Probabilistic Warp Consistency for Weakly-Supervised Semantic Correspondences

Prune Truong Martin Danelljan Fisher Yu Luc Van Gool
Computer Vision Lab, ETH Zurich, Switzerland
{prune.truong, martin.danelljan, vangool}@vision.ee.ethz.ch i@yf.io

Abstract

We propose Probabilistic Warp Consistency, a weakly-supervised learning objective for semantic matching. Our approach directly supervises the dense matching scores predicted by the network, encoded as a conditional probability distribution. We first construct an image triplet by applying a known warp to one of the images in a pair depicting different instances of the same object class. Our probabilistic learning objectives are then derived using the constraints arising from the resulting image triplet. We further account for occlusion and background clutter present in real image pairs by extending our probabilistic output space with a learnable unmatched state. To supervise it, we design an objective between image pairs depicting different object classes. We validate our method by applying it to four recent semantic matching architectures. Our weakly-supervised approach sets a new state-of-the-art on four challenging semantic matching benchmarks. Lastly, we demonstrate that our objective also brings substantial improvements in the strongly-supervised regime, when combined with keypoint annotations.

1. Introduction

The semantic matching problem entails finding pixel-wise correspondences between images depicting instances of the same semantic category of object or scene, such as ‘cat’ or ‘bird’. It has received growing interest, due to its applications in *e.g.*, semantic segmentation [35,38] and image editing [1, 6, 9, 22]. The task nevertheless remains extremely challenging due to the large intra-class appearance and shape variations, view-point changes, and background-clutter. These issues are further complicated by the inherent difficulty to obtain ground-truth annotations.

While a few current datasets [10, 11, 30] provide manually annotated keypoints matches, these are often ill-defined, ambiguous and scarce. Strongly-supervised approaches relying on such annotations therefore struggle to generalize across datasets, as demonstrated in recent works [4, 31]. As a prominent alternative, unsupervised approaches [27, 31–33, 37, 39, 41] often train the network with synthetically generated dense ground-truth and image data. While benefiting from direct supervision, the lack of real image pairs often leads to poor generalization to real data. *Weakly-supervised* methods [13,16,31,33,34] thus appear as an attractive paradigm, leveraging supervision from real image pairs by only exploiting image-level class labels, which are inexpensive compared to keypoint annotations.

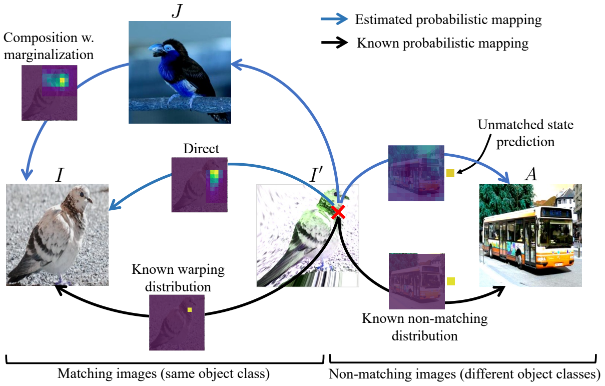

J
Estimated probabilistic mapping
Known probabilistic mapping
Composition w. marginalization
Direct
I
I'
Unmatched state prediction
A
Known warping distribution
Known non-matching distribution
Matching images (same object class)
Non-matching images (different object classes)

Figure 1. From a real image pair (I, J) representing the same object class, we generate a new image I' by warping I according to a randomly sampled transformation. We further extend the image triplet with an additional image A, that depicts a different object class. For each pixel in I', we introduce two consistency objectives by enforcing the conditional probability distributions obtained either from the composition $I' \rightarrow J \rightarrow I$, or directly through $I' \rightarrow I$, to be equal to the known warping distribution. We further model occlusion and unmatched regions by introducing a learnable unmatched state. It is trained by enforcing the predicted distribution between the non-matching images (I', A) to be mapped to the unmatched state for all pixels.

Previous weakly-supervised alternatives introduce objectives on the predicted dense correspondence volume, which encapsulates the matching confidences for all pairwise matches between the image pair. The most common strategy is to maximize the maximum scores [13, 34] or negative entropy [31] of the correspondence volume computed between images of the same class, while minimizing the same quantity for images of different classes. However, these strategies only provide very limited supervision due to their weak and indirect learning signal. While these approaches act directly on the predicted dense correspondence volume, Truong *et al.* [42] recently introduced Warp Consistency, a weakly-supervised learning objective for dense flow regression. The objective is derived from flow constraints obtained when introducing a third image, constructed by randomly warping one of the images in the original pair. While it achieves impressive results, the warp consistency objective is limited to the learning of flow regression. As such an approach predicts a single match for each pixel without any confidence measure, it struggles to handle occlusions and background clutter, which are prominent in the semantic matching task.

We propose Probabilistic Warp Consistency, a weakly-supervised learning objective for semantic matching. Following [4,13,34] and unlike [42], we employ a *probabilistic mapping* representation of the predicted dense correspondences, encoding the transitional probabilities from every pixel in one image to every pixel in the other. Starting from a real image pair (I, J), we consider the image triplet introduced in [42], where the synthetic image I' is related to I by a randomly sampled warp (Fig. 1). We derive our probabilistic consistency objective based on predicting the *known* probabilistic mapping relating I' to I with the composition through the image J. The composition is obtained by marginalizing over all the intermediate paths that link pixels in image I' to pixels in I through image J.

Since the constraints employed to derive our objective are only valid in mutually visible object regions, we further tackle the problem of identifying pixels that can be matched. This is particularly challenging in the presence of background clutter and occlusions, common in semantic matching. We explicitly model occlusion and unmatched regions, by introducing a learnable unmatched state into our probabilistic mapping formulation. To train the model to detect unmatched regions, we design an additional probabilistic loss that is applied on pairs of images depicting different object classes, as illustrated in Fig. 1. Further, we also employ a visibility mask, which constrains our introduced consistency loss to visible object regions.

We extensively evaluate and analyze our approach by applying it to four recent semantic matching architectures, across four benchmark datasets. In particular, we train SF-Net [21] and NC-Net [34] with our weakly-supervised Probabilistic Warp Consistency objective. Our approach brings relative gains of 4.3% and 5.8% on PF-Pascal [11] and PF-Willow [10] respectively, for SF-Net, and $+22.6\%$ and $+14.8\%$ for NC-Net on SPair-71K [30] and TSS [38], respectively. This leads to a new state-of-the-art on all four datasets. Finally, we extend our approach to the strongly-supervised regime, by combining our probabilistic objectives with keypoint supervision. When integrated in SF-Net, NC-Net, DHPF [31] and CATs [4], it leads to substantially better generalization properties across datasets, setting a new state-of-the-art on three benchmarks. Code is available at github.com/PruneTruong/DenseMatching

2. Related Work

Semantic matching architectures: Most semantic matching pipelines include 3 main steps, namely feature extraction, cost volume construction, and displacement estimation. Multiple works focus on the latter, through either predicting the global geometric transformation parameters [2, 16, 18, 32, 33, 37], or directly regressing the flow field [19, 39–42] relating an image pair. Nevertheless, most methods instead predict a cost volume as the final network output, which is further transposed to point-to-point correspondences with argmax or soft-argmax [21] operations. Recent methods thus focus on improving the cost volume aggregation stage, through formulating the semantic matching task as an optimal transport problem [26] or leveraging multi-resolution features and cost volumes [4,21,29,31,45]. Another line of work deals with refining the cost volume, with 4D [13, 23, 24, 34] or 6D [28] convolutions, an online optimization-based module [39], an encoder-decoder style architecture [17] or a Transformer module [4].

Unsupervised and weakly-supervised semantic matching: A common technique for unsupervised learning of semantic correspondences is to rely on synthetically warped versions of images [2, 17, 32, 37, 41]. It nevertheless comes at the cost of poorer generalization abilities to real data. Some methods instead use real image pairs, by leveraging additional annotations in the form of 3D CAD models [44, 46], segmentation masks [3, 21], or by jointly learning semantic matching with attribute transfer [19]. Most related to our work are approaches that use proxy losses on the cost volume constructed between real image pairs, with image labels as the only supervision [13,16,18,33,34]. Jeon *et al.* [16] identify correct matches from forward-backward consistency. NC-Net [34] and DCC-Net [13] are trained by maximizing the mean matching scores over all hard assigned matches from the cost volume. Min *et al.* [31] instead encourage low and high correlation entropy for image pairs depicting the same or different classes, respectively. In this work, we instead construct an image triplet by warping one of the original images with a known warp, from which we derive our probabilistic losses.

Unsupervised learning from videos: Our approach is also related to [14], which proposes a self-supervised approach for learning features, by casting matches as predictions of links in a space-time graph constructed from videos. Recent works [8, 15, 43] further leverage the temporal consistency in videos to learn a representation for feature matching.

3. Background: Warp Consistency

We derive our approach based on the warp consistency constraints introduced by [42]. They propose a weakly-supervised loss, termed Warp Consistency, for learning correspondence regression networks. We therefore first review relevant background and introduce the notation that we use.

We define the mapping $M_{I \leftarrow J} : \mathbb{R}^2 \rightarrow \mathbb{R}^2$, which encodes the absolute location $M_{I \leftarrow J}(\mathbf{j}) \in \mathbb{R}^2$ in I corresponding to the pixel location $\mathbf{j} \in \mathbb{R}^2$ in image J. We consistently use the hat $\widehat{\cdot}$ to denote an estimated or predicted quantity.

Warp Consistency graph: Truong *et al.* [42] first build an image triplet, which is used to derive the constraints. From a real image pair (I, J), an image triplet (I, I', J) is constructed, by creating I' through warping of I with a randomly sampled mapping M_W, as $I' = I \circ M_W$. Here, $\circ$ denotes function composition. The resulting triplet (I, I', J) gives rise to a warp consistency graph (Fig. 2a), from which a family of mapping-consistency constraints is derived.

Mapping-consistency constraints: Truong *et al.* [42] analyse the possible mapping-consistency constraints arising from the triplet and identify two of them as most suitable when designing a weakly-supervised learning objective for dense correspondence regression. Particularly, the proposed objective is based on the W-bipath constraint, where the mapping M_W is computed through the composition $I' \rightarrow J \rightarrow I$ via image J, formulated as,

$$M_W = M_{I \leftarrow J} \circ M_{J \leftarrow I'} \,. \tag{1}$$

It is further combined with the warp-supervision constraint,

$$M_W = M_{I \leftarrow I'} \,, \tag{2}$$

derived from the graph by the direct path $I' \rightarrow I$.

In [42], these constraints were used to derive a weakly-supervised objective for correspondence regression. However, regressing a mapping vector $M_{I \leftarrow J}(\mathbf{j})$ for each position $\mathbf{j}$ only retrieves the position of the match, without any information on its uncertainty or multiple hypotheses. We instead aim at predicting a matching conditional probability distribution for each position $\mathbf{j}$. The distribution encapsulates richer information about the matching ability of this location $\mathbf{j}$, such as confidence, uniqueness, and existence of the correspondence. In this work, we thus generalize the mapping constraints (1)- (2) extracted from the warp consistency graph to conditional probability distributions.

4. Method

We address the problem of estimating the pixel-wise correspondences relating an image pair (J, I), depicting semantically similar objects. The dense matches are encapsulated in the form of a conditional probability matrix, referred to as probabilistic mapping. The goal of this work is to design a weakly-supervised learning objective for probabilistic mappings, applied to the semantic matching task.

4.1. Probabilistic Formulation

In this section, we first introduce our probabilistic representation and define a typical base predictive architecture. We let $\mathbf{j} \in \mathbb{R}^2$ denote the 2D pixel location in a grid of dimension $h_J \times w_J$, corresponding to image J. We refer to $j \in \mathbb{R}$ as the index $j = \{1, ..., h_J w_J\}$ corresponding to $\mathbf{j}$ when the spatial dimensions $h_J \times w_J$ are vectorized into one dimension $h_J w_J$. Following [4, 31, 34], we aim at predicting the probabilistic mapping $P_{I \leftarrow J} \in \mathbb{R}^{h_I w_I \times h_J w_J}$ relating J to I. Given a position j in frame J, $P_{I \leftarrow J}(i|j)$ gives the probability that j is mapped to location i in image I. $P_{I \leftarrow J}(\cdot|j) \in \mathbb{R}^{h_I w_I}$ thus encodes the entire discrete conditional probability distribution of where j is mapped in image I. We can see $P_{I \leftarrow J}$ as a matrix, where each column at index j encapsulates the distribution $P_{I \leftarrow J}(\cdot|j)$. Also note that the probabilistic mapping $P_{I \leftarrow J}$ is asymmetric.

Probabilistic mapping prediction: We here describe a standard architecture predicting the probabilistic mapping P relating an image pair. We let $D^I \in \mathbb{R}^{h_I w_I \times d}$ and $D^J \in \mathbb{R}^{h_J w_J \times d}$ denote the d-channel feature maps extracted from the images I and J, respectively.

A cost volume $C_{I \leftarrow J} \in \mathbb{R}^{h_I w_I \times h_J w_J}$ is then constructed, which encodes the pairwise deep feature similarities between all locations in the two feature maps, as,

$$C_{I \leftarrow J}(i, j) = D^I(i)^T D^J(j) \,. \tag{3}$$

The cost volume is finally converted to a probabilistic mapping $P_{I \leftarrow J} \in \mathbb{R}^{h_I w_I \times h_J w_J}$ by simply applying the SoftMax operation over the first dimension,

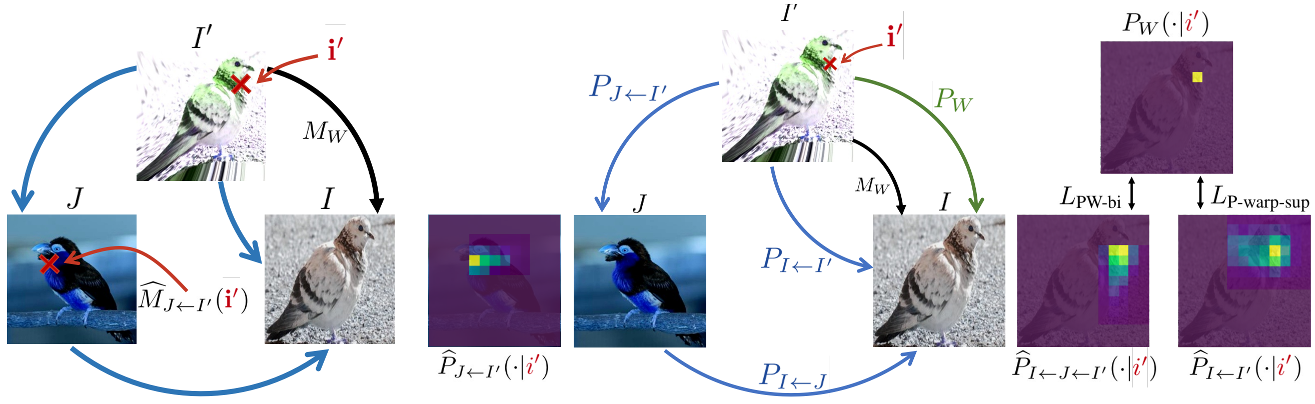
$$P_{I \leftarrow J}(i|j) = \frac{\exp(C_{I \leftarrow J}(i, j))}{\sum_k \exp(C_{I \leftarrow J}(k, j))} \tag{4}$$

Note that extensions of this basic approach can also be considered, by *e.g.* adding post-processing convolutional layers [13, 34] or a Transformer module [4]. The goal of this work is to design a weakly-supervised learning objective to train a neural network f_θ, with parameters θ, that predicts the probabilistic mapping $\widehat{P}_{I \leftarrow J} = f_\theta(J, I)$ relating J to I.

4.2. Probabilistic Warp Consistency Constraints

We set out to design a weakly-supervised loss for probabilistic mappings. To this end, we consider the consistency graph introduced in [42] and generalize the mapping constraints (1)- (2) to their corresponding probabilistic form.

Probabilistic W-bipath constraint: We start from the W-bipath constraint (1) extracted from the Warp Consistency graph Fig. 2a and extend it to its probabilistic matrix counterpart, which we denote as PW-bipath. It states that we


I'
$\mathbf{i}'$
M_W
J
I
$\widehat{M}_{J\leftarrow I'}(\mathbf{i}')$
$P_{J\leftarrow I'}$
P_W
M_W
$P_{I\leftarrow I'}$
$P_{I\leftarrow J}$
$P_W(\cdot|i')$
$L_{\text{PW-bi}}$
$L_{\text{P-warp-sup}}$
$\widehat{P}_{J\leftarrow I'}(\cdot|i')$
$\widehat{P}_{I\leftarrow J\leftarrow I'}(\cdot|i')$
$\widehat{P}_{I\leftarrow I'}(\cdot|i')$

(a) Warp Consistency Graph [42]

(b) Our probabilistic PW-bipath (6) and PWarp-supervision constraints, with corresponding losses (7)-(8)

Figure 2. Mapping and probabilistic mapping constraints derived from the warp consistency graph between the images (I, I', J). I' is generated by warping I according to a randomly sampled mapping M_W (black arrow). (a) The W-bipath (1) and warp-supervision (2) mapping constraints [42] predict M_W by the composition $I' \rightarrow J \rightarrow I$, and directly by $I' \rightarrow I$ respectively. (b) Our probabilistic PW-bipath and PWarp-supervision constraints are derived by enforcing the composition $\widehat{P}_{I\leftarrow J\leftarrow I'}$ of the predicted distributions, and the direct prediction $\widehat{P}_{I\leftarrow I'}$ respectively, to be equal to the known warping distribution P_W.

obtain the same conditional probability distribution by proceeding through the path $I' \rightarrow I$, which is determined by the randomly sampled warp M_W, or by taking the detour through image J. In the latter case, the resulting probability distribution is derived by marginalizing over the intermediate paths that link pixels in I' to pixels in I through J as,

$$P_W(i|i') = \sum_j P_{I\leftarrow J}(i|j) \cdot P_{J\leftarrow I'}(j|i')\,. \tag{5}$$

The above equality is expressed in matrix form as,

$$P_W = P_{I\leftarrow J} \otimes P_{J\leftarrow I'}\,. \tag{6}$$

where $\otimes$ represents matrix multiplication. This constraint is schematically represented in Fig. 2b.

PW-bipath training objective: We aim at formulating an objective based on the PW-bipath constraint (6). Crucially, in our setting, the mapping $M_{I\leftarrow I'} = M_W$ is known by construction, from which we can derive the ground-truth probabilistic mapping $P_{I\leftarrow I'} = P_W \in \mathbb{R}^{h_I w_I \times h_{I'} w_{I'}}$. To measure the distance between the right and the left side of (6), the KL divergence appears as a natural choice. Since P_W is a constant, it simplifies to the familiar cross-entropy,

$$L_{\text{PW-bi}} = \sum_{i'} \mathcal{H}\left([\widehat{P}_{I\leftarrow J} \otimes \widehat{P}_{J\leftarrow I'}](\cdot|i')\ ,\ P_W(\cdot|i')\right) \tag{7}$$

Here, $\mathcal{H}$ is the cross-entropy loss. To simplify notations, we sometimes refer to the marginalization as $\widehat{P}_{I\leftarrow J\leftarrow I'} = \widehat{P}_{I\leftarrow J} \otimes \widehat{P}_{J\leftarrow I'}$. Supervising $\widehat{P}_{I\leftarrow J\leftarrow I'}$ with the label P_W provides an implicit learning signal for the predicted intermediate distributions $\widehat{P}_{J\leftarrow I'}$ and $\widehat{P}_{I\leftarrow J}$.

PWarp-supervision constraint and objective: Similarly, we generalize the warp-supervision constraint (2) to its probabilistic matrix form, as $P_W = P_{I\leftarrow I'}$. As previously, by exploiting the fact that P_W is known, we derive the corresponding training objective,

$$L_{\text{P-warp-sup}} = \sum_{i'} \mathcal{H}\left(\widehat{P}_{I\leftarrow I'}(\cdot|i')\ ,\ P_W(\cdot|i')\right) \tag{8}$$

The PW-bipath constraint (6) and its loss (7) assume that all pixels of image I' have a match in both I and J. However, due to the occlusions introduced by the triplet creation and the non-matching backgrounds of the images in the semantic matching task, this assumption is partly invalidated.

4.3. Modelling Unmatched Regions

The semantic matching task aims to estimate correspondences between different image instances of the same object class. However, even in that case, the backgrounds of each image do not match. As a result, the *common visible regions* only represent a fraction of the images (see the birds in Fig. 2). Nevertheless, the distribution $P_{I\leftarrow J}(\cdot|j)$ is unable to model the no-match case for pixel j.

Moreover, the construction of our image triplet (I, I', J) introduces occluded areas, for which the constraint (6) is undefined. In fact, it is only valid in *non-occluded object regions*. However, in our setting, the locations of the objects in the real image pairs (I, J) are unknown. In this section, we derive our visibility-aware learning objective. We additionally introduce explicit modelling of occlusion and unmatchable regions into our probabilistic formulation.

Visibility-aware training objective: In general, the PW-bipath constraint (6) is only valid in regions of I' that are visible in both images J and I. That is, only in *non-occluded object regions*, as illustrated in Fig. 3. Applying the loss (7) in non-matching regions, such as in background areas, or in occluded objects regions (blue area in

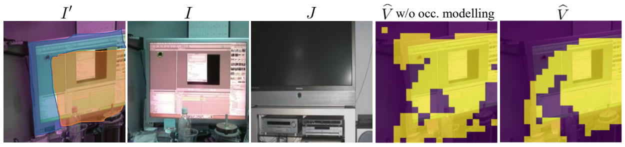


Figure 3. Triplet of images for training, and the visibility mask $\widehat{V}$ (yellow is $\widehat{V} = 1$). The shaded blue region on I' represent object pixels visible in both I' and I, but occluded in J, for which our PW-bipath loss (7) is not valid. It is only valid in object regions visible in all three images, *i.e.* the orange shaded region. Explicitly modelling occlusions further helps to identify them.

Fig. 3), bares the risk to confuse the network by enforcing matches in non-matching areas. As a result, we extend the introduced loss (7) by further integrating a visibility mask $V \in [0, 1]^{w_{I'} h_{I'}}$. The mask V takes a value $V(i') = 1$ for any pixel i' belonging to the non-occluded common object (roughly the orange area in Fig. 3) and $V(i') = 0$ otherwise. The loss (7) is then extended as,

$$L_{\text{vis-PW-bi}} = \sum_{i'} \widehat{V}(i')\ \mathcal{H}\left(\widehat{P}_{I\leftarrow J\leftarrow I'}(\cdot|i'), P_W(\cdot|i')\right) \quad (9)$$

Since we do not know the true V, we aim to find an estimate $\widehat{V}$, also visualized in Fig. 3. We consider the predicted probability value $\widehat{P}_{I\leftarrow J\leftarrow I'}(M_W(i')|i') \in [0, 1]$ of a pixel i' of I' to be mapped to position $M_W(i')$ in I, according to the known mapping M_W. We assume that this value should be higher in matching regions, *i.e.* the object, than in non-matching regions, *i.e.* the background, where the constraint (6) doesn't hold. We therefore compute our visibility mask by taking the highest γ percent of $\widehat{P}_{I\leftarrow J\leftarrow I'}(M_W(i')|i')$ over all i' of I'. The scalar γ is a hyperparameter controlling the sensitivity of the mask estimation. While we do not know the actual coverage of the object in the image, which might vary across training images, we found that taking a high estimate for γ is sufficient in practise, as it simply removes the obvious non-matching regions. Moreover, while we could have instead computed $\widehat{V}$ by thresholding the probabilities as $\widehat{V}(i') = \mathbb{1}\left[\widehat{P}_{I\leftarrow J\leftarrow I'}(M_W(i')|i') > \beta\right]$, our approach avoids tedious continuous tuning of the β parameter during training, necessary to follow the evolution of the probabilities. While valid as it is, the accuracy of the estimate $\widehat{V}$ can further be improved through explicit occlusion modelling.

Occlusion modelling: In order to explicitly model occlusion and non-matching regions into our probabilistic mapping $P_{I\leftarrow J}$, we *predict* the probability of a pixel to be occluded or unmatched in one image, given that it is visible in the other. This can, for example, be achieved by augmenting the cost volume C in (3) with an unmatched bin [7, 36] $\varnothing$, such as $C(\varnothing, j) = z \in \mathbb{R}$, where z is a single learnable parameter. After converting the cost volume C into a probabilistic mapping P through (4), $P_{I\leftarrow J}(\varnothing|j)$ encodes the probability of pixel j of image J to map to the unmatched or occluded state $\varnothing$, *i.e.* to have no match in image I. We further specify the matching distribution given an unmatched state, to always be mapped to the unmatched state. Specifically, we augment $\widehat{P}$ with a fixed column, forcing the distribution given an unmatched state to be as $\widehat{P}(\varnothing|\varnothing) = 1$.

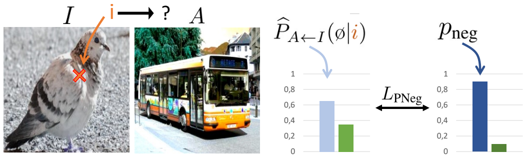


Figure 4. Learning objective on non-matching images (I, A).

Occlusion aware PW-bipath: Our modelling of the unmatched state given the unmatched state, as $\widehat{P}_{I\leftarrow J}(\varnothing|\varnothing) = 1$ naturally ensures that the following scheme is respected. If a pixel i' in image I' is predicted as unmatched in image J, such as $\widehat{P}_{J\leftarrow I'}(\varnothing|i') = 1$, it will also be predicted unmatched in image I, *i.e.* $\widehat{P}_{I\leftarrow J\leftarrow I'}(\varnothing|i') = 1$. This prevents enforcing (9) on $\widehat{P}_{I\leftarrow J\leftarrow I'}$ for pixels of image I' which are visible in I, but occluded in image J (blue area in Fig. 3). Moreover, predicting a high probability for the occluded state $\widehat{P}_{I\leftarrow J\leftarrow I'}(\varnothing|i')$ allows to identify occluded and non-matching areas i' in I'. It further ensures that these regions are not selected in $\widehat{V}$, and therefore not supervised with (9).

Supervision of the unmatched state: Our introduced objectives (8)-(9) do not impact the unmatched state $\varnothing$. We thus propose an additional loss to supervise it. Particularly, we aim at encouraging background and occluded object regions in images (I, I', J) depicting the same object class, to be predicted as unmatchable. Nevertheless, since the locations of the object in (I, J) are unknown during training, we cannot get direct supervision. To overcome this, we introduce an image A, depicting a different semantic content than the triplet. We then supervise the unmatched state by guiding the mode of the distribution between A and I to be in the unmatched state for all pixels of the images. The corresponding learning objective on the non-matching image pair (I, A) is defined as follows, and illustrated in Fig. 4,

$$L_{\text{PNeg}} = \sum_{i} \mathcal{B}(\widehat{P}_{A\leftarrow I}(\varnothing|i)\ ,\ p_{\text{neg}}) \quad (10)$$

$\mathcal{B}$ denotes the binary cross-entropy and we set $p_{\text{neg}} = 0.9$.

4.4. Final Training Objectives

Finally, we introduce our final weakly-supervised objective, the Probabilistic Warp Consistency, as a combination of our previously introduced PW-bipath (9), PWarp-supervision (8) and PNeg (10) objectives. We additionally propose a strongly-supervised approach, benefiting from our losses while also leveraging keypoint annotations.

Weak supervision: In this setup, we assume that only image-level class labels are given, such that each image pair is either positive, *i.e.* depicting the same object class, or negative, *i.e*. representing different classes, following [13, 31, 34]. We obtain our final weakly-supervised objective by combining the PW-bipath (9) and PWarp-supervision (8) losses applied to positive image pairs, with our negative probabilistic objective (10) on negative image pairs.

$$L_{\text{weak}} = L_{\text{vis-PW-bi}} + \lambda_{\text{P-warp-sup}} L_{\text{P-warp-sup}} + \lambda_{\text{PNeg}} L_{\text{PNeg}} \quad (11)$$

Here, $\lambda_{\text{P-warp-sup}}$ and λ_{PNeg} are weighting factors.

Strong supervision: We extend our approach to the strongly-supervised regime, where keypoint match annotations are given for each training image pair. Previous approaches [4, 24, 28] leverage these annotations by training semantic networks with a keypoint objective L_{kp}. Our final strongly-supervised objective is defined as the combination of the keypoint loss with our PW-bipath (9) and PWarp-supervision (8) objectives. Note that we do not include our explicit occlusion modelling, *i.e*. the unmatched state and its corresponding loss (10) on negative image pairs. This is to ensure fair comparison to previous strongly-supervised approaches, which solely rely on keypoint annotations, and not on image-level labels, required for our loss (10).

$$L_{\text{strong}} = L_{\text{vis-PW-bi}} + \lambda_{\text{P-warp-sup}} L_{\text{P-warp-sup}} + \lambda_{\text{kp}} L_{\text{kp}} \quad (12)$$

Here, $\lambda_{\text{vis-PW-bi}}$ and λ_{kp} also are weighting factors.

5. Experimental Results

We evaluate our weakly-supervised learning approach for two semantic networks. The benefits brought by the combination of our probabilistic losses with keypoint annotations are also demonstrated for four recent networks. We extensively analyze our method and compare it to previous approaches, setting a new state-of-the-art on multiple challenging datasets.

5.1. Networks and Implementation Details

For weak supervision, we integrate our approach (11) into baselines SF-Net [21] and NC-Net [34]. It leads to our weakly-supervised **PWarpC-SF-Net** and **PWarpC-NC-Net** respectively. We also apply our strongly-supervised loss (12) to baselines SF-Net, NC-Net, DHPF [31] and CATs [4], resulting in respectively **PWarpC-SF-Net***, **PWarpC-NC-Net***, **PWarpC-DHPF** and **PWarpC-CATs**. For fair comparison, we additionally train a strongly-supervised baseline for both SF-Net and NC-Net, referred to as SF-Net* and NC-Net*. Note that for all methods, the strongly-supervised baseline is trained with only L_{kp}, which is defined as the cross-entropy loss for SF-Net*, NC-Net* and DHPF, and the End-Point-Error objective after applying soft-argmax [21] for CATs. To convert the predicted probabilistic mapping to point-to-point matches for evaluation, all networks trained with our PWarpC objectives employ the argmax operation, except for PWarpC-CATs where we adopt the same soft-argmax as in the baseline CATs [4]. Additional details on the integration of our objectives for each architecture are provided in the appendix, Sec. A-F. We train all networks on PF-Pascal [11], using the splits of [12]. The results when trained on SPair-71K are further presented in the appendix, Sec. G.1.

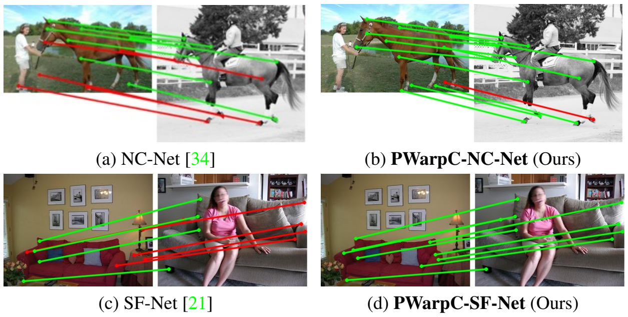

(a) NC-Net [34]
(b) **PWarpC-NC-Net** (Ours)
(c) SF-Net [21]
(d) **PWarpC-SF-Net** (Ours)

Figure 5. Example predictions of baselines NC-Net [34] and SF-Net [21], compared to our weakly-supervised PWarpC-NC-Net and PWarpC-SF-Net. Green and red line denotes correct and wrong predictions, respectively, with respect to the ground-truth.

5.2. Experimental Settings

We evaluate our networks on four standard benchmark datasets for semantic matching, namely PF-Pascal [11], PF-Willow [10], SPair-71K [30] and TSS [38]. Results on Caltech-101 [20] are further shown in appendix H.6.

Datasets: The **PF-Pascal**, **PF-Willow** and **SPair-71K** are keypoint datasets, which respectively contain 1341, 900 and 70958 image pairs from 20, 4 and 18 categories. Images have dimensions ranging from 102 × 300 to 500 × 500. **TSS** is the only dataset providing dense flow field annotations for the foreground object in each pair. It contains 400 image pairs, divided into three groups: FG3DCAR, JODS, and PASCAL, according to the origins of the images.

Metrics: We adopt the standard metric, Percentage of Correct Keypoints (PCK), with a pixel threshold of $\alpha_\tau \cdot \max(h_s^\tau, w_s^\tau)$. Here, h_s and w_s are either the dimensions of the source image or the dimensions of the object bounding box in the source image, such as $\tau \in \{\text{img}, \text{bbox}\}$.

5.3. Results

We present results on PF-Pascal, PF-Willow, SPair-71K and TSS in Tab. 1. A few previous approaches compute the PCK metrics after resizing the annotations to a different resolution than the original. Nevertheless, we found that in practise, the annotation resolution can lead to notable variations in results, as evidenced for DHPF or CATs in Tab. 1. For fair comparison, we thus compute the metrics on the standard setting, *i.e*. the original image size, and re-compute the PCK in this setting for baseline works if necessary. We also indicate the annotation size used, whenever reported by the authors or provided in their public implementation.

			PF-Pascal			**PF-Willow**			**Spair-71K**		**TSS**			
			PCK @ α_{img}			PCK @ α_{bbox}			PCK @ α_{bbox}		PCK @ α_{img} , $\alpha = 0.05$			
	Methods	Reso	0.05	0.10	0.15	0.05	0.10	0.15	0.05	0.10	FG3DCar	JODS	Pascal	Avg.
S	UCN$_{\text{res101}}$ [5]	-	-	75.1	-	-	-	-	-	17.7	-	-	-	-
	SCNet$_{\text{VGG16}}$ [12]	-	36.2	72.2	82.0	-	-	-	-	-	-	-	-	-
	HPF$_{\text{res101}}$ [29]	max 300	60.1	84.8	92.7	45.9	74.4	85.6	-	-	93.6	79.7	57.3	76.9
	SCOT$_{\text{res101}}$ [26]	max 300	63.1	85.4	92.7	47.8	76.0	87.1	-	-	95.3	81.3	57.7	78.1
	ANC-Net$_{\text{res101}}$ [24]	-	-	86.1	-	-	-	-	-	28.7	-	-	-	-
	CHM$_{\text{res101}}$ [28]	240	80.1	91.6	-	-	-	-	-	-	-	-	-	-
	PMD$_{\text{res101}}$ [25]	-	-	90.7	-	-	75.6	-	-	-	-	-	-	-
	PMNC$_{\text{res101}}$ [23]	-	**82.4**	90.6	-	-	-	-	-	28.8	-	-	-	-
	MMNet$_{\text{res101}}$ [45]	224 × 320	77.7	89.1	94.3	-	-	-	-	-	-	-	-	-
	DHPF$_{\text{res101}}$ [31]	240	75.7	90.7	95.0	41.4 †	67.4 †	81.8 †	15.4 †	27.4	-	-	-	-
	CATs$_{\text{res101}}$ [4]	256	67.5	89.1	94.9	37.4 †	65.8 †	79.7 †	10.9 †	22.4 †	-	-	-	-
	CATs-ft-features$_{\text{res101}}$ [4]	256	75.4	92.6	96.4	40.9 †	69.5 †	83.2 †	13.6 †	27.0 †	-	-	-	-
	CATs$_{\text{res101}}$ [4]	ori †	67.3	88.6	94.6	41.6	68.9	81.9	10.8	22.1	89.5	76.0	58.8	74.8
	PWarpC-CATs$_{\text{res101}}$	ori	67.1	88.5	93.8	44.2	71.2	83.5	12.2	23.3	93.2	83.4	70.7	82.4
	CATs-ft-features$_{\text{res101}}$ [4]	ori †	76.8	*92.7*	**96.5**	45.2	73.2	85.2	13.7	26.8	92.1	78.9	64.2	78.4
	PWarpC-CATs-ft-features$_{\text{res101}}$	ori	*79.8*	92.6	*96.4*	*48.1*	75.1	86.6	15.4	27.9	*95.5*	*85.0*	*85.5*	*88.7*
	DHPF$_{\text{res101}}$ [31]	ori	77.3	91.7	95.5	44.8	70.6	83.2	15.3	27.5	88.2	71.9	56.6	72.2
	PWarpC-DHPF$_{\text{res101}}$	ori	79.1	91.3	96.1	**48.5**	74.4	85.4	16.4	28.6	89.1	74.1	59.7	74.3
	NC-Net*$_{\text{res101}}$	ori	78.6	91.7	95.3	43.0	70.9	83.9	*17.3*	32.4	92.3	76.9	57.1	75.3
	PWarpC-NC-Net*$_{\text{res101}}$	ori	79.2	92.1	95.6	48.0	*76.2*	*86.8*	**21.5**	**37.1**	**97.5**	**87.8**	**88.4**	**91.2**
	SF-Net*$_{\text{res101}}$	ori	78.7	**92.9**	96.0	43.2	72.5	85.9	13.3	27.9	88.0	75.1	58.4	73.8
	PWarpC-SF-Net*$_{\text{res101}}$	ori	78.3	92.2	96.2	47.5	**77.7**	**88.8**	*17.3*	*32.5*	94.9	83.4	74.3	84.2
U	CNNGeo$_{\text{res101}}$ [32]	ori	41.0	69.5	80.4	36.9	69.2	77.8	-	18.1	90.1	76.4	56.3	74.4
	PARN$_{\text{res101}}$ [16]	ori	-	-	-	-	-	-	-	-	89.5	75.9	71.2	78.8
	GLU-Net$_{\text{vgg16}}$ [41]	ori	42.2	69.1	83.1	30.4	57.7	72.9	-	-	93.2	73.3	71.1	79.2
	Semantic-GLU-Net$_{\text{vgg16}}$ [41,42]	ori	48.3	72.5	85.1	39.7	67.5	82.1	7.6	16.5	95.3	82.2	78.2	85.2
	A2Net$_{\text{res101}}$ [37]	-	42.8	70.8	83.3	36.3	68.8	84.4	-	20.1	-	-	-	-
	PMD$_{\text{res101}}$ [25]	-	-	80.5	-	-	73.4	-	-	-	-	-	-	-
M	SF-Net$_{\text{res101}}$ [21]	288 / ori	53.6	81.9	90.6	46.3	74.0	84.2	-	-	-	-	-	-
	SF-Net$_{\text{res101}}$ [21]	ori †	59.0	84.0	92.0	46.3	74.0	84.2	11.2	24.0	90.8	78.6	58.0	75.8
W	**PWarpC-SF-Net**$_{\text{res101}}$	ori	**65.7**	**87.6**	*93.1*	*47.5*	**78.3**	**89.0**	17.6	33.5	95.1	84.7	76.8	85.5
	WarpC-SF-Net$_{\text{res101}}$ ◇ [21,42]	ori	*64.9*	*86.1*	92.2	46.9	*76.6*	87.9	13.1	26.6	95.7	82.3	68.8	82.2
	WeakAlign$_{\text{res101}}$ [33]	ori/ ori / -	49.0	75.8	84.0	38.2	71.2	85.8	-	21.1	90.3	76.4	56.5	74.4
	RTNs$_{\text{res101}}$ [18]	-	55.2	75.9	85.2	41.3	71.9	86.2	-	-	90.1	78.2	63.3	77.2
	DCCNet$_{\text{res101}}$ [13]	240 / ori / -	55.6	82.3	90.5	43.6	73.8	86.5	-	26.7	93.5	82.6	57.6	77.9
	SAM-Net$_{\text{vgg19}}$ [19]	-	60.1	80.2	86.9	-	-	-	-	-	*96.1*	82.2	67.2	81.8
	DHPF$_{\text{res101}}$ [31]	240	56.1	82.1	91.1	40.5 †	70.6 †	83.8 †	14.7 †	28.5	-	-	-	-
	DHPF$_{\text{res101}}$ [31]	ori †	61.2	84.1	92.4	45.1	73.6	85.0	14.7	27.8	-	-	-	-
	GSF$_{\text{res101}}$ [17]	-	62.8	84.5	**93.7**	47.0	75.8	*88.9*	-	33.5	-	-	-	-
	PMD$_{\text{res101}}$ [25]	-	-	81.2	-	-	74.7	-	-	26.5	-	-	-	-
	WarpC-SemGLU-Net$_{\text{vgg16}}$ [42]	ori	62.1	81.7	89.7	**49.0**	75.1	86.9	13.4 †	23.8 †	**97.1**	84.7	*79.7*	*87.2*
	NC-Net$_{\text{res101}}$ [34]	240 / ori / -	54.3	78.9	86.0	44.0	72.7	85.4	-	26.4	-	-	-	-
	WarpC-NC-Net$_{\text{res101}}$ ◇ [34,42]	ori	59.1	75.0	81.2	44.6	70.1	81.3	*18.0*	*35.0*	95.8	*87.5*	79.3	87.0
	NC-Net$_{\text{res101}}$ [34]	ori †	60.5	82.3	87.9	44.0	72.7	85.4	13.9	28.8	94.5	81.4	57.1	77.7
	PWarpC-NC-Net$_{\text{res101}}$	ori	64.2	84.4	90.5	45.0	75.9	87.9	**18.2**	**35.3**	95.9	**88.8**	**82.9**	**89.2**

Table 1. PCK [%] obtained by different state-of-the-art methods on the PF-Pascal [11], PF-Willow [10], SPair-71K [30] and TSS [38] datasets. All approaches are trained on the training set of PF-Pascal, except for [41]. **S** denotes strong supervision using keypoint match annotations, **M** refers to using ground-truth object segmentation mask, **U** is fully unsupervised requiring only single images, and **W** refers to weakly-supervised with image-level class labels. Each method evaluates with ground-truth annotations resized to a specific resolution. However, using different ground-truth resolutions leads to slightly different results. We therefore use the standard setting of evaluating on the original resolution (**ori**) and gray the results computed with the ground-truth annotations at a different size. When needed, we re-compute metrics of baselines using the provided pre-trained weights, indicated by †. For each of our PWarpC networks, we compare to its corresponding baseline within the dashed-lines. For completeness, we also train the baseline networks using the weakly-supervised mapping-based Warp Consistency objective [42], indicated with ◇. Best and second best results are in red and blue respectively.

Weak supervision (W): In Tab. 1, bottom part, we compare approaches trained with weak-supervision in the form of image labels. In this setting, our PWarpC networks are trained with L_{weak} in (11). While bringing improvements on the PF-Pascal dataset itself, our approach PWarpC-NC-Net most notably achieves widely better generalization properties, with impressive 4.4% (+ 3.2), 22.6% (+ 6.5) and 14.8% (+ 11.5) relative (and absolute) gains compared to the baseline NC-Net on PF-Willow ($\alpha = 0.1$), SPair-71K ($\alpha = 0.1$) and TSS ($\alpha = 0.05$) respectively. Our PWarpC-NC-Net thus sets a new state-of-the-art on SPair-71K and TSS among weakly-supervised methods trained on PF-Pascal.

Even though it utilizes a lower degree of supervision, our approach PWarpC-SF-Net also significantly outperforms the baseline SF-Net, which is trained with mask supervision (M), on all datasets. In particular, it shows a relative (and absolute) gain of 4.3% (+ 3.6), 5.8% (+ 4.3) and 39.6% (+ 9.5) on respectively PF-Pascal, PF-Willow and SPair-71K for $\alpha = 0.1$, and of 10.9% (+ 8.3) on TSS for $\alpha = 0.05$. This makes our PWarpC-SF-Net the new state-of-the-art across all unsupervised (U), weakly-supervised (W) and mask-supervised (M) approaches on PF-Pascal and PF-Willow. Example predictions are shown in Fig. 5

Strong supervision (S): In the top part of Tab. 1, we evaluate networks trained with strong supervision, in the form of key-point annotations. Our strongly-supervised PWarpC approaches are trained with our L_{strong} (12). For all networks, while the results are on par with the baselines on PF-Pascal, the PWarpC networks show drastically better performance on PF-Willow, SPair-71K and TSS compared to their respective baselines. PWarpC-SF-Net* and PWarpC-NC-Net* thus set a new state-of-the-art on respectively PF-Willow, and the SPair-71K and TSS datasets, across all strongly-supervised approaches trained on PF-Pascal. Finally, while most works focus on designing novel semantic architectures, we here show that the right training strategy bridges the gap between architectures.

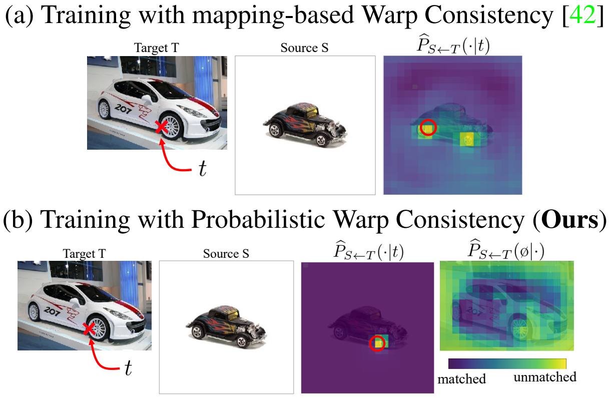
5.4. Method Analysis

We here perform a comprehensive analysis of our approach in Tab. 2. We adopt SF-Net as the base architecture.

Ablation study: In the top part of Tab. 2, we analyze key components of our approach. The version denoted as (II) is trained using our PW-bipath objective (7), without the visibility mask. Further introducing our visibility mask (9) in (III) significantly boosts the results since it enables to supervise only in the common visible regions. Note that this version (III) already outperforms the baseline SF-Net (I), while using less annotation (class instead of mask). In (IV), we add our probabilistic warp-supervision (8), leading to a small improvement for all thresholds and on all datasets. From (IV) to (V), we further introduce our explicit occlusion modelling associated with our negative loss (10), which results in drastically better performance. This version corresponds to our final weakly-supervised PWarC-SF-Net, trained with (11). An example of the regions identified as unmatched by PWarpC-SF-Net is shown in Fig. 6b.

		PF-Pascal		PF-Willow		Spair-71K	TSS
		α_{img}		α_{bbox}		α_{bbox}	α_{img}
	Methods	0.05	0.10	0.05	0.10	0.10	0.05
I	SF-Net baseline	59.0	84.0	46.3	74.0	24.0	75.8
II	PW-bipath (7)	59.1	82.3	44.9	74.3	28.0	83.4
III	+ visibility mask (9)	61.2	83.7	46.1	75.8	28.5	78.4
IV	+ PWarp-supervision (8)	63.0	84.9	47.0	76.9	30.7	83.5
V	+ PNeg (10) (**PWarpC-SF-Net**)	**65.7**	**87.6**	**47.5**	**78.3**	**33.5**	**85.5**
V	**PWarpC-SF-Net** (Ours)	**65.7**	**87.6**	**47.5**	**78.3**	**33.5**	**85.5**
VI	Mapping Warp Consistency [42]	64.9	86.1	46.9	76.6	26.6	82.2
VII	PWarp-supervision only (8)	52.9	74.3	38.0	66.6	27.9	79.4
VIII	Max-score [34]	52.4	76.7	31.2	59.5	24.6	74.8
IX	Min-entropy [31]	44.7	74.4	25.4	57.8	20.6	69.6

Table 2. Ablation study (top part) and comparison to alternative objectives (bottom part) for PWarpC-SF-Net.


(a) Training with mapping-based Warp Consistency [42]

(b) Training with Probabilistic Warp Consistency (**Ours**)

Target T
Source S
$\widehat{P}_{S\leftarrow T}(\cdot|t)$
t
Target T
Source S
$\widehat{P}_{S\leftarrow T}(\cdot|t)$
$\widehat{P}_{S\leftarrow T}(\o|\cdot)$
t
matched
unmatched

Figure 6. In (a), SF-Net is trained using the mapping-based Warp Consistency approach [42], after converting the cost volume to a mapping through soft-argmax [21]. It predicts ambiguous matching scores, struggling to differentiate between the car wheels. Our probabilistic approach (b) instead directly predicts a Dirac-like distribution, whose mode is correct. Here, we also show that our approach identifies most of the background areas as unmatched.

Comparison to other losses: In Tab. 2, bottom part, we first compare our probabilistic approach (11) corresponding to (V) with the mapping-based warp consistency objective [42], denoted as (VI). Our approach (V) leads to better performance than warp consistency (VI), with a particularly impressive 6.9% absolute gain on the challenging SPair-71K dataset. We further illustrate the benefit of our approach on an example in Fig. 6. Moreover, using *only* the PWarp-supervision loss (8) in (VII) results in much worse performance than our Probabilistic Warp Consistency (V). Finally, we compare our approach (V) to previous losses applied on cost volumes. The versions (VIII) and (IX) are trained with respectively maximizing the max scores [34], and minimizing the cost volume entropy [31]. Both approaches lead to poor results, likely caused by the very indirect supervision signal that these objectives provide.

6. Conclusion

We propose Probabilistic Warp Consistency, a weakly-supervised learning objective for semantic matching. We introduce multiple probabilistic losses derived both from a triplet of images generated based on a real image pair, and from a pair of non-matching images. When integrated into four recent semantic networks, our approach sets a new state-of-the-art on four challenging benchmarks.

Limitations: Since our approach acts on cost volumes, which are memory expensive, it is limited to relatively coarse resolution. This might in turn impact its accuracy.

Acknowledgements: This work was supported by a Huawei Gift and the ETH Future Computing Laboratory (EFCL) financed by a gift from Huawei Technologies.

References

[1] Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan B. Goldman. Patchmatch: a randomized correspondence algorithm for structural image editing. *ACM Trans. Graph.*, 28(3):24, 2009. 1

[2] Jianchun Chen, Lingjing Wang, Xiang Li, and Yi Fang. Arbicon-net: Arbitrary continuous geometric transformation networks for image registration. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 3410–3420, 2019. 2

[3] Yun-Chun Chen, Yen-Yu Lin, Ming-Hsuan Yang, and Jia-Bin Huang. Show, match and segment: Joint weakly supervised learning of semantic matching and object co-segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(10):3632–3647, 2021. 2

[4] Seokju Cho, Sunghwan Hong, Sangryul Jeon, Yunsung Lee, Kwanghoon Sohn, and Seungryong Kim. Semantic correspondence with transformers. *NeurIPS*, 2021. 1, 2, 3, 6, 7

[5] Christopher Bongsoo Choy, JunYoung Gwak, Silvio Savarese, and Manmohan Krishna Chandraker. Universal correspondence network. In *NIPS*, pages 2406–2414, 2016. 7

[6] Kevin Dale, Micah K. Johnson, Kalyan Sunkavalli, Wojciech Matusik, and Hanspeter Pfister. Image restoration using online photo collections. In *IEEE 12th International Conference on Computer Vision, ICCV 2009, Kyoto, Japan, September 27 - October 4, 2009*, pages 2217–2224. IEEE Computer Society, 2009. 1

[7] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *2018 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 224–236, 2018. 5

[8] Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. Temporal cycle-consistency learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 1801–1810, 2019. 2

[9] Yoav HaCohen, Eli Shechtman, Dan B. Goldman, and Dani Lischinski. Non-rigid dense correspondence with applications for image enhancement. *ACM Trans. Graph.*, 30(4):70, 2011. 1

[10] Bumsub Ham, Minsu Cho, Cordelia Schmid, and Jean Ponce. Proposal flow. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 1, 2, 6, 7

[11] Bumsub Ham, Minsu Cho, Cordelia Schmid, and Jean Ponce. Proposal flow: Semantic correspondences from object proposals. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(7):1711–1725, 2018. 1, 2, 6, 7

[12] Kai Han, Rafael S. Rezende, Bumsub Ham, Kwan-Yee K. Wong, Minsu Cho, Cordelia Schmid, and Jean Ponce. Scnet: Learning semantic correspondence. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 1849–1858, 2017. 6, 7

[13] Shuaiyi Huang, Qiuyue Wang, Songyang Zhang, Shipeng Yan, and Xuming He. Dynamic context correspondence network for semantic alignment. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 2010–2019. IEEE, 2019. 1, 2, 3, 6, 7

[14] Allan Jabri, Andrew Owens, and Alexei A. Efros. Space-time correspondence as a contrastive random walk. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 2

[15] Allan Jabri, Andrew Owens, and Alexei A. Efros. Space-time correspondence as a contrastive random walk. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 2

[16] Sangryul Jeon, Seungryong Kim, Dongbo Min, and Kwanghoon Sohn. PARN: pyramidal affine regression networks for dense semantic correspondence. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part VI*, pages 355–371, 2018. 1, 2, 7

[17] Sangryul Jeon, Dongbo Min, Seungryong Kim, Jihwan Choe, and Kwanghoon Sohn. Guided semantic flow. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXVIII*, volume 12373 of *Lecture Notes in Computer Science*, pages 631–648. Springer, 2020. 2, 7

[18] Seungryong Kim, Stephen Lin, Sangryul Jeon, Dongbo Min, and Kwanghoon Sohn. Recurrent transformer networks for semantic correspondence. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada.*, pages 6129–6139, 2018. 2, 7

[19] Seungryong Kim, Dongbo Min, Somi Jeong, Sunok Kim, Sangryul Jeon, and Kwanghoon Sohn. Semantic attribute matching networks. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 12339–12348, 2019. 2, 7

[20] R. Fergus L. Fei-Fei and P. Perona. One-shot learning of object categories. *IEEE Trans. Pattern Recognition and Machine Intelligence. In press.* 6

[21] Junghyup Lee, Dohyung Kim, Jean Ponce, and Bumsub Ham. Sfnet: Learning object-aware semantic correspondence. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 2278–2287, 2019. 2, 6, 7, 8

[22] Junsoo Lee, Eungyeup Kim, Yunsung Lee, Dongjun Kim, Jaehyuk Chang, and Jaegul Choo. Reference-based sketch image colorization using augmented-self reference and dense semantic correspondence. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*

2020, Seattle, WA, USA, June 13-19, 2020, pages 5800–5809. Computer Vision Foundation / IEEE, 2020. 1

[23] Jae Yong Lee, Joseph DeGol, Victor Fragoso, and Sudipta N. Sinha. Patchmatch-based neighborhood consensus for semantic correspondence. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 13153–13163. Computer Vision Foundation / IEEE, 2021. 2, 7

[24] Shuda Li, Kai Han, Theo W. Costain, Henry Howard-Jenkins, and Victor Prisacariu. Correspondence networks with adaptive neighbourhood consensus. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 10193–10202. Computer Vision Foundation / IEEE, 2020. 2, 6, 7

[25] Xin Li, Deng-Ping Fan, Fan Yang, Ao Luo, Hong Cheng, and Zicheng Liu. Probabilistic model distillation for semantic correspondence. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 7505–7514. Computer Vision Foundation / IEEE, 2021. 7

[26] Yanbin Liu, Linchao Zhu, Makoto Yamada, and Yi Yang. Semantic correspondence as an optimal transport problem. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 4462–4471. Computer Vision Foundation / IEEE, 2020. 2, 7

[27] Iaroslav Melekhov, Aleksei Tiulpin, Torsten Sattler, Marc Pollefeys, Esa Rahtu, and Juho Kannala. DGC-Net: Dense geometric correspondence network. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2019. 1

[28] Juhong Min and Minsu Cho. Convolutional hough matching networks. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 2940–2950. Computer Vision Foundation / IEEE, 2021. 2, 6, 7

[29] Juhong Min, Jongmin Lee, Jean Ponce, and Minsu Cho. Hyperpixel flow: Semantic correspondence with multi-layer neural features. In *ICCV*, 2019. 2, 7

[30] Juhong Min, Jongmin Lee, Jean Ponce, and Minsu Cho. Spair-71k: A large-scale benchmark for semantic correspondence. *CoRR*, abs/1908.10543, 2019. 1, 2, 6, 7

[31] Juhong Min, Jongmin Lee, Jean Ponce, and Minsu Cho. Learning to compose hypercolumns for visual correspondence. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XV*, pages 346–363, 2020. 1, 2, 3, 6, 7, 8

[32] Ignacio Rocco, Relja Arandjelovic, and Josef Sivic. Convolutional neural network architecture for geometric matching. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 39–48, 2017. 1, 2, 7

[33] Ignacio Rocco, Relja Arandjelovic, and Josef Sivic. End-to-end weakly-supervised semantic alignment. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 6917–6925, 2018. 1, 2, 7

[34] Ignacio Rocco, Mircea Cimpoi, Relja Arandjelovic, Akihiko Torii, Tomás Pajdla, and Josef Sivic. Neighbourhood consensus networks. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada.*, pages 1658–1669, 2018. 1, 2, 3, 6, 7, 8

[35] Michael Rubinstein, Armand Joulin, Johannes Kopf, and Ce Liu. Unsupervised joint object discovery and segmentation in internet images. In *2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, June 23-28, 2013*, pages 1939–1946. IEEE Computer Society, 2013. 1

[36] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 4937–4946, 2020. 5

[37] Paul Hongsuck Seo, Jongmin Lee, Deunsol Jung, Bohyung Han, and Minsu Cho. Attentive semantic alignment with offset-aware correlation kernels. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part IV*, pages 367–383, 2018. 1, 2, 7

[38] Tatsunori Taniai, Sudipta N. Sinha, and Yoichi Sato. Joint recovery of dense correspondence and cosegmentation in two images. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 4246–4255, 2016. 1, 2, 6, 7

[39] Prune Truong, Martin Danelljan, Luc Van Gool, and Radu Timofte. GOCor: Bringing globally optimized correspondence volumes into your neural network. In *Annual Conference on Neural Information Processing Systems, NeurIPS*, 2020. 1, 2

[40] Prune Truong, Martin Danelljan, Luc Van Gool, and Radu Timofte. Learning accurate dense correspondences and when to trust them. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, 2021. 2

[41] Prune Truong, Martin Danelljan, and Radu Timofte. GLU-Net: Global-local universal network for dense flow and correspondences. In *2020 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2020*, 2020. 1, 2, 7

[42] Prune Truong, Martin Danelljan, Fisher Yu, and Luc Van Gool. Warp consistency for unsupervised learning of dense correspondences. In *IEEE/CVF International Conference on Computer Vision, ICCV*, 2021. 2, 3, 4, 7, 8

[43] Xiaolong Wang, Allan Jabri, and Alexei A. Efros. Learning correspondence from the cycle-consistency of time. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 2566–2576, 2019. 2

[44] Yang You, Chengkun Li, Yujing Lou, Zhoujun Cheng, Lizhuang Ma, Cewu Lu, and Weiming Wang. Semantic correspondence via 2d-3d-2d cycle. *CoRR*, abs/2004.09061, 2020. 2

[45] Dongyang Zhao, Ziyang Song, Zhenghao Ji, Gangming Zhao, Weifeng Ge, and Yizhou Yu. Multi-scale matching networks for semantic correspondence. 2021. 2, 7

[46] Tinghui Zhou, Philipp Krähenbühl, Mathieu Aubry, Qi-Xing Huang, and Alexei A. Efros. Learning dense correspondence via 3d-guided cycle consistency. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 117–126, 2016. 2