Neurocomputing 447 (2021) 257–271

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Weakly supervised object-aware convolutional neural networks for semantic feature matching

Wei Lyu, Lang Chen, Zhong Zhou *, Wei Wu

State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing 100191, China

ARTICLE INFO

Article history:
Received 30 May 2020
Revised 6 March 2021
Accepted 13 March 2021
Available online 20 March 2021
Communicated by Zidong Wang

Keywords:
Convolutional neural network
Semantic feature matching
Nearest-neighbor searching
Semantic perception
Cycle consistency
Image alignment

ABSTRACT

We address the task of establishing visual correspondences between two images depicting main objects of the same semantic category. This task encounters various challenges such as background clutter, intra-class variation, and viewpoint variations. Existing works are dominated by end-to-end training methods that rely on redundant calculation or large amounts of manual annotations, and cannot generalize to unseen object categories. In this paper, we propose to construct a weakly supervised object-aware convolutional neural network architecture for semantic feature matching, while being trainable end-to-end without the requirement for manual annotations. The main component of this architecture is a similarity filter module containing a trainable neural nearest neighbors network. Since training data for semantic feature matching is rather limited, we introduce a simple and effective foreground selection strategy to produce the foreground masks. Using these masks as a form of weak supervision signal for correspondence task and tackle the background clutter. Extensive experiments illustrate that the proposed approach outperforms the state-of-the-art methods for semantic feature matching on multiple public standard benchmark datasets.

1. Introduction

Establishing correspondences, which is traditionally defined as calculating the associations among multiple images depicting the same scene or object, is one of the fundamental problems in computer vision and graphics. This has been widely used in a variety of graphics fields such as image stitching [1,5], 3D reconstruction [3,4], and stereo matching [2]. They search for the correspondences with different handcrafted features, typically Scale-Invariant Feature Transform (SIFT) [10], Histogram of Oriented Gradient (HOG) [11], Speeded Up Robust Features (SURF) [36], and some improved descriptors [49]. Some researchers have also been committed to seeking better matching techniques [51,52,54,55]. With the breakthrough of strong representation capabilities of the Convolutional Neural Networks (CNNs), many excellent matching algorithms are proposed [47,48,50,53], and semantic understanding-based matching has been also developed in the latest years [13,14]. Essentially, it is the basis for some rising fields such as semantic object segmentation [43], object detection [6,37], and Re-identification [7].

Semantic feature matching is concerned with estimating the correspondences between two objects of the same semantic category in different images, which can be roughly divided into two branches. The first branch aims to construct a post-processor [21,14,22]. The extracted handcrafted features [10,22] or the learned CNN features are taken as inputs to the designed processor [14,15]. Matching constraints are used to minimize appearance matching cost and enforce geometric consistency between all candidate feature pairs. However, it generally obtains low accuracy performance that cannot meet the requirements for further applying, resulting in rarely used for semantic feature matching. Another branch of the methods is based on a correlation filter and CNNs. A similarity filter is trained by encoding the spatial consistency and semantic associations between intra-class objects. Existing methods develop different convolutional neural network architectures for correspondence task which are trainable end-to-end to improve the accuracy [18–20]. But they generally prefer to estimate the parameters of the geometric transformation relating the input images instead of the matched features, resulting in a narrower applicability. Meanwhile, they are sensitive to the interference factors present in the images, such as background clutter, and intra-class variation. Besides, shallow neural network and large-kernel convolution increase the computational complexity [18]. And [19] strongly relies on the synthetic datasets which reduces the

* Corresponding author.
E-mail address: zz@buaa.edu.cn (Z. Zhou).

https://doi.org/10.1016/j.neucom.2021.03.052
0925-2312/

generalization capabilities of the model for unseen object categories.

In this work we establish sparse feature associations between intra-class semantic objects, as shown in Fig. 1. It is challenging in background clutter, intra-class variation, changes in viewpoint and illumination, and non-overlapping of scenes or objects. Inspired by the state-of-the-art semantic feature matching method, i.e., NCNet [18], we construct a weakly supervised convolutional matching network for correspondence task. The key is to search for sufficient salient features and estimate the correspondences between two objects by fully exploiting their similar semantics. In contrast to the original version [18], we aim to design an object-aware matching mechanism to alleviate the background clutter. Essentially, our approach adopts a salient foreground selection strategy to produce the foreground masks, which provides a form of weak supervision signal to train a re-ranking convolutional neural network. This mechanism can effectively constrain the nearest-neighbor searching scope, and perceive main semantic regions. Specifically, we introduce a common 2-D re-ranking network instead of a complicated 4-D neighbourhood consensus module.

We propose a weakly supervised object-aware convolutional neural network architecture for semantic feature matching, consisting of three main modules: feature extraction, similarity measurer, and similarity filter, as shown in Fig. 2. Given two input images depicting main objects of the same semantic category, we first adopt a very weak supervision in the form of ImageNet pretrained feature representations [28] for each image, which are analogous to dense local descriptors and readily available. This obtains the object-specific attribute representations and low-level contexts such as colors and edges. Then we implement an attribute transfer process to eliminate the interference caused by the differences in color space. The purpose of this process is to simultaneously alleviate the confusion caused by the low-level visual features, and provide normalized data for further filtering. Further, a common correlation layer is used to match the feature representations across images into the tentative correlation maps, namely the initial correlation maps.

Finally, a similarity filter module is introduced to produce the resulting correspondences. We first introduce a cycle consistency constraint to weight the initial correlation maps. This can initially distinguish between the inliers and outliers from the collection of the correlated features, encourage one-to-one matching, and reduces the computational load of the filter network. Then we develop an object-aware matching mechanism. A neural nearest neighbors network (3N-Network) is driven by designing a semantic perception loss function. Motivated by the notion of the classical *k*-nearest neighbor matching strategy, this module enforces the nearest-neighbor searching process under a confidential salient constraint, which effectively mitigates the interference caused by the background clutter. Specifically, the filter module is used to accelerate calculations, and detect the positive correspondences by fully exploiting the local associations between objects. Analogously to a mutual nearest-neighbor matching process, this module can parse more local nonrepresentational features from images. The main contributions of this work are three-folds:

- We propose to construct a weakly supervised object-aware convolutional matching network architecture, while being trainable end-to-end without the requirement for manual annotations.
- We develop an object-aware matching mechanism. A simple and effective foreground selection strategy is incorporated into a semantic perception loss to enable weakly-supervised learning. This enforces the nearest-neighbor searching process in the main semantic regions, reduces the computational load, and enhances the capability of extracting the salient features.
- Extensive experiments thoroughly validate the effectiveness of the proposed approach on multiple public standard benchmark datasets, where it also outperforms state-of-the-art methods for semantic feature matching.

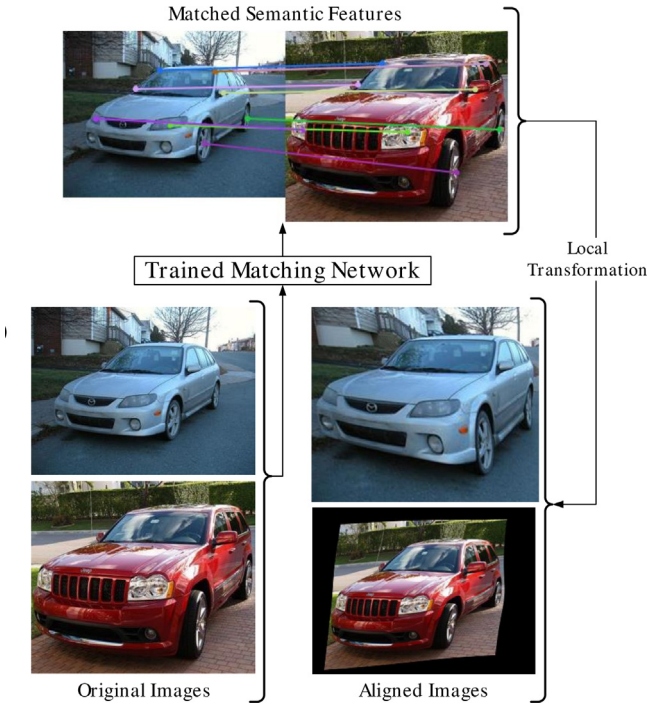

Fig. 1. The original images are passed through the proposed matching network architecture, which is trainable end-to-end without the requirement for annotations, to produce the matched semantic features. Furthermore, we locally deform and align the original images with appearance differences using the resulting correspondences. The keypoint pairs should be clustered in the main semantic regions and located at the salient positions.

2. Related work

Semantic feature matching has gained rising attention in the last several years. Recent works are concerned with learning-based matching, and continue to make new advances.

2.1. Flow-based methods

Early works are mainly based on the notion of flow and the handcrafted feature descriptor [10]. The first version calculates the displacement vectors of discrete pixel-points using a hierarchical optimization strategy. The main idea is to enforce geometric consistency to minimize the appearance matching cost [8,9]. Further, a spatial pyramid matching framework is presented by Kim et al. [21]. They regularize the correspondence consistency from an entire image, to coarse grids, to each pixel rather than only pixel, which improves matching accuracy in the face of challenging intra-class variations. But these are all limited to the complexity of matching the scenes or objects with background clutter. It is difficult to effectively distinguish between the main semantic region and background. To tackle this problem, object detection method is introduced to narrow down the search regions [22,23]. Most of these works rely on geometric constrains equivariance to transformations. However, since handcrafted descriptors are sensitive to appearance variations and originally designed for the same scene or object, they are not suitable for semantic correspondence.

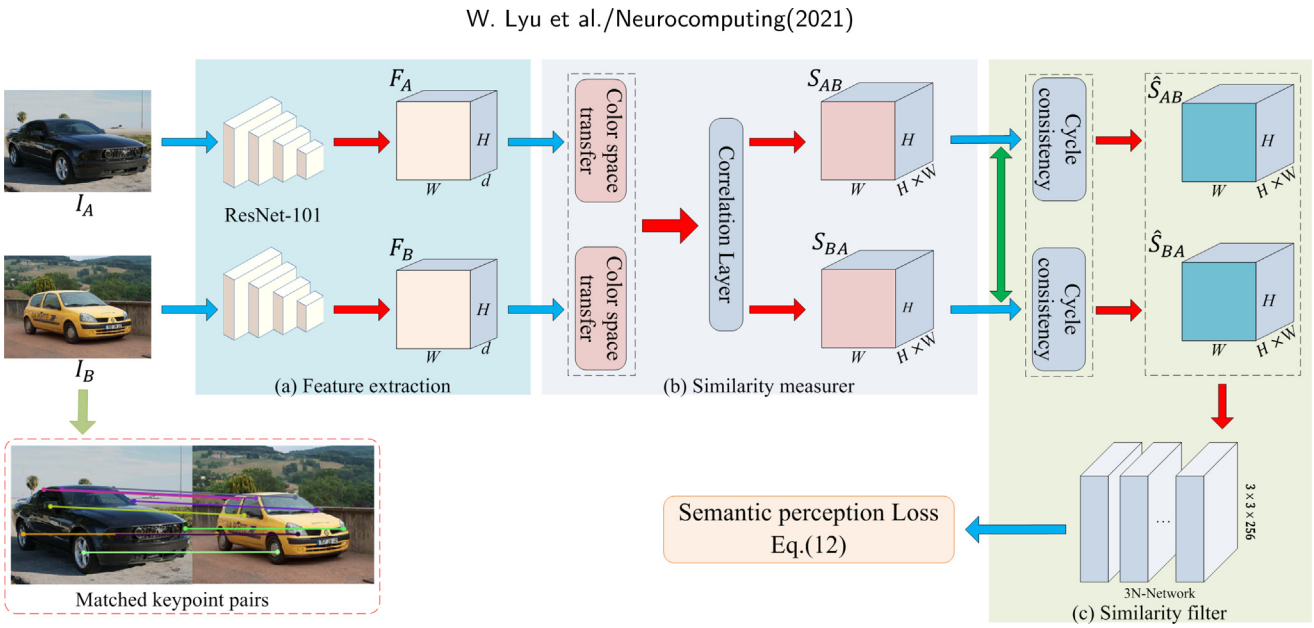

Fig. 2. Overview of our approach. Given two input images I_A and I_B depicting main objects of the same semantic category, which are passed through two identical convolutional networks which share parameters for feature extraction in (a). Then we propose to perform an attribute transfer process, and then introduce a correlation layer to produce the initial plausible correlation maps S_{AB} and S_{BA} shown in part (b), which contain the scores of all pairwise. Further, a cycle consistency constraint is used to weight the initial correlation maps, and encourage one-to-one matching. Finally, the resulting correspondences are established by developing a neural nearest neighbors network (3N-Network), and a salient foreground selection strategy is adopted to produce the foreground masks, which provides a form of weak supervision signal for correspondence task, as shown in (c). Specifically, the proposed matching network architecture aims to locate salient features and produce the matched keypoint pairs instead of only learning the transformation parameters.

2.2. CNN features based post-processing mechanisms

More recently, some works are concerned with determining correspondences using the learned CNN feature representations. They first adopt some pre-trained convolutional neural network architectures for classification, such as AlexNet [24] and VGG-19 [25], for feature extraction. Then a post-processor is designed for inlier detection, and to produce the dense correspondences. [14] formulated this task as solving a sparse Markov Random Field (MRF) model to enforce geometry consistency and appearance consistency between intra-class objects. [15,16] introduced a hierarchical optimization strategy to further improve the correspondences. They both generate a convolutional feature map pyramid using a pre-trained VGG-19 model [25], and the correspondences are found at each level from the top pyramid levels to the bottom ones. [35] combined convolutional pyramid and geometric consistency to perform a reverse mapping hierachical correspondence process. The CNN network is only used to extract the semantic features, and the whole process is not trainable. These works show better performance of using the CNN features for semantic feature matching compared to the handcrafted features, and they also have no requirement for additional annotations. But the utilization of deep neural networks still needs to be improved and the accuracy performance is insufficient for further applying.

2.3. End-to-end trainable CNNs

End-to-end trainable matching network is driven by the powerful information mining and fitting capabilities of deep neural network architectures. Rocco et al. [20] presented a geometric matching network which is trainable end-to-end on the handcrafted datasets. Their architecture performs the standard steps of feature extraction, feature matching, and simultaneous inlier detection and model parameter estimation. Generally, the generalization capabilities of their method among different scenes is weak, and the matching model strongly relies on the synthetic datasets that are rather costly. To mitigate these problems, a scoring mechanism is used for outlier rejection [19], and a pyramid regression network architecture is constructed by stacking the previously mentioned matching module [39]. Similarly, an adaptive learning model is designed to produce an effective CNN feature descriptor [27,17]. [19,20] both focus on estimating the parameters of the geometric transformation relating the input images, and aligning the salient semantic regions rather than establishing pairs of the matched features. We aim to implement semantic feature matching which provides sparse feature pairs for a variety of applications.

Kim et al. [40] presented a recurrent transformer matching network. Their main idea is to adopt a self-supervision training mechanism to avoid manual annotations, whereas resulting in high complexity. Rocco et al. [18] proposed a neighbourhood consensus network to find sparse correspondences between a pair of images, which is trainable end-to-end in a weakly-supervised manner. However, they parse all possible correspondences by traversing all the feature representations, increasing the amount of unnecessary calculations and the possibility of the outliers. In addition, Lee et al. [42] constructed their matching network framework by combining the notion of flow and deep neural network. They learn the semantic flow by estimating the transformation parameters with the need for additional foreground masks. Specifically, existing works are concentrated on handling the intra-class variation by designing different matching networks, and our approach aims to alleviate the background clutter developing an object-aware matching mechanism.

3. Proposed approach

In this section, we describe the proposed framework for semantic feature matching in detail. As shown in Fig. 2, given a pair of

images depicting main objects of the same semantic category, a pre-trained CNNs model is first used for feature extraction. Then we introduce a color space homogenization method to perform an attribute transfer process, and a correlation layer is adopted to produce the initial tentative correspondence maps aross images. Furthermore, a cycle consistency constraint is used to enforce the one-to-one matching constraint. Finally, a neural nearest neighbors network is developed to produce the resulting correspondences. The resulting pipeline can be trained in an end-to-end manner for correspondence task.

3.1. Feature extraction

The first step of the proposed approach is feature extraction, for which we adopt a standard CNN model without fully connected layers. We formulate it as a siamese architecture such that the two input images are passed through two identical convolutional networks which share parameters. This module extracts discriminative image features through multiple convolutional layers, and produces the corresponding feature maps for each image. Given an image pair (I_A, I_B), each image is taken as an input to the ResNet-101 model [28] which has superior performance of parsing high-level semantics [12,13]. Specifically, feature representations (F_A, F_B) are produced from the $Conv4$ layer of the CNN model initialized on ImageNet [31] for the task of image classification. They are $H \times W \times d$ tensors, which are denoted as dense $H \times W$ spatial grids of d-dimensional local features and a feature map $F \in \mathbb{R}^{H\times W\times d}$.

3.2. Similarity measurer

Below, a similarity measurer is designed to determine the initial plausible correspondences for further inlier detection. Firstly we adopt a color space homogenization method to alleviate the interference caused by the differences in color space, and normalize the discrete feature representations. Then a correlation layer is introduced to measure the similarities between the normalized CNN features, and produce the initial correlation maps that contain the scores of all pairwise. Analogously to the classical matching method [10], only descriptor similarities and the corresponding spatial positions should be considered instead of the original descriptors themselves.

Generally, there are color differences and changes in illumination between the input images with natural scenery. Meanwhile, the robustness of matching can be first enhanced by integrating the regularization theory. A color space homogenization method, which is formulated as an attribute transfer model, is introduced to simultaneously handle these. Given the feature representations $F_A, F_B \in \mathbb{R}^{H\times W\times d}$, we adopt a Z-Score normalization method with introducing additional balance factors, as detailed

$$\widehat{F}_A = \beta_{AB} \cdot \frac{F_A - \mu(F_A)}{\sigma(F_A)} + \gamma_{AB} \tag{1}$$

where $\mu(\cdot), \sigma(\cdot) \in \mathbb{R}^d$represent the spatial mean and standard deviation of the feature representations in each channel seperately, the coefficients $\beta_{AB}, \gamma_{AB} \in \mathbb{R}^d$ are used to eliminate the interference caused by the differences in color space between images. Please refer to [38] for further details.

Next, a correlation layer is introduced to measure the similarities between the normalized feature representations $\widehat{F}_A$ and $\widehat{F}_B$. We adopt a cosine similarity metric function to generate the initial correlation map S_{AB} that contains the one-way scores of all pairwise from the image I_A to I_B, denoted as

$$S : \mathbb{R}^{H\times W\times d} \times \mathbb{R}^{H\times W\times d} \rightarrow \mathbb{R}^{H\times W\times (H\times W)} \tag{2}$$

$$S_{AB} = S_{AB}(p,q) = \frac{\widehat{F}_A(p)\ \widehat{F}_B^T(q)}{\left\|\widehat{F}_A(p)\right\| * \left\|\widehat{F}_B(q)\right\|} \tag{3}$$

where p and q are the positions of the two feature representations in the input images I_A and I_B, respectively. $\|\cdot\|$ denotes the L2 norm. The same method can be used for S_{BA}.

3.3. Similarity filter

This module aims to filter out most of the outliers from the initial correlation maps S_{AB} and S_{BA}, and produce the resulting correspondences. Firstly we adopt a cycle consistency constraint to weight all the candidates, and preliminarily distinguish between the inliers and outliers. Then a neural nearest neighbors network (3N-Network) is designed for inlier detection, which is driven by developing an object-aware matching mechanism. The details are as follows.

Cycle consistency constraint. For our semantic correspondence task, the resulting correspondences should agree with the one-to-one mapping constraint. Whereas the initial similarity metric follows one-to-many rules, each feature representation corresponds to $H \times W$ scores in the initial correlation maps. Furthermore, the correspondences with high scores can be considered to be inliers, and each feature may match multiple features shown in Fig. 3(a). Thus we introduce a cycle consistency strategy to encourage one-to-one matching. Analogously to a mutual nearest-neighbor matching process among images [26], we estimate the association relationship between two feature representations with each other to enforce their one-to-one matching constraint, as well as calculate the associations $p \rightarrow q$ and $p \leftarrow q$, as shown in Fig. 3(b).

To effectively filter out the outliers, a cycle consistency constraint is adopted to weight the initial correlation maps, which is interpreted as the evaluation coefficients of the correlation maps. We take S_{AB} as the example in the following steps, and the weights for each candidate is denoted as

$$w_{A\rightarrow B} = \frac{S_{AB}(p_i, q)}{max\{S_{AB}(p_i, q)\}} \tag{4}$$

$$w_{B\rightarrow A} = \frac{S_{BA}(q_j, p)}{max\{S_{BA}(q_j, p)\}} \tag{5}$$

where the denominator, $max\{S_{AB}(p_i, q)\}$, searches for the maximum scores from the image I_A to I_B. Each feature representation p_i has N

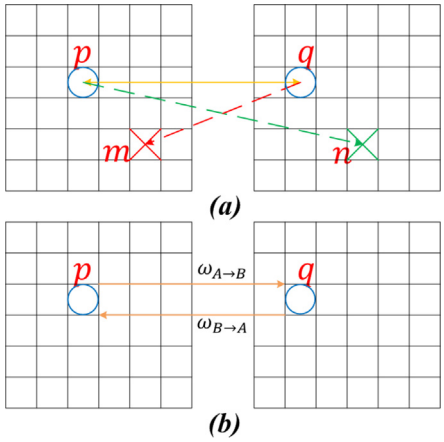

Fig. 3. Each initial correlation map contains the one-way scores of all pairwise from the original image to the target image. A feature representation p may match multiple features such as q, and n shown in part (a). A cycle consistency constraint is used to encourage one-to-one matching in (b).

scores corresponding to N feature representations q in I_B. The correlation map is reinterpreted as

$$\widehat{S}_{AB} = w_{A\to B} \cdot w_{B\to A} \cdot S_{AB} \tag{6}$$

For a candidate pair (p_i, q_j), $w_{A\to B}(p_i, q_j) = w_{B\to A}(q_j, p_i) \approx 1$ when obeying the cycle consistency constraint shown in Fig. 3(b). Otherwise, $w_{A\to B}(p_i, q_j) \cdot w_{B\to A}(q_j, p_i) \to 0$. This is equivalent to zeroing the scores of the outliers to ensure one-to-one matching.

The same method is used for S_{BA}, and to produce the corresponding $\widehat{S}_{BA}$. This strategy preliminarily distinguishes between the positive and negative correspondences, reduces the computational complexity of the filter network, and improves the fitting and inlier selection capabilities of the matching network. Specifically, this constraint is only used to encourage one-to-one matching, which can be considered as a preprocessing operation of the filter module.

Neural nearest neighbors network. The correlation maps $\widehat{S}_{AB}$ and $\widehat{S}_{BA}$ contain the scores of all pairwise, and rich contextual information. The neighbourhood of a positive correspondence covers other positive correspondences with high scores, because there is a local correlation between the CNN features. In contrast, no sufficient correspondences are used to support the outliers [29]. Therefore, we further parse the context and the correlations among features using a deep neural network. The idea presented by Rocco et al. [18] can be used, but they horizontally increase the dimension of each convolutional layer based on a shallow convolution framework (i.e., three layers for semantic correspondence). A key observation that, a deeper neural network model has better mining capabilities of the salient features [28] and can implement more discriminately mapping, compared to the corresponding shallow counterpart. To simultaneously reduce the computational load and enhance the fitting capabilities of the neural network, we construct our neural nearest neighbors network (3N-Network) using a common convolutional layer based on a deeper convolution model.

We implement 2-D convolution for the correspondence filtering task, as shown in Fig. 4. Concretely, we stack six blocks of convolutional layers with 3×3 kernels, followed by batch normalization and ReLU non-linearity. For the weighed correlation maps $\widehat{S}_{AB}$ and $\widehat{S}_{BA}$, we use the 3N-Network to produce the improved correlation maps C_{AB} and C_{BA}, which is driven by designing an object-aware matching mechanism. Essentially, 3N-Network weights all matching scores following the one-to-one mapping constraint, which is based on a classical k-nearest neighbor algorithm [10]. Each weighed correlation map is passed through the 3N-Network. According to the scores of matching the corresponding neighbors, a 3×3 kernel is used to upweight and downweight all the candidates. A candidate pair is considered to be inliers when establishing enough positive correspondences between its neighbors, as well as obtaining high score. Finally, the output has the same dimensions as the 2D input correspondences. A re-ranking process for pairwise matching and the corresponding non-linear mapping operations are iteratively implemented to exploit the global correspondence information.

Essentially, we adopt a common 2-D re-ranking network model, which counts the matching scores of the neighbourhood of each correspondence through convolution, so as to re-rank the correspondences. Compared to neighbourhood consensus network [18], 3N-Network achieves improvements in performance and efficiency while ensuring the same size of perception field. There are two advantages as follows. On the one hand, we adopt a common 2-D convolution module with 3×3 kernels for the correspondence filting task. The number of the corresponding parameters is dropped from original $25N$ to $18N$, compared to their complicated 4-D convolution network with 5×5 kernels. These simplify the matching model and contribute to the convergence of the network. On the other hand, smaller convolution kernel facilitates the capability of extracting salient features which can be transferred into high-level semantic information with subsequent convolution.

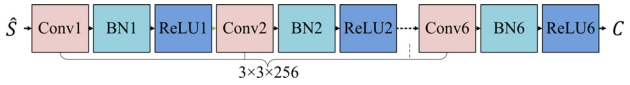

Fig. 4. Architecture of the neural nearest neighbors network. It consists of six convolutional layers with 3×3 kernels, followed by batch normalization and ReLU.

3.4. Semantic perception Loss

We observe that, pixels belonging to main semantic region in an image are usually matched to some pixels in other regions (e.g., background) in another image. To simultaneously tackle this issue and avoid the requirement for manual annotations such as the ground-truth correspondences or object proposal [39,42], we develop an object-aware matching mechanism, which designs a semantic perception loss function to drive the training of 3N-Network. We utilize the notion of the foreground detection method to select the salient features, and focus matching on the candidate objects, which approximately perceives the main semantic regions.

A salient foreground selection strategy is developed to serve the loss function for the training of the proposed 3N-Network. Firstly a scoring scheme is used to evaluate the C_{AB} and C_{BA} produced by the similarity filter. Then a simple and effective threshold selection strategy is introduced to select the salient features, to produce the foreground masks. Using these masks as a form of weak supervision signal for correspondence task. Furthermore, a restrictive loss is designed to maximize the scores of the positive correspondences. We take C_{AB} as the example in the following steps.

For the correlation map C_{AB}, a normalization process is first implemented by utilizing a soft-max function:

$$N_{AB}(p, q) = \frac{exp(C_{AB}(p, q))}{\sum exp(C_{AB}(p_i, q))} \tag{7}$$

where the maximum value is determined as the resulting score of the feature representation p_j, which is denoted as

$$\widehat{C}_{AB}(p_j) = max\{N_{AB}(p_j, q)\} \tag{8}$$

where $\widehat{C}_{AB} \in \mathbb{R}^{H\times W}$ contains the resulting scores of matching each CNN feature in image I_A to the best candidate in image I_B. Furthermore, it is used as a metric to filter out the outliers and select the salient features, which approximately focuses matching on the main semantic objects. Concretely, we estimate the mask of the candidate region, $Mask_{AB} \in \mathbb{R}^{h\times w}$,

$$Mask_{AB}(P) = \begin{cases} 1 & if\ T_{AB}(P) > \tau \\ 0 & otherwise \end{cases} \tag{9}$$

where 1 and 0 represents the main semantic region and background separately, the threshold τ is set to 0.5 in our experiments, and T_{AB} represents the evaluation function for the saliency of the feature representations, which is denoted as

$$T_{AB}(P) = \frac{\widehat{C}_{AB}(P) - min_i\left\{\widehat{C}_{AB}(p_i)\right\}}{\varepsilon + max_i\left\{\widehat{C}_{AB}(p_i)\right\} - min_i\left\{\widehat{C}_{AB}(p_i)\right\}} \tag{10}$$

where ε is a constant. Further, a one-way loss function, $Loss_{AB}(I_A, I_B)$, is constructed as

$$Loss_{AB}(I_A, I_B) = \frac{1}{H * W} \sum_{P} \widehat{C}_{AB}(P) * Mask_{AB}(P) \tag{11}$$

where $Loss_{BA}(I_B, I_A)$ is constructed the same as $Loss_{AB}$. Finally we incorporate them to the final training loss as

$$LOSS = Label(I_A, I_B) * (Loss_{AB} + Loss_{BA}) \tag{12}$$

To effectively avoid the potential overfitting, we add some noise data, typically negative image pairs containing the objects belonging to different categories, into training sets.

$$Label(I_A, I_B) = \begin{cases} -1 & if\ (I_A, I_B)\ are\ Positive\ Pairs \\ +1 & if\ (I_A, I_B)\ are\ Negative\ Pairs \end{cases} \tag{13}$$

In the training process, we aim to maximize the confidence of the main semantic region of the positive examples, and minimize the corresponding confidence of the negative examples. This facilitates continued attention to the positive correspondences and upweights them. In contrast, the weights of the outliers and non-salient features are reduced. They both contribute to perceive the candidate regions. Finally we try to enhance the capabilities of nearest-neighbor searching by exchanging the input order of the images I_A and I_B. Specifically, we introduce a empirical threshold selection strategy to produce the foreground mask using a important hyperparameter τ. This provides a form of weak supervision signal for correspondence task. Experiments have verified its effectiveness, and the threshold selection also follows a stable rule, shown in 7. In addition, since main semantic object occupies most of the region in the image used for training, it is beneficial to the implementation of our method.

Algorithm 1. Training Procedure Using Standard DataSet

Require: Image dataset *DS*, CNN model *M*
Ensure: Trained CNN model *M*
initial $\tau = 0.5$;
for *training epochs* **do**
  for *I in DS* **do**
    $I_A, I_B \leftarrow I$;
    $C_{AB}, C_{BA} \leftarrow M(I_A, I_B), M(I_B, I_A)$;
    $L \leftarrow LOSS(I_A, I_B)$;
    $W \leftarrow update(W, \frac{\delta L}{\delta W})$;
  end For
end For

4. Implementation and evaluation

In this section, we evaluate the performance of the proposed approach on several publicly available benchmark dat-asets for semantic feature matching. Meanwhile, the implementation details, results, analyses, and the comparisons to the state-of-the-art methods are provided in details.

4.1. Implementation detail

The proposed matching network framework is impleme-nted with PyTorch [32], and we train the network on an Intel Core i7-7700 CPU with an NVIDIA GeForce GTX 1080Ti GPU. For feature extraction, we adopt the ResNet-101 model [28] with up to $Conv4\_23$ layer, whose initial parameters are analogous to the pre-trained parameters of the ResNet-101 model on ImageNet [31] for the task of image classification. Input images are resized to 250×250 producing 16×16 feature maps that are passed into the matching layer. The matching network model is trainable end-to-end in a weakly supervised manner using the Adam optimizer [33] with learning rate 5×10^{-8}, and a batch size of 4. To avoid the potential overfitting, we swap the source and target images, and add negative image pairs into the training datasets. Specifically, we combine the learned CNN features themselves and a salient foreground object selection strategy to provide weak supervision information for training instead of manual annotations. Furthermore, the early stopping scheme is introduced to select best parameters of the matching network during training. The training algorithm is detailed in Algorithm 1.

We evaluate our approach with three parts which is organized as follows. Accuracy evaluation of the proposed approach is provided in Sections 4.2.1 and 4.2.3. Robustness evaluation is provided in Section 4.2.2. Finally we also present the qualitative evaluation in Section 4.2.4.

4.2. Matching results

Both of quantitative and qualitative are performed on three publicly available benchmark datasets generally used for this task. Accuracy evaluation of the proposed matching architecture is implemented on the Proposal Flow-PASCAL dataset [22] and Caltch-101 dataset [34], and robustness evaluation is on the Proposal Flow-WILLOW dataset [22] containing more challenging examples. Meanwhile, the proposed approach is comprehensively evaluated through comparisons to state-of-the-art methods for semantic feature ma-tching, including UCNet [17], PF-LOW [22], SCNet-AG+ [27], CNNGeo-R [20], End-to-End [19], NCNet [18], CAT-FCSS [45], SFNet [42], CC-DCTM [46], and MaCoSNet [41]. To validate the effectiveness of the proposed components, we implement the ablation experiments between our approach and NCNet [18]. Note that all the comparisons are based on the same training and test sets.

4.2.1. Results on Proposal Flow-PASCAL dataset

We first evaluate our approach on the Proposal Flow-PASCAL dataset [22], which contains image pairs depicting different instances of the same category, such as persons and cars. Images from each pair are manually selected to ensure that objects have similar poses. This dataset contains 20 semantic categories with totaling approximately 1300 image pairs. We utilize the data partitioning method used in [27]. Approximately 700 image pairs are used for training, 300 image pairs are used as the validation set, and the remaining 300 image pairs are used as the test set for the proposed matching network.

Evaluation Metric. The Proposal Flow-PASCAL dataset [22] provides manual annotations in each image pair as ground-truth, which are represented as the matched pairs of keypoints on intra-class semantic objects. So we implement the quantitative evaluation of the proposed approach using the standard evaluation metric for this benchmark, i.e., the percentage of correct key-points transfer (PCK) metric [27,30]. Specifically, the annotations are not used for the training of the proposed matching model, but only for testing. PCK is obtained by measuring the offset between the ground-truth position and the real location of transferring the candidate keypoint. A correspondence is considered to be inliers when the corresponding offset is less than a predefined distance threshold. For a sparse set of correspondences, $\left\{\left(P_S^i, P_T^i\right)\right\}_{i=1}^{n}$, between the source and target images, the annotated keypoints are warped from source image to target image using the estimated transformation ϕ with the resulting correspondences. The PCK is calculated as follows:

$$PCK = \frac{\left|\left\{P_S^i \in \boldsymbol{P_S}, \boldsymbol{d}\left(\phi\left(\boldsymbol{P_S^i}\right), \boldsymbol{P_T^i}\right) < \boldsymbol{Dist_{thre}}\right\}\right|}{n} \quad (14)$$

where $d(\cdot)$ is the Euclidean distance function. $Dist_{thre} = \theta \cdot max(h, w), \theta$ is a tolerance factor, and h and w is the height and width of the bounding box, respectively.

In addition, we also measure the mean intersection over union (mIoU) for different correspondence methods on this dataset benchmark. This metric measures the degree of overlap between the predicted object segmentations and ground truths. The average inference time is also measured and includes of all the pipelines of the matching framework.

Results. Calculating PCK relies on the density and accuracy of pairwise matching, and the estimated transformation, we evaluate these on the Proposal Flow-PASCAL dataset [22]. The average PCK is calculated for various matching techniques, as shown in Table 1. We summarize the matching accuracy for state-of-the-art matching techniques, and the larger PCK corresponds to more accurate matching and transformation. Comparisons are implemented with $\theta = 0.05, 0.10, 0.15$, and the corresponding tolerance error for matching is approximately 10, 20 and 30 pixels, respectively.

As it can be observed, compared to CNN feature based post-processing methods [22,45,46], end-to-end trainable CNNs [17,27,20,19,18,41,42] obtain better accuracy performance even though they do not incorporate geometric consistency into their matching model. In addition, some existing methods [17,27,20,19,18] aim to perform the classical matching process using the CNNs, but do not consider the essential problems, such as background clutter. The method of SFNet [42] manually synthesizes the foreground mask as supervisory signal to enable fully supervised learning. MaCoSNet [41] jointly trains an object segmentation network to provide weak supervision signal for correspondence network. In contrast to these methods, we utilize the learned features themselves to perform an object-aware matching mechanism without the requirement for annotations. Our approach achieves more competitive performance on the mainly benchmark dataset for semantic feature matching. Running time (average time per pair) for each method is shown in the last column in Table 1. Compared to [18], it takes less time to perform our matching framework. Besides, our mIoU value exceeds other methods. This verify the effectiveness of the proposed object-aware matching mechanism, which alleviates the background clutter.

Quantitative evaluation is also performed on the Proposal Flow-PASCAL dataset [22] with $\theta \in [0.06, 0.16]$, as shown in Fig. 5. Compared to the fully supervised learning model [20] which requires the parameters of the ground-truth geometric transformation, weakly supervised matching networks [18,19] provide higher accuracy in most cases. Our approach and [41] also obtain better performance over [42] shown in Table 1. These show that limited datasets are generally more suitable for weakly supervised learning model, and the generalization capabilities of the matching networks is stronger. It also shows that our approach achieves real performance improvement over the other methods, which is consistent with the results in Table 1. Note that we mainly implement several representative methods previously mentioned based on the publicly available official codes [18–20]. The other results are achieved from corresponding literatures due to some belong to proprietary projects as shown in Tables 1, 3 and 4.

Ablation study. To be more convincing, we provide insightful ablation study by incorporating NCNet [18] and the proposed framework to reconstruct two new matching network frameworks. We analyse the accuracy performance of all proposed modules on the Proposal Flow-PASCAL benchmark [22]. We mainly implement on two variants of our approach: NCNet+OurLoss combines our semantic perception loss and the matching network used by [18], and NCNet+Ournetwork incorporates our base matching network and their loss function. Meanwhile, we evaluate the effectiveness of the proposed 3N-Network by constructing different variants of our matching network with different numbers of layers (NoL): Ours-Fo, Ours-Fi, and Ours-S respectively contain a four-layer, five-layer, and seven-layer convolution module. The results are presented in Table 2. Our approach and all its variants achieve consistently improvement over the NCNet [22]. This clearly shows that our proposed two modules are effective. This might be explained by the fact that our object-aware matching model mitigates some essential interference factors, and deals with images with the background clutter better.

Essentially, NCNet [18] determines that each correspondence with the maximum score is a positive match, which contains the correspondences in the background. our approach mainly selects the positive correspondences in the salient semantic regions using an object-aware strategy. NCNet+OurLoss utilizes our semantic perception loss to train their consensus network, which provides performance improvement over the original version. This verifies the effectiveness of the proposed object-aware matching mechanism. Experimental results also show that our six-layer convolution module achieves best performance over other variants. Six-layer structure is a critical point. Too shallow convolutional network cannot produce the desired results, and the parameters increase as the increase of the NoL and the performance cannot be effectively improved. Finally we construct a six-layer convolution model for correspondence task.

4.2.2. Results on Proposal Flow-WILLOW dataset

We implement the robustness evaluation on the Proposal Flow-WILLOW dataset [22]. This dataset is composed of 4 semantic categories, which is further divided into 10 subsets, for a total of approximately 900 image pairs for testing. And 10 keypoints are annotated as ground truth for each image. Besides, this dataset

Table 1
Quantitative results compared to state-of-the-art correspondence techniques on the Proposal Flow-PASCAL dataset [22].

Methods	mIoU	Mean PCK			Time (ms)
		$\theta = 0.05$	$\theta = 0.10$	$\theta = 0.15$	
UCNet [17]	0.502	0.241	0.493	0.621	>1000
PF-LOM [22]	0.511	0.242	0.451	0.640	>1000
CAT-FCSS [45]	0.591	0.270	0.472	0.646	–
CC-DCTM [46]	0.652	0.268	0.473	0.643	–
SCNet-AG+ [27]	0.534	0.362	0.722	0.820	>1000
CNNGeo-R [20]	0.579	0.403	0.693	0.846	40
End-to-End [19]	0.663	0.442	0.748	0.863	41
NCNet [18]	0.713	–	0.771	0.860	261
SFNet [42]	0.691	0.459	0.787	0.855	51
MaCoSNet [41]	0.739	0.487	0.790	0.881	>1000
Ours	**0.743**	**0.490**	**0.794**	**0.887**	158

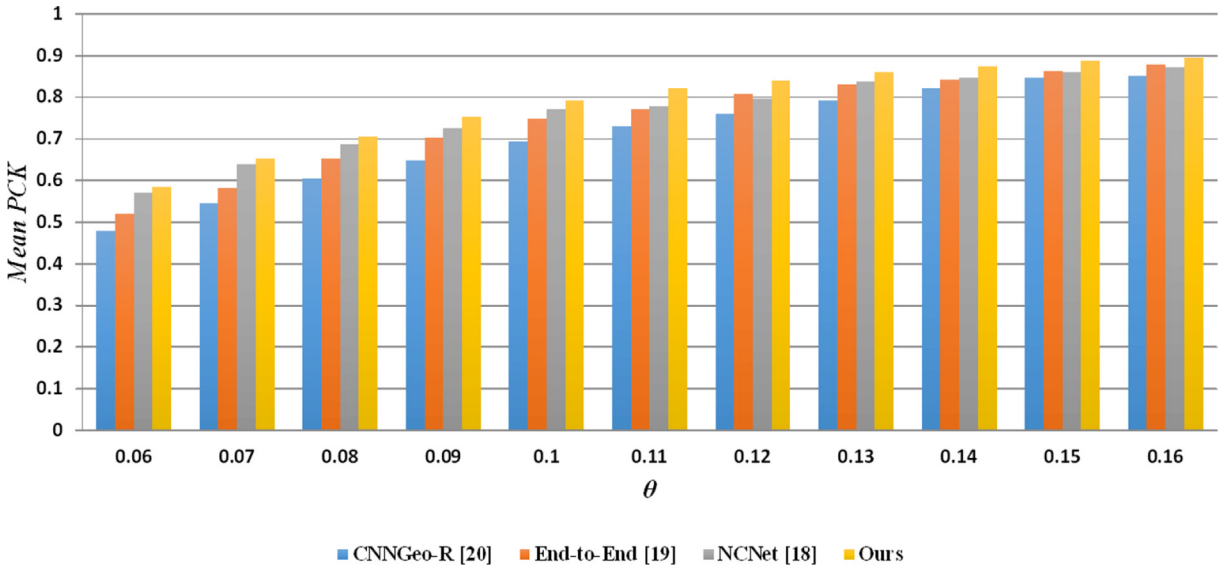

Fig. 5. Quantitative results compared to state-of-the-art correspondence techniques on the Proposal Flow-PASCAL dataset [22] with different tolerance factors: $\theta \in [0.06, 0.16]$.

Table 2
Ablation study experimental results on the Proposal Flow-PASCAL dataset [22].

Methods	NoL	mIoU	Mean PCK ($\theta = 0.1$)
NCNet [18]	3	0.713	0.771
NCNet+OurLoss	3	0.729	0.784
OurModel+NCNetLoss	6	0.725	0.785
Ours	**6**	**0.743**	**0.794**
Ours-Fo	4	0.702	0.777
Ours-Fi	5	0.728	0.792
Ours-S	7	0.735	0.790

contains more challenging examples such as intra-class variation, background, and viewpoint variations, which are suitable to evaluate the robustness of the matching methods. For the evaluation metric, we measure the PCK value between the ground truth keypoint and the transfered keypoint [27,30].

Results. We evaluate our matching framework and calculate the mean PCK with $\theta = 0.05, 0.10, 0.15$. As shown in Table 3, we provide comparisons to recent state-of-the-art methods for semantic feature matching. The results illustrate that our approach outperforms most other methods and is effective in cases of severe appearance and viewpoint variations.

Generally, the key of semantic feature matching is to accurately extract the salient feature representations from the main objects of interest in the examples with background clutter, and to handle the appearance differences to estimate the transformation model between the examples with viewpoint variations. The proposed matching modules exhibit superior performance. This might be explained by the fact that our salient foreground selection strategy narrows down the searching scope, and concentrates the candidates on the main semantic objects. This mitigates the background clutter, compared to the other methods. And, the cycle consistency constraint ensures one-to-one matching and outliers rejection. Overall, the proposed approach is more robust with respect to the camera, pose and appearance variations than the other works on semantic feature matching.

Threshold selection. A semantic perception loss function is designed to drive the training of our matching network with the requirement for a threshold τ. We evaluate its correctness and effectiveness on the Proposal Flow-PASCAL dataset and PF-WILLOW dataset [22] by measuring the mean PCK [27,30], as shown in Fig. 7. The mean PCK has a tendency to grow first and then flat, and finally decline. The smaller threshold leads to weaker mining and filtering capabilities of the matching layer for the candidate pairs, and larger number of feature representations. Further, more outliers are obtained and the matching tends to be redundant. Obviously, the saturation tends to be saturated at around 0.5, which gives a good compromise between matching accuracy and the number of the correspondences. The number of the feature representations is less and less when the threshold continues to increase, and the matched pairs are discretely distributed, which is not conducive to the performance evaluation. Finally we select a reasonable threshold $\tau = 0.5$.

4.2.3. Results on Caltch101 dataset

Lastly, we also evaluate our matching framework on the Caltch-101 dataset [34], which contains 101 semantic categories. We select 15 image pairs from each category, with a total of 1515 image pairs analogously to [27] in our experiments. There are keypoint annotations that can be used for semantic object segmentation, but not the matched keypoint pairs, which is not suitable to calculate the PCK. So we evaluate the quality of segmentation mask alignment using the following three metrics used by [21]: a) Label Transfer Accuracy (LT-ACC). b) Intersection-over-Union (IoU). c)

Table 3
Quantitative results compared to state-of-the-art correspondence techniques on the PF-WILLOW dataset [22].

Methods	Mean PCK		
	$\theta = 0.05$	$\theta = 0.10$	$\theta = 0.15$
UCNet [17]	0.291	0.417	0.513
PF-LOM [22]	0.284	0.568	0.682
CAT-FCSS [45]	0.311	0.579	0.725
CC-DCTM [46]	0.386	0.621	0.730
SCNet-AG+ [27]	0.386	0.704	0.853
CNNGeo-R [20]	0.448	0.777	0.899
End-to-End [19]	0.477	0.812	0.917
NCNet [18]	-	0.844	0.923
SFNet [42]	0.459	0.735	0.855
MaCoSNet [41]	**0.538**	0.854	0.939
Ours	0.534	**0.867**	**0.948**

Object Localization Error (LOC-ERR). Specifically, LT-ACC and IoU both evaluate the accuracy performance by measuring the degree of overlap between the warped objects using the estimated correspondences. However, LT-ACC identifies all the correctly aligned pixels, and IoU only considers the foreground region. And, the LOC-ERR measures the relative offset between the warped keypoint and corresponding point in the target image.

Results. Table 4 presents quantitative results on the Cal-tech-101 dataset [34]. We implement the comparisons to state-of-the-art correspondence methods under general settings. Note that the second and third columns indicate that the larger the value, the higher the accuracy, whereas the LOC-ERR metric values are the opposite. As can be seen, our approach achieves competitive performance on this benchmark, and outperforms most other correpondence methods.

4.2.4. Qualitative results

Qualitative evaluation is implemented on the Proposal Flow-PASCAL dataset [22] and Caltch-101 dataset [34]. The details are as follows.

Object-aware Visualization. We also implement the qualitative evaluation of the proposed object-aware matching mechanism on the Proposal Flow-PASCAL dataset [22]. In our experiments, a key observation that the learned CNN feature representations (positions) are uniformly distributed throughout the convolutional feature map as dense $H \times W$ grids. To effectively implement the evaluation, we first uniformly partition the input images into a grid of 16×16 cells. Then the vertices of grids are determined as the keypoints that are passed through the similarity measurer and filter network.

The qualitative results are presented in Fig. 6. As it can be observed, most of the candidate keypoints are distributed in the main semantic regions in Fig. 6(b), and the pairwise matching of the salient positions is approximately accurate. The experiments illustrate that our object-aware matching method achieves the expected results. However, the vertices of grids are directly matched from the source image to the target image, resulting in coordinate mapping error.

To effectively handle it, we adopt an interpolation mapping scheme to produce the sparse matched semantic features, as shown in Fig. 6(c). Analogously to the grid-based interpolation method [44], the coordinates of each keypoint are formed by linear interpolation with the enclosed vertices. The correspondences previously mentioned are first used to interpolate the annotated keypoints. Then a transformation model can be estimated based on these correspondences, as shown in Fig. 6(b). Further, the corresponding keypoints in the source image are transferred to the target image, and the warpped positions are established, as shown in Fig. 6(c). Compared to the recent state-of-the-art matching method [18], our results are approximately consistent with the ground-truth correspondences shown in Fig. 6(c) and (e), whereas visible errors are presented as shown in the red boxes in Fig. 6(d). These results verify the accuracy of our approach, and visualize the principle of calculating the coefficient of PCK. In addition, we also provide some qualitative results on the Caltch-101 dataset [34], as shown in Fig. 8. Specifically, both men and women belong to the same semantic category (persons) in the public benchmark datasets [22,34], which have similar semantics such as eyes, nose, and ears.

Table 4
Quantitative results compared to state-of-the-art correspondence techniques on the Caltech-101 dataset [34].

Methods	LT-ACC	IoU	LOC-ERR
UCN [17]	–	–	–
PF-LOM [22]	0.78	0.50	0.26
CAT-FCSS [45]	0.84	0.55	0.20
CC-DCTM [46]	0.85	0.56	0.21
SCNet-AG+ [27]	0.79	0.51	0.25
CNNGeo-R [20]	0.83	0.61	0.25
End-to-End [19]	0.85	0.63	0.24
NCNet [18]	0.87	0.69	0.21
SFNet [42]	0.88	0.67	–
MaCoSNet [41]	0.86	0.74	0.19
Ours	**0.90**	**0.73**	**0.17**

Image alignment. We consider that, most existing methods on semantic correspondence estimate the parameters of the transformation, typically homography, affine, or thin-plate spline transformation, which are used to globally deform and align the objects. However, it is sensitive to large viewpoint variations between the images, and ignores some salient details. Thus we utilize the notion of local deformation [44] to improve it. Analogously to the scheme previously mentioned, we first uniformly partition the 2D domain images into dense $C \times C$ grids, and each grid corresponds to an estimated transformation using the resulting correspondences. Pixels within the same grid are deformed using the same transformation. And then the original images are locally deformed and aligned to each other.

Following the same procedure as in [27,20], we mainly evaluate the quality of segmentation masks aligement on the Caltech-101 dataset [34] using three metrics previously mentioned: LT-ACC, IoU, and LOC-ERR. In addition, the evaluation is also performed on the Proposal Flow dataset [22] containing annotated matched keypoint pairs by calculating the average endpoint error (AEE), and mIoU. For image alignment, LT-ACC calculates the differences between the foreground mask of transfering source image to target image using dense correspondences and ground-truth segmentation mask, and counts the number of correctly annotated pixels. IoU (or mIoU) mainly focuses on the correctly aligned foreground annotations. Both of them measure the degree of overlap between the warped objects, which is based on foreground and background segmentation. Contrary to LT-ACC and IoU, the LOC-ERR metric prefers to concentrate on the details, and estimates the relative offsets between the positions of transferring the keypoints using dense correspondences and corresponding points in the target image. The AEE estimates the actual matching error between the matched keypoints, which is also considered to be a metric for image alignment.

Table 5 summarizes the alignment accuracy compared to recent state-of-the-art matching methods [45,46,27,20,19,18,42,41]. We implement image alignment on two variants of our approach: global ($Ours_G$) and local deformation ($Ours_L$) scheme. As it can be observed, both of our models improve the overall aearage score. Meanwhile, the local deformation model achieves good results on several public benchmark datasets and shows a significant improvement over most previously published results. Compared $Ours_G$ with other methods illustrates that our approach produces best correspondences. This is because the quality of image alignment is mainly determined by the matching accuracy, and strongly relies on the estimated transformation model.

In Fig. 9, we present the qualitative results on the Calte-ch-101 dataset [34]. As it can be observed, the proposed approach can produce good alignment results, which are close to the target image. The geometric pose between intra-class objects remains approximately consistent, such as the size of bounding box, orientation, and geometric shape.

5. Applications

Having established a sparse set of correspondences between a pair of images depicting the main objects of the same semantic cat-

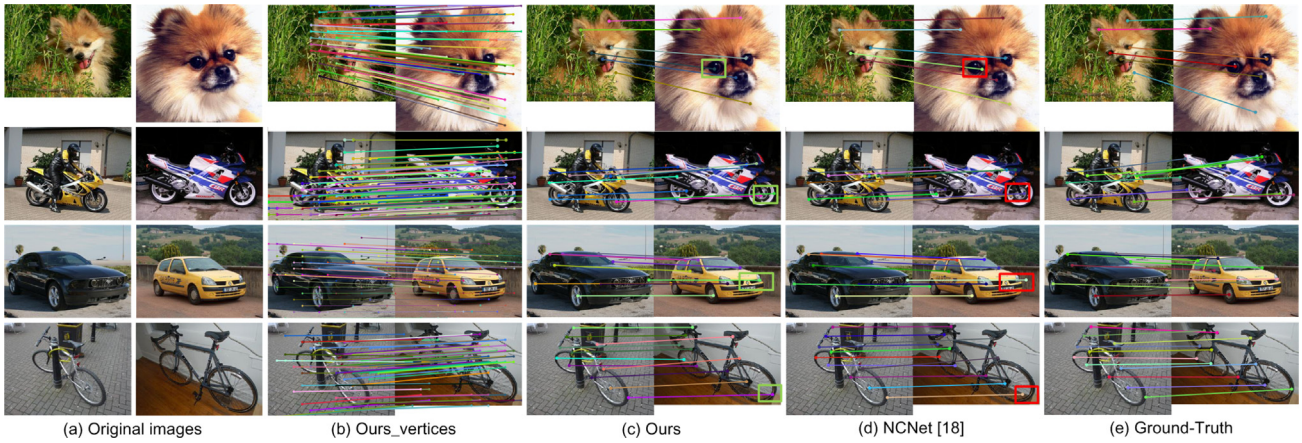

Fig. 6. Qualitative results on the Proposal Flow-PASCAL dataset [22]. (a) Original images are uniformly partitioned into a grid of 16×16 cells, and the vertices are determined as the keypoints that are passed through the trained matching model to produce the sparse matched keypoint pairs (b). (c) The sparse correspondences are obtained by interpolating the annotated keypoints with the dense vertices of grids. The results are obtained by using NCNet [18] (d).

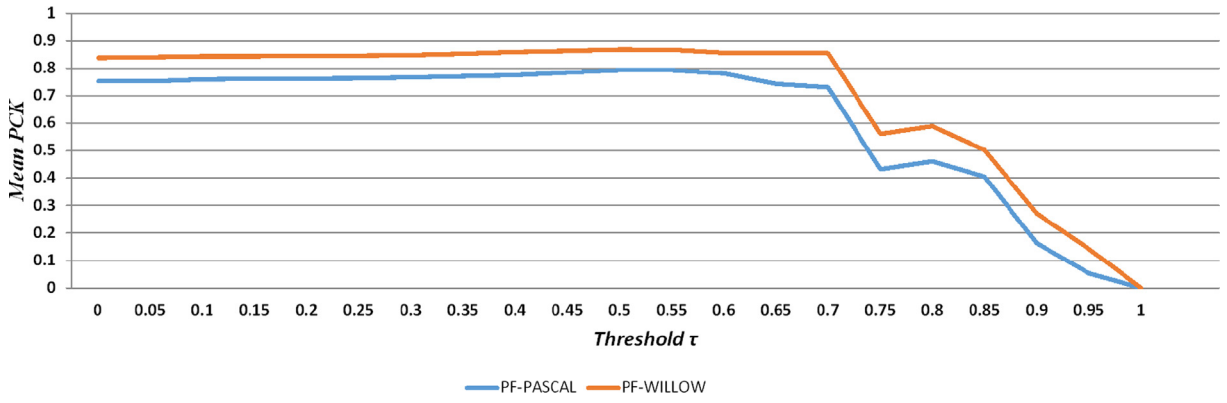

Fig. 7. Evaluation results of the selected thresholds on the Proposal Flow-PASCAL and Proposal Flow-WILLOW datasets [22].

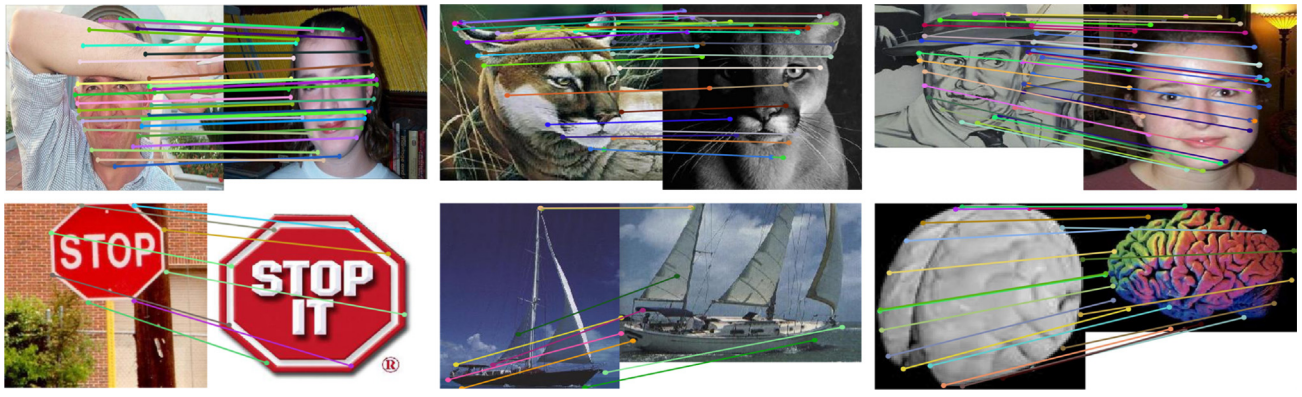

Fig. 8. Qualitative results on the Caltech-101 dataset [34].

egory, which can be generalized to guide the alignment of two overlapping images containing the same object, as well as to estimate a dense correspondence field between the two images. Actually, high-level semantics are more robust than low-level visual features for matching. This can facilitate a variety of graphics applications, one of which is discussed below.

Instance matching. We implement our approach on several sets of instance images depicting the same scene or object, as shown in Fig. 10. In our experiments, we observe that, position displacement would be produced by directly mapping a point across the convolutional layers. The matched keyponts are inversely mapped to the specific positions in the original images, as they are propagated through multiple convolutional feature pyramid layers. This further results in inaccurate locations on the low-level features and the loss of valuable information. To mitigate it, we take the output of our similarity filter as the input to a hierarchical inverse mapping process [15].

In Table 6, we compare our approach with the classical hand-crafted feature [10] by counting the number of the matched keypoint pairs. We implement on two variants of our approach:

Table 5
Quantitative results compared to state-of-the-art correspondence techniques on the Proposal Flow dataset [22] and Caltech-101 dataset [34].

Methods	Caltech-101			Proposal Flow-PASCAL		Proposal Flow-WILLOW
	LT-ACC	IoU	LOC-ERR	AEE	mIoU	AEE
CAT-FCSS [45]	0.84	0.55	0.20	–	0.591	–
CC-DCTM [46]	0.85	0.56	0.21	–	0.652	–
SCNet-AG+ [27]	0.79	0.51	0.25	22.8	0.534	19.2
CNNGeo-R [20]	0.83	0.61	0.25	22.3	0.579	18.9
End-to-End [19]	0.85	0.63	0.24	21.0	0.663	17.1
NCNet [18]	0.87	0.69	0.21	19.3	0.713	15.6
SFNet [42]	0.88	0.67	–	–	0.691	–
MaCoSNet [41]	0.86	0.74	0.19	-	0.739	–
$Ours_G$	0.88	0.70	0.20	18.9	0.733	15.2
$Ours_L$	**0.90**	**0.73**	**0.17**	**18.4**	**0.743**	**14.8**

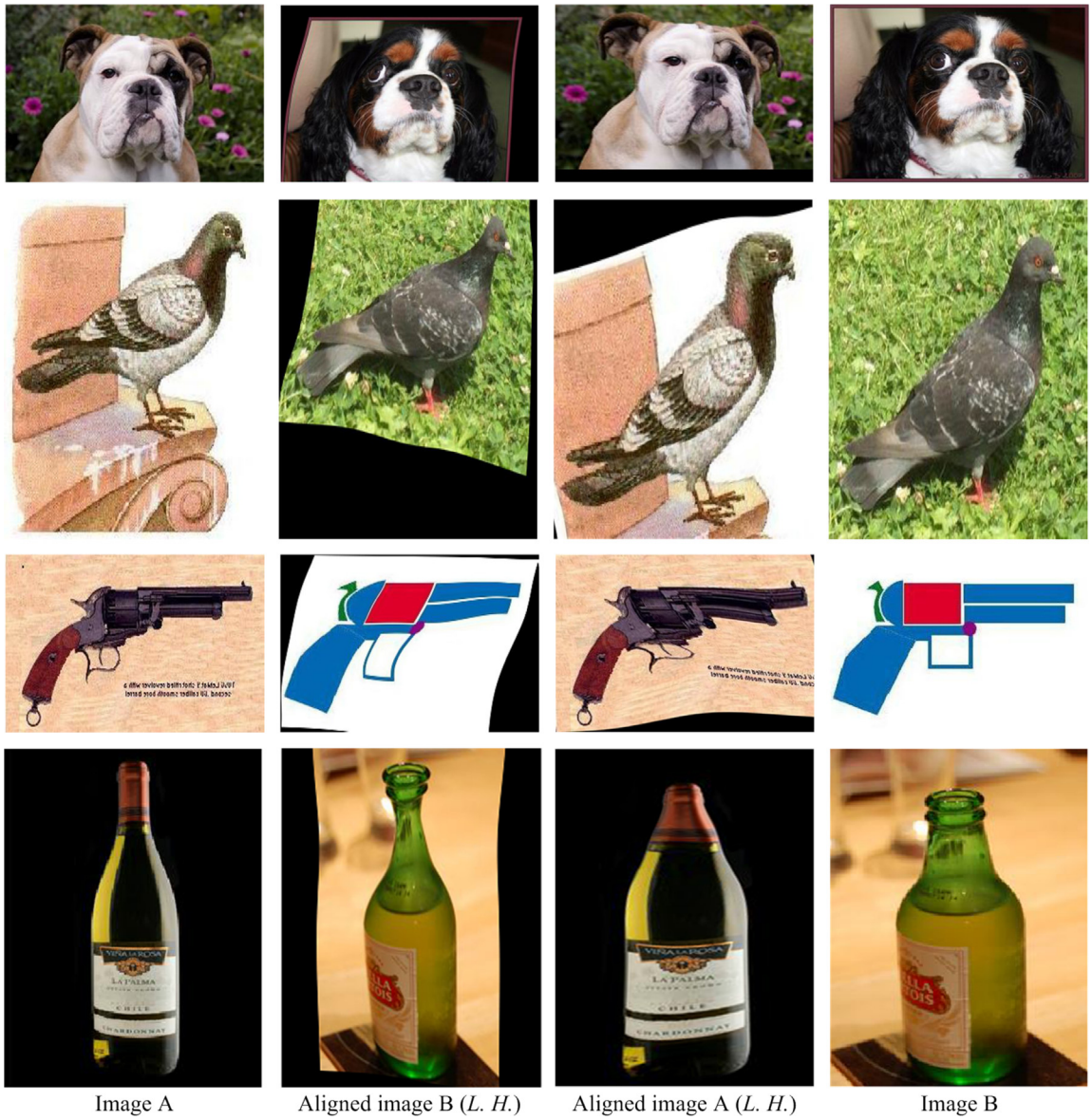

Fig. 9. Qualitative results on the Caltech-101 dataset [34]. The input images are locally deformed and aligned to each other according to the estimated homography transformation model.

Direct Mapping (DM) directly maps the candidate keypoints to original images across the convolutional layers, which is expanded through Hierarchical Mapping (HM) [15]. As it can be observe, our approach outperforms the method [10] in most cases. However, the opposite occurs in the "Park" case. As can be seen in Fig. 10 (d), no salient main objects appear across the image, which makes it difficult to search for the correspondences using CNN-based semantic correspondence model.

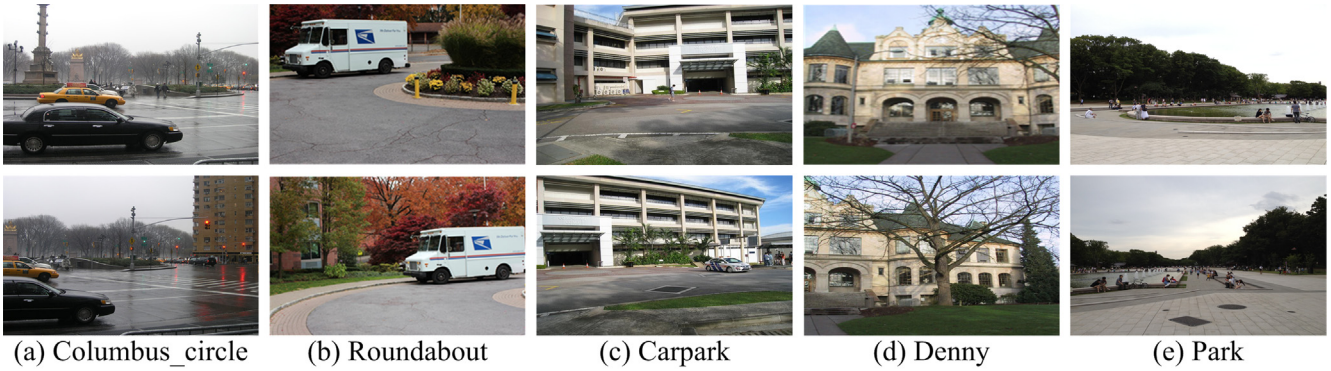

Fig. 10. Instance dataset. (a), (b) Salient foreground objects are presented in the image pairs. (c), (d) Artificial buildings occupy most of the image scene. (e) There is no main object of interest.

Table 6
Evaluation results on the overlapping image pairs.

Images	Resolution	DM	HM	SIFT [10]
Columbus-circle	640 × 480	81	371	197
Roundabout	1533 × 1022	65	345	293
Carpark	653 × 490	46	255	219
Denny	480 × 640	50	283	268
Park	1442 × 542	41	183	318

Qualitative results on the instance image pairs are presented in Fig. 12. The proposed approach is mainly suitable for the images with salient objects. They generally contain abundant high-level semantics such as edges, lines, and curves, with being robust to low-level visual features, e.g., color and illumination. The matched keypoints are mainly distributed throughout the objects shown in Fig. 12(b). On the contrary, the keypoints are discretely distributed in the image scene using the classical SIFT matching, as shown in Fig. 12(a). Besides, we align and stitch the corresponding images using the established correspondences, as shown in Fig. 11. Good synthesis results without ghosting or artifacts are obtained, and the images are aligned perfectly. This, in turn, verifies the accuracy of the correspondences.

Discussion. Some limitations still exist even though our approach establishes correct correspondences in many challenging cases. Our model has strong requirements for artificial scenes or salient foreground objects. But these still illustrate that the proposed approach can be effectively extended to solve some traditional graphics issues.

6. Conclusions

We have developed a semantic feature matching network framework, while being trainable end-to-end without the requirement for annotations. Our approach is based on an object-aware convolutional neural network architecture. The framework is

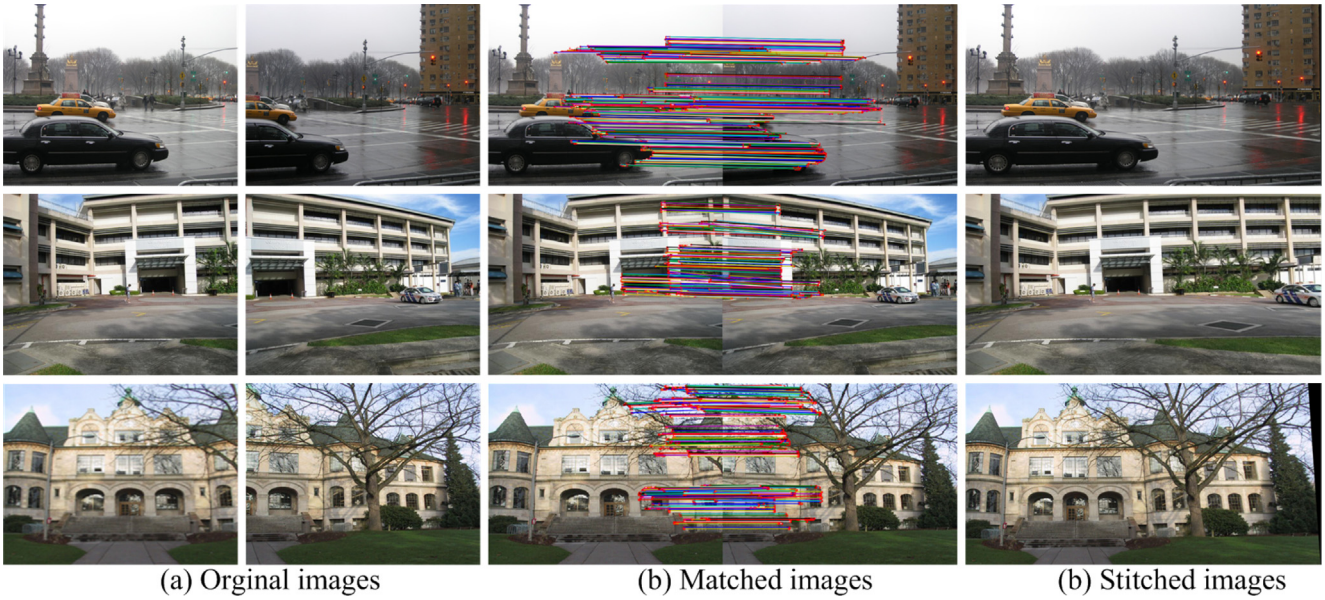

Fig. 11. Qualitative results on the instance images.

(a) SIFT [10]

(b) Ours (*HM*)

Fig. 12. Qualitative results on the instance images.

simple and effective, and achieves superior performance. Experiments have clearly shown that our approach outperforms most state-of-the-art methods for semantic feature matching on several standard benchmark datasets. Meanwhile, Extensive experiments illustrate that the proposed approach is also suitable for instance matching, obtaining confident results for some challenging instance images.

CRediT authorship contribution statement

Wei Lyu: Writing - original draft, Conceptualization, Writing - review & editing, Investigation, Methodology. **Lang Chen:** Software, Validation, Conceptualization. **Zhong Zhou:** Conceptualization, Supervision, Funding acquisition. **Wei Wu:** Conceptualization, Supervision, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work was supported in part by National Key R & D Program of China (Grant No. 2018YFB2100601), and in part by the National Natural Science Foundation of China (Grant No. 61872023).

References

[1] S. Richard, Image alignment and stitching: a tutorial, Foundations and Trends in Computer Graphics and Vision 2 (1) (2006) 1–104.

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

[3] R.-A. Newcombe, S.-J. Lovegrove, A.-J. Davison, DTAM: Dense tracking and mapping in real-time, in: Proceedings of the IEEE International Conference on Computer Vision, 2011, pp. 2320-2327.

[4] N. Snavely, S.-M. Seitz, R. Szeliski, Photo tourism: exploring photo collections in 3D, ACM Transactions on Graphics (SIGGRAPH Proceedings) 25 (3) (2006) 835–846.

[5] F. Zhang, F. Liu, Parallax-tolerant image stitching, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 3262–3269.

[6] T. Yu, J. Meng, J. Yuan, Multi-view harmonized bilinear network for 3D object recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 186–194.

[7] J. Si, H. Zhang, C. Li, J. Kuen, X. Kong, A.C. Kot, G. Wang, Dual attention matching network for context-aware feature sequence based person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5363–5372.

[8] B.K.P. Horn, B.G. Schunck, Determining optical flow, Artificial Intelligence 17 (1–3) (1981) 185–203.

[9] C. Liu, J. Yuen, A. Torralba, Sift flow: dense correspondence across scenes and its applications, IEEE Transactions on Pattern Analysis and Machine Intelligence 33 (5) (2011) 978–994.

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

[11] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2005, pp. 886–893.

[12] P. Fischer, A. Dosovitskiy, T. Brox, Descriptor matching with convolutional neural networks: a comparison to sift, arXiv preprint arXiv:1405.5769 2014..

[13] J. Long, N. Zhang, T. Darrell, Do convnets learn correspondence?, in: Advances in Neural Information Processing Systems, 2014, pp 1601–1609.

[14] N. Ufer, B. Ommer, Deep semantic feature matching, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 5929–5938.

[15] K. Aberman, J. Liao, M. Shi, D. Lischinski, B. Chen, D. Cohen-Or, Neural best-buddies: sparse cross-domain correspondence, ACM Transactions on Graphics (SIGGRAPH Proceedings) 37 (4) (2018) 1–14.

[16] J. Liao, Y. Yao, L. Yuan, G. Hua, S.B. Kang, Visual attribute transfer through deep image analogy, ACM Transactions on Graphics 36 (4) (2017) 1–15.

[17] C.B. Choy, J. Gwak, S. Savarese, M. Chandraker, Universal correspondence network, in: Advances in Neural Information Processing Systems, 2016, pp. 2406–2414.

[18] I. Rocco, M. Cimpoi, R. ArandjeloviĂG, A. Torii, T. Pajdla, J. Sivic, NCNet: neighbourhood consensus networks for estimating image correspondences, IEEE Transactions on Pattern Analysis and Machine Intelligence (2020) 1.

[19] I. Rocco, R. ArandjeloviĂG, J. Sivic, End-to-end weakly-supervised semantic alignment, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6917–6925.

[20] I. Rocco, R. ArandjeloviĂG, J. Sivic, Convolutional neural network architecture for geometric matching, IEEE Transactions on Pattern Analysis and Machine Intelligence 41 (11) (2018) 2553–2567.

[21] J. Kim, C. Liu, F. Sha, K. Grauman, Deformable spatial pyramid matching for fast dense correspondence, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 2307–2314.

[22] B. Ham, M. Cho, C. Schmid, J. Ponce, Proposal flow: semantic correspondences from object proposals, IEEE Transactions on Pattern Analysis and Machine Intelligence 40 (7) (2018) 1711–1725.

[23] F. Yang, X. Li, H. Cheng, J. Li, L. Chen, Object-aware dense semantic correspondence, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4151–4159.

[24] A. Krizhevsky, I. Sutskever, G. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in Neural Information Processing Systems, 2012, pp. 1097–1105..

[25] K. Simonyan, A. Vedaldi, A. Zisserman, Learning local feature descriptors using convex optimisation, IEEE Transactions on Pattern Analysis and Machine Intelligence 36 (8) (2014) 1573–1585.

[26] T. Dekel, S. Oron, M. Rubinstein, S. Avidan, W.-T. Freeman, Best-buddies similarity for robust template matching, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 2021–2029.

[27] K. Han, R.S. Rezende, B. Ham, K.K. Wong, M. Cho, C. Schmid, J. Ponce, SCNet: learning semantic correspondence, in: Proceedings of the International Conference on Computer Vision, 2017, pp. 1849–1858.

[28] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[29] J.W. Bian, W.Y. Lin, Y. Liu, L. Zhang, S.K. Yeung, M.M. Cheng, I. Reid, GMS: grid-based motion statistics for fast, ultra-robust feature correspondence, International Journal of Computer Vision 128 (2020) 1580–1593.

[30] T. Zhou, Y.J. Lee, S.X. Yu, A.A. Efros, Flowweb: Joint image set alignment by weaving consistent, pixel-wise correspondences, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1191–1200.

[31] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 248–255.

[32] PyTorch. http://pytorch.org/..

[33] D.P. Kingma, J. Ba, Adam: a method for stochastic optimization, in: Proceedings of the International Conference on Learning Representations, 2015.

[34] L. Fei-Fei, R. Fergus, P. Perona, One-shot learning of object categories, IEEE Transactions on Pattern Analysis and Machine Intelligence 28 (4) (2006) 594–611.

[35] W. Lyu, L. Chen, Z. Zhou, W. Wu, Deep semantic feature matching using confidential correspondence consistency, IEEE Access 8 (2020) 12802–12814.

[36] H. Bay, A. Ess, T. Tuytelaars, L. Van Gool, Speeded-up robust features (SURF), Computer Vision and Image Understanding 110 (3) (2008) 346–359.

[37] X. Wang, A. Jabri, A.A. Efros, Learning correspondence from the cycle-consistency of time, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 2561–2571.

[38] X. Huang, S. Belongie, Arbitrary style transfer in real-time with adaptive instance normalization, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 1510–1519.

[39] S. Jeon, S. Kim, D. Min, K. Sohn, PARN: pyramidal affine regression networks for dense semantic correspondence, Proceedings of the European Conference on Computer Vision (2018) 355–371.

[40] S. Kim, S. Lin, S. Jeon, D. Min, K. Sohn, Recurrent transformer networks for semantic correspondence, in: Advances in Neural Information Processing Systems, 2018, pp. 6129–6139.

[41] Y.C. Chen, Y.Y. Lin, M.H. Yang, J.B. Huang, Show, match and segment: joint weakly supervised learning of semantic matching and object co-segmentation, IEEE Transactions on Pattern Analysis and Machine Intelligence (2020) 1.

[42] J. Lee, D. Kim, J. Ponce, B. Ham, SFNet: learning object-aware semantic correspondence, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 2273–2282.

[43] K.-J. Hsu, Y.-Y. Lin, Y.-Y. Chuang, DeepCO3: deep instance co-segmentation by co-peak search and co-saliency detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 8838–8847.

[44] J. Zaragoza, T.-J. Chin, Q.-H. Tran, M.S. Brown, D. Suter, As-projective-as-possible image stitching with moving DLT, IEEE Transactions on Pattern Analysis and Machine Intelligence 36 (7) (2014) 1285–1298.

[45] S. Kim, D. Min, B. Ham, S. Lin, K. Sohn, FCSS: fully convolutional self-similarity for dense semantic correspondence, IEEE Transactions on Pattern Analysis and Machine Intelligence 41 (3) (2019) 581–595.

[46] S. Kim, D. Min, S. Lin, K. Sohn, Discrete-continuous transformation matching for dense semantic correspondence, IEEE Transactions on Pattern Analysis and Machine Intelligence 42 (1) (2020) 59–73.

[47] S.C. Pang, A. Du, M.A. Orgun, H.C. Chen, Weakly supervised learning for image keypoint matching using graph convolutional networks, Knowledge-Based Systems 197 (2020) 105871.

[48] Q. Zhou, W.B. Yang, G.W. Gao, W.H. Ou, H.M. Lu, J. Chen, L.J. Latecki, Multi-scale deep context convolutional neural networks for semantic segmentation, World Wide Web-Internet and Web Information Systems 22 (2) (2019) 555–570.

[49] H.M. Liu, Q.Q. Zhang, B. Fan, Z.H. Wang, J.W. Han, Features combined binary descriptor based on voted ring-sampling pattern, IEEE Transactions on Circuits and Systems for Video Technology 30 (10) (2020) 3675–3687.

[50] B. Fan, H. Liu, H. Zeng, J. Zhang, J. Han, Deep unsupervised binary descriptor learning through locality consistency and self distinctiveness, IEEE Transactions on Multimedia (2020) 1.

[51] B. Fan, Q.Q. Kong, B.Q. Zhang, H.M. Liu, C.M. Pan, J.W. Lu, Efficient nearest neighbor search in high dimensional hamming space, Pattern Recognition 99 (2020).

[52] X.Y. Jiang, J.Y. Ma, J.J. Jiang, X.J. Guo, Robust feature matching using spatial clustering with heavy outliers, IEEE Transactions on Image Processing 29 (2020) 736–746.

[53] J.Y. Ma, X.Y. Jiang, J.J. Jiang, J. Zhao, X.J. Guo, LMR: learning a two-class classifier for mismatch removal, IEEE Transactions on Image Processing 28 (8) (2019) 4045–4059.

[54] J.Y. Ma, J. Zhao, J.J. Jiang, H.B. Zhou, X.J. Guo, Locality preserving matching, International Journal of Computer Vision 127 (2019) 512–531.

[55] J.Y. Ma, J. Zhao, J.W. Tian, A.-L. Yuille, Z.W. Tu, Robust point matching via vector field consensus, IEEE Transactions on Image Processing 23 (4) (2014) 1706–1721.

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, 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 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, Beihang University, Beijing, China. His main research interest is computer vision and image processing, including image instance matching and image semantic matching.

Zhong Zhou 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 Ph.D. adviser with the State Key Lab 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.

Wei Wu received the PhD degree from Harbin Institute of Technology, Harbin, China, in 1995. He is currently a professor with the State Key Laboratory of Virtual Reality Technology and Systems with Beihang University. He is chair of the Technical Committee on Virtual Reality and Visualization, China Computer Federation. His current research interests include virtual reality, wireless networking, and distributed interactive system.