Guided Semantic Flow

Sangryul Jeon[1], Dongbo Min[2], Seungryong Kim[3], Jihwan Choe[4], and Kwanghoon Sohn[1(✉)]

[1] Yonsei University, Seoul, South Korea
{cheonjsr,khsohn}@yonsei.ac.kr
[2] Ewha Womans University, Seoul, South Korea
dbmin@ewha.ac.kr
[3] Korea University, Seoul, South Korea
seungryong_kim@korea.ac.kr
[4] Samsung, Suwon, South Korea
jihwan.choe@samsung.com

Abstract. Establishing dense semantic correspondences requires dealing with large geometric variations caused by the unconstrained setting of images. To address such severe matching ambiguities, we introduce a novel approach, called guided semantic flow, based on the key insight that sparse yet reliable matches can effectively capture non-rigid geometric variations, and these confident matches can guide adjacent pixels to have similar solution spaces, reducing the matching ambiguities significantly. We realize this idea with learning-based selection of confident matches from an initial set of all pairwise matching scores and their propagation by a new differentiable upsampling layer based on moving least square concept. We take advantage of the guidance from reliable matches to refine the matching hypotheses through Gaussian parametric model in the subsequent matching pipeline. With the proposed method, state-of-the-art performance is attained on several standard benchmarks for semantic correspondence.

Keywords: Dense semantic correspondence · Matching confidence · Moving least square

1 Introduction

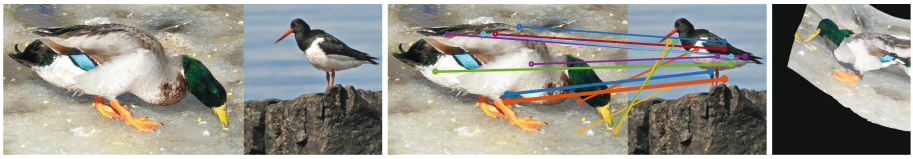
Finding pixel-level correspondences across *semantically* similar images facilitates a variety of computer vision applications, including non-parametric scene parsing [22,30,52], image manipulation [10,26,51], visual localization [41,47], and to name a few.

Classical approaches for dense correspondence take *visually* similar images taken under constraint settings, such as 1D epipolar line for stereo matching [43,

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-030-58604-1_38) contains supplementary material, which is available to authorized users.

A. Vedaldi et al. (Eds.): ECCV 2020, LNCS 12373, pp. 631–648, 2020.
https://doi.org/10.1007/978-3-030-58604-1_38

50] and 2D small motion for optical flow estimation [1,9]. Contrarily, semantic correspondence has no such constraints on the input image pairs except that two images describe the same object or scene category, posing additional challenges due to large appearance and geometric intra-class variations. Recent state-of-the-art methods [17,19,20,23,26,28,39–41,44] have attempted to address these challenges by carefully designing convolutional neural networks (CNNs) that mimic the classical matching pipeline [36]: feature extraction, similarity score computation, and correspondence estimation.



(a) image pair (b) guidance displacements (c) our result

Fig. 1. Visualization of our intuition: (a) image pair, (b) selected confident matches, and (c) warped image using the correspondences from our method. The proposed method, *guided semantic flow*, establishes reliable dense semantic correspondences by leveraging the guidance from confident matches to reduce matching ambiguities.

Since no viewpoint constraint is imposed on the source and target images, the search space for each pixel on the source image have to be defined with all pixels of the target image. However, searching over the full set of pairwise matching candidates inevitably increases the uncertainty in the matching pipeline, especially in the presence of non-rigid deformations and repetitive patterns.

One possible approach to this issue is to design additional modules that can vote for plausible transformation candidates from the full set of pairwise matches [17,39–41,44]. Following the pioneering work of [39], several methods [40,44] attempted to directly regress an image-level global transformation (*e.g.* affine or thin plate spline) between images. However, all matching scores are equally treated regardless of how confident they are, thus these approaches are inherently vulnerable to inaccurate matching scores that are often produced under severe intra-class variations. Without the need of global geometry, some methods [17,41] recently proposed to identify locally consistent matches by analyzing neighborhood consensus patterns. They down-weight ambiguous matches by assessing the confidence of matching scores, but this is performed only with a hand-crafted criterion (*e.g.* mutual consistency) that may often produces high confidence scores even for unconfident pixels.

Alternatively, similar to stereo matching and optical flow estimation [9,50], one can simply discard ambiguous matches by constraining the search space within a predefined local region centered at the querying pixel [20,26], but these approaches disregard the possibility of non-local matches that often appear across the semantically similar images. To address this issue, dilation technique [49] was utilized in [23], but the number of ambiguous matches increases at

the same time. Some methods alleviated this by limiting the search space based on the heuristic matching cues, *e.g.* computing the discrete argmax [28] or starting with an image-level global transformation [19] estimated from a full set of pairwise similarity scores. However, such heuristics are often violated under large intra-class variations where the feature representations are quite inconsistent to measure accurate matching similarity or non-rigid geometric deformations that cannot be modeled with a global transformation model.

In this paper, we propose a novel approach, dubbed as *guided semantic flow*, that reliably infers dense semantic correspondence fields under large intra-class variations, as illustrated in Fig. 1. Our key idea is based on two observations: sparse yet reliable matches can effectively capture non-rigid geometric variations, and these confident matches can guide the adjacent pixels to have similar solution spaces, reducing the matching ambiguities significantly. Our method realizes this idea through three different modules consisting of pruning, propagation, and matching. We first select confident matches from a complete set of pairwise matching candidates through deep networks, and then propagate their reliable information to invalid neighborhoods through a new differentiable upsampling layer inspired by moving least square (MLS) approach [42]. Lastly, dense correspondence fields are reliably inferred from the refined correlation volume by constraining the search space with Gaussian parametric model that is centered at the interpolated displacement vector. Experimental results on various benchmarks demonstrate the effectiveness of the proposed model over the latest methods for dense semantic correspondence.

2 Related Works

Stereo Matching and Optical Flow Estimation. There have been numerous efforts on reducing the matching ambiguitiy for classical dense correspondence problems, *i.e.* stereo matching and optical flow estimation.

Based on the seminal work of PatchMatch [2], the randomized search scheme has been utilized and extended in numerous literature thanks to its effectiveness in pruning the search space [7,15,16]. Another popular idea is to leverage the spatial pyramid of an image, naturally imposing the hierarchical smoothness constraint in a coarse-to-fine manner [5,38,45]. Also, in order to enhance matching scores, recent approaches for depth estimation [35,37] additionally exploit sparse yet reliable measurements retrieved from an external source (*e.g.* LiDAR). However, since these approaches are tailored to the specific problem constraints such as epipolar geometry and relatively small motion, they are not directly applicable to the semantic correspondence task where two images may have large variations in terms of appearance and geometry.

Semantic Correspondence. Most conventional methods for semantic correspondence that use hand-crafted features and regularization terms [22,30,32] have provided limited performance due to a low discriminative power. Recent state-of-the-art approaches have used deep CNNs to extract their features [11,

25,27] and/or spatially regularize correspondence fields in an end-to-end manner [19,23,39,44].

To deal with large geometric deformations, several approaches [17,39–41,44] first computed similarity scores with respect to all possible pairwise matching candidates and then predicted the semantic correspondence through deep networks. As a pioneering work, Rocco et al. [39,40] estimates a global geometric model such as an affine and thin plate spline (TPS) transformation through CNN architecture mimicking the traditional matching pipeline. Seo et al. [44] proposed an offset-aware correlation kernel to put more attention to reliable similarity scores. Without the need of global geometric model, Rocco et al. [41] proposed to identify sets of spatially consistent matches by analyzing neighborhood consensus patterns. Huang et al. [17] extended this architecture by leveraging context-aware semantic representation to further resolve local ambiguities.

Rather than considering all possible matching candidates, some methods [19,20,23,26,28] constrain matching candidates within pre-defined local regions, like stereo matching and optical flow approaches [9,50]. In [20,23,26], locally-varying affine transformation fields are iteratively estimated within locally constrained cost volume. More recently, Lee et al. [28] proposed to leverage a kernel soft argmax function to deal with multi-modal distribution within a correlation volume.

The most relevant method to ours is [19] that utilizes intermediate results from the previous level to constrain the search space of the current level in a coarse-to-fine manner. However, they start with the global affine transformation estimation that often fails to capture reliable matches under large geometric variations with non-rigid transformation.

3 Problem Statement

Let us denote *semantically* similar source and target images as I^s and I^t, respectively. The objective is to establish a two dimensional correspondence field $\tau_i = [u_i, v_i]^T$ between the two images that is defined for each pixel $i = [i_{\mathbf{x}}, i_{\mathbf{y}}]^T$ in I^s.

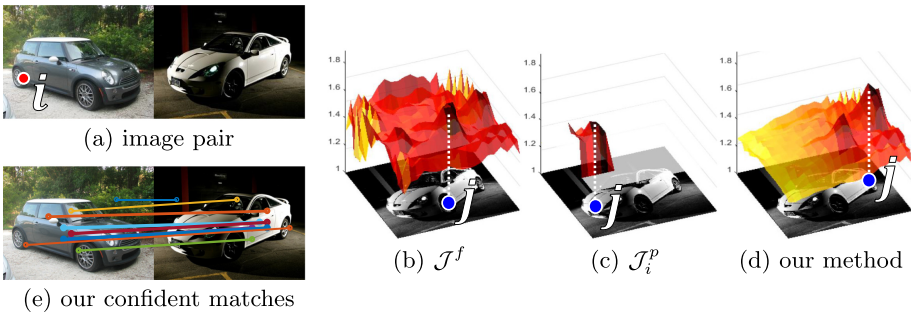
Analogously to the classical matching pipeline [36], this objective involves first extracting dense feature maps from I^s and I^t, denoted by $F^s, F^t \in \mathbb{R}^{h \times w \times d}$ where (h, w) denotes the spatial resolution of the image, and d the dimensionality of feature. Then, given two dense feature maps, a correlation volume C is computed by encoding the similarity as cosine distance:

$$C_{ij}(F^s, F^t) = \langle F_i^s, F_j^t \rangle / \|F_i^s\|_2 \|F_j^t\|_2 \tag{1}$$

where i and j indicate the individual feature position in the source and target images, respectively.

In this stage, several methods [17,39–41,44] construct a full correlation volume C^f considering a set of all possible matching candidates $\mathcal{J}_i^f$, such that

$$\mathcal{J}_i^f = \{j | j_{\mathbf{x}} \in [1, ..., w], j_{\mathbf{y}} \in [1, ..., h]\}. \tag{2}$$



(a) image pair (b) $\mathcal{J}^f$ (c) $\mathcal{J}_i^p$ (d) our method (e) our confident matches

Fig. 2. (a) Given an image pair and a reference pixel i, we visualize its corresponding match ($j = \text{argmax}_l(C_{il})$) and correlation score map (C_{il}), computed with: (b) matching candidates $\mathcal{J}^f$ [17,39–41,44], (c) matching candidates $\mathcal{J}_i^p$ [19,20,23,26,28], and (d) the proposed method. (e) Our key observation is that sparse yet reliable matches can guide the adjacent pixels to have similar solution spaces, reducing matching ambiguities significantly.

Note that $\mathcal{J}_i^f$ is independent to pixel i and identical for all i pixels. However, as exemplified in Fig. 2 (a), the similarity scores in C^f are not guaranteed to be accurate due to inconsistent feature representations under large semantic variations. To address this, several approaches [39,40,44] design an additional module that can vote for the transformation candidates by regressing an image-level single transformation, but they treat the matching scores of all pixels evenly regardless of their confidence. While some methods [17,41] alleviate this by filtering the correlation volume with mutual consistency constraint, they assess the confidences based on a simple criterion such as maximum normalization which may lack the robustness that is attainable with deep CNNs.

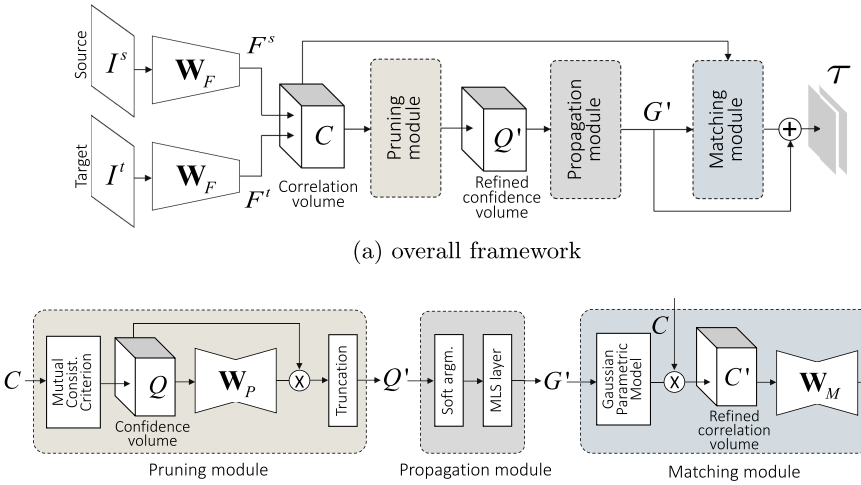
Meanwhile, as shown in Fig. 2(b), some approaches [19,20,23,26,28] construct a partial correlation volume C^p by constraining the search space of each reference pixel i as the restricted local region $\mathcal{N}_k$ centered at the pixel k on the target image. Formally, denoting the pixel k that is dependent on pixel i as $k(i)$, the constrained matching candidates $\mathcal{J}_i^p$ can be defined as

$$\mathcal{J}_i^p = \{j | j \in \mathcal{N}_{k(i)}\}. \tag{3}$$

The center of the local region, $k(i)$, is determined in various ways; as a reference pixel i itself ($k(i) = i$) [20,23,26] or by finding the matching cues from the fully constructed correlation volume through applying the discrete argmax function [28] ($k(i) = \text{argmax}_j(C_{ij}^f)$) or estimating an image-level coarse transformation $\tau^g(C^f)$ [19] ($k(i) = i + \tau_i^g(C^f)$). However, as exemplified in Fig. 2(b), these approaches often fail to constrain the search space correctly under the large intra-class variations where the feature representations between two input images are quite inconsistent to measure accurate matching scores or complex geometric deformations cannot be modeled with a global affine transformation model.

4 Guided Semantic Flow

The proposed method leverages guidance cues from the confident matches to generate reliable likelihood matching hypotheses, as illustrated in Fig. 2(c). Unlike the existing methods that alleviate matching ambiguities with inaccurately assessed matching confidences [17,41] or with the heuristically constrained search spaces [19,20,23,26,28], we address this issue with a learning-based selection of confident matches and their propagation, reducing matching ambiguities significantly while maintaining the robustness to large geometric variations.



(a) overall framework

(b) illustration of each module

Fig. 3. (a) Our overall framework consists of pruning, propagation, and matching modules. (b) The pruning module takes a full correlation volume C as an input and predicts pairwise confidence scores Q' from it by retaining confident matches and rejecting ambiguous ones with the parameters $\mathbf{W}_P$. The propagation module converts this volume Q' into a dense guidance map G' in a fully differentiable manner. The matching module refines the initial correlation volume C with the guidance map G' and then estimates a dense correspondence field τ with the pararmeters $\mathbf{W}_M$.

4.1 Network Architecture

The proposed method consists of three modules as illustrated in Fig. 3: *pruning module* that estimates the confidence probability volume Q', *propagation module* that converts the confidence probability volume into a guidance displacement map G', and *matching module* that refines the initial correlation volume and estimates dense correspondence fields τ from it.

To extract convolutional feature maps of source and target images, the input images are passed through the shared feature extraction networks with parameters $\mathbf{W}_F$ such that $F = \mathcal{F}(I; \mathbf{W}_F)$ where $\mathcal{F}$ denotes a feed-forward operation.

The initial correlation volume C^f is then constructed considering all possible pairwise matching candidates, following (1) and (2), to consider the large intra-class geometric deformations.

Pruning Module. To establish an initial set of confidence probabilities over all pairwise matches, we adopt a differentiable mutual consistency criterion [17,41], such that

$$Q_{ij} = \frac{(C_{ij})^2}{\max_i C_{ij} \cdot \max_j C_{ij}} \tag{4}$$

where Q_{ij} equals one if and only if the match between i and j satisfies the mutual consistency constraint, and becomes smaller than 1 otherwise. Recent works [17,41] utilized this confidence volume Q to filter their similarity scores C (*e.g.* $Q \cdot C$), but the confidence of each pixel is assessed only with the handcrafted criterion as in (4), thus often producing a high confidence score even for an unconfident pixel as exemplified in Fig. 4(a).

In this work, we propose to refine the initial confidence volume with the pruning networks that consist of an encoder-decoder style architecture and a sigmoid function, yielding a value in $(0, 1)$ to suppress false positives, as exemplified in Fig. 4(b). Formally, the refined confidence probability volume Q' can be obtained by

$$Q'_{ij} = T(Q_{ij} \cdot [\mathcal{F}(Q; \mathbf{W}_P)]_{ij}, \rho) \tag{5}$$

where $\mathbf{W}_P$ is the parameters of the pruning networks and $T(\cdot, \rho)$ is a truncation function that discards a probability lower than a threshold ρ to retain only confident matches, such that $T(X, \rho) = X$ if $X > \rho$ and $T(X, \rho) = 0$ otherwise.

It should be noted that several works have also attempted to find the reliable correspondences from the full pairwise similarity scores by thresholding [40], the correspondence consistency [19], or learning with the probabilisitic model [20]. However, these constraints are used in the loss functions only as a supervision for training their deep networks, and are not explicitly used to refine the correlation volume.

Propagation Module. Taking the refined confidence volume Q' as an input, our propagation module first extracts the displacement vectors of the confident matches that can guide nearby ambiguous ones to have similar solution space. Specifically, given a set of the collected confident pixels $\mathcal{S} = \{i | \sum_j Q'_{ij} \neq 0\}$, our propagation module converts the confidence volume Q' into 2-dimensional displacement map G through a soft argmax layer [21], such that

$$G_i = \begin{cases} \sum_j j \cdot \exp(Q'_{ij}) / \sum_l \exp(Q'_{il}) - i, & \text{if } \ i \in \mathcal{S} \\ \text{invalid}, & \text{otherwise.} \end{cases} \tag{6}$$

The displacement map G can then be used to constrain the plausible search range from all possible matching candidates, but this guidance is valid only for confident pixels ($i \in \mathcal{S}$). To guide the search space of the invalid pixels

(a) image pair (b) confident matches in Q (c) confident matches in Q'

Fig. 4. The effectiveness of the pruning networks: (a) matches that satisfy the mutual consistency criterion (*i.e.* $Q_{ij} = 1$), and (b) matches from the refined confidence volume Q' (*i.e.* $Q'_{ij} > \rho$). Our pruning networks effectively suppress the false positive confidence matches that often occur at ambiguous regions.

($i \notin \mathcal{S}$) with the help of confident pixels, we attempted to interpolate the sparse displacement map G using the existing bilinear upsampler of [18]. However, this cannot be directly realized since the confident matches in $\mathcal{S}$ are sparsely and irregularly distributed in the spatial dimension. In this work, we introduce a new differentiable upsampling layer that interpolates the sparse displacement map G into a dense guidance map G'. Concretely, inspired by moving least square approach [42], the displacement vector G'_i at a pixel i can be computed with a spatially-varying weight function w as

$$G'_i = \sum\nolimits_{s \in \mathcal{S}} G_s \cdot w(s-i) / \sum\nolimits_{s \in \mathcal{S}} w(s-i) \tag{7}$$

where $w(z) = \exp(-||z||^2 / 2{c_P}^2)$ is formed with a coefficient c_P. The differentiability of this operator G'_i with respect to G_i can be easily derived similar to [18].

Matching Module. With a favor of densely interpolated guidance displacements G', we refine the initial correlation volume C by maintaining only the similarity scores of highly probable matches. To be specific, we compute the refined correlation volume C' by modulating the original volume C with Gaussian parametric model centered at the guidance displacement vector G':

$$C'_{ij} = \exp(-(j - G'_i)^2 / 2{c_M}^2) \cdot C_{ij} \tag{8}$$

where c_M adjust the distribution of Gaussian model. Unlike the existing methods [19,20,23,26,28] that constrain the search space with simple heuristics, our method leverages the reliable information propagated from the confident matches to effectively deal with large intra-class geometric variations.

With the resulting uni-modal likelihood hypotheses where matching ambiguities are significantly reduced, we subsequently formulate matching networks to regress residual displacements at sub-pixel level, facilitating fine-grained localization. The final dense correspondence field τ is computed as

$$\tau_i = G'_i + [\mathcal{F}(C'; \mathbf{W}_M)]_i \tag{9}$$

where $\mathbf{W}_M$ is the parameters of our matching networks.

4.2 Objective Functions

To overcome the limitation of insufficient training data for semantic correspondence, our matching networks are learned using weak image-level supervision in a form of matching image pairs. Additionally, we expedite the learning process by allowing only the gradients of the foreground pixels to be backpropagated within object masks of the source and target images, similar to [19,23,24,28].

Pruning Networks. To train the pruning networks with the parameter $\mathbf{W}_P$, we define a novel loss function that consists of silhouette consistency loss and geometry consistency loss, such that

$$\mathcal{L}_{\mathrm{P}} = \mathcal{L}_{\mathrm{sil}} + \lambda \mathcal{L}_{\mathrm{geo}} \tag{10}$$

where λ is the weighting parameter.

With the intuition that local structures between source and target image features should be similar at the correct confident correspondences, we encourage the pruning networks to automatically discard the matches that do not satisfy the following local geometry consistency constraint

$$\mathcal{L}_{\mathrm{geo}} = \sum\nolimits_{i \in \mathcal{S}} \sum\nolimits_{l \in \mathcal{N}_i} ||F_l^s - [G' \circ F^t]_l||_F^2 \tag{11}$$

where $\mathcal{N}_i$ is a local window centered at the pixel i, $\circ$ is a warping operator, and $||\cdot||_F^2$ denotes Frobenius norm. By aggregating the contextual information of $\mathcal{N}_i$ through the parameters $\mathbf{W}_P$, we can predict more accurate confidence scores than the handcrafted criterion of (4) that relies only on the pixel-level similarity scores.

Additionally, we formulate the silhouette consistency loss that encourages the refined confidence volume Q' to lie within the silhouette of the initial volume Q:

$$\mathcal{L}_{\mathrm{sil}} = \sum\nolimits_{\{i,j\} \in \mathcal{S}^*} |\log(Q'_{ij}/Q_{ij})| \tag{12}$$

where $\mathcal{S}^* = \{i, j | Q_{ij} > \rho\}$, hence Q'_{ij}/Q_{ij} becomes $[\mathcal{F}(Q; \mathbf{W}_P)]_{ij}$. Note that similar loss function is used in the object landmark detection literature [46] to encourage the landmarks to lie within the silhouette of the object of interest.

Matching Networks. Thanks to the guidance displacements G', most of geometric deformations are already resolved, and thus computing the residual transformation field $\mathcal{F}(C'; \mathbf{W}_M)$ with the weakly-supervised loss function of [23] is tractable, such that

$$\mathcal{L}_{\mathrm{M}} = \sum\nolimits_i -\log(P_i(\tau)) \tag{13}$$

where $P(\tau)$ is the softmax matching probability defined with a local neighborhood $\mathcal{M}_i$ as

$$P_i(\tau) = \frac{\exp(< F_i^s, [\tau \circ F^t]_i >)}{\sum_{l \in \mathcal{M}_i} \exp(< F_i^s, [\tau \circ F^t]_l >)}. \tag{14}$$

This objective allows us to consider both positive and negative samples by maximizing the similarity score at the correct transformation while minimizing the scores of remaining candidates within local neighborhood $\mathcal{M}_i$.

Final Objective Function. We additionally utilize L_1 regularization loss $\mathcal{L}_{\text{sm}}$ for the spatial smoothness in the final correspondence field τ [26,28]. A final objective is defined as a weighted summation of the presented three losses:

$$\mathcal{L}_{\text{final}} = \lambda_{\text{P}}\mathcal{L}_{\text{P}} + \lambda_{\text{M}}\mathcal{L}_{\text{M}} + \lambda_{\text{sm}}\mathcal{L}_{\text{sm}}. \tag{15}$$

4.3 Training Details

Inspired by recent works on finding good matches for wide-baseline stereo [4,34], we first freeze the network parameters $\mathbf{W}_F$, $\mathbf{W}_M$ and learn the pruning networks $\mathbf{W}_P$ only with the gradients from $\mathcal{L}_{\text{P}}$. This allows the pruning networks to be converged stably by fixing the values Q of silhouette consistency loss (12). In second stage, we train the whole networks in an end-to-end manner with $\mathcal{L}_{\text{final}}$ where the properly selected confident matches from the pruning networks boost the convergence of the feature extraction and matching networks by providing well-defined negative samples within the neighborhood $\mathcal{M}_i$ of matching loss (14).

Following [20,26,40], this two-stage learning procedure first utilizes synthetically generated image pairs, by applying random synthetic transformations to a single image of PASCAL VOC 2012 segmentation dataset [8] using the split in [28]. Then, our networks are finetuned with semantically similar image pairs from PF-PASCAL dataset [12] using the split in [40].

5 Experimental Results

5.1 Implementation Details

For feature extraction, we used two CNNs as main backbone networks; ImageNet [6]-pretrained ResNet 101 [14] and PASCAL VOC 2012 [8]-pretrained SFNet [28], where activations are sampled at 'conv4-23' and 'conv5-3'. The activations adapted from 'conv5-3' are upsampled using bilinear interpolation. We denote these backbone networks in the following evaluations as "Ours w/ResNet" and "Ours w/SFNet". We set threshold ρ to 0.9, the variances $\{c_P, c_M\}$ to $\{7, 5\}$, and Referring to the ablation study of [23], the radius of local window $\mathcal{M}_i$ is set to 5. More details about the implementation and the performance analysis with respect to the hyper-parameters are provided in the supplemental material.

5.2 Results

PF-WILLOW and PF-PASCAL Dataset. PF-WILLOW dataset [11] includes 10 object sub-classes with 10 keypoint annotations for each image, providing 900 image pairs. PF-PASCAL dataset [12] contains 1,351 image pairs over

20 object categories with PASCAL keypoint annotations [3]. Following the split in [13,40], we used only 300 testing image pairs for the evaluation. We used a common metric of the percentage of correct keypoint (PCK) by computing the distance between flow-warped keypoints and the ground-truth ones [31]. The warped keypoints are determined to be correct if they lie within $\alpha \cdot \max(h, w)$ pixels from the ground-truth keypoints for $\alpha \in [0, 1]$, where h and w are the height and width of either an image (α_{img}) or an object bounding box (α_{bb}). PCK with α_{bb} is more stringent metric than that of α_{img} [33]. In line with the previous works, we used α_{bb} for PF-WILLOW [11] and α_{img} for PF-PASCAL [12].

Table 1. Matching accuracy compared to state-of-the-art correspondence techniques on PF-WILLOW dataset [11], PF-PASCAL dataset [12], and Caltech-101 dataset [29]. Results of [13,39–41,44] are borrowed from [33].

Methods		PF-PASCAL (PCK@α_{img})			PF-WILLOW (PCK@α_{bb})			Caltech-101	
		$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.15$	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.15$	LT-ACC	IoU
Unsupervised	CNNgeo [39]	41.0	69.5	80.4	36.9	69.2	77.8	0.79	0.56
	A2Net [44]	42.8	70.8	83.3	36.3	68.8	84.4	0.80	0.57
Fully supervised	SCNet [13]	36.2	72.2	82.0	38.6	70.4	85.3	0.79	0.51
	HPF [33]	60.1	84.8	92.7	45.9	74.4	85.6	0.87	0.63
Weakly supervised	CNNinlier [40]	49.0	74.8	84.0	37.0	70.2	79.9	0.85	0.63
	NCNet [41]	54.3	78.9	86.0	33.8	67.0	83.7	0.85	0.60
	RTNs [23]	55.2	75.9	85.2	41.3	71.9	86.2	0.86	0.65
	SFNet [28]	50.0	78.7	88.9	37.5	71.1	88.5	0.88	0.67
	SAMNet [26]	60.1	80.2	86.9	–	–	–	–	–
	DCCNet [17]	–	82.3	–	43.6	73.8	86.5	–	–
	Ours w/ResNet	62.8	84.5	93.7	47.0	75.8	88.9	0.88	**0.69**
	Ours w/SFNet	**65.6**	**87.8**	**95.9**	**49.1**	**78.7**	**90.2**	**0.89**	**0.69**

The average PCK scores are summarized in Table 1 showing that our model ("Ours w/ResNet") exhibits a competitive performance to the latest weakly-supervised and even fully-supervised techniques for semantic correspondence, demonstrating the benefits of generating highly probable hypotheses based on the confident matches. When combined with sophisticate CNN features ("Ours w/SFNet"), the outstanding performance was attained.

Caltech-101 Dataset. We also evaluated our method on Caltech-101 dataset [29] which provides the images of 101 object categories with ground-truth object masks. For the evaluation, we used the 1,515 image pairs used in [13,40], *i.e.* 15 image pairs for each object category. Compared to other datasets described above, the Caltech-101 dataset [29] enable us to evaluate the performances under more general settings with the image pairs from more diverse classes. Following the experimental protocol in [22], the matching accuracy was evaluated with two metrics: the label transfer accuracy (LT-ACC), and the intersection-over-union (IoU) metric.

In Table 1, our method achieves a competitive performance compared to state-of-the-art methods in terms of both LT-ACC and IoU metrics. In particular, our results show better performances with significant margins compared to the methods [39–41,44] that consider all possible matching scores.

This reveals the effectiveness of the proposed pruning and propagation modules where only reliable information is propagated and leveraged to reduce the matching ambiguity.

(a) image pair and our result (b) RTNs [23], NCNet [41], SFNet [28]

Fig. 5. Qualitative results of the semantic alignment on the testing pair of SPair-71k benchmark [33]: (a) input image pairs and warped source images using correspondences obtained from our method, and (b) warped source images from state-of-the-art-methods; (left) RTNs [23], (middle) NCNet [41], (right) SFNet [28].

SPair-71k Benchmark. The evaluation was also performed on the SPair-71k benchmark [33] that includes 70,958 image pairs of 18 object categories from PASCAL 3D+ [48] and PASCAL VOC 2012 [8], providing 12,234 pairs for testing. This benchmark is more challenging than other datasets [11,12,29] for semantic correspondence evaluation, as it covers significantly large variations of 4 factors as shown in Table 2. For the evaluation metric, we used the PCK setting the threshold with respect to the object bounding box to $\alpha_{\text{bb}} = 0.1$.

Table 2. Matching accuracy compared to the state-of-the-art techniques on SPair-71k benchmark [33]. Difficulty levels of viewpoints and scales are labeled 'easy', 'medium', and 'hard', while those of truncation and occlusion are indicated by 'none', 'source', 'target', and 'both'. The performances are evaluated by fixing the levels of other variations as 'easy' and 'none'. Results of [39–41,44] are borrowed from [33].

Methods	Viewpoint			Scale			Truncation				Occlusion				All
	easy	medi	hard	easy	medi	hard	none	src	tgt	both	none	src	tgt	both	
CNNgeo [39]	25.2	10.7	5.9	22.3	16.1	8.5	21.1	12.7	15.6	13.9	20.0	14.9	14.3	12.4	18.1
A2Net [44]	27.5	12.4	6.9	24.1	18.5	10.3	22.9	15.2	17.6	15.7	22.3	16.5	15.2	14.5	20.1
CNNinlier [40]	29.4	12.2	6.9	25.4	19.4	10.3	24.1	16.0	18.5	15.7	23.4	16.7	16.7	14.8	21.1
NCNet [41]	34.0	18.6	12.8	31.7	23.8	14.2	29.1	22.9	23.4	21.0	29.0	21.1	21.8	19.6	26.4
RTNs [23]	34.8	18.2	11.7	33.4	24.7	14.3	30.1	20.9	22.7	20.5	28.8	19.5	20.9	18.8	25.7
HPF [33]	35.6	20.3	15.5	33.0	26.1	15.8	31.0	24.6	24.0	23.7	30.8	23.5	22.8	21.8	28.2
Ours w/ResNet	40.6	22.3	17.8	39.5	30.1	18.7	37.0	28.7	27.1	27.7	36.4	27.8	27.5	23.7	33.5
Ours w/SFNet	**42.1**	**25.7**	**20.1**	**42.3**	**34.0**	**20.8**	**39.8**	**31.1**	**30.0**	**29.9**	**38.8**	**29.3**	**28.3**	**26.9**	**36.1**

Table 2 reports the quantitative performance with respect to different levels of four variation factors. The qualitative results are visualized in Fig. 5. As shown in Table 2 and Fig. 5, our results have shown highly improved performances qualitatively and quantitatively compared to the state-of-the-art techniques on all variation factors. In contrast to the methods [23,28] that cannot capture large geometric variations due to the simple heuristics used to constrain the search space, a large PCK gain for difficult image pairs in Table 2 indicates that our method is effective especially in the presence of severe appearance and shape variations thanks to the guidance by the confident matches learned from all matching candidates. Though the performance was evaluated only on the sparsely annotated keypoints provided from the benchmark, the qualitative results in Fig. 5 indicates that the objective measure can be significantly boosted if dense ground-truth annotations are given for evaluation.

5.3 Ablation Study

Lastly, we conducted an ablation study on different modules and losses in our model of "Ours w/ResNet" evaluating on the testing image pairs of SPair-71k benchmark [33].

Network Architecture. We report the quantitative assessment when one of our modules is removed from the network architecture in Table 3(a) in terms of average PCK at $\alpha_{\text{bb}} = 0.1$. Interestingly, the guidance displacement map G', which is the result obtained with only the pruning and propagation modules, already outperforms state-of-the-art methods by a large margin as shown in Table 2. The performance degradation due to the lack of the pruning or propagation modules highlights the importance of the learning-based selection of confident matches and the MLS layer. Figure 6 shows the intermediate results of our method.

Training Loss. To validate the effectiveness of the utilized losses, we examined the performance of our model when learned with different loss functions. In Table 3(b), the first three rows compare the performances for the variants of the pruning networks. The performance gain from 25.1 to 28.5 with respect to $\mathcal{L}_{\text{geo}}$ indicates the effectiveness of imposing local geometry consistency constraint by aggregating the contextual information. On the other hand, with respect to $\mathcal{L}_{\text{sil}}$, the degraded performance from 28.5 to 24.3 demonstrates the importance of regularizing the refined confidence scores to be similar with the initial ones, so that the retained confident matches also satisfy mutual consistency.

Table 3. Ablation study on the testing pairs of SPair-71k benchmark [33] for (a) different components and (b) different loss functions. Note that, in (a), when the 'MLS layer' in the propagation module is removed, the refined correlation volume C' is computed by applying Gaussian parametric model only on the confident pixels[a].

Pruning ($Q \to Q'$)	MLS layer ($G \to G'$)	Matching ($C, G' \to \tau$)	PCK ($\alpha_{\text{bb}} = 0.1$)
✓	✓	✗	29.3
✓	✗	✓	26.8
✗	✓	✓	25.1
✓	✓	✓	**33.5**

(a) network architecture

$\mathcal{L}_{\text{sil}}$	$\mathcal{L}_{\text{geo}}$	$\mathcal{L}_{\text{M}}$	$\mathcal{L}_{\text{sm}}$	Training stage	PCK
–	✓	–	—	1st	24.3
✓	–	–	–	1st	25.1
✓	✓	–	–	1st	28.5
✓	✓	✓	✓	only 2nd	30.2
✓	✓	✓	✓	1st & 2nd	**33.5**

(b) training loss

[a] $C'_{ij} = \begin{cases} \exp(-(j - G_i)^2 / 2{c_M}^2) \cdot C_{ij}, & \text{if } \ i \in \mathcal{S} \\ C_{ij}, & \text{otherwise.} \end{cases}$

(a) confident matches in Q' (b) matching result with G' (c) matching result with τ

Fig. 6. The visualization of the intermediate results: (a) source and target images, (b) the selected confident matches Q', (c) matching results with the guidance displacements G', and (d) matching results with the final correspondence fields τ.

The last two rows in Table 3(b) reveal the effect of the used two-stage learning process. The performance drop from 33.5 to 30.2 by removing the first stage highlights that the properly selected confident matches from the pruning networks can boost the convergence of our training by allowing only well-defined matching candidates to be utilized during the second stage.

6 Conclusion

We presented a novel framework, guided semantic flow, that reliably infers dense semantic correspondences under large appearance and spatial variations. Taking advantage of the reliable information of confident matches, we effectively handle severe non-rigid geometric deformations and reduce matching ambiguities. The outstanding performance was validated through extensive experiments on various benchmarks.

Acknowledgements. 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 and ICT (NRF2017M3C4A7069370).

References

1. Baker, S., Scharstein, D., Lewis, J., Roth, S., Black, M.J., Szeliski, R.: A database and evaluation methodology for optical flow. Int. J. Comput. Vision **92**(1), 1–31 (2011)
2. Barnes, C., Shechtman, E., Finkelstein, A., Goldman, D.B.: PatchMatch: a randomized correspondence algorithm for structural image editing. ACM Trans. Graph. (ToG) **28**, 24 (2009)
3. Bourdev, L., Malik, J.: Poselets: body part detectors trained using 3D human pose annotations. In: 2009 IEEE 12th International Conference on Computer Vision, pp. 1365–1372. IEEE (2009)
4. Brachmann, E., et al.: DSAC-differentiable RANSAC for camera localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6684–6692 (2017)
5. Chang, J.R., Chen, Y.S.: Pyramid stereo matching network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5410–5418 (2018)
6. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255. IEEE (2009)
7. Duggal, S., Wang, S., Ma, W.C., Hu, R., Urtasun, R.: DeepPruner: learning efficient stereo matching via differentiable PatchMatch. In: The IEEE International Conference on Computer Vision (ICCV), October 2019
8. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (VOC) challenge. Int. J. Comput. Vision **88**(2), 303–338 (2010)
9. Fischer, P., et al.: FlowNet: learning optical flow with convolutional networks. In: ICCV (2015)
10. 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)
11. Ham, B., Cho, M., Schmid, C., Ponce, J.: Proposal flow. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3475–3484 (2016)
12. Ham, B., Cho, M., Schmid, C., Ponce, J.: Proposal flow: semantic correspondences from object proposals. IEEE Trans. PAMI **40**(7), 1711–1725 (2018)

13. Han, K., et al.: SCNet: learning semantic correspondence. In: ICCV (2017)
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
15. Heise, P., Klose, S., Jensen, B., Knoll, A.: PM-Huber: PatchMatch with Huber regularization for stereo matching. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2360–2367 (2013)
16. Hu, Y., Song, R., Li, Y.: Efficient coarse-to-fine PatchMatch for large displacement optical flow. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5704–5712 (2016)
17. Huang, S., Wang, Q., Zhang, S., Yan, S., He, X.: Dynamic context correspondence network for semantic alignment. In: The IEEE International Conference on Computer Vision (ICCV), October 2019
18. Jaderberg, M., Simonyan, K., Zisserman, A., et al.: Spatial transformer networks. In: Advances in Neural Information Processing Systems, pp. 2017–2025 (2015)
19. Jeon, S., Kim, S., Min, D., Sohn, K.: PARN: pyramidal affine regression networks for dense semantic correspondence. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11210, pp. 355–371. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01231-1_22
20. Jeon, S., Min, D., Kim, S., Sohn, K.: Joint learning of semantic alignment and object landmark detection. In: The IEEE International Conference on Computer Vision (ICCV), October 2019
21. Kendall, A., et al.: End-to-end learning of geometry and context for deep stereo regression. In: Proceedings of the International Conference on Computer Vision (ICCV) (2017)
22. Kim, J., Liu, C., Sha, F., Grauman, K.: Deformable spatial pyramid matching for fast dense correspondences. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2307–2314 (2013)
23. Kim, S., Lin, S., Jeon, S., Min, D., Sohn, K.: Recurrent transformer networks for semantic correspondence. In: Advances in Neural Information Processing Systems (2018)
24. Kim, S., Min, D., Ham, B., Jeon, S., Lin, S., Sohn, K.: FCSS: fully convolutional self-similarity for dense semantic correspondence. In: CVPR (2017)
25. Kim, S., Min, D., Ham, B., Lin, S., Sohn, K.: FCSS: fully convolutional self-similarity for dense semantic correspondence. IEEE Trans. Pattern Anal. Mach. Intell. **41**, 581–595 (2018)
26. Kim, S., Min, D., Jeong, S., Kim, S., Jeon, S., Sohn, K.: Semantic attribute matching networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 12339–12348 (2019)
27. Kim, S., Min, D., Lin, S., Sohn, K.: DCTM: discrete-continuous transformation matching for semantic flow. In: ICCV (2017)
28. Lee, J., Kim, D., Ponce, J., Ham, B.: SFNet: learning object-aware semantic correspondence. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2278–2287 (2019)
29. Li, F.F., Fergus, R., Perona, P.: One-shot learning of object categories. IEEE Trans. PAMI **28**(4), 594–611 (2006)
30. Liu, C., Yuen, J., Torralba, A.: SIFT Flow: dense correspondence across scenes and its applications. IEEE Trans. Pattern Anal. Mach. Intell. **33**(5), 978–994 (2010)
31. Long, J.L., Zhang, N., Darrell, T.: Do convnets learn correspondence? In: Advances in Neural Information Processing Systems, pp. 1601–1609 (2014)

32. Lu, J., Li, Y., Yang, H., Min, D., Eng, W., Do, M.N.: PatchMatch filter: edge-aware filtering meets randomized search for visual correspondence. IEEE Trans. PAMI **39**(9), 1866–1879 (2017)
33. Min, J., Lee, J., Ponce, J., Cho, M.: Hyperpixel flow: semantic correspondence with multi-layer neural features. In: The IEEE International Conference on Computer Vision (ICCV), October 2019
34. Moo Yi, K., Trulls, E., Ono, Y., Lepetit, V., Salzmann, M., Fua, P.: Learning to find good correspondences. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2666–2674 (2018)
35. Park, K., Kim, S., Sohn, K.: High-precision depth estimation with the 3D lidar and stereo fusion. In: 2018 IEEE International Conference on Robotics and Automation (ICRA), pp. 2156–2163. IEEE (2018)
36. Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A.: Object retrieval with large vocabularies and fast spatial matching. In: 2007 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8. IEEE (2007)
37. Poggi, M., Pallotti, D., Tosi, F., Mattoccia, S.: Guided stereo matching. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 979–988 (2019)
38. Ranjan, A., Black, M.J.: Optical flow estimation using a spatial pyramid network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4161–4170 (2017)
39. Rocco, I., Arandjelović, R., Sivic, J.: Convolutional neural network architecture for geometric matching. In: CVPR (2017)
40. Rocco, I., Arandjelović, R., Sivic, J.: End-to-end weakly-supervised semantic alignment. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6917–6925 (2018)
41. Rocco, I., Cimpoi, M., Arandjelović, R., Torii, A., Pajdla, T., Sivic, J.: Neighbourhood consensus networks. In: Advances in Neural Information Processing Systems, pp. 1658–1669 (2018)
42. Schaefer, S., McPhail, T., Warren, J.: Image deformation using moving least squares. ACM Trans. Graph. (TOG) **25**, 533–540 (2006)
43. Scharstein, D., Szeliski, R.: A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. Int. J. Comput. Vision **47**(1–3), 7–42 (2002)
44. Seo, P.H., Lee, J., Jung, D., Han, B., Cho, M.: Attentive semantic alignment with offset-aware correlation kernels. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11208, pp. 367–383. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01225-0_22
45. Sun, D., Yang, X., Liu, M.Y., Kautz, J.: PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8934–8943 (2018)
46. Suwajanakorn, S., Snavely, N., Tompson, J.J., Norouzi, M.: Discovery of latent 3D keypoints via end-to-end geometric reasoning. In: Advances in Neural Information Processing Systems, pp. 2059–2070 (2018)
47. Taira, H., et al.: Is this the right place? Geometric-semantic pose verification for indoor visual localization. In: The IEEE International Conference on Computer Vision (ICCV), October 2019
48. Xiang, Y., Mottaghi, R., Savarese, S.: Beyond PASCAL: a benchmark for 3D object detection in the wild. In: IEEE Winter Conference on Applications of Computer Vision, pp. 75–82. IEEE (2014)
49. Yu, F., Koltun, V.: Multi-scale context aggregation by dilated convolutions. In: ICLR (2016)

50. Zbontar, J., LeCun, Y.: Stereo matching by training a convolutional neural network to compare image patches. JMLR **17**(1), 2287–2318 (2016)
51. Zhang, B., et al.: Deep exemplar-based video colorization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8052–8061 (2019)
52. Zhou, T., Jae Lee, Y., Yu, S.X., Efros, A.A.: FlowWeb: joint image set alignment by weaving consistent, pixel-wise correspondences. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1191–1200 (2015)