Deep Semantic Matching with Foreground Detection and Cycle-Consistency

Yun-Chun Chen[1,2], Po-Hsiang Huang[2], Li-Yu Yu[2], Jia-Bin Huang[3], Ming-Hsuan Yang[4,5], and Yen-Yu Lin[1]

[1]Academia Sinica   [2]National Taiwan University   [3]Virginia Tech
[4]University of California, Merced   [5]Google Cloud

Abstract. Establishing dense semantic correspondences between object instances remains a challenging problem due to background clutter, significant scale and pose differences, and large intra-class variations. In this paper, we address weakly supervised semantic matching based on a deep network where only image pairs without manual keypoint correspondence annotations are provided. To facilitate network training with this weaker form of supervision, we 1) explicitly estimate the foreground regions to suppress the effect of background clutter and 2) develop cycle-consistent losses to enforce the predicted transformations across multiple images to be geometrically plausible and consistent. We train the proposed model using the PF-PASCAL dataset and evaluate the performance on the PF-PASCAL, PF-WILLOW, and TSS datasets. Extensive experimental results show that the proposed approach performs favorably against the state-of-the-art methods. The code and model will be available at https://yunchunchen.github.io/WeakMatchNet/.

1   Introduction

Dense correspondence matching is an important and active research topic in computer vision. Optical flow estimation [1, 2] and stereo matching [3, 4] aim to estimate per-pixel correspondences to match across images depicting the same scene or object instance. While correspondence estimation has been extensively studied, there has been a growing trend to extend the idea of matching the same objects across images to matching images covering *different instances* of an object category. This progress not only attracts substantial attention but also facilitates many real-world applications ranging from object recognition [5], object co-segmentation [6–8], to 3D reconstruction [9]. However, due to the presence of background clutter, ambiguity induced by large intra-class variations, and the limited scalability of obtaining large-scale datasets with manually annotated correspondences, semantic matching remains challenging.

Conventional methods for semantic matching rely on hand-crafted descriptors such as SIFT [5] or HOG [10] as well as an effective geometric regularizer. However, these hand-crafted descriptors cannot be adapted to the given visual domains, leading to a sub-optimal performance of semantic matching. Driven by the recent success of convolutional neural networks (CNNs), several
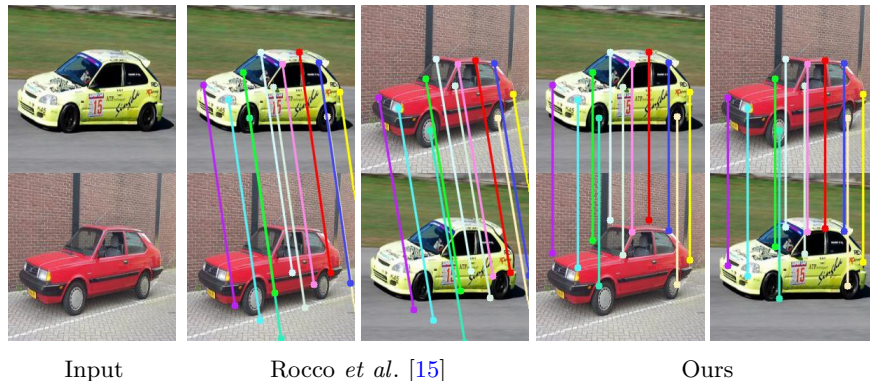
Input   Rocco *et al.* [15]   Ours

Fig. 1: **Comparisons with the state-of-the-art semantic matching algorithm [15].** Existing semantic matching methods often suffer from background clutter and may produce inconsistent matching results when swapping the source/target image. Through integrating foreground detection and cycle-consistent checking into semantic matching, our method produces more accurate and consistent matching results in both directions.

learning-based approaches have been proposed for tackling the problem of semantic matching [11–15]. While promising results have been shown, these approaches still suffer from the following limitations. The methods in [11–14] require a vast amount of supervised data for training the network. Collecting large-scale and diverse data, however, is expensive and labor-intensive. While weakly supervised methods such as [15] have been recently proposed to relax the issue, these approaches implicitly match the background features from both images to be similar. Thus, they often suffer from the unfavorable effect of background clutter.

In this paper, we address these challenges by performing foreground detection and enforcing cycle consistency constraints in semantic matching. To suppress the negative impacts caused by background clutter, we develop a foreground detection module that allows the model to exclude background regions and focus on matching the detected foreground regions. Therefore, the effect of background clutter can be alleviated. To address the matching difficulties caused by complex appearance and large intra-class variations, we focus on filtering out geometrically inconsistent correspondences. Our key insight is that correct correspondence should be *cycle-consistent* meaning that when matching a particular point from one image to the other and then performing reverse matching, we should arrive at the same spot. To exploit this property, we introduce a cycle-consistency loss that provides additional supervisory signals for network training. We further extend this idea to explore transitivity consistency across multiple images. We build upon the model by Rocco *et al.* [15] for a weakly-supervised and end-to-end trainable network and evaluate the effectiveness of the proposed approach on three benchmarks. Experimental results demonstrate

that our approach improves the baseline model [15], as shown in Fig. 1, and performs favorably against the state-of-the-art methods.

Our contributions are summarized as follows. First, we present a weakly-supervised learning framework that integrates foreground detection into semantic matching. With a module for explicit foreground detection, the proposed network suppresses the unfavorable effect of background clutter. Second, our model implicitly tackles the ambiguity induced by a vast matching space via inferring bi-directional geometric transformations during matching. With these transformations, we explicitly enforce the inferred geometric transformations to be cycle-consistent by introducing the *forward-backward consistency loss*. In addition, we explore the property of transitivity consistency and introduce the *transitivity consistency loss* to further enhance the matching performance. We train our network with the image pairs of the PF-PASCAL dataset [16]. We then evaluate the proposed model on several benchmark datasets for semantic matching, including the PF-PASCAL [16], PF-WILLOW [16], and TSS [7] datasets. Extensive comparisons with existing semantic matching algorithms demonstrate that the proposed approach achieves the state-of-the-art performance.

2   Related Work

Semantic matching has been extensively studied in the literature. Here, we review several related topics.

Semantic matching. Conventional methods for semantic matching [17–19] leverage hand-crafted descriptors such as SIFT [5] or HOG [10] along with geometric matching models. These methods find keypoint correspondences across images through energy minimization. The SIFT Flow [5] method aligns two images with SIFT features [5] using a similar formulation as an optical flow algorithm. Kim *et al.* [20] compute dense correspondence efficiently using the deformable spatial pyramid. Ham *et al.* [16] use the object proposals as the matching primitives and leverage the HOG descriptor to establish semantic correspondence. With the use of object proposals, the Proposal Flow method is robust to scaling and background clutter. Taniai *et al.* [7] propose a hierarchical Markov random field model to jointly perform object co-segmentation and dense correspondence. However, the aforementioned methods adopt hand-crafted descriptors, which are pre-defined and not optimized for the given images.

Semantic matching via deep learning. Convolutional neural networks have been successfully applied to semantic matching. Choy *et al.* [11] propose the universal correspondence network (UCN) and a correspondence contrastive loss for network training. The UCN method adopts a convolutional spatial transformer for feature transformations, making their method robust to scaling and rotations. Kim *et al.* [13] propose the fully convolutional self-similarity (FCSS) descriptor and integrate the descriptor into the Proposal Flow framework [16] for image matching. The SCNet [12] method learns a geometrically plausible

model for semantic correspondence by incorporating geometric consistency constraints into its loss function. While the methods in [12, 13] employ trainable descriptors for semantic correspondence, the feature matching is learned at the *object-proposal* level. Consequently, these methods are not end-to-end trainable since a fusion step is required to produce the final results. Rocco *et al.* [14] present an end-to-end trainable CNN architecture for estimating parametric geometric transformations. While these methods [11–14, 16] perform better than those based on hand-crafted features, the dependence on supervised training data with manually labeled keypoint correspondences limits the applicability.

Several recent CNN-based methods [15, 21, 22] have carried out weakly supervised semantic correspondence. The AnchorNet [21] learns a set of filters whose response is geometrically consistent across different object instances. The AnchorNet model, however, is not end-to-end trainable due to the use of the hand-engineered alignment model. The WarpNet [22] learns fine-grained image matching with small-scale and pose variations via aligning objects across images through known deformation. Inspired by the inlier scoring procedure of RANSAC, Rocco *et al.* [15] propose an end-to-end trainable alignment network which computes dense semantic correspondence while aligning two images.

Our proposed method differs from these methods [15, 21] in two aspects. First, our approach further takes into account foreground detection. Our network learns feature embedding to enhance inter-image foreground similarity while alleviating the unfavorable effects caused by complex background. Second, our model simultaneously infers bi-directional transformations. We explicitly enforce cycle-consistent constraints on the predicted transformations, resulting in more accurate and consistent matching results.

Cycle consistency. Exploiting cycle consistency property to regularize learning has been extensively studied. In the context of motion analysis, computing bi-directional optical flow has been shown to be useful to reason about occlusion for learning optical flow [23, 24] and enforcing temporal consistency [25, 26]. In the context of image-to-image translation, enforcing cycle consistency enables learning mapping between domains with unpaired data [27, 28]. Several methods exploit the idea of cycle consistency for semantic matching. Zhou *et al.* [29] tackle the problem of matching multiple images by jointly optimizing feature matching and enforcing cycle consistency. The FlowWeb [30] method learns image alignment by establishing globally-consistent dense correspondences with cycle consistency constraints. However, these methods [29,30] employ hand-engineered descriptors which cannot adapt to an arbitrary object category given for matching. Zhou *et al.* [31] establish dense correspondences by using an additional 3D CAD model to form a cross-instance loop between synthetic data and real images. However, the cycle consistency loss in [31] requires four images at a time. In contrast, we develop two loss functions to enforce cycle consistency and do not need additional data to guide the training. Experimental results demonstrate that by exploiting cycle consistency constraints, the proposed method produces consistent matching results and improves the performance.
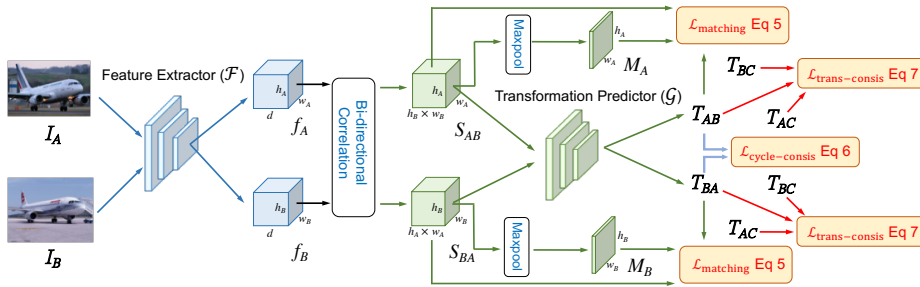
Fig. 2: **Semantic matching network.** Our model is composed of two CNN modules, including a feature extractor $\mathcal{F}$ for extracting features and a transformation predictor $\mathcal{G}$ for estimating the geometric transformations between a given image pair. We train the model with three loss functions, including the foreground-guided matching loss $\mathcal{L}_{\text{matching}}$, the forward-backward consistency loss $\mathcal{L}_{\text{cycle}-\text{consis}}$, and the transitivity consistency loss $\mathcal{L}_{\text{trans}-\text{consis}}$ (if given 3 input images).

3   Proposed Algorithm

In this section, we first provide an overview of our approach. We then describe each loss in our objective function in detail and the implementation details.

3.1   Framework overview

Let $\mathcal{D} = \{I_i\}_{i=1}^{N}$ denote a set of images consists of instances of the same object category, where I_i is the ith image and N is the number of images. Our goal is to learn a CNN-based model that can estimate the keypoint correspondences between each image pair (I_A, I_B) in $\mathcal{D}$ *without knowing the object class in advance.* Our formulation for semantic matching is *weakly-supervised* since training our model requires only weak image-level supervision in the form of training image pairs containing common objects. No ground truth keypoint correspondences are used.

To accomplish this task, we present an end-to-end trainable network which is composed of two modules: 1) the feature extractor $\mathcal{F}$ and 2) the transformation predictor $\mathcal{G}$. The feature extractor $\mathcal{F}$ extracts features for each image in a given image pair. The transformation predictor $\mathcal{G}$ predicts the transformation that warps an image so that the warped image can better align the other image.

As shown in Fig. 2, the proposed network architecture takes an image pair as input. For a given image pair (I_A, I_B), we use the feature extractor $\mathcal{F}$ to extract their feature maps f_A and f_B, respectively. We then compute correlation from f_A to f_B to generate the correlation map S_{AB}. The other correlation map S_{BA} is symmetrically obtained. The transformation predictor $\mathcal{G}$ then estimates the geometric transformation T_{AB} which warps I_A so that the warped image $\tilde{I}_A$ can align I_B. In the following section, we describe our objective function used

to optimize the feature extractor $\mathcal{F}$ and the transformation predictor $\mathcal{G}$. After optimizing the objective function, the matching between an image pair (I_A, I_B) can be performed via the predicted transformation T_{AB} or T_{BA}.

3.2   Objective function

The overall training objective consists of three loss functions. First, the foreground-guided matching loss $\mathcal{L}_{\text{matching}}$ minimizes the distance between the corresponding features based on the estimated geometric transformations. Unlike existing semantic matching methods [14, 15], our model predicts foreground masks to suppress the effect of background clutter by excluding background matching. Second, the forward-backward consistency loss $\mathcal{L}_{\text{cycle-consis}}$ and the transitivity consistency loss $\mathcal{L}_{\text{trans-consis}}$ enforce the predicted transformations across multiple images to be geometrically plausible and consistent. Both losses regularize the network training. Specifically, our training objective is

$$\mathcal{L} = \mathcal{L}_{\text{matching}} + \lambda_C \cdot \mathcal{L}_{\text{cycle-consis}} + \lambda_T \cdot \mathcal{L}_{\text{trans-consis}}, \tag{1}$$

where λ_C and λ_T are hyper-parameters used to control the relative importance of the respective loss functions. Below we outline the details of each loss function.

3.3   Foreground-guided matching loss

To reduce the effect of background clutter and enforce only foreground regions to be similar, our model minimizes the foreground-guided matching loss $\mathcal{L}_{\text{matching}}$. Given an image pair (I_A, I_B), the feature extractor $\mathcal{F}$ extracts their respective feature maps $f_A \in \mathbb{R}^{h_A \times w_A \times d}$ and $f_B \in \mathbb{R}^{h_B \times w_B \times d}$, where d is the number of channels. We correlate f_A with f_B to generate the correlation map $S_{AB} \in \mathbb{R}^{h_A \times w_A \times h_B \times w_B}$. Each element $S_{AB}(i, j, s, t) = S_{AB}(\mathbf{p}, \mathbf{q})$ records the normalized inner product between the feature vectors stored at two spatial locations $\mathbf{p} = [i, j]^\top$ in f_A and $\mathbf{q} = [s, t]^\top$ in f_B. The other correlation map $S_{BA} \in \mathbb{R}^{h_B \times w_B \times h_A \times w_A}$ can be computed symmetrically. The correlation map S_{AB} is reshaped to a three-dimensional tensor with dimensions h_A, w_A, and $(h_B \times w_B)$, i.e., $S_{AB} \in \mathbb{R}^{h_A \times w_A \times (h_B \times w_B)}$. As such, the reshaped correlation map S_{AB} can be interpreted as a dense $h_A \times w_A$ grid with $(h_B \times w_B)$-dimensional local features. We apply the reshaping operation to S_{BA} as well. With the reshaped S_{AB}, we use the transformation predictor $\mathcal{G}$ [14] to estimate a geometric transformation T_{AB} which warps I_A to $\tilde{I}_A$ so that $\tilde{I}_A$ aligns well to I_B.

Since the correlation map $S_{AB}(\mathbf{p}, \mathbf{q})$ records the normalized inner product between two feature vectors located at $\mathbf{p}$ in f_A and $\mathbf{q}$ in f_B. Our model estimates the foreground mask $M_A \in \mathbb{R}^{h_A \times w_A}$ by

$$M_A(\mathbf{p}) = \max_{\mathbf{q}}(S_{AB}(\mathbf{p}, \mathbf{q})). \tag{2}$$

Note that both the correlation maps S_{AB} and S_{BA} are compiled through a rectified linear unit (ReLU) to eliminate negative matching values in advance.

Therefore, the value of the estimated foreground masks at each pixel will be bounded between 0 and 1. Intuitively, the mask $M_A(\mathbf{p})$ has a low value (i.e., location $\mathbf{p}$ is likely to belong to background) if none of the feature vectors in f_B matches well with $f_A(\mathbf{p})$. The mask M_B can be obtained following a similar procedure.

With the estimated geometric transformation T_{AB}, we can identify and filter out geometrically inconsistent correspondences. Consider a correspondence with endpoints $(\mathbf{p} \in \mathcal{P}_A, \mathbf{q} \in \mathcal{P}_B)$, where $\mathcal{P}_A$ and $\mathcal{P}_B$ are the sets of all spatial coordinates of f_A and f_B, respectively. The distance $\|T_{AB}(\mathbf{p}) - \mathbf{q}\|$ represents the projection error of this correspondence with respect to transformation T_{AB}. Following Rocco $et\ al.$ [15], we introduce a correspondence mask m_A to determine if the correspondences are geometrically consistent with transformation T_{AB}. Specifically, m_A is of the form

$$m_A(\mathbf{p}, \mathbf{q}) = \begin{cases} 1, & \text{if } \|T_{AB}(\mathbf{p}) - \mathbf{q}\| \leq \varphi, \\ 0, & \text{otherwise.} \end{cases}, \text{ for } \mathbf{p} \in \mathcal{P}_A \text{ and } \mathbf{q} \in \mathcal{P}_B, \qquad (3)$$

where $\varphi = 1$ is the number of pixels.

Given the geometric transformation T_{AB} and the correspondence mask m_A, we compute matching score of each spatial location $\mathbf{p} \in \mathcal{P}_A$ as

$$s_A(\mathbf{p}) = \sum_{\mathbf{q} \in \mathcal{P}_B} m_A(\mathbf{p}, \mathbf{q}) \cdot S_{AB}(\mathbf{p}, \mathbf{q}). \qquad (4)$$

To suppress the effect of background clutter, we incorporate the estimated foreground masks to focus on matching the detected foreground regions. We define the foreground-guided matching loss $\mathcal{L}_{\text{matching}}$ as

$$\mathcal{L}_{\text{matching}}(I_A, I_B; \mathcal{F}, \mathcal{G}) = -\left(\sum_{\mathbf{p} \in \mathcal{P}_A} s_A(\mathbf{p}) \cdot M_A(\mathbf{p}) + \sum_{\mathbf{q} \in \mathcal{P}_B} s_B(\mathbf{q}) \cdot M_B(\mathbf{q}) \right). \ (5)$$

Note that the negative sign in (5) is used in the objective function, since maximizing the matching score corresponds to minimizing the foreground-guided matching loss $\mathcal{L}_{\text{matching}}$.

3.4   Cycle consistency

For a pair of images I_A and I_B, the transformation predictor $\mathcal{G}$ estimates a geometric transformation T_{AB} which maps pixel coordinates from I_A to I_B. However, the large capacity of the transformation predictor $\mathcal{G}$ often leads to a circumstance where various transformations can warp I_A to $\tilde{I}_A$ such that $\tilde{I}_A$ aligns I_B very well. This phenomenon implies that using the foreground-guided matching loss $\mathcal{L}_{\text{matching}}$ alone is insufficient to reliably train the transformation predictor $\mathcal{G}$ in the weakly supervised setting since no ground truth correspondences are available to constrain the learning of predicting transformations. We address this issue by simultaneously estimating T_{AB} and T_{BA} and enforce the

predicted transformations to be geometrically plausible and consistent across multiple images. As such, exploiting the cycle consistency constraint greatly reduces the feasible space of transformations and can serve as a regularization term in training the transformation predictor $\mathcal{G}$. To this end, we develop two loss functions where cycle-consistency checking is performed in conjunction with the proposed method such that the model is end-to-end trainable. The developed loss functions are described in the following.

Forward-backward consistency loss. Consider the correlation maps S_{AB} and S_{BA} generated from images I_A and I_B. The forward consistency states that property $T_{BA}(T_{AB}(\mathbf{p})) \approx \mathbf{p}$ holds for any $\mathbf{p} \in \mathcal{P}_A$. By the same token, the backward consistency means $T_{AB}(T_{BA}(\mathbf{q})) \approx \mathbf{q}$ for any $\mathbf{q} \in \mathcal{P}_B$. The resultant forward-backward consistency loss $\mathcal{L}_{\text{cycle−consis}}$ is then defined by

$$
\begin{aligned}
\mathcal{L}_{\text{cycle−consis}}(I_A, I_B; \mathcal{F}, \mathcal{G}) = &\sum_{\mathbf{p} \in \mathcal{P}_A} \|T_{BA}(T_{AB}(\mathbf{p})) - \mathbf{p}\| \\
&+ \sum_{\mathbf{q} \in \mathcal{P}_B} \|T_{AB}(T_{BA}(\mathbf{q})) - \mathbf{q}\|,
\end{aligned}
\tag{6}
$$

where $\|T_{BA}(T_{AB}(\mathbf{p})) - \mathbf{p}\|$ is the reprojection error between coordinate $\mathbf{p}$ and the reprojected coordinate $T_{BA}(T_{AB}(\mathbf{p}))$.

Transitivity consistency loss. We further extend the forward-backward consistency between a pair of images to the transitivity consistency across multiple images. Considering the case of three images I_A, I_B, and I_C, we first estimate three geometric transformations T_{AB}, T_{BC}, and T_{AC}. Transitivity consistency in this case states that the coordinate transformation from I_A to I_C should be path invariant. That is, for any coordinate $\mathbf{p} \in \mathcal{P}_A$, the property, $T_{BC}(T_{AB}(\mathbf{p})) \approx T_{AC}(\mathbf{p})$, holds. We can thus introduce the transitivity consistency loss $\mathcal{L}_{\text{trans−consis}}$ as

$$
\begin{aligned}
\mathcal{L}_{\text{trans−consis}}(I_A, I_B, I_C; \mathcal{F}, \mathcal{G}) = &\sum_{\mathbf{p} \in \mathcal{P}_A} \|T_{BC}(T_{AB}(\mathbf{p})) - T_{AC}(\mathbf{p})\| \\
&+ \sum_{\mathbf{q} \in \mathcal{P}_B} \|T_{AC}(T_{BA}(\mathbf{q})) - T_{BC}(\mathbf{q})\|.
\end{aligned}
\tag{7}
$$

3.5   Network selection and initialization

We adopt the semantic matching network proposed by Rocco *et al.* [15] as our feature extractor $\mathcal{F}$ due to its state-of-the-art performance for image alignment. The network employs the ResNet-101 [32] model. The extracted features are those generated by layer `conv4-23`. For the transformation predictor $\mathcal{G}$, we use the same architecture as that in [14]. The transformation predictor $\mathcal{G}$ is a cascade of two modules predicting an affine transformation and a thin plate spline (TPS)

transformation. Given an image pair, the model first estimates an affine transformation with 6 degrees of freedom to obtain a rough alignment. The model then performs a second-stage geometric estimation based on the roughly aligned image pair to predict TPS transformation for alignment refinement. Similar to Rocco *et al.* [14], we use a uniform 3×3 grid of control points for TPS, which corresponds to $3 \times 3 \times 2 = 18$ degrees of freedom. We initialize the feature extractor $\mathcal{F}$ and the transformation predictor $\mathcal{G}$ from the parameters pre-trained in [15] and fine-tune the feature extractor $\mathcal{F}$ and the transformation predictor $\mathcal{G}$ by using the proposed objective function. We note that there may exist degenerate solutions to the foreground-guided matching loss $\mathcal{L}_{\text{matching}}$ since no annotated correspondences are used to guide the network training. In this work, we build our model upon Rocco *et al.* [15], which is pre-trained on a large-scale synthetic dataset. The pre-trained model provides good enough initialization for predicting the geometric transformations, reducing the chance of falling into degenerate solutions. In addition, the foreground-guided matching loss $\mathcal{L}_{\text{matching}}$ and the cycle-consistency losses work jointly. The three adopted loss terms regularize the network training and avoid degenerate solutions.

4   Experimental Results

Experiments are conducted in this section. Here, we first describe the implementation details and the experimental setting. We evaluate and compare the proposed approach with the state-of-the-art, following analyzing the relative contributions of individual components in the proposed model.

4.1   Implementation details

We implement our model using PyTorch. We use the training and validation image pairs from the PF-PASCAL dataset [16]. All images are resized to the resolution of 240×240. We perform data augmentation by horizontal flipping, random cropping the input images, and swapping the order of images in the image pair. We train our model using the ADAM optimizer [33] with an initial learning rate of 5×10^{-8}. For transitivity consistency loss, the input triplets are randomly selected within a mini-batch. We sample $10 \times 10 = 100$ spatial coordinates for computing the forward-backward consistency loss and the transitivity consistency loss. The training process takes about 2 hours on a single NVIDIA GeForce GTX 1080 GPU.

4.2   Evaluation metric and datasets

We conduct the evaluation on the PF-PASCAL [16], PF-WILLOW [16], and TSS [7] benchmark datasets.

Table 1: **Per-class PCK on the PF-PASCAL dataset with $\tau = 0.1$.**

Method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	d.table	dog	horse	moto	person	plant	sheep	sofa	train	tv	mean
HOG+PF-LOM [16]	73.3	74.4	54.4	50.9	49.6	73.8	72.9	63.6	46.1	79.8	42.5	48.0	68.3	66.3	42.1	62.1	65.2	57.1	64.4	58.0	62.5
UCN [11]	64.8	58.7	42.8	59.6	47.0	42.2	61.0	45.6	49.9	52.0	48.5	49.5	53.2	72.7	53.0	41.4	83.3	49.0	**73.0**	66.0	55.6
VGG-16+SCNet-A [12]	67.6	72.9	69.3	59.7	74.5	72.7	73.2	59.5	51.4	78.2	39.4	50.1	67.0	62.1	**69.3**	68.5	78.2	63.3	57.7	59.8	66.3
VGG-16+SCNet-AG [12]	83.9	81.4	70.6	62.5	60.6	81.3	81.2	59.5	53.1	81.2	**62.0**	58.7	65.5	73.3	51.2	58.3	60.0	69.3	61.5	**80.0**	69.7
VGG-16+SCNet-AG+ [12]	85.5	84.4	66.3	70.8	57.4	82.7	82.3	71.6	**54.3**	**95.8**	55.2	59.5	**68.6**	75.0	56.3	60.4	60.0	**73.7**	66.5	76.7	72.2
VGG-16+CNNGeo [14]	79.5	80.9	69.9	61.1	57.8	77.1	84.4	55.5	48.1	83.3	37.0	54.1	58.2	70.7	51.4	41.4	60.0	44.3	55.3	30.0	62.6
ResNet-101+CNNGeo(S) [14]	83.0	82.2	81.1	50.0	57.8	79.9	92.8	77.5	44.7	85.4	28.1	69.8	65.4	77.1	64.0	65.2	**100.0**	50.8	44.3	54.4	69.5
ResNet-101+CNNGeo(W) [15]	84.7	88.9	80.9	55.6	76.6	89.5	**93.9**	79.6	52.0	85.4	28.1	71.8	67.0	75.1	66.3	70.5	**100.0**	62.1	62.3	61.1	74.8
Ours	**85.6**	**89.6**	**82.1**	**83.3**	**85.9**	**92.5**	**93.9**	**80.2**	52.2	85.4	55.2	**75.2**	64.0	**77.9**	67.2	**73.8**	**100.0**	65.3	69.3	61.1	**78.0**

Evaluation metric. We evaluate the performance of the proposed method on a semantic correspondence task. To assess the performance, we adopt the percentage of correct keypoints (PCK) metric [34] which measures the percentage of keypoints whose reprojection errors are below the given threshold. The reprojection error is the Euclidean distance $d(\phi(\mathbf{p}), \mathbf{p}^*)$ between the locations of the warped keypoint $\phi(\mathbf{p})$ and the ground truth keypoint $\mathbf{p}^*$. The threshold is defined as $\tau \cdot \max(h, w)$ where h and w are the height and width of the annotated object bounding box on the image, respectively.

PF-PASCAL [16]. The PF-PASCAL dataset is selected from the PASCAL 2011 keypoint annotations [35] containing 1,351 semantically related image pairs from 20 object categories. For images of a category, they contain different object instances of that category with similar poses but different appearances. In addition, the presence of background clutter makes it a challenging dataset on semantic matching. We divide the dataset into 735 pairs for training, 308 pairs for validation, and 308 pairs for testing. Manually annotated correspondences are provided for each image pairs. However, under the weakly supervised setting, we do not use the keypoint annotations for training. The annotations are used only for evaluation. We compute the PCK for each object category with τ equals to 0.1.

PF-WILLOW [16]. The PF-WILLOW dataset is composed of 100 images with 900 image pairs divided into four semantically related subsets: car, duck, motorbike, and wine bottle. Each subset contains images with large intra-class variations and background clutters. For each image, there are 10 keypoint annotations. We follow Han *et al.* [12] and compute the PCK at three different thresholds with τ equals to 0.05, 0.1, and 0.15, respectively.

TSS [7]. The TSS dataset comprises 400 semantically related image pairs divided into three groups, including FG3DCar, JODS, and PASCAL. FG3DCar contains 195 image pairs of automobiles. JODS is composed of 81 image pairs of airplanes, cars, and horses. There are 124 image pairs of trains, cars, buses, bikes, and motorbikes form the group of PASCAL. Ground truth flows for each image pair are provided. Following Taniai *et al.* [7], we compute the PCK over foreground object by setting τ to 0.05.

Table 2: **Ablation experiments on PF-PASCAL with $\tau = 0.1$.**

Method	mean
Rocco *et al.* [15]	74.8
Rocco *et al.* [15] + $\mathcal{L}_{\mathrm{matching}}$	75.5
Rocco *et al.* [15] + $\mathcal{L}_{\mathrm{cycle-consis}}$	77.4
Rocco *et al.* [15] + $\mathcal{L}_{\mathrm{trans-consis}}$	77.6
Ours	**78.0**
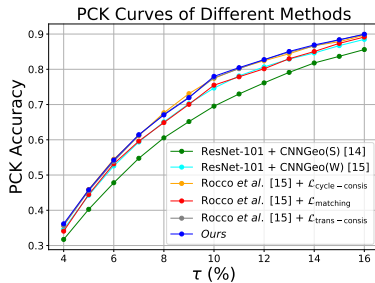
Fig. 3: **Semantic correspondence results on the PF-PASCAL dataset.** The matched coordinates are linked with color lines.

4.3 Experimental results on the PF-PASCAL dataset

In the following, we compare the performance of the proposed method with the state-of-the-art approaches. Note that many of the existing methods require manually annotated correspondences while our model can be trained using only image-level supervision.

Performance evaluation. We compare our method with the Proposal Flow [16], the UCN [11], different versions of the SCNet [12], the CNNGeo [14] with different feature extractors, and a weakly supervised approach proposed by Rocco *et al.* [15]. Table 1 presents the experimental results for the PF-PASCAL dataset. Our results show that the proposed approach compares favorably against state-of-the-art methods, achieving an overall PCK of 78.0% (outperforming the previous best method [15] by 3.2%). The advantage of incorporating foreground detection and enforcing cycle consistency constraints can be observed by comparing our method with ResNet-101+CNNGeo(W) [15] since both methods utilize the same feature descriptor and are trained with image-level supervision only.

Fig. 3 presents the qualitative results of semantic correspondence on the PF-PASCAL dataset. To further highlight the importance of each component of the proposed method, we present an ablation study of our method.

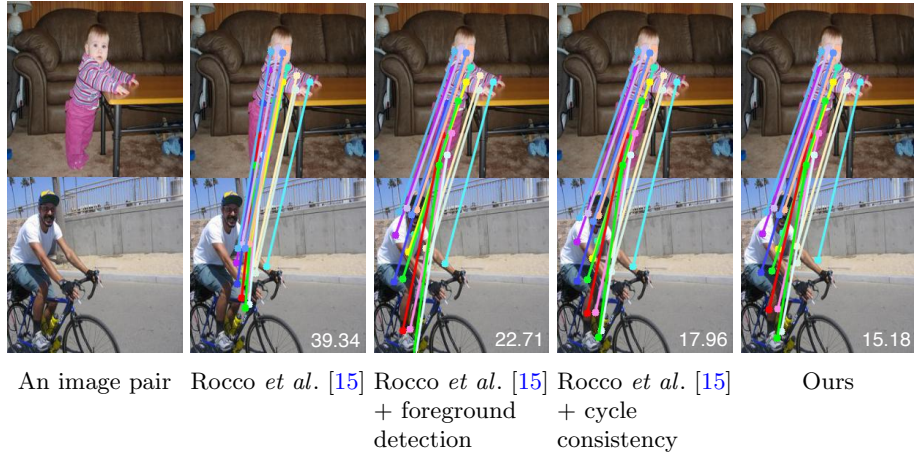| An image pair | Rocco *et al.* [15] | Rocco *et al.* [15]
+ foreground
detection | Rocco *et al.* [15]
+ cycle
consistency | Ours |

Fig. 4: **Visualization of the effect of each component.** Given an image pair, existing methods often suffer from the negative impacts due to background clutter. Integrating foreground detection into semantic matching alleviates the unfavorable effects of background clutters. Enforcing cycle consistency improves the matching result. Our method integrates foreground detection and exploits cycle-consistency property in semantic matching, resulting in more accurate results. The bottom right corners display the errors, namely the average distances between the predicted keypoints and the ground truth.

Ablation study. To analyze the importance of each loss function, we conduct ablation experiments on the PF-PASCAL [16] dataset. Table 2 presents the mean PCK value of the variants of our approach evaluated on the PF-PASCAL dataset with τ equals to 0.1. Our results show that both $\mathcal{L}_{\mathrm{cycle-consis}}$ and $\mathcal{L}_{\mathrm{trans-consis}}$ substantially improve the performance when comparing with Rocco *et al.* [15]. We visualize the effect of each component in Fig. 4. To demonstrate the effectiveness of forward-backward consistency property, we visualize an example in Fig. 6 where the red points indicate the key points and the green points represent the reprojected points. The length of the yellow line represents the distance (loss) between the corresponding points. We observe that enforcing cycle consistency effectively encourages the network to produce geometrically consistent predictions. However, the performance gain of using only the foreground-guided matching loss $\mathcal{L}_{\mathrm{matching}}$ is modest. We believe that the reason is due to the evaluation protocol of datasets considers only the matching on the foreground region. Namely, matching a background pixel in the source image to a foreground pixel in the target image will not be penalized. To demonstrate the effect of foreground-guided matching loss $\mathcal{L}_{\mathrm{matching}}$, we compute the percentage of correctly warped pixels (i.e., pixels in the foreground/background regions that are correctly warped into foreground/background region) over the entire dataset. As shown in Fig. 5, our method effectively reduces the errors in matching pixels from

	Ground Truth Label	
Warped Label Foreground	82.3 %	29.1 %
Background	17.7 %	70.9 %
	Foreground	Background

Rocco *et al.* [15]

	Ground Truth Label	
Warped Label Foreground	84.7 %	27.8 %
Background	15.3 %	72.2 %
	Foreground	Background

Ours

	Baseline	Ours
Accuracy	79.1 %	81.2%
Variance	10.3%	6.73%

Comparison

Fig. 5: Effect of using the foreground-guided matching loss $\mathcal{L}_{\mathrm{matching}}$.

Table 3: **Results on PF-WILLOW.**

Method	$\tau = 0.05$	$\tau = 0.1$	$\tau = 0.15$
SIFT Flow [5]	0.247	0.380	0.504
DAISY w/SF [36]	0.324	0.456	0.555
DeepC w/SF [37]	0.212	0.364	0.518
LIFT w/SF [38]	0.224	0.346	0.489
VGG w/SF [39]	0.324	0.456	0.555
FCSS w/SF [13]	0.354	0.532	0.681
LOM HOG [16]	0.284	0.568	0.682
UCN [11]	0.291	0.417	0.513
DSFM [40]	-	0.680	-
SCNet-A [12]	0.390	0.725	0.873
SCNet-AG [12]	0.394	0.721	0.871
SCNet-AG+ [12]	0.386	0.704	0.853
ResNet-101+CNNGeo(S) [14]	0.448	0.777	0.899
ResNet-101+CNNGeo(W) [15]	0.477	0.812	0.917
Ours	**0.491**	**0.819**	**0.922**

Table 4: **Results on TSS.** Marker * indicates that the method uses extra images from the PASCAL VOC 2007 dataset.

Method	FG3DCar	JODS	PASCAL	Avg.
HOG+PF-LOM [16]	0.786	0.653	0.531	0.657
HOG+TSS [7]	0.830	0.595	0.483	0.636
FCSS+SIFT Flow [13]	0.830	0.656	0.494	0.660
FCSS+PF-LOM [13]	0.839	0.635	0.582	0.685
HOG+OADSC [41]*	0.875	0.708	0.729	0.771
FCSS+DCTM [42]	0.891	0.721	**0.610**	0.740
VGG-16+CNNGeo [14]	0.835	0.656	0.527	0.673
ResNet-101+CNNGeo(S) [14]	0.886	0.758	0.560	0.735
ResNet-101+CNNGeo(W) [15]	0.892	0.758	0.562	0.737
Ours	**0.898**	**0.768**	0.560	**0.742**

foreground to background and vice versa. The improvement here is important in real-world applications but is not reflected in the metric used in the standard datasets. We also note that our method may not produce a clear figure-ground separation when the background contains visually similar regions to the foreground object in the other image. However, this case is also challenging for most methods.

The ablation study shows that all of the proposed components play crucial roles in producing accurate matching results. From Fig. **??**, we observe that the proposed method outperforms the best competitor [15] with a significant margin at multiple thresholds.

4.4 Experimental results on the PF-WILLOW and TSS datasets

To evaluate the generalization capability, we apply the learned model trained on the PF-PASCAL dataset to test directly on the PF-WILLOW and TSS datasets without finetuning on these two datasets.

Results on the PF-WILLOW dataset. Table 3 presents the quantitative results for the PF-WILLOW dataset. We compare the performance with several recent methods [11–15] as well as conventional approaches [5, 36–39] using hand-crafted features. The results are directly taken from [12] except [14,15]. For [14] and [15], we run the code provided by the authors to obtain the results. From Table 3, we observe that our model achieves the state-of-the-art performance with all three thresholds.

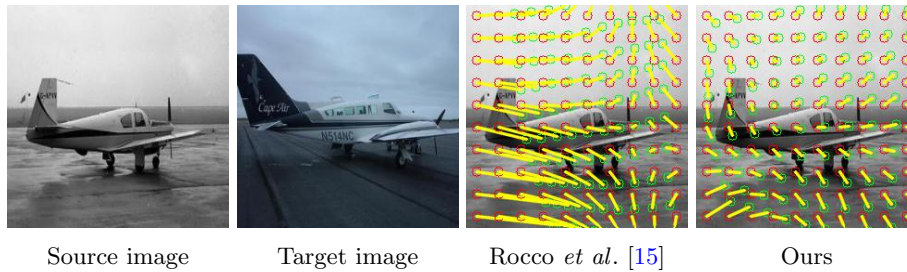| Source image | Target image | Rocco *et al.* [15] | Ours |

Fig. 6: **Cycle consistency property.** We present the visualization that demonstrates the effect of forward-backward consistency loss where the red points indicate the keypoints while the green points denote the reprojected points. Yellow line represents the distance (loss) between the linked points.

Results on the TSS dataset. We also evaluate the performance on the TSS dataset. Table 4 presents the quantitative results. We observe that our method achieves the state-of-the-art performance on two of the three groups of the TSS dataset: FG3DCar and JODS. Our results are slightly worse than that in [41] in the PASCAL group. However, the method in [41] uses additional images from the PASCAL VOC 2007 dataset. We report their results for completeness. Under the same experimental settings, the proposed method performs favorably against existing approaches.

5   Conclusions

In this work, we present an effective approach to improve semantic matching. The core technical novelty of our approach lies in the explicit modeling of a *foreground detection* module to suppress the effect of background clutter and exploiting the *cycle consistency* constraints so that the predicted geometric transformations are geometrically plausible and consistent across multiple images. The network training requires only training image pairs with image-level supervision and thus significantly alleviates the cost of constructing and labeling large-scale training datasets. Experimental results demonstrate that our approach performs favorably against existing semantic matching algorithms on several standard benchmarks. Moving forward, we believe that the semantic matching network can be further integrated to other computer vision tasks, e.g., supporting 3D semantic object reconstruction and fine-grained visual recognition.

References

1. Berthold KP Horn and Brian G Schunck. Determining optical flow. *Artificial intelligence*, 1981.
2. Bruce D Lucas, Takeo Kanade, et al. An iterative image registration technique with an application to stereo vision. In *IJCAI*, 1981.
3. Daniel Scharstein and Richard Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *IJCV*, 2002.
4. Zeng-Fu Wang and Zhi-Gang Zheng. A region based stereo matching algorithm using cooperative optimization. In *CVPR*, 2008.
5. Ce Liu, Jenny Yuen, and Antonio Torralba. Sift flow: Dense correspondence across scenes and its applications. *TPAMI*, 2011.
6. Hsin-Yi Chen, Yen-Yu Lin, and Bing-Yu Chen. Co-segmentation guided hough transform for robust feature matching. *TPAMI*, 2015.
7. Tatsunori Taniai, Sudipta N Sinha, and Yoichi Sato. Joint recovery of dense correspondence and cosegmentation in two images. In *CVPR*, 2016.
8. Kuang-Jui Hsu, Yen-Yu Lin, and Yung-Yu Chuang. Co-attention cnns for unsupervised object co-segmentation. In *IJCAI*, 2018.
9. Armin Mustafa and Adrian Hilton. Semantically coherent co-segmentation and reconstruction of dynamic scenes. In *CVPR*, 2017.
10. Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.
11. Christopher B Choy, JunYoung Gwak, Silvio Savarese, and Manmohan Chandraker. Universal correspondence network. In *NIPS*, 2016.
12. Kai Han, Rafael S Rezende, Bumsub Ham, Kwan-Yee K Wong, Minsu Cho, Cordelia Schmid, and Jean Ponce. Scnet: Learning semantic correspondence. In *ICCV*, 2017.
13. Seungryong Kim, Dongbo Min, Bumsub Ham, Sangryul Jeon, Stephen Lin, and Kwanghoon Sohn. Fcss: Fully convolutional self-similarity for dense semantic correspondence. In *CVPR*, 2017.
14. Ignacio Rocco, Relja Arandjelović, and Josef Sivic. Convolutional neural network architecture for geometric matching. In *CVPR*, 2017.
15. Ignacio Rocco, Relja Arandjelović, and Josef Sivic. End-to-end weakly-supervised semantic alignment. In *CVPR*, 2018.
16. Bumsub Ham, Minsu Cho, Cordelia Schmid, and Jean Ponce. Proposal flow: Semantic correspondences from object proposals. *TPAMI*, 2017.
17. Yuan-Ting Hu and Yen-Yu Lin. Progressive feature matching with alternate descriptor selection and correspondence enrichment. In *CVPR*, 2016.
18. Yuan-Ting Hu, Yen-Yu Lin, Hsin-Yi Chen, Kuang-Jui Hsu, and Bing-Yu Chen. Matching images with multiple descriptors: An unsupervised approach for locally adaptive descriptor selection. *TIP*, 2015.
19. Kuang-Jui Hsu, Yen-Yu Lin, Yung-Yu Chuang, et al. Robust image alignment with multiple feature descriptors and matching-guided neighborhoods. In *CVPR*, 2015.
20. Jaechul Kim, Ce Liu, Fei Sha, and Kristen Grauman. Deformable spatial pyramid matching for fast dense correspondences. In *CVPR*, 2013.
21. David Novotny, Diane Larlus, and Andrea Vedaldi. Anchornet: A weakly supervised network to learn geometry-sensitive features for semantic matching. In *CVPR*, 2017.
22. Angjoo Kanazawa, David W Jacobs, and Manmohan Chandraker. Warpnet: Weakly supervised matching for single-view reconstruction. In *CVPR*, 2016.

23. Simon Meister, Junhwa Hur, and Stefan Roth. Unflow: Unsupervised learning of optical flow with a bidirectional census loss. In *AAAI*, 2018.
24. Yuliang Zou, Zelun Luo, and Jia-Bin Huang. Df-net: Unsupervised joint learning of depth and flow using cross-task consistency. In *ECCV*, 2018.
25. Jia-Bin Huang, Sing Bing Kang, Narendra Ahuja, and Johannes Kopf. Temporally coherent completion of dynamic video. *ACM Transactions on Graphics (TOG)*, 2016.
26. Wei-Sheng Lai, Jia-Bin Huang, Oliver Wang, Eli Shechtman, Ersin Yumer, and Ming-Hsuan Yang. Learning blind video temporal consistency. In *ECCV*, 2018.
27. Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *CVPR*, 2017.
28. Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *ECCV*, 2018.
29. Xiaowei Zhou, Menglong Zhu, and Kostas Daniilidis. Multi-image matching via fast alternating minimization. In *ICCV*, 2015.
30. Tinghui Zhou, Yong Jae Lee, Stella X Yu, and Alyosha A Efros. Flowweb: Joint image set alignment by weaving consistent, pixel-wise correspondences. In *CVPR*, 2015.
31. Tinghui Zhou, Philipp Krahenbuhl, Mathieu Aubry, Qixing Huang, and Alexei A Efros. Learning dense correspondence via 3d-guided cycle consistency. In *CVPR*, 2016.
32. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.
33. Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.
34. Yi Yang and Deva Ramanan. Articulated human detection with flexible mixtures of parts. *TPAMI*, 2013.
35. Lubomir Bourdev and Jitendra Malik. Poselets: Body part detectors trained using 3d human pose annotations. In *ICCV*, 2009.
36. Engin Tola, Vincent Lepetit, and Pascal Fua. Daisy: An efficient dense descriptor applied to wide-baseline stereo. *TPAMI*, 2010.
37. Sergey Zagoruyko and Nikos Komodakis. Learning to compare image patches via convolutional neural networks. In *CVPR*, 2015.
38. Kwang Moo Yi, Eduard Trulls, Vincent Lepetit, and Pascal Fua. Lift: Learned invariant feature transform. In *ECCV*, 2016.
39. Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.
40. Nikolai Ufer and Björn Ommer. Deep semantic feature matching. In *CVPR*, 2017.
41. Fan Yang, Xin Li, Hong Cheng, Jianping Li, and Leiting Chen. Object-aware dense semantic correspondence. In *CVPR*, 2017.
42. Seungryong Kim, Dongbo Min, Stephen Lin, and Kwanghoon Sohn. Dctm: Discrete-continuous transformation matching for semantic flow. In *CVPR*, 2017.