



PARN: Pyramidal Affine Regression Networks for Dense Semantic Correspondence

Sangryul Jeon¹, Seungryong Kim¹, Dongbo Min², and Kwanghoon Sohn¹(✉)

¹ Yonsei University, Seoul, South Korea

{cheonjsr, srkim89, khsohn}@yonsei.ac.kr

² Ewha Womans University, Seoul, South Korea

dbmin@ewha.ac.kr

Abstract. This paper presents a deep architecture for dense semantic correspondence, called pyramidal affine regression networks (PARN), that estimates locally-varying affine transformation fields across images. To deal with intra-class appearance and shape variations that commonly exist among different instances within the same object category, we leverage a pyramidal model where affine transformation fields are progressively estimated in a coarse-to-fine manner so that the smoothness constraint is naturally imposed within deep networks. PARN estimates residual affine transformations at each level and composes them to estimate final affine transformations. Furthermore, to overcome the limitations of insufficient training data for semantic correspondence, we propose a novel weakly-supervised training scheme that generates progressive supervisions by leveraging a correspondence consistency across image pairs. Our method is fully learnable in an end-to-end manner and does not require quantizing infinite continuous affine transformation fields. To the best of our knowledge, it is the first work that attempts to estimate dense affine transformation fields in a coarse-to-fine manner within deep networks. Experimental results demonstrate that PARN outperforms the state-of-the-art methods for dense semantic correspondence on various benchmarks.

Keywords: Dense semantic correspondence · Hierarchical graph model

1 Introduction

Establishing dense correspondences across semantically similar images is essential for numerous computer vision and computational photography applications, such as scene parsing, semantic segmentation, and image editing [1–5].

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-030-01231-1_22) contains supplementary material, which is available to authorized users.

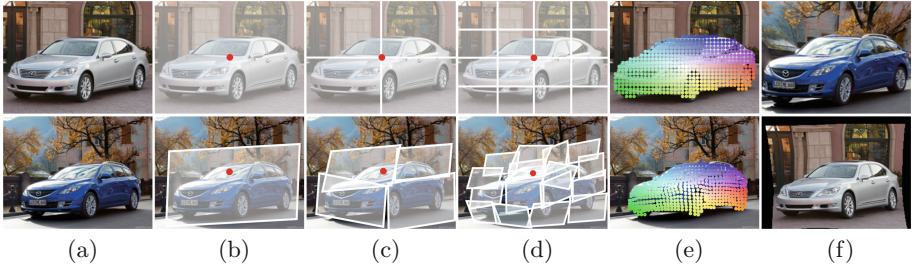


Fig. 1. Visualization of pyramidal model in the PARN: (a) source and target images, estimated affine field at (b) level 1, (c) level 2, (d) level 3, (e) pixel-level, and (f) warped images. In each grid at each level, PARN estimates corresponding affine transformation field regularized with the estimated transformation field at previous level.

Unlike classical dense correspondence tasks such as stereo matching [6] or optical flow estimation [7] that have been dramatically advanced, semantic correspondence estimation still remains unsolved due to severe intra-class appearance and shape variations across images. Several recent approaches [8, 9] have been proposed by leveraging deep convolutional neural networks (CNNs), providing satisfactory performances in capturing reliable matching evidences under intra-class appearance variations. However, they still consider geometric variations in just a limited manner such as those used for stereo matching or optical flow estimation [6, 7]. In some approaches [9, 10], more complex geometric variations such as scale or rotation were addressed, but they seek the labeling solution from only a set of scales and/or rotations quantized within pre-defined ranges. Recently, the discrete-continuous transformation matching (DCTM) framework [10] combined with the fully convolutional self-similarity (FCSS) [8] descriptor exhibits much improved performance by estimating locally-varying affine transformation fields on continuous and discrete domains in an alternative manner. Although DCTM has shown the state-of-the-art performance in dealing with non-rigid shape deformations, it is formulated with handcrafted smoothness constraint model and optimization technique, and thus it cannot guarantee optimal results when the geometric variation is relatively large.

In addition to the effort at measuring reliable matching evidences across images under intra-class appearance variations, recent CNN-based approaches have begun directly regressing geometric deformation fields through deep networks [11, 12]. As pioneering works, spatial transformer networks (STNs) [13] and its variant, inverse compositional spatial transformer networks (IC-STNs) [14], offer a way to deal with geometric variations within CNNs. Rocco et al. [12] and Schneider et al. [15] developed a CNN architecture for geometry-invariant matching that estimates transformation parameters across semantically similar images and different modalities. However, these methods assume the global transformation model, and thus they cannot deal with spatially-varying geometric variations, which frequently appear in dense semantic correspondence. More recently, some methods such as universal correspondence network (UCN)

[9] and deformable convolutional networks (DCN) [16] were proposed to encode locally-varying geometric variations in CNNs, but they do not have smoothness constraints with neighboring points, and cannot guarantee reliable performance under relatively large geometric variations. An additional challenge lies in the lack of training data with ground-truth for semantic correspondence, making the use of supervised training approaches difficult.

In this paper, we present a novel CNN architecture, called pyramidal affine regression networks (PARN), that estimates locally-varying affine transformation fields across semantically similar images in a coarse-to-fine fashion, as shown in Fig. 1. Inspired by pyramidal graph models [3, 17] that impose the hierarchical smoothness constraint on labeling results, our approach first estimates a global affine transformation over an entire image, and then progressively increases the degree of freedom of the transformation in a form of quad-tree, finally producing pixel-wise continuous affine transformation fields. The regression networks estimate residual affine transformations at each level and these are composed to provide final affine transformation fields. To overcome the limitations of insufficient training data for semantic correspondence, we propose a novel weakly-supervised training scheme that generates progressive supervisions by leveraging the correspondence consistency. Our method works in an end-to-end manner, and does not require quantizing the search space, different from conventional methods [17, 18]. To the best of our knowledge, it is the first attempt to estimate the locally-varying affine transformation fields through deep network in a coarse-to-fine manner. Experimental results show that the PARN outperforms the latest methods for dense semantic correspondence on several benchmarks including Taniai dataset [19], PF-PASCAL [20], and Caltech-101 [21].

2 Related Works

Dense Semantic Correspondence. Liu et al. [2] pioneered the idea of dense correspondence across different scenes, and proposed SIFT Flow. Inspired by this, Kim et al. [3] proposed the deformable spatial pyramid (DSP) which performs multi-scale regularization within a hierarchical graph. More recently, Yang et al. [22] proposed the object-aware hierarchical graph (OHG) to regulate matching consistency over whole objects. Among other methods are those that take an exemplar-LDA approach [23], employ joint image set alignment [5], or jointly solve for cosegmentation [19]. As all of these techniques use handcrafted descriptors such as SIFT [24] or DAISY [18], they lack the robustness to deformations that is possible with deep CNNs.

Recently CNN-based descriptors have been used to establish dense semantic correspondences because of their high invariance to appearance variations. Zhou et al. [25] proposed a deep network that exploits cycle-consistency with a 3-D CAD model [26] as a supervisory signal. Choy et al. [9] proposed the universal correspondence network (UCN) based on fully convolutional feature learning. Novotny et al. [27] proposed AnchorNet that learns geometry-sensitive features for semantic matching with weak image-level labels. Kim et al. [8] proposed the

FCSS descriptor that formulates local self-similarity within a fully convolutional network. However, none of these methods is able to handle severe non-rigid geometric variations.

Transformation Invariance. Several methods have aimed to alleviate geometric variations through extensions of SIFT Flow, including scale-less SIFT Flow (SLS) [28], scale-space SIFT Flow (SSF) [29], and generalized DSP [17]. However, these techniques have a critical and practical limitation that their computational cost increases linearly with the search space size. HaCohen et al. [1] proposed in a non-rigid dense correspondence (NRDC) algorithm, but it employs weak matching evidence that cannot guarantee reliable performance. Geometric invariance to scale and rotation is provided by DAISY Filer Flow (DFF) [4], but its implicit smoothness constraint often induces mismatches. Recently, Ham et al. [30] presented the Proposal Flow (PF) algorithm to estimate correspondences using object proposals. Han et al. [31] proposed SCNet to learn the similarity function and geometry kernel of PF algorithm within deep CNN. While these aforementioned techniques provide some amount of geometric invariance, none of them can deal with affine transformations across images, which frequently occur in dense semantic correspondence. More recently, Kim et al. [10] proposed DCTM framework where dense affine transformation fields are inferred using a handcrafted energy function and optimization.

STNs [13] offer a way to deal with geometric variations within CNNs by warping features through a global parametric transformation. Lin et al. [14] proposed IC-STNs that replaces the feature warping with transformation parameter propagation. Rocco et al. [12] proposed a CNN architecture for estimating a geometric model such as an affine transformation for semantic correspondence estimation. However, it only estimates globally-varying geometric fields, and thus exhibits limited performance for dealing with locally-varying geometric deformations. Some methods such as UCN [9] and DCN [16] were proposed to encode locally-varying geometric variations in CNNs, but they do not have the smoothness constraints with neighboring points and cannot guarantee reliable performance for images with relatively large geometric variations [10].

3 Method

3.1 Problem Formulation and Overview

Given a pair of images I and I' , the objective of dense correspondence estimation is to establish a correspondence i' for each pixel $i = [i_x, i_y]$. In this work, we infer a field of affine transformations, each represented by a 2×3 matrix

$$\mathbf{T}_i = \begin{bmatrix} \mathbf{T}_{i,x} \\ \mathbf{T}_{i,y} \end{bmatrix} \quad (1)$$

that maps pixel i to $i' = \mathbf{T}_i \mathbf{i}$, where \mathbf{i} is pixel i represented in homogeneous coordinates such that $\mathbf{i} = [i, 1]^T$.

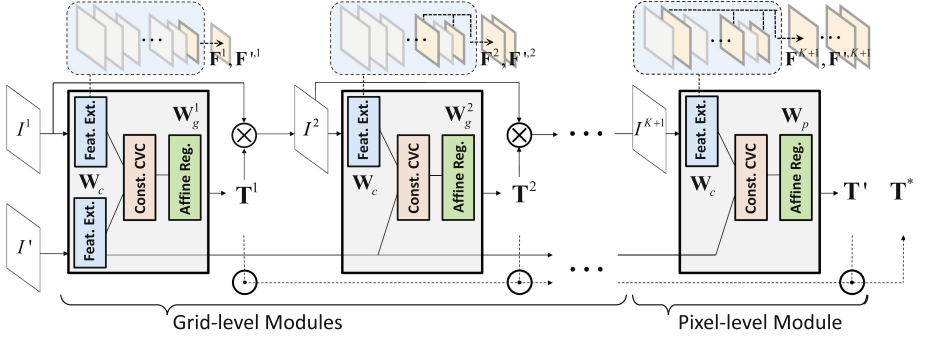


Fig. 2. Network configuration of the PARN, which is defined on the pyramidal model and consists of several grid-level modules and a single pixel-level module. Each module is designed to mimic the standard matching process within a deep architecture, including feature extraction, cost volume construction, and regression.

Compared to the constrained geometric transformation model (i.e. only translational motion) commonly used in the stereo matching or optical flow estimation, the affine transformation fields can model the geometric variation in a more principled manner. Estimating the pixel-wise affine transformation fields, however, poses additional challenges due to its infinite and continuous solution space. It is well-known in stereo matching literatures that global approaches using the smoothness constraint defined on the Markov random field (MRF) [32] tend to achieve higher accuracy on the labeling optimization, compared to local approaches based on the structure-aware cost aggregation [33]. However, such global approaches do not scale very well to our problem in terms of computational complexity, as the affine transformation is defined over the 6-D continuous solution space. Additionally, it is not easy to guarantee the convergence of affine transformation fields estimated through the discrete labeling optimization due to extremely large label spaces. Though randomized search and propagation strategy for labeling optimization [32, 34] may help to improve the convergence of labeling optimization on high-dimensional label space, most approaches just consider relatively lower-dimensional label space, e.g. 4-D label space consisting of translation, rotation, and scale.

Inspired by the pyramidal graph model [3, 17, 35] and the parametric geometry regression networks [11, 12], we propose a novel deep architecture that estimates dense affine transformation fields in a coarse-to-fine manner. Our key observation is that affine transformation fields estimated at a coarse scale tend to be robust to geometric variations while the results at a fine scale preserve fine-grained details of objects better. While conventional approaches that employ the coarse-to-fine scheme in dense correspondence estimation [2, 36] focus on *image scales*, our approach exploits *semantic scales* within the hierarchy of deep convolutional networks. Our method first estimates an image-level affine transformation using the deepest convolutional activations and then progressively localizes the affine transformation field additionally using the shallower convolutional acti-

vations in a quad-tree framework, producing the pixel-level affine transformation fields as the final labeling results.

As shown in Fig. 2, our method is defined on the *pyramidal model* (see Fig. 1) that consists of two kind of networks, several grid-level modules and a single pixel-level module, similar to [3, 17]. Each module within two networks is designed to mimic the standard matching process within a deep architecture [12]: feature extraction, correlation volume construction, and regression. Concretely, when two images I and I' are given, convolutional features are first extracted as multi-level intermediate activations through the feature network (with \mathbf{W}_c) in order to provide fine-grained localization precision ability at each level while preserving robustness to deformations. Then, the correlation volume is constructed between these features at the cost volume construction layer of Fig. 2. Finally the affine transformation fields are inferred by passing the correlation volume to the regression network (with $\mathbf{W}_g^k, \mathbf{W}_p$ of Fig. 2). This procedure is repeated for K grid-level modules and a single pixel-level module.

3.2 Pyramidal Affine Regression Networks

Each module of our pyramidal model has three main components. The first one extracts *hierarchically* concatenated features from the input images and the second computes a cost volume within *constrained* search windows. Lastly, from the third one, a *locally-varying* affine field is densely estimated for all pixels.

Feature Extraction. While conventional CNN-based descriptors have shown the excellent capabilities in handling intra-class appearance variations [37, 38], they have difficulties in yielding both semantic robustness and matching precision ability at the same time. To overcome this limitation, our networks are designed to leverage the inherent hierarchies of CNNs where multi-level intermediate convolutional features are extracted through a shared siamese network. We concatenate some of these convolutional feature maps such that

$$\mathbf{F}^k = \bigcup_{n \in M(k)} \mathcal{F}(I^k; \mathbf{W}_c^n) \quad (2)$$

where \bigcup denotes the concatenation operator, \mathbf{W}_c^n is the feature extraction network parameter until n -th convolutional layer and $M(k)$ is the sampled indices of convolutional layers at level k . This is illustrated by the upper of Fig. 2.

Moreover, iteratively extracting the features along our pyramidal model provides evolving receptive fields which is a key ingredient for the geometric invariance [4, 10]. By contrast, existing geometry regression networks [11, 12] face a tradeoff between appearance invariance and localization precision due to the fixed receptive field of extracted features. Note that we obtained I^k with the outputs from the previous level by warping I^{k-1} with \mathbf{T}^{k-1} through bilinear samplers [13] which facilitate an end-to-end learning framework.

Constrained Cost Volume Construction. To estimate geometry between image pairs I^k and I' , the matching cost according to search spaces should be

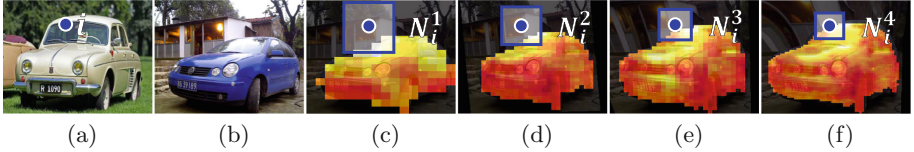


Fig. 3. Visualization of the constrained search window N_i^k : (a) source image and a reference pixel (blue colored). The matching costs are visualized as the heat maps for the reference pixel at (c) level 1, (d) level 2, (e) level 3, and (f) pixel-level. (color figure online)

computed using extracted features \mathbf{F}^k and \mathbf{F}'^k . Unlike conventional approaches that quantize search spaces for estimating depth [6], optical flow [7], or similarity transformations [17], quantizing the 6-D affine transformation defined over an infinite continuous solution space is computationally expensive and also degenerates the estimation accuracy. Instead, inspired by traditional robust geometry estimators such as RANSAC [39] or Hough voting [24], we first construct the cost volume computed with respect to translational motion only, and then determine the affine transformation for each block by passing it through subsequent convolutional layers to reliably prune incorrect matches.

Concretely, the matching costs between extracted features \mathbf{F}^k , \mathbf{F}'^k are computed within a search window as a rectified cosine similarity, such that

$$\mathbf{C}^k(i, j) = \max(0, \mathbf{F}'^k(i) \cdot \mathbf{F}^k(j)), \quad \text{where } j \in N_i^k. \quad (3)$$

A constrained search window N_i^k is centered at pixel i with the radius $r(k)$ as exemplified in Fig. 3. In our pyramidal model, a relatively large radius is used at coarser levels to estimate a rough yet reliable affine transform as a guidance at subsequent finer levels. The radius becomes smaller as the level goes deeper where the regression network is likely to avoid local minima thanks to the guidance of affine transformation fields estimated on the previous level. Thus only reliable matching candidates are provided as an input to the following regression network where even fine-scaled geometric transformations can be estimated at deeper level. The constructed cost volume can be further utilized for generating the supervisions with correspondence consistency check as described in Sect. 3.3.

Grid-Level Regression. The constrained cost volume \mathbf{C}^k is passed through successive CNNs and bilinear upsampling layer to estimate the affine transformation field such that $\mathbf{T}^k = \mathcal{F}(\mathbf{C}^k; \mathbf{W}_g^k)$, where \mathbf{W}_g^k is the grid-level regression network parameter at the level k . Since each level in the pyramid has a simplified task (it only has to estimate residual transformation field), the regression networks can be simple to have 3-6 convolutional layers.

Within the hierarchy of the pyramidal model, our first starts to estimate the transformation from an entire image and then progressively increase the degree of freedom of the transformation by dividing each grid into four rectangular grids, yielding $2^{k-1} \times 2^{k-1}$ grid of affine fields at level k . However, the estimated coarse

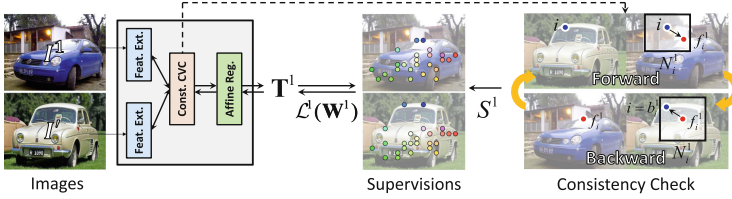


Fig. 4. Training the grid-level module at level 1. By using the correspondence consistency, tentative sparse correspondences are determined and used to train the network.

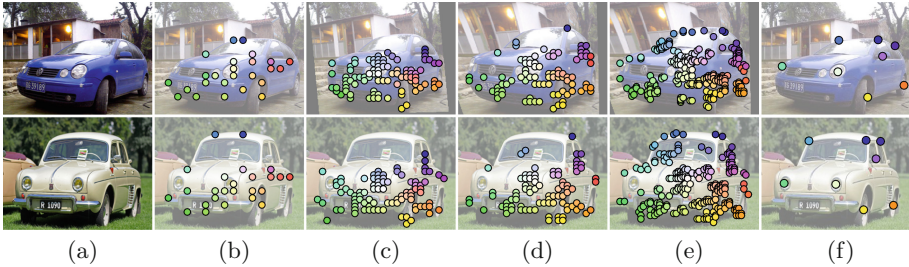


Fig. 5. Visualization of the generated supervisions at each level: (a) source and target images, (b) level 1, (c) level 2, (d) level 3, (e) pixel level, (f) GT keypoints. The tentative positive samples are color-coded. (Best viewed in color.)

affine field has the discontinuities between nearby affine fields occurring blocky artifacts around grid boundaries as shown in (d) and (f) of Fig. 6. To alleviate this, a bilinear upsampler [13] is applied at the end of successive CNNs, upsampling a coarse grid-wise affine field to the original resolution of the input image I . This simple strategy regularizes the affine field to be smooth, suppressing the artifacts considerably as exemplified in Fig. 6.

Note that the composition of the estimated affine fields from level 1 to k can be computed as multiplications of augmented matrix in homogeneous coordinates such that

$$\mathbf{M}(\mathbf{T}_i^{[1,k]}) = \prod_{n \in \{1, \dots, k\}} \mathbf{M}(\mathbf{T}_i^n) \tag{4}$$

where $\mathbf{M}(\mathbf{T})$ represents \mathbf{T} in homogeneous coordinates as $[\mathbf{T}; [0, 0, 1]]$.

Pixel-Level Regression. To improve the matching ability localizing fine-grained object boundaries, we additionally formulate a pixel-level module. Similar to the grid-level modules, it also consists of feature extraction, constrained cost volume construction, and regression network. The main difference is that an encoder-decoder architecture is employed for the regression network, which has been adopted in many pixel-level prediction tasks such as disparity estimation [40], optical flow [41], or semantic segmentation [42]. Taking a warped image I^{K+1} as an input, a constrained cost volume \mathbf{C}^{K+1} is computed and the

Algorithm 1. Pyramidal Affine Regression Network

Input: images I, I'
Output: network parameters $\mathbf{W}_c, \mathbf{W}_g, \mathbf{W}_p$, affine fields \mathbf{T}^*

```

1 :   Compute convolutional activations of target image  $I'$ 
      for  $k = 1 : K$  do
2 :       Compute image  $I^k$  by warping  $I^{k-1}$  with  $\mathbf{T}^{k-1}$  when  $k > 1$ 
3 :       [Only when training] : Initialize affine fields as  $\mathbf{T}_i^k = [\mathbf{I}_{2 \times 2}, \mathbf{0}_{2 \times 1}]$ 
          /* Feature Extraction */
4 :       Compute convolutional activations of  $I^k$  and extract features  $\mathbf{F}^k, \mathbf{F}'^k$ 
          /* Constrained Correlation Volume */
5 :       Construct the constrained cost volume  $\mathbf{C}^k$  with radius  $r(k)$ 
6 :       [Only when training] : Generate supervisions  $S^k$  and train the network
          /* Affine Transformation Field Regression */
7 :       [Only when testing] : Estimate affine fields  $\mathbf{T}^k = \mathcal{F}(\mathbf{C}^k; \mathbf{W}_g^k)$ 
      end for
8 :   Estimate pixel-level affine fields  $\mathbf{T}' = \mathcal{F}(\mathbf{C}^{K+1}; \mathbf{W}_p)$ 
9 :   Compute final affine fields  $\mathbf{M}(\mathbf{T}_i^*) = \prod_{n \in \{1, \dots, K\}} \mathbf{M}(\mathbf{T}^n) \cdot \mathbf{M}(\mathbf{T}'_i)$ 
    
```

pixel-level affine field is regressed through the encoder-decoder network such that $\mathbf{T}' = \mathcal{F}(\mathbf{C}^{K+1}; \mathbf{W}_p)$, where \mathbf{W}_p is the pixel-level regression network parameter. The final affine transformation field between source and target image can be computed as $\mathbf{M}(\mathbf{T}_i^*) = \mathbf{M}(\mathbf{T}_i^{[1, K]}) \cdot \mathbf{M}(\mathbf{T}'_i)$.

3.3 Training

Generating Progressive Supervisions. A major challenge of semantic correspondence with CNNs is the lack of ground-truth correspondence maps for training data. A possible approach is to synthetically generate a set of image pairs transformed by applying random transformation fields to make the pseudo ground-truth [11, 12], but this approach cannot reflect the *realistic* appearance variations and geometric transformations well.

Instead of using synthetically deformed imagery, we propose to generate supervisions directly from the *semantically related* image pairs as shown in Fig. 4, where the correspondence consistency check [35, 48] is applied to the constructed cost volume of each level. Intuitively, the correspondence relation from a source image to a target image should be consistent with that from the target image to the source image. Given the constrained cost volume \mathbf{C}^k , the best match f_i^k is computed by searching the maximum score for each point i , $f_i^k = \operatorname{argmax}_j \mathbf{C}^k(i, j)$. We also compute the backward best match b_i^k for f_i^k such that $b_i^k = \operatorname{argmax}_m \mathbf{C}^k(m, f_i)$ to identify that the best match f_i^k is consistent or not. By running this consistency check along our pyramidal model, we actively collect the tentative positive samples at each level such that $S^k = \{i | i = b_i^k, i \in \Omega\}$. We found that the generated supervisions are qualitatively and quantitatively superior to the sparse ground-truth keypoints as exemplified in Fig. 5.

For the accuracy of supervisions, we limit the correspondence candidate regions using object location priors such as bounding boxes or masks contain-

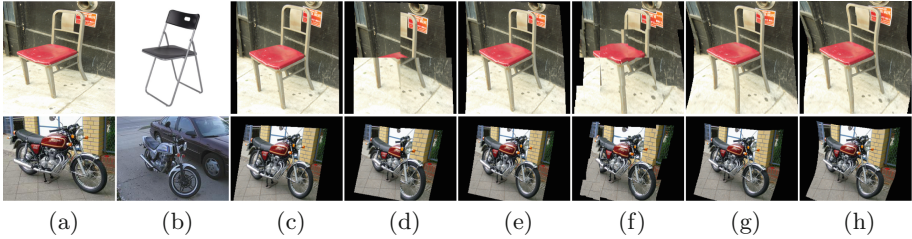


Fig. 6. Qualitative results of the PARN at each level: (a) source image, (b) target image, warping result at (c) level 1, (d) level 2 without upsampling layer, (e) level 2, (f) level 3 without upsampling layer, (g) level 3, and (h) pixel-level.

ing the target object to be matched, which are provided in most benchmarks [21, 43, 44]. Note that our approach is conceptually similar to [8], but we generate the supervisions from the constrained cost volume in a hierarchical manner so that the false positive samples are avoided which is critical to train the geometry regression network.

Loss Function. To train the module at level k , the loss function is defined as a distance between the flows at the positive samples and the flow fields computed by applying estimated affine transformation field such that

$$\mathcal{L}^k(\mathbf{W}^k) = \frac{1}{N} \sum_{i \in S^k} \|\mathbf{T}_i^k \mathbf{i} - (i - f_i^k)\|^2, \quad (5)$$

where \mathbf{W}^k is the parameters of feature extraction network and regression network at level k and N is the number of training samples. Algorithm 1 provides an overall summary of PARN.

4 Experimental Results

4.1 Experimental Settings

For feature extraction networks in each regression module, we used the ImageNet pretrained VGGNet-16 [45] and ResNet-101 [38] with their network parameters. For the grid-level regressions, we used three grid-level modules ($K = 3$), followed by a single pixel-level module. For $M(k)$ in the feature extraction step, we sampled convolutional activations after intermediate pooling layers such as ‘conv5-3’, ‘conv4-3’, and ‘conv3-3’. The radius of search space $r(k)$ is set to the ratio of the whole search space, and decreases as the level goes deeper such that $\{1/10, 1/10, 1/15, 1/15\}$.

In the following, we comprehensively evaluated PARN through comparisons to state-of-the-art dense semantic correspondences, including SIFT Flow [24], DSP [3], and OHG [22]. Furthermore, geometric-invariant methods including PF [30], SCNet [31], CNNGM [12], DCTM [10]. The performance was measured on Taniai benchmark [19], PF-PASCAL dataset [20], and Caltech-101 [21].

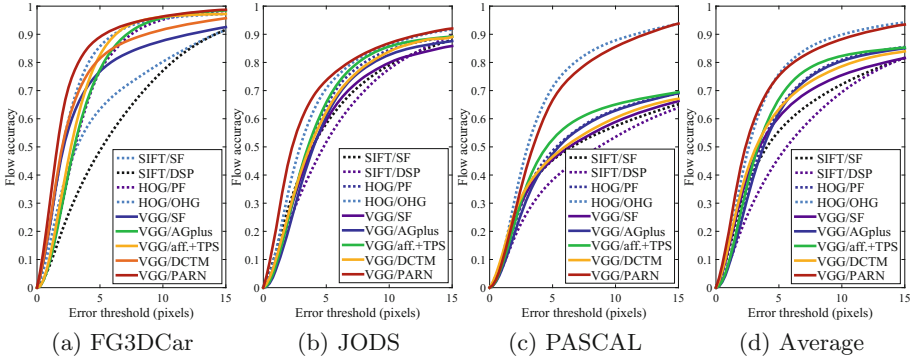


Fig. 7. Average matching accuracy with respect to endpoint error threshold on the Taniai benchmark [19]: (from left to right) FG3DCar, JODS, PASCAL, and average.

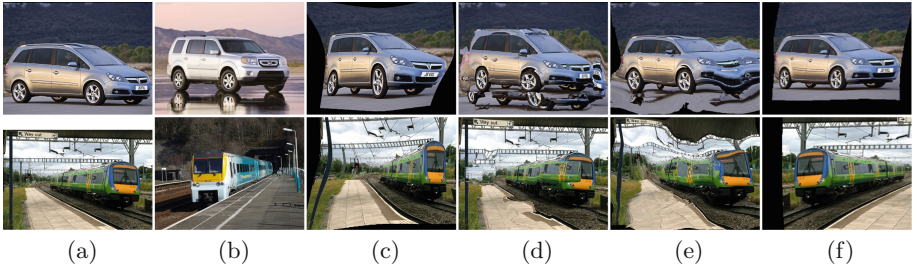


Fig. 8. Qualitative results on the Taniai benchmark [19]: (a) source image, (b) target image, (c) CNNGM-Aff.TPS [12], (d) SCNet-AG+ [31], (e) DCTM [10], (f) PARN. The source images were warped to the target images using correspondences.

4.2 Training Details

For training, we used the PF-PASCAL dataset [20] that consists of 1,351 image pairs selected from PASCAL-berkeley keypoint annotations of 20 object classes. We did not use the ground-truth keypoints at all to learn the network, but we utilized the masks for the accuracy of generated supervisions. We used 800 pairs as a training data, and further divide the rest of PF-PASCAL data into 200 validation pairs and 350 testing pairs. Additionally, we synthetically augment the training pair 10 times by applying randomly generated geometric transformations including horizontal flipping [12]. To generate the most accurate supervisions in the first level, we additionally apply M-estimator Sample and Consensus (MSAC) [46] to build the initial supervisions \mathbf{T}^0 and restrict the search space with the estimated transformation. We sequentially trained the regression modules for 120k iterations each with a batch size of 16 and further finetune all the regression networks in an end-to-end manner [14]. The more details of experimental settings and training are provided in the supplemental material.

Table 1. Matching accuracy compared to state-of-the-art correspondence techniques on the Taniai benchmark [19].

Methods	Descriptor	Matching	FG3D.	JODS	PASC.	Avg.
SF [2]	SIFT	SF	0.632	0.509	0.360	0.500
DSP [3]	SIFT	DSP	0.487	0.465	0.382	0.445
PF [30]	HOG	LOM	0.786	0.653	0.531	0.657
OHG [22]	HOG	OHG	0.875	0.708	0.729	0.771
SCNet [31]	VGG-16	A	0.774	0.574	0.476	0.608
		AG	0.764	0.600	0.463	0.609
		AGplus	0.776	0.608	0.474	0.619
CNNGM [12]	VGG-16	Aff.	0.771	0.662	0.501	0.644
		Aff.+TPS	0.835	0.656	0.527	0.672
DCTM [10]	VGG-16	DCTM	0.790	0.611	0.528	0.630
	Affine-FCSS		0.891	0.721	0.610	0.740
Baseline	VGG-16	SF	0.756	0.490	0.360	0.535
		PARN-Lv1	0.783	0.668	0.641	0.697
		PARN-Lv2	0.837	0.689	0.656	0.739
		PARN-Lv3	0.869	0.707	0.681	0.752
Proposed	VGG-16	PARN	0.876	0.716	0.688	0.760
	ResNet-101		0.895	0.759	0.712	0.788

4.3 Ablation Study

To validate the components within PARN, we additionally evaluated it at each level such as ‘PARN-Lv1’, ‘PARN-Lv2’, and ‘PARN-Lv3’ as shown in Fig. 6 and Table 1. For quantitative evaluations, we used the matching accuracy on the Taniai benchmark [19], which is described in details in the following section. As expected, even though the global transformation was estimated roughly well in the coarsest level (i.e. level 1), the fine-grained matching details cannot be achieved reliably, thus showing the limited performance. However, as the levels go deeper, the localization ability has been improved while maintaining globally estimated transformations.

The performance of the backbone network was also evaluated with a standard SIFT flow optimization [2]. Note that the evaluation of the pixel-level module only in our networks is impracticable, since it requires a pixel-level supervision that does not exist in the current public datasets for semantic correspondence.

4.4 Results

See the supplemental material for more qualitative results.

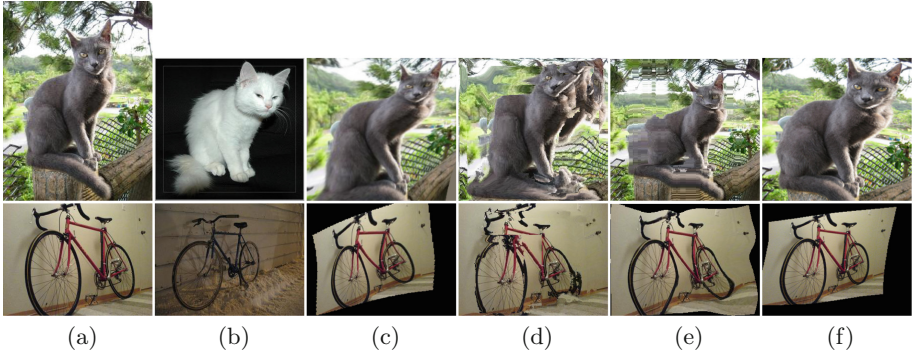


Fig. 9. Qualitative results on the PF-PASCAL benchmark [20]: (a) source image, (b) target image, (c) CNNGM-Aff.+TPS [12], (d) SCNet-AG+ [31], (e) DCTM [10], (f) PARN. The source images were warped to the target images using correspondences.

Taniai Benchmark. We evaluated PARN compared to other state-of-the-art methods on the Taniai benchmark [19], which consists of 400 image pairs divided into three groups: FG3DCar, JODS, and PASCAL. Flow accuracy was measured by computing the proportion of foreground pixels with an absolute flow endpoint error that is smaller than a certain threshold T , after resizing images so that its larger dimension is 100 pixels. Figure 7 shows the flow accuracy with varying error threshold T . Our method outperforms especially when the error threshold is small. This clearly demonstrates the advantage of our hierarchical model in terms of both localization precision and appearance invariance.

Table 1 summarizes the matching accuracy for various dense semantic correspondence techniques at the fixed threshold ($T = 5$ pixels). The quantitative results of ‘PARN-Lv1’ and ‘CNNGM-Aff’ in Table 1 verify the benefits of our weakly supervised training scheme. Whereas ‘CNNGM-Aff.’ is also trained in weakly supervised manner, it relies only on the synthetically deformed image pairs while our method employs semantically sensitive supervisions. Note that we implemented our regression module at level 1 in the same architecture of ‘CNNGM-Aff.’. From the qualitative results of Fig. 8, while DCTM is trapped in local minima unless an appropriate initial solution is given, our method progressively predicts locally-varying affine transformation fields and able to handle relatively large semantic variations including flip variations without handcrafted parameter tuning. The superiority of PARN can be seen by comparing to the correspondence techniques with the same ‘VGG-16’ descriptor in Table 1 and Fig. 7 and even outperforms the supervised learning based method of [31]. We also evaluated with ResNet-101 [38] as a backbone network to demonstrate the performance boosting of our method with more powerful features, where our method achieves the best performance on average.

PF-PASCAL Benchmark. We also evaluated PARN on the testing set of PF-PASCAL benchmark [30]. For the evaluation metric, we used the probability of

Table 2. Matching accuracy compared to state-of-the-art correspondence techniques on the PF-PASCAL benchmark [30] and Caltech-101 dataset [21].

Dataset Methods	PF-PASCAL			Caltech-101		
	PCK			LT-ACC	IoU	LOC-ERR
	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.15$			
SF [2]	0.192	0.334	0.492	0.75	0.48	0.32
DSP [3]	0.198	0.372	0.414	0.77	0.47	0.35
PF [30]	0.235	0.453	0.621	0.78	0.50	0.25
OHG [22]	-	-	-	0.81	0.55	0.19
SCNet [31]	0.260	0.482	0.658	0.79	0.51	0.25
CNNGM [12]	0.254	0.461	0.641	0.80	0.56	0.25
DCTM [10]	0.257	0.477	0.648	0.84	0.53	0.18
PARN	0.268	0.491	0.662	0.87	0.65	0.21

correct keypoint (PCK) between flow-warped keypoints and the ground-truth. The warped keypoints are deemed to be correctly predicted if they lie within $\alpha \cdot \max(h, w)$ pixels of the ground-truth keypoints for $\alpha \in [0, 1]$, where h and w are the height and width of the object bounding box, respectively. Figure 9 shows qualitative results for dense flow estimation.

Without ground-truth annotations, our PARN has shown the outperforming performance compared to other methods in Table 2 where [31] is trained in fully supervised manner. The relatively modest gain may come from the limited evaluation only on the sparsely annotated keypoints of PF-PASCAL benchmark. However, the qualitative results of our method in Fig. 9 indicates that the performance can be significantly boosted when dense annotations are given for evaluation. Although [31] estimates the sparse correspondences in a geometrically plausible model, they compute the final dense semantic flow by linearly interpolating them which may not consider the semantic structures of target image. By contrast, our method leverages a pyramidal model where the smoothness constraint is naturally imposed among semantic scales within deep networks (Fig. 10).

Caltech-101 Dataset. Our evaluations also include the Caltech-101 dataset [21]. Following the experimental protocol in [21], we randomly selected 15 pairs of images for each object class, and evaluated matching accuracy with three metrics: label transfer accuracy (LT-ACC), the IoU metric, and the localization error (LOC-ERR) of corresponding pixel positions. Note that compared to other benchmarks described above, the Caltech-101 dataset provides image pairs from more diverse classes, enabling us to evaluate our method under more general correspondence settings. For the results, our PARN clearly outperforms the semantic correspondence techniques in terms of LT-ACC and IoU metrics. Table 2 summarizes the matching accuracy compared to state-of-the-art methods.

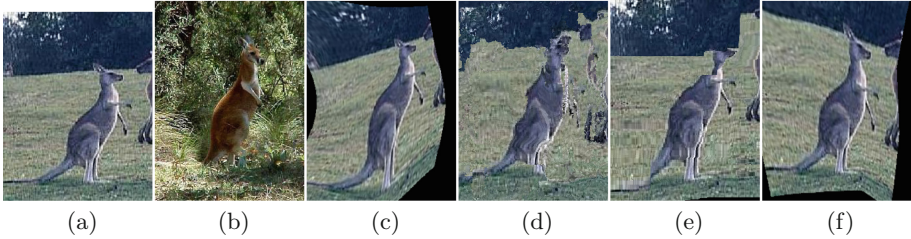


Fig. 10. Qualitative result on the Caltech-101 benchmark [21]: (a) source image, (b) target image, (c) CNGM-Aff.+TPS [12], (d) SCNet-AG+ [31], (e) DCTM [10], (f) PARN. The source image was warped to the target images using correspondences.

5 Conclusion

We presented a novel CNN architecture, called PARN, which estimates locally-varying affine transformation fields across semantically similar images. Our method defined on pyramidal model first estimates a global affine transformation over an entire image and then progressively increases the transformation flexibility. In contrast to previous CNN based methods for geometric field estimations, our method yields locally-varying affine transformation fields that lie in the continuous solution space. Moreover, our network was trained in a weakly-supervised manner, using correspondence consistency within object bounding boxes in the training image pairs. We believe PARN can potentially benefit instance-level object detection and segmentation, thanks to its robustness to severe geometric variations.

Acknowledgement. This research was supported by Next-Generation Information Computing Development Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Science, ICT (NRF-2017M3C4A7069370).

References

1. HaCohen, Y., Shechtman, E., Goldman, D.B., Lischinski, D.: Non-rigid dense correspondence with applications for image enhancement. *ACM Trans. Graph. (TOG)* **30**(4), 70 (2011)
2. Liu, C., Yuen, J., Torralba, A.: Sift flow: dense correspondence across scenes and its applications. *IEEE Trans. PAMI* **33**(5), 815–830 (2011)
3. Kim, J., Liu, C., Sha, F., Grauman, K.: Deformable spatial pyramid matching for fast dense correspondences. In: *CVPR* (2013)
4. Yang, H., Lin, W.Y., Lu, J.: Daisy filter flow: a generalized discrete approach to dense correspondences. In: *CVPR* (2014)
5. Zhou, T., Lee, Y.J., Yu, S.X., Efros, A.A.: Flowweb: joint image set alignment by weaving consistent, pixel-wise correspondences. In: *CVPR* (2015)
6. Scharstein, D., Szeliski, R.: A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *IJCV* **47**(1), 7–42 (2002)

7. Butler, D.J., Wulff, J., Stanley, G.B., Black, M.J.: A naturalistic open source movie for optical flow evaluation. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012. LNCS, vol. 7577, pp. 611–625. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33783-3_44
8. Kim, S., Min, D., Ham, B., Jeon, S., Lin, S., Sohn, K.: FCSS: fully convolutional self-similarity for dense semantic correspondence. In: CVPR (2017)
9. Choy, C.B., Gwak, J., Savarese, S., Chandraker, M.: Universal correspondence network. In: NIPS (2016)
10. Kim, S., Min, D., Lin, S., Sohn, K.: DCTM: discrete-continuous transformation matching for semantic flow. In: ICCV (2017)
11. DeTone, D., Malisiewicz, T., Rabinovich, A.: Deep image homography estimation. arXiv preprint [arXiv:1606.03798](https://arxiv.org/abs/1606.03798) (2016)
12. Rocco, I., Arandjelović, R., Sivic, J.: Convolutional neural network architecture for geometric matching. In: CVPR (2017)
13. Jaderberg, M., Simonyan, K., Zisserman, A., Kavukcuoglu, K.: Spatial transformer networks. In: NIPS (2015)
14. Lin, C.H., Lucey, S.: Inverse compositional spatial transformer networks. In: CVPR (2017)
15. Schneider, N., Piewak, F., Stiller, C., Franke, U.: RegNet: multimodal sensor registration using deep neural networks. In: IV (2017)
16. Dai, J., et al.: Deformable convolutional networks. In: ICCV (2017)
17. Hur, J., Lim, H., Park, C., Ahn, S.C.: Generalized deformable spatial pyramid geometry-preserving dense correspondence estimation. In: CVPR (2015)
18. Tola, E., Lepetit, V., Fua, P.: Daisy: an efficient dense descriptor applied to wide-baseline stereo. *IEEE Trans. PAMI* **32**(5), 815–830 (2010)
19. Taniai, T., Sinha, S.N., Sato, Y.: Joint recovery of dense correspondence and cosegmentation in two images. In: CVPR (2016)
20. Ham, B., Cho, M., Schmid, C., Ponce, J.: Proposal flow: semantic correspondences from object proposals. *IEEE Trans. PAMI* **40**, 1711–1725 (2017)
21. Li, F.F., Fergus, R., Perona, P.: One-shot learning of object categories. *IEEE Trans. PAMI* **28**(4), 594–611 (2006)
22. Yang, F., Li, X., Cheng, H., Li, J., Chen, L.: Object-aware dense semantic correspondence. In: CVPR (2017)
23. Bristow, H., Valmadre, J., Lucey, S.: Dense semantic correspondence where every pixel is a classifier. In: ICCV (2015)
24. Lowe, D.: Distinctive image features from scale-invariant keypoints. *IJCV* **60**(2), 91–110 (2004)
25. Zhou, T., Krahenbuhl, P., Aubry, M., Huang, Q., Efros, A.A.: Learning dense correspondence via 3D-guided cycle consistency. In: CVPR (2016)
26. <http://www.shapenet.org/>
27. Novotny, D., Larlus, D., Vedaldi, A.: AnchorNet: a weakly supervised network to learn geometry-sensitive features for semantic matching. In: CVPR (2017)
28. Hassner, T., Mayzels, V., Zelnik-Manor, L.: On sifts and their scales. In: CVPR (2012)
29. Qiu, W., Wang, X., Bai, X., Yuille, A., Tu, Z.: Scale-space sift flow. In: WACV (2014)
30. Ham, B., Cho, M., Schmid, C., Ponce, J.: Proposal flow. In: CVPR (2016)
31. Han, K., et al.: SCNet: learning semantic correspondence. In: ICCV (2017)
32. Li, Y., Min, D., Brown, M.S., Do, M.N., Lu, J.: SPM-BP: Sped-up patchmatch belief propagation for continuous MRFs. In: ICCV (2015)

33. Hosni, A., Rhemann, C., Bleyer, M., Rother, C., Gelautz, M.: Fast cost-volume filtering for visual correspondence and beyond. *IEEE Trans. PAMI* **35**(2), 504–511 (2013)
34. Lu, J., Yang, H., Min, D., Do, M.N.: Patchmatch filter: efficient edge-aware filtering meets randomized search for fast correspondence field estimation. In: *CVPR* (2013)
35. Revaud, J., Weinzaepfel, P., Harchaoui, Z., Schmid, C.: Deepmatcing: Hierarchical deformable dense matching. *IJCV* **120**, 300–323 (2015)
36. Ranjan, A., Black, M.J.: Optical flow estimation using a spatial pyramid network. In: *CVPR* (2017)
37. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: *NIPS* (2012)
38. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: *CVPR* (2016)
39. Fischler, M.A., Bolles, R.C.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM* **24**(6), 381–395 (1981)
40. Godard, C., Mac Aodha, O., Brostow, G.J.: Unsupervised monocular depth estimation with left-right consistency. In: *CVPR* (2017)
41. Fischer, P., et al.: FlowNet: learning optical flow with convolutional networks. In: *ICCV* (2015)
42. Noh, H., Hong, S., Han, B.: Learning deconvolution network for semantic segmentation. In: *ICCV* (2015)
43. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The pascal visual object classes (VOC) challenge. *IJCV* **88**(2), 303–338 (2010)
44. Chen, X., Mottaghi, R., Liu, X., Fidler, S., Urtasum, R., Yuille, A.: Detect what you can: detecting and representing objects using holistic models and body parts. In: *CVPR* (2014)
45. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: *ICLR* (2015)
46. Torr, P.H., Zisserman, A.: MLESAC: a new robust estimator with application to estimating image geometry. *Comput. Vis. Image Underst.* **78**(1), 138–156 (2000)