Dynamic Context Correspondence Network for Semantic Alignment

Shuaiyi Huang, Qiuyue Wang, Songyang Zhang, Shipeng Yan, Xuming He
ShanghaiTech University
{huangshy1, wangqy2, zhangsy1, yanshp, hexm}@shanghaitech.edu.cn

Abstract

Establishing semantic correspondence is a core problem in computer vision and remains challenging due to large intra-class variations and lack of annotated data. In this paper, we aim to incorporate global semantic context in a flexible manner to overcome the limitations of prior work that relies on local semantic representations. To this end, we first propose a context-aware semantic representation that incorporates spatial layout for robust matching against local ambiguities. We then develop a novel dynamic fusion strategy based on attention mechanism to weave the advantages of both local and context features by integrating semantic cues from multiple scales. We instantiate our strategy by designing an end-to-end learnable deep network, named as Dynamic Context Correspondence Network (DCCNet). To train the network, we adopt a multi-auxiliary task loss to improve the efficiency of our weakly-supervised learning procedure. Our approach achieves superior or competitive performance over previous methods on several challenging datasets, including PF-Pascal, PF-Willow, and TSS, demonstrating its effectiveness and generality.

1. Introduction

Estimating dense correspondence across related images is a fundamental task in computer vision [33, 13, 15]. While early works have focused on correspondence between images depicting the same object or scene, semantic alignment aims to find dense correspondence between different objects belonging to the same category [25]. Such semantic correspondence has attracted much attention recently [10, 31, 18] due to its potential use in a broad range of real-world applications such as image editing [7], co-segmentation [35], 3D reconstruction and scene recognition [1, 27]. However, this task remains extremely challenging because of large intra-class variations, viewpoint changes, background clutters and lack of data with dense annotation [30, 31].

There has been tremendous progress in semantic correspondence recently, thanks to learned feature representations based on convolutional neural networks (CNNs) and the adoption of weak supervision strategy in network training [32, 31, 30, 18, 23, 19]. Most existing approaches learn a convolutional feature embedding so that similar image patches are mapped close to each other in the feature space, and use nearest neighbor search or geometric models for correspondence estimation [30, 31, 18, 23]. In order to achieve localization precision and robustness against deformations, such feature representations typically capture local image patterns which are insufficient to encode global semantic cues. Consequently, they are particularly sensitive to large intra-class variations and the presence of repetitive patterns. While recent efforts [19, 32] introduce local neighborhood cues to improve the matching quality, their effectiveness is limited by the local operations and short-range context.

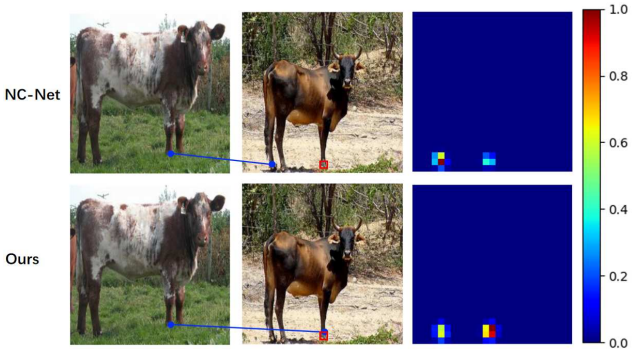

Figure 1. Given a point in the source image (blue dots in column 1), our goal is to match the corresponding point (red squares in column 2) in the target image. The values of correlation score maps (column 3) indicate the likelihood of the corresponding point locating at each location in the target image. Our model (row 2) predicts correspondence with higher precision than the baseline model [32] (row 1), demonstrating its robustness despite the repetitive patterns (blue dots in column 2).

In this work, we aim to address the aforementioned limitations by incorporating global context information and a fusion mechanism that weaves the advantages of both local and spatial features for accurate semantic matching, as shown in Fig. 1. To this end, we first introduce a context-aware semantic representation that integrates appearance features with a self-similarity pattern descriptor, which enables us to capture global semantic context with spatial lay-

out cues. In addition, we propose a pixel-wise attention mechanism that dynamically combines correlation maps derived from local features and context-aware semantic features. The key idea of our approach is to reduce matching ambiguities and to improve localization accuracy simultaneously by the dynamic blending of information from multiple spatial scales.

Concretely, we develop a novel Dynamic Context Correspondence Network (DCCNet), which consists of three main modules: a spatial context network, a correlation network and an attention fusion network. Given an input image pair, we first compute their convolutional (conv) features using a backbone CNN (e.g., ResNet [12]). The conv features are fed into our first module, the spatial context network, which computes the context-aware semantic features that are robust against repetitive patterns and ambiguous matching. Our second module, the correlation network, has two shared branches that generates two correlation score maps for the context-aware semantic and the original conv features respectively. The third module, attention fusion net, predicts a pixel-wise weight mask to fuse two correlation score maps for final correspondence prediction. Our network is fully differentiable and is trained with a weakly-supervised strategy in an end-to-end manner. To improve the training efficiency, we propose a new hybrid loss with multiple auxiliary tasks.

We evaluate our method by extensive experiments on three public benchmarks, including PF-Willow [8], PF-PASCAL [9] and TSS datasets [35]. The experimental results demonstrate the strong performance of our model, which outperforms the prior state-of-the-art approaches in most cases. We also conduct a detailed ablation study to illustrate the benefits of our proposed modules.

The main contributions of this work can be summarized as follows:

- We propose a context-aware semantic representation to generate robust matching against repetitive patterns and local ambiguities in the semantic correspondence problem.
- We develop a novel dynamic fusion strategy based on an attention mechanism to integrate multiple levels of feature representation. To the best of our knowledge, we are the first to adaptively combine context spatial information with local appearance in the semantic correspondence task.
- We design a multi-auxiliary task loss to regularize the training process for weakly-supervised semantic correspondence task and achieve superior or competitive performance on public benchmarks.

2. Related Work

Semantic Correspondence Traditional methods of semantic matching mostly utilize hand-crafted features to find similar image patches with additional spatial smoothness constraints in their alignment models [25, 36, 35]. SIFT Flow [25] extends classical optical flow to establish correspondences across similar scenes using dense SIFT descriptors. Taniai *et al*. [35] adopt HOG descriptors to jointly perform co-segmentation and dense alignment. Due to lack of semantics in feature representations, those approaches often suffer from inaccurate matching when facing large appearance changes from intra-class variations.

Recently, CNNs have been successfully applied to semantic matching thanks to their learned feature representations, which are more robust to appearance or shape variations. Early attempts [28, 19] employ learnable feature descriptors with hand-drafted alignment models, while other approaches [10, 19] requires external modules to generate object proposals for feature extraction, all of which are hence not end-to-end trainable. More recent work tends to use fully trainable network to learn the feature and alignment jointly. Rocco *et al*. [30] proposes a network architecture for geometric matching using a self-supervised strategy from synthetic images, and further improves it with weakly-supervised learning in [31]. The follow-up work extends this strategy in several directions by improving the global transformation model [14], developing cycle-consistency loss [5], estimating locally-varying geometric fields [18, 16], or exploiting neighborhood consensus to produce consistent flow [32]. However, most CNN-based approaches rely on dense matching of conv features, which are incapable of encoding global context [26, 3].

Spatial Context in Correspondence Spatial context has been explored for semantic matching in the literature before deep learning era. Particularly, Irani *et al*. propose the Local Self Similarity (LSS) descriptor [34] to capture self-similarity structure, which has been extended to deep learning based correspondence estimation [21, 22]. More recent work, FCSS [19] and its extension [23], reformulate LSS as a CNN module, computing local self-similarity with learned sparse sampling pattern in object proposals. In contrast, our method exploits a larger spatial context and computes a dense self-similarity descriptor, which is more robust against repetitive patterns and encodes richer context. We also combine this descriptor with local conv features, further improving the discriminative capability of our feature and stabilizing training.

Dynamic Fusion Attention mechanism has been widely used in computer vision tasks to focus on relevant information. For instance, attention-based dynamic fusion is adopted for confidence measure in stereo matching [17]. In semantic segmentation, Chen *et al*. [4] propose an attention mechanism that learns to fuse multi-scale features at each pixel location. In semantic correspondence, recent methods design attention modules for suppressing background

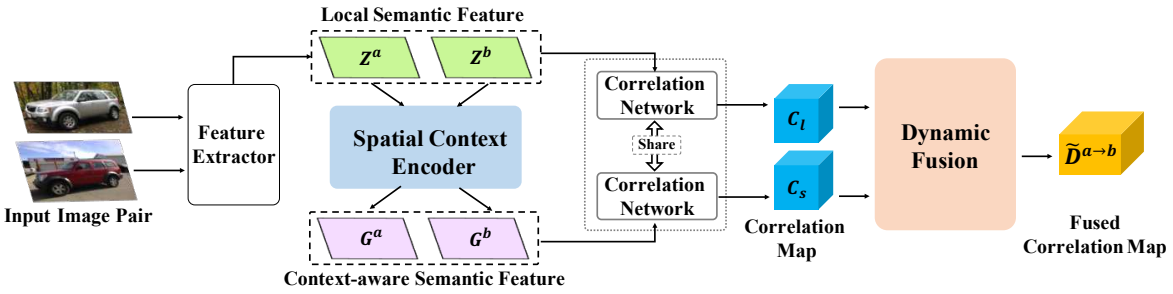

Figure 2. **Overview of DCCNet.** Our proposed DCCNet consists of three main modules: a spatial context encoder, a correlaton network and a dynamic fusion network, which are used to produce a fused correlation map.

regions in images [5, 14]. By contrast, our work addresses the challenge of integrating local and context cues in semantic matching, for which, to the best of our knowledge, dynamic fusion has not been explored before.

3. Method

We now describe our method for estimating a robust and accurate semantic correspondence between two images. Our goal is to seek a flexible feature representation that enables us to capture global semantic contexts as well as informative local features. To this end, we introduce a learnable context-aware semantic representation that augments each local convolutional feature with a global context descriptor. Such a context-aware feature is integrated into the correlation computation by a dynamic fusion mechanism, which combines correlation scores from the context-aware feature and the local conv feature in a selective manner to generate high-quality matching predictions.

Below we start a brief introduction to the semantic correspondence task and an overview of our framework in Sec. 3.1. We then present our proposed context-aware semantic feature and its encoder network in Sec. 3.2, followed by a dynamic fusion module in Sec. 3.4. Finally, we describe our multi-auxiliary task loss in Sec. 3.5.

3.1. Problem Setting and Overview

Given an input image pair $(\mathbf{I}^a, \mathbf{I}^b)$, the goal of semantic alignment is to estimate a dense correspondence between pixels in two images. A common strategy is to infer the correspondence from a correlation map $\mathbf{C}^I$, which describes the matching similarities between any two locations from different images. Formally, let $\mathbf{I}^a \in \mathbb{R}^{3\times h^a\times w^a}$, $\mathbf{I}^b \in \mathbb{R}^{3\times h^b\times w^b}$, where h^a, h^b and w^a, w^b are the height and width of those two images, respectively. The correlation map is denoted as $\mathbf{C}^I \in \mathbb{R}^{h^a\times w^a\times h^b\times w^b}$ and $\mathbf{C}^I_{(i,j,m,n)} = f(I^a_{(i,j)}, I^b_{(m,n)})$ where f is a similarity function. To achieve point-to-point spatial correspondence between two images, we can perform a hard assignment in either of two possible directions, from $\mathbf{I}^a$ to $\mathbf{I}^b$, or vice versa (cf. [32]). Specifically, we have the following mapping from a to b:

$$\begin{aligned} &\mathbf{I}^a_{(i,j)} \text{ correspond to a given } \mathbf{I}^b_{(m,n)} \\ \Leftrightarrow\quad &(i,j) = \underset{1\le i'\le h^a, 1\le j'\le w^a}{\arg\max} \mathbf{C}^I_{(i',j',m,n)} \end{aligned} \tag{1}$$

By doing so, we convert the semantic correspondence problem to a correlation map prediction task, in which our goal is to find a functional mapping from the image pair to an optimal correlation map that generates the accurate pixel-wise correspondences.

A typical deep learning based approach aims to build a high-quality correlation map based on learned feature representation. Formally, we first compute the conv features of the images $\mathbf{I}^a, \mathbf{I}^b$ by an embedding network, which is pretrained on a large dataset (e.g., ImageNet). Denoting the embedding network as $\mathcal{F}$, we generate the image feature maps as follows,

$$\mathbf{Z}^a = \mathcal{F}(\mathbf{I}^a), \quad \mathbf{Z}^b = \mathcal{F}(\mathbf{I}^b), \tag{2}$$

where $\mathbf{Z}^a \in \mathbb{R}^{d\times h^a_f\times w^a_f}$ and $\mathbf{Z}^b \in \mathbb{R}^{d\times h^b_f\times w^b_f}$ are the normalized conv feature representations of the input image pair $(\mathbf{I}^a, \mathbf{I}^b)$, d is the number of feature channel.

Given the conv features, we then build a correlation network that learns a mapping from the feature pair to their correlation map $\mathbf{C} \in \mathbb{R}^{h^a_f\times w^a_f\times h^b_f\times w^b_f}$. Formally,

$$\mathbf{C} = \mathcal{G}(\mathbf{Z}^a, \mathbf{Z}^b; \Theta^{ab}) \tag{3}$$

where $\mathcal{G}$ is the mapping function implemented by the deep network and Θ_{ab} is its parameters. Given a feature-wise correspondence, we can derive the pixel-wise correspondences in Eq. (1) by interpolation on the image plane.

While such deep correspondence networks (e.g. [32]) provide a powerful framework to learn a flexible representation for matching, in practice they are sensitive to large intra-class variations and repetitive patterns in images due to lack of global context. In this work, we propose a novel correspondence network to tackle those challenges in semantic correspondence. Our network is capable of capturing global context of each feature location and dynamically

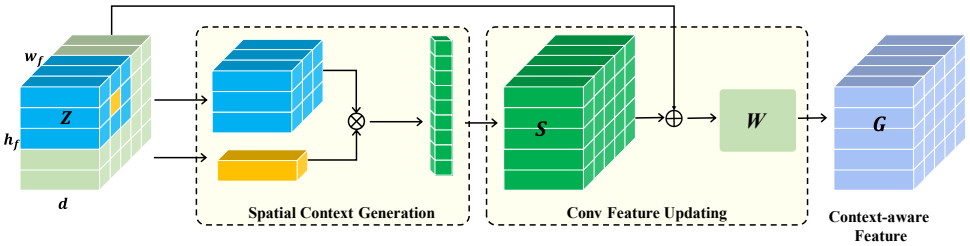

Figure 3. **Overview of Spatial Context Encoder.** Spatial context encoder generates context-aware features from local conv features.

integrating context-aware semantic cues with local semantic information to reduce the matching ambiguities. Hence we refer to our network as Dynamic Context Correspondence Network (DCCNet). Our DCCNet network is composed of three main modules: 1) a *spatial context* encoder, 2) a *correlation* network and 3) a *dynamic fusion* network. Below we will introduce the details of each module and an overview of our network is illustrated in Fig. 2.

3.2. Spatial Context Encoder

Taking as input the conv features of the image pairs, the first component of DCCNet is a spatial context encoder that incorporates global semantic context into the conv feature. To achieve this, we propose a self-similarity based operator to describe the spatial context, as shown in Fig. 3. Specifically, the spatial context encoder consists of two modules: a) *spatial context generation*, b) *context-aware semantic feature generation*, which will be detailed below.

Spatial Context Generation Inspired by LSS [34], we design a novel self-similarity based descriptor on top of deep conv features to encode spatial context at each location in an image. Concretely, given the conv feature map $\mathbf{Z} = \{\mathbf{z}_{(i,j)}\} \in \mathbb{R}^{d\times h_f\times w_f}$ of an image $\mathbf{I}$ (omit superscript here for clarity), we first apply a zero padding of size $(k-1)/2$ (k is odd) on the feature map $\mathbf{Z}$ to get the padded feature map $\widetilde{\mathbf{Z}} \in \mathbb{R}^{d\times(h_f+k-1)\times(w_f+k-1)}$. For location (i,j) in $\mathbf{Z}$, its spatial context descriptor is defined as a self-similarity vector computed between its own local feature $\mathbf{z}_i$ and the features in its neighboring region of size $k\times k$ centered at (i,j). Specifically, we compute the self-similarity features as follows:

$$\mathbf{s}_{(i,j)} = [\mathbf{z}_{i,j}^{\intercal}\tilde{\mathbf{z}}_{(i,j)}, \cdots, \mathbf{z}_{i,j}^{\intercal}\tilde{\mathbf{z}}_{(i+k-1,j+k-1)}], \quad (4)$$

$$\mathbf{S} = \{\mathbf{s}_{(1,1)}, \cdots, \mathbf{s}_{(h_f,w_f)}\}, \quad (5)$$

$$\mathbf{s}_{(i,j)} \in \mathbb{R}^{k^2\times 1}, \quad \mathbf{S} \in \mathbb{R}^{k^2\times h_f\times w_f}, \quad (6)$$

where $\mathbf{s}_{(i,j)}$ is the spatial context descriptor of location (i,j) and $\mathbf{S}$ denotes spatial context of the image I. We refer the neighborhood size k as the kernel size of the context descriptor. With varying kernel sizes, the descriptor is able to encode the spatial context at different scales.

It is worth noting that our spatial context descriptor differs from non-local graph networks [37] in encoding context information, as our descriptor maintains spatial structure, which is important for matching, while graph propagation typically uses aggregation operators to integrate out spatial cues. Our representation also differs from FCSS [19] in several aspects. First, we use a large context to compute self-similarity instead of a local window in order to achieve robustness toward repetitive patterns. Second, FCSS [19] relies on object proposals to remove background while we learn to select informative semantic cues. Moreover, we empirically find that the spatial context descriptor alone is insufficient for high-quality matching, and therefore combine it with local conv features, which will be described below.

Context-aware Semantic Feature The second module of our spatial context encoder computes a context-aware semantic feature for each location on the conv feature map. While the spatial context descriptor encodes second-order statistics in a neighborhood of feature location, it lacks local semantic cues represented by the original conv feature. In order to capture different aspects of semantic objects, we employ a simple fusion step to generate a context-aware semantic representation which provides us better matching quality. Concretely, we apply a non-linear transformation over the concatenation of $\mathbf{Z}$ and $\mathbf{S}$ as below:

$$\mathbf{G}_{(i,j)} = \sigma(\mathbf{W}^{\intercal}[\mathbf{s}_{(i,j)}^{\intercal}, \mathbf{z}_{(i,j)}^{\intercal}]^{\intercal}) \quad (7)$$

$$\mathbf{G} = \{\mathbf{g}_{(1,1)}, \cdots, \mathbf{g}_{(h_f,w_f)}\} \quad (8)$$

$$\mathbf{g}_{(i,j)} \in \mathbb{R}^{l}, \quad \mathbf{G} \in \mathbb{R}^{l\times h_f\times w_f} \quad (9)$$

where σ is a nonlinear function (ReLU) and the weight matrix $\mathbf{W} \in \mathbb{R}^{(d+k^2)\times l}$ transforms the features into l dimensional space. We use $\mathbf{G}$ to denote the context-aware semantic features of image $\mathbf{I}$, and add superscript to represent context-aware semantic feature $\mathbf{G}^a$ and $\mathbf{G}^b$ from the image $\mathbf{I}^a$ and $\mathbf{I}^b$, respectively.

3.3. Correlation Network

The second module of DCCNet is a correlation network that takes in feature representations of an image pair and produces a correlation map. While any correlation computation module can be used here, we adopt the neighborhood consensus module [32] in this work for its superior performance. Specifically, for each type of feature representations

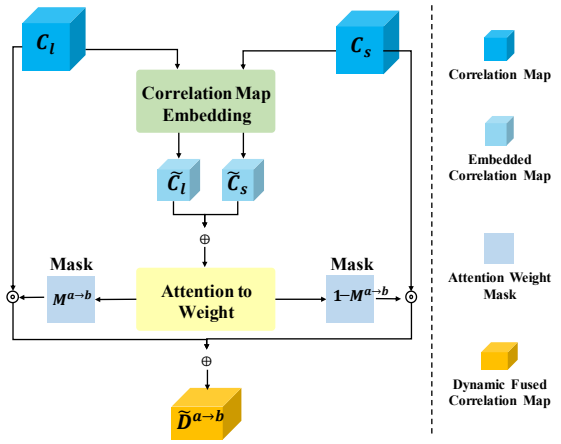

Figure 4. **Overview of Dynamic Fusion Network.** The network employs correlation map embedding and attention-based fusion to combine context and local semantic cues.

of an image pair, say the context-aware semantic feature $(\mathbf{G}^a, \mathbf{G}^b)$ or the local semantic feature $(\mathbf{Z}^a, \mathbf{Z}^b)$, we feed them into the correlation network to generate their corresponding correlation map:

$$\mathbf{C}_l = \mathcal{H}(\mathbf{Z}^a \circledast \mathbf{Z}^b), \quad \mathbf{C}_s = \mathcal{H}(\mathbf{G}^a \circledast \mathbf{G}^b) \tag{10}$$

$$\mathbf{C}_l, \mathbf{C}_s \in \mathbb{R}^{h_f^a \times w_f^a \times h_f^b \times w_f^b} \tag{11}$$

where $\mathcal{H}$ is the neighborhood consensus operator, $\circledast$ is the correlation operation. We use $\mathcal{H}$ to refine the correlation maps based on local neighborhood information. In addition, mutual nearest neighbor consistency constraint [32] is applied before and after $\mathcal{H}$, which is merged into $\mathcal{H}$ for simplicity as it does not contain learnable parameters. We refer the reader to [32] for more details. We now have two correlation maps, $\mathbf{C}_s$ and $\mathbf{C}_l$, that describes the pixelwise correspondence using context-aware semantic cues and local semantic features, respectively.

3.4. Dynamic Fusion Network

While the context-aware semantic feature allows us to encode more global visual patterns, the spatial context encoder in Sec. 3.2 adopts a spatial-invariant fusion mechanism (i.e., a global embedding) to combine local cues and spatial context, which turns out to be sub-optimal for feature locations with distracting neighboring region. An effective solution is to introduce a spatially varying fusion mechanism to balance the context and local conv features specifically for each location. To that end, we propose a dynamic fusion strategy to achieve adaptive fusion for different locations in each image pair. Our fusion utilizes scores from two correlation maps computed in Sec. 3.3 for each location and determines which one is more trustworthy using a location-specific weight.

Specifically, given two correlation maps, $\mathbf{C}_s$ and $\mathbf{C}_l$, we introduce the third module of DCCNet, a dynamic fusion network, to integrate two correlation scores. Motivated by [4], we exploit an attention mechanism to generate a location-aware weight mask for correlation map fusion. The attention-based dynamic fusion consists of the following two modules: 1) *correlation map embedding*, 2) *attention-based fusion*, which will be described below.

Our dynamic fusion strategy is associated with the matching direction. Here we describe the dynamic fusion in the direction from image $\mathbf{I}^a$ to image $\mathbf{I}^b$ for clarity, as the other direction is similar, as shown in Fig. 4.

Correlation Map Embedding In order to predict the attention mask, we first compute a feature representation from the correlation maps. Concretely, we apply an embedding function $\mathcal{E}$ to produce a correlation map embedding:

$$\tilde{\mathbf{C}}_s = \sigma(\mathcal{E}(\mathbf{C}_s; \theta_\mathcal{E})), \quad \tilde{\mathbf{C}}_l = \sigma(\mathcal{E}(\mathbf{C}_l; \theta_\mathcal{E})) \tag{12}$$

where $\mathcal{E}$ is implemented by 4D convolutional neural network, and $\theta_\mathcal{E}$ is the learnable parameter of $\mathcal{E}$. $\tilde{\mathbf{C}}_s, \tilde{\mathbf{C}}_l$ are at the same dimension with $\mathbf{C}_s, \mathbf{C}_l$, in $\mathbb{R}^{h_f^a \times w_f^a \times h_f^b \times w_f^b}$. By this module, we extract those 4D correlation features $\tilde{\mathbf{C}}_l$, $\tilde{\mathbf{C}}_s$, before reshaping them in the next attention module that produces the weight mask and fusion result.

Attention-based Fusion To compute the attention weight mask, we first reshape $\tilde{\mathbf{C}}_s, \tilde{\mathbf{C}}_l$ into a tensor form $\mathbf{D}_l \in \mathbb{R}^{N_b \times h_f^a \times w_f^a}$ and $\mathbf{D}_s \in \mathbb{R}^{N_b \times h_f^a \times w_f^a}$, where $N_b = h_f^b \times w_f^b$. We then compute a fusion weight map for each image pair, which indicates whether the local conv feature is more informative than the context-aware semantic feature for each location. For the direction of image $\mathbf{I}^a$ to $\mathbf{I}^b$, we stack the reshaped correlation maps $\mathbf{D}_l$ and $\mathbf{D}_s$ along the first axis followed by an attention network to predict the fusion weights:

$$\mathbf{D}^{a\to b} = \mathbf{D}_l \oplus \mathbf{D}_s, \quad \mathbf{D} \in \mathbb{R}^{(2N_b) \times h_f^a \times w_f^a} \tag{13}$$

$$\mathbf{M}^{a\to b} = \mathcal{M}(\mathbf{D}^{a\to b}), \quad \mathbf{M}^{a\to b} \in \mathbb{R}^{1 \times h_f^a \times w_f^a} \tag{14}$$

where $\oplus$ is concatenation operator along the first dimension, and $\mathbf{M}^{a\to b}$ is the attention weight mask for $\tilde{\mathbf{C}}_l$. The attention network $\mathcal{M}(\cdot)$ is implemented by a fully convolution layer followed by a softmax operator to normalize the attention weights. Given the attention mask, we fuse the correlation maps in an adaptive way as follows,

$$\tilde{\mathbf{D}}^{a\to b} = \mathbf{D}_l \circ \mathbf{M}^{a\to b} + \mathbf{D}_s \circ (1 - \mathbf{M}^{a\to b}) \tag{15}$$

$$\tilde{\mathbf{D}}^{a\to b} \in \mathbb{R}^{N_b \times h_f^a \times w_f^a} \tag{16}$$

where $\circ$ is the element-wise multiplication with broadcasting for producing the weighted correlation maps. The output correlation $\tilde{\mathbf{C}}^{a\to b}$ is generated by reshaping $\tilde{\mathbf{D}}^{a\to b}$ into the 4D form $\mathbb{R}^{h_f^a \times w_f^a \times h_f^b \times w_f^b}$. Similarly, the adaptively fused correlation $\tilde{\mathbf{C}}^{b\to a} \in \mathbb{R}^{h_f^a \times w_f^a \times h_f^b \times w_f^b}$ from the other direction can also be computed by this module. Finally, those two refined correlation map $\tilde{\mathbf{C}}^{a\to b}$ and $\tilde{\mathbf{C}}^{b\to a}$ are used to find semantic correspondence (cf. [32]).

Methods	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	d.table	dog	horse	moto	person	plant	sheep	sofa	train	tv	all
HOG+PF-LOM[8]	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[6]	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
VGG-16+SCNet-A[11]	67.6	72.9	69.3	59.7	74.5	72.7	73.2	59.5	51.4	78.2	39.4	50.1	67.0	62.1	69.3	68.5	78.2	63.3	57.7	59.8	66.3
VGG-16+SCNet-AG[11]	83.9	81.4	70.6	62.5	60.6	81.3	81.2	59.5	53.1	81.2	**62.0**	58.7	65.5	73.3	51.2	58.3	60.0	69.3	61.5	80.0	69.7
VGG-16+SCNet-AG+[11]	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
VGG-16+CNNGeo[30]	79.5	80.9	69.9	61.1	57.8	77.1	84.4	55.5	48.1	83.3	37.0	54.1	58.2	70.7	51.4	41.4	60.0	44.3	55.3	30.0	62.6
ResNet-101+CNNGeo(S)[30]	82.4	80.9	85.9	47.2	57.8	83.1	92.8	86.9	43.8	91.7	28.1	76.4	70.2	76.6	68.9	65.7	80.0	50.1	46.3	60.6	71.9
ResNet-101+CNNGeo(W)[31]	83.7	88.0	83.4	58.3	68.8	**90.3**	92.3	83.7	47.4	91.7	28.1	76.3	**77.0**	76.0	71.4	**76.2**	80.0	59.5	62.3	63.9	75.8
RTN [18]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	75.9
NC-Net[32]	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
Our Method	**87.3**	**88.6**	82.0	**66.7**	**84.4**	89.6	**94.0**	**90.5**	**64.4**	91.7	51.6	**84.2**	74.3	**83.5**	**72.5**	72.9	60.0	68.3	**81.8**	81.1	**82.3**

Table 1. **Performance on the PF-Pascal dataset [9]**. Per-class and overall PCK are shown in the table and the best results are in bold.

3.5. Learning with Multi-auxiliary Task Loss

We learn the model parameters of our DDCNet in a weakly-supervised manner from a set of matched images. Given two images $\mathbf{I}^a$ and $\mathbf{I}^b$, the outputs of our model are $\tilde{\mathbf{C}}^{a\to b}$ and $\tilde{\mathbf{C}}^{b\to a}$. We first adopt the weakly-supervised training loss proposed in NC-Net [32], which has a functional form :

$$\mathcal{L}(\tilde{\mathbf{C}}^{b\to a}, \tilde{\mathbf{C}}^{a\to b}, y) = -y\left(\overline{s}^a + \overline{s}^b\right) \tag{17}$$

where y denotes the groundtruth label of the image pair $\left(I^a, I^b\right)$ with $y = +1$ for positive matching, and $y = -1$ for negative. $\overline{s}^a$ and $\overline{s}^b$ are the mean matching scores over all hard assigned matches of a given image pair $\left(I^a, I^b\right)$ in both matching directions. To minimize this loss, the model should maximizes the scores of positive and minimizes the scores of negative matching pairs, respectively. We denote this loss term as $\mathcal{L}_{fuse}(\mathbf{I}^a, \mathbf{I}^b)$.

To learn an effective dynamic fusion strategy, we further introduce additional supervision from two auxiliary tasks. Specifically, we also use the correlation map $\mathbf{C}_l$ of local semantic feature and the correlation map $\mathbf{C}_s$ of context-aware semantic feature to generate the matching results, and denote their correspondence losses as $\mathcal{L}_{local}$ and $\mathcal{L}_{context}$, respectively. Here we compute the auxiliary task losses $\mathcal{L}_{local}$ and $\mathcal{L}_{context}$ following the same procedure as in $\mathcal{L}_{fuse}$. The overall training loss is then defined as,

$$\mathcal{L}(\mathbf{I}^a, \mathbf{I}^b) = \mathcal{L}_{fuse} + \lambda\mathcal{L}_{local} + \gamma\mathcal{L}_{context} \tag{18}$$

where λ and γ are the hyper-parameter to balance the main and auxiliary task losses.

4. Experiments

We evaluate our DCCNet on the weakly-supervised semantic correspondence task by conducting a series of experiments on three public datasets, including PF-PASCAL [9], PF-WILLOW [8] and TSS [35]. In this section, we introduce our experiment settings and report evaluation results in detail. We first describe the implementation details in Sec.4.1, followed by the quantitative results of the three datasets in Sec.4.2, Sec.4.3 and Sec.4.4, respectively. Finally, ablation study and comprehensive analysis are provided in Sec.4.5.

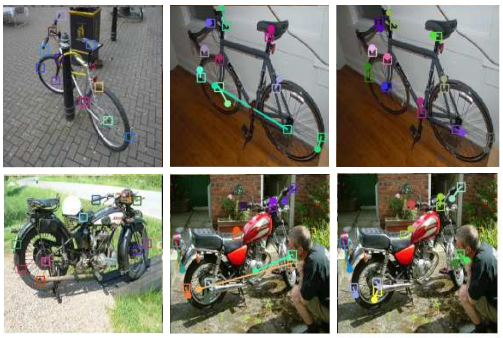

Figure 5. **Qualitative comparisons on the PF-PASCAL benchmark [9].** The leftmost column shows source images. The second and third columns show predictions from Nc-Net [32] and our proposed DCCNet respectively. We show the ground truth keypoints in squares and the predicted keypoints in dots, with their distance in target images depicting the matching error. It is clear that our model is robust to repetitive patterns.

4.1. Implementation details

We implement our DCCNet with the PyTorch framework [29]. For the feature extractor, we use the ResNet-101 [12] pre-trained on ImageNet with the parameters fixed and truncated at the conv4_23 layer. The spatial context encoder adopts a kernel size $k = 25$ and the output dimension l of the context-aware semantic features is set to 1024, which are determined by validation. For the correlation network, we follow [32] and stack three 4D convolutional layers with the kernel size at 5×5×5×5 and set the channel number of the intermediate layer to be 16. For the dynamic fusion net, we choose the same 4D conv layers as in the correlation network for the correlation embedding module, and the attention mask prediction layer is implemented with a 1×1 conv layer.

To train the model, we set λ and γ in the multi-auxiliary task loss to 1 by validation. The model parameters are randomly initialized except for feature extractor. The model is trained for 5 epochs on 4 GPUs with early stopping to avoid overfitting. We use Adam optimizer [24] with a learning rate of 5×10^{-4}.

Images of all three datasets are first resized into the size of 400×400. Our model is trained on the PF-PASCAL benchmark [9]. To further validate generalization capac-

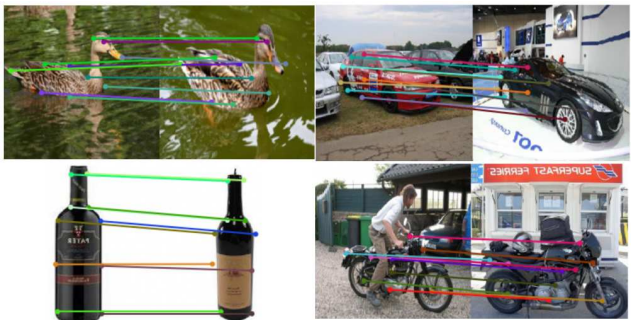

Figure 6. **Semantic alignment examples on PF-WILLOW.** Our model can produce reasonable matching results despite large background clutters and viewpoint changes.

ity of our model, we test the trained model with the PF-WILLOW dataset [8] and the TSS dataset [8] without any further finetuning. Finally, we conduct the ablation study on the PF-PASCAL dataset [9].

4.2. PF-Pascal Benchmark

Dataset and Evaluation Metric The PF-PASCAL [9] benchmark is built from the PASCAL 2011 keypoint annotation dataset [2], which consists of 20 object categories. Following the dataset split in [11], we partition the total 1351 image pairs into a training set of 735 pairs, validation set of 308 pairs and test set of 308 pairs, respectively. The model learning is performed in a weakly-supervised manner where keypoint annotations are not used for training but for evaluation only. We report the percentage of the correct keypoints (PCK) metric [39] which measures the percentage of keypoints whose transfer errors below a given threshold. In line with previous work, we report PCK ($\alpha = 0.1$) w.r.t. image size.

Experimental Results As shown in Table 1, we compare our proposed method with previous methods including NC-Net [32], WeakAlign [31], RTN [18], CNNGeo [30], Proposal Flow [9], UCN [6] and different versions of SC-Net [11]. Our approach achieves an overall PCK of 82.3% , outperforming the prior state of the art [32] by 3.4%.

Visualization Results Fig. 5 shows qualitative comparisons with Nc-Net [32]. We can see that our model is robust against repetitive patterns thanks to our proposed context-aware semantic representation and dynamic fusion. More qualitative results can be found in the suppl. material.

4.3. PF-WILLOW Benchmark

Dataset and evaluation metric The PF-WILLOW dataset consists of 900 image pairs selected from a total of 100 images [8]. We report the PCK scores with multiple thresholds ($\alpha = 0.05, 0.10, 0.15$) w.r.t. bounding box size in order to compare with prior methods.

Methods	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.15$
HOG+PF-LOM [9]	28.4	56.8	68.2
DCTM [23]	38.1	61.0	72.1
UCN-ST [6]	24.1	54.0	66.5
CAT-FCSS [20]	36.2	54.6	69.2
SCNet [11]	38.6	70.4	85.3
ResNet-101+CNNGeo [30]	36.9	69.2	77.8
ResNet-101+CNNGeo(W) [31]	38.2	71.2	85.8
RTN [18]	41.3	71.9	86.2
NC-Net [32]	**44.0**	72.7	85.4
Our Method	43.6	**73.8**	**86.5**

Table 2. **Evaluation results on PF-WILLOW [8].** We report the PCK scores with three thresholds and the best results are in bold.

Figure 7. **Qualitative results on the TSS benchmark [35].** The first column depicts source image and target image respectively. From the second column to the last column is results from WeakAlign [31], NC-Net [32] and our model respectively.

Experimental Results Table 2 compares the PCK accuracies of our DCCNet to those of the state-of-the-art semantic correspondence techniques. Our proposed method improves the PCK accuracies over the previously published best performance by 1.1% when $\alpha = 0.10$ and $\alpha = 0.15$. Our model also achieves a competitive PCK ($\alpha = 0.05$) of 43.6% which is merely 0.4% lower than the state-of-the-art result, partially due to the large scale variation in this dataset unseen in the training. Fig. 6 shows qualitative results on the PF-WILLOW dataset, which further demonstrate the strength of our method.

4.4. TSS Benchmark

Dataset and evaluation metric The TSS dataset contains 400 image pairs in total, divided into three groups, including FG3DCAR, JODS, and PASCAL. Ground truth flows and foreground masks for image pair are provided, where we only use it for evaluation in the weak supervision setting. Following Taniai *et al*. [35], we report the PCK over foreground object by setting α to 0.05 w.r.t. image size.

Experimental Results Table 3 presents quantitative results on the TSS benchmark. We observe that our method outperforms previous methods on one of the three groups of the TSS dataset and our average performance over three groups on the TSS dataset achieves new state of the art. This shows our method can generalize to novel datasets despite the moderate change of data distribution. Qualitative results are presented in Fig. 7.

Methods	FG3D.	JODS	PASC.	avg.
HOG+PF-LOM [9]	78.6	65.3	53.1	65.7
HOG+TSS [35]	83.0	59.5	48.3	63.6
FCSS+SIFT Flow [19]	83.0	65.6	49.4	66.0
FCSS+PF-LOM [19]	83.9	63.5	58.2	68.5
HOG+OADSC [38]	87.5	70.8	72.9	77.1
FCSS+DCTM [23]	89.1	72.1	61.0	74.0
VGG-16+CNNGeo [30]	83.9	65.8	52.8	67.5
ResNet-101+CNNGeo(S) [30]	83.9	76.4	56.3	74.3
ResNet-101+CNNGeo(W) [31]	90.3	76.4	56.5	74.4
RTN [18]	90.1	78.2	**63.3**	77.2
NC-Net [32]	**94.5**	81.4	57.1	77.7
Our Method	93.5	**82.6**	57.6	**77.9**

Table 3. **Evaluation results on the TSS dataset [35].** We report the PCK scores with $\alpha = 0.05$ and the best results are in bold.

4.5. Ablation Study

To understand the effectiveness of our model components, we conduct a series of ablation studies focusing on: 1) *effects of individual modules*, 2) *kernel sizes in spatial context*, 3) *different fusion methods* and 4) *multi-auxiliary task losses*. We select NC-Net [32] as our baseline and report PCK ($\alpha = 0.1$) on the PF-PASCAL [9] test split.

Effects of Individual Modules We consider five different ablation settings and the overall results are shown in Table 4. First, we note that applying our proposed spatial context encoder (Baseline+S) generates large performance improvement (2.0%) over NC-Net [32]. Second, adding dynamic fusion with auxiliary loss (Baseline+SDA) provides a further boost of 1.4%. Below we introduce detailed analysis for each module via the other three ablation settings.

Spatial Context Encoder Table 5 shows the effects of incorporating context with different kernel sizes. For using our spatial context encoder alone (Baseline+S), the performance increases first and then drops with increasing kernel sizes, which is due to degradation of context-aware features as more background clutters are included. Our dynamic fusion and auxiliary loss (Baseline+SDA) can effectively alleviate the degradation problem.

Fusion method We study the effects of our dynamic fusion by simple average fusion of two correlation maps, referring to the resulting model as Baseline+SAA. From Table 4 we can see that our dynamic fusion model (Baseline+SDA) yields significant better results (82.3%) than average fusion (80.2%), showing the necessity of our attention module. Moreover, Baseline+SAA underperforms the model setting with context-aware semantic feature alone (Baseline+S) due to its global averaging. In contrast, the pixel-wise weight mask from attention net enables each location to adaptively merge different scales of semantic cues. We also evaluate the model setting without correlation map embedding during dynamic fusion (Baseline+SCA), which generates worse results, indicating the efficacy of 4D correlation map features in the dynamic fusion network.

Models	SCE	Fusion	Auxiliary Loss	PCK
NC-Net [32]	-	-	-	78.9
Baseline+S	✓	-	-	80.9
Baseline+SCA	✓	Dynamic w/o Corr Embedding	✓	79.9
Baseline+SAA	✓	Average w/ Corr Embedding	✓	80.2
Baseline+SD	✓	Dynamic w/ Corr Embedding	✗	81.0
Baseline+SDA	✓	Dynamic w/ Corr Embedding	✓	**82.3**

Table 4. **Analysis of individual modules of DCCNet on the PF-PASCAL [9] dataset.** NC-Net [32] is used as our baseline. Our ablation includes whether using spatial context encoder, fusion method adopted, and whether using multi-auxiliary task loss.

Models	Kernel size	PCK
NC-Net [32]	-	78.9
Baseline+S	11	78.9
Baseline+S	25	**80.9**
Baseline+S	31	77.1
Baseline+SDA	25	**82.3**
Baseline+SDA	31	80.7

Table 5. **Effect of kernel sizes in our spatial context on the PF-PASCAL [9] dataset.** NC-Net [32] is used as our baseline.

Multi-auxiliary task loss To validate the effect of our proposed auxiliary task loss, we train a model without two additional loss terms, which is referred to as Baseline+SD. Table 4 shows that our model with auxiliary loss terms (Baseline+SDA) attains 1.3% higher PCK scores than the Baseline+SD model, reaching the state-of-the-art result of 82.3%. This improvement indicates the effectiveness of our multi-auxiliary task loss in regularizing the training process for weakly-supervised semantic correspondence task. With the multi-auxiliary task loss, our local feature and context-aware semantic feature branches have stronger supervision signals, which in turn benefits the fusion branch and produces better overall matching results.

5. Conclusion

In this work, we have proposed an effective deep correspondence network, DCCNet, for the semantic alignment problem. Compared to the prior work, our approach has several innovations in semantic matching. First, we develop a learnable context-aware semantic representation that is robust against repetitive patterns and local ambiguities. In addition, we design a novel dynamic fusion module to adaptively combine semantic cues from multiple spatial scales. Finally, we adopt a multi-auxiliary task loss to better regularize the learning of our dynamic fusion strategy. We demonstrate the efficacy of our approach by extensive experimental evaluations on the PF-PASCAL, PF-WILLOW and TSS datasets. The results evidently show that our DCCNet achieves the superior or comparable performances over the prior state-of-the-art approaches on all three datasets.

Acknowledgments This work was supported in part by the NSFC Grant No.61703195 and the Shanghai NSF Grant No.18ZR1425100.

References

[1] Sameer Agarwal, Yasutaka Furukawa, Noah Snavely, Ian Simon, Brian Curless, Steven M Seitz, and Richard Szeliski. Building rome in a day. *Communications of the ACM*, 54(10):105–112, 2011.

[2] Lubomir Bourdev and Jitendra Malik. Poselets: Body part detectors trained using 3d human pose annotations. In *Proceedings of the International Conference on Computer Vision(ICCV)*, 2009.

[3] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint*, 2017.

[4] Liang-Chieh Chen, Yi Yang, Jiang Wang, Wei Xu, and Alan L Yuille. Attention to scale: Scale-aware semantic image segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, 2016.

[5] Yun-Chun Chen, Po-Hsiang Huang, Li-Yu Yu, Jia-Bin Huang, Ming-Hsuan Yang, and Yen-Yu Lin. Deep semantic matching with foreground detection and cycle-consistency. In *Proceedings of the Asian Conference on Computer Vision(ACCV)*, 2018.

[6] Christopher B Choy, JunYoung Gwak, Silvio Savarese, and Manmohan Chandraker. Universal correspondence network. In *Advances in Neural Information Processing Systems(NeurIPS)*, 2016.

[7] Kevin Dale, Micah K Johnson, Kalyan Sunkavalli, Wojciech Matusik, and Hanspeter Pfister. Image restoration using online photo collections. In *Proceedings of the International Conference on Computer Vision(ICCV)*, 2009.

[8] Bumsub Ham, Minsu Cho, Cordelia Schmid, and Jean Ponce. Proposal flow. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, 2016.

[9] Bumsub Ham, Minsu Cho, Cordelia Schmid, and Jean Ponce. Proposal flow: Semantic correspondences from object proposals. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.

[10] Kai Han, Rafael S Rezende, Bumsub Ham, Kwan-Yee K Wong, Minsu Cho, Cordelia Schmid, and Jean Ponce. Scnet: Learning semantic correspondence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, 2017.

[11] Kai Han, Rafael S. Rezende, Bumsub Ham, Kwan-Yee K. Wong, Minsu Cho, Cordelia Schmid, and Jean Ponce. Scnet: Learning semantic correspondence. In *Proceedings of the International Conference on Computer Vision(ICCV)*, 2017.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, 2016.

[13] Heiko Hirschmuller. Stereo processing by semiglobal matching and mutual information. *IEEE Transactions on pattern analysis and machine intelligence*, 30(2):328–341, 2007.

[14] Paul Hongsuck Seo, Jongmin Lee, Deunsol Jung, Bohyung Han, and Minsu Cho. Attentive semantic alignment with offset-aware correlation kernels. In *Proceedings of the European Conference on Computer Vision(ECCV)*, 2018.

[15] Berthold KP Horn and Brian G Schunck. Determining optical flow. *Artificial intelligence*, 17(1-3):185–203, 1981.

[16] Sangryul Jeon, Seungryong Kim, Dongbo Min, and Kwanghoon Sohn. Parn: Pyramidal affine regression networks for dense semantic correspondence. In *Proceedings of the European Conference on Computer Vision(ECCV)*, 2018.

[17] Sunok Kim, Seungryong Kim, Dongbo Min, and Kwanghoon Sohn. Laf-net: Locally adaptive fusion networks for stereo confidence estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, 2019.

[18] Seungryong Kim, Stephen Lin, SANG RYUL JEON, Dongbo Min, and Kwanghoon Sohn. Recurrent transformer networks for semantic correspondence. In *Advances in Neural Information Processing Systems(NeurIPS)*, 2018.

[19] Seungryong Kim, Dongbo Min, Bumsub Ham, Sangryul Jeon, Stephen Lin, and Kwanghoon Sohn. Fcss: Fully convolutional self-similarity for dense semantic correspondence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, 2017.

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

[21] Seungryong Kim, Dongbo Min, Bumsub Ham, Seungchul Ryu, Minh N Do, and Kwanghoon Sohn. Dasc: Dense adaptive self-correlation descriptor for multi-modal and multi-spectral correspondence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, 2015.

[22] Seungryong Kim, Dongbo Min, Stephen Lin, and Kwanghoon Sohn. Deep self-correlation descriptor for dense cross-modal correspondence. In *Proceedings of the European Conference on Computer Vision(ECCV)*, 2016.

[23] Seungryong Kim, Dongbo Min, Stephen Lin, and Kwanghoon Sohn. Dctm: Discrete-continuous transformation matching for semantic flow. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, 2017.

[24] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations(ICLR)*, 2014.

[25] Ce Liu, Jenny Yuen, and Antonio Torralba. Sift flow: Dense correspondence across scenes and its applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2011.

[26] Wenjie Luo, Yujia Li, Raquel Urtasun, and Richard Zemel. Understanding the effective receptive field in deep convolutional neural networks. In *Advances in Neural Information Processing Systems(NeurIPS)*, 2016.

[27] Ekaterina Nikandrova and Ville Kyrki. Category-based task specific grasping. *Robotics and Autonomous Systems*, 2015.

[28] David Novotny, Diane Larlus, and Andrea Vedaldi. Anchornet: A weakly supervised network to learn geometry-sensitive features for semantic matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, 2017.

[29] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[30] Ignacio Rocco, Relja Arandjelović, and Josef Sivic. Convolutional neural network architecture for geometric matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, 2017.

[31] 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(CVPR)*, 2018.

[32] Ignacio Rocco, Mircea Cimpoi, Relja Arandjelović, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Neighbourhood consensus networks. In *Advances in Neural Information Processing Systems(NeurIPS)*, 2018.

[33] Daniel Scharstein and Richard Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International journal of computer vision*, 47(1-3):7–42, 2002.

[34] Eli Shechtman and Michal Irani. Matching local self-similarities across images and videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, 2007.

[35] Tatsunori Taniai, Sudipta N Sinha, and Yoichi Sato. Joint recovery of dense correspondence and cosegmentation in two images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, 2016.

[36] Engin Tola, Vincent Lepetit, and Pascal Fua. Daisy: An efficient dense descriptor applied to wide-baseline stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010.

[37] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7794–7803, 2018.

[38] Fan Yang, Xin Li, Hong Cheng, Jianping Li, and Leiting Chen. Object-aware dense semantic correspondence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, 2017.

[39] Yi Yang and Deva Ramanan. Articulated human detection with flexible mixtures of parts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013.