consistencies, and hierarchical mapping relationships. Given two input images containing objects of the same semantic category, we first extract the feature representations based on a pre-trained classification CNN model [29] and generate a five-layer feature map pyramid for each image. The top layer, which is defined as the coarset layer, holds semantic information covering object-specific attribute representation and the bottom layer defined as the finest layer describes low-level contexts including color and texture. Then, we develop a sparse feature matching mechanism to produce the initial feature correspondences based on the top-layer representation in the pyramid. Analogously to traditional feature descriptor and matching [10], the best candidate match for each feature representation can be established by identifying its nearset neighbor in the database of features from the original images. We select as many representational and salient features as possible and establish inter-image correspondences based on semantic and geometric consistencies via the nearest-neighbor searching.

To ensure approximate accuracy, it enforces the nearest-neighbor searching under confidential correspondence consistency constraints which encode the geometric and semantic associations between intra-class objects. Furthermore, since the feature point mapping from the top pyramid layers to the bottoms ones has position offsets, directly projecting the correspondence from a high-layer pyramid, which describe correspondence between paired feature representations from different images, to a low-layer one produces inaccurate low-level feature point correspondences and may drop a lot of salient feature information. To overcome the misaligned mapping from the top layer to the bottom one and improve the correspondence, we introduce a hierarchical optimization strategy, which is designed to re-target the corresponding keypoints across layers. We directly map the corresponding points of the higher layer to the lower layer and treat the center of the mapped patches at the lower pyramid layer as the candidate corresponding keypoints. It searches for the mutual nearest neighbors of the candidate keypoints to ensure each potential salient position. Meanwhile, we introduce a narrowing scheme to further improve the correspondence at each pyramid layer. An imitation foreground detection method is adopted to reduce the non-salient features and remove mismatches. Our approach not only ensures the robustness of the intra-class variation between objects, but also accurately locates sparse feature correspondences. The contributions of our work are mainly three-fold:

- We integrate high-level semantic information and object-level geometric information of the images into candidate generation, and establish correspondences with semantic contents of the objects.
- We introduce a simple yet effective narrowing strategy, i.e., the imitation foreground detection method, for feature selection to improve the exclusivity among candidates. It can reduce the searching scope at the lower pyramid layer and mitigate the error accumulation during hierarchical optimization.
- Experiments illustrate that the proposed approach obtains competitive performance on standard benchmarks for semantic correspondence.

II. RELATED WORK

Semantic correspondence has gained rising attention in the last years. Especially, more and more works are concerned with deep feature matching, and continue to make new advances.

The first version establishes correspondences using the handcrafted features which is based on semantic flow [13]. It constructs a hierarchical optimization structure to solve the displacement vectors of discrete pixel-points. The main idea is to introduce a matching constraint to minimize the appearance matching cost. To perform more effective matching, Kim *et al.* [14] design a spatial pyramid matching model. They regularize the corresponding consistency from an entire image, to meshes, to each pixel rather than only pixel-based in SIFT-Flow [13], and enable faster dense matching. Besides, Zhou *et al.* [21] improve both feature affinities and cycle consistency of the correspondence by solving a low-rank matrix recovery problem [23]. Wang *et al.* [22] add another matching constraint to preserve the geometric consistency. These methods obtain effective correspondences based on the notion of matching SIFT [10], even better than most of the work on semantic correspondence. However, accurate initial inputs are important.

Recently, some works focus on deep semantic feature matching due to the breakthrough of deep neural networks. These works are generally divided into two categories: end-to-end alignment methods and post-processing based methods. The former utilizes the powerful information mining and fitting capabilities of deep neural networks, and also inherits its restrictions such as strongly dependence on manual annotations and additional training. SCNet is presented by Han *et al.* [42], which utilizes the CNNs to learn a geometrically plausible model for semantic correspondence. It incorporates the geometric consistency and uses region proposals as matching primitives. Rocco *et al.* [24] train their neural network architecture for geometric matching in a supervised manner to mimic the traditional matching [10]. To avoid the manual annotations, they further develop a weakly supervised

matching network [7] by adopting a scoring strategy of inliers [25] as loss function. Besides, Kim *et al.* [44] construct a self-similarity CNN feature descriptor. They leverage object candidate priors provided in selected datasets and combine a matching consistency to mitigate the restrictions, with enabling a weakly-supervised training. It is further improved by introducing a discrete local labeling optimization [45], with estimating dense affine transformation model between semantically similar images. Besides, they also train a pyramidal affine regression networks to estimate locally-varying transformation field across images [46]. These methods are concerned with designing complicated network architectures to calculate the parameters of transformation model between objects or scenes for alignment, and obtain superior performance for semantic correspondence with a certain degree of supervision, such as additional prior information or manual annotations for their learning procedure.

In contrast to the end-to-end alignment methods, the latter aims to establish the correspondences by employing the CNN feature representations and constructing an effective post-processing model. It generally adopts pre-trained classification networks to extract image features without any annotations or additional training. Ufer and Ommer [5] design a complicated matching system based on optical flow [12]. CNN features are extracted from images by using pre-trained AlexNet model [26] and a convolutional pyramid model is formulated [20], with each input image corresponding to a Gaussian image pyramid. The Gaussian image at each image pyramid layer is used to generate a feature pyramid. Furthermore, a cross-domain correspondence method (NBB) is proposed by Kfir *et al.* [27]. It utilizes the notion of Deep Image Analogy [28], [43] to find the correspondences between main objects of interest belonging to different semantic categories in different images, while sharing similar geometry. A hierarchical matching method is adopted to search for candidate correspondences according to the constructed convolutional feature pyramid. It follows a simple matching rule, which directly measures the similarity between CNN feature descriptor, resulting in lower robustness. Error accumulation is caused by its feature selection scheme and mismatches are ignored at each pyramid layer. Besides, geometric information may be lost since only semantic association between objects are used for their matching.

To mitigate these, we adopt the notion of hierarchical optimization method and propose a narrowing scheme to improve the method. An imitation foreground detection method is adopted to improve the correspondences at each pyramid layer. It rejects the outliers and focuses the correspondences on objects. Thus unnecessary mappings are avoided, and the search scope is reduced at the lower pyramid layer. It mitigates background clutter and reduces computational complexity. The learned CNN features are directly used without constructing additional descriptors. Meanwhile, geometric and semantic associations between intra-class objects are encoded to enforce the nearest-neighbor searching. The proposed method is implemented without any training or additional annotations in our experiment.

III. PROPOSED APPROACH

This section describes the proposed framework for semantic correspondence in detail. Given two images depicting objects of the same semantic category, we first extract the feature representations based on a pre-trained CNNs model and produce a five-layer feature map pyramid for each image. And then a salient feature selection scheme is employed to filter out most of the non-salient features. Furthermore, a coarse-to-fine matching process is performed. Firstly, a sparse feature matching mechanism is introduced to produce the initial correspondences based on the top-level representation in the pyramid. It encodes both the geometric and semantic associations between intra-class objects as confidential correspondence consistency constraints, which is formulated as minimizing an objective function consisting of semantic consistency term, distance consistency term, and orientation consistency term. To further improve the correspondences, we introduce a hierarchical optimization strategy, which is designed to re-target the corresponding feature points across layers. We directly map the corresponding points of the higher layer to the lower layer and treat the center of the mapped patches at the lower pyramid layer as the candidate corresponding keypoints. Then the mutual nearest-neighbors of the candidate pairs are found by using sliding windows in the corresponding patch pair at each pyramid layer, and we adopt an imitation foreground detection method, consisting of a dynamic threshold selection scheme and an outliers rejection model [25], to further improve the hierarchical optimization process. Finally, the resulting correspondences are obtained at the bottom pyramid layer, as shown in Figure. 1.

A. PRETREATMENT

In this section we use a pre-trained CNN classification network to extract multi-level features and produce a five-layer feature map pyramid. Then a salient feature selection process is performed to filter out most of the non-salient features by using a regularization algorithm and a threshold selection scheme. These operations ensure the matching in subsequent section.

1) DEEP CONVOLUTIONAL FEATURE PYRAMID

Below, we use a standard CNN architecture for feature extraction. It generally extracts discriminative image features through multiple convolutional layers, and generates the corresponding feature maps with different scales by combing activation layer and pooling layer.

Given an image pair (I^i, I^j), they are fed forward through the VGG-19 model [29] pre-trained on ImageNet [30]. We extract the CNN features and produce a five-layer feature map pyramid ($L = 1, 2, 3, 4, 5$) with scaling factor $1/2$. Specifically, the feature map F^L is extracted from the Relu$L$_1 layer of VGG-19. The feature maps (F_i^L, F_j^L) for two images are $h_L \times w_L \times d_L$ tensors, which are

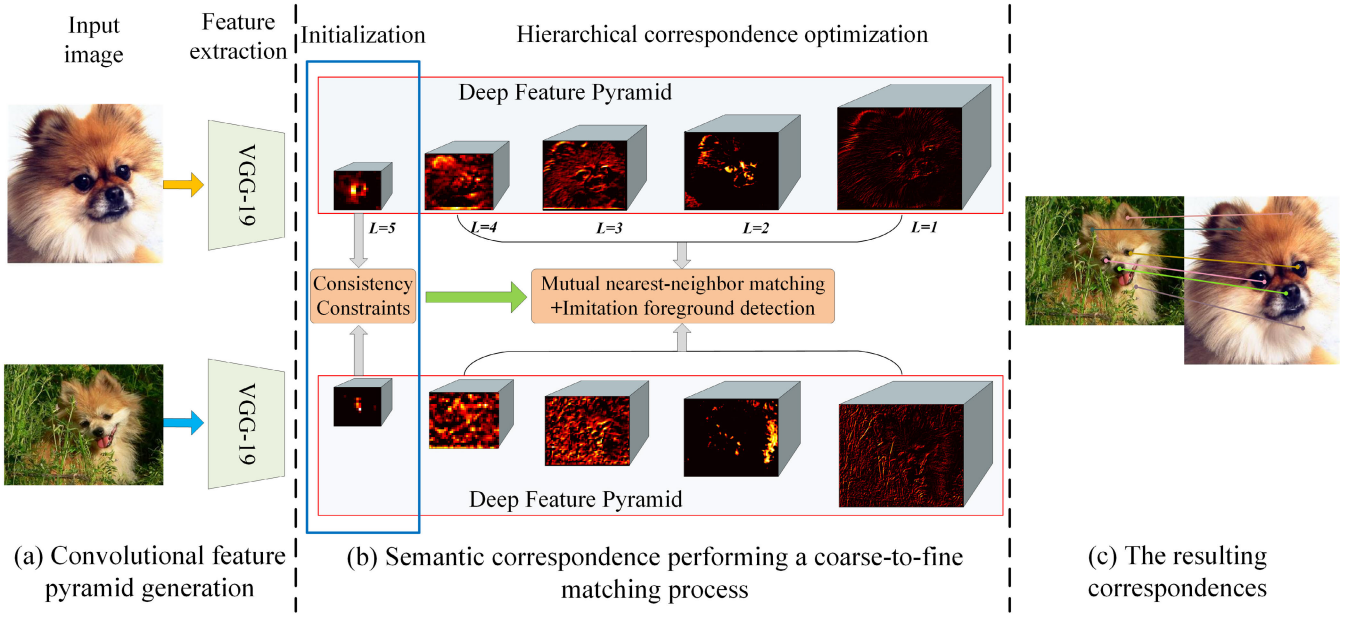


FIGURE 1. Overview of our approach. Given two images depicting the main objects of the same semantic category, firstly a pre-trained VGG-19 Network model [29] is used for feature extraction and to produce a five-level convolutional feature pyramid as shown in (a). Then we propose to perform a coarse-to-fine matching process to obtain accurate correspondences in (b). Initial correspondence is formulated as minimizing an objective function consisting of three consistency constraints, which encodes geometric and semantic associations between intra-class objects, shown as blue block in part (b). Then we explore a hierarchical correspondence optimization strategy to improve the correspondence from the top pyramid layers to the bottom ones. It performs a mutual nearest-neighbor matching model with additional imitation foreground detection at each pyramid layer. Then the resulting correspondences are produced at the bottom pyramid layer. Specifically, the orientation of the pyramid structure given in (b) is the inverse of the convolutional feature pyramid generated in (a).

denoted as dense $h \times w$ grids of d-dimension CNN features: $F^L \in \mathbb{R}^{h_L \times w_L \times d_L}$.

2) SALIENT FEATURE SELECTION

One of the challenges of semantic feature matching is the background clutter, resulting in incorrectly matching certain regions of the background to semantic features. To effectively match the salient semantic features, a salient feature selection scheme is employed to filter out most of the non-salient features at each pyramid layer. In other word, the salient features within the foreground semantic objects are maintained, while features in the background are rejected. Generally, the VGG-19 model is used for object recognition and classification. It can effectively locate the semantic objects and produce different feature maps or feature representations, which is very beneficial for semantic feature matching. Specifically, this process is performed for all the feature presentations at the top pyramid layer, and only for the corresponding features within the candidate patches at the other pyramid layer.

Firstly the extracted image features are taken as the input to a filter that performs a salient feature selection process at the top pyramid layer. We utilize the notion of the keypoint localization of SIFT [10] and construct a filter to select the salient features by comparing a point to its neighbors in the corresponding 3×3 region at the current scale. The robustness of matching can be first enhanced by integrating the regularization theory. The discrete CNN feature F^L with $L = 5$, which corresponds to the feature maps at the top pyramid layer, is regularized to a normalized interval (i.e. [0, 1]). Then we adopt a min-max normalization function to measure the saliency of each feature point, which is denoted as

$$s^L(p) = \frac{\left\|F^L(p)\right\| - min\left\|F^L\right\|}{max\left\|F^L\right\| - min\left\|F^L\right\|} \tag{1}$$

where $\|\cdot\|$ represents the L_2 norm and $s^L(p)$ represents the probability score of being salient feature at position p at the current pyramid layer L. $max\left\|F^L\right\|$ and $min\left\|F^L\right\|$ are the maximal and minimal normalization value of all feature points on the tensor feature map F^L, respectively. Then we perform a salient selection process to filter out the low-response points and preserve the salient features, which is defined as

$$1\_s^L(p) := \begin{cases} 1 \ if \ s(p) > \tau^L \\ 0 \ otherwise \end{cases} \tag{2}$$

where τ^L is a threshold for layer L. 1_s^L is an indicator matrix and $1\_s^L \in \mathbb{R}^{h_L \times w_L}$. The new tensor feature map $\widetilde{F}^L$ is generated by preserving the feature points whose indicator values equal to 1 and assigning others with 0. It contains all regularized salient features from the input image.

B. INITIALIZATION FOR MATCHING

The goal of initialization stage is to obtain as many correspondences as possible while ensuring approximate accuracy. One of the challenges is that there are no same scenes or

objects between images. To avoid matching those in compatible regions, we enforce the nearest-neighbor searching under confidential correspondence consistency constraints which encode the geometric and semantic associations between intra-class objects. Meanwhile, the search scope is constrained within the main objects of interest. Inspired by [5], [12], we develop a simple and straightforward matching model at the top pyramid layer, which has high response on objects and suppresses irrelevant backgrounds, to formulate the initial correspondences of selected salient features.

In the matching model, exploring the initial correspondences is formulated as minimizing an objective function consisting of three constraint terms, i.e., the semantic consistency constraint ensuring correspondences with similar appearance, the distance consistency constraint and orientation consistency constraint enforcing them with similar geometry. The consistency constraints are defined in detail in the following.

1) SEMANTIC CONSTRAINT

Generally, the description of the appearance is an important and fundamental feature representation of an standard RGB image. A semantic consistency constraint is introduced to ensure the matched candidates sharing similar semantic appearance. Given a feature pair (p, q) extracted from the corresponding image pair (I_i, I_j), a cosine similarity function, $sim(\cdot)$, is adopted to measure the semantic distance between them. This term is defined as

$$E_S(V) = \sum_{p,q} e^{(1-sim(f_i^L(p),f_j^L(q)))^2/\sigma_S^2} - 1 \quad (3)$$

$$sim(f_i^L(p), f_j^L(q)) = \frac{f_i^L(p) \cdot (f_j^L(p))^T}{\|f_i^L(p)\| * \|f_j^L(p)\|} \quad (4)$$

where σ_S is a constant factor, and $\boldsymbol{f}_i^L(p)$ is the normalized feature descriptor in image I_i. To reduce the computational complexity, it searches for K nearest neighbors as candidates in image I_j for each keypoint in another image I_i. Note that K is set to 5 in our experiment.

2) DISTANCE CONSTRAINT

Given a pair of salient feature points for each object, we consider that there is relative positional consistency between the two pairs. For example, different saloon cars share similar relative position between the front wheel and headlight, as shown in Figure. 2. To effectively utilize this characteristic, the difference between the relative positions of feature pairs from two images, which is equivalent to the relative Euclidean distance, is calculated. The smaller difference indicates the larger the matching probability. Then we introduce a distance constraint to enforce distance consistency by minimizing the difference, which is defined as

$$E_D(V) = \sum_{p,p',q,q'} e^{d^2(p,p',q,q')/\sigma_D^2} - 1 \quad (5)$$

FIGURE 2. The geometric correlations between the positions of salient features. (p, q), (p', q') are two keypoint pairs and the relative distance and orientation between directed line vector pp' and qq' are all intended to be consistency.

where (p, p'), (q, q') represent the point pairs selected from the images i and j separately and σ_D is a constant factor. $d(p, p', q, q')$ encodes the difference between the relative distances $\overrightarrow{pp'}$ and $\overrightarrow{qq'}$ as

$$d(p, p', q, q') = abs\left(\frac{\left|\overrightarrow{pp'}\right|}{B_i^{diag}} - \frac{\left|\overrightarrow{qq'}\right|}{B_j^{diag}}\right) \quad (6)$$

where $|\cdot|$ is the module of the corresponding line vector and B is the diagonal distance of the bounding box around object.

3) ORIENTATION CONSTRAINT

Similarly, each pair of feature points selected from an image can be connected by a straight line. The orientations of intra-lines from two images should be consistency as shown in Figure. 2. Thus an orientation consistency constraint is introduced and defined as

$$E_O(V) = \sum_{p,p',q,q'} e^{r^2(p,p',q,q')/\sigma_O^2} - 1 \quad (7)$$

where $r(\cdot)$ is denoted as an inverse cosine function, which measures the angle of intersection of the intra-object lines, and defined as

$$r(p, p', q, q') = arc\cos\left(\frac{\overrightarrow{pp'} * \overrightarrow{qq'}}{\left|\overrightarrow{pp'}\right| * \left|\overrightarrow{qq'}\right|}\right) \quad (8)$$

4) OBJECTIVE FUNCTION

The essence of matching is to search for two candidate feature points with minimal consistency. We formulate this task as minimizing an objective function, which is defined as

$$E(V) = E_S(V) + \lambda_D E_D(V) + \lambda_O E_O(V) \quad (9)$$

where λ_D and λ_O are the weights of the distance constraint and orientation constraint separately, $E_S(V)$ constrains the appearance similarity, and $E_D(V)$ and $E_O(V)$ enforce geometric consistency. Finally, $E(V)$ is iteratively solved to produce the initial correspondences V.

5) DISCUSSION

Different from [5], [12], our proposed approach has several advantages: (a) It requires a general and simpler convolutional feature pyramid, and has more robust salient feature

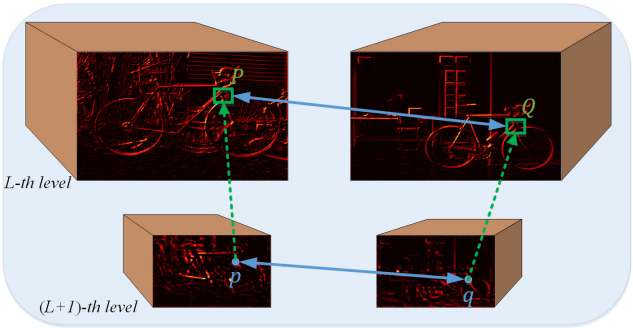


FIGURE 3. The mapping relationship between point and patch at the neighboring pyramid layers. (p, q) is a matched keypoint pair at the ($L+1$)-th pyramid layer, which corresponds to a patch pair (P, Q) at the L-th pyramid layer according to the receptive field.

selection scheme motivated by SIFT [10]. (b) It adopts a nearest-neighbor searching scheme during the matching step, which obeys an error tolerance mechanism. It only searches for K nearest neighbors, without additional penalty, to obtain as many correspondences as possible while ensuring approximate accuracy. (c) The keypoint selection scheme is based on geometric information such as edge rather than object proposal.

C. OPTIMIZATION

Since the points at the top pyramid layer have a large receptive field to the input image, the initial correspondences describe relationships between patches. To propagate the patch-level correspondence to pixel-level, we develop a novel hierarchical mapping scheme to gradually mapping the initial correspondences from the top layer to the bottom layer. Firstly, each point at the higher pyramid layer is projected to a patch at the lower layer. A point-to-patch mapping process is performed to transfer the point-level correspondences at the higher layer to the patch-level correspondences at the lower layer. Then the descriptors contained in the patches are normalized by using a revised Z-Score normalization method. And then a mutual nearest-neighbor matching model is introduced to search for the potential correspondences in each candidate patch pair at each pyramid layer. Finally we adopt a simple yet effective scheme, which imitates the foreground detection, to mitigate the background clutter. The details are follows.

1) POINT-TO-PATCH MAPPING

For a set of matched keypoint pairs, $V^{L+1}=\{(p_m^{L+1}, q_m^{L+1})\}_{m=1}^{M}$ ($L=1, 2, 3, 4$), at the $(L+1)$-th pyramid layer, a point-to-patch mapping process $V^{L+1} \rightarrow U^L$ aims to generate a set of corresponding patch pairs, $U^L=\{(P_m^L, Q_m^L)\}_{m=1}^{M}$, at the L-th pyramid layer. As shown in Figure. 3, point p is inversely mapped to patch P according to the coordinate mapping of the VGG-19 model [29], i.e., up-sampling in double. For each matched keypoint pair (p_m^{L+1}, q_m^{L+1}), the point $p_m^{L+1}(x, y)$ (or $q_m^{L+1}(x, y)$) is mapped to the center point p'^L_m (or q'^L_m) of a patch P_m^L (or Q_m^L), in which x and y are the coordinates of point p_m^{L+1}. The mapped patch P_m^L is denoted as

$$P_m^L=\left[p'^L_{m\_x}-\zeta, p'^L_{m\_x}+\zeta\right]\times\left[p'^L_{m\_y}-\zeta, p'^L_{m\_y}+\zeta\right] \tag{10}$$

where $(p'^L_{m\_x}, p'^L_{m\_y})$ is the two-dimensional coordinate of point p'^L_m with corresponding to $(2x, 2y)$, Q_m^L is represented the same as P_m^L. ζ is a constant and set to 2 in our experiment. Q_m^L and P_m^L are then formulated as a matched patch pair in U^L.

2) DESCRIPTOR NORMALIZATION

To eliminate obvious luminance and color differences before further matching, we adopt the Z-Score normalization method with introducing additional parameters [27], [33], [34], which is mainly used for style transferring. It locally normalizes the feature descriptors and globally balances the differences in luminance and color. For the selected keypoint pairs in each matched patch pair (P^L, Q^L), corresponds to the feature maps (F_i^L, F_j^L), which are normalized as

$$N_i^L(p^L)=\mu(P^L, Q^L)\cdot\frac{F_i^L(p^L)-\mu(P^L)}{\sigma(P^L)}+\sigma(P^L, Q^L) \tag{11}$$

where

$$\begin{aligned}\mu(P^L, Q^L)&=\frac{\mu(P^L)+\mu(Q^L)}{2}\\ \sigma(P^L, Q^L)&=\frac{\sigma(P^L)+\sigma(Q^L)}{2}\end{aligned} \tag{12}$$

where $\mu(\cdot) \in R^d$, $\sigma(\cdot) \in R^d$ are the corresponding spatial mean and standard deviation respectively and $N_j^L(q^L)$ is defined similarly.

3) MUTUAL NEAREST-NEIGHBOR MATCHING

Given the matched patch pairs $U^L=\{(P_m^L, Q_m^L)\}_{m=1}^{M}$ for the image pair (I_i, I_j) and normalized feature representations (N_i, N_j) for keypoints in the patches, a mutual nearest-neighbor matching process, i.e., matching from I_i to I_j and that inversely from I_j to I_i, is implemented at the L-pyramid layer for both images. It seeks to locally explore a set of corresponding keypoints within each matched patches, with the assistance of a number of neighbouring keypoints, and produce the resulting keypoint correspondences $S^L=\{(p_n^L, q_n^L)\}_{n=1}^{N^L}$.

Firstly, we compute the local keypoint correspondences in a matched patch pair (P_m^L, Q_m^L) holding a pair of candidate keypoints $(p^L \in P_m^L, q^L \in Q_m^L)$. For correspondences from P_m^L to Q_m^L taken as the example in the following steps, the correspondences $S_{P_m^L \rightarrow Q_m^L}(p_n^L)$ is evaluated with the keypoint similarity as follows,

$$Corr(p^L)=argmax\ \widehat{SIM}(p^L, q^L, P_m^L, Q_m^L), \tag{13}$$

where $\widehat{SIM}(\cdot)$ is the similarity metric function measured by introducing a weighted nearest-neighbor metric scheme. It is defined as

$$\widehat{SIM}(p, q, P, Q)=\sum_{p''\in P, q''\in Q} w\cdot\frac{N_i(p'')\cdot N_j(q'')}{\|N_i(p'')\|\cdot\|N_j(q'')\|}, \tag{14}$$

where p'' and q'' represent the neighbors of the keypoints p_n^L and q_n^L, separately. $N_i(p'')$ is the normalized descriptor of p''. The weight w is defined as $w = w_i \cdot w_j$ with $w_i = 1/d^2(p, p'')$ and $w_j = 1/d^2(q, q'')$. Here $d(\cdot)$ represents the Euclidean distance between two points. We restrict the neighbor region to 3×3 at the higher pyramid layers ($L = 3, 4$), and to 5×5 at the lower pyramid layers ($L = 1,2$).

The resulting correspondence $(p^L, Corr(P^L))$ is added to $S_{P_m^L \to Q_m^L}$. Note that $S_{P_m^L \to Q_m^L}$ contains all the correspondences with maximal similarity value. The correspondences $S_{Q_m^L \to P_m^L}(q_n^L)$ is established the same as $S_{P_m^L \to Q_m^L}$. Then the final correspondences S^L is obtained by performing the intersection of $S_{P_m^L \to Q_m^L}(p_n^L)$ and $S_{Q_m^L \to P_m^L}(q_n^L)$.

Furthermore, we develop an imitation foreground detection method for further optimization, which is implemented at each pyramid layer. Firstly a feature selection process is performed with selecting the dynamic threshold in Eq.(2) to adjust the rejection mechanism of salient features at the L-th pyramid layer according to the distribution of CNN features. Meanwhile, an outliers rejection scheme is used to further remove the mismatches [25]. A transformation model, e.g., homography, is first estimated according to a set of observed keypoints containing outliers. The corresponding parameters are calculated and a threshold is set to distinguish inliers and outliers. These approximately select the foreground objects which mitigates the background clutter, narrows the search scope at the lower pyramid layer, and reduces the error accumulation during hierarchical optimization. Finally, the resulting correspondences are obtained at the bottom pyramid layer. The algorithm flow is as shown in Algorithm 1.

4) DISCUSSION

We utilize the notion of hierarchical optimization [27] inspired by Liao *et al.* [43]. Specifically, they use a direct mapping scheme, in which each candidate is mapped to a patch at the lower pyramid layer and they only utilize the semantic information. It results in complicated calculation, error accumulation, and more mismatches. To mitigate these, we propose an imitation foreground detection method to improve the correspondences at each pyramid layer. It rejects the outliers and focuses the correspondences on objects. Thus unnecessary mappings are avoided, and the search scope is reduced at the lower pyramid layer. It mitigates the background clutter and reduces computational complexity. Meanwhile, geometric and semantic associations between intra-class objects are encoded to enforce the nearest-neighbor searching.

Algorithm 1 Deep Semantic Feature Matching

Require:
Two RGB images I_i, I_j;
Ensure:
A set of matched point pairs $V^1 = \{(p_n^1, q_n^1)\}_{n=1}^{N^1}$;
Pretreatment and Initialization:
1. Extract $\{F_i^L\}_{L=1}^5$ and $\{F_j^L\}_{L=1}^5$ using the VGG-19 network model pre-trained on ImageNet;
2. Generate a set of normalized features, S^*, solving Eq. (2);
3. Establish the initial correspondences $V^5 = \{(p_m^5, q_m^5)\}_{m=1}^M$ using Eq. (9).
Optimization:
for L = 4 to 1 **do**
1. Map keypoint pairs $\{(p_m^{L+1}, q_m^{L+1})\}_{m=1}^M$ to the patch pairs $\{(P_m^L, Q_m^L)\}_{m=1}^M$, using Eq. (10);
2. Calculate the normalized feature descriptors $N_i^L(p^L)$ and $N_j^L(q^L)$ using Eq. (11);
3. Estimate the correspondences $S_{P_m^L \to Q_m^L}(p_n^L)$ and $S_{Q_m^L \to P_m^L}(q_n^L)$, using Eq. (13);
4. Generate a set of correspondences, $S_{P_m^L \to Q_m^L}(p_n^L) \cap S_{Q_m^L \to P_m^L}(q_n^L)$, at the L-layer pyramid;
5. Produce the final correspondences, $\{(p_n^L, q_n^L)\}_{n=1}^{N^L}$, at the L-th pyramid layer by using an imitation foreground detection method.
end for

IV. EXPANSION

Objectively, most existing methods on semantic matching globally estimate a transformation model, typically homography, affine, or thin-plate spline transformation, to reject the outliers and align the objects. However, it is sensitive to images with large differences in viewpoints and ignores some salient details. Thus we introduce a local deformation method to improve it. Firstly the input image are divided into uniform grids, with each grid corresponding to an estimated homography. Then the outliers are eliminated according to the different grids. Furthermore, the input images are deformed and aligned. The homography estimation is as example in detail in the following.

A. HOMOGRAPHY TRANSFORMATION

Given a matched keypoint pair $X = [x, y]$ and $X' = [x', y']$, a homography transformation relationship between the keypoints is estimated as calculating an linear transformation with homogeneous coordinates, which is denoted as

$$\begin{bmatrix} x' \\ y' \\ z' \end{bmatrix} \sim H \begin{bmatrix} x \\ y \\ z \end{bmatrix} \tag{15}$$

where $\sim$ indicates equality up to a scale factor, H represents a 3×3 homography matrix with eight parameters in our experiment defined as

$$H = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix} \tag{16}$$

The mapping between X and X' is as

$$\begin{cases} x' = \dfrac{h_{11}x + h_{12}y + h_{13}}{h_{31}x + h_{32}y + 1} \\ y' = \dfrac{h_{21}x + h_{22}y + h_{23}}{h_{31}x + h_{32}y + 1} \end{cases} \tag{17}$$

where is linearised as

$$\begin{bmatrix} -x & -y & -1 & 0 & 0 & 0 & xx' & yx' & x' \\ 0 & 0 & 0 & -x & -y & -1 & xy' & yy' & y' \end{bmatrix} h = 0 \tag{18}$$

which can be written as

$$A \cdot h = 0 \tag{19}$$

where $h = [h_{11}, h_{12}, h_{13}, h_{21}, h_{22}, h_{23}, h_{31}, h_{32}, 1]$, it has 8 degree-of-freedom.

$$A = \begin{bmatrix} -x & -y & -1 & 0 & 0 & 0 & xx' & yx' & x' \\ 0 & 0 & 0 & -x & -y & -1 & xy' & yy' & y' \end{bmatrix} \tag{20}$$

Generally, four matched keypoint pairs are selected to estimate the eight parameters, and the input images can be globally deformed and aligned according to the estimated homography.

B. LOCAL WEIGHTED HOMOGRAPHY TRANSFORMATION

The input images are first divided into uniform grids. Then a local weighted homography transformation model is adopted to estimate the homography for each grid, which is denoted as

$$\begin{bmatrix} \widehat{x}' \\ \widehat{y}' \\ 1 \end{bmatrix} \sim \widehat{H} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \tag{21}$$

where $\widehat{X}_i = [\widehat{x}_i, \widehat{y}_i]^T$ is the center of each grid, $\widehat{H}$ is estimated as solving a weighted problem,

$$\widehat{h} = argmin \sum_{i=1}^{N} \|\omega_i A_i h\|^2 \quad s.t. \ \|h\| = 1 \tag{22}$$

where $\{\omega_i\}_{i=1}^N$ are the weights of assigning higher weights to the points that are closer to $\widehat{X}_i$, which is defined as

$$\omega_i = exp\left(\frac{-\|\widehat{X} - X_i\|^2}{\sigma^2}\right) \tag{23}$$

where σ is a scale factor. Please refer to [39] for further details.

V. IMPLEMENTATION AND EVALUATION

In this section, we evaluate the performance of the proposed approach on the publicly available benchmark datasets. Meanwhile, the implementation details, results, analyses, and comparisons to the state-of-the-art methods are provided in detail.

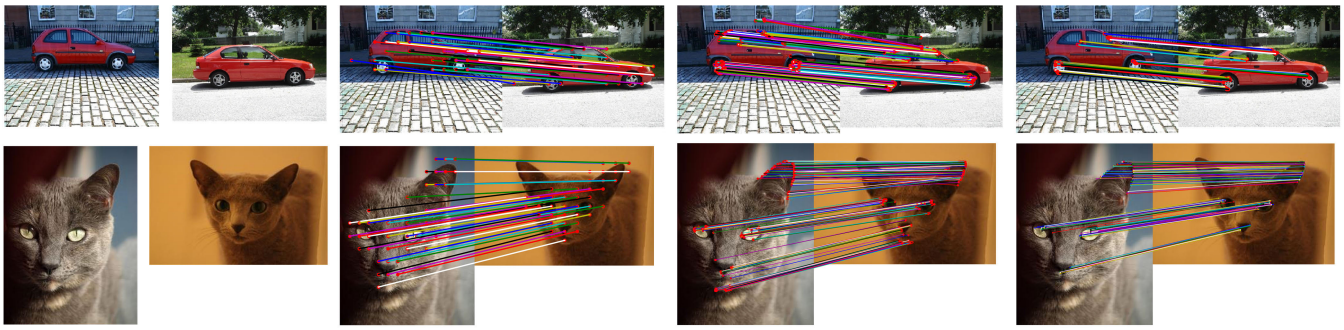
A. EVALUATION DATASETS AND PERFORMANCE MEASURE

Quantitative evaluation of our approach is implemented on two benchmark datasets: PF-PASCAL dataset and PF-WILLOW dataset which are the subsets of the Proposal Flow dataset [35]. The former contains 20 semantic categories with totaling approximately 1300 image pairs. The latter includes 4 semantic categories, which is divided into 10 subsets according to the background distribution and different viewpoints, for a total of approximately 900 image pairs for testing. Note that both provide the keypoint annotations on the semantic objects in the corresponding image pairs as ground truth for evaluation, and several matched keypoint pairs are assigned onto the salient positions of the corresponding object pairs. Generally, a percentage of correct key-points transfer (PCK) metric [36], [37] is adopted for the evaluation of our approach on both benchmark datasets. It is calculated by measuring the offsets between the ground-truth and the practical positions of transferring the matched keypoints. For a sparse set of correspondences between source and target images selected from datasets, the annotated keypoints are warped from source image to target image according to an estimated transformation. A correspondence is determined as inliers when the corresponding offset is less than $\theta \cdot max(H, W)$, where θ is the tolerance factor, and H and W is the height and width of the bounding box respectively, which is provided by datasets.

B. IMPLEMENTATION DETAIL

Essentially, the aim of the proposed approach is to establish the correspondences located at the salient postions and the key module is the constructed post-processor, it will not have much influcence on the results using the CNN models with different depths. We use a VGG-19 [29] model pre-trained on ImageNet [30] without fully connected layers. In the initialization stage, the objective function relies on the geometric and semantic associations between intra-class objects. It is iteratively solved for initial correspondence at the top pyramid layer. In our experiments, we empirically set $\lambda_D = 0.5$ and $\lambda_O = 0.5$, and σ_A, σ_D, σ_O are all set to 5. Note that the input images are resized to 224×224. For the generated convolutional feature pyramid, few salient features and correspondences are produced at the higher pyramid layer ($L = 4,5$), we lowered the quantitative criteria by selecting a smaller threshold. Meanwhile, the number of the candidate pairs is sufficient for matching at the lower pyramid layer ($L = 1, 2, 3$), a larger threshold was selected to reduce the computational complexity and negative matches. Thus a dynamic threshold feature selection scheme is adopted, and the corresponding threshold is empirically set to 0.3 at the L-th pyramid layer ($L = 4, 5$) and to 0.4 ($L = 1, 2, 3$). Finally the projective transformation model is estimated, according to the resulting correspondences, to deform and align the input images.

We evaluate our approach with three parts which is organized as follows. Section 4.3 evaluates the imitation foreground detection method on the PF-PASCAL dataset. Accuracy and robustness evaluation of the proposed approach are provided in Section 4.4. Based on the resulting correspondences, semantic alignment is introduced in Section 4.5.



(a) Source and Target Images (b) Semantic Correspondence Initialization (c) Fine Semantic Correspondence (d) Salient Semantic Correspondence

FIGURE 4. Pairwise semantic object matching using our approach.

TABLE 1. PSCK on the PF-PASCAL dataset.

Methods	aero	bike	bird	boat	bottle	bus	car	cat	chair	dog	mobike	tv
Original	0.44	0.59	0.48	0.27	0.42	0.35	0.70	0.56	0.45	0.39	0.44	0.56
RANSAC + Threshold	0.71	0.87	0.72	0.42	0.64	0.53	0.86	0.75	0.66	0.65	0.71	0.77

C. KEYPOINT SELECTION

For a set of correspondences, $U^L = \{(p_i^L, q_i^L)\}_{i=1}^{u^L}$, obtained after performing the nearest-neighbor matching at the L-th pyramid layer, a set of optimized correspondences, $M^L = \{(p_j^L, q_j^L)\}_{j=1}^{v^L}$, is produced by using an imitation foreground detection method, where $v \leq u$. In order to evaluate this method, we introduce a percentage of salient and correct key-points (PSCK) metric inspired by SIFT [10]. It is defined as PSCK $= v/u$ to determine the percentage of original and positive correspondences. Note that the larger PSCK indicates the better the correspondences. From Tab.1, it is clear that the proposed strategy effectively improves the correspondences, achieving an approximatively 25% improvement over original without using the method on the PF-PASCAL dataset [35]. Generally, we adopt this simple and effective strategy to improve the correspondences at each pyramid layer. It effectively rejects the outliers and focuses the correspondences on objects. Thus unnecessary mappings are avoided, and the search scope is reduced at the lower pyramid layer. Besides, It mitigates background clutter and reduces computational complexity.

Besides, some parameters are selected and proved in our experiments. The dynamic threshold feature selection scheme is adopted to further reject outliers at each pyramid layers, where the threshold is set to 0.3 at the L-th pyramid layer ($L = 4, 5$) and to 0.4 ($L = 1, 2, 3$). According to different thresholds (i.e., $\tau \in [0, 1]$), the corresponding tests on the PF-PASCAL and PF-WILLOW datasets are implemented, as shown in Figure. 5. The result shows that the mean PCK has a tendency to grow first and then flat, and finally decline. The smaller threshold leads to the weaker filter capacity of the matched feature pairs and the larger number of feature representations. It further results in producing more outliers and the matching tends to be redundant. Obviously, the saturation tends to be saturated at around 0.5, the number and

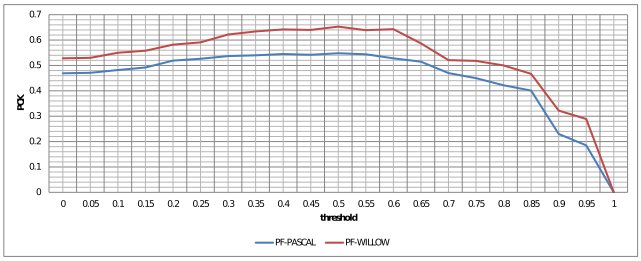


FIGURE 5. The test result of the selected thresholds on the PF-PASCAL and PF-WILLOW datasets.

accuracy of the matched feature pairs are balanced. The number of feature representations is less and less when the threshold continues to increase, and the matched feature pairs are unevenly distributed, which is not conducive to the performance evaluation. It shows that the trend of the mean PCK is relatively smooth in the range of 0.3 to 0.55, where the accuracy of the matching and the number of feature representations are relatively balanced in this interval. Let us consider that the scale of the feature representations at the higher pyramid layers is smaller ($L = 4, 5$), and the number of feature representations is very small. And the scale is larger and the candidates are more at the lower pyramid layer ($L = 1, 2, 3$). So we select the smaller threshold to search for more feature representations at the higher pyramid layer ($L = 4, 5$), and the larger threshold to reduce the redundancy of the feature representations at the lower pyramid layer ($L = 1, 2, 3$).

D. KEYPOINT MATCHING

Accuracy evaluation of the proposed method is implemented on the PF-PASCAL dataset, and the robustness evaluation is on the PF-WILLOW dataset containing more challenging examples with intra-class variation, background clutter,

TABLE 2. PCK ($\theta = 0.1$) on the PF-PASCAL dataset.

Methods	aero	bike	bird	boat	bot	bus	car	cat	chair	cow	horse	mobike	pers	plant	tv	*Avg.*
Deep Flow [37]	0.55	0.31	0.10	0.19	0.24	0.36	0.31	0.12	0.22	0.10	0.11	0.32	0.10	0.08	0.17	0.219
GMK [15]	0.61	0.49	0.15	0.21	0.29	0.47	0.52	0.14	0.23	0.23	0.13	0.39	0.12	0.16	0.22	0.291
SIFT Flow [13]	0.61	0.56	0.20	0.34	0.32	0.54	0.56	0.26	0.29	0.21	0.23	0.43	0.18	0.17	0.34	0.349
DSP [14]	0.64	0.56	0.17	0.27	0.38	0.51	0.55	0.20	0.23	0.24	0.23	0.41	0.15	0.11	0.28	0.329
Zhou et al. [38]	0.58	0.35	0.15	0.27	0.36	0.40	0.42	0.23	0.26	0.29	0.13	0.33	0.16	0.18	0.28	0.293
PF [35]	**0.75**	**0.76**	0.34	**0.41**	**0.55**	**0.71**	**0.73**	0.32	**0.41**	0.41	0.38	0.57	**0.29**	0.17	**0.46**	**0.484**
Liao et al. [43]	0.68	0.71	**0.35**	0.24	0.40	0.64	0.52	0.25	0.30	0.60	0.41	0.56	0.25	0.18	0.37	0.431
NBB [27]	0.71	0.73	0.32	0.28	0.45	0.68	0.61	**0.35**	0.38	**0.80**	**0.45**	**0.65**	0.28	**0.20**	0.33	0.481
$Ours_{global}$	**0.84**	0.73	0.46	0.25	0.36	0.70	0.76	**0.56**	0.26	0.83	0.58	**0.71**	0.41	0.33	0.40	0.545
$Ours_{local}$	0.83	0.74	**0.54**	0.30	0.38	**0.78**	**0.80**	**0.56**	0.28	**0.89**	**0.60**	**0.71**	**0.51**	**0.35**	0.35	**0.575**

(a) Source image (b) PF (c) DSFM (d) NBB (e) ours (f) target image

FIGURE 6. The visual comparisons of matching results based on the ground truth of Pascal 3D+ dataset. It was done between the proposed approach (e) to different methods, i.e., (b) PF [35], DSFM [5], and NBB [27].

or difference in viewpoint. To be more convincing, we compare the proposed approach to the state-of-the-art methods on semantic correspondence.

1) RESULT ON THE PF-PASCAL DATASET [35]

Since the estimation of PCK relies on the density and accuracy of correspondences, and the corresponding transformation, we evaluate these on the PF-PASCAL dataset, with the larger PCK corresponding to more accurate correspondence and transformation. Concretely, the evaluation is implemented with $\theta = 0.1$ and the tolerance error for correspondence is approximately 20 pixels, and $\theta = 0.05$ and the tolerance error is about 10 pixels, respectively. Experiments illustrate that the proposed approach outperforms the other pairwise correspondence methods, and more detailed comparison per-class is shown in Tab. 2 and 3.

The results show that traditional handcrafted features or other non-feature based matching are insufficient for semantic object matching, such as SIFT-Flow [13], DSP [14], and Zhou *et al.* [38]. They cannot be effectively used to extract enough salient features from original images with non-overlapping regions. Some other works, such as DSFM [5], PF [35], Liao *et al.* [43], NBB [27], and our approach, build on pre-trained CNN features and obtain better correspondences. Note that we utilize the notion of

TABLE 3. Mean PCK ($\theta = 0.05$) evaluation on the PF-PASCAL dataset.

Methods	PCK
DSP [14]	0.174
Collection Flow [40]	0.125
RASL [41]	0.158
PF [35]	0.170
SCNet [42]	0.180
DSFM [5]	0.249
Liao et al. [43]	0.178
NBB [27]	0.211
PARN [46]	**0.268**
$Ours_{global}$	0.329
$Ours_{local}$	**0.335**

hierarchical optimization [27] with introducing an imitation foreground detection method, and our approach outperforms these other works. Tab.2 shows that our method outperforms the other methods, obtaining an overall PCK of 54.5%, which is a 6.1% improvement over the best competitor [35]. These indicate that the proposed consistency constraints and optimization strategy are effective, and the coarse-to-fine matching process works well on semantic feature matching.

However, the results show that the generalization ability of other matching methods and our approach cannot meet our expectations. It is represented as binarization as shown in Tab. 2, such as the categories of bird, pers, and plant

TABLE 4. PCK($\theta = 0.1$) comparison on the PF-WILLOW dataset.

Methods	car(S)	car(G)	car(M)	duck(S)	mot(S)	mot(G)	mot(M)	win(w/o C)	win(w/ C)	win(M)	*Avg.*
Deep Flow [37]	0.33	0.13	0.22	0.20	0.20	0.08	0.13	0.46	0.08	0.18	0.201
GMK [15]	0.48	0.25	0.34	0.27	0.31	0.12	0.15	0.41	0.17	0.18	0.268
SIFT Flow [13]	0.54	0.37	0.36	0.32	0.41	0.20	0.23	0.83	0.16	0.33	0.375
DSP [14]	0.46	0.30	0.32	0.25	0.31	0.15	0.14	0.85	0.25	0.64	0.367
Zhou et al. [38]	0.77	0.34	0.52	0.42	0.34	0.19	0.20	0.78	0.19	0.38	0.413
PF [35]	**0.86**	**0.60**	**0.53**	**0.64**	**0.49**	0.25	0.29	**0.91**	0.37	**0.65**	**0.559**
Liao et al. [43]	0.78	0.53	0.45	0.43	0.35	0.32	0.33	0.82	0.21	0.56	0.469
NBB [27]	0.81	0.59	0.52	0.53	0.42	**0.34**	**0.36**	0.84	**0.38**	0.62	0.541
$Ours_{global}$	0.89	**0.73**	0.65	0.55	0.57	0.59	0.54	0.81	0.44	0.57	0.634
$Ours_{local}$	**0.93**	**0.73**	**0.68**	**0.70**	**0.65**	**0.63**	**0.60**	0.85	**0.46**	0.63	**0.686**

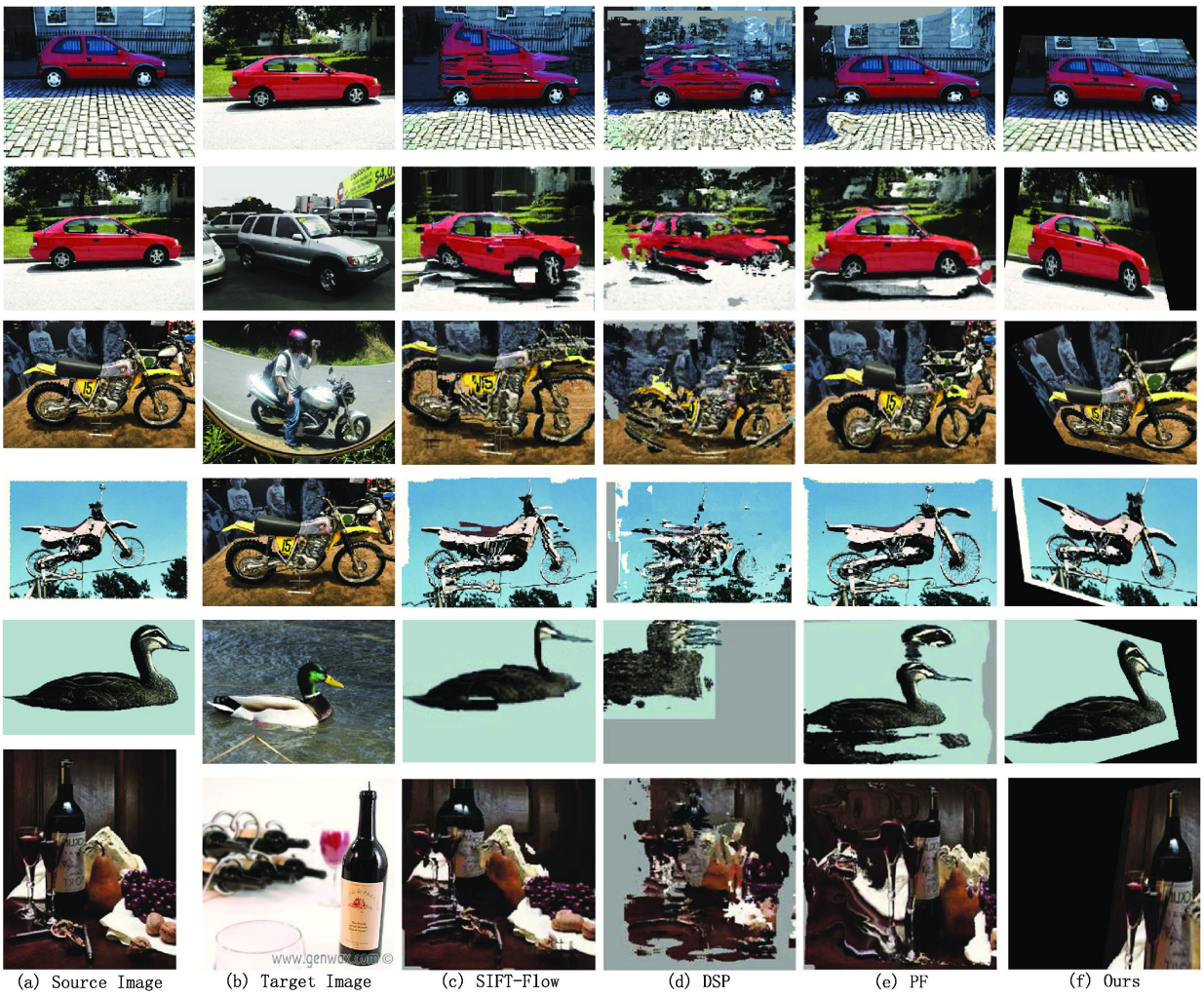


FIGURE 7. Alignment examples on the Proposal Flow datasets [35]. (a) Source images are deformed and aligned to (b) Target images according to the estimated transformations using different methods, i.e., (c) SIFT-Flow [13], (d) DSP [14], (e) PF [35], and (f) Our approach.

with low accuracy. It is mainly caused by strong occlusion, background cluster, and single geometric feature contained in original images. Typically, the bird in the grass and the flowerpot filled with plants, bringing more challenges and resulting in insufficient keypoints for correspondence. Besides, another result with $\theta = 0.05$ is achieved from the corresponding literatures due to some belong to proprietary projects as shown in Tab. 3. It clearly shows that our approach gains an obvious improvement over other matching methods.

2) RESULT ON THE PF-WILLOW DATASET [35]

Next, we evaluate the robustness of the proposed approach on the PF-WILLOW dataset by calculating the PCK with $\theta = 0.1$. More challenging examples are selected for evaluation. The key of matching semantic features is to accurately extract the salient features from the main objects of interest in the examples with background clutter, and to address differences in appearance to estimate the transformation model between the examples with changes in viewpoint. The results

illustrate that our approach outperforms the other methods as shown in Tab. 4. Overall, the proposed approach is more robust than other matching methods.

3) DISCUSSION

The results of semantic correspondence using global and local transformation estimation are represented as $Ours_{global}$ and $Ours_{local}$ separetely, as shown in Tab. 2, 3, and 4. Specifically, since the size of the feature maps at the lower pyramid layers ($L = 3, 4, 5$) is insufficient to locally calculate the transformation, the difference between two schemes is concentrated in the first two layers of pyramid ($L = 1, 2$), as shown in Figure 1 (b). In addition, the example of the matching is shown in Figure. 4 and visual comparison of matching results is implemented on a comprehensive dataset, i.e., Pascal 3D+ dataset, as shown in Figure. 6. The annotations in the original image are transformed to the target image according to the estimated transformations using different methods. To effectively perform the comparison, some results are achieved from the NBB [27] and our methods outperforms these other works.

E. IMAGE ALIGNMENT

Early, some existing works, such as SIFT Flow [13] and DSP [14], use handcrafted features for semantic correspondence. They build on the notion of flow and implement matching and alignment between images by calculating the vector displacement among pixels. Furthermore, some other methods are concerned with establishing better transformation models for alignment [5], [7], and it is also an extension and development trend of semantic correspondence. Our approach establishes correspondences by searching for CNN feature keypoints and estimates the corresponding transformation between images, which is based on sparse features.

The source image is deformed and aligned to the target image according to the estimated transformations using the established correspondences [39]. To effectively evaluate the proposed approach, we select some challenging examples with side viewpoints, mixed viewpoints (i.e., general viewpoints + side viewpoints), and background clutter. Experiments illustrate that obvious distortion appears in the alignments [13], [14], [35], and the proposed approach performs well on image alignment, as shown in Figure. 7. Overall the proposed approach establishes successful semantic feature matching.

Discussion: The resulting correspondences are evaluated on benchmark datasets as shown in Tab. 2 and 3, and the results of the robustness evaluation are shown in Tab. 4. Our approach performs well on semantic feature matching and alignment, and obtains the competitive performance to many methods, such as SIFT-Flow [13], DSP [14], and PF [35] regarded as the state-of-the-art methods for semantic feature matching, and Zhou *et al.* [38], DSFM [5], SCNet [42], and NBB [27] proposed in the latest years. The alignment results are shown in Figure. 7.

VI. CONCLUSION

We have proposed a novel approach for semantic feature matching based on pre-trained CNN features, with enforcing the nearest-neighbor searching under additional consistency constraints and guiding correspondence improvements with a hierarchical optimization strategy. According to the characteristics of CNN features and intra-class objects, we perform a coarse-to-fine process by minimizing an objective function and introducing the corresponding optimization scheme. The results clearly demonstrate the competitive performance of the proposed approach for semantic correspondence on standard benchmark datasets.

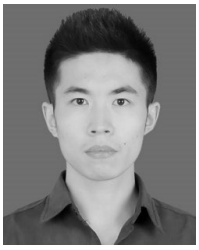
ACKNOWLEDGMENT

(Wei Lyu and Lang Chen contributed equally to this work.) W. Lyu would like to thank Ms. C. Wang for her assistance and contributions to the article, including grammar modification, logical suggestions, and proof-reading.

REFERENCES

[1] R. Szeliski, "Image alignment and stitching: A tutorial," *FNT Comput. Graph. Vis.*, vol. 2, no. 1, pp. 1–104, Jan. 2007.
[2] N. Snavely, S. M. Seitz, and R. Szeliski, "Photo tourism: Exploring photo collections in 3d," *ACM Trans. Graph.*, vol. 25, no. 3, pp. 835–846, Jul. 2006.
[3] R. A. Newcombe, S. J. Lovegrove, and A. J. Davison, "DTAM: Dense tracking and mapping in real-time," in *Proc. 2011 Int. Conf. Comput. Vis.*, Barcelona, Spain, Nov. 2011, pp. 2320–2327.
[4] A. Hosni, C. Rhemann, M. Bleyer, C. Rother, and M. Gelautz, "Fast cost-volume filtering for visual correspondence and beyond," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 2, pp. 504–511, Feb. 2013.
[5] N. Ufer and B. Ommer, "Deep semantic feature matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5929–5938.
[6] F. Yang, X. Li, H. Cheng, J. Li, and L. Chen, "Object-aware dense semantic correspondence," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2777–2785.
[7] I. Rocco, R. Arandjelovic, and J. Sivic, "End-to-end weakly-supervised semantic alignment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6917–6925.
[8] T. Yu, J. Meng, and J. Yuan, "Multi-view harmonized bilinear network for 3D object recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 186–194.
[9] J. Si, H. Zhang, C.-G. Li, J. Kuen, X. Kong, A. C. Kot, and G. Wang, "Dual attention matching network for context-aware feature sequence based person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5363–5372.
[10] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004.
[11] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (SURF)," *Comput. Vis. Image Understand.*, vol. 110, no. 3, pp. 346–359, Jun. 2008.
[12] B. K. Horn and B. G. Schunck, "Determining optical flow," *Artif. Intell.*, vol. 17, nos. 1–3, pp. 185–203, Aug. 1981.
[13] C. Liu, J. Yuen, and A. Torralba, "Sift flow: Dense correspondence across scenes and its applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 5, pp. 978–994, May 2011.
[14] J. Kim, C. Liu, F. Sha, and K. Grauman, "Deformable spatial pyramid matching for fast dense correspondences," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 2307–2314.
[15] O. Duchenne, A. Joulin, and J. Ponce, "A graph-matching kernel for object categorization," in *Proc. Int. Conf. Comput. Vis.*, Barcelona, Spain, Nov. 2011, pp. 1792–1799.
[16] C. Barnes, E. Shechtman, A. Finkelstein, and D. B. Goldman, "PatchMatch: A randomized correspondence algorithm for structural image editing," *ACM Trans. Graph.*, vol. 28, no. 3, p. 24, Aug. 2009.
[17] C. Barnes, E. Shechtman, D. B. Goldman, and A. Finkelstein, "The generalized patchmatch correspondence algorithm," in *Proc. Eur. Conf. Comput. Vis.*, Berlin, Germany, 2010, pp. 29–43.

[18] P. Fischer, A. Dosovitskiy, and T. Brox, "Descriptor matching with convolutional neural networks: A comparison to sift," May 2014, *arXiv:1405.5769*. [Online]. Available: https://arxiv.org/abs/1405.5769
[19] J. L. Long, N. Zhang, and T. Darrell, "Do convnets learn correspondence?" in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 1601–1609.
[20] R. Girshick, F. Iandola, T. Darrell, and J. Malik, "Deformable part models are convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 205, pp. 437–446.
[21] X. Zhou, M. Zhu, and K. Daniilidis, "Multi-image matching via fast alternating minimization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4032–4040.
[22] Q. Wang, X. Zhou, and K. Daniilidis, "Multi-image semantic matching by mining consistent features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*, Jun. 2018, pp. 685–694.
[23] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Commun. ACM*, vol. 55, no. 6, p. 111, Jun. 2012.
[24] I. Rocco, R. Arandjelovic, and J. Sivic, "Convolutional neural network architecture for geometric matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 6148–6157.
[25] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381–395, Jun. 1981.
[26] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.
[27] A. Kfir, L. Jing, S. Mingyi, L. Dani, C. Baoquan, and O. and C. Daniel, "Neural best-buddies: Sparse cross-domain correspondence," *ACM Trans. Graph.*, vol. 37, no. 4, p. 69, Aug. 2018.
[28] T. Dekel, S. Oron, M. Rubinstein, S. Avidan, and W. T. Freeman, "Best-buddies similarity for robust template matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2021–2029.
[29] K. Simonyan, A. Vedaldi, and A. Zisserman, "Learning local feature descriptors using convex optimisation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 8, pp. 1573–1585, Aug. 2014.
[30] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. 2009 IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
[31] T. Tuytelaars and K. Mikolajczyk, "Local invariant feature detectors: A survey," *FNT Comput. Graph. Vis.*, vol. 3, no. 3, pp. 177–280, 2007.
[32] V. Dumoulin and F. Visin, "A guide to convolution arithmetic for deep learning," Mar. 2016, *arXiv:1603.07285*. [Online]. Available: https://arxiv.org/abs/1603.07285
[33] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," Feb. 2015, *arXiv:1502.03167*. [Online]. Available: https://arxiv.org/abs/1502.03167
[34] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1501–1510.
[35] B. Ham, M. Cho, C. Schmid, and J. Ponce, "Proposal flow," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 3475–3484.
[36] Y. Yang and D. Ramanan, "Articulated human detection with flexible mixtures of parts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 12, pp. 2878–2890, Dec. 2013.
[37] J. Revaud, P. Weinzaepfel, Z. Harchaoui, and C. Schmid, "DeepMatching: Hierarchical deformable dense matching," *Int. J. Comput. Vis.*, vol. 120, no. 3, pp. 300–323, Dec. 2016.
[38] T. Zhou, P. Krahenbuhl, M. Aubry, Q. Huang, and A. A. Efros, "Learning dense correspondence via 3D-guided cycle consistency," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 117–126.
[39] J. Zaragoza, T.-J. Chin, M. S. Brown, and D. Suter, "As-projective-as-possible image stitching with moving DLT," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 2339–2346.
[40] I. Kemelmacher-Shlizerman and S. Seitz, "Collection flow," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 1792–1799.
[41] Y. Peng, A. Ganesh, J. Wright, W. Xu, and Y. Ma, "RASL: Robust alignment by sparse and low-rank decomposition for linearly correlated images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2233–2246, Nov. 2012.
[42] K. Han, R. S. Rezende, B. Ham, K.-Y.-K. Wong, M. Cho, C. Schmid, and J. Ponce, "SCNet: Learning semantic correspondence," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1831–1840.
[43] J. Liao, Y. Yao, L. Yuan, G. Hua, and S. B. Kang, "Visual attribute transfer through deep image analogy," *TOGACM Trans. Graph.*, vol. 36, no. 4, pp. 1–15, Jul. 2017.
[44] S. Kim, D. Min, B. Ham, S. Jeon, S. Lin, and K. Sohn, "FCSS: Fully convolutional self-similarity for dense semantic correspondence," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 6560–6569.
[45] S. Kim, D. Min, S. Lin, and K. Sohn, "DCTM: Discrete-continuous transformation matching for semantic flow," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4529–4538.
[46] S. Jeon, S. Kim, D. Min, and K. Sohn, "PARN: Pyramidal affine regression networks for dense semantic correspondence," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany, Sep. 2018, pp. 351–366.



WEI LYU received the B.S. degree in computer science from Sichuan Agricultural University, Yaan, China, in 2011, and the M.E. degree in computer science from Guizhou University, Guiyang, China, in 2015. He is currently pursuing the Ph.D. degree with the State Key Laboratory of Virtual Reality Technology and Systems, College of Computer Science, Beihang University, Beijing, China. His research interests include semantic matching, semantic segmentation, geometric modeling, and virtual reality.

LANG CHEN received the B.S. degree in computer science from the Beijing Institute of Technology, Beijing, China, in 2017. He is currently pursuing the M.S. degree with the State Key Laboratory of Virtual Reality Technology and Systems, College of Computer Science, Beihang University, Beijing. His main research interests are computer vision and image processing, including image instance matching and image semantic matching.

ZHONG ZHOU (Member, IEEE) received the B.S. degree from Nanjing University, Nanjing, China, in 1999, and the Ph.D. degree from Beihang University, Beijing, China, in 2005. He is currently a Professor and a Ph.D. Adviser with the State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing, China. His main research interests include virtual reality/ augmented reality/mixed reality, computer vision, and artificial intelligence. He is a member of ACM and CCF.

WEI WU received the Ph.D. degree from the Harbin Institute of Technology, Harbin, China, in 1995. He is currently a Professor with the State Key Laboratory of Virtual Reality Technology and Systems, Beihang University. His current research interests include virtual reality, wireless networking, and distributed interactive systems. He is the Chair of the Technical Committee on Virtual Reality and Visualization, China Computer Federation.