Show, Match and Segment: Joint Learning of Semantic Matching and Object Co-segmentation

Yun-Chun Chen, Yen-Yu Lin, Ming-Hsuan Yang, and Jia-Bin Huang

Abstract—We present an approach for jointly matching and segmenting object instances of the same category within a collection of images. In contrast to existing algorithms that tackle the tasks of semantic matching and object co-segmentation in isolation, our method exploits the complementary nature of the two tasks. The key insights of our method are two-fold. First, the estimated dense correspondence field from semantic matching provides supervision for object co-segmentation by enforcing consistency between the predicted masks from a pair of images. Second, the predicted object masks from object co-segmentation in turn allow us to reduce the adverse effects due to background clutters for improving semantic matching. Our model is end-to-end trainable and does not require supervision from manually annotated correspondences and object masks. We validate the efficacy of our approach on four benchmark datasets: TSS, Internet, PF-PASCAL, and PF-WILLOW, and show that our algorithm performs favorably against the state-of-the-art methods on both semantic matching and object co-segmentation tasks.

Index Terms—Semantic matching, object co-segmentation, weakly-supervised learning.

1 INTRODUCTION

We address the problem of jointly aligning and segmenting different object instances of the same category from a collection of images. These two tasks, known as semantic matching and object co-segmentation (see Figure 1), are fundamental and active research topics in computer vision with applications ranging from object recognition [1], semantic segmentation [2], 3D reconstruction [3], content-based image retrieval [4], to interactive image editing [5]. Nevertheless, due to the presence of background clutters, large intra-class appearance variations, and drastic diversities of scales, poses, and viewpoints, both semantic matching and object co-segmentation remain challenging.

Existing approaches and their drawbacks. Numerous methods have been proposed to address the problems of semantic matching or object co-segmentation. Earlier approaches for semantic matching rely on hand-engineered features and a geometric alignment model in an energy minimization framework [1], [6], [7], [8]. Similarly, conventional object co-segmentation algorithms do not involve feature learning [9], [10], [11], [12]. The lack of end-to-end trainable features and inference pipelines often leads to limited performance. In light of this, recent methods leverage trainable descriptors and models for semantic matching [7], [13], [14], [15] and object co-segmentation [16], [17]. While promising results have been reported, training these models [7], [13], [14], [15], [16], [17] requires strong supervision in the form of manually labeled ground truth such as keypoint correspondences for semantic matching and object masks for object co-segmentation. However, constructing large-scale and diverse datasets is difficult since the labeling process is often expensive and labor-intensive. The dependence on manual supervision restricts the scalability of such approaches.

To alleviate this issue, several weakly supervised methods for semantic matching [18], [19], [20] and object co-segmentation [21] have been proposed. While these weakly supervised methods alleviate the need for collecting manually labeled datasets, two issues remain. First, existing algorithms for semantic matching [18], [19], [20] implicitly enforce the background features from both images to be similar, suffering from the negative impact caused by background clutters. Second, existing approaches for object co-segmentation [21] often resort to off-the-shelf object proposals algorithms to circumvent the need of manually annotated object masks. Nevertheless, generating a saliency map as pseudo supervision for each single image independently without considering the contents in other images is often error-prone (considering the case where multiple object instances are present in the given image). Furthermore, existing methods tend to segment only the discriminative regions rather than the entire objects.

Our work. In this paper, we propose to jointly tackle both semantic matching and object co-segmentation with a two-stream network in an end-to-end trainable fashion. Our key insights are two-fold. First, to suppress the effect of background clutters, the predicted object masks by object co-segmentation allow the model to focus on matching the segmented foreground regions while excluding background matching. Second, the estimated dense correspondence fields by semantic matching provide supervision for enforcing the model to generate geometrically consistent object masks across images. Therefore, we exploit the interdependency between the two network outputs, i.e., the estimated dense correspondence fields and the predicted foreground object masks, by introducing the cross-network consistency loss. Incorporating this loss improves both tasks since it encourages
two networks to generate more consistent explanations of the given image pair as shown in Figure 2.

The proposed training objective requires only weak image-level supervision (i.e., image pairs containing common objects). To facilitate the network training with such weak supervision, for semantic matching we develop cycle-consistent losses that make the predicted image transformations more geometrically plausible. For object co-segmentation, motivated by the classic idea of enforcing the foreground histograms of different images to be similar, i.e., histogram matching [22], we propose a perceptual contrastive loss that enhances the foreground appearance similarity between images while enforcing the figure-ground dissimilarity within each image. As shown in Figure 2, our model carries out joint learning to address both tasks simultaneously, producing more accurate and consistent semantic matching and object co-segmentation results.

Our contributions. First, we present a weakly-supervised and end-to-end trainable algorithm for joint semantic matching and object co-segmentation. Second, we propose a cross-network consistency loss that enforces consistency between the estimated correspondence fields and the predicted object masks, resulting in significant performance improvement for both tasks. Third, motivated by the histogram matching idea, we propose a perceptual contrastive loss that allows the model to segment the co-occurrent objects from an image collection. Fourth, we conduct extensive experiments on four benchmark datasets including the TSS [6], Internet [23], PF-PASCAL [7], and PF-WILLOW [7]. Extensive evaluation with existing semantic matching and object co-segmentation methods demonstrate that the proposed algorithm achieves the state-of-the-art performance on both tasks.

2 RELATED WORK

Semantic matching and object co-segmentation have been extensively studied in the literature. In this section, we review several topics relevant to our approach.

Semantic matching. Semantic matching algorithms can be grouped into two categories depending on the adopted feature descriptors: (1) hand-crafted descriptor based methods [1], [6], [7], [8], [24], [25], [26], [27], [28], [29], [30], [31], [32], [33] and (2) trainable descriptor based methods [13], [14], [15], [18], [34], [35], [36], [37], [38], [39], [40], [41], [42], [43], [44], [45], [46]. Hand-crafted descriptor based methods often leverage SIFT [47] or HOG [48] features along with geometric matching models to solve correspondence matching by energy minimization. However, hand-crafted descriptors are pre-defined and cannot adapt to various tasks. Trainable descriptor based methods either use pre-trained [31] or trainable CNN features for semantic matching [13], [14], [15], [34], [35], [36], [37], [42], [45], [46]. While these methods demonstrate significantly performance gain over those using hand-crafted features, they require manual correspondence annotations for training [13], [14], [31], [34], [35], [36], [37], [45], [46] or need to be learned in a self-supervised fashion [15], [42], [44].

Recently, several weakly supervised approaches [18], [29], [38], [39], [40], [41], [43] have been proposed to relax the dependence on keypoint-based supervision. The AnchorNet [29] learns a set of filters with geometrically consistent responses across different object instances to establish inter-image correspondences. The AnchorNet model, however, is not end-to-end trainable due to the use of the hand-engineered alignment model. The WarpNet [41] considers fine-grained image matching with small-scale and pose variations via aligning objects across images through known deformation. However, the application domain is relatively ideal since the objects are located in the image centers with limited translations, scale variations, and background clutters. Gaur et al. [43] propose an optimization algorithm that learns a latent space to cluster semantically related object parts. While this method provides geometric invariance to some degree, their approach cannot handle affine transformations across images which frequently occur in the context of semantic matching. To address this issue, several methods are proposed. Rocco et al. [18] present a weakly supervised semantic matching network using a differentiable soft inlier scoring module. The PARN [38] estimates locally-varying affine transformation fields across semantically similar images in a coarse-to-fine manner. Motivated by the procedure of non-rigid image registration between an image pair, the RTNs [40] use recurrent networks to progressively compute dense correspondences between two images. In addition to estimating geometric transformations between an image pair [18], [38], [40], another line of research focuses on establishing dense per-pixel correspondences without using any geometric models [39]. Rocco et al. [39] establish dense
Object co-segmentation. Object co-segmentation algorithms can be categorized into two groups: (1) graph-based [9], [10], [11], [12] and (2) clustering-based [49], [50], [51] approaches. Graph-based methods first construct a graph to encode the relationships between object instances from different images and formulate object co-segmentation as a labeling problem. Clustering-based methods, on the other hand, assume that common objects share similar appearances and achieve co-segmentation by finding tight clusters. Existing methods in both groups use hand-crafted features such as SIFT [47], HOG [48], or texton [52] for describing a set of object candidates extracted from super-pixels or region-based proposals. Recently, learning based methods [16], [17], [21], [53] have been developed for object co-segmentation. While significant improvement has been shown, these methods [16], [17], [53] require costly foreground masks for training and are not applicable to unseen object categories. To alleviate this issue, Hsu et al. [21] leverage unsupervised object proposals algorithms to produce a saliency map as pseudo ground-truth for each image. However, the pseudo ground truth is generated independently for each image and may not accurately highlight the co-occurrent objects in an image collection. Similar to the scheme by Hsu et al. [21], our method does not require manually labeled object masks and can segment objects of unseen categories. Our algorithm differs from them [21] in three aspects. First, our method does not need pseudo ground-truth for each image as supervision to guide the network training. Second, the estimated geometric transformations from semantic matching provide supervision for object co-segmentation by enforcing the predicted object masks from the given image pair to be geometrically consistent. Third, our model further takes into account the correlation map, i.e., cross-image information, when performing object co-segmentation.

Joint semantic matching and object co-segmentation. Several methods explore joint semantic correspondence and object co-segmentation. Rubinstein et al. [23] carry out object co-segmentation by exploiting the dense correspondence fields derived by the SIFT flow [1]. Taniai et al. [6] develop a hierarchical Markov random field (MRF) model for joint dense matching and object co-segmentation. However, these methods employ hand-crafted descriptors. Motivated by these methods [6], [23], our approach leverages the complementary nature of the two tasks and develop cross-network loss functions that couple the two networks during optimization for improving the performance of the individual tasks. After optimization, the learned semantic matching and the object co-segmentation models can be applied jointly or independently.

Meta-supervision via coupled network training. Enforcing consistency across different network outputs has been used in several vision applications including image translations [54], [55], [56], depth and ego-motion [57], depth and optical flow [58], and shape reconstruction [59]. In this work, we design a cross-network objective function to make the semantic matching and object co-segmentation networks complementary to each other, and demonstrate that coupled training of the two heterogeneous networks leads to significant performance improvement on both tasks.

Cycle consistency. Cycle consistency constraints have been

Fig. 2: Separate learning vs. joint learning. Addressing semantic matching (left) or object co-segmentation (right) in isolation often suffers from the effect of background clutters (for semantic matching) or only focuses on segmenting the discriminative parts (for object co-segmentation). In this work, we exploit the property that the predicted object masks allow the model to suppress the negative impact due to background clutters while the estimated dense correspondence fields provide supervision for object co-segmentation. We couple the learning of both tasks through a cross-network consistency loss and show that joint learning improves the performance of both tasks.
used to regularize network training for numerous vision tasks. In image-to-image translation, enforcing cycle consistency allows the model to learn the mappings between domains without paired data [55], [56], [60]. In unsupervised domain adaptation, exploiting cross-domain invariance in the label space results in more consistent task predictions for unlabeled images of different domains [61]. In motion analysis, enforcing forward-backward consistency constraints has been shown effective for detecting occlusion while learning optical flow [58], [62] or enforcing temporal consistency in videos [63]. Similar to these methods, the idea of cycle consistency is also extensively applied to semantic matching. The FlowWeb [64] enforces cycle consistency constraints to establish globally-consistent dense correspondences. Zhou et al. [65] address multi-image matching by jointly learning feature matching and enforcing cycle consistency. However, these methods [64], [65] adopt hand-crafted descriptors, which may not adapt to unseen object category given for matching. While trainable descriptor-based methods [66] are proposed to alleviate this limitation, this method [66] utilizes an additional 3D CAD model to form a cross-instance loop between synthetic and real images for establishing dense correspondences. Namely, this method [66] requires four images for computing the cycle consistency loss. In contrast, our method does not need any additional data to guide network training. Experimental results show that with the two developed cycle consistency losses, our model produces consistent matching results, resulting in significant performance gain.

3 Proposed Algorithm

In this section, we first provide an overview of the proposed approach. We then describe the objective functions for joint semantic matching and object co-segmentation followed by the implementation details.

3.1 Algorithmic overview

Given a set of \( N \) images \( \mathcal{I} = \{I_j\}_{j=1}^N \) containing objects of a specific category, our goal is to learn a model that can determine the geometric correspondences between the input image pairs while segmenting the common objects in \( \mathcal{I} \) without knowing the object class a priori. Our formulation for joint semantic matching and object co-segmentation is weakly-supervised since it requires only weak image-level supervision in the form of training image pairs containing objects of one particular class. No ground-truth keypoint correspondences and object masks are used in the training stage.

Network. As shown in Figure 3, our model is composed of four CNN sub-networks: an encoder \( E \), a transformation predictor \( G \), a decoder \( D \), and an ImageNet-pretrained ResNet-50 feature extractor \( F \). Given an input image pair, we first use the encoder \( E \) to encode the content of each image. We then apply a correlation layer for computing matching scores for every pair of features from two images. The correlation layer has been extensively applied in the other context, including optical flow [67], stereo [68], [69] video object segmentation [70], [71]. Here, taking two tensors of matching scores as inputs, we apply a transformation predictor \( G \) to estimate the geometric transformation that aligns the two images. To generate object masks, we use the fully convolutional network decoder \( D \) for object co-segmentation. To capture the co-occurrence information, we concatenate the encoded image features with the correlation maps. Our decoder then takes the concatenated features as inputs to generate object segmentation masks. The use of both feature maps and correlation maps for object co-segmentation has been proposed in [17]. However, the method in [17] requires manually annotated object masks for training. To enable the decoder to segment the co-occurrence objects from the given image pair without supervision from ground truth object masks, we leverage the perceptual contrastive loss \( L_{\text{contrast}} \) using the ImageNet-pretrained ResNet-50 feature extractor \( F \) to enforce the appearance similarity between the segmented foreground across images and the dissimilarity between the masked foreground and background within each image.

Training losses. Our training objective consists of five loss functions. First, the foreground-guided matching loss \( L_{\text{matching}} \) minimizes the distance between corresponding features based on the estimated geometric transformation. Unlike existing feature learning methods for semantic matching [14], [15], [18], our model explicitly takes the predicted object masks into account to suppress the negative impacts caused by background clutters. Second, the cross-network consistency loss \( L_{\text{task-consis}} \) penalizes the inconsistency of the predicted object masks of an input image pair and the estimated geometric transformation between that pair. Such a cross-network loss couples the networks during training and provides supervisory signals for both semantic matching and object co-segmentation. Third, the perceptual contrastive loss \( L_{\text{contrast}} \) guides the decoder \( D \) to produce object co-segments with higher inter-image foreground similarity and large intra-image figure-ground separation. Fourth, both the forward-backward consistency loss \( L_{\text{cycle-consis}} \) and transitivity consistency loss \( L_{\text{trans-consis}} \) (applied to three input images at a time) regularize the network training by enforcing the predicted geometric transformations to be consistent across multiple images. Specifically, the training objective \( \mathcal{L} \) is defined by

\[
\mathcal{L} = L_{\text{matching}} + \lambda_{\text{cycle}} \cdot L_{\text{cycle-consis}} + \lambda_{\text{trans}} \cdot L_{\text{trans-consis}} + \lambda_{\text{contrast}} \cdot L_{\text{contrast}} + \lambda_{\text{task}} \cdot L_{\text{task-consis}},
\]

where \( \lambda_{\text{cycle}}, \lambda_{\text{trans}}, \lambda_{\text{contrast}}, \) and \( \lambda_{\text{task}} \) are the hyper-parameters used to control the relative importance of the respective loss terms.

3.2 Semantic matching

Foreground-guided matching loss \( L_{\text{matching}} \). Given an image pair \( (I_A, I_B) \), the encoder \( E \) represents the images with the 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 apply a correlation layer to \( f_A \) and \( f_B \), and obtain the correlation map \( S_{AB} \in \mathbb{R}^{h_A \times w_A \times h_B \times w_B} \), where the element \( S_{AB}(i, j, s, t) = S_{AB}(p, q) \) records the normalized inner product between the feature vectors extracted at two locations \( p = [i, j]^\top \) in \( f_A \) and \( q = [s, t]^\top \) in \( f_B \). The correlation map \( S_{AB} \) can be 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
Fig. 3: Overview of the proposed model. Our model is a two-stream network: (top) semantic matching network and (bottom) object co-segmentation network. Our model consists of four main CNN sub-networks: an encoder $\mathcal{E}$ (for extracting features from the input images), a transformation predictor $\mathcal{G}$ (for estimating the geometric transformation between an input image pair), a decoder $\mathcal{D}$ (for producing object masks), and an ImageNet-pretrained and fixed ResNet-50 feature extractor $\mathcal{F}$ (for computing the perceptual contrastive loss). The model training is driven by four loss functions, including the foreground-guided matching loss $\mathcal{L}_{\text{matching}}$, forward-backward consistency loss $\mathcal{L}_{\text{cycle-consist}}$, perceptual contrastive loss $\mathcal{L}_{\text{contrast}}$, and cross-network consistency loss $\mathcal{L}_{\text{task-consist}}$.

such, $S_{AB}$ can be interpreted as a dense $h_A \times w_A$ grid with $(h_B \times w_B)$-dimensional local features. With the reshaped $S_{AB}$, we use the transformation predictor $\mathcal{G}$ [15] to estimate a geometric transformation $T_{AB}$ which warps $I_A$ to $\tilde{I}_A$ so that $\tilde{I}_A$ and $I_B$ can be well aligned.

With the geometric transformation $T_{AB}$, we can identify and remove geometrically inconsistent correspondences. Consider a correspondence with the endpoints $(p \in \mathcal{P}_A, q \in \mathcal{P}_B)$, where $\mathcal{P}_A$ and $\mathcal{P}_B$ are the domains of all spatial coordinates of $f_A$ and $f_B$, respectively. We refer the distance $||T_{AB}(p) - q||$ as the projection error of this correspondence with respect to transformation $T_{AB}$. Following Rocco et al. [18], we introduce a correspondence mask $m_A \in \mathbb{R}^{h_A \times w_A \times (h_B \times w_B)}$ to determine if the correspondences are geometrically consistent with transformation $T_{AB}$. Specifically, $m_A$ is of the form

$$m_A(p, q) = \begin{cases} 1, & \text{if } ||T_{AB}(p) - q|| \leq \varphi, \\ 0, & \text{otherwise,} \end{cases}$$

(2)

where $\varphi$ is a predefined threshold. In (2), a correspondence is considered geometrically consistent with transformation $T_{AB}$ if its projection error is not larger than the threshold $\varphi$. Empirically, we set the threshold $\varphi$ to 1 in all experiments.

For the correspondence with the endpoints $(p \in \mathcal{P}_A, q \in \mathcal{P}_B)$, the correlation map $S_{AB}(p, q)$ and the correspondence mask $m_A(p, q)$ capture its appearance and geometric consensus, respectively. When focusing on point $p \in \mathcal{P}_A$, we compute the matching score of location $p$ by

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

(3)

Semantic matching often suffers from false positives caused by background clutters and false negatives caused by large intra-class variations. In this work, we exploit object co-segmentation, where object-level similarity complements patch-level similarity in semantic matching, to address the aforementioned issues. Specifically, we consider the object mask $M_A$ estimated by the decoder $\mathcal{D}$ for object co-segmentation, and resize it to resolution $h_A \times w_A$ for guiding the matching loss $\mathcal{L}_{\text{matching}}$. Our foreground-guided matching loss is formulated as

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

(4)

where $s_B$ and $M_B$ are similarly defined as $s_A$ and $M_A$, respectively. The negative sign in (4) indicates that maximizing the matching score is equivalent to minimizing the foreground-guided matching loss $\mathcal{L}_{\text{matching}}$. The loss $\mathcal{L}_{\text{matching}}$ encourages the transformation predictor $\mathcal{G}$ to generate transformations $T_{AB}$ and $T_{BA}$ with which the corresponding foreground features across the two images are as similar as possible.

**Cycle consistency.** For an image pair $I_A$ and $I_B$, the transformation predictor $\mathcal{G}$ estimates a geometric transformation $T_{AB}$ which can warp $I_A$ to $\tilde{I}_A$ such that $\tilde{I}_A$ aligns $I_B$ well. However, the large capacity of the transformation predictor $\mathcal{G}$ often leads to situations where various transformations can warp $I_A$ to $\tilde{I}_A$ such that $\tilde{I}_A$ aligns $I_B$ well. Namely, multiple points on $I_A$ can match well a single point on $I_B$. These cases implies that using the foreground-guided matching loss $\mathcal{L}_{\text{matching}}$ alone is insufficient to reliably train the transformation predictor $\mathcal{G}$ under the weakly supervised setting since no ground-truth correspondences
where the two aforementioned criteria are respectively imposed on \( d_{AB}^t \) and \( d_{AB}^o \):

\[
d_{AB}^t = \frac{1}{c} \left\| F(i_A^o) - F(i_B^o) \right\|^2 \quad \text{and} \quad d_{AB}^o = \max \left( 0, m - \frac{1}{2c} \left( \| F(i_A^o) - F(i_B^o) \|^2 + \| F(i_B^t) - F(i_B^o) \|^2 \right) \right)
\]

In (10), the constant \( c \) is set to be 2,048 which is the dimension of the semantic features produced by \( F \) [72] and the margin \( m \) is set to be 2 is the cutoff threshold.

As shown in Figure 4, minimizing the perceptual contrastive loss \( L_{\text{contrast}} \) in (8) entails minimizing the intra-image foreground object distinctness in (9) while maximizing the inter-image foreground-background discrepancy in (10). We note that minimizing \( d_{AB}^t \) in (9) is equivalent to trainable matching foreground histograms [22] between a pair of images. To prevent the perceptual contrastive loss \( L_{\text{contrast}} \) from being dominated by minimizing (10) (i.e., maximizing the figure-ground dissimilarity within each image) since the object image \( i_A^t \) and background image \( i_B^t \) are inherently dissimilar, we introduce a cutoff threshold \( m \).

Cross-network consistency loss \( L_{\text{task-consis}} \). Using the perceptual contrastive loss \( L_{\text{contrast}} \) alone for object co-segmentation may generate object masks that highlight only the discriminative parts rather than the entire objects. As shown in the right example of separate learning for object co-segmentation in Figure 2, the windows of the top bus
We implement our model using PyTorch. We adopt the works for the two tasks are coupled during training but can be applied independently for each task during inference.

The cross-network consistency loss $L_{\text{task-consis}}$ that bridges the outputs of the semantic matching network and the object co-segmentation network. This loss enforces the consistency between the learned geometric transformations $T_{AB}$ and $T_{BA}$ and the predicted object masks $M_A$ and $M_B$. To this end, we use $T_{AB}$ to warp $M_A$ and encourage that the warped mask $\tilde{M}_A$ and $M_B$ highly overlap. We compute the symmetric binary cross-entropy loss and define the cross-network consistency loss as

$$L_{\text{task-consis}}(I_A, I_B; E, G, D) = L_{\text{bce}}(\tilde{M}_A, M_B) + L_{\text{bce}}(\tilde{M}_B, M_A),$$

(11)

where $L_{\text{bce}}(\tilde{M}_A, M_B)$ computes the binary cross-entropy loss between $M_A$ and $M_B$, and is defined by

$$L_{\text{bce}}(\tilde{M}_A, M_B) = -\frac{1}{H_B \times W_B} \left( \sum_{i,j} \tilde{M}_A(i,j) \log(M_B(i,j)) \right)$$

$$+ \sum_{i,j} \left(1 - \tilde{M}_A(i,j)\right) \log \left(1 - M_B(i,j)\right).$$

(12)

The cross-network consistency loss $L_{\text{task-consis}}$ provides supervisory signals for both tasks without the need of ground-truth keypoint correspondences and object masks. While the model consists of four individual CNN subnetworks, our method end-to-end and jointly optimizes the training objective in (1) using weak supervision. The networks for the two tasks are coupled during training but can be applied independently for each task during inference.

### 3.4 Implementation details

We implement our model using PyTorch. We adopt the ResNet-101 [72] as the encoder $E$ and use the feature activations from the conv4–23 layer as our feature map. Similar to [15], our transformation predictor $G$ is a cascade of two modules predicting an affine transformation and a thin plate spline (TPS) transformation, respectively. We initialize the encoder $E$ and the transformation predictor $G$ from those in [73]. We construct the decoder $D$ with a siamese structure using four blocks, each of which contains one deconvolutional layer and two convolutional layers. The decoder $D$ is randomly initialized. We add skip connections between each block of the encoder $E$ and the decoder $D$. Our network $F$ is an ImageNet-pretrained ResNet-50 [72] and remains fixed during training. All images are resized to the resolution of $240 \times 240$ in advance. 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 [74] with an initial learning rate of $5 \times 10^{-8}$. 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.

### 4 Experimental Results

In this section, we first describe the experimental settings, and then present the quantitative and qualitative evaluation with comparisons to the state-of-the-art methods on four benchmark datasets for semantic matching and object co-segmentation. The source code, the pre-trained models, and additional results are available at https://yunchunen.github.io/MaCoSNet/.

#### 4.1 Evaluation metrics and datasets

Here we describe the evaluation metrics for semantic matching and object co-segmentation as well as the four datasets.

**Evaluation metrics.** We evaluate our proposed method on both semantic matching and object co-segmentation tasks. To measure the performance of semantic matching, we use the commonly used percentage of correct keypoints (PCK) metric [75] which calculates the percentage of keypoints whose reprojection errors are less than a given threshold. The reprojection error is the Euclidean distance $d(T_{AB}(p), p')$ between the locations of the warped keypoint $T_{AB}(p)$ and the ground-truth keypoint $p'$. The threshold is defined by $\alpha \cdot \max(H, W)$ where $H$ and $W$ are the height and width of the annotated object bounding box on the image, respectively. We report the performance under different values of $\alpha$.

For object co-segmentation, we adopt the precision $P$ and the Jaccard index $J$. The precision $P$ measures the percentage of correctly classified pixels. The Jaccard index $J$ is the ratio of the intersection area of the predicted foreground objects and the ground truth to their union area. Since background pixels are taken into account in the precision metric, this measure may not precisely reflect the quality of object co-segmentation results. In contrast, the Jaccard index $J$ is considered more reliable since it focuses on foreground objects.

**Datasets.** We conduct the experiments on four public benchmarks, including the TSS [6], Internet [23], PF-PASCAL [7],
and PF-WILLOW [7] datasets. We use the TSS dataset [6] for evaluating joint semantic matching and object co-segmentation as the TSS dataset contains the ground-truth annotations for both tasks. For object co-segmentation, we use a more challenging Internet dataset [23]. For semantic matching, we use the PF-PASCAL [7] and PF-WILLOW [7] datasets. The details of these datasets can be found the project website of this work.

### 4.2 Joint matching and co-segmentation

**Results on the TSS dataset.** Table 1 shows the quantitative results of semantic matching on the TSS [6] dataset. In this experiment, we set the hyper-parameters as follows: $\lambda_{\text{cycle-cons}} = 5$, $\lambda_{\text{trans-cons}} = 5$, $\lambda_{\text{contrast}} = 10$, and $\lambda_{\text{task-cons}} = 10$. Overall, the proposed method achieves the state-of-the-art performance on all three categories. Although our method performs slightly worse than the OHG [30] on the PASCAL category, the OHG method uses additional images from the PASCAL VOC 2007 dataset. In the bottom block of Table 1, we follow the PARN [38] and RTNs [40], and resize all images to the larger dimension to 100 (i.e., resizing $(H, W)$ to 100). The proposed method also performs favorably against all competing methods.

Table 2 shows the quantitative results of object co-segmentation on the TSS [6] dataset. The proposed method achieves a precision of 49.7% and a Jaccard index of 84%, and performs favorably against the state-of-the-art methods by a large margin 4.9% in precision and 12% in Jaccard index against the best competitor [10]. We attribute the significant performance gains to two factors. First, unlike most existing methods, our proposed approach tackles object co-segmentation with an end-to-end trainable model. Second, integrating semantic matching further improves co-segmentation. Figure 5 presents visual comparisons of object co-segmentation with existing methods. The proposed method generates more accurate co-segmentation results, particularly when images contain drastic background clutters and large intra-class appearance variations.

**Effect of joint learning.** We conduct an ablation study on joint learning of matching and co-segmentation. We investigate two variants: 1) Ours w/o co-seg: disabling the object co-segmentation network stream (falling back to our preliminary results WeakMatchNet [73]) and 2) Ours w/o matching: disabling the semantic matching stream. Table 1 and Table 2 show the quantitative results of these two variant methods. For semantic matching, our model suffers a performance loss of 1.8% in average PCK at $\alpha = 0.05$. For object co-segmentation, our results show a drop of 4.8% in precision and a 11% drop in Jaccard index. The proposed coupled training approach significantly improves the performance for both tasks. In particular, the improvement over Ours w/o co-seg [73] indicates the benefits of explicit object mask estimation using object co-segmentation.

Figure 6 shows four example results of the qualitative object co-segmentation examples. The co-segmentation model (Ours w/o matching) may focus only the most discriminative parts as reflected by the motorbike example (the model focuses on segmenting the wheels of the motorbike). Many false positives and false negatives are generated due to drastic appearance variations. With the guidance of geometric transformations inferred from semantic matching, our joint training model significantly alleviates these unfavorable false positives and false negatives, resulting in more accurate and consistent object co-segmentation results.

**4.3 Object co-segmentation**

Table 3 reports the quantitative results on the challenging Internet dataset [23]. In this experiment, we set the hyper-parameters as follows: $\lambda_{\text{cycle-cons}} = 5$, $\lambda_{\text{trans-cons}} = 5$, $\lambda_{\text{contrast}} = 20$, and $\lambda_{\text{task-cons}} = 10$. Our results show that our method compares favorably against existing weakly-supervised methods and achieves competitive performance when compared with a strongly supervised approach [17]. The performance gain over the best competitor under the same experimental setting [21] is 2.6% in precision and 2%
in Jaccard index. Our results demonstrate that our method is capable of adapting itself well to unseen object categories by training with weak image-level supervision provided by the dataset. Figure 7 presents the visual comparisons of object co-segmentation with existing methods. From the visual results, we observe that our method is more robust to intra-class appearance variations and viewpoint changes, and produces more accurate and consistent co-segmentation results when comparing with existing methods.

4.4 Semantic matching

To evaluate the proposed method on semantic matching, we conduct experiments on the PF-PASCAL [7] and PF-WILLOW [7] datasets. We set the hyper-parameters as follows: $\lambda_{\text{cycle-cons}} = 20$, $\lambda_{\text{trans-cons}} = 10$, $\lambda_{\text{contrast}} = 2.5$, and $\lambda_{\text{task-cons}} = 2.5$.

Results on the PF-PASCAL dataset. Table 4 shows the quantitative results of semantic matching on the PF-PASCAL [7] dataset. The proposed approach performs favorably against the state-of-the-art methods, achieving an overall PCK of 79.0%. The advantage of integrating object co-segmentation over performing foreground detection on the feature maps can be assessed by comparing the proposed method with

### TABLE 3: Experimental results of object co-segmentation on the Internet dataset [23]

| Method          | Descriptor | Airplane | Car | Horse | Avg. |
|-----------------|------------|----------|-----|-------|------|
| Ours            | VGG-16 [84]| 0.689    | 0.941| 0.928 | 0.75 |
| et al. [86]     | VGG-16 [84]| 0.403    | 0.670| 0.73  | 0.71 |
| Joulin et al. [77]| SIFT [47] | 0.467    | 0.702| 0.85  | 0.76 |
| Joulin et al. [49]| SIFT [47] | 0.475    | 0.962| 0.85  | 0.77 |
| Kim et al. [87] | SIFT [47] | 0.402    | 0.869| 0.900| 0.75 |
| Robinson et al. [10] | SIFT [47] | 0.390    | 0.854| 0.844| 0.72 |
| Chen et al. [88] | SIFT [47] | 0.902    | 0.976| 0.85  | 0.90 |
| Que et al. [11] | SIFT [47] | 0.910    | 0.985| 0.92  | 0.94 |
| Hati et al. [85] | SIFT [47] | 0.777    | 0.721| 0.65  | 0.71 |
| Chang et al. [9] | SIFT [47] | 0.726    | 0.759| 0.76  | 0.72 |
| MSRC [59]       | SIFT [47] | 0.525    | 0.647| 0.73  | 0.65 |
| [87]            | SIFT [47] | 0.903    | 0.880| 0.67  | 0.89 |
| Hsu et al. [21] | VGG-16 [84]| 0.815  | 0.864| 0.75  | 0.86 |
| Ours            | VGG-16 [84]| 0.822  | 0.877| 0.85  | 0.87 |
| Ours            | ResNet-101 [72]| 0.848  | 0.877| 0.85  | 0.88 |
TABLE 4: Experimental results of semantic matching on the PF-PASCAL dataset [7]. The bold and underlined numbers indicate the top two results, respectively.

| Method | Descriptor | aero | bike | bird | boat | bottle | bus | car | cat | chair | cow | dtable | dog | horse | moto | person | plant | sheep | sofa | train | tv | mean |
|--------|------------|------|------|------|------|--------|-----|-----|-----|-------|-----|--------|-----|-------|------|--------|------|-------|-----|-------|-----|------|
| Proposal Flow + LOM [7] | HOG [48] | 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 [14] | GoogleNet [83] | 68.8 | 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 | 53.3 | 49.0 | 72.0 | 66.0 | 55.6 |
| SCNet-AG [13] | VGG-16 [84] | 59.7 | 72.3 | 48.3 | 59.7 | 74.5 | 72.7 | 73.2 | 59.5 | 51.4 | 79.2 | 39.4 | 50.1 | 67.0 | 62.1 | 49.3 | 66.5 | 78.2 | 63.3 | 57.7 | 59.4 | 66.3 |
| SCNet-AG+ [13] | VGG-16 [84] | 75.5 | 84.4 | 66.3 | 70.8 | 57.4 | 82.7 | 92.3 | 71.6 | 54.2 | 98.8 | 52.2 | 59.5 | 66.6 | 75.0 | 56.3 | 60.4 | 60.0 | 75.7 | 66.5 | 77.7 | 72.2 |
| CNNGeo [15] | ResNet-101 [72] | 79.5 | 80.9 | 69.9 | 61.1 | 57.8 | 77.1 | 84.4 | 53.5 | 58.1 | 83.3 | 37.0 | 54.1 | 58.2 | 70.7 | 51.4 | 41.4 | 60.0 | 44.3 | 55.3 | 30.0 | 62.6 |
| CNNGeo [15] | ResNet-101 [72] | 83.0 | 82.2 | 81.1 | 50.0 | 57.8 | 79.9 | 92.8 | 77.5 | 44.7 | 85.8 | 28.1 | 69.8 | 65.4 | 77.1 | 64.0 | 65.2 | 100.0 | 50.8 | 44.3 | 54.4 | 69.5 |
| CNNGeo w/ Inlier [18] | ResNet-101 [72] | 84.7 | 85.8 | 80.9 | 55.6 | 76.6 | 88.5 | 92.9 | 79.6 | 52.0 | 85.4 | 28.1 | 71.8 | 67.0 | 75.1 | 66.3 | 70.5 | 100.0 | 62.1 | 62.5 | 61.3 | 74.8 |
| SCNet-A [1] | ResNet-101 [72] | 86.8 | 80.5 | 86.7 | 55.5 | 82.4 | 88.6 | 93.8 | 87.1 | 54.3 | 87.5 | 43.2 | 82.0 | 64.1 | 79.2 | 71.1 | 71.0 | 60.4 | 54.2 | 75.0 | 82.8 | 78.9 |
| SCNet-AG [1] | ResNet-101 [72] | 87.6 | 89.6 | 82.1 | 83.3 | 85.9 | 82.3 | 93.9 | 80.2 | 52.3 | 85.4 | 58.2 | 75.2 | 64.0 | 79.2 | 67.2 | 73.8 | 100.0 | 65.3 | 69.3 | 61.1 | 78.0 |
| SCNet-AG+ [1] | ResNet-101 [72] | 87.6 | 87.4 | 85.5 | 82.2 | 76.6 | 94.2 | 94.7 | 86.9 | 54.8 | 88.4 | 57.5 | 83.2 | 70.6 | 79.2 | 75.3 | 70.5 | 100.0 | 63.0 | 66.3 | 64.4 | 79.0 |
| WeakMatchNet [73] | Ours | 87.6 | 89.6 | 82.1 | 83.3 | 85.9 | 82.3 | 93.9 | 80.2 | 52.3 | 85.4 | 58.2 | 75.2 | 64.0 | 79.2 | 67.2 | 73.8 | 100.0 | 65.3 | 69.3 | 61.1 | 78.0 |

Fig. 7: Qualitative results of object co-segmentation on the Internet [23] dataset. Our method is capable of delineating accurate co-occurring object masks under large intra-class variations and background clutter.

TABLE 5: Experimental results of semantic matching on the PF-WILLOW dataset [7]. The bold and underlined numbers indicate the top two results, respectively.

| Method | Descriptor | aero | bike | bird | boat | bottle | bus | car | cat | chair | cow | dtable | dog | horse | moto | person | plant | sheep | sofa | train | tv | mean |
|--------|------------|------|------|------|------|--------|-----|-----|-----|-------|-----|--------|-----|-------|------|--------|------|-------|-----|-------|-----|------|
| SIFT Flow [1] | SIFT [47] | 0.247 | 0.380 | 0.504 |
| SIFT Flow [1] | VGG-16 [84] | 0.324 | 0.456 | 0.595 |
| CNNGeo [15] | ResNet-101 [72] | 0.448 | 0.777 | 0.899 |
| CNNGeo w/ Inlier [18] | ResNet-101 [72] | 0.477 | 0.812 | 0.917 |
| Proposal Flow + LOM [7] | HOG [48] | 0.284 | 0.568 | 0.682 |
| UCN [14] | GoogleNet [83] | 0.291 | 0.417 | 0.513 |
| SCNet-A [1] | VGG-16 [84] | 0.390 | 0.725 | 0.873 |
| SCNet-AG [1] | VGG-16 [84] | 0.394 | 0.721 | 0.871 |
| SCNet-AG+ [1] | VGG-16 [84] | 0.386 | 0.704 | 0.853 |
| WeakMatchNet [73] | ResNet-101 [72] | 0.454 | 0.816 | 0.918 |
| RTRs [40] | ResNet-101 [72] | 0.413 | 0.719 | 0.862 |
| NC-Net [39] | ResNet-101 [72] | 0.514 | 0.818 | 0.927 |
| Ours | ResNet-101 [72] | 0.534 | 0.854 | 0.938 |

the WeakMatchNet [73]. The proposed method improves the performance by 1.0% in terms of PCK evaluated at \( \alpha = 0.1 \). The top row of Figure 8 shows semantic matching results of evaluated methods. Estimating geometric transformations leads to more geometrically consistent matching results than approaches that establishes correspondences without using any geometric transformation models (i.e., NC-Net [39]). The advantage of incorporating object co-segmentation can be observed in the fourth example of Figure 8 by comparing between our approach and [73]. Our method generates more accurate matching results.

Results on the PF-WILLOW dataset. To evaluate the generalization capability of the proposed method, we evaluate the proposed model trained on the PF-PASCAL dataset [7] to the PF-WILLOW dataset [7] without fine-tuning. Table 5 shows the quantitative results of semantic matching on the PF-WILLOW [7] dataset. Our method performs favorably against existing methods on all three evaluated PCK thresholds. The performance gain over the second best method [39] is 2.0% at \( \alpha = 0.05 \) or 3.6% at \( \alpha = 0.1 \). The results suggest that sufficient generalization ability in establishing dense correspondences can be exhibited by our model. Figure 8 shows two examples of visual results of the evaluated
methods. The matching results by our method are more accurate and geometrically consistent.

### 4.5 Ablation study

**Removing one loss at a time.** To analyze the importance of each adopted loss function, we conduct an ablation study by turning off one of the loss terms at a time. For object co-segmentation, we carry out experiments on the Internet dataset [23]. Table 6 shows the experimental results. For semantic matching, without the foreground-guided matching loss $L_{\text{matching}}$, there is no explicit supervision to maximize the similarity between the corresponding features of an image pair. Thus, our model suffers from a significant performance drop by 4.4% at $\alpha = 0.05$. For object co-segmentation, while the performance drops are moderate when the foreground-guided matching loss $L_{\text{matching}}$ is turned off, the results indicate that having better ability in predicting geometric transformations can further improves object co-segmentation.

Without the forward-backward consistency loss $L_{\text{cycle-consis}}$, our model only enforces the consistency across multiple images through the transitivity consistency loss $L_{\text{trans-consis}}$. Experimental results show that the performance drops by 0.9% at $\alpha = 0.05$ for semantic matching, and 0.7% in precision and 1% in Jaccard index for object co-segmentation.

Without the transitivity consistency loss $L_{\text{trans-consis}}$, the performance drops by 4.4% at $\alpha = 0.05$ for semantic matching, and 0.7% in precision and 1% in Jaccard index for object co-segmentation. The bold and underlined numbers indicate the top two results.

### Table 7: Sensitivity analysis of the cutoff threshold $m$ on object co-segmentation on the TSS dataset [6].

The bold and underlined numbers indicate the top two results.

| Threshold $m$ | FG3D | Car | Horse | Avg. |
|--------------|------|-----|-------|------|
| 0            | 0.923 | 0.64 | 0.796 | 0.72 |
| 0.5          | 0.927 | 0.64 | 0.887 | 0.72 |
| 1            | 0.931 | 0.64 | 0.930 | 0.72 |
| 2            | 0.931 | 0.64 | 0.930 | 0.72 |
| 5            | 0.931 | 0.64 | 0.930 | 0.72 |
| 10           | 0.931 | 0.64 | 0.930 | 0.72 |

### Table 8: Ablation study of semantic matching on the PF-WILLOW dataset [7] under three different PCK thresholds $\alpha$. The bold and underlined numbers indicate the top two results.

| Method | $\alpha = 0.05$ | $\alpha = 0.10$ | $\alpha = 0.15$ |
|--------|-----------------|-----------------|-----------------|
| Ours (full model) | 0.538 | 0.854 | 0.939 |
| Ours w/o $L_{\text{matching}}$ | 0.494 | 0.822 | 0.927 |
| Ours w/o $L_{\text{cycle-consis}}$ | 0.529 | 0.847 | 0.938 |
| Ours w/o $L_{\text{trans-consis}}$ | 0.532 | 0.851 | 0.930 |
| Ours w/o $L_{\text{task-consis}}$ | 0.514 | 0.842 | 0.928 |
| WeakMatchNet [73] | 0.502 | 0.823 | 0.922 |

Experimental results show that the performance drops by 0.9% at $\alpha = 0.05$ for semantic matching, and 0.7% in precision and 1% in Jaccard index for object co-segmentation.

Without the transitivity consistency loss $L_{\text{trans-consis}}$, the performance drops by 4.4% at $\alpha = 0.05$ for semantic matching, and 0.7% in precision and 1% in Jaccard index for object co-segmentation.
our model only enforces the consistency on the estimated geometric transformations between an image pair (i.e., the forward-backward consistency loss $L_{\text{cycle-consis}}$ is still in effect). Experimental results show that the performance drops by 0.6% at $\alpha = 0.05$ for semantic matching, and 0.5% in precision and 1% in Jaccard index.

Without the perceptual contrastive loss $L_{\text{contrast}}$, there is no other loss to explicitly guide the object co-segmentation network to predict object masks. For object co-segmentation, significant performance drops of 51% in precision and 32% in Jaccard index occur since our model no longer segments the co-occurrent objects in an image collection even though the cross-network consistency loss $L_{\text{task-consis}}$ facilitates supervision (i.e., dense correspondence field) for the output of the decoder $D$. For semantic matching, since our model does not learn to perform object co-segmentation, our model suffers from the negative impact caused by background clutters, resulting in a 3.6% performance drop at $\alpha = 0.05$.

When the cross-network consistency loss $L_{\text{task-consis}}$ is turned off, the is no explicit supervision to enforce the predicted object masks to be geometrically consistent across images. For object co-segmentation, the model thus tends to segment only the discriminative parts of the objects as reflected in Figure 6, resulting in performance drops of 2.3% in precision and 2% in Jaccard index. For semantic matching, since the predicted object masks may not precisely highlight the entire objects, our model may not effectively suppress the impact caused by background clutters when incorporating such object masks. A 2.4% performance drop by our method when $\alpha$ is set to 0.05.

The ablation study for object co-segmentation demonstrates that the proposed cross-network consistency loss $L_{\text{task-consis}}$ and the perceptual contrastive loss $L_{\text{contrast}}$ are crucial to achieving high performance. On the other hand, the foreground-guided matching loss $L_{\text{matching}}$, forward-backward consistency loss $L_{\text{cycle-consis}}$, and transitivity consistency loss $L_{\text{trans-consis}}$ facilitate object co-segmentation. For semantic matching, the foreground-guided matching loss $L_{\text{matching}}$ and perceptual contrastive loss $L_{\text{contrast}}$ are important to our proposed method. On the other hand, the cross-network consistency loss $L_{\text{task-consis}}$, forward-backward consistency loss $L_{\text{cycle-consis}}$, and transitivity consistency loss $L_{\text{trans-consis}}$ are helpful for enhancing the generalization ability in semantic matching.

**Effect of cutoff threshold $m$.** To analyze the sensitivity of our model against the cutoff threshold $m$ in (8), we conduct sensitivity analysis on the TSS dataset [6] by varying the value of the cutoff threshold $m$. Table 7 shows the experimental results. When the cutoff threshold $m$ is set to 0, i.e., $d_{AB} = 0$ in (10), the model enforces only the inter-image foreground similarity in (9). Without enforcing intra-image figure-ground dissimilarity in (10), the model may not produce clean foreground-background separation. We observe the performance of the proposed model drops by 10.5% in precision $P$ and 15% in Jaccard index $J$. When increasing the cutoff threshold $m$ to 2, the results in both precision $P$ and Jaccard index $J$ are significantly improved. When further increasing the cutoff threshold $m$ from 2 to 5 or 10, minimizing the perceptual contrastive loss $L_{\text{contrast}}$ is dominated by maximizing the foreground-background distinctness. The performance of our model drops instead. Introducing the cutoff threshold $m$ can considerably enhance the perceptual contrastive loss $L_{\text{contrast}}$ by setting the cutoff threshold $m$ to an appropriate value.

**Sensitivity analysis.** We analyze the performance of the proposed model by varying the value of each hyper-parameter on the PF-PASCAL [7] validation set for semantic matching, and on the TSS validation set for object co-segmentation. Figure 9 presents the experimental results on sensitivity analysis. For semantic matching, we report the results at PCK threshold $\alpha = 0.1$. When each of the hyper-parameter is set to 0 (i.e., the corresponding loss function is turned off), our model suffers from performance drops. When the individual hyper-parameters are set within a reasonable range, the performance is improved significantly. These results show that each loss function contributes to our method. However, when the hyper-parameter is set to a large value, e.g., 1,000, the corresponding loss term dominates the full training objective in (1), leading to significant performance drop. Similar conclusions can be drawn on the object co-segmentation task.

**4.6 Run-time analysis**

Given 800 images of the TSS dataset for joint semantic matching and object co-segmentation, it takes 280 minutes on a machine with an Intel i7 3.4 GHz processor and a single NVIDIA GeForce GTX 1080 graphics card with 11GB memory. The average run-time for processing each image in the set is 21 seconds.
We propose a weakly-supervised and end-to-end trainable network for joint semantic matching and object co-segmentation.

The core technical novelty lies in the coupled training of both tasks. We introduce a cross-network consistency loss to encourage the two-stream network to produce a consistent explanation of the given image pair. The network training requires only weak image-level supervision, making the proposed method scalable to real-world applications.

Through joint optimization, semantic matching is improved owing to the object masks revealed by object co-segmentation, while object co-segmentation is enhanced by referring to cross-image geometric transformations estimated during semantic matching. Experimental results demonstrate that our approach achieves the state-of-the-art performance for semantic matching and object co-segmentation.

5 CONCLUSIONS

We propose a weakly-supervised and end-to-end trainable network for joint semantic matching and object co-segmentation.

The core technical novelty lies in the coupled training of both tasks. We introduce a cross-network consistency loss to encourage the two-stream network to produce a consistent explanation of the given image pair. The network training requires only weak image-level supervision, making the proposed method scalable to real-world applications. Through joint optimization, semantic matching is improved owing to the object masks revealed by object co-segmentation, while object co-segmentation is enhanced by referring to cross-image geometric transformations estimated during semantic matching. Experimental results demonstrate that our approach achieves the state-of-the-art performance for semantic matching and object co-segmentation.

REFERENCES

[1] C. Liu, J. Yuen, and A. Torralba, “Sift flow: Dense correspondence across scenes and its applications,” TPAMI, 2011.

[2] T. Shen, G. Lin, L. Liu, C. Shen, and I. Reid, “Weakly supervised semantic segmentation based on co-segmentation,” in BMVC, 2017.

[3] A. Mustafa and A. Hilton, “Semantically coherent co-segmentation and reconstruction of dynamic scenes,” in CVPR, 2017.

[4] W. Zhou, H. Li, and Q. Tian, “Recent advance in content-based image retrieval: A literature survey,” arXiv, 2017.

[5] K. Dale, M. K. Johnson, K. Sunkavalli, W. Matusik, and H. Pfister, “Image restoration using online photo collections,” in ICCV, 2009.

[6] T. Tanai, S. N. Sinha, and Y. Sato, “Joint recovery of dense correspondence and cosegmentation in two images,” in CVPR, 2016.

[7] B. Ham, M. Cho, C. Schmid, and J. Ponce, “Proposal flow: Semantic correspondences from object proposals,” TPAMI, 2017.

[8] E. Tola, V. Lepetit, and P. Fua, “Daisy: An efficient dense descriptor applied to wide-baseline stereo,” TPAMI, 2010.

[9] H.-S. Chang and Y.-C. Wang, “Optimizing the decomposition for multiple foreground cosegmentation,” CVIU, 2015.

[10] K. Jerripothula, J. Cai, and J. Yuan, “Image co-segmentation via saliency co-fusion,” TMM, 2016.

[11] R. Quan, J. Han, D. Zhang, and F. Nie, “Object co-segmentation via graph optimized-flexible manifold ranking,” in CVPR, 2016.

[12] C. Wang, H. Zhang, L. Yang, X. Cao, and H. Xiong, “Multiple semantic matching on augmented n-partite graph for object co-segmentation,” TIP, 2017.

[13] K. Han, R. S. Rezende, B. Ham, K.-Y. K. Wong, M. Cho, C. Schmid, and J. Ponce, “S noe: Learning semantic correspondence,” in ICCV, 2017.

[14] B. C. Choy, J. Gwak, S. Savarese, and M. Chandraker, “Universal correspondence network,” in NeurIPS, 2016.

[15] I. Rocco, R. Arandjelovic, and J. Sivic, “Convolutional neural network architecture for geometric matching,” in CVPR, 2017.

[16] Z. Yuan, T. Lu, and Y. Wu, “Deep-dense conditional random fields for object co-segmentation,” in IJCAI, 2017.

[17] W. Li, O. H. Jafari, and C. Rother, “Deep object co-segmentation,” in ACCV, 2018.

[18] I. Rocco, R. Arandjelovic, and J. Sivic, “End-to-end weakly-supervised semantic alignment,” in CVPR, 2018.

[19] I. Rocco, M. Cimpoi, R. Arandjelovic, A. Torii, T. Pajdla, and J. Sivic, “Neighbourhood consensus networks,” in NeurIPS, 2018.

[20] S. Kim, S. Lin, S. R. JEON, D. Min, and K. Sohn, “Recurrent transformer networks for semantic correspondence,” in NeurIPS, 2018.

[21] K.-J. Hsu, Y.-Y. Lin, and Y.-Y. Chuang, “Co-attention cnns for unsupervised object co-segmentation,” in IJCAI, 2018.

[22] C. Rother, T. Minka, A. Blake, and V. Kolmogorov, “Cosegmentation of image pairs by histogram matching-incorporating a global constraint into mfs,” in CVPR, 2006.

[23] M. Rubinstein, A. Joulin, J. Kopf, and C. Liu, “Unsupervised joint object discovery and segmentation in internet images,” in CVPR, 2013.

[24] B. Ham, M. Cho, C. Schmid, and J. Ponce, “Proposal flow,” in CVPR, 2016.

[25] Y.-T. Hu and Y.-Y. Lin, “Progressive feature matching with alternate descriptor selection and correspondence enrichment,” in CVPR, 2016.

[26] Y.-T. Hu, Y.-Y. Lin, H.-Y. Chen, K.-J. Hsu, and B.-Y. Chen, “Matching images with multiple descriptors: An unsupervised approach for locally adaptive descriptor selection,” TIP, 2015.

[27] K.-J. Hsu, Y.-Y. Lin, and Y.-Y. Chuang, “Robust image alignment with multiple feature descriptors and matching-guided neighborhoods,” in CVPR, 2015.

[28] J. Kim, C. Liu, F. Sha, and K. Grauman, “Deformable spatial pyramid matching for fast dense correspondences,” in CVPR, 2013.

[29] D. Novotny, D. Larlus, and A. Vedaldi, “Anchornet: A weakly-supervised network to learn geometry-sensitive features for semantic matching,” in CVPR, 2017.

[30] F. Yang, X. Li, H. Cheng, J. Li, and L. Chen, “Object-aware dense semantic correspondence,” in CVPR, 2017.

[31] N. Ufer and B. Ommer, “Deep semantic feature matching,” in CVPR, 2017.

[32] O. Duchenne, A. Joulin, and J. Ponce, “A graph-matching kernel for object categorization,” in ICCV, 2011.

[33] H. Yang, W.-Y. Lin, and J. Lu, “Daisy filter flow: A generalized discrete approach to dense correspondences,” in CVPR, 2014.

[34] S. Kim, D. Min, B. Ham, S. Jeon, S. Lin, and K. Sohn, “Fcss: Fully convolutional self-similarity for dense semantic correspondence,” in CVPR, 2017.

[35] S. Kim, D. Min, B. Ham, S. Lin, and K. Sohn, “Fcss: Fully convolutional self-similarity for dense semantic correspondence,” in TPAMI, 2018.

[36] S. Kim, D. Min, S. Lin, and K. Sohn, “Discrete-continuous transformation matching for dense semantic correspondence,” TPAMI, 2018.

[37] S. Kim, D. Min, S. Lin, and K. Sohn, “Dctm: Discrete-continuous transformation matching for semantic flow,” in ICCV, 2017.

[38] S. Jeon, S. Kim, D. Min, and K. Sohn, “Parn: Pyramid affine regression networks for dense semantic correspondence estimation,” in ECCV, 2018.

[39] I. Rocco, M. Cimpoi, R. Arandjelovic, A. Torii, T. Pajdla, and J. Sivic, “Neighbourhood consensus networks,” in NeurIPS, 2018.

[40] S. Kim, S. Lin, S. R. JEON, D. Min, and K. Sohn, “Recurrent transformer networks for semantic correspondence,” in NeurIPS, 2018.

[41] A. Kanazawa, D. W. Jacobs, and M. Chandraker, “Warpnet: Weakly supervised matching for single-view reconstruction,” in CVPR, 2016.

[42] P. Hongshuck Seo, J. Lee, D. Jung, B. Han, and M. Cho, “Attentive semantic alignment with offset-aware correlation kernels,” in ECCV, 2018.

[43] U. Gaur and B. Manjunath, “Weakly supervised manifold learning for dense semantic object correspondence,” in ICCV, 2017.

[44] D. Novotny, S. Albanie, D. Larlus, and A. Vedaldi, “Self-supervised learning of geometrically stable features through probabilistic introspection,” in CVPR, 2018.

[45] U. Rfi, J. Gall, and B. Leibe, “Direct shot correspondence matching,” in BMVC, 2018.

[46] Z. Laskar and J. Kannala, “Semi-supervised semantic matching,” in ECCV, 2018.

[47] D. Lowe, “Object recognition from local scale-invariant features,” in ICCV, 1999.

[48] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in CVPR, 2005.

[49] J. Shotton, J. Winn, C. Rother, and A. Criminisi, “Textonboost for object categorization,” in TPAMI, 2015.

[50] C. Lee, W.-D. Jang, J.-Y. Sim, and C.-S. Kim, “Multiple random walkers and their application to image cosegmentation,” in TPAMI, 2017.
[54] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” in ICCV, 2017.

[55] H.-Y. Lee, H.-Y. Tseng, J.-B. Huang, M. Singh, and M.-H. Yang, “Difference image-to-image translation via disentangled representations,” in ECCV, 2018.

[56] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, “Multimodal unsupervised image-to-image translation,” in ECCV, 2018.

[57] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, “Unsupervised learning of depth and ego-motion from video,” in CVPR, 2017.

[58] Y. Zou, Z. Luo, and J.-B. Huang, “Di-net: Unsupervised joint learning of depth and flow using cross-task consistency,” in ECCV, 2018.

[59] S. Tulsiani, A. A. Efros, and J. Malik, “Multi-view consistency as supervisory signal for learning shape and pose prediction,” in CVPR, 2018.

[60] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” in CVPR, 2017.

[61] Y.-C. Chen, Y.-Y. Lin, M.-H. Yang, and J.-B. Huang, “Crdoco: Pixel-level domain transfer with cross-domain consistency,” in CVPR, 2019.

[62] S. Meister, J. Hur, and S. Roth, “Unflow: Unsupervised learning of optical flow with a bidirectional census loss,” in AAAI, 2018.

[63] W.-S. Lai, J.-B. Huang, O. Wang, E. Shechtman, E. Yumer, and M.-H. Yang, “Learning blind video temporal consistency,” in ECCV, 2018.

[64] T. Zhou, Y. Jae Lee, S. X. Yu, and A. A. Efros, “Flowweb: Joint image set alignment by weaving consistent, pixel-wise correspondences,” in CVPR, 2015.

[65] X. Zhou, M. Zhu, and K. Daniilidis, “Multi-image matching via fast alternating minimization,” in ICCV, 2015.

[66] T. Zhou, P. Krähenbühl, M. Aubry, Q. Huang, and A. A. Efros, “Learning dense correspondence via 3d-guided cycle consistency,” in CVPR, 2016.

[67] P. Fischer, A. Dosovitskiy, I. Ilg, P. Häusser, C. Hazırbaş, V. Golkov, P. Van der Smagt, D. Cremers, and T. Brox, “Flownet: Learning optical flow with convolutional networks,” in ICCV, 2015.

[68] A. Kendall, H. Martirosyan, S. Dasgupta, P. Henry, R. Kennedy, A. Bachrach, and A. Bry, “End-to-end learning of geometry and context for deep stereo regression,” in ICCV, 2017.

[69] P.-H. Huang, K. Matzen, J. Kopf, N. Ahuja, and J.-B. Huang, “Deepmvs: Learning multi-view stereopsis,” in CVPR, 2018.

[70] Y.-T. Hu, J.-B. Huang, and A. G. Schwing, “Videomatch: Matching based video object segmentation,” in ECCV, 2018.

[71] P. Voigtlaender, Y. Chai, F. Schroff, H. Adam, B. Leibe, and L.-C. Chen, “Feelvos: Fast end-to-end embedding learning for video object segmentation,” in CVPR, 2019.

[72] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in CVPR, 2016.

[73] Y.-C. Chen, P.-H. Huang, L.-Y. Yu, J.-B. Huang, M.-H. Yang, and Y.-Y. Lin, “Deep semantic matching with foreground detection and cycle-consistency,” in ACCV, 2018.

[74] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in ICLR, 2014.

[75] Y. Yang and D. Ramanan, “Articulated human detection with flexible mixtures of parts,” TPAMI, 2013.

[76] A. Faktor and M. Irani, “Co-segmentation by composition,” in ICCV, 2013.

[77] A. Joulin, F. Bach, and J. Ponce, “Discriminative clustering for image co-segmentation,” in CVPR, 2010.

[78] K. Jerripothula, J. Cai, J. Lu, and J. Yuan, “Object co-skeletonization with co-segmentation,” in CVPR, 2017.

[79] E. Shechtman and M. Irani, “Matching local self-similarities across images and videos,” in CVPR, 2007.

[80] S. Kim, D. Min, B. Ham, S. Ryu, M. N. Do, and K. Sohn, “Dasc: Dense adaptive self-correlation descriptor for multi-modal and multi-spectral correspondence,” in CVPR, 2015.

[81] X. Han, T. Leung, Y. Jia, R. Sukthankar, and A. C. Berg, “Matchnet: Unifying feature and metric learning for patch-based matching,” in CVPR, 2015.

[82] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in NeurIPS, 2012.

[83] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” in CVPR, 2015.

[84] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” arXiv, 2014.

[85] A. Hati, S. Chaudhuri, and R. Velmurugan, “Image co-segmentation using maximum common subgraph matching and region growing,” in ECCV, 2018.

[86] J. Sun and J. Ponce, “Learning dictionary of discriminative part detectors for image categorization and cosegmentation,” in IJCV, 2016.

[87] G. Kim, E. P. Xing, L. Fei-Fei, and T. Kanade, “Distributed cosegmentation via submodular optimization on anisotropic diffusion,” in ICCV, 2011.

[88] X. Chen, A. Shrivastava, and A. Gupta, “Enriching visual knowledge bases via object discovery and segmentation,” in CVPR, 2014.

Yun-Chun Chen received the B.S. degree in electrical engineering from National Taiwan University, Taipei, Taiwan in 2018. His current research interests include computer vision, deep learning, and machine learning.

Yen-Yu Lin received the B.B.A. degree in information management, and the M.S. and Ph.D. degrees in computer science and information engineering from National Taiwan University, in 2001, 2003, and 2010, respectively. He is currently an Associate Research Fellow with the Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan. His current research interests include computer vision, machine learning, and artificial intelligence.

Ming-Hsuan Yang is a Professor in Electrical Engineering and Computer Science at University of California, Merced. He served as an associate editor of the IEEE Transactions on Pattern Analysis and Machine Intelligence from 2007 to 2011, and is an associate editor of the International Journal of Computer Vision, Image and Vision Computing and Journal of Artificial Intelligence Research. Yang received the NSF CAREER award in 2012 and the Google Faculty Award in 2009. He is a Fellow of the IEEE.

Jia-Bin Huang is an assistant professor in the Bradley Department of Electrical and Computer Engineering at Virginia Tech. He received the B.S. degree in Electronics Engineering from National Chiao-Tung University, Hsinchu, Taiwan and his Ph.D. degree in the Department of Electrical and Computer Engineering at University of Illinois, Urbana-Champaign in 2016.