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

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

IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), 2020

June 14, 2020

Outline

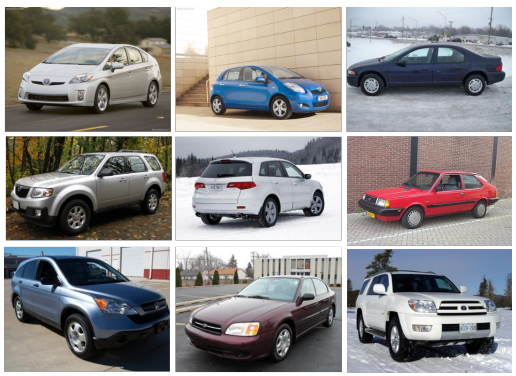
- Introduction
- Related work
- Proposed method
- Experimental results
- Conclusions

Outline

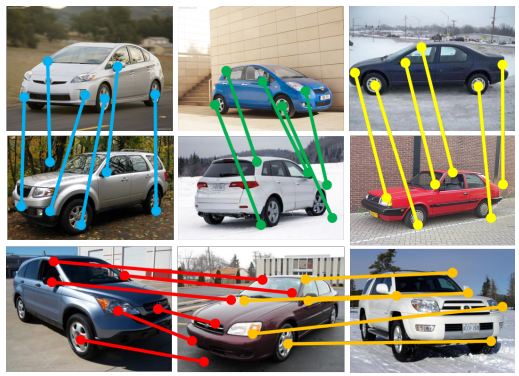
- Introduction
- Related work
- Proposed method
- Experimental results
- Conclusions

Joint semantic matching and object co-segmentation

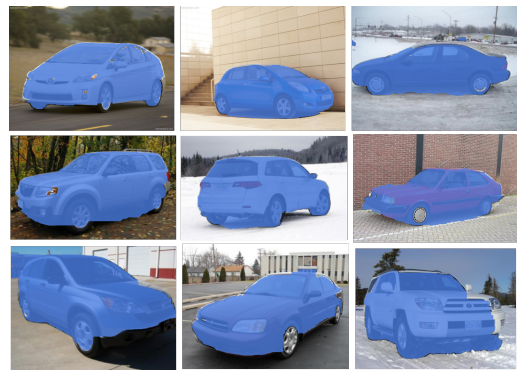
- Input: a collection of images containing objects of a specific category.
- Goal: establish correspondences between object instances and segment them out.
- Setting: weakly supervised (no ground-truth keypoint correspondences and object masks are used for training).



A collection of images



Semantic matching

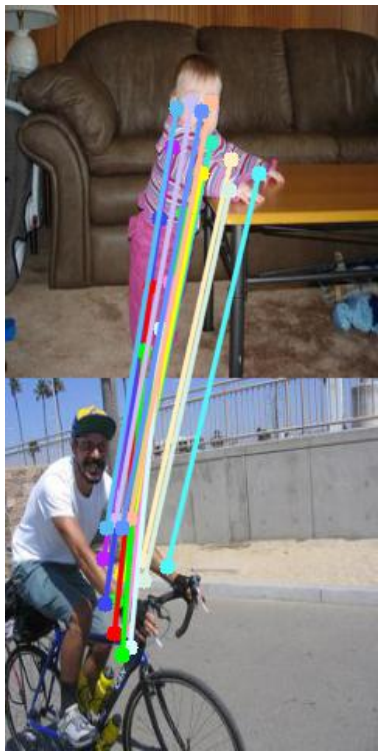


Object co-segmentation

Issues with semantic matching and object co-segmentation

- Semantic matching: suffer from background clutters.
- Object co-segmentation: segment only the most discriminative regions.

Input



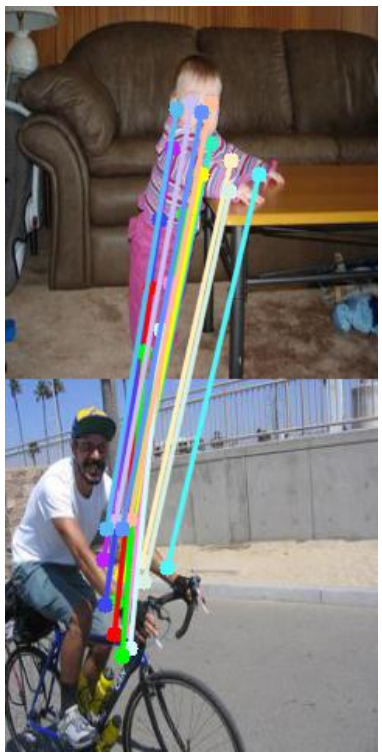
Semantic matching

Input

Co-segmentation

Motivation of joint learning

- Semantic matching: dense correspondence fields provide supervision by enforcing consistency between the predicted object masks.
- Object co-segmentation: object masks allow the model to focus on matching the foreground regions.



Separate learning

Joint learning (Ours)

Separate learning

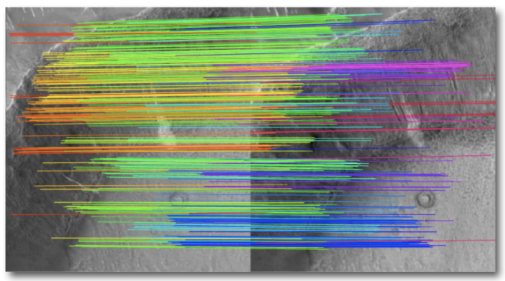
Joint learning (Ours)

Outline

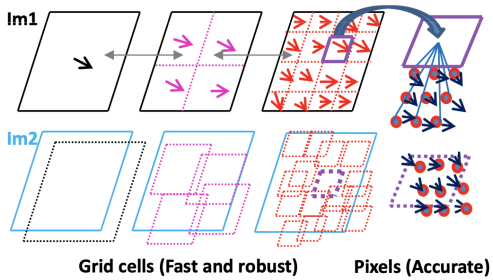
- Introduction
- Related work
- Proposed method
- Experimental results
- Conclusions

Semantic matching - early methods

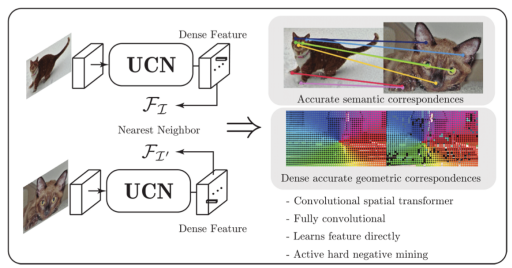
- Hand-crafted descriptor based methods: leverage SIFT or HOG features along with geometric matching models to solve correspondence matching by energy minimization.
- Trainable descriptor based approaches: adopt trainable CNN features for semantic matching.
- Limitation: require manual correspondence annotations for training.



SIFT Flow [1]



DSP [2]



UCN [3]

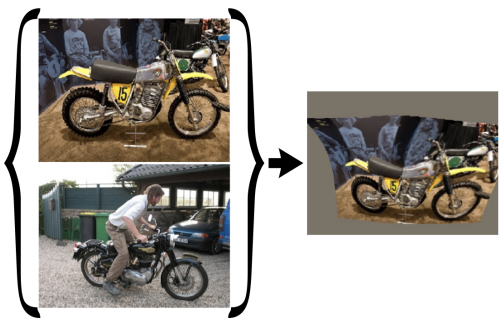
[1] Liu et al. SIFT Flow: Dense Correspondence across Scenes and its Applications. TPAMI'11.

[2] Kim et al. Deformable Spatial Pyramid Matching for Fast Dense Correspondences. CVPR'13.

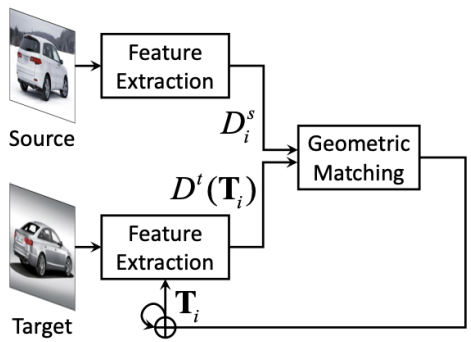
[3] Choy et al. Universal Correspondence Network. NeurIPS'16.

Semantic matching - recent approaches

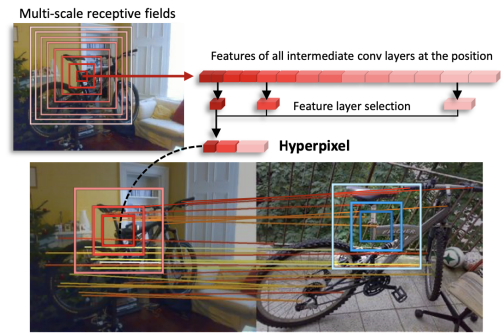
- Estimate geometric transformations (affine or TPS) using CNN or RNN for semantic alignment.
- Adopt multi-scale features for establishing semantic correspondences.
- Limitation: suffer from background clutters and inconsistent bidirectional matching.



CNNGeo [4]



RTNs [5]



HPF [6]

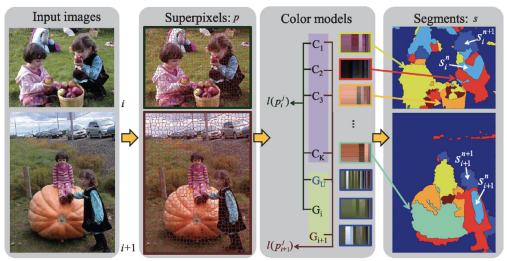
[4] Rocco et al. Convolutional neural network architecture for geometric matching. CVPR'17.

[5] Kim et al. Recurrent Transformer Networks for Semantic Correspondence. NeurIPS'18.

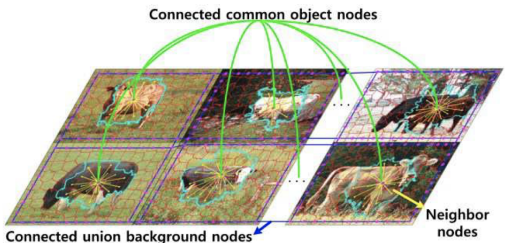
[6] Min et al. Hyperpixel Flow: Semantic Correspondence with Multi-layer Neural Features. ICCV'19.

Object co-segmentation - early methods

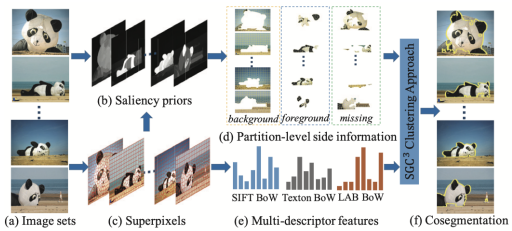
- Graph based methods: construct a graph to encode the relationships between object instances.
- Clustering based approaches: assume that common objects share similar appearances and achieve co-segmentation by finding tight clusters.
- Limitation: lack of an end-to-end trainable pipeline.



MFC [7]



GO-FMR [8]



SGC^3 [9]

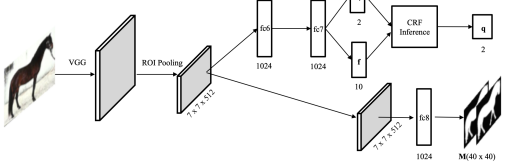
[7] Chang et al. Optimizing the decomposition for multiple foreground cosegmentation. CVIU'15.

[8] Quan et al. Object Co-segmentation via Graph Optimized-Flexible Manifold Ranking. CVPR'16.

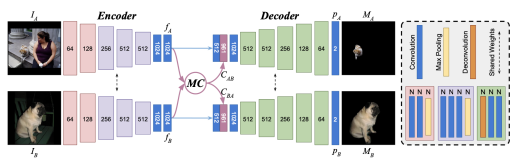
[9] Tao et al. Image Cosegmentation via Saliency-Guided Constrained Clustering with Cosine Similarity. AAAI'17.

Object co-segmentation - recent approaches

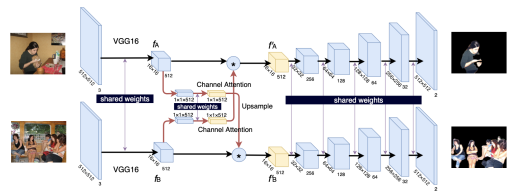
- Leverage CNN models with CRF or attention mechanisms to achieve object co-segmentation.
- Limitation: require foreground masks for training and not applicable to unseen object categories.



DDCRF [10]



DOCS [11]



CA [12]

[10] Yuan et al. Deep-dense Conditional Random Fields for Object Co-segmentation. IJCAI'17.

[11] Li et al. Deep object co-segmentation. ACCV'18.

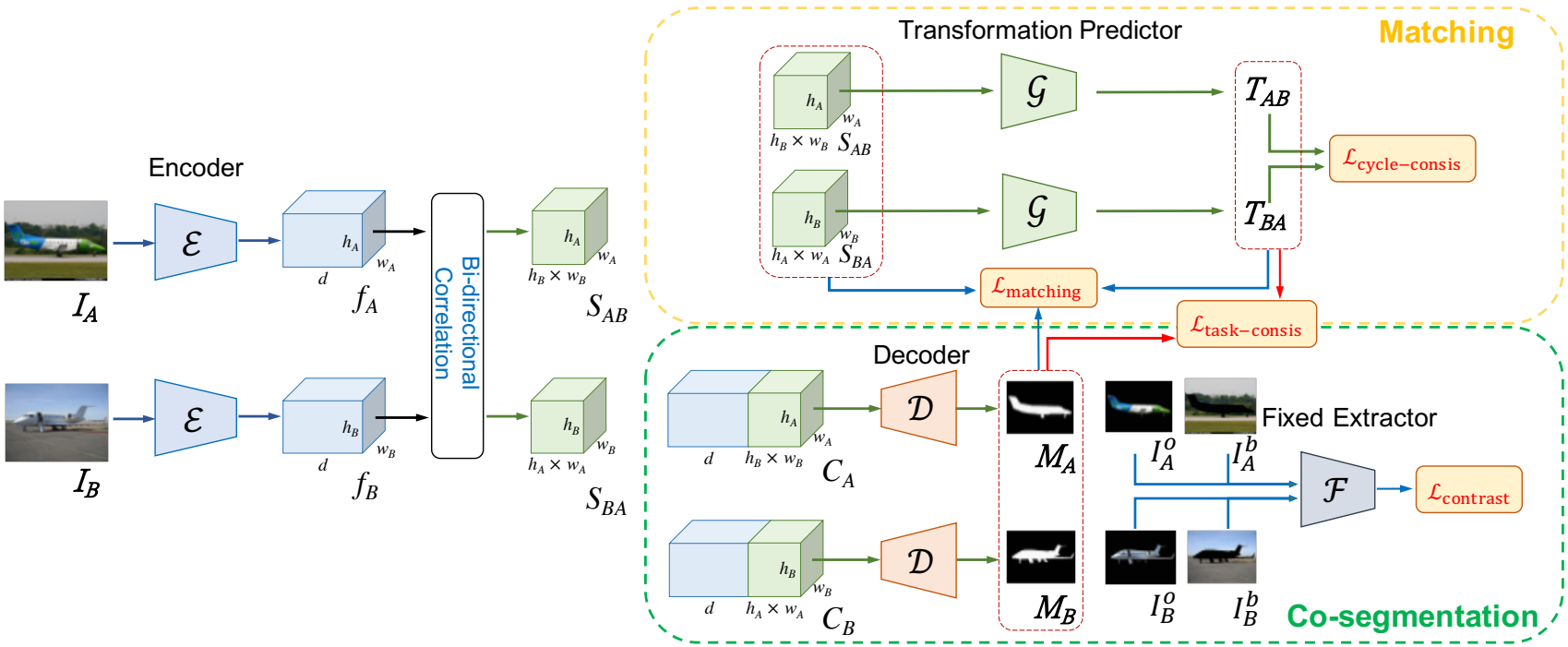
[12] Chen et al. Semantic Aware Attention Based Deep Object Co-segmentation. ACCV'18.

Outline

- Introduction
- Related work
- Proposed method
- Experimental results
- Conclusions

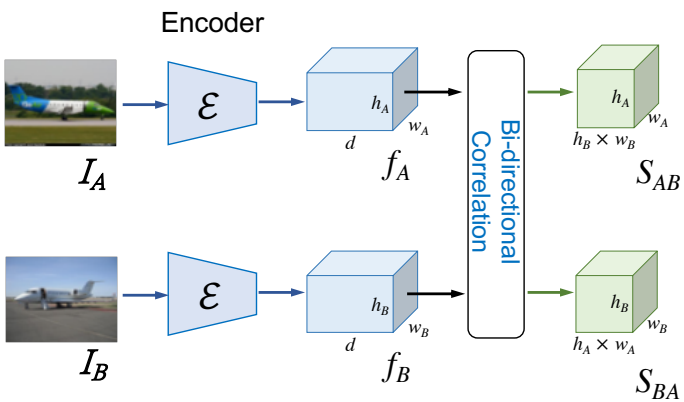
Overview of the MaCoSNet

- A two-stream network:
  - (*top*) semantic matching network.
  - (*bottom*) object co-segmentation network.
- Input: an image pair containing objects of a specific category.
- Goal: establish correspondences between object instances and segment them out.
- Supervision: image-level supervision (i.e., weakly supervised).



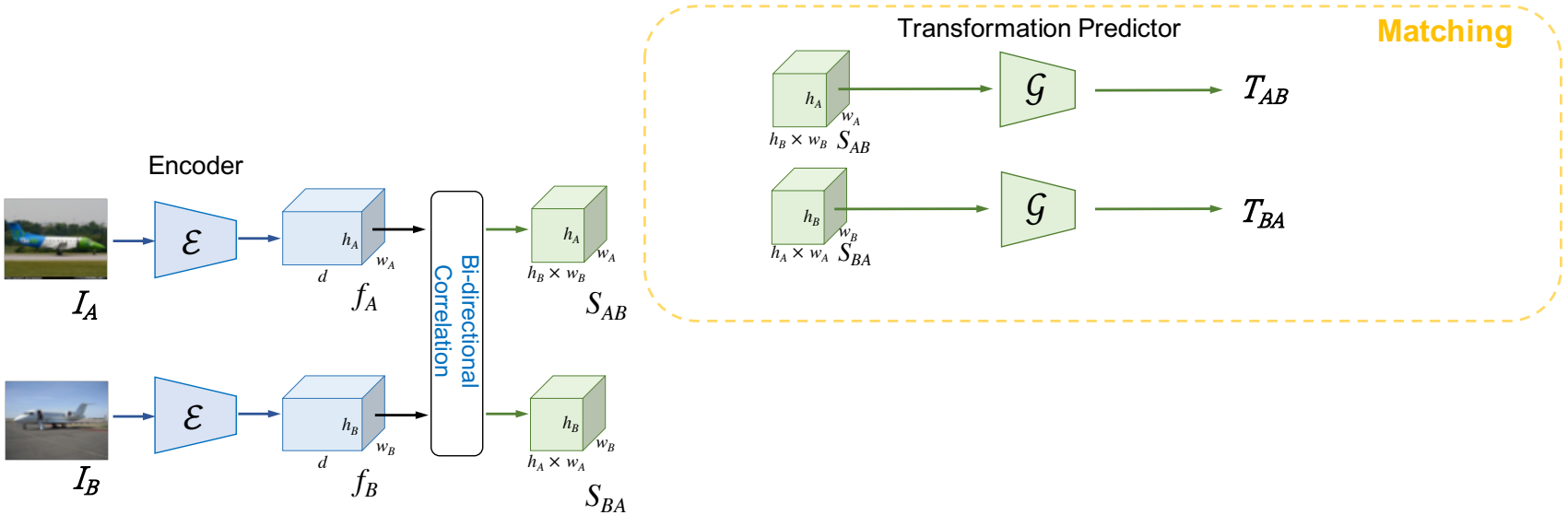
Shared feature encoder

- Given an input image pair, we first use the feature encoder $\mathcal{E}$ 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.



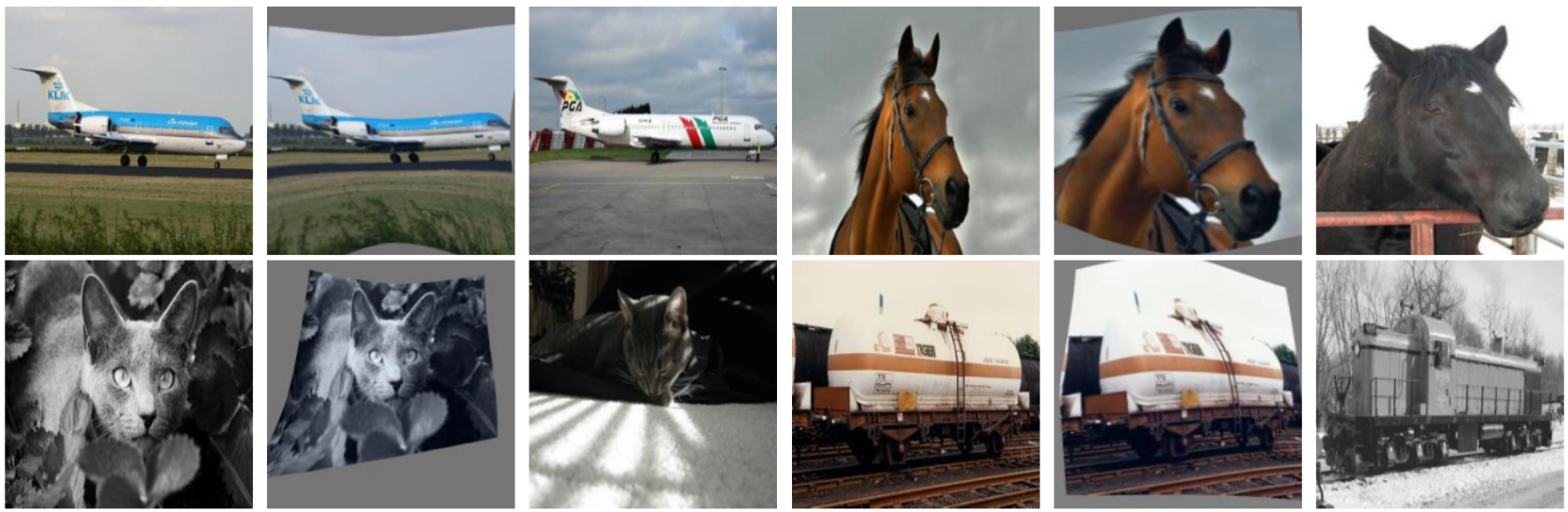
Overview of the semantic matching network

- Our semantic matching network is composed of a transformation predictor $\mathcal{G}$.
- The transformation predictor $\mathcal{G}$ takes the correlation maps as inputs and estimates the geometric transformations that align the two images.



Geometric transformation

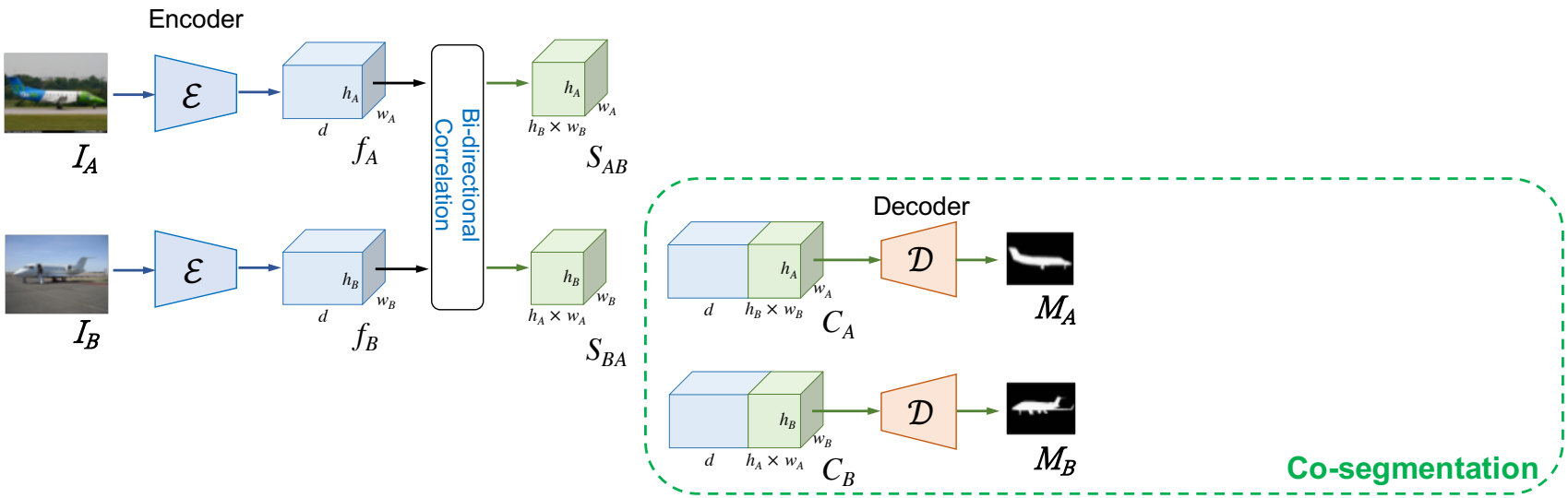
- Our transformation predictor $\mathcal{G}$ is a cascade of two modules predicting an affine transformation and a thin plate spline (TPS) transformation, respectively [4].
- The estimated geometric transformation allows our model to warp a source image so that the warped source image aligns well with the target image.



[4] Rocco et al. Convolutional neural network architecture for geometric matching. CVPR'17.

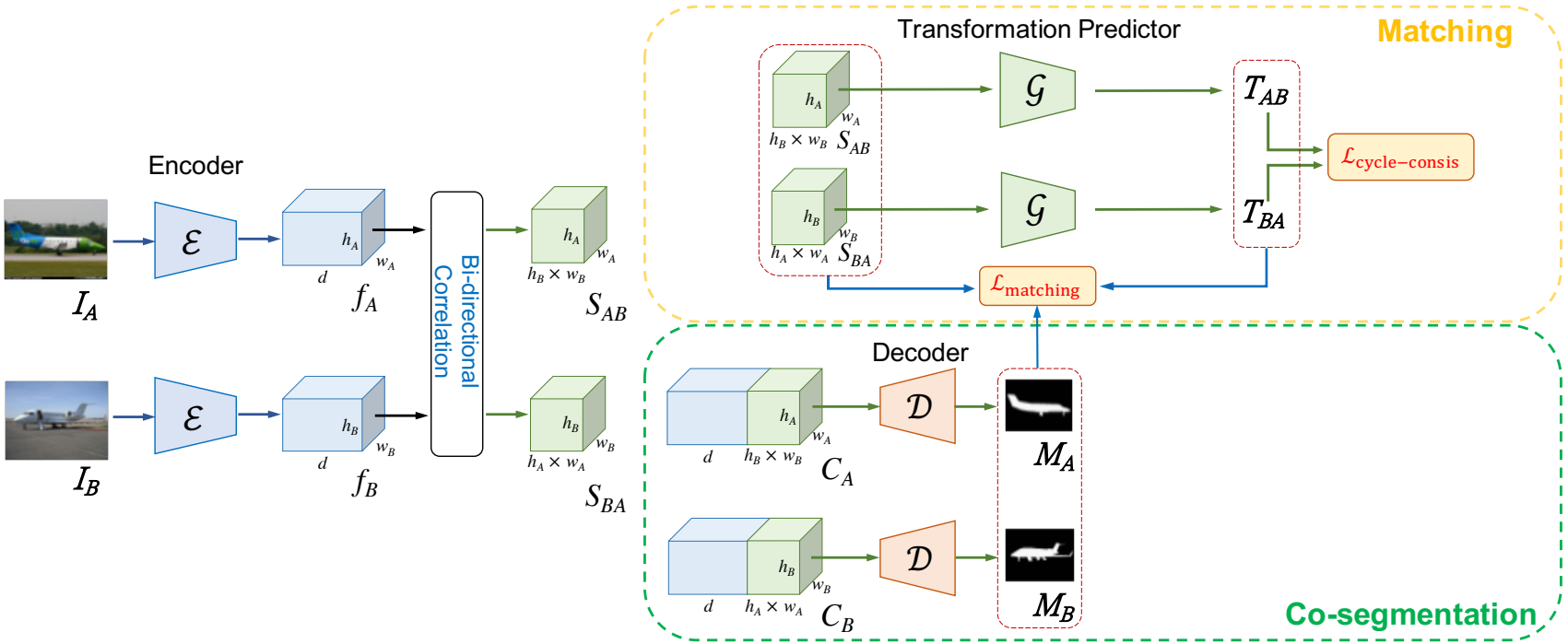
Overview of the object co-segmentation network

- We use the fully convolutional network decoder $\mathcal{D}$ for generating object masks.
- To capture the co-occurrence information, we concatenate the encoded image features with the correlation maps.
- The decoder $\mathcal{D}$ then takes the concatenated features as inputs to generate object segmentation masks.



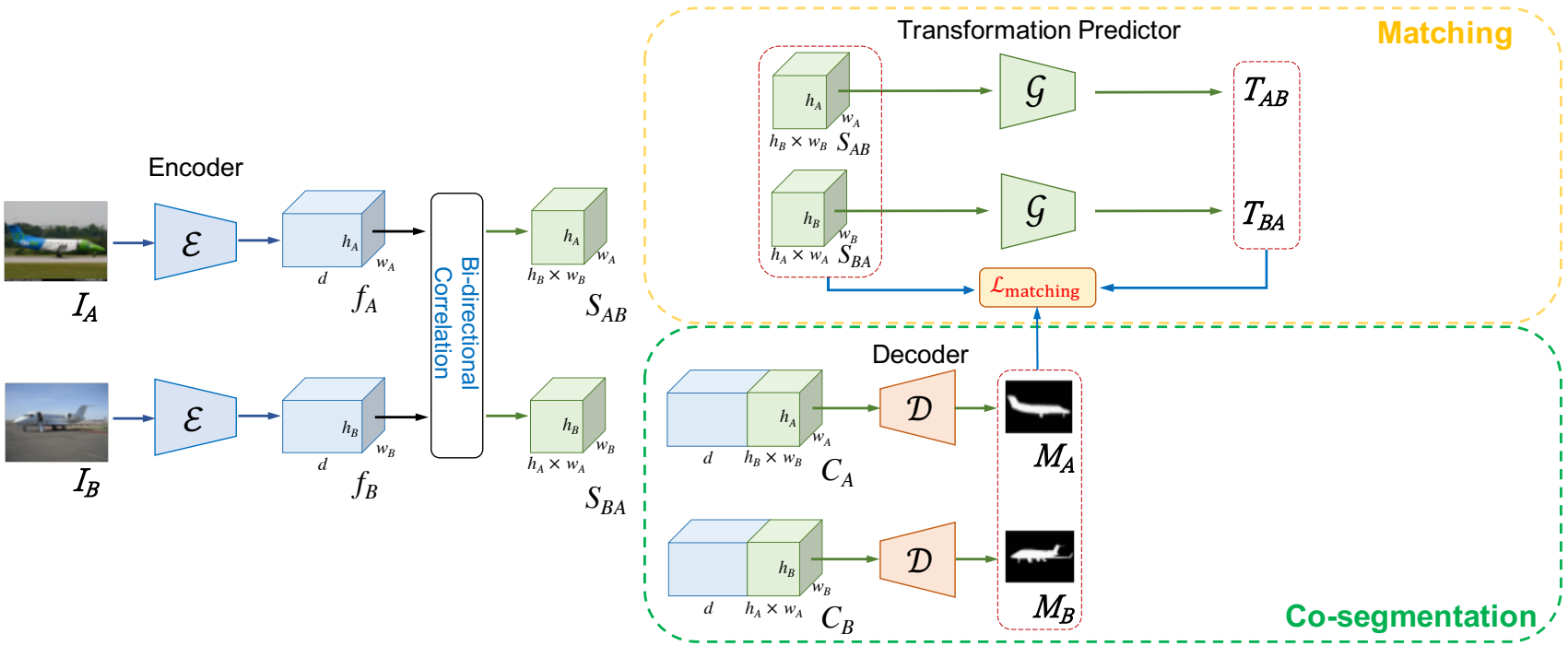
Training the semantic matching network

- There are two losses to train the semantic matching network:
  - foreground-guided matching loss $\mathcal{L}_{\text{matching}}$.
  - forward-backward consistency loss $\mathcal{L}_{\text{cycle-consis}}$.



Foreground-guided matching loss $\mathcal{L}_{\text{matching}}$

- Minimize the distance between corresponding features based on the estimated geometric transformation.
- Leverage the predicted object masks to suppress the negative impacts caused by background clutters.



Foreground-guided matching loss $\mathcal{L}_{\text{matching}}$

- Given the estimated geometric transformation T_{AB}, we can identify and remove geometrically inconsistent correspondences.
- Consider a correspondence with the endpoints $(\mathbf{p} \in \mathcal{P}_A, \mathbf{q} \in \mathcal{P}_B)$, where $\mathcal{P}_A$ and $\mathcal{P}_B$ are the domains of all spatial coordinates of f_A and f_B, respectively.
- We 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}.

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

- A correspondence $(\mathbf{p}, \mathbf{q})$ is considered geometrically consistent with transformation T_{AB} if its projection error $\| T_{AB}(\mathbf{p}) - \mathbf{q} \|$ is not larger than the threshold φ.

Foreground-guided matching loss $\mathcal{L}_{\text{matching}}$

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

$$s_A(\mathbf{p}) = \sum_{\mathbf{q} \in \mathcal{P}_B} m_A(\mathbf{p}, \mathbf{q}) \cdot S_{AB}(\mathbf{p}, \mathbf{q}). \tag{2}$$

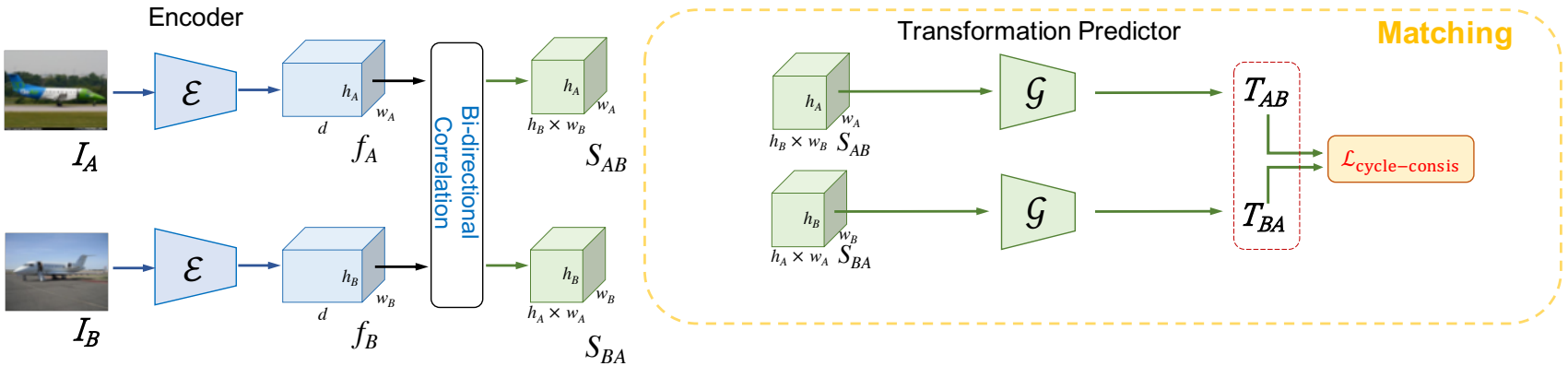
- To suppress the effect of background clutters, we leverage the object masks M_A and M_B estimated by the decoder $\mathcal{D}$ to focus on matching the foreground regions.
- The foreground-guided matching loss $\mathcal{L}_{\text{matching}}$ is defined as

$$\mathcal{L}_{\text{matching}} = -\left(\sum_{\mathbf{p} \in \mathcal{P}_A} s_A(\mathbf{p}) \cdot M_A(\mathbf{p}) + \sum_{\mathbf{q} \in \mathcal{P}_B} s_B(\mathbf{q}) \cdot M_B(\mathbf{q}) \right). \tag{3}$$

- The negative sign indicates that maximizing the matching score is equivalent to minimizing the foreground-guided matching loss.

Forward-backward consistency loss $\mathcal{L}_{\text{cycle-consis}}$

- Regularize the network training by enforcing the predicted geometric transformations to be consistent between an image pair.
- Enforce the property $T_{BA}(T_{AB}(\mathbf{p})) \approx \mathbf{p}$ for any coordinate $\mathbf{p} \in \mathcal{P}_A$.



$$\mathcal{L}_{\text{cycle-consis}} = \frac{1}{\|\mathcal{P}_A\|} \sum_{\mathbf{p} \in \mathcal{P}_A} \| T_{BA}(T_{AB}(\mathbf{p})) - \mathbf{p} \| + \frac{1}{\|\mathcal{P}_B\|} \sum_{\mathbf{q} \in \mathcal{P}_B} \| T_{AB}(T_{BA}(\mathbf{q})) - \mathbf{q} \|. \tag{4}$$

Transitivity consistency loss $\mathcal{L}_{\text{trans}-\text{consis}}$

- The idea of forward-backward consistency between an image pair can be extended to the transitivity consistency across multiple images, e.g., three images.
- Given three images I_A, I_B, and I_C, we first estimate three geometric transformations T_{AB}, T_{BC}, and T_{CA}.
- We then enforce the property $T_{CA}(T_{BC}(T_{AB}(\mathbf{p}))) \approx \mathbf{p}$ for any coordinate $\mathbf{p} \in \mathcal{P}_A$.

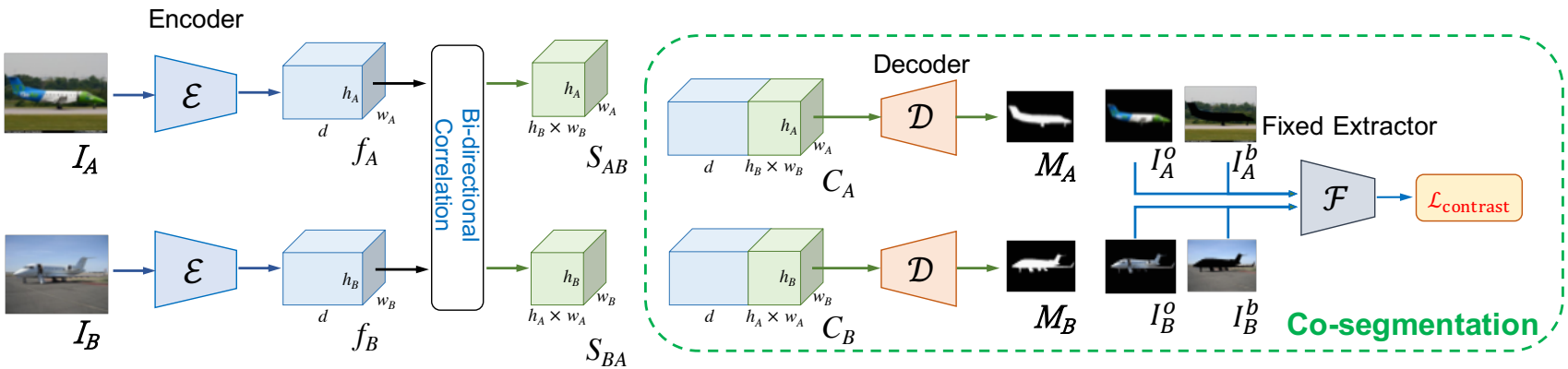
$$\mathcal{L}_{\text{trans}-\text{consis}} = \frac{1}{\|\mathcal{P}_A\|} \sum_{\mathbf{p} \in \mathcal{P}_A} \| T_{CA}(T_{BC}(T_{AB}(\mathbf{p}))) - \mathbf{p} \|. \tag{5}$$

Details of the consistency losses

- For the transitivity consistency loss $\mathcal{L}_{\text{trans}-\text{consis}}$, 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 $\mathcal{L}_{\text{cycle}-\text{consis}}$ and the transitivity consistency loss $\mathcal{L}_{\text{trans}-\text{consis}}$.

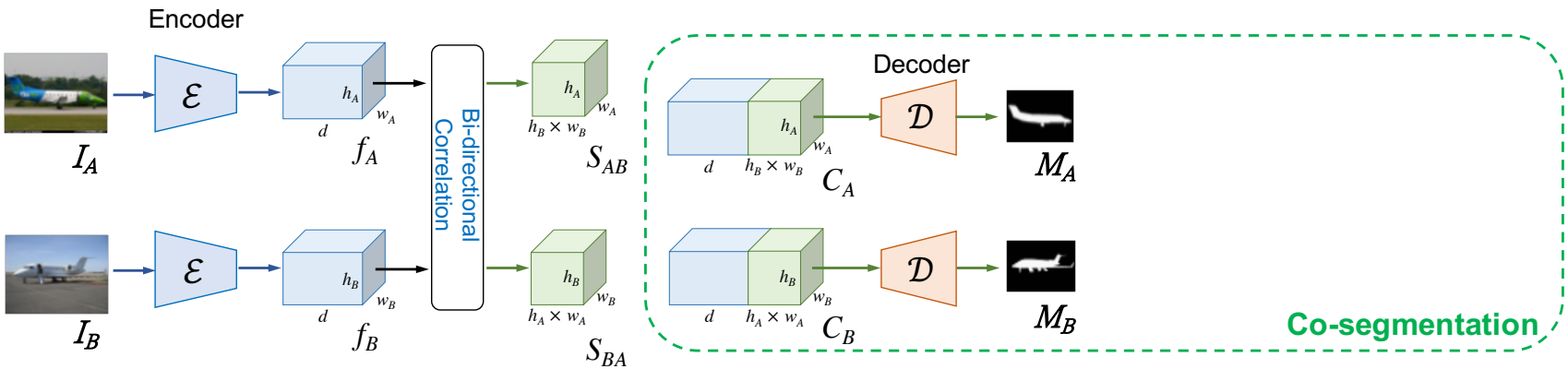
Training the object co-segmentation network

- There is one loss to train the object co-segmentation network:
  - perceptual contrastive loss $\mathcal{L}_{\text{contrast}}$.



Perceptual contrastive loss $\mathcal{L}_{\text{contrast}}$

- Given the feature maps f_A and f_B and the correlation maps S_{AB} and S_{BA}, we first generate the concatenated features $C_A = [f_A, S_{AB}]$ and $C_B = [f_B, S_{BA}]$.
- The decoder $\mathcal{D}$ then takes the concatenated feature maps C_A and C_B as inputs and produces object masks M_A and M_B for input images I_A and I_B, respectively.



Perceptual contrastive loss $\mathcal{L}_{\text{contrast}}$

- To facilitate the decoder $\mathcal{D}$ segmenting the co-occurrent objects, we exploit two properties:
  - high foreground object similarity *across* images.
  - high foreground-background discrepancy *within* each image.
- We first generate the object image I_i^o and the background image I_i^b for each image I_i by

$$I_i^o = M_i \otimes I_i \quad \text{and} \quad I_i^b = (1 - M_i) \otimes I_i \text{ for } i \in \{A, B\}, \tag{6}$$

  where $\otimes$ denotes the pixel-wise multiplication between the two operands.
- We apply an ImageNet-pretrained ResNet-50 network $\mathcal{F}$ to I_i^o and I_i^b to extract their semantic feature vectors $\mathcal{F}(I_i^o)$ and $\mathcal{F}(I_i^b)$, respectively.

Perceptual contrastive loss $\mathcal{L}_{\text{contrast}}$

- The perceptual contrastive loss $\mathcal{L}_{\text{contrast}}$ is defined as

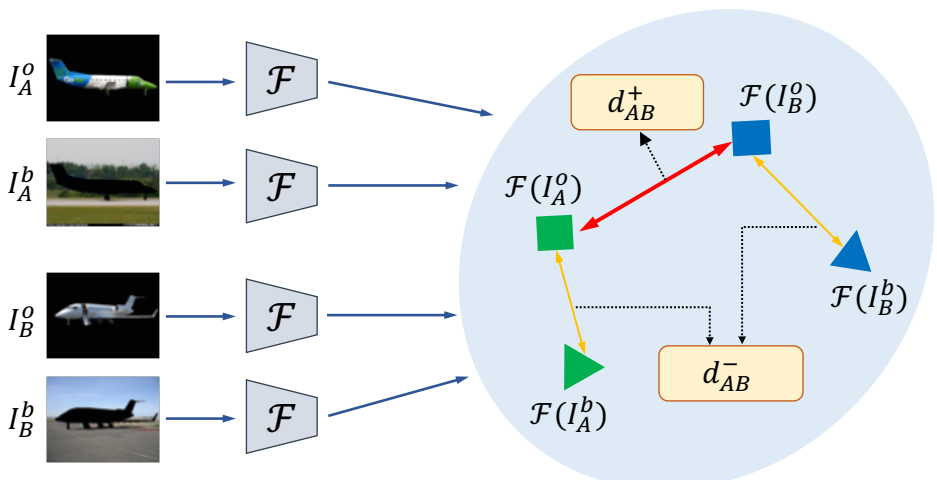
$$\mathcal{L}_{\text{contrast}} = d_{AB}^{+} + d_{AB}^{-}, \tag{7}$$

where the two criteria are respectively imposed on d_{AB}^{+} and d_{AB}^{-}:

$$d_{AB}^{+} = \frac{1}{c}\|\mathcal{F}(I_A^o) - \mathcal{F}(I_B^o)\|^2 \text{ and} \tag{8}$$

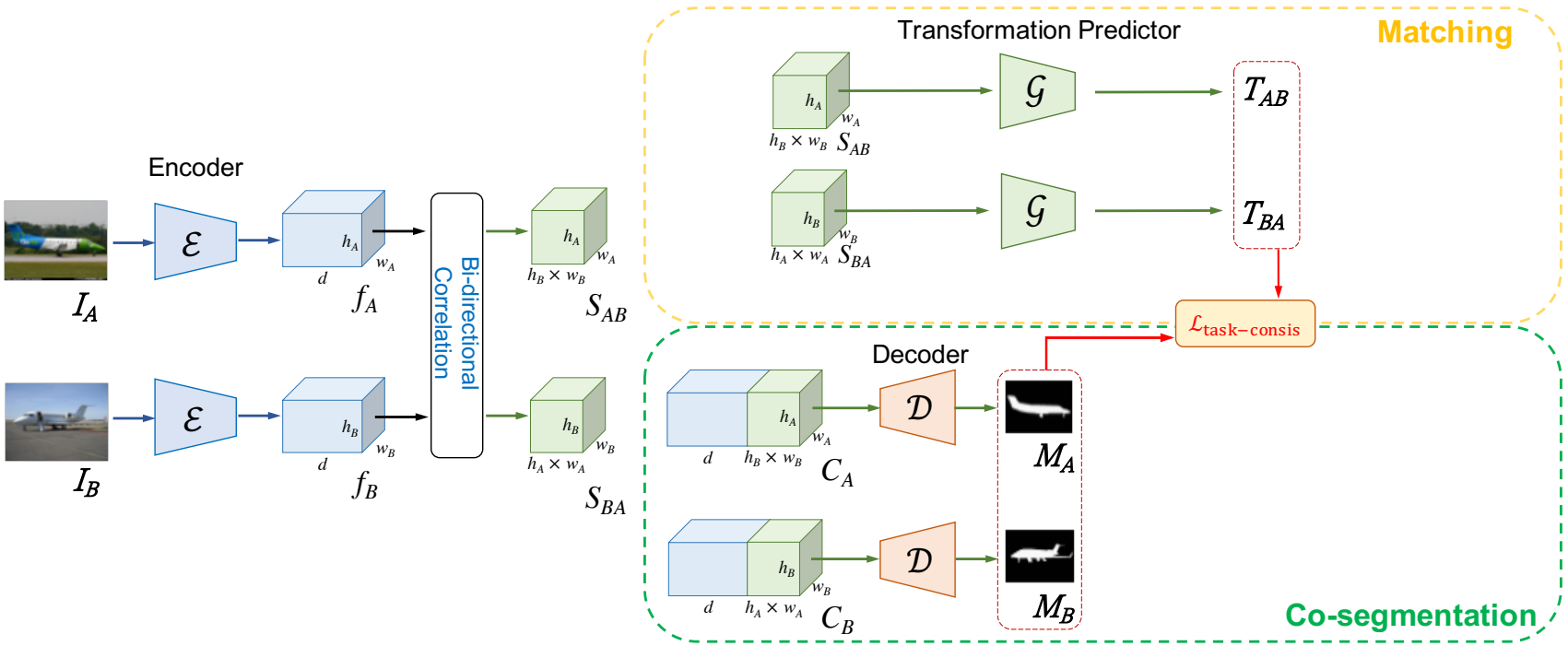
$$d_{AB}^{-} = \max\left(0, m - \frac{1}{2c}\left(\|\mathcal{F}(I_A^o) - \mathcal{F}(I_A^b)\|^2 + \|\mathcal{F}(I_B^o) - \mathcal{F}(I_B^b)\|^2\right)\right). \tag{9}$$

- The constant c is the dimension of the semantic features produced by $\mathcal{F}$, and the margin m is the cutoff threshold.



Cross-network training

- Using the perceptual contrastive loss $\mathcal{L}_{\text{contrast}}$ alone for object co-segmentation may generate object masks that highlight only the discriminative parts rather than the entire objects.
- We leverage the dense correspondence fields estimated from semantic matching to provide supervision for object co-segmentation.



Cross-network consistency loss $\mathcal{L}_{\text{task}-\text{consis}}$

- Propose a cross-network consistency loss $\mathcal{L}_{\text{task}-\text{consis}}$ that bridges the outputs of the semantic matching co-segmentation networks.
- Predicted object masks M_A and M_B should be geometrically consistent with the learned geometric transformations T_{AB} and T_{BA}: apply T_{AB} to M_A and obtain $\tilde{M}_A$ to match M_B
- The cross-network consistency loss $\mathcal{L}_{\text{task}-\text{consis}}$ is defined as

$$\mathcal{L}_{\text{task}-\text{consis}} = \mathcal{L}_{\text{bce}}(\tilde{M}_A, M_B) + \mathcal{L}_{\text{bce}}(\tilde{M}_B, M_A), \tag{10}$$

where $\mathcal{L}_{\text{bce}}(\tilde{M}_A, M_B)$ computes the binary cross-entropy loss between $\tilde{M}_A$ and M_B, and is defined as

$$\begin{aligned}\mathcal{L}_{\text{bce}}(\tilde{M}_A, M_B) = -\frac{1}{H_B \times W_B}\Bigg(&\sum_{i,j} \tilde{M}_A(i,j) \log\Big(M_B(i,j)\Big) \\ &+ \sum_{i,j}\Big(1 - \tilde{M}_A(i,j)\Big) \log\Big(1 - M_B(i,j)\Big)\Bigg).\end{aligned} \tag{11}$$

Full training loss $\mathcal{L}$

- The full training loss $\mathcal{L}$ is composed of five loss functions defined by

$$\begin{aligned}\mathcal{L} = \mathcal{L}_{\text{matching}} + \lambda_{\text{cycle}} \cdot \mathcal{L}_{\text{cycle-consis}} + \lambda_{\text{trans}} \cdot \mathcal{L}_{\text{trans-consis}} \\ + \lambda_{\text{contrast}} \cdot \mathcal{L}_{\text{contrast}} + \lambda_{\text{task}} \cdot \mathcal{L}_{\text{task-consis}},\end{aligned} \tag{12}$$

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

Outline

- Introduction
- Related work
- Proposed method
- Experimental results
- Conclusions

Evaluation metrics and datasets

- Evaluation metrics:
  - semantic matching:
    - ★ the percentage of correct keypoints (PCK).
  - object co-segmentation:
    - ★ the precision $\mathcal{P}$.
    - ★ the Jaccard index $\mathcal{J}$.
- Datasets:
  - joint semantic matching and object co-segmentation:
    - ★ TSS.
  - semantic matching:
    - ★ PF-PASCAL.
    - ★ PF-WILLOW.
    - ★ SPair-71k.
  - object co-segmentation:
    - ★ Internet.

Evaluation of joint matching and co-segmentation

Table: Experimental results of semantic matching on the TSS dataset.

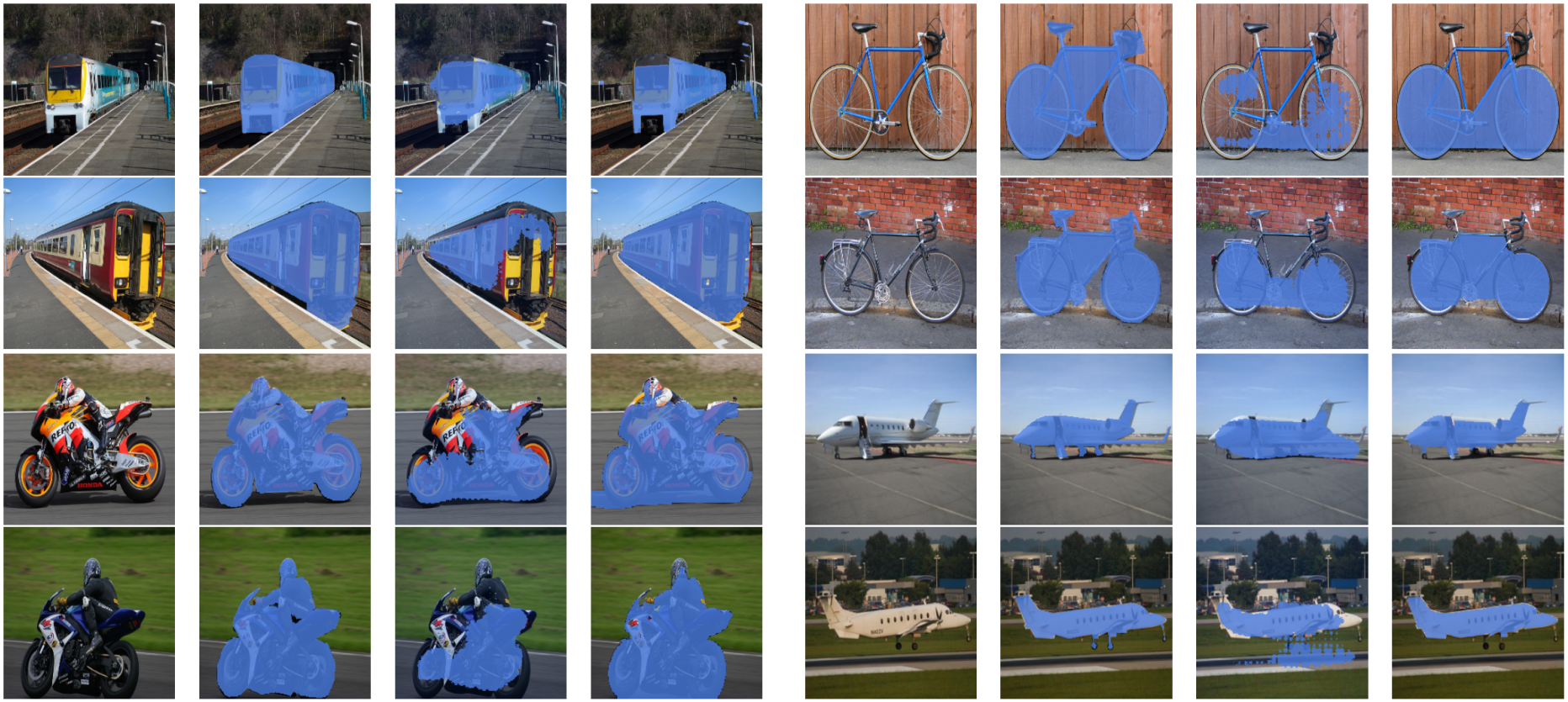
Method	Descriptor	Supervision	FG3DCar	JODS	PASCAL	Avg.
SIFT Flow	SIFT	-	0.632	0.509	0.360	0.500
DSP	SIFT	-	0.487	0.465	0.382	0.445
TSS	HOG	-	0.829	0.595	0.483	0.636
DAISY	DAISY	-	0.636	0.373	0.338	0.449
UCN	GoogLeNet	Strong	0.853	0.672	0.511	0.679
FCSS	FCSS	Strong	0.830	0.656	0.494	0.660
Proposal Flow	FCSS	Strong	0.839	0.635	0.582	0.685
DCTM	FCSS	Strong	0.891	0.721	0.610	0.740
SCNet-AG+	VGG-16	Strong	0.776	0.608	0.474	0.619
CNNGeo	ResNet-101	Strong	0.886	0.758	0.560	0.735
CNNGeo w/ Inlier	ResNet-101	Weak	0.892	0.758	0.562	0.737
Ours w/o co-seg	ResNet-101	Weak	<u>0.907</u>	<u>0.781</u>	0.565	0.751
Ours	ResNet-101	Weak	**0.908**	**0.783**	<u>0.615</u>	<u>0.769</u>

Evaluation of joint matching and co-segmentation

Table: Experimental results of object co-segmentation on the TSS dataset.

Method	Descriptor	FG3DCar		JODS		PASCAL		Avg.	
		$\mathcal{P}$	$\mathcal{J}$	$\mathcal{P}$	$\mathcal{J}$	$\mathcal{P}$	$\mathcal{J}$	$\mathcal{P}$	$\mathcal{J}$
SIFT Flow	SIFT	0.661	0.42	0.557	0.24	0.628	0.41	0.615	0.36
DSP	SIFT	0.502	0.29	0.454	0.22	0.496	0.34	0.484	0.28
Hati et al.	SIFT	0.785	0.47	0.778	0.31	0.701	0.31	0.755	0.36
Chang et al.	SIFT	0.872	0.67	0.851	0.52	0.723	0.40	0.815	0.53
Jerripothula et al.	SIFT	0.913	0.78	0.900	0.65	0.880	0.73	0.898	0.72
Faktor et al.	HOG	0.873	0.69	0.859	0.54	0.771	0.50	0.834	0.58
Joulin et al.	SIFT	0.651	0.46	0.626	0.32	0.587	0.40	0.621	0.39
MRW	SIFT	0.784	0.63	0.730	0.46	0.804	0.66	0.773	0.58
DFF	DAISY	0.704	0.33	0.696	0.21	0.601	0.21	0.667	0.25
TSS	HOG	0.877	0.76	0.761	0.50	0.778	0.65	0.805	0.63
Ours w/o matching	ResNet-101	0.958	0.88	0.911	0.71	0.829	0.61	0.899	0.73
Ours	ResNet-101	**0.963**	**0.90**	**0.940**	**0.77**	**0.939**	**0.86**	**0.947**	**0.84**

Visual results of joint learning vs. separate learning



Evaluation of co-segmentation on Internet

Table: Experimental results of object co-segmentation on the Internet dataset.

Method	Descriptor	Airplane		Car		Horse		Avg.	
		$\mathcal{P}$	$\mathcal{J}$	$\mathcal{P}$	$\mathcal{J}$	$\mathcal{P}$	$\mathcal{J}$	$\mathcal{P}$	$\mathcal{J}$
DOCS	VGG-16	0.946	0.64	0.940	0.83	0.914	0.65	0.933	0.70
Sun et al.	HOG	0.886	0.36	0.870	0.73	0.876	0.55	0.877	0.55
Joulin et al.	SIFT	0.475	0.12	0.592	0.35	0.642	0.30	0.570	0.24
Kim et al.	SIFT	0.802	0.08	0.689	0.0004	0.751	0.06	0.754	0.05
Rubinstein et al.	SIFT	0.880	0.56	0.854	0.64	0.828	0.52	0.827	0.43
Chen et al.	HOG	0.902	0.40	0.876	0.65	0.893	0.58	0.890	0.54
Quan et al.	SIFT	0.910	0.56	0.885	0.67	0.893	0.58	0.896	0.60
Hati et al.	SIFT	0.777	0.33	0.621	0.43	0.738	0.20	0.712	0.32
Chang et al.	SIFT	0.726	0.27	0.759	0.36	0.797	0.36	0.761	0.33
MRW	SIFT	0.528	0.36	0.647	0.42	0.701	0.39	0.625	0.39
Jerripothula et al.	SIFT	0.818	0.48	0.847	0.69	0.813	0.50	0.826	0.56
Hsu et al.	VGG-16	0.936	**0.66**	0.914	0.79	0.876	0.59	0.909	0.68
Ours	ResNet-101	**0.941**	0.65	**0.940**	**0.82**	**0.922**	**0.63**	**0.935**	**0.70**

Visual comparisons of object co-segmentation

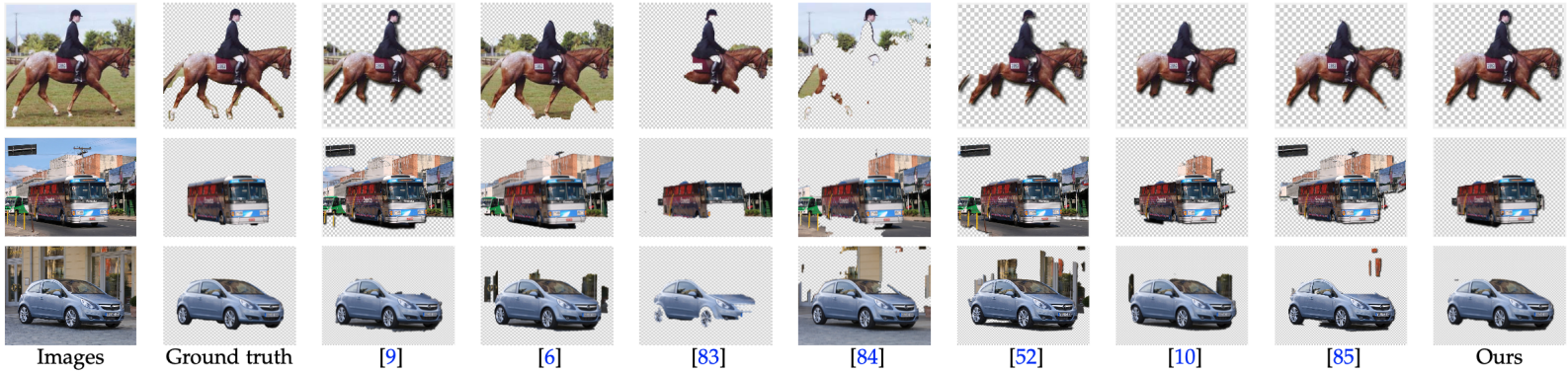


Figure: Visual comparisons on the TSS dataset.

Figure: Visual comparisons on the Internet dataset.

Evaluation of semantic matching on PF-PASCAL

Table: Experimental results of semantic matching on the PF-PASCAL dataset.

Method	Descriptor	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	d.table	dog	horse	moto	person	plant	sheep	sofa	train	tv	mean
Proposal Flow+LOM	HOG	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	GoogLeNet	64.8	58.7	42.8	59.6	47.0	42.2	61.0	45.6	49.9	52.0	48.5	49.5	53.2	72.7	53.0	41.4	83.3	49.0	73.0	66.0	55.6
A2Net	ResNet-101	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	59.0
GSF	ResNet-50	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	66.5
SCNet-AG+	VGG-16	85.5	84.4	66.3	70.8	57.4	82.7	82.3	71.6	54.3	**95.8**	55.2	59.5	68.6	75.0	56.3	60.4	60.0	**73.7**	66.5	76.7	72.2
CNNGeo	ResNet-101	83.0	82.2	81.1	50.0	57.8	79.9	92.8	77.5	44.7	85.4	28.1	69.8	65.4	77.1	64.0	65.2	**100.0**	50.8	44.3	54.4	69.5
CNNGeo w/ Inlier	ResNet-101	84.7	88.9	80.9	55.6	76.6	89.5	93.9	79.6	52.0	85.4	28.1	71.8	67.0	75.1	66.3	70.5	**100.0**	62.1	62.3	61.1	74.8
NC-Net	ResNet-101	**86.8**	86.7	**86.7**	55.6	82.8	88.6	93.8	**87.1**	54.3	87.5	43.2	**82.0**	64.1	**79.2**	71.1	71.0	60.0	54.2	**75.0**	**82.8**	78.9
WeakMatchNet	ResNet-101	85.6	**89.6**	82.1	**83.3**	**85.9**	92.5	93.9	80.2	52.2	85.4	55.2	75.2	64.0	77.9	67.2	**73.8**	**100.0**	65.3	69.3	61.1	78.0
Ours	ResNet-101	83.4	87.4	85.3	72.2	76.6	**94.6**	**94.7**	86.6	**54.9**	89.6	52.6	80.2	**70.6**	**79.2**	**73.3**	70.5	**100.0**	63.0	66.3	64.4	**79.0**

Evaluation of semantic matching on PF-WILLOW

Table: Experimental results of semantic matching on the PF-WILLOW dataset.

Method	Descriptor	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.15$
SIFT Flow	VGG-16	0.324	0.456	0.555
CNNGeo	ResNet-101	0.448	0.777	0.899
CNNGeo w/ Inlier	ResNet-101	0.477	0.812	0.917
Proposal Flow + LOM	HOG	0.284	0.568	0.682
UCN	GoogLeNet	0.291	0.417	0.513
SCNet-AG+	VGG-16	0.386	0.704	0.853
A2Net	ResNet-101	-	0.680	-
WeakMatchNet	ResNet-101	0.484	0.816	0.918
RTNs	ResNet-101	0.413	0.719	0.862
NC-Net	ResNet-101	<u>0.514</u>	<u>0.818</u>	<u>0.927</u>
Ours	ResNet-101	**0.538**	**0.854**	**0.939**

Evaluation of semantic matching on SPair-71k

Table: Experimental results of semantic matching on the SPair-71k dataset.

Method	Fine-tune	Avg.
CNNGeo		18.1
A2Net		20.1
CNNGeo w/ Inlier		21.1
NC-Net		<u>26.4</u>
Ours		25.8
CNNGeo	✓	20.6
A2Net	✓	22.3
CNNGeo w/ Inlier	✓	20.9
NC-Net	✓	20.1
Ours	✓	**26.6**

Visual comparisons of semantic matching

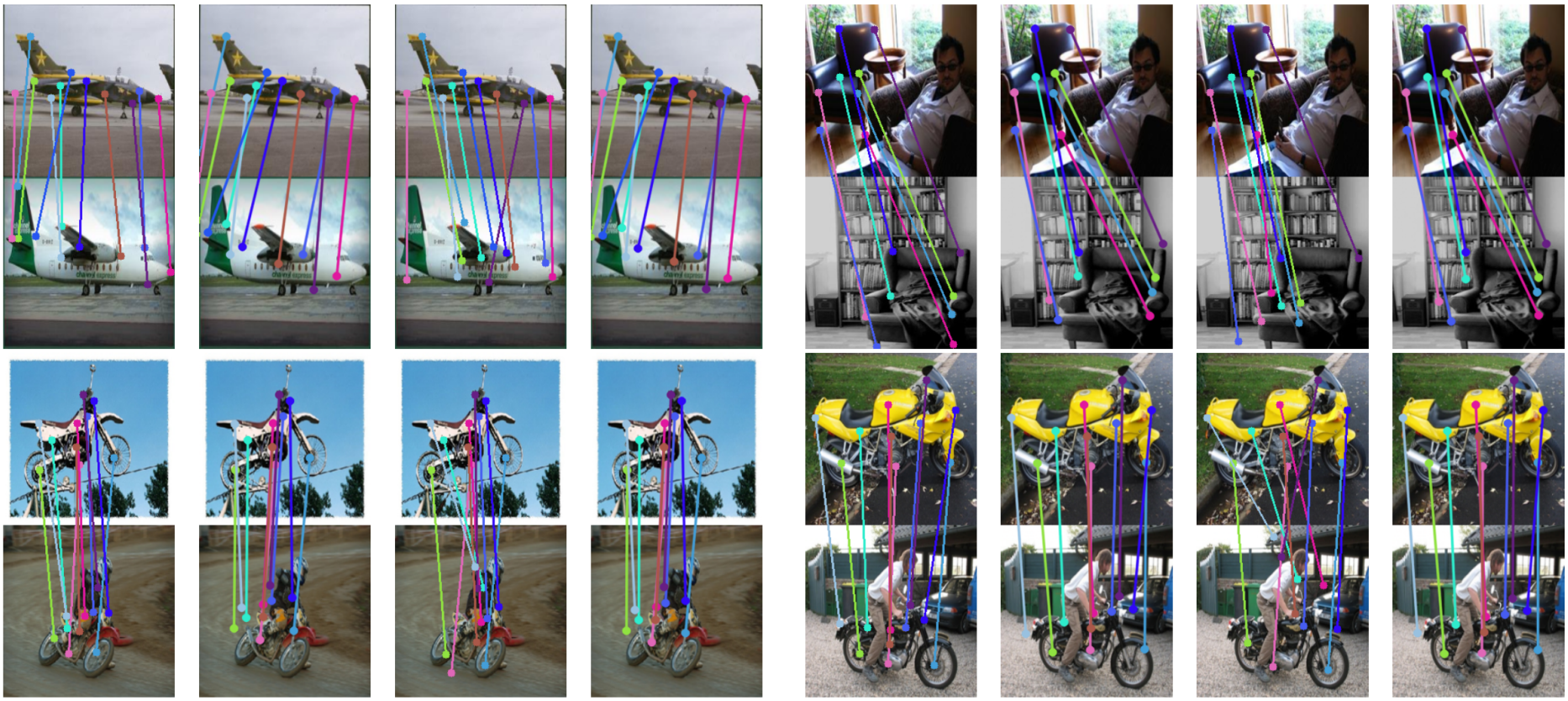
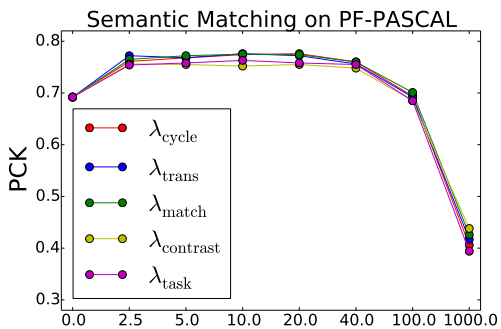


Figure: Visual comparisons on the PF-PASCAL (*top row*) and PF-WILLOW (*bottom row*) datasets.

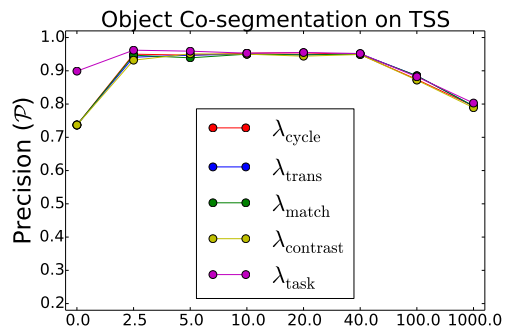
Sensitivity analysis on hyperparameters for training loss

- We analyze the sensitivity of our model by varying the value of each hyperparameter in the full training loss.

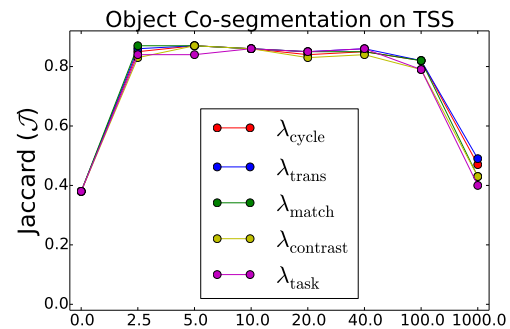
$$\mathcal{L} = \mathcal{L}_{\text{matching}} + \lambda_{\text{cycle}} \cdot \mathcal{L}_{\text{cycle-consis}} + \lambda_{\text{trans}} \cdot \mathcal{L}_{\text{trans-consis}} + \lambda_{\text{contrast}} \cdot \mathcal{L}_{\text{contrast}} + \lambda_{\text{task}} \cdot \mathcal{L}_{\text{task-consis}}, \tag{13}$$



Semantic matching (PCK)



Co-segmentation ($\mathcal{P}$)



Co-segmentation ($\mathcal{J}$)

- For semantic matching, the three most important hyperparameters are $\lambda_{\text{matching}}$, λ_{cycle}, and λ_{trans}.
- For object co-segmentation, the two most important hyperparameters are $\lambda_{\text{contrast}}$ and λ_{task}.

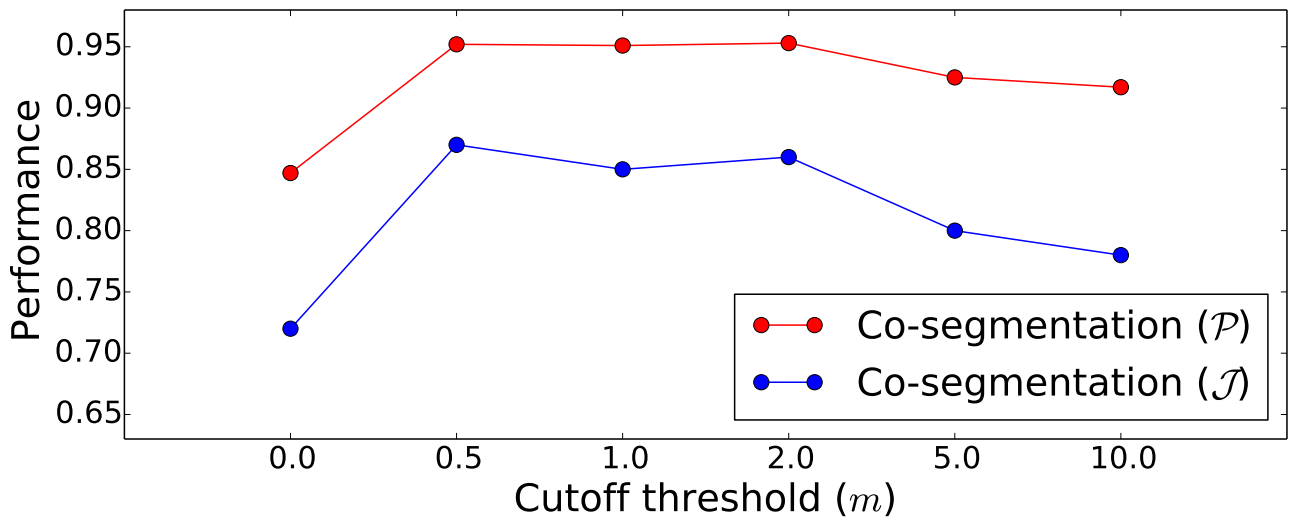
Sensitivity analysis on the cutoff threshold m

- We analyze the sensitivity of our model against the cutoff threshold m in the perceptual contrastive loss $\mathcal{L}_{\text{contrast}}$.

$$\mathcal{L}_{\text{contrast}} = d_{AB}^{+} + d_{AB}^{-}, \tag{14}$$

$$d_{AB}^{+} = \frac{1}{c}\|\mathcal{F}(I_A^o) - \mathcal{F}(I_B^o)\|^2 \text{ and} \tag{15}$$

$$d_{AB}^{-} = \max\left(0, m - \frac{1}{2c}\left(\|\mathcal{F}(I_A^o) - \mathcal{F}(I_A^b)\|^2 + \|\mathcal{F}(I_B^o) - \mathcal{F}(I_B^b)\|^2\right)\right). \tag{16}$$

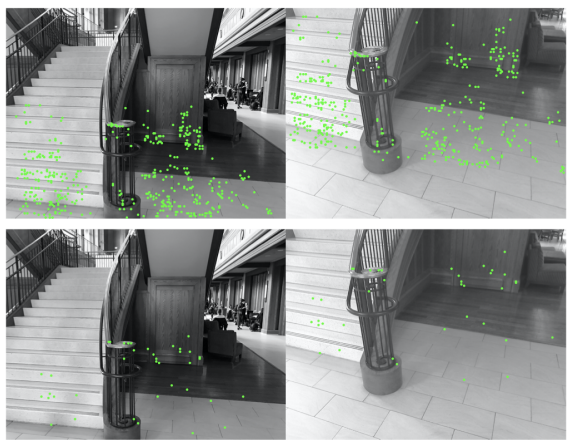


Limitations

- Our method may not work for images that contain multiple object instances.
- For semantic matching, our method predicts only one transformation matrix for an image pair. When multiple object instances are present, our method may not work well since multiple geometric transformations are required.
- For object co-segmentation, our method may fail if there exist background patches that are visually similar to the foreground objects.

Future work

- Joint semantic matching and object co-segmentation from images containing multiple object instances can potentially be addressed by instance-level semantic matching methods and instance co-segmentation approaches.



NC-Net [13]

DeepCO^3 [14]

[13] Rocco et al. Neighbourhood Consensus Networks. NeurIPS'18.

[14] Hsu et al. DeepCO^3: Deep Instance Co-segmentation by Co-peak Search and Co-saliency Detection. CVPR'19.

Outline

- Introduction
- Related work
- Proposed method
- Experimental results
- Conclusions

Conclusions

- We propose a weakly-supervised and end-to-end trainable network for joint semantic matching and object co-segmentation.
- To couple the 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 our method scalable to real-world applications.
- Experimental results demonstrate that our approach performs favorably against the state-of-the-art methods on both semantic matching and object co-segmentation tasks.

UNIVERSITY OF CALIFORNIA, MERCED

Show, Match and Segment: Joint Weakly-Supervised Learning of Semantic Matching and Object Co-Segmentation

Ming-Hsuan Yang
UC Merced / Google
http://vllab.ucmerced.edu

UCMERCED
www.ucmerced.edu

Weak or self supervision from images

- Exploit visual information at different levels
  - Within one image: pixels and regions
  - Between images: two or multiple views
- Exploit consistency
  - Appearance
  - Geometry
  - Semantics
  - Color
  - Forward/backward (cycle) matching
- Solve two or more tasks simultaneously
- Transfer learned models
- Exploit other image or video level information

Topics

- Show, match and segment [CVPR 19, PAMI 20]
  - Semantic matching and co-segmentation
- Joint-task self-supervised learning for temporal correspondence [NeurIPS 19]
  - Region and pixel correspondence
- Self-supervised co-part segmentation [CVPR 19]
  - Appearance, geometry, semantic
- Weakly-supervised semantic segmentation by iterative affinity learning [IJCV 20]
  - Caption information
- Video object segmentation via transferable representation [IJCV 20]
  - Adapt learned models to unseen objects

Weak or self supervision from images

- Exploit visual information at different levels
  - Within one image: pixels and regions
  - Between images: two or multiple views
- Exploit consistency
  - Appearance, geometry, semantics, color
- Solve two or more tasks simultaneously
- Transfer learned models
- Exploit other image or video level information