Joint Learning of Semantic Alignment and Object Landmark Detection

Sangryul Jeon[1], Dongbo Min[2], Seungryong Kim[3], Kwanghoon Sohn[1,*]
[1]Yonsei University, [2]Ewha Womans University, [3]École Polytechnique Fédérale de Lausanne (EPFL)
{cheonjsr,khsohn}@yonsei.ac.kr
dbmin@ewha.ac.kr,seungryong.kim@epfl.ch

Abstract

Convolutional neural networks (CNNs) based approaches for semantic alignment and object landmark detection have improved their performance significantly. Current efforts for the two tasks focus on addressing the lack of massive training data through weakly- or unsupervised learning frameworks. In this paper, we present a joint learning approach for obtaining dense correspondences and discovering object landmarks from semantically similar images. Based on the key insight that the two tasks can mutually provide supervisions to each other, our networks accomplish this through a joint loss function that alternatively imposes a consistency constraint between the two tasks, thereby boosting the performance and addressing the lack of training data in a principled manner. To the best of our knowledge, this is the first attempt to address the lack of training data for the two tasks through the joint learning. To further improve the robustness of our framework, we introduce a probabilistic learning formulation that allows only reliable matches to be used in the joint learning process. With the proposed method, state-of-the-art performance is attained on several standard benchmarks for semantic matching and landmark detection, including a newly introduced dataset, JLAD, which contains larger number of challenging image pairs than existing datasets.

1. Introduction

Establishing dense correspondences and discovering object landmarks over *semantically* similar images can facilitate a variety of computer vision and computational photography applications [3, 21, 22, 33, 2]. Both tasks aim to understand the underlying structure of an object that is geometrically consistent across different but semantically related instances.

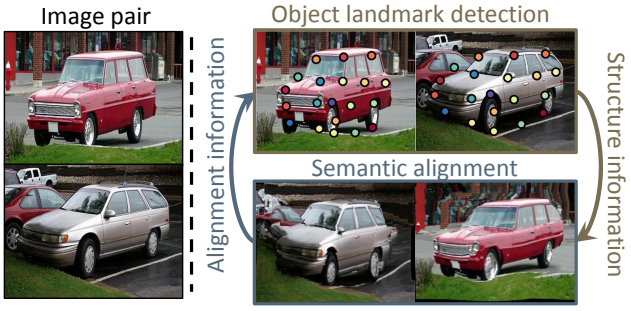


Figure 1. Illustration of the proposed joint learning framework: given only semantically similar image pairs, we address the crucial drawbacks of current weakly- or unsupervised models for the object landmark detection and semantic alignment task by alternatively leveraging mutual guidance information between them.

Recently, numerous approaches for the semantic alignment [24, 25, 26, 9, 27, 15] and object landmark detection [30, 29, 33, 8] have been proposed to tackle each problem with deep convolutional neural networks (CNNs) in an end-to-end manner. However, supervised training for such tasks often involves in constructing large-scale and diverse annotations of dense semantic correspondence maps or object landmarks. Collecting such annotations under large intra-class appearance and shape variations requires a great deal of manual works and is prone to error due to its subjectiveness. Consequently, current efforts have focused on using additional constraints or assumptions that help their networks to automatically learn each task in a weakly- or unsupervised setting.

To overcome the limitations of insufficient training data for semantic correspondence, several works [24, 27] have been proposed to utilize a set of sparse corresponding points between source and target images as an additional cue for supervising their networks. The key idea is to regulate the densely estimated transformation fields to be consistent with the given sparse corresponding points. A possible approach is to synthetically generate the corresponding points from an image itself, i.e., by uniformly sampling grid points from a source image and then globally deforming them with

This research was supported by R&D program for Advanced Integrated-intelligence for Identification (AIID) through the National Research Foundation of KOREA (NRF) funded by Ministry of Science and ICT (NRF-2018M3E3A1057289).
*Corresponding author

random transformations [10]. However, these synthetic supervisions do not consider photometric variations at all and have difficulties in capturing realisitic geometric deformations. Alternatively, several methods [25, 9] alleviate this issue by collecting tentative correspondence samples from real image pairs during training, but this is done in a simple manner, e.g., by thresholding [25] or checking consistency [9] with the matching scores. More recently, instead of using sparsely collected samples, some methods [15, 26] have employed a complete set of dense pixel-wise matches to estimate locally-varying transformation fields, outperforming previous methods based on a global transfomation model [24, 25, 27]. However, they often show limited performances in handling relatively large geometric variations due to their weak implicit smoothness constraints such as constraining transformation candidates within local window [15] and analyzing local neighbourhood patterns [26].

Meanwhile, to automatically discover object landmarks without the need of ground-truth labels, following a pioneering work of Thewlis *et al*. [30], dense correspondence information across the different instances have been used to impose the equivariance constraint, such that the landmarks should be consistently detectable with respect to given spatial deformations [30, 29, 33, 28]. However, while semantically meaningful and highly accurate correspondences are required to meet the full equivariance, existing techniques mostly rely on synthetic supervisions in a way of generating dense correspondence maps with randomly transformed imagery. Similar to existing semantic alignment approaches that leverage synthetic supervision [24, 27], as shown in [25, 9], they usually do not generalize well to real image pairs and often fail to detect landmarks at semantically meaningful locations of the object.

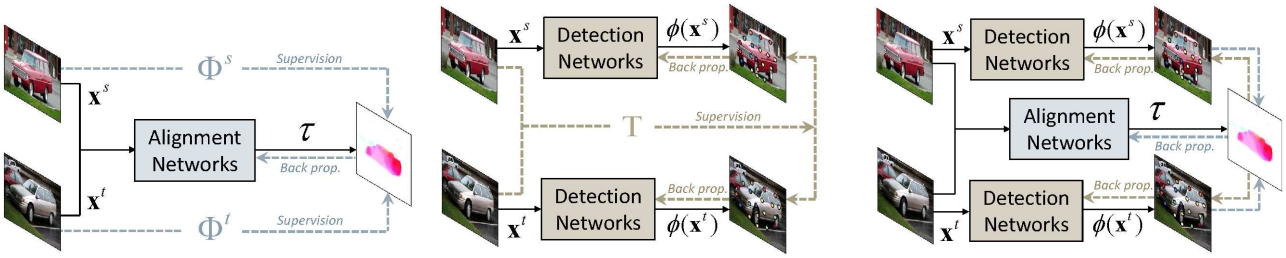
In this paper, we present a method for jointly learning object landmark detection and semantic alignment to address the aforementioned limitations of current weakly- or unsupervised learning models of each task. As illustrated in Fig. 1, our key observation is that the two tasks are mutually complementary to each other since more realistic and informative supervisions can be provided from their counterparts. To be specific, the detected landmarks can offer structure information of an object for the semantic alignment networks where the estimated correspondence fields are encouraged to be consistent with provided object structures. At the same time, densely estimated correspondences across semantically similar image pairs facilitate the landmarks to be consistently localized even under large intra-class variations. Our networks accomplish this by introducing a novel joint objective function that alternatively imposes the consistency constraints between the two tasks, thereby boosting the performance and addressing the lack of training data in a principled manner. We further improve the robustness of our framework by allowing only reliable matches to be used in the joint learning process through a probabilistic learning formulation of the semantic alignment networks. Experimental results on various benchmarks demonstrate the effectiveness of the proposed model over the latest methods for object landmark detection and semantic alignment.

2. Related Work

Semantic alignment Recent state-of-the-art techniques for semantic alignment generally regress the transformation parameters directly through an end-to-end CNN model [24, 25, 9, 15, 27], outperforming conventional methods based on hand-crafted descriptor or optimization [14, 21, 4]. Rocco *et al*. [24, 25] proposed a CNN architecture that estimates image-level transformation parameters mimicking traditional matching pipeline, such that feature extraction, matching, and parameter regression. Seo *et al*. [27] extended this idea with an offset-aware correlation kernel to focus on reliable correlations, filtering out distractors. While providing the robustness against semantic variations to some extent, they have difficulties in yielding fine-grained localization due to the assumption of a global transformation model. To address this issue, Jeon *et al*. [9] proposed a pyramidal graph model that estimates locally-varying geometric fields with coarse-to-fine scheme. Kim *et al*. [15] presented recurrent transformation networks that iteratively align features of source and target and finally obtain precisely refined local translational fields. Rocco *et al*. [26] proposed to analyze neighbourhood consensus patterns by imposing local constraints to find reliable matches among correspondence candidates. However, they rely on weak implicit smoothness constraints such as coarse-to-fine inference [9], constrained local search spaces [15], and local neighbourhood consensus [26]. In contrast, we explicitly regularize the estimated transformation fields to be consistent with the detected object landmarks through the joint learning process.

Object landmark detection Methods for unsupervised landmark detection generally rely on the equivariance property such that the object landmarks should be consistently detected with respect to given image deformations. As a pioneering work, Thewlis *et al*. [30] proposed to randomly synthesize the image transformations for learning to discover the object landmarks that are equivariant with respect to those transformations. They further extended this idea to learn dense object-centric coordinate frame [29]. Both of them rely on the synthetically generated supervisory signals and thus provide inherently limited performance when substantial intra-class variations are given.

Afterward, several works [33, 8] proposed an autoencoding formulation to discover landmarks as explicit structural representations in a way of generating new images conditioned on them. Zhang *et al*. [33] proposed to take object



(a) (b) (c)

Figure 2. Summary of the methods for: (a) semantic alignment [24, 25, 27, 9], (b) object landmark detection [30, 29, 33], and (c) the proposed joint learning framework. Our key observation is that each task can provide an essential supervisory signals to another one. With this motivation, we seamlessly weave both techniques to overcome the lack of training data.

landmarks as an intermediate learnable latent variable for reproducing the input image. Jakab *et al*. [8] proposed to generate images combining the appearance of the source image and the geometry of the target one by minimizing the perceptual distance. However, the ablation studies reported in [33, 8] show that they still rely on the supervision from an image itself such as synthesized image pairs or adjacent frames in videos instead of considering rich appearance variation between different object instances, thus yielding limited performance.

3. Method

3.1. Problem Statement and Overview

Let us denote *semantically* similar source and target images as $\mathbf{x}^s$ and $\mathbf{x}^t \in \mathbb{R}^{H\times W\times 3}$ where H and W denotes height and width of an image. We are interested in learning two mapping functions, $\phi : \mathbf{x} \rightarrow \mathbb{R}^{K\times 2}$ that extracts the spatial coordinates of K keypoints from an image $\mathbf{x}$ and $\tau : (\mathbf{x}^s, \mathbf{x}^t) \rightarrow \mathbb{R}^{H\times W\times 2}$ that infers a dense correspondence field from source to target image defined for each pixel in $\mathbf{x}^s$. We specifically learn the two functions through the joint prediction model using only weak supervision in the form of semantically similar image pairs. To address the insufficient training data for semantic correspondence, several methods [24, 25, 27, 9] utilized a set of sparse corresponding points on the source and target images, called anchor pairs, as an additional cue for supervising their networks. The key intuition is that the networks automatically learn to estimate geometric transformation fields over a set of transformation candidates by minimizing the discrepancy between given sparse correspondences. Specifically, denoting anchor pairs on source and target image as Φ^s and Φ^t, they define the semantic alignment loss as

$$\mathcal{L}_{\mathrm{A}}(\tau) = \sum\nolimits_n ||\Phi^t_n - \tau(\mathbf{x}^s, \mathbf{x}^t) \circ \Phi^s_n||^2, \qquad (1)$$

where n is the number of anchor pairs and $\circ$ is an warping operation. This is illustrated in Fig. 2(a). Meanwhile, to address the lack of training data for the landmark detection, the state-of-the-art techniques [29, 30, 33] generally employ dense correspondences between the training image pairs. The main idea lies in the equivariance constraint such that the detected landmarks should be equivariant with respect to given geometric deformation. Formally, denoting a dense correspondence map between source and target images as T, they aim to learn the landmark detection networks through a siamese configuration by minimizing

$$\mathcal{L}_{\mathrm{D}}(\phi) = \sum\nolimits_m ||\phi_m(\mathbf{x}^t) - \mathrm{T} \circ \phi_m(\mathbf{x}^s)||^2, \qquad (2)$$

where m is the number of detected landmarks. This is illustrated in Fig. 2(b).

However, current weakly- or unsupervised learning models for both tasks still suffer from the lack of supervisions of good quality, which may not fully satisfy their consistency constraints. To overcome this, we propose to leverage guidance information from each task for supervising another networks, as exemplified in Fig. 2(c). The proposed method offers a principled solution that overcomes the lack of massive training data by jointly learning the object landmark detection and semantic alignment in an end-to-end and boosting manner. To this end, we introduce a novel joint loss function that alternatively imposes the consistency constraints between the two tasks. To further enhance the joint learning process, we propose a probabilistic formulation that predicts and penalizes unreliable matches in the semantic alignment networks.

3.2. Network Architectures

The proposed networks consist of three sub-networks, including feature extraction networks with parameters $\mathbf{W}_F$ to extract feature maps from input images, landmark detection networks with parameters $\mathbf{W}_D$ to detect probability maps of landmarks, and semantic alignment networks with parameters $\mathbf{W}_A$ and $\mathbf{W}_C$ to infer a geometric transformation field and a uncertainty map, as illustrated in Fig. 3.

Feature extraction and similarity score computation To extract convolutional feature maps of source and target images, the input images are passed through a fully-

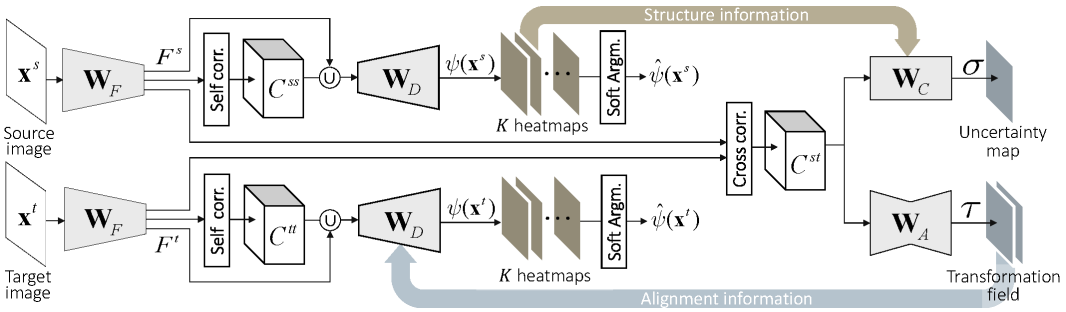


Figure 3. Network configuration of our framework, consisting of feature extraction networks, landmark detection networks, semantic alignment networks. We alternatively leverage the outputs from each landmark detection and semantic alignment networks as a guidance information for supervising the another one.

convolutional feature extraction networks with shared parameters $\mathbf{W}_F$ such that $F = \mathcal{F}(\mathbf{x}; \mathbf{W}_F) \in \mathbb{R}^{H \times W \times C}$. We share the parameters $\mathbf{W}_F$ for both feature extractions. After extracting the features, we normalize them using L_2 norm along the C channels.

The similarity between two extracted features is then computed as the cosine similarity with L_2 normalization:

$$C_i^{AB} = < F_i^A, F_j^B > / \sqrt{\sum_l < F_i^A, F_l^B >^2}, \quad (3)$$

where $j, l \in \mathcal{N}_i$ belong to the search window $\mathcal{N}_i$ centered at pixel i. Different from [24] that consider all possible samples within an image, we constrain search candidates within a local window to reduce matching ambiguity and runtime. The similarity score is finally normalized over the search candidates to reliably prune incorrect matches by down-weighting the influence of features that have multiple high scores [25]. Note that A and B represents source (s) or target (t) images. For instance, C^{ss} and C^{tt} indicate *self*-similarities computed from the source and target images, respectively. C^{st} is the *cross* similarity between source and target images.

Semantic alignment networks Our semantic alignment networks consist of two modules: an alignment module that estimates geometric transformation fields, and an uncertainty module that identifies which regions in an image are likely to be mismatched.

Taking the cross similarity scores between source and target images as an input, the alignment module based on an encoder-decoder architecture with parameters $\mathbf{W}_A$ estimates locally-varying transformation fields to deal with non-rigid geometric deformations more effectively, such that $\tau = \mathcal{F}(C^{st}; \mathbf{W}_A) \in \mathbb{R}^{H \times W \times 2}$. Different from recent semantic alignment approaches [15, 26] that estimate local geometric transformations, our alignment module employs the detected landmarks as an additional guidance information to focus more on the salient parts of the objects.

Additionally, inspired by the probabilistic learning model [12, 11], we formulate an uncertainty module that predicts how accurately the correspondences will be established at a certain image location. The predicted unreliable matches are prevented from being utilized during joint learning process to improve the robustness of our model against possible occlusions or ambiguous matches. Unlike existing methods [23, 22, 19, 12] where the uncertainty map is inferred from an input image, our uncertainty module leverages the matching score volume C^{st} to provide more informative cues, as in the approaches for confidence estimation in stereo matching [17]. Concretely, a series of convolutional layers with parameters $\mathbf{W}_C$ are applied to predict the uncertainty map σ from matching similarity scores C^{st} such that $\sigma = \mathcal{F}(C^{st}; \mathbf{W}_C) \in \mathbb{R}^{H \times W \times 1}$.

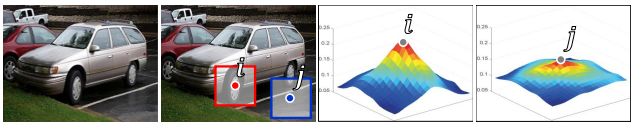
Landmark detection networks To enable our landmark detection networks to focus on more discriminative regions of the object, we explicitly supply local structures of an image by leveraging self-similarity scores C^{ss} and C^{tt} computed within a local window, as examplified in Fig. 5. This is different from existing methods [33, 8] that employ only convolutional features of images and thus often fail to detect semantically meaningful landmarks under challenging conditions.

Formally, we concatenate the extracted features F^s and F^t with self-similarity scores C^{ss} and C^{tt} respectively, and then pass them through the decoder style networks with parameters $\mathbf{W}_D$ to estimate $K+1$ detection score maps for K landmarks and one background, such that $\phi = \mathcal{F}(F \bigcup C; \mathbf{W}_D) \in \mathbb{R}^{H \times W \times (K+1)}$ where $\bigcup$ denotes a concatenation operator. The softmax layer is applied at the end of the networks to transform raw score maps into probability maps by normalizing across the $K+1$ channels,

$$\psi_i^k = \exp(\phi_i^k) / \sum_{m=0}^{K} \exp(\phi_i^m), \quad (4)$$

where ϕ^k is the score map of the k^{th} landmark. The spatial coordinate of the k^{th} landmark is then computed as an expected value over the spatial coordinate i weighted by its probability ψ_i^k, similar to the soft argmax operator in [13]:

$$\hat{\psi}^k = \sum_i i \cdot \psi_i^k / \sum_i \psi_i^k. \quad (5)$$



(a) (b) (c) (d)

Figure 4. Visualization of the effectiveness of self-similarity: (a) an image, (b) arbitrary two coordinates, i and j, (c) C_i^{ss}, and (d) C_j^{ss}. C^{ss} has high variance at more discriminative regions, providing local structure information to landmark detection networks.

This layer is differentiable, enabling us to formulate loss functions with respect to the landmark coordinates, which will be described in the next section.

3.3. Objective Functions

Loss for semantic alignment networks Our semantic alignment networks are learned using weak image-level supervision in a form of matching image pairs. Concretely, we start with recent weakly-supervised learning techniques proposed in [16, 15]. Under the assumption that corresponding features of source and target images are identical, they cast the semantic alignment into a classification task such that the networks can learn the geometric field as a hidden variable over a set of transformation candidates. However, this strict assumption is often violated, *e.g.* around occlusions, textureless regions and background clutters, thus requiring additional object location priors to penalize regions where the assumption is invalid.

To address this, we propose to identify *unreliable matches* through the probabilistic formulation of cross-entropy loss such that

$$\mathcal{L}_{\mathrm{A}}(\tau,\sigma)=\sum_{i}(-\sum_{j\in\mathcal{M}_i}\frac{s_j^*}{\sigma_i}\log(s_{i,j}(\tau_i))+\log\sigma_i), \quad (6)$$

where σ is the predicted uncertainty map with parameters $\mathbf{W}_C$ and $s_{i,j}(\tau)$ is a softmax probability defined as

$$s_{i,j}(\tau)=\frac{\exp(<F_i^s,[\tau\circ F^t]_j>)}{\sum_{l\in\mathcal{M}_i}\exp(<F_i^s,[\tau\circ F^t]_l>)}. \quad (7)$$

For $j\in\mathcal{M}_i$, a class label s_j^* is set to 1 if $j=i$, and 0 otherwise such that a center point i becomes a positive sample while other points within $\mathcal{M}_i$ are negative samples. By dividing the cross entropy loss with the predicted uncertainty map σ, we can penalize unreliable matches and avoid them to disrupt the loss function. The $\log\sigma$ serves as a regularization term to prevent σ them from becoming too large.

Losses for landmark detection networks Following [30, 33, 28], our landmark detection networks are designed to meet the two common characteristics of landmarks, such that each probability map $\hat{\psi}$ should concentrate on a discriminative local region and, at the same time, distributed at different parts of an object.

(a) (b) (c) (d)

Figure 5. Visualization of the effectiveness of probabilistic learning formulation: warped results using correspondences learned (a) from (1) with synthetic supervisions, (b) from (6) without probabilistic formulation, (c) from (6), and (d) uncertainty map where the darker pixels represent high degree of uncertainty.

The first constraint is used to define a concentration loss $\mathcal{L}_{\mathrm{con}}(\psi)$ that minimizes the variance over the spatial coordinate i with respect to the landmark coordinate ϕ [33]:

$$\mathcal{L}_{\mathrm{con}}(\psi)=\sum_k(\sum_i(i-\hat{\psi}^k)^2\cdot\psi_i^k/\sum_i\psi_i^k). \quad (8)$$

For the second constraint, we define a hinge embedding loss that encourages the landmarks to be far away than a margin c [28], such that

$$\mathcal{L}_{\mathrm{sep}}(\psi)=\sum_k\sum_{k'\neq k}\max(0,c-||\hat{\psi}^k-\hat{\psi}^{k'}||^2). \quad (9)$$

A final loss for the landmark detection networks is defined as a weighted sum of concentration and separation loss, such that $\mathcal{L}_{\mathrm{D}}(\psi)=\lambda_{\mathrm{con}}\mathcal{L}_{\mathrm{con}}(\psi)+\lambda_{\mathrm{sep}}\mathcal{L}_{\mathrm{sep}}(\psi)$.

Note that similar loss functions are used in the landmark detection literatures [30, 33], but our method is different in that more realistic supervisory signals for training the landmark detection networks are provided from the semantic alignment networks.

Loss for joint training Here, we integrate two independent learning processes into a single model by formulating an additional constraint for joint training. We apply the outputs of two tasks to a joint distance function as a form of

$$L_{\mathrm{J}}(\psi,\tau,\sigma)=\sum_k\sum_i\frac{1}{\sigma_i}||\psi_i^k(\mathbf{x}^s)-\tau\circ\psi_i^k(\mathbf{x}^t)||^2. \quad (10)$$

By imposing the consistency constraint between the landmark detection and semantic alignment, the joint loss function allows us to mutually take advantage of guidance information from both tasks, boosting the performance and addressing the lack of training data in a principled manner. Furthermore, we mitigate the adverse impact of unreliable matches in the joint learning process by discounting the contributions of them with the predicted uncertainty map σ_i. Note that instead of landmark coordinates $\hat{\psi}$ in (10), the probability map ψ is utilized for a stronger spatial consistency between two tasks. A final objective can be defined as a weighted summation of the presented three losses:

$$\mathcal{L}_{\mathrm{JDA}}(\psi,\tau,\sigma)=\lambda_{\mathrm{D}}\mathcal{L}_{\mathrm{D}}(\psi)+\lambda_{\mathrm{A}}\mathcal{L}_{\mathrm{A}}(\tau,\sigma)+\lambda_{\mathrm{J}}\mathcal{L}_{\mathrm{J}}(\psi,\tau,\sigma). \quad (11)$$

(a) (b) (c) (d)

Figure 6. The effectiveness of the proposed joint learning framework: detected landmarks and aligned images (a), (b) when learned separately, and (c), (d) when learned jointly .

3.4. Training

Alternative Optimization To optimize the landmark detection and semantic alignment networks in a mutually reinforced way, we learn the landmark detection networks and semantic alignment networks in an alternating fashion. For better initialization, we first pretrain both networks independently with synthetically generated image pairs, similar to [25]. Randomly perturbed images are generated by applying global affine or TPS transformation to the original images from the Pascal VOC 2012 segmentation dataset [1], and utilize these image pairs for learning each networks with loss functions (2) and (1). In sequence, we finetune both pretrained networks in an end-to-end manner for semantically similar images pairs from the JLAD dataset described in the following section. Specifically, the network parameters $\{\mathbf{W}_F, \mathbf{W}_A, \mathbf{W}_C\}$ are optimized for semantic alignment by setting $\{\lambda_D, \lambda_A, \lambda_J\}$ as $\{1, 10, 10\}$, and $\{\mathbf{W}_F, \mathbf{W}_D\}$ for landmark detection by setting $\{\lambda_D, \lambda_A, \lambda_J\}$ as $\{10, 1, 100\}$. We iterate this process until the final objective converges.

JLAD Dataset To learn our networks with the proposed consistency constraint (11), large-scale semantically similar image pairs are required, but existing public datasets are limited quantitatively. To overcome this, we introduce a new dataset that contains a larger number of challenging image pairs, called JLAD dataset. The images and keypoint annotations are sampled and refined from the original ones of PASCAL 3D benchmark [31] and MAFL dataset [34]. For each object category in PASCAL 3D dataset [31] which provides about 36,000 images for 12 categories, we first preprocessed their images to contain only a single object. Specifically, the images are cropped according to the provided object bounding box annotations, including margins for background clutters. Then using the ground-truth viewpoint annotations such as azimuth and elevation angles, we sampled about 1,000 image pairs for each category. For human faces, we sampled image pairs randomly from MAFL dataset [34] excluding testing set without considering geometric constraints since their images are already cropped and aligned. We used the split which divides the collected image pairs into roughly 70 percent for training, 20 percent for validation, and 10 percent for testing.

Methods	Alignment acc.	Detection acc.
	PCK@$\alpha = 0.1$	IOD
Separate learning	63.2	7.97
Iteration 1	67.0	7.36
Iteration 2	70.2	7.16
Iteration 3	72.1	7.05
Ours	**72.7**	**6.92**

Table 1. Ablation study for the effectiveness of the proposed joint learning framework on the JLAD dataset. The accuracies for semantic alignment and object landmark detection are reported with PCK and IOD metrics, respectively.

4. Experimental Results

4.1. Experimental Settings

For feature extraction, we used the ImageNet-pretrained ResNet [7], where the activations are sampled after pooling layers such as 'conv4-23' for ResNet-101 [7]. Margin c is set to be 0.05, 0.03, 0.02 for detecting 10, 15, 30 landmarks respectively. The radius of the search space for $\mathcal{N}_i$ is set to 5, which is equivalent to 40×40 window at the original resolution. Following [12], our uncertainty networks are formulated to predict log variance of uncertainty, *i.e.* $\log\sigma$, to avoid a potential division of (6) by zero. During alternative optimization, we set the maximum number of alternation to 4 to avoid overfitting. We used ADAM optimizer [18] with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. We set the training batch size to 16. A learning rate initially set to 10^{-3} and decreases to 10^{-4} and 10^{-5} later.

In the following, we comprehensively evaluated our framework in comparison to state-of-the-art methods for landmark detection, including FPE [30], DEIL [29], StrucRep [33], CIG [8], and for semantic alignment, including CNNgeo [24], CNNinlier [25], A2Net [27] and NC-Net [26]. Performance was measured on JLAD dataset and PF-PASCAL [5] for 12 object categories, and on MAFL dataset [34] and AFLW dataset [20] for human faces. See the supplemental material for more details on the implementation of our system and more qualitative results.

4.2. Ablation Study

We first analyze the effectiveness of the components within our method. The performances of landmark detection and semantic alignment are examined for different numbers of alternative iterations. The qualitative and quantitatve assessments are conducted on the testing image pairs of JLAD dataset. As shown in Table 1 and Fig. 6, the results of our joint learning model show significant improvements in comparision to separate estimation models that rely on synthetic transformations. We also conducted

Methods	aero.	bicy.	boat	bott.	bus	car	chair	d.table	motor.	sofa	train	tv.	All
CNNgeo [24]	71.3	74.4	44.4	60.9	79.6	83.8	63.9	36.6	72.1	43.8	42.5	48.0	60.1
CNNinlier [25]	79.6	82.9	54.4	68.7	89.5	88.5	70.7	39.2	79.4	48.2	49.4	51.1	66.8
A2Net [27]	80.9	81.4	53.6	69.5	88.6	89.5	71.3	41.2	78.1	51.8	52.0	51.7	67.5
RTNs [15]	81.5	85.4	56.3	70.8	87.4	92.7	72.3	43.6	84.3	59.8	55.2	53.5	70.2
NCNet [26]	82.4	85.2	57.9	71.2	88.8	93.1	**75.8**	**46.9**	87.8	57.7	57.1	**56.5**	71.7
Ours	**84.7**	**89.1**	**62.5**	**74.5**	**90.3**	**93.3**	73.3	46.7	**89.4**	**60.7**	**62.1**	56.3	**73.6**

Table 2. Matching accuracy compared to the state-of-the-art semantic alignment techniques over various object categories on the JLAD dataset. The distance threshold of PCK α is set to 0.01.

Methods	PCK		
	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.15$
CNNgeo [24]	36.9	62.3	71.4
CNNinlier [25]	44.1	68.2	74.8
A2Net [27]	43.1	68.4	74.1
RTNs [15]	49.2	69.3	76.2
NCNet [26]	50.7	70.9	78.1
Ours wo/UM	49.4	68.2	76.9
Ours	**52.8**	**72.7**	**79.2**

Table 3. Matching accuracy compared to the state-of-the-art correspondence techniques on the PF-PASCAL benchmark [5].

Methods	K	MAFL	ALFW	K	JLAD
FPE [30]	50	6.67	10.53	20	13.32
DEIL [29]	-	5.83	8.80	-	10.76
StrucRep [33]	30	3.16	6.58	20	7.33
CIG [8]	30	3.08	6.98	20	12.87
Ours wo/SS	30	3.58	7.72	20	8.16
Ours	10	3.33	7.17	10	7.54
	30	**2.98**	**6.51**	20	**6.92**

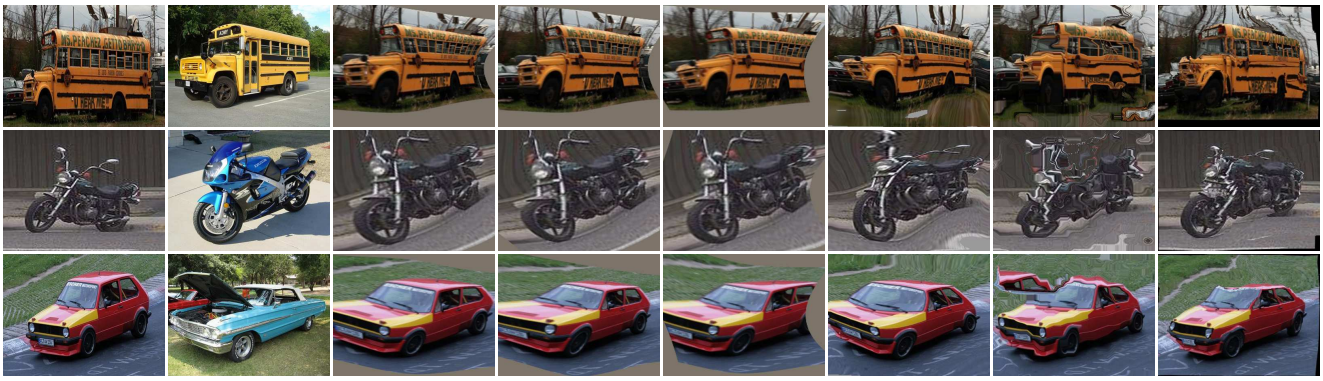
Table 4. Comparison with state-of-the-art landmark detection techniques on the MAFL [34], ALFW [20], and JLAD dataset. K denotes the number of used landmarks for linear regressor.

an ablation study by removing the uncertainty prediction model within semantic alignment networks (Ours wo/UM) and the correlation layer within landmark detection network that computes local self-similarties (Ours wo/SS). Degraded performance of "Ours wo/SS" and "Ours wo/UM" in Table 1 and Table 2 highlights the importance of encoding local structure through self-similarities for landmark detection and considering possible ambiguous matches for semantic alignment.

4.3. Results

Semantic alignment We evaluated our semantic alignment networks over 12 object categories on the JLAD dataset and the PF-PASCAL benchmark [5]. For the evaluation metric, we used the percentage of correct keypoints (PCK) metric [32] which counts the number of keypoints having a transfer error below a given threshold α, following the procedure employed in [6]. Table 2 and Table 3 summarize the PCK values, and Fig. 7 shows qualitative results. The results of detected landmarks of each image pair in Fig. 7 are visualized in Fig. 8. As shown in Table 2, Table 3, Fig. 7 our results have shown highly improved performance qualitatively and quantitatively compared to the methods [24, 25, 9, 27] that rely on synthetically or heuristically collected correspondence samples. This reveals the effect of the proposed joint learning technique where the structural smoothness is naturally imposed with respect to the detected object landmarks. This is in contrast to the methods that employ weak implicit smoothness constraints, such as image-level global transformation model [24, 25, 27], locally constrained transformation candidates [15], or local neighbourhood consensus [26].

Object landmark detection We evaluated our landmark detection networks for human faces on MAFL and AFLW benchmarks [34, 20], including various objects on JLAD dataset. For the evaluation on MAFL benchmark [34], we trained our model with facial image pairs in the CelebA training set excluding those appearing in the MAFL test set. For AFLW benckmark [20], we further finetune the pretrained networks on AFLW training image sets, simlar to [33, 30]. To evaluate our discovered landmarks quality, we use a linear model without a bias term to regress from the discovered landmarks to the human-annotated landmarks [33, 30, 29, 8]. Ground-truth landmark annotations of testing image pairs are provided to train this linear regressor. We follow the standard MSE metric in [34] and report performances in inter-ocular distance (IOD). Fig. 8 shows qualitative results on JLAD dataset and Fig. 9 for MAFL benchmark [34]. Table 4 shows that our method achieves the state-of-the-art performance compared with existing models [33, 30] that use synthesized image deformations for training their networks. The relatively modest gain on human faces compared to other object catogories may come from the limited appearance and geometric variations on MAFL and AFLW benchmarks, where the faces are cropped and aligned including little background clutters. A visual comparison of Fig. 8 and quantitative results of Table 4 demonstrate the benefits of joint learning with semantic alignment networks. Unlike existing methods [33, 30, 29, 8] that do not consider rich variations from the real image pairs, our method consistently discovers semantically meaningful landmarks over various object cate-

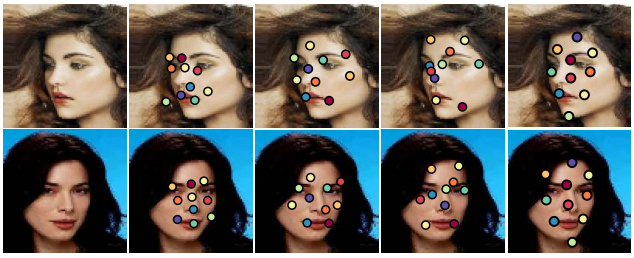


(a) (b) (c) (d) (e) (f) (g) (h)

Figure 7. Qualitative results of the semantic alignment on the JLAD dataset: (a) source image, (b) target image, (c) CNNgeo [24], (d) CNNinlier [25], (e) A2Net [27], (f) RTNs [15], (g) NCNet [26], and (h) Ours. The source images were warped to the target images using correspondences.

(a) (b) (c) (d) (e) (f) (g) (h)

Figure 8. Qualitative results of the object landmark detection on the JLAD dataset: (a), (b) ground-truth landmarks, the image pairs of Fig. 7 are used to discover landmarks with (c), (d) CIG [8], (e), (f) StrucRep [33], and (g), (h) Ours.



(a) (b) (c) (d) (e)

Figure 9. Qualitative results of the object landmark detection on the MAFL benchmark [34]: (a) ground-truth landmarks, (b) FPE [30], (c) StrucRep [33], (d) CIG [8], (e) Ours.

(a) (b) (c) (d) (e)

Figure 10. Qualitative results of our semantic aligment networks on the MAFL benchmark: (a) source image, (b) target image, (c), (d) detected landmarks on source and target image, (e) warped image using correspondences.

gories even under large appearance and shape variations.

5. Conclusion

We presented a joint learning framework for the landmark detection and semantic correspondence that utilizes the complementary interactions between the two tasks to overcome the lack of training data by alternatively imposing the consistent constraints. Experimental results on various benchmarks, including a newly introduced JLAD dataset, demonstrate the effectiveness of our method, such that the image pairs can be precisely aligned with the intrinsic structures of detected landmarks, and at the same time the landmarks can be consistently discovered with estimated semantic correspondence fields.

References

[1] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

[2] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *IEEE transactions on pattern analysis and machine intelligence*, 32(9):1627–1645, 2010.

[3] Yoav HaCohen, Eli Shechtman, Dan B Goldman, and Dani Lischinski. Non-rigid dense correspondence with applications for image enhancement. *ACM transactions on graphics (TOG)*, 30(4):70, 2011.

[4] Bumsub Ham, Minsu Cho, Cordelia Schmid, and Jean Ponce. Proposal flow. *In: CVPR*, 2016.

[5] Bumsub Ham, Minsu Cho, Cordelia Schmid, and Jean Ponce. Proposal flow: Semantic correspondences from object proposals. *IEEE Trans. PAMI*, 2017.

[6] Kai Han, Rafael S Rezende, Bumsub Ham, Kwan-Yee K Wong, Minsu Cho, Cordelia Schmid, and Jean Ponce. Scnet: Learning semantic correspondence. *In: ICCV*, 2017.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Sun. Jian. Deep residual learning for image recognition. *In: CVPR*, 2016.

[8] Tomas Jakab, Ankush Gupta, Hakan Bilen, and Andrea Vedaldi. Conditional image generation for learning the structure of visual objects. *Proceedings of the 32nd Conference on Neural Information Processing Systems*, 2018.

[9] Sangryul Jeon, Seungryong Kim, Dongbo Min, and Kwanghoon Sohn. Pyramidal affine regression networks for dense semantic correspondence. In *ECCV*, 2018.

[10] Angjoo Kanazawa, David W Jacobs, and Manmohan Chandraker. Warpnet: Weakly supervised matching for single-view reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3253–3261, 2016.

[11] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems (NIPS)*, 2017.

[12] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[13] Alex Kendall, Hayk Martirosyan, Saumitro Dasgupta, Peter Henry, Ryan Kennedy, Abraham Bachrach, and Adam Bry. End-to-end learning of geometry and context for deep stereo regression. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2017.

[14] Jaechul Kim, Ce Liu, Fei Sha, and Kristen Grauman. Deformable spatial pyramid matching for fast dense correspondences. *In: CVPR*, 2013.

[15] Seungryong Kim, Stephen Lin, Sangryul Jeon, Dongbo Min, and Kwanghoon Sohn. Recurrent transformer networks for semantic correspondence. In *Advances in Neural Information Processing Systems*, 2018.

[16] Seungryong Kim, Dongbo Min, Bumsub Ham, Stephen Lin, and Kwanghoon Sohn. Fcss: Fully convolutional self-similarity for dense semantic correspondence. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.

[17] Sunok Kim, Dongbo Min, Seungryong Kim, and Kwanghoon Sohn. Unified confidence estimation networks for robust stereo matching. *IEEE Transactions on Image Processing*, 28(3):1299–1313, 2019.

[18] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *In: ICLR*, 2015.

[19] Maria Klodt and Andrea Vedaldi. Supervising the new with the old: Learning sfm from sfm. In *European Conference on Computer Vision*, pages 713–728. Springer, 2018.

[20] Martin Koestinger, Paul Wohlhart, Peter M Roth, and Horst Bischof. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In *2011 IEEE international conference on computer vision workshops (ICCV workshops)*, pages 2144–2151. IEEE, 2011.

[21] Ce Liu, Jenny Yuen, and Antonio Torralba. Sift flow: Dense correspondence across scenes and its applications. *IEEE Trans. PAMI*, 33(5):815–830, 2011.

[22] David Novotny, Samuel Albanie, Diane Larlus, and Andrea Vedaldi. Self-supervised learning of geometrically stable features through probabilistic introspection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[23] David Novotny, Diane Larlus, and Andrea Vedaldi. Learning 3d object categories by looking around them. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 5228–5237. IEEE, 2017.

[24] Ignacio Rocco, Relja Arandjelović, and Josef Sivic. Convolutional neural network architecture for geometric matching. *In: CVPR*, 2017.

[25] Ignacio Rocco, Relja Arandjelović, and Josef Sivic. End-to-end weakly-supervised semantic alignment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6917–6925, 2018.

[26] Ignacio Rocco, Mircea Cimpoi, Relja Arandjelović, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Neighbourhood consensus networks. In *Advances in Neural Information Processing Systems*, pages 1658–1669, 2018.

[27] Paul Hongsuck Seo, Jongmin Lee, Deunsol Jung, Bohyung Han, and Minsu Cho. Attentive semantic alignment with offset-aware correlation kernels. In *ECCV*, 2018.

[28] Supasorn Suwajanakorn, Noah Snavely, Jonathan J Tompson, and Mohammad Norouzi. Discovery of latent 3d keypoints via end-to-end geometric reasoning. In *Advances in Neural Information Processing Systems*, pages 2063–2074, 2018.

[29] James Thewlis, Hakan Bilen, and Andrea Vedaldi. Unsupervised learning of object frames by dense equivariant image labelling. In *Advances in Neural Information Processing Systems*, pages 844–855, 2017.

[30] James Thewlis, Hakan Bilen, and Andrea Vedaldi. Unsupervised learning of object landmarks by factorized spatial embeddings. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5916–5925, 2017.

[31] Yu Xiang, Roozbeh Mottaghi, and Silvio Savarese. Beyond pascal: A benchmark for 3d object detection in the wild. In *IEEE Winter Conference on Applications of Computer Vision*, pages 75–82. IEEE, 2014.

[32] Yi Yang and Deva Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *CVPR 2011*, pages 1385–1392. IEEE, 2011.

[33] Yuting Zhang, Yijie Guo, Yixin Jin, Yijun Luo, Zhiyuan He, and Honglak Lee. Unsupervised discovery of object landmarks as structural representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2694–2703, 2018.

[34] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Facial landmark detection by deep multi-task learning. In *European conference on computer vision*, pages 94–108. Springer, 2014.