Multiview Co-segmentation for Wide Baseline Images using Cross-view Supervision

Yuan Yao
University of Minnesota
yaoxx340@umn.edu

Hyun Soo Park
University of Minnesota
hspark@umn.edu

Figure 1: This paper presents a semi-supervised learning framework for an object co-segmentation task from multiview images. In particular, we consider *wide baseline* images where photometric matching does not apply. We formulate a novel cross-view self-supervision method to transfer a segmentation mask from one view to the other. This allows us to effectively segment foreground objects with the limited number of labeled images instances including monkey, Indian dancer, and a public dataset of social videos captured by handheld cameras.

Abstract

This paper presents a method to co-segment an object from wide baseline multiview images using cross-view self-supervision. A key challenge in the wide baseline images lies in the fragility of photometric matching. Inspired by shape-from-silhouette that does not require photometric matching, we formulate a new theory of shape belief transfer—the segmentation belief in one image can be used to predict that of the other image through epipolar geometry. This formulation is differentiable, and therefore, an end-to-end training is possible. We analyze the shape belief transfer to identify the theoretical upper and lower bounds of the unlabeled data segmentation, which characterizes the degenerate cases of co-segmentation. We design a novel triple network that embeds this shape belief transfer, which is agnostic to visual appearance and baseline. The resulting network is validated by recognizing a target object from realworld visual data including non-human species and a subject of interest in social videos where attaining large-scale annotated data is challenging.

1. Introduction

This paper addresses the problem of co-segmentation for a novel object class from a set of multiview images. In particular, we consider *wide baseline*[1] images where photometric matching across views is highly fragile without a non-trivial scene assumption [10, 37, 64]. This problem setting of the wide baseline reflects the nature of the practical multi-camera deployment in our daily lives. For instance, there is an emerging trend of social videos [3, 5, 22, 53]—a collection of videos that record an activity of interest (e.g., political rally, concert, and wedding) from social members at the same time[2]. The camera placement of such social videos are, by definition, driven by mobile users who behave in accordance with the social norm of proxemics [27], which naturally produces multiview images with wide baseline (bottom row of Fig. 1). Further, the capability of co-segmenting a novel object class for wide baseline cameras

[1]The baseline is defined in relation to the depth of an object, i.e., incident angle of triangulation [56].

[2]There exist multiple online repositories such as Rashomon Project [1] and CrowdSync cellphone app [2] that host the social videos.

enables the volumetric reconstruction of non-human species such as monkeys, which gives rise to significant scientific impact [35, 81] (top row of Fig. 1).

However, co-segmentation of wide baseline images involves with three matching challenges. (1) Local visual features are highly fragile for establishing correspondences between wide baseline images, and thus, geometric constraints (e.g., epipolar geometry, multiview stereo [37], and 3D volumetric reasoning [10, 64]) cannot be used to validate multiview segmentation. (2) Graph matching using the spatial relationship between superpixels [11, 36, 57, 59, 69] is prone to severe self-occlusion caused by the huge view difference. Further, the spurious superpixels due to moving occluding boundary makes the photometric matching unreliable. (3) Recognition based matching, e.g., semantic segmentation [14, 48] can only apply to the object classes that belong to the existing datasets such as MS COCO [47], i.e., a novel object (e.g., monkey) co-segmentation is impossible without a major modification of the dataset.

To address these matching challenges, we propose to learn the shape of object through cross-view self-supervision. A key innovation is that integrating the multiview geometric constraint into the segmentation task in a differentiable fashion, resulting in end-to-end training. We derive a new formulation of *shape belief transfer*—the segmentation belief in one image can be used to predict that of the other image through epipolar geometry. In fact, this is an inverse problem of shape-from-silhouette [21, 26, 41, 42] that reconstructs a 3D object volume (visual hull) from the foreground segmentation of multiview images without explicit photometric matches [39, 43, 50]. The shape belief transfer is a composition of two belief transfers: (a) 3D shape reconstruction by triangulating the segmentation map (confidence) in multiview source images; and (b) 2D projection of the reconstructed 3D shape onto a target view to approximate its segmentation map.

We characterize the shape belief transfer, providing the theoretical upper and lower bounds of unlabeled data segmentation: its gap approaches asymptotically to zero as the number of labeled views increases. We further show that the shape belief transfer can be implemented by transforming the operation of 2D projection to max-pooling operation, which allows bypassing 3D shape reconstruction that has been used in existing approaches [7, 8, 34, 63, 68, 70, 78] for cross-view supervision. Based on the theory, we design a triplet network that takes as input multiview image with the limited number of the labeled data and outputs the object segmentation on unlabeled data as shown in Fig. 1. The network is trained by minimizing the geometric inconsistency of multiview segmentation.

This framework is flexible: (1) segmentations can be customized as it does not require a pre-trained model, i.e., we train a segmentation model from scratch with manual annotations for each sequence; (2) it can be built on any segmentation network design such as DeepLab [13], SegNet [4], and Mask R-CNN [29] that outputs an object segmentation confidence; (3) it can apply to general multi-camera systems including social videos (e.g., different multi-camera rigs, number of cameras, and intrinsic parameters).

The core contributions of this paper include (I) a new formulation of a differentiable shape belief transfer to integrate multiview geometry into the object segmentation task; (II) theoretical analysis of the shape belief transfer that characterizes degenerate cases; (III) a unique triple network design that embeds the shape belief transfer to perform cross-view supervision for unlabeled data in an end-to-end fashion; and (IV) application to realworld challenging visual data captured from wide baseline cameras, including non-human species and a subject of interest in social videos where attaining large-scale annotation data is infeasible. We quantitatively show that our approach with cross-view supervision consistently outperforms the the existing models.

2. Related Work

This work lies in the intersection between object co-segmentation and cross-view self-supervision, which enables learning from a small set of the labeled data possible. While there exist a large volume of literature on self-supervised segmentation such as temporal supervision on videos [19, 20, 51], we will focus on cross-view self-supervision.

Semi-supervised Segmentation To use image segmentation in practice requires a wide variety of object classes and a large number of annotations for each class. Moreover, the process of pixel-level labeling requires substantial manual efforts. This problem can be alleviated by semi-supervised settings, in which segmentation model is trained with weak labels that are much easier to obtain such as scene class [52, 54, 55] or bounding box [18] with a small amount of labeled data. An encoder-decoder framework is trained with large number of scene class level annotated data and a few fully-annotated data [30]. The adversarial discriminators [31] is used to differentiate the predicted probability maps from ground truth instead of being used to classify the input as real or fake. Therefore, this discriminator enables semi-supervised learning by finding the trustworthy regions in prediction of unlabeled data.

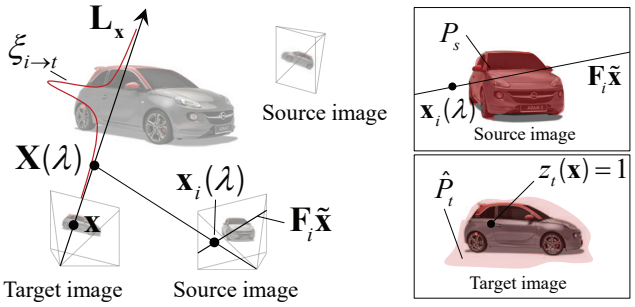
Co-segmentation Object co-segmentation is the task of detecting and segmenting the common objects from a group of images [74], which segments common parts in an image pair and by extension to more images [33, 69]. A deep dense conditional random field framework is applied on co-segmentation task in [75]. They use co-occurrence map to measure the objectness for object proposals, and the similarity evidence for proposals is generated by selective search which uses SIFT feature. Therefore, this is

not end-to-end training. An end-to-end training framework for co-segmentation is proposed in [45]. They present a CNN-based method to jointly detect and segment the common object from a pair of images. An attention based co-segmentation model is proposed in [12]. The model consists of an encoder, a semantic attention learner, and a decoder. The semantic attention learner takes the encoded features to learn to pay attention to the common objects.

Multiview Self-supervision Learning a view invariant representation is a long-standing goal in visual recognition research, which requires to predict underlying 3D structure from a single view image. Geometrically, it is an ill-posed problem while two data driven approaches have made promising progress. (1) Direct 3D-2D supervision: for a few representative objects such as furniture [46], vehicles [71], and human body [49], their 3D models (e.g., CAD, point cloud, and mesh) exist where the 3D-2D relationship can be directly regressed. The 3D models can produce a large image dataset by projecting onto all possible virtual viewpoints where the object's pose and shape can be learned from 3D-2D pairs. This 3D model projection can be generalized to scenes measured by RGBD data [8, 24, 32, 40, 65, 70] and graphically generated photo-realistic scenes [15, 58] where visual semantics associated with 3D shape can be encoded. (2) Indirect supervision via non-rigid graph matching: to some extent, it is possible to infer the common shape and appearance from a set of single view image instances without 3D supervision. For instance, tables have a common shape expressed by four legs and planar top. Such holistic spatial relationship can be unveiled by casting it as a graph matching problem where local shape rigidity and appearance models can describe the relationship between nodes and edges [6, 9, 16, 44, 66, 77]. Further, leveraging a underlying geometric constraint between instances (e.g., cyclic consistency [79, 80], volumetric projection [17, 67, 68], and kinematic chain [62, 66, 72]) can extend the validity of graph matching. These existing approaches require many correspondences between domains that are established by manual annotations. In contrast, our approach will leverage self-supervision via multiview geometry to adapt to a novel scene with minimal manual efforts.

3. Multiview Cross-view Supervision

We present a semi-supervised learning framework to train an object segmentation model (network) by leveraging unlabeled multiview images with wide baseline where the amount of unlabeled data is larger than that of labeled data ($< 4\%$ of unlabeled data). Consider a segmentation network that takes an input image $\mathcal{I}$ and outputs a per-pixel object confidence, i.e., $\phi(\mathcal{I}; \mathbf{w}) \in [0, 1]^{W \times H \times 2}$ where W and H are the width and height of the output distribution, respectively. This is equivalent to a binary segmentation



(a) Cross-view supervision (b) Shape transfer

Figure 2: (a) Inspired by shape-from-silhouette, we supervise object segmentation in the target image from source images. (b) The shape is transferred through epipolar geometry to the target image. Note that the transferred shape $\widehat{P}_t$ is always bigger than the true shape.

(object and background). The network is parametrized by the weight $\mathbf{w}$ learned by minimizing the following loss:

$$\underset{\mathbf{w}}{\text{minimize}} \;\; \mathcal{L}_L + \lambda_s \mathcal{L}_S + \lambda_p \mathcal{L}_P, \tag{1}$$

where $\mathcal{L}_L$, $\mathcal{L}_S$, and $\mathcal{L}_P$ are the losses for labeled supervision, cross-view supervision, and bootstrapping prior, and λ_s and λ_p are the weights that control their importance.

For the labeled data $\mathcal{D}_L$, we use the sum of pixelwise cross entropy to measure the segmentation loss:

$$\mathcal{L}_L = -\sum_{j \in \mathcal{D}_L} \sum_{\mathbf{x} \in X} y_j(\mathbf{x}) \log \phi(\mathcal{I}_j)|_\mathbf{x}\,, \tag{2}$$

where $y_j(\mathbf{x}) \in \{0, 1\}$ is the ground truth label of the j^{th} labeled data at pixel location $\mathbf{x}$, and X is the domain of $\mathbf{x}$.

3.1. Shape Transfer

Inspired by the image-based shape-from-silhouette [50], we formulate a method of cross-view supervision for co-segmentation using 3D *shape belief transfer*. Consider a point $\mathbf{x} \in \mathbb{R}^2$ in the target image $\mathcal{I}_t$. Without loss of generality, the camera projection matrix of the target image is set to $\mathbf{P} = \mathbf{K} \begin{bmatrix} \mathbf{I}_3 & \mathbf{0} \end{bmatrix}$ where $\mathbf{K}$ is the intrinsic parameter. The point in an image is equivalent to a 3D ray $\mathbf{L}_\mathbf{x} = \mathbf{K}^{-1}\tilde{\mathbf{x}}$ emitted from the target camera where $\tilde{\mathbf{x}}$ is the homogeneous representation of $\mathbf{x}$ [28]. A 3D point along the ray can be represented as $\mathbf{X}(\lambda) = \lambda \mathbf{L}_\mathbf{x}$ where any scalar depth $\lambda > 0$.

A series of projections of $\mathbf{X}(\lambda)$ onto a source image, $\mathcal{I}_{s_1}$ form the epipolar line $\mathbf{l}_1 = \mathbf{F}_1\tilde{\mathbf{x}}$ where $\mathbf{F}_1$ is the fundamental matrix between the target and source image. This indicates the point on the epipolar line can be parametrized by λ as shown in Fig. 2(a), i.e., $\mathbf{x}_1(\lambda) \in \mathbf{l}_1$[3]. Likewise a point $\mathbf{x}_i$ in the i^{th} source image $\mathcal{I}_i$ can be described accordingly.

[3]We use an abuse of notation: $\mathbf{x} \in \mathbf{l}$ is equivalent to $\tilde{\mathbf{x}}^\mathsf{T}\mathbf{l} = 0$, i.e., the point $\mathbf{x}$ belongs to the line $\mathbf{l}$

The image-based shape-from-silhouette computes a binary map $z_t : \mathbb{R}^2 \rightarrow \{0, 1\}$ that determines the pixel belong to object if one, and zero otherwise. This binary map can be approximated by the logical operations between the binary maps from the n source images $(z_{s_1}, \cdots, z_{s_n})$:

$$\hat{z}_t(\mathbf{x}) = \begin{cases} 1 & \text{if } \exists\, \lambda > 0 \text{ s.t. } \bigwedge_i z_{s_i}(\mathbf{x}_i(\lambda)) = 1 \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

The geometric interpretation of Eq. (3) is that the object map for $\mathbf{x}$ is computed by sweeping across all 3D points along the ray $\mathbf{L}_\mathbf{x}$ to see if the ray intersects with the 3D volumetric shape defined by the foreground maps from n views. This shape-from-silhouette $\hat{z}_t(\mathbf{x})$ from n views is always inclusive of the true object $z_t(\mathbf{x})$, i.e., $\{\mathbf{x}|z_t(\mathbf{x}) = 1\} \subseteq \{\mathbf{x}|\hat{z}_t(\mathbf{x}) = 1\}$ as shown in Fig. 2(b).

The implication of Eq. (3) is significant for cross-view supervision for unlabeled data because it is possible to transfer a shape belief in one image to another. Let $P_i : \mathbb{R}^2 \rightarrow [0, 1]$ be the segmentation map (confidence) of the i^{th} source image, i.e., $P_i(\mathbf{x}) = \phi(\mathcal{I}_i; \mathbf{w})|_\mathbf{x}$. The distribution over the ray $\mathbf{L}_\mathbf{x}$ emitted from the target image can be computed by projecting the ray onto the i^{th} image:

$$\xi_{i\rightarrow t}(\lambda; \mathbf{L}_\mathbf{x}) = P_i(\mathbf{x}_i(\lambda)) \;\text{ where }\; \mathbf{x}_i(\lambda) \in \mathbf{F}_i\widetilde{\mathbf{x}}, \tag{4}$$

where $\xi_{i\rightarrow t}(\lambda; \mathbf{L}_\mathbf{x})$ is object confidence (distribution) over the ray parametrized by the depth λ.

From Eq. (4), the object confidence in the target image $P_t : \mathbb{R}^2 \rightarrow [0, 1]$ can be approximated by a 3D line max-pooling over joint probability over n views:

$$\hat{P}_t(\mathbf{x}) = \sup_{\lambda>0} \prod_{i=1}^{n} \xi_{i\rightarrow t}(\lambda; \mathbf{L}_\mathbf{x}), \tag{5}$$

where $\hat{P}_t(\mathbf{x})$ is the object confidence transferred from n views. Eq. (5) is equivalent to Eq. (3) where it takes the probability of a 3D point most likely being in the volumetric shape (Fig. 2).

Note that similar to $\hat{z}_t$, the $\hat{P}_t$ provides the upper bound of the P_t, i.e., $\{\mathbf{x}|P_t(\mathbf{x}) > \epsilon\} \subseteq \{\mathbf{x}|\hat{P}_t(\mathbf{x}) > \epsilon\}$. Therefore, direct distribution matching using KL divergence [38] does not apply. Instead, we formulate a new loss D_S using one-way relative cross-entropy as follow:

$$\mathcal{L}_S = D_S(P_t||\hat{P}_t) = \sum_{\mathbf{x}\in X} (1 - \hat{P}_t(\mathbf{x}))P_t(\mathbf{x}), \tag{6}$$

where X is the range of the target image coordinate. Note that the distance measure is not symmetric. It penalizes only the set of pixels $\{\mathbf{x}|\hat{P}_t(\mathbf{x}) < P_t(\mathbf{x})\}$.

The main benefits of Eq. (6) are threefold. (1) Multi-view segmentation involves two processes: 3D reconstruction of the shape with source views and 2D projection onto the target view. The requirement of 3D reconstruction introduces an additional estimation such as multiview [20,37] or single view depth prediction [34,68,78] where the accuracy of the segmentation is bounded by the reconstruction quality. Eq. (6) integrates the 3D reconstruction and projection through the joint probability over the epipolar lines and supremum operation, which bypass the 3D reconstruction. (2) By minimizing Eq. (6), it can provide a *pseudo*-label for unlabeled data transferred from labeled data. As the number of labeled data increases, the transferred segmentation label approaches to the true label of unlabeled data [39,50], which allows cross-view supervision, i.e., segmentation in a label image can supervise that in an unlabeled image. (3) Not only for unlabeled data, but also it can correct the geometrically inconsistent object segmentation for labeled data. This is a significant departure from the existing object co-segmentation that cannot recover erroneous segmentation label, which often arises from per-pixel manual annotations.

3.2. Cross-view Supervision via Shape Transfer

In practice, embedding Eq. (6) into an end-to-end neural network is not trivial because (a) a new max-pooling operation over oblique epipolar lines needs to be defined; (b) sampling interval for max-pooling along the line is arbitrary, i.e., uniform sampling does not encode geometric meaning such as depth; and (c) sampling interval across different epipolar line parameters is also arbitrary, which may introduce sampling artifacts.

We introduce a new operation inspired by stereo rectification, which warps the segmentation confidence such that the epipolar lines become parallel (horizontal). This rectification allows converting the oblique line max-pooling into regular row-wise max-pooling.

Eq. (4) can be re-written by rectifying the segmentation confidence of the source view with respect to the target view:

$$\overline{\xi}_{1\rightarrow t}(u; \mathbf{L}_\mathbf{x}) = \overline{P}_1\left(\begin{bmatrix} u \\ v_1 \end{bmatrix}\right), \text{ s.t. } \mathbf{K}\mathbf{R}_1\mathbf{K}^{-1}\widetilde{\mathbf{x}} \propto \begin{bmatrix} x \\ v_1 \\ 1 \end{bmatrix},$$

where $\mathbf{K}\mathbf{R}_1\mathbf{K}^{-1}\widetilde{\mathbf{x}}$ is the rectified coordinate of the target view, $\mathbf{R}_1 \in SO(3)$ is the relative rotation for the rectification. See Appendix for more details. Note that ξ is no longer a function of the depth scale λ but the x coordinate (disparity), which eliminates irregular sampling across pixels with the y coordinate v_1.

The key advantage of this rectification is that the x coordinate of the i^{th} view can still be parametrized by the same u, i.e., the coordinate is linearly transformed to from the first view to the rest views:

$$\overline{\xi}_{i\rightarrow t}(a_i u + b_i; \mathbf{L}_\mathbf{x}) = \overline{P}_i\left(\begin{bmatrix} a_i u + b_i \\ v_i \end{bmatrix}\right)$$

where a_i and b_i are the linear re-scaling factor and bias between the first and i^{th} views accounting for camera intrinsic

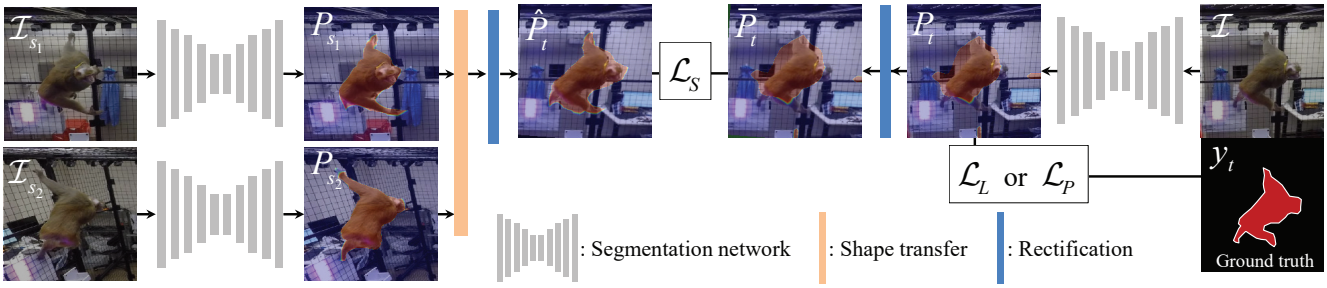


Figure 3: We design a novel triplet network to perform cross-view supervision via shape transfer. Two source images $\mathcal{I}_{s_1}$ and $\mathcal{I}_{s_2}$ are fed into a segmentation network to produce the object confidence maps P_{s_1} and P_{s_2}, respectively. These shape are transferred to the target image with rectification to form $\widehat{P}_t$. This transferred shape supervises the object confidence map of the targe image P_t. The label loss is measured for the target image if labeled. Otherwise, the bootstrapping loss is used.

and cropping parameters. $\overline{\phi}_i$ is computed by the rectified segmentation confidence of the i^{th} view $\overline{P}_i$ with respect to the target view. See Appendix for more details. This simplifies the supremum operation over the 3D ray in Eq. (5) to the max operation over the x coordinates:

$$\hat{P}_t(\mathbf{x}) = \max_{u\in[0,W]} \overline{\xi}_1(u;\mathbf{L}_\mathbf{x}) \prod_{i=2}^{n} \overline{\xi}_i(a_i u + b_i;\mathbf{L}_\mathbf{x}). \quad (7)$$

3.3. Bootstrapping Prior

Eq. (3) is often highly effective to generate a prior for 3D shape given the binary label. Inspired by multiview bootstrapping [63, 73], we approximate the 3D shape using the pre-trained neural network ϕ. Note that unlike keypoint detection, RANSAC [23] outlier rejection approaches cannot be applied because pixel correspondences are not available for semantic segmentation. We binarize the probability of the foreground segment to compute the i^{th} source binary map $z_{s_i}(\mathbf{x}) = 1$ if $P_i(\mathbf{x}) > 0.5$, and zero otherwise. Using all source binary maps, a pseudo-binary map for the j^{th} unlabeled data $\hat{z}_j$ can be computed and used for the bootstrapping prior, i.e.,

$$\mathcal{L}_P = \sum_{j\in\mathcal{D}_U} \sum_{\mathbf{x}\in X} (1 - \hat{z}_j(\mathbf{x}))P_j(\mathbf{x}) \quad (8)$$

Similar to Eq. (6), $\hat{z}_j$ provides the superset of the ground truth, which requires the one-way relative cross entropy as a prior loss.

3.4. Network Design

We design a novel triplet network that allows measuring three losses: $\mathcal{L}_L$, $\mathcal{L}_S$, and $\mathcal{L}_P$. Fig. 3 illustrates the overall design of the triplet. All subnetworks share their weights $\mathbf{w}$. Two source images $\mathcal{I}_{s_1}$ and $\mathcal{I}_{s_2}$ are fed into a segmentation network to produce the object confidence maps P_{s_1} and P_{s_2}, respectively. These two confidence maps are transferred to the target image by applying stereo rectification (Eq. (7)). This transferred and rectified confidence map $\widehat{P}_t$ is used to supervise the confidence map computed from the target image P_t. The cross-view loss $\mathcal{L}_S$ is measured by using Eq. (6). The label loss $\mathcal{L}_L$ is measured by comparing with the ground truth y_t if it is available. If the ground truth label is not available for the target image, the bootstrapping loss $\mathcal{L}_P$ is measured instead. All operations in the network is differentiable, and therefore, end-to-end training is possible. Fig. 4 is an example which shows that one target view can be supervised by multiple different image pairs during training in practice.

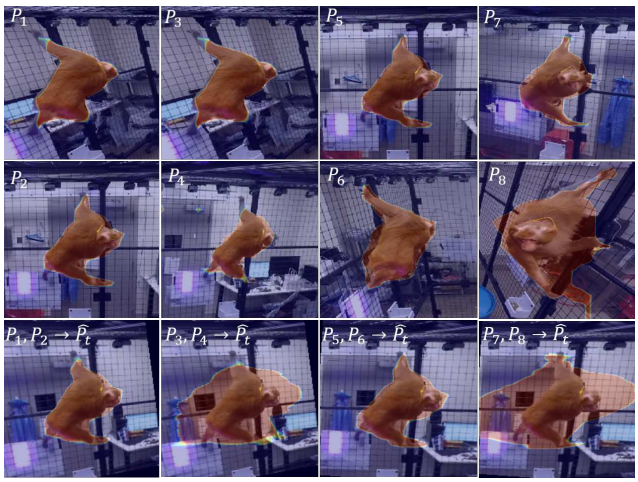


Figure 4: Different image pairs (top two rows) can be used to supervise one target view (bottom). We use such multiple triplets to supervise each other's view.

4. Degenerate Case Analysis

Eq. (6) has a degenerate case: a trivial solution $P_t = 0$ is the global minimizer. Therefore, when the unlabeled data sample is used for the target view, the cross-view supervision via shape transfer based on the labeled data is not possible, i.e., $\hat{P}_U = P_U^+ > P_U$.

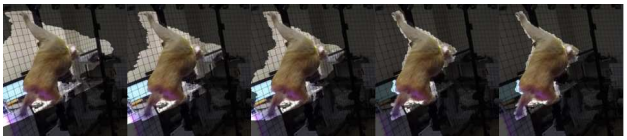


Figure 5: The upper bound of the probability for the unlabeled data becomes tighter as the number of the labeled data increases.

Theorem 1. *There exists the lower bound of the probability of the unlabeled data sample, P_U^-.*

Proof. Consider an inverse shape transfer for the unlabeled data in Eq. (5), $\phi_U(\lambda; \mathbf{L_x})$, to explain the first labeled data sample $\hat{P}_{L_1}$:

$$\hat{P}_{L_1}(\mathbf{x}) = \sup_{\lambda>0} \xi_U(\lambda; \mathbf{L_x}) \prod_{i=2}^{n} \xi_{L_i}(\lambda; \mathbf{L_x}), \tag{9}$$

where $\hat{P}_{L_i}$ is the probability of the i^{th} labeled data. Since the supremum in Equation (9) is a non-decreasing function with respect to $\xi_U(\lambda; \mathbf{L_x})$, there exists $\xi_U^-(\lambda; \mathbf{L_x}) < \xi_U(\lambda; \mathbf{L_x})$ that cannot explain $\hat{P}_{L_1}(\mathbf{x})$:

$$\hat{P}_{L_1}(\mathbf{x}) > \sup_{\lambda>0} \xi_U^-(\lambda; \mathbf{L_x}) \prod_{i=2}^{n} \xi_{L_i}(\lambda; \mathbf{L_x}). \tag{10}$$

Therefore, there exists the lower bound of P_U. □

From Theorem 1, Eq. (6) can provide both upper and lower bounds of the unlabeled data if used as the target and source views, i.e., $P_U^- < P_U \leq P_U^+$, and P_U^- asymptotically approaches to P_U^+ as the number of labeled views increases [39, 50], i.e., $\lim_{|\mathcal{D}_L|\to\infty}(P_U^+ - P_U^-) = 0$. Fig. 5 shows the upper bound becomes tighter as the number of labeled data increases

We leverage this asymptotic convergence of the shape transfer to self-supervise the unlabeled data, i.e., the unlabeled data are fed into both the target and source views, which allows the gradient induced by the error in the loss function of Eq. (6) can be backpropagated through the neural network to reduce the gap between P_U^+ and P_U^-.

5. Result

We evaluate our semi-supervised learning approach for an object co-segmentation task on realworld data where the number of annotations is limited.

Implementation We build a model per subject without a pre-trained model. The DeepLab v3 [13] network is used for our base network (segmentation network in Fig. 3). Each network takes as an input RGB image (200×200×3), and outputs two confidence maps (object and background) with the input size. We use the batch size 5, learning rate 10^{-5}, batch normalization with epsilon 10^{-5} and 0.9997. We use the ADAM optimizer of TensorFlow trained on a single NVIDIA GTX 1080. Fig. 6 illustrates the progression of training process for unlabeled data at every 2,000 iterations. The cross-view supervisionary signals are propagated through unlabeled data and eventually recognize the correct segment of monkey and Indian dancer.

Datasets We validate our semi-supervised semantic segmentation framework on multiple sequences of diverse realworld subjects and environments including monkeys, Indian dancer, and social videos captured in multi-camera systems. These cameras form wide baseline images, i.e., while physical distance between two cameras is small, the foreground object appears significantly different due to the short distance to the object as shown in Fig. 4. We used synchronized images extracted from multiview videos. These videos contain both dynamic and relatively static background, and the background subtraction methods [25, 76] fail on these videos. (1) **Monkey subject** 35 GoPro HD cameras running at 60 fps are installed in a large cage ($9' \times 12' \times 9'$) that allows the free-ranging behaviors of monkeys. There are diverse monkey activities include grooming, hanging, and walking. The monkey body was stretched to a variety of shapes and background constantly changed during activities. The camera produces 1280 × 960 images. We uniformly sampled 15 frames from total 1,800 frames, and for each frame, we randomly annotate 8 views as labeled data an the rest views are unlabeled data. (2) **Indian dancer** Multi-camera system composed of 69 synchronized HD cameras (1024×1280 at 30 fps) in three layers with different heights are used to capture the peformance of an Indian dancer. We uniformly sampled 17 frames from total 2,000 frames, and for each frame, we randomly annotate 8 views as labeled data an the rest views are unlabeled data. This dataset has most number of camera views and static background. (3) **Social videos** A public social video dataset [3] is used to validate wide baseline image co-segmentation. We focus on a sequence of break dancers who were surround by 16 audiences. They freely move around the dancers: the average distance between cameras is approximately 2m. GoPro cameras (1024×960 at 30 fps) are used for capture the performance. We use standard structure-from-motion algorithm [60, 61] to reconstruct the scene geometry and camera intrinsic and extrinsic. We uniformly sampled 24 frames from total 1,300 frames, and for each frame, we randomly annotate 7 views as labeled data an the rest views are unlabeled data.

Baselines We compare our approach with four different baseline algorithms. Since multiview object co-segmentation on wide baseline images is a new task, we adapt existing algorithms to our task with a minor modification. Note that the work by Kowdle et al. [37] is not compared because it builds upon multiview stereo, which is not applicable to wide baseline co-segmentation. For all

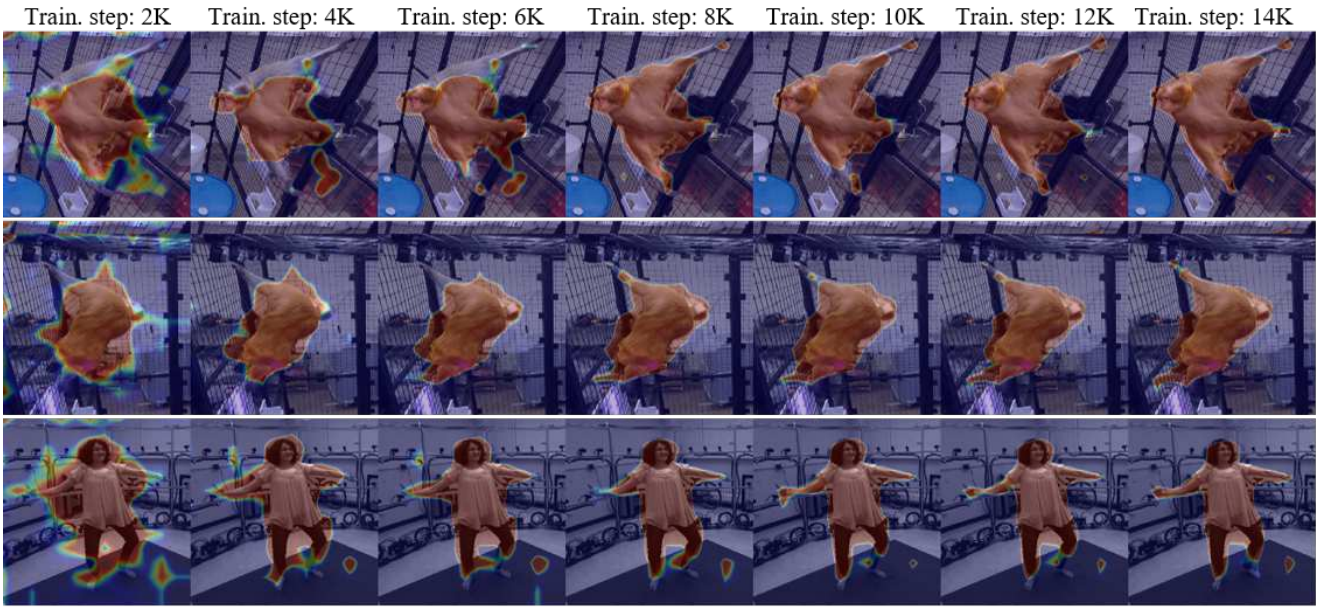


Figure 6: We visualize the prediction result of our semi-supervised framework on unlabeled data every 2000 training iterations.

	Monkey (IoU)			Dance (IoU)			Social (IoU)			Monkey (Pixel Acc.)			Dance (Pixel Acc.)			Social (Pixel Acc.)		
Number of labeled data	2	5	8	2	5	8	2	5	7	2	5	8	2	5	8	2	5	7
Supervised learning	0.73	0.76	0.77	0.66	0.71	0.74	0.70	0.73	0.73	0.91	0.92	0.93	0.84	0.87	0.89	0.85	**0.87**	0.87
Bootstrapping [63]	0.76	0.83	0.85	0.67	0.76	0.76	0.53	0.55	0.54	0.92	0.95	0.96	0.82	0.89	0.90	0.73	0.77	0.76
Attention-based [12]	0.72	0.77	0.83	0.17	0.31	0.33	0.31	0.51	0.44	0.93	0.94	**0.96**	0.68	0.71	0.73	0.70	0.72	0.73
Adversarial network [31]	0.78	0.81	0.82	0.55	0.78	**0.83**	0.35	0.52	0.60	0.83	0.84	0.84	0.72	0.74	0.75	0.68	0.71	0.72
Ours	**0.82**	**0.85**	**0.87**	**0.77**	**0.80**	0.81	**0.71**	**0.73**	**0.74**	**0.94**	**0.95**	0.96	**0.90**	**0.91**	**0.92**	**0.87**	**0.87**	**0.88**

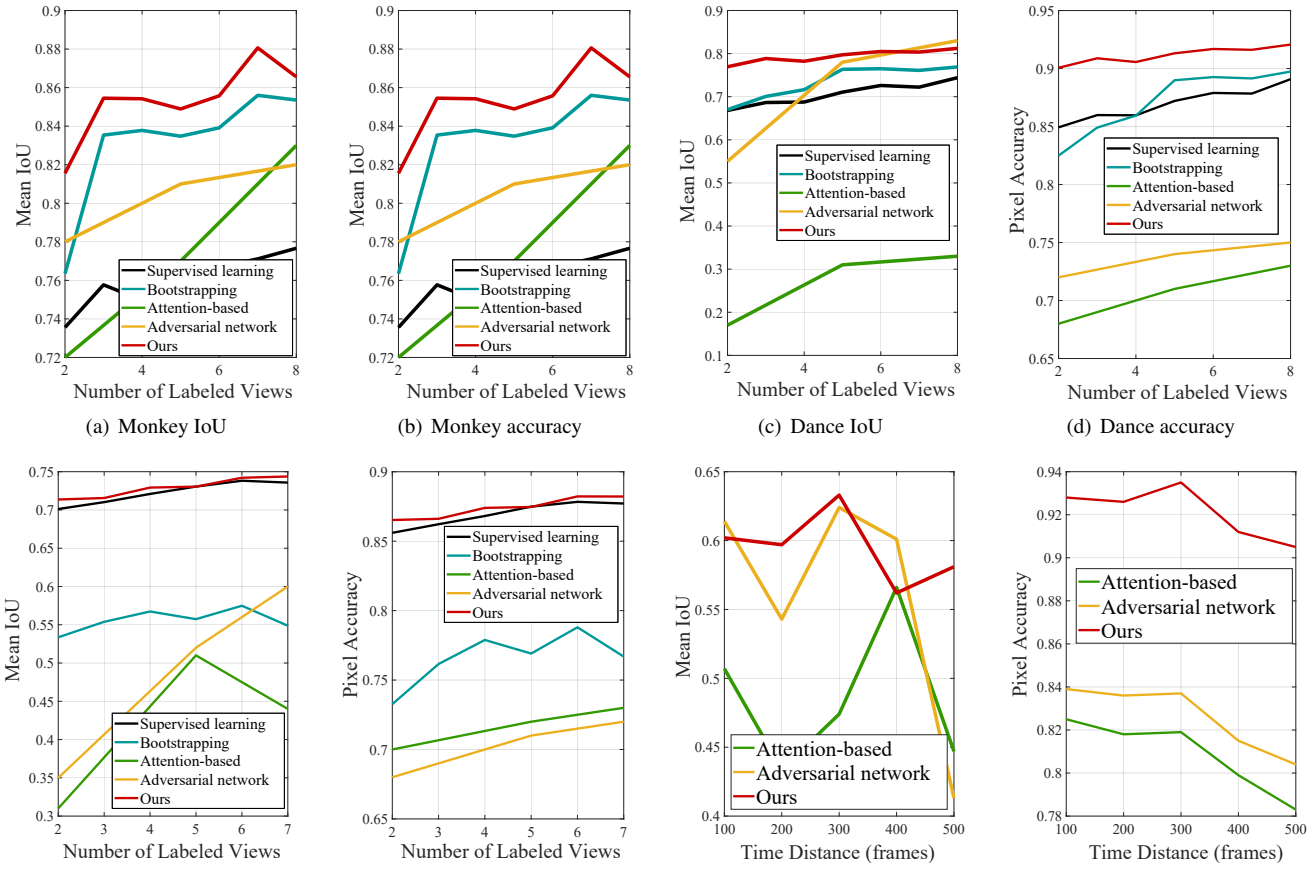
Table 1: Mean IoU and pixel accuracy result on different datasets with different number of labeled views

algorithms, we evaluate the performance on unlabeled data. (1) **Supervised learning**: we use the limited labeled data to train our base network [13]. This algorithm is not accessible to unlabeled data during the training. (2) **Bootstrapping**: we leverage labeled data to provide a bootstrapping prior [63] to unlabeled data (Section 3.3). This algorithm has an access to unlabeled data during training while it is highly biased to the bootstrapping. (3) **Attention-based co-segmentation**: state-of-the-art attention network is used to perform object co-segmentation task [12]. This network does not encode cross-view supervision. (4) **Adversarial segmentation**: state-of-the-art adversarial network [31] is used to segment an object in a semi-supervised fashion. For last two baselines, we use their publicly available algorithms without the pre-trained model.

Metric We evaluate our approach based on two metrics: mean IoU (intersection over union) and mean pixel accuracy.

Accuracy For each dataset, we manually annotate all the views in about 20 randomly selected frames sampled from videos as the test data. Note that sampled frames are not used for training. Figs. 7(a)–7(f) illustrate the performance comparison of our cross-view supervision with the baseline algorithms, and Table 1 summarizes mean IoU and accuracy. Since all semi-supervised learning have an access to unlabeled data during training, their performance on unlabeled data is superior to the supervised learning. Among semi-supervised learning frameworks, our approach that leverages cross-view supervision to transfer shape outperforms other approaches with a large margin. The attention-based co-segmentation and adversarial segmentation shows inferior performance comparing to bootstrapping approach as the baseline of the cameras are fairly wide where learning common visual semantics is difficult. We also evaluate the impact of the labeled data on performance. Note that in Figs. 7(e) and 7(f), for social camera data where cameras are moving, the supervised method is nearly on a par without our approach. We identified that the main source of performance degradation was the geometric inconsistency caused by errors in synchronization.

Label Data Sensitivity We conduct an experiment to identify the label data sensitivity, i.e., how the choice of labeled data matters. We measure the segmentation accuracy with respect to the distance to the labeled data in time. For in-



(a) Monkey IoU (b) Monkey accuracy (c) Dance IoU (d) Dance accuracy

(e) Social video IoU (f) Social video accuracy (g) Label sensitivity IoU (h) Label sensitivity accuracy

Figure 7: (a-f) We compare our approach with four baseline algorithms: supervised learning, bootstrapping [63], attention-based co-segmentation [12], and adversarial segmentation [31]. Our approach outperforms these baselines in terms of IoU and accuracy. (g-h) We test label data sensitivity.

	IoU					Pixel Accuracy				
Time distance (frames)	100	200	300	400	500	100	200	300	400	500
Attention-based [12]	0.51	0.43	0.47	0.57	0.45	0.83	0.82	0.82	0.80	0.78
Adversarial network [31]	**0.61**	0.54	0.62	**0.60**	0.41	0.84	0.84	0.83	0.82	0.80
Ours	0.60	**0.60**	**0.63**	0.56	**0.58**	**0.93**	**0.93**	**0.94**	**0.91**	**0.90**

Table 2: Mean IoU and pixel accuracy result of different time distance

stance, the appearance on monkey changes significantly as moving, i.e., the unlabeled data closer to the labeled data in time are more likely to look similar. We sample unlabeled data at every 100 frame (3 seconds) to compare the performance on the unlabeled data. Figs. 7(g) and 7(h) show that both IoU and pixel accuracy decrease as the difference between two time instances increases. However, our cross supervision have better ,performance than the semi-supervised learning baselines [12, 31] in both metrics, and the performance degradation is much milder than two methods. The numerical results can be found in Table 2.

Qualitative Result The qualitative result is shown in Fig. 1. Our cross-view supervision via shape transfer can handle wide baseline multiview images and correct the segmentation errors in the baselines by leveraging multiview images jointly. This becomes more evident on the boundaries or protruding body parts, e.g., monkey's paws and tails, human's legs and hands. See Appendix for more result.

6. Summary

We present a semi-supervised framework to train an object co-segmentation network by leveraging multi-view images. The key novelty is a method of shape belief transfer—using segmentation belief in one image to predict that of the other image through epipolar geometry analogous to shape-from-silhouette. The shape belief transfer provides the upper and lower bounds of the segmentation for the unlabeled data. We introduce a triplet network which embeds computing of transferred shape. We also use multi-view images to bootstrap the unlabeled data for training data augmentation.

7. Acknowledgment

This work is supported by NSF IIS 1846031 and NSF IIS 1755895.

References

[1] http://rieff.ieor.berkeley.edu/rashomon/.

[2] http://www.crowdsyncapp.com/.

[3] I. Arev, H. S. Park, Y. Sheikh, J. K. Hodgins, and A. Shamir. Automatic editing of footage from multiple social cameras. *SIGGRAPH*, 2014.

[4] V. Badrinarayanan, A. Handa, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling. *arXiv:1505.07293*, 2015.

[5] T. D. Basha, Y. Moses, and S. Avidan. Photo sequencing. In *ECCV*, 2012.

[6] A. C. Berg, T. L. Berg, and J. Malik. Shape matching and object recognition using low distortion correspondences. In *CVPR*, 2005.

[7] G. Bertasius, S. X. Yu, H. S. Park, and J. Shi. Exploiting visual-spatial first-person co-occurrence for action-object detection without labels. In *ICCV*, 2017.

[8] A. Byravan and D. Fox. SE3-nets: Learning rigid body motion using deep neural networks. In *ICRA*, 2016.

[9] T. S. Caetano, J. J. McAuley, L. Cheng, Q. V. Le, and A. J. Smola. Learning graph matching. *TPAMI*, 2009.

[10] N. Campbell, G. Vogiatzis, C. Hernandez, and R. Cipolla. Automatic 3d object segmentation in multiple views using volumetric graph-cuts. In *Image and Vision Computing*, 2010.

[11] N. Campbell, G. Vogiatzis, C. Hernandez, and R. Cipolla. Automatic object segmentation from calibrated images. In *Conference for Visual Media Production*, 2011.

[12] H. Chen, Y. Huang, and H. Nakayama. Semantic aware attention based deep object co-segmentation. In *ACCV*, 2018.

[13] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *TPAMI*, 2018.

[14] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018.

[15] Q. Chen and V. Koltun. Photographic image synthesis with cascaded refinement networks. In *ICCV*, 2017.

[16] M. Cho, K. Alahari, and J. Ponce. Learning graphs to match. In *ICCV*, 2013.

[17] C. B. Choy, D. Xu, J. Gwak, K. Chen, and S. Savarese. 3D-r2n2: A unified approach for single and multi-view 3D object reconstruction. In *ECCV*, 2016.

[18] J. Dai, K. He, and J. Sun. Boxsup: Exploiting bounding boxes to supervise convolutionalnetworks for semantic segmentation. In *ICCV*, 2015.

[19] A. Djelouah, J. Franco, E. Boyer, P. Prez, and G. Drettakis. Cotemporal multi-view video segmentation. In *3DV*, 2016.

[20] A. Djelouah, J.-S. Franco, and E. Boyer. Multi-view object segmentation in space and time. In *ICCV*, 2013.

[21] A. Djelouah, J.-S. Franco, E. Boyer, F. Le Clerc, and P. Pérez. Sparse Multi-View Consistency for Object Segmentation. *TPAMI*, 2015.

[22] A. Fathi, J. K. Hodgins, and J. M. Rehg. Social interaction: A first-person perspective. In *CVPR*, 2012.

[23] M. A. Fischler and R. C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *ACM Comm.*, 1981.

[24] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *CVPR*, 2012.

[25] A. Godbehere, A. Matsukawa, and K. Goldberg. Visual tracking of human visitors under variable-lighting conditions for a responsive audio art installation. In *American Control Conference*, 2012.

[26] L. Guan, S. Sinha, J.-S. Franco, and M. Pollefeys. Visual hull construction in the presence of partial occlusion. In *3D Data Processing, Visualization, and Transmission*, 2006.

[27] E. T. Hall. A system for the notation of proxemic behaviour. *American Anthropologist*, 1963.

[28] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, second edition, 2004.

[29] K. He, G. Gkioxari, P. Dollar, and R. Girshick. Mask R-CNN. In *ICCV*, 2017.

[30] S. Hong, H. Noh, and B. Han. Decoupled deep neural networkfor semi-supervised semantic segmentation. In *NIPS*, 2015.

[31] W.-C. Hung, Y.-H. Tsai, Y.-T. Liou, Y.-Y. Lin, and M.-H. Yang. Adversarial learning for semi-supervised semantic segmentation. In *BMVC*, 2018.

[32] S. Jeong, J. Lee, B. Kim, Y. Kim, and J. Noh. Object segmentation ensuring consistency across multi-viewpoint images. In *CVPR*, 2018.

[33] A. Joulin, F. Bach, and J. Ponce. Discriminative clustering for image co-segmentation. In *CVPR*, 2010.

[34] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik. End-to-end recovery of human shape and pose. In *CVPR*, 2018.

[35] A. Kanazawa, S. Tulsiani, A. A. Efros, and J. Malik. Learning category-specific mesh reconstruction from image collections. In *ECCV*. 2018.

[36] G. Kim and E. P. Xing. On multiple foreground cosegmentation. In *CVPR*, 2012.

[37] A. Kowdle, S. N. Sinha, , and R. Szeliski. Multiple view object cosegmentation using appearance and stereo cues. In *ECCV*, 2012.

[38] S. Kullback and R. A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 1951.

[39] K. N. Kutulakos and S. M. Seitz. A theory of shape by space carving. *IJCV*, 2000.

[40] L. Ladicky, J. Shi, and M. Pollefeys. Pulling things out of perspective. In *CVPR*, 2014.

[41] A. Laurentini. The visual hull concept for silhouette-based image understanding. *TPAMI*, 1994.

[42] A. Laurentini. How many 2D silhouettes does it take to reconstruct a 3D object? *CVIU*, 1997.

[43] J. Lee, W. Woo, and E. Boyer. Large displacement optical flow: descriptor matching in variational motion estimation. *TPAMI*, 2011.

[44] M. Leordeanu, M. Hebert, and R. Sukthankar. Beyond local appearance: Category recognition from pairwise interactions of simple features. In *CVPR*, 2007.

[45] W. Li, O. H. Jafari, and C. Rother. Deep object co-segmentation. In *arXiv:1804.06423*, 2018.

[46] J. J. Lim, A. Khosla, and A. Torralba. FPM: Fine pose parts-based model with 3D cad models. In *ECCV*, 2014.

[47] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollàr, and C. L. Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014.

[48] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. *TPAMI*, 2017.

[49] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. SMPL: A skinned multi-person linear model. *SIGGRAPH Asia*, 2015.

[50] W. Matusik, C. Buehler, R. Raskar, S. J. Gortler, and L. McMillan. Image-based visual hulls. *SIGGRAPH*, 2000.

[51] A. Mustafa and A. Hilton. Semantically coherent co-segmentation and reconstruction of dynamic scenes. In *CVPR*, 2017.

[52] G. Papandreou, L.-C. Chen, K. Murphy, , and A. L. Yuille.l. Weakly-and semi-supervisedlearning of a dcnn for semantic image segmentation. In *ICCV*, 2015.

[53] H. S. Park, E. Jain, and Y. Shiekh. 3D social saliency from head-mounted cameras. In *NIPS*, 2012.

[54] D. Pathak, E. Shelhamer, J. Long, and T. Darrel. Fully convolutional multi-classmultiple instance learning. In *ICLR*, 2015.

[55] P. O. Pinheiro and R. Collobert. Weakly supervised semantic segmentation with convolutionalnetworks. In *ICCV*, 2015.

[56] P. Pritchett and A. Zisserman. Wide baseline stereo matching. In *ICCV*, 1998.

[57] R. Quan, J. Han, D. Zhang, and F. Nie. Object co-segmentation via graph optimized-flexible manifold ranking. In *CVPR*, 2016.

[58] S. R. Richter, V. Vineet, S. Roth, and V. Koltun. Playing for data: Ground truth from computer games. In *ECCV*, 2016.

[59] J. Rubio, J. Serrat, A. Lopez, and N. Paragios. Unsupervised co-segmentation through region matching. In *CVPR*, 2012.

[60] J. Schönberger and J. Frahm. Structure-from-motion revisited. In *CVPR*, 2016.

[61] J. Schönberger, E. Zheng, M. Pollefeys, and J. Frahm. Pixelwise view selection for unstructured multi-view stereo. In *ECCV*, 2016.

[62] A. Shaji, A. Varol, L. Torresani, and P. Fua. Simultaneous point matching and 3D deformable surface reconstruction. In *CVPR*, 2010.

[63] T. Simon, H. Joo, I. Matthews, and Y. Sheikh. Hand keypoint detection in single images using multiview bootstrapping. In *CVPR*, 2017.

[64] M. Sormann, C. Zach, and K. Karner. Graph cut based multiple view segmentation for 3d reconstruction. In *3DPVT*, 2006.

[65] H. Su, C. Qi, K. Mo, and L. Guibas. PointNet: Deep learning on point sets for 3D classification and segmentation. In *CVPR*, 2017.

[66] L. Torresani, V. Kolmogorov, and C. Rother. Feature correspondence via graph matching: Models and global optimization. In *ECCV*, 2008.

[67] S. Tulsiani, S. Gupta, D. Fouhey, A. A. Efros, and J. Malik. Factoring shape, pose, and layout from the 2D image of a 3D scene. In *CVPR*, 2018.

[68] S. Tulsiani, T. Zhou, A. A. Efros, and J. Malik. Multi-view supervision for single-view reconstruction via differentiable ray consistency. In *CVPR*, 2017.

[69] S. Vicente, C. Rother, and V. Kolmogorov. Object cosegmentation. In *CVPR*, 2011.

[70] S. Vijayanarasimhan, S. Ricco, C. Schmid, R. Sukthankar, and K. Fragkiadaki. SfM-Net: Learning of structure and motion from video. In *arXiv:1704.07804*, 2017.

[71] Y. Xiang, R. Mottaghi, and S. Savarese. Beyond pascal: A benchmark for 3D object detection in the wild. In *WACV*, 2014.

[72] J. Yan and M. Pollefeys. Automatic kinematic chain building from feature trajectories of articulated objects. In *CVPR*, 2006.

[73] Y. Yao, Y. Jafarian, and H. S. Park. MONET: Multiview semi-supervised keypoint detection via epipolar divergence. In *ICCV*, 2019.

[74] R. Yonetani, K. M. Kitani, and Y. Sato. Cosegmentation of image pairs by histogram matching. In *CVPR*, 2006.

[75] Z. Yuan and Y. Lu, T.and Wu. Deep-dense conditional random fields for object co-segmentation. In *IJCAI*, 2018.

[76] Z. Z and F. Heijden. Efficient adaptive density estimation per image pixel for the task of background subtraction. *Pattern Recogn. Lett.*, 2006.

[77] F. Zhou and F. D. la Torre. Deformable graph matching. In *CVPR*, 2013.

[78] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe. Unsupervised learning of depth and ego-motion from video. In *CVPR*, 2017.

[79] T. Zhou, P. Krähenbühl, M. Aubry, Q. Huang, and A. A. Efros. Learning dense correspondence via 3D-guided cycle consistency. In *CVPR*, 2016.

[80] T. Zhou, Y. J. Lee, S. X. Yu, and A. A. Efros. Flowweb: Joint image set alignment by weaving consistent, pixel-wise correspondences. In *ICCV*, 2015.

[81] S. Zuffi, A. Kanazawa, D. Jacobs, and M. J. Black. 3d menagerie: Modeling the 3d shape and pose of animals. In *CVPR*, 2017.