Warp Consistency for Unsupervised Learning of Dense 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

The key challenge in learning dense correspondences lies in the lack of ground-truth matches for real image pairs. While photometric consistency losses provide unsupervised alternatives, they struggle with large appearance changes, which are ubiquitous in geometric and semantic matching tasks. Moreover, methods relying on synthetic training pairs often suffer from poor generalisation to real data.

We propose Warp Consistency, an unsupervised learning objective for dense correspondence regression. Our objective is effective even in settings with large appearance and view-point changes. Given a pair of real images, we first construct an image triplet by applying a randomly sampled warp to one of the original images. We derive and analyze all flow-consistency constraints arising between the triplet. From our observations and empirical results, we design a general unsupervised objective employing two of the derived constraints. We validate our warp consistency loss by training three recent dense correspondence networks for the geometric and semantic matching tasks. Our approach sets a new state-of-the-art on several challenging benchmarks, including MegaDepth, RobotCar and TSS. Code and models are at github.com/PruneTruong/DenseMatching.

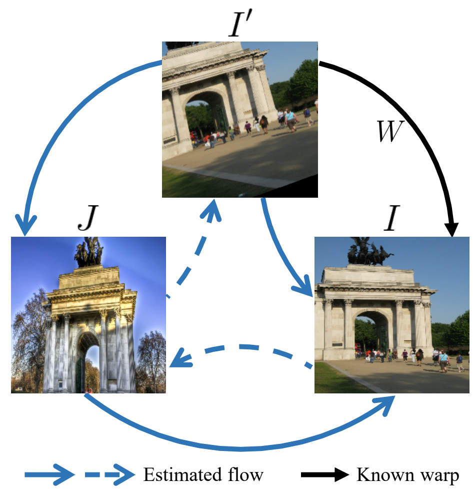


Figure 1. We introduce the warp consistency graph of the image triplet (I, I', J). The image I' is constructed by warping I according to a randomly sampled flow W (black arrow). The blue arrows represent predicted flows. Our unsupervised loss is derived from the two constraints represented by the solid arrows, which predict W by the composition $I' \rightarrow J \rightarrow I$ and directly by $I' \rightarrow I$.

1. Introduction

Finding dense correspondences between images continues to be a fundamental vision problem, with many applications in video analysis [44], image registration [48, 42], image manipulation [7, 25], and style transfer [19, 24]. While supervised deep learning methods have achieved impressive results, they are limited by the availability of ground-truth annotations. In fact, collecting dense ground-truth correspondence data of real scenes is extremely challenging and costly, if not impossible. Current approaches therefore resort to artificially rendered datasets [4, 14, 45, 13], sparsely computed matches [5, 55], or sparse manual annotations [3, 34, 10]. These strategies lack realism, accuracy, or scalability. In contrast, there is a virtually endless source of unlabelled image and video data, which calls for the design of effective unsupervised learning approaches.

Photometric objectives, relying on the brightness constancy assumption, have prevailed in the context of unsupervised optical flow [35, 57, 31]. However, in the more general case of geometric matching, the images often stem from radically different views, captured at different occasions, and under different conditions. This leads to large appearance transformations between the frames, which significantly undermine the brightness constancy assumption. It is further invalidated in the semantic matching task [25], where the images depict different instances of the same object class. As a prominent alternative to photometric objectives, warp-supervision [50, 49, 36, 32], also known as self-supervised learning [37, 40, 34], trains the network on synthetically warped versions of an image. While benefiting from direct supervision, the lack of real image pairs often leads to poor generalization to real data.

We introduce *Warp Consistency*, an unsupervised learning objective for dense correspondence regression. Our loss

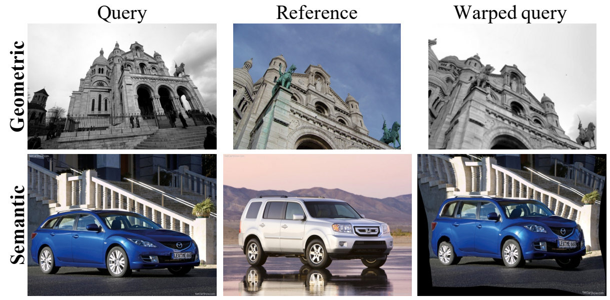


Figure 2. Warped query image (right) according to our predicted flow. Geometric and semantic matching applications pose highly challenging appearance and geometric transformations.

leverages real image pairs without invoking the photometric consistency assumption. Unlike previous approaches, it is capable of handling large appearance and view-point changes, while also generalizing to unseen real data. From a real image pair (I, J), we construct a third image I' by warping I with a known flow field W, that is created by randomly sampling *e.g*. homographies, from a specified distribution. We then consider the consistency graph arising from the resulting image triplet (I, I', J), visualized in Fig. 1. It is used to derive a family of new flow-consistency constraints. By carefully analyzing their properties, we propose an unsupervised loss based on predicting the flow W by the composition $I' \to J \to I$ via image J (Fig. 1). Our final warp consistency objective is then obtained by combining it with the warp-supervision constraint, also derived from our consistency graph by the direct path $I' \to I$.

We perform comprehensive empirical analysis of the objectives derived from our warp consistency graph and compare them to existing unsupervised alternatives. In particular, our warp consistency loss outperforms approaches based on photometric consistency and warp-supervision on multiple geometric matching datasets. We further perform extensive experiments for two tasks by integrating our approach into three recent dense matching architectures, namely GLU-Net [50] and RANSAC-Flow [41] for geometric matching, and SemanticGLU-Net [50] for semantic matching. Our unsupervised learning approach brings substantial gains: $+18.2\%$ PCK-5 on MegaDepth [23] for GLU-Net, $+2.8\%$ PCK-5 for RANSAC-Flow on RobotCar [20, 29], as well as $+16.1\%$ and $+4.4\%$ PCK-0.05 on PF-Pascal [9] and TSS [46] respectively, for SemanticGLU-Net. This leads to a new state-of-the-art on all four datasets. Example predictions are shown in Fig. 2.

2. Related work

Unsupervised optical flow: While supervised optical flow networks need carefully designed synthetic datasets for their training [4, 30], unsupervised approaches do not require ground-truth annotations. Inspired by classical optimization-based methods [11], they instead learn deep models based on brightness constancy and spatial smoothness losses [35, 57]. The predominant technique mainly relies on photometric losses, *e.g*. Charbonnier penalty [57], census loss [31], or SSIM [54, 52]. Such losses are often combined with forward-backward consistency [31] and edge-aware smoothness regularization [53]. Occlusion estimation techniques [16, 31, 53] are also employed to mask out occluded or outlier regions from the objective. Recently, several works [27, 28, 26] use a data distillation approach to improve the flow predictions in occluded regions. However, all aforementioned approaches rely on the assumption of limited appearance changes between two consecutive frames. While this assumption holds to a large degree in optical flow data, it is challenged by the drastic appearance changes encountered in geometric or semantic matching applications, as visualised in Fig. 2.

Unsupervised geometric matching: Geometric matching focuses on the more general case where the geometric transformations and appearance changes between two frames may be substantial. Methods either estimate a dense flow field [32, 50, 49, 41] or output a cost volume [39, 55], which can be further refined to increase accuracy [38, 22, 47]. The later approaches train the feature embedding, which is then used to compute dense similarity scores. Recent works further leverage the temporal consistency in videos to learn a suitable representation for feature matching [6, 15, 51]. Our work focuses on the first class of methods, which directly learn to regress a dense flow field. Recently, Xen *et al*. [41] use classical photometric and forward-backward consistency losses to train RANSAC-Flow. They partially alleviate the sensitivity of photometric losses to large appearance changes by pre-aligning the images with Ransac. Several methods [32, 50, 49] instead use a warp-supervision loss. By posing the network to regress a randomly sampled warp during training, a direct supervisory signal is obtained, but at the cost of poorer generalization abilities to real data.

Semantic correspondences: Semantic matching poses additional challenges due to intra-class appearance and shape variations. Manual annotations in this context are ill-defined and ambiguous, making it crucial to develop unsupervised objectives. Methods rely on warp-supervision strategies [36, 37, 3, 40, 50], use proxy losses on the cost volume [12, 39, 37, 34], identify correct matches from forward-backward consistency of the cost volumes [17], or jointly learn semantic correspondence with attribute transfer [19] or segmentation [21]. Most related to our work are [58, 56, 59]. Zhou *et al*. [58] learn to align multiple images using 3D-guided cycle-consistency by leveraging the ground-truth matches between multiple CAD models. However, the need for 3D CAD models greatly limits its applicability in practice. In FlowWeb [59], the authors optimize online pre-existing pair-wise correspondences using the cycle consistency of flows between images in a collection. Unlike these approaches, we require pairs of images as unique supervision and propose a general loss formulation, learning to regress dense correspondences directly.

3. Method

3.1. Problem formulation and notation

We address the problem of finding pixel-wise correspondences between two images $I \in \mathbb{R}^{h\times w\times 3}$ and $J \in \mathbb{R}^{h\times w\times 3}$. Our goal is to estimate a dense displacement field $F_{I\to J} \in \mathbb{R}^{h\times w\times 2}$, often referred to as flow, relating pixels in I to J. The flow field $F_{I\to J}$ represents the pixel-wise 2D motion vectors in the coordinate system of image I. It is directly related to the mapping $M_{I\to J} \in \mathbb{R}^{h\times w\times 2}$, which encodes the absolute location $M_{I\to J}(\mathbf{x}) \in \mathbb{R}^2$ in J corresponding to the pixel location $\mathbf{x} \in \mathbb{R}^2$ in image I. It is thus related to the flow through $M_{I\to J}(\mathbf{x}) = \mathbf{x} + F_{I\to J}(\mathbf{x})$. It is important to note that the flow and mapping representations are asymmetric. $M_{I\to J}$ parametrizes a mapping from each pixel in image I, which is not necessarily bijective.

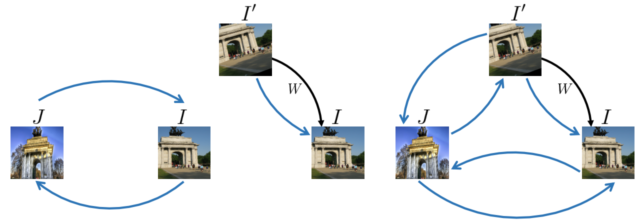
With a slight abuse of notation, we interchangeably view $F_{I\to J}$ and $M_{I\to J}$ as either elements of $\mathbb{R}^{h\times w\times 2}$ or as functions $F_{I\to J}, M_{I\to J} : \mathbb{R}^2 \to \mathbb{R}^2$. The latter is generally obtained by a bilinear interpolation of the former, and the interpretation will be clear from context when important. We define the *warping* $\Phi_F(T)$ of a function $T : \mathbb{R}^2 \to \mathbb{R}^d$ by the flow F as $\Phi_F(T)(\mathbf{x}) = T(\mathbf{x} + F(\mathbf{x}))$. This is more compactly expressed as $\Phi_F(T) = T \circ M_F$, where M_F is the mapping defined by F and $\circ$ denotes function composition. Lastly, we let $\mathbb{I} : \mathbb{R}^2 \to \mathbb{R}^2$ be the identity map $\mathbb{I}(\mathbf{x}) = \mathbf{x}$.

The goal of this work is to learn a neural network f_θ, with parameters θ, that predicts an estimated flow $\widehat{F}_{I\to J} = f_\theta(I, J)$ relating I to J. We will consistently use the hat $\widehat{\cdot}$ to denote an estimated or predicted quantity. The straightforward approach to learn f_θ is to minimize the discrepancy between the estimated flow $\widehat{F}_{I\to J}$ and the ground-truth flow $F_{I\to J}$ over a collection of real training image pairs (I, J). However, such supervised training requires large quantities of densely annotated data, which is extremely difficult to acquire for real scenes. This motivates the exploration of unsupervised alternatives for learning dense correspondences.

3.2. Unsupervised data losses

To develop our approach, we first briefly review relevant existing alternatives for unsupervised learning of flow. While there is no general agreement in the literature, we adopt a *practical* definition of unsupervised learning in our context. We call a learning formulation 'unsupervised' if it does not require any information (*i.e.* supervision) other than pairs of images (I, J) depicting the same scene or object. Specifically, unsupervised methods do not require any annotations made by humans or other matching algorithms.

Photometric losses: Most unsupervised approaches train



(a) Forw.-backw. (2) (b) Warp-superv. (3) (c) Warp consistency

Figure 3. Alternative unsupervised strategies.

the network using a photometric loss [57, 31, 53, 41]. Under the photometric consistency assumption, it minimizes the difference between image I and image J warped according to the estimated flow field $\widehat{F}_{I\to J}$ as,

$$L_{\text{photo}} = \rho\left(I\,,\,\Phi_{\widehat{F}_{I\to J}}(J)\right)\,. \tag{1}$$

Here, $\rho(\cdot,\cdot)$ is a function measuring the difference between two images, *e.g.* L_2 [57], SSIM [54], or census [31].

Forward-backward consistency: By constraining the backward flow $\widehat{F}_{J\to I}$ to yield the reverse displacement of its forward counterpart $\widehat{F}_{I\to J}$, we achieve the forward-backward consistency loss [31],

$$L_{\text{fb}} = \left\|\widehat{F}_{I\to J} + \Phi_{\widehat{F}_{I\to J}}(\widehat{F}_{J\to I})\right\|. \tag{2}$$

Here, $\|\cdot\|$ denotes a suitable norm. While well motivated, (2) is enforced by the trivial degenerate solution of always predicting zero flow $\widehat{F}_{I\to J} = \widehat{F}_{J\to I} = 0$. It therefore bares the risk of degrading performance by biasing the prediction towards zero, even if combined with a photometric loss (1). Both aforementioned losses are most often used together with a visibility mask that filters out the influence of occluded regions from the objective.

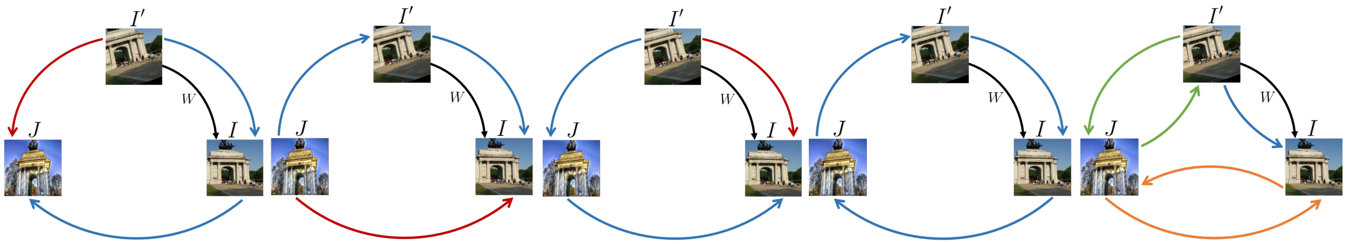
Warp-supervision: Another approach relies on synthetically generated training pairs, where the ground-truth flow is obtained by construction [50, 36, 32]. Given only a single image I, a training pair (I, I') is created by applying a randomly sampled transformation W, *e.g.* a homography, to I as $I' = \Phi_W(I)$. Here, W is the synthetic flow field, which serves as direct supervision through a regression loss,

$$L_{\text{warp}} = \left\|\widehat{F}_{I'\to I} - W\right\|. \tag{3}$$

While this results in a strong and direct training signal, warp supervision methods struggle to generalize to real image pairs (I, J). This can lead to over-smooth predictions and instabilities in the presence of unseen appearance changes.

3.3. Warp consistency graph

We set out to find a new unsupervised objective suitable for scenarios with large appearance and view-point changes, where photometric based losses struggle. While the photometric consistency assumption is avoided in the forward-backward consistency (Fig. 3a) and warp-supervision (Fig. 3b) objectives, these methods suffer from



(a) $I'J$-bipath (b) JI-bipath (c) W-bipath (d) Cycle consistency (e) Pair-wise consistency

Figure 4. Consistency relations derived from our warp consistency graph constructed between images (I, I', J). For the bipaths constraints a, b and c, the red and blue arrows indicate the paths used for the left and right hand side, respectively, of the constraints in (4)-(5).

severe drawbacks in terms of degenerate solutions and lack of realism, respectively. To address these issues, we consider all possible consistency relations obtained from the three images involved in both aforementioned objectives. Using this generalization, we not only retrieve forward-backward and warp-supervision as special cases, but also derive a *family of new consistency relations*.

From an image pair (I, J), we first construct an image triplet (I, I', J) by warping I with a known flow-field W in order to generate the new image $I' = \Phi_W(I)$. We now consider the full consistency graph, visualized in Fig. 3c, encompassing all flow-consistency constraints derived from the triplet of images (I, I', J). Crucially, we exploit the fact that the transformation $F_{I' \to I} = W$ is known. The goal is to find consistency relations that translate to suitable learning objectives. Particularly, we wish to improve the network prediction between the real image pair (I, J). We therefore first explore the possible consistency constraints that can be derived from the graph shown in Fig. 3c. For simplicity, we do not explicitly denote visible or valid regions of the stated consistency relations. They should be interpreted as an equality constraint for all pixel locations $\mathbf{x}$ where both sides represent a valid, non-occluded mapping or flow.

Pair-wise constraints: We first consider the consistency constraints recovered from pairs of images, as visualized in Fig. 4e. From the pair (I, J), and analogously (J, I'), we recover the standard forward-backward consistency constraint $\mathbb{I} = M_{J \to I} \circ M_{I \to J}$, from which we derive (2). Furthermore, from the pair (I', I) we can derive the warp-supervision constraint (3) $F_{I' \to I} = W$.[1]

Bipath constraints: The novel consistency relations stem from constraints that involve all three images in the triplet (I, I', J). These appear in two distinct types, here termed *bipath* and *cycle* constraints, respectively. We first consider the former, which have the form $M_{1 \to 2} = M_{3 \to 2} \circ M_{1 \to 3}$. That is, we obtain the same mapping by either proceeding directly from image 1 to 2 or by taking the detour through image 3. We thus compute the same mapping by two different *paths*: $1 \to 2$ and $1 \to 3 \to 2$, from which we derive the name of the constraint. The images 1, 2, and 3 represent any enumeration of the triplet (I, I', J) that respects the direction $I' \to I$, specified by the known warp W. There thus exist three different bipath constraints, detailed in Sec. 3.4.

Cycle constraints: The last category of constraints is formulated by starting from any of the three images in Fig. 4d and composing the mappings in a full cycle. Since we return to the starting image, the resulting composition is equal to the identity map. This is expressed in a general form as $\mathbb{I} = M_{3 \to 1} \circ M_{2 \to 3} \circ M_{1 \to 2}$, where we have proceeded in the cycle $1 \to 2 \to 3 \to 1$. Again constraining the direction $I' \to I$, we obtain three different constraints, as visualized in Fig. 4d. Compared to the bipath constraints, the cycle variants require two consecutive warping operations, stemming from the additional mapping composition. Each warp reduces the valid region and introduces interpolation noise and artifacts in practice. Constraints involving fewer warping operations are thus desirable, which is an advantage of the class of bipath constraints. In the next parts, we therefore focus on the later class to find a suitable unsupervised objective for dense correspondence estimation.

3.4. Bipath constraints

As mentioned in the previous section, there exist three different bipath constraints that preserve the direction of the known warp W. These are stated in terms of mappings as,

$$M_{I' \to J} = M_{I \to J} \circ M_W \tag{4a}$$
$$M_{J \to I} = M_W \circ M_{J \to I'} \tag{4b}$$
$$M_W = M_{J \to I'} \circ M_{I \to J} \,. \tag{4c}$$

From (4), we can derive the equivalent flow constraints as,

$$F_{I' \to J} = W + \Phi_W(F_{I \to J}) \tag{5a}$$
$$F_{J \to I} = F_{J \to I'} + \Phi_{F_{J \to I'}}(W) \tag{5b}$$
$$W = F_{I' \to J} + \Phi_{F_{I' \to J}}(F_{J \to I}) \,. \tag{5c}$$

Each constraint is visualized in Fig. 4a, b and c respectively. At first glance, any one of the constraints in (5) could be used as an unsupervised loss by minimizing the error between the left and right hand side. However, by separately

[1]While $\mathbb{I} = M_{I \to I'} \circ M_W$ and $\mathbb{I} = M_W \circ M_{I \to I'}$ are also possible, they offer no advantage over standard warp-supervision: $M_{I' \to I} = M_W$.

analyzing each constraint in (4)-(5), we will find them to have radically different properties which impact their suitability as an unsupervised learning objective.

$I'J$**-bipath:** The constraint (4a), (5a) is derived from the two possible paths from I' to J (Fig. 4a). While not obvious from (5a), it can be directly verified from (4a) that this constraint has a degenerate trivial solution. In fact, (4a) is satisfied for any W by simply mapping all inputs $\mathbf{x}$ to a constant pixel location $\mathbf{c} \in \mathbb{R}^2$ as $\widehat{M}_{I' \to J}(\mathbf{x}) = \widehat{M}_{I \to J}(\mathbf{x}) = \mathbf{c}$. In order to satisfy this constraint, the network can thus learn to predict the same flow $\widehat{F} = \mathbf{c} - \mathbb{I}$ for any input image pair.

JI**-bipath:** From the paths $J \to I$ in Fig. 4b, we achieve the constraint (4b), (5b). The resulting unsupervised loss is formulated as

$$L_{J \to I} = \left\| \widehat{F}_{J \to I'} + \Phi_{\widehat{F}_{J \to I'}}(W) - \widehat{F}_{J \to I} \right\|. \tag{6}$$

Unfortunately, this objective suffers from another theoretical disadvantage. Due to the cancellation effect between the estimated flow terms $\widehat{F}_{J \to I'}$ and $\widehat{F}_{J \to I}$, the objective (6) is insensitive to a constant bias in the prediction. Specifically, if a small constant bias $\mathbf{b} \in \mathbb{R}^2$ is added to all flow predictions in (6), it can be shown that the increase in the loss (6) is approximately bounded by $\left\| \Phi_{\widehat{F}_{J \to I'}}(DW\mathbf{b}) \right\|$. Here, the bias error $\mathbf{b}$ is scaled with the Jacobian DW of the warp W. Since a smooth and invertible warp W implies a generally small Jacobian DW, the change in the loss will be negligible. The resulting insensitivity of (6) to a prediction bias is further confirmed empirically by our experiments. We provide derivations in the suppl. A.1. To further understand and compare the bipath constraints (5), it is also useful to consider the limiting case of reducing the magnitude of the warps $\|W\| \to 0$. By setting $W = 0$ it can be observed that (6) becomes zero, *i.e.* no learning signal remains.

W**-bipath:** The third bipath constraint (4c), (5c) is derived from the paths $I' \to I$, which is determined by W (Fig. 4c). It leads to the W-bipath consistency loss,

$$L_W = \left\| \widehat{F}_{I' \to J} + \Phi_{\widehat{F}_{I' \to J}}(\widehat{F}_{J \to I}) - W \right\|. \tag{7}$$

We first analyze the limiting case $\|W\| \to 0$ by setting $W = 0$, which leads to standard forward-backward consistency (2) since $I' = I$. The W-bipath is thus a direct generalization of the latter constraint. Importantly, by randomly sampling non-zero warps W, degenerate solutions are avoided, effectively solving the one fatal issue of forward-backward consistency objectives. In addition to avoiding degenerate solutions, W-bipath does not experience cancellation of prediction bias, as in (6). Furthermore, compared to warp-supervision (3), it enables to directly learn the flow prediction $\widehat{F}_{J \to I}$ between the real pair (I, J). In the next section, we therefore develop our final unsupervised objective based on the W-bipath consistency.

3.5. Warp consistency loss

In this section, we develop our warp consistency loss, an unsupervised learning objective for dense correspondence estimation, using the consistency constraints derived in Sec. 3.3 and 3.4. Specifically, following the observations in Sec. 3.4, we base our loss on the W-bipath constraint.

W**-bipath consistency term:** To formulate an objective based on the W-bipath consistency constraint (5c), we further integrate a visibility mask $V \in [0,1]^{w \times h}$. The mask V takes a value $V(\mathbf{x}) = 1$ for any pixel $\mathbf{x}$ where both sides of (4c), (5c) represent a valid, non-occluded mapping, and $V(\mathbf{x}) = 0$ otherwise. The loss (7) is then extended as,

$$L_{\text{W-vis}} = \left\| \widehat{V} \cdot \left(\widehat{F}_{I' \to J} + \Phi_{\widehat{F}_{I' \to J}}(\widehat{F}_{J \to I}) - W \right) \right\|. \tag{8}$$

Since we do not know the true V, we replace it with an estimate $\widehat{V}$. While there are different techniques for estimating visibility masks [16, 31, 53], we base our strategy on [31]. Specifically, we compute our visibility mask as,

$$\begin{aligned} \widehat{V} = \mathbb{1}\Big[& \left| \widehat{F}_{I' \to J} + \Phi_{\widehat{F}_{I' \to J}}(\widehat{F}_{J \to I}) - W \right|_2^2 < \alpha_2 + \\ & \alpha_1 \left(\left| \widehat{F}_{I' \to J} \right|_2^2 + \left| \Phi_{\widehat{F}_{I' \to J}}(\widehat{F}_{J \to I}) \right|_2^2 + |W|_2^2 \right) \Big]. \end{aligned} \tag{9}$$

Here, $\mathbb{1}[\cdot]$ takes the value 1 or 0 if the input statement is true or false, respectively. The scalars α_1 and α_2 are hyperparameters controlling the sensitivity of the mask estimation. For the warp operation $\Phi_{\widehat{F}_{I' \to J}}(\widehat{F}_{J \to I})$, we generally found it beneficial not to back-propagate gradients through the flow $\widehat{F}_{I' \to J}$ used for warping. We believe that this better encourages the network to directly adjust the flow $\widehat{F}_{J \to I}$, rather than 'move' the flow vectors using the warp $\Phi_{\widehat{F}_{I' \to J}}$.

Warp-supervision term: In addition to our W-bipath objective (8), we use the warp-supervision (3), found as a pairwise constraint in our consistency graph (Fig. 4e). Benefiting from the strong and direct supervision provided by the synthetic flow W, the warp-supervision term increases convergence speed and helps in driving the network towards higher accuracy. Further, by the direct regression loss against the flow W, which is smooth by construction, it also acts as a smoothness constraint. On the other hand, through the W-bipath loss (8), the network learns the realistic motion patterns and appearance changes present between real images (I, J). As a result, both loss terms are mutually beneficial. From a practical perspective, the warp-supervision loss can be integrated at a low computational and memory cost, since the backbone feature extraction for the three images I, I', J can be shared between the two loss terms.

Adaptive loss balancing: Our final unsupervised objective combines the losses (8) and (3) as $L = L_{\text{W-vis}} + \lambda L_{\text{warp}}$. This raises the question of how to set the trade-off λ. Instead of resorting to manual tuning, we eliminate this hyper-parameter by automatically balancing the weights over each training batch as $\lambda = L_{\text{W-vis}}/L_{\text{warp}}$.

3.6. Sampling warps W

The key element of our warp consistency objective is the sampled warp W. During training, we randomly sample it from a distribution $W \sim p_W$, which we need to design. As discussed in Sec. 3.4, the W-bipath loss (8) approaches the forward-backward consistency loss (2) when the magnitude of the warps decreases $\|W\| \to 0$. Exclusively sampling too small warps $W \approx 0$ therefore risks biasing the prediction towards zero. On the other hand, too large warps would render the estimation of $\widehat{F}_{I' \to J}$ challenging and introduce unnecessary invalid image regions. As a rough guide, the distribution p_W should yield warps of similar magnitude as the real transformations $\|F_{J \to I}\|$, thus giving similar impact to all three terms in (8). Fortunately, as analyzed in the supplementary Sec. G, our approach is not sensitive to these settings as long as they are within reasonable bounds.

We construct W by sampling homography, Thin-plate Spline (TPS) and affine-TPS transformations randomly, following a procedure similar to previous approaches using warp-supervision [36]. **(i)** Homographies are constructed by randomly translating the four image corner locations. The magnitudes of the translations are chosen independently through Gaussian or uniform sampling, with standard-deviation or range equal to σ_H. **(ii)** For TPS, we randomly jitter a 3×3 grid of control points by independently translating each point. We use the same standard deviation or range σ_H as for our homographies. **(iii)** To generate larger scale and rotation changes, we also compose affine and TPS. We first sample affine transformations by selecting scale, rotation, translation and shearing parameters according to a Gaussian or uniform sampling. The TPS transform is then sampled as explained above and the final synthetic flow W is a composition of both flows.

To make the warps W harder, we optionally also compose the flow obtained from (i), (ii) and (iii) with randomly sampled elastic transforms. Specifically, we generate an elastic deformation motion field, as described in [43] and apply it in multiple regions selected randomly. Detailed settings are provided in the supplementary Sec. C, D and E.

4. Experiments

We evaluate our unsupervised learning approach for three dense matching networks and two tasks, namely GLU-Net [50] and RANSAC-Flow [41] for geometric matching, and SemanticGLU-Net [50] for semantic matching. We extensively analyze our method and compare it to earlier unsupervised objectives, defining a new state-of-the-art on multiple datasets. Further results, analysis, visualizations and implementation details are provided in the supplementary.

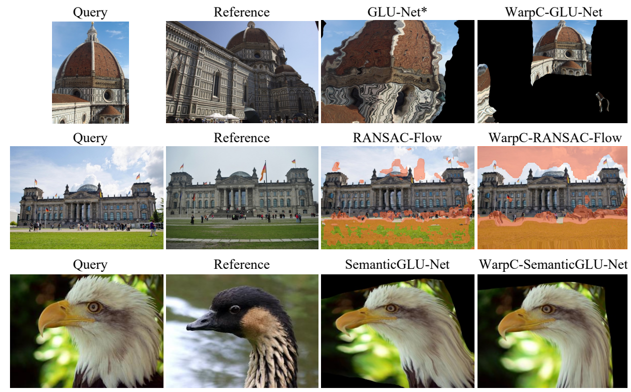


Figure 5. Warped query according to baseline network and our approach. In the middle row, we visualize the predicted mask by RANSAC-Flow based networks in red (unmatchable regions).

4.1. Method analysis

We first perform a comprehensive analysis of our approach. We adopt GLU-Net [50] as our base architecture. It is a 4-level pyramidal network operating at two image resolutions to estimate dense flow fields.

Experimental set-up for GLU-Net: We slightly simplify the GLU-Net [50] architecture by replacing the dense decoder connections with standard residual blocks, which drastically reduces the number of network parameters with negligible impact on performance. As in [50], the feature extraction network is set to a VGG-16 [2] with ImageNet pre-trained weights. We train the rest of the architecture from scratch in two stages. We first train GLU-Net using our unsupervised objective, described in Sec. 3.5, but without the visibility mask $\widehat{V}$. As a second stage, we add the visibility mask and employ stronger warps W, with elastic transforms. For both stages, we use the training split of the MegaDepth dataset [23], which comprises diverse internet images of 196 different world monuments.

Datasets and metrics: We evaluate on standard datasets with sparse ground-truth, namely **RobotCar** [29, 20] and **MegaDepth** [23]. For the latter, we use the test split of [41], which consists of 19 scenes not seen during training. Images in Robotcar depict outdoor road scenes and are particularly challenging due to their many textureless regions. MegaDepth images show extreme view-point and appearance variations. In line with [41], we use the Percentage of Correct Keypoints at a given pixel threshold T (PCK-T) as the evaluation metric (in %). We also employ the 59 sequences of the homography dataset **HPatches** [1]. We evaluate with the Average End-Point-Error (AEPE) and PCK.

Warp consistency graph losses: In Tab. 1 we empirically compare the constraints extracted from our warp consistency graph (Sec. 3.3). All networks are trained with only the first stage, on the same synthetic transformations W. Since we observed it to give a general improvement, we stop gradients through the flow used for warping (but not the flow that is warped). The $I'J$-bipath (II) and JI-bipath (III) losses lead to a degenerate solution and a large predicted bias respectively, which explains the very poor performance of the networks. The cycle loss (V) obtains much better results but does not reach the performance of the W-bipath constraint (IV). We only show the cycle starting from I' here (V), since it performs best among all cycle losses (see suppl. A.3). While the warp-supervision loss (I) results in a better accuracy on all datasets (PCK-1 and PCK-5 for HPatches), it is significantly less robust to large view-point changes than the W-bipath objective (IV), as evidenced by results in PCK-10 and AEPE. These two losses have complementary behaviors and combining them (VIII) leads to a significant gain in both accuracy and robustness. Combining the warp-supervison loss (I) with $I'J$-bipath (II) in (VI) or with JI-bipath (III) in (VII) instead results in drastically lower performance than (VIII). The cycle loss (V) with the warp-supervision (I) in (IX) is also slightly worse.

Ablation study: In Tab. 2 we analyze the key components of our approach. We first show the importance of not back-propagating gradients in the warp operation. Adding the warp-supervision objective with constant weights of $\lambda = 1$ increases both the network's accuracy and robustness for all datasets. Further using adaptive loss balancing (Sec. 3.5) provides a significant improvement in accuracy (PCK-1) for MegaDepth with only minor loss on other thresholds. Including our visibility mask $\widehat{V}$ in the second training stage drastically improves all metrics for all datasets. Finally, further sampling harder transformations results in better accuracy, particularly for PCK-1 on MegaDepth. We therefore use this as our standard setting in the following experiments, where we denote it as **WarpC**.

Comparison to alternative losses: Finally, in Tab. 3 we compare and combine our proposed objective with alternative losses. The census loss [31] (I), popular in optical flow, does not have sufficient invariance to appearance changes and thus leads to poor results on geometric matching datasets. The SSIM loss [54] (II) is more robust to the large appearance variations present in MegaDepth. Further combining SSIM with the forward-backward consistency loss (III) leads to a small improvement. Compared to SSIM (III) on MegaDepth, our WarpC approach (VI) achieves superior PCK-5 (+7.8%) and PCK-10 (+10.2%) at the cost of a slight reduction in sub-pixel accuracy. Furthermore, our approach demonstrates superior generalization capabilities by outperforming all other alternatives on the RobotCar and HPatches datasets. For completeness, we also evaluate the combination (VII) of our loss with the photometric SSIM loss. This leads to improved PCK-1 on MegaDepth but degrades other metrics compared to WarpC (VI). Nevertheless, adding WarpC significantly improves upon SSIM (II) for all thresholds and datasets. Moreover, combining the warp-supervision (IV) with the forward-backward loss in (V) leads to an improvement compared to (IV). It is however significantly worse than combining the warp-supervision with our W-bipath loss in (VI), which can be seen as a generalization of the forward-backward loss. Finally, we compare with using the sparse ground-truth supervision provided by SfM reconstruction of the MegaDepth training images. Interestingly, training the dense prediction network from scratch with solely sparse annotations (VIII) leads to inferior performance compared to our unsupervised objective (VI). Lastly, we fine-tune (IX) our proposed network (VI) with sparse annotations. While this leads to a moderate gain on MegaDepth, it comes at the cost of worse generalization properties on RobotCar and HPatches.

		MegaDepth			RobotCar			HPatches	
		PCK-1	PCK-5	PCK-10	PCK-1	PCK-5	PCK-10	AEPE	PCK-5
I	Warp-supervision (3)	**35.98**	57.21	63.88	**2.43**	33.63	54.50	28.50	**76.76**
II	$I'J$-bipath (5a)	0.00	0.05	0.21	0.00	0.00	0.13	370.80	0.01
III	JI-bipath (5b),(6)	0.00	0.06	0.21	0.00	0.05	0.21	162.50	0.04
IV	W-bipath (5c),(7)	29.55	**67.70**	**74.42**	2.25	**33.88**	**55.38**	**26.13**	70.51
V	I'-cycle	25.04	64.44	71.75	2.19	32.79	54.55	27.51	66.16
VI	$I'J$-bipath + warp-sup.	0.00	0.11	0.45	0.01	0.35	1.52	255.40	0.02
VII	JI-bipath + warp-sup.	33.72	61.10	67.44	2.26	34.06	55.07	28.91	71.52
VIII	W-bipath + warp-sup.	**43.47**	**69.90**	**75.23**	2.49	35.28	56.45	**22.83**	**78.60**
IX	I'-cycle + warp-sup.	42.11	68.84	74.28	**2.52**	**35.75**	**56.96**	24.16	78.58

Table 1. Analysis of warp consistency graph losses (Sec. 3.3-3.4).

	MegaDepth			RobotCar			HPatches	
	PCK-1	PCK-5	PCK-10	PCK-1	PCK-5	PCK-10	AEPE	PCK-5
W-bipath (7), grad in warp	20.06	58.57	67.83	2.04	31.70	53.57	29.37	60.40
W-bipath (7)	29.55	67.70	74.42	2.25	33.88	55.38	26.13	70.51
+ warp-supervision (3)	39.66	70.38	76.06	2.45	34.92	56.37	22.52	78.65
+ adaptive loss balancing	43.47	69.90	75.23	2.49	35.28	56.45	22.83	78.60
+ visibility mask $\widehat{V}$ (8)	48.86	77.58	82.27	**2.51**	35.78	57.19	21.63	82.55
+ harder warps W	**50.61**	**78.61**	**82.94**	**2.51**	**35.92**	**57.44**	**21.00**	**83.24**

Table 2. Ablation study by incrementally adding each component.

		MegaDepth			RobotCar			HPatches	
		PCK-1	PCK-5	PCK-10	PCK-1	PCK-5	PCK-10	AEPE	PCK-5
I	Census (1)	33.49	58.44	61.42	1.85	28.25	48.37	59.85	48.15
II	SSIM (1)	51.93	69.58	71.58	2.18	31.48	51.65	38.62	62.61
III	SSIM (1) + f-b (2)	52.59	70.78	72.78	2.12	31.86	52.13	35.79	64.48
IV	Warp-superv. (3)	38.50	59.60	66.21	2.36	33.28	54.47	25.04	78.60
V	Warp-superv. + f-b (2)	45.62	71.36	75.92	2.50	**36.04**	57.13	23.10	79.64
VI	**WarpC** ((8) + (3))	50.61	**78.61**	**82.94**	**2.51**	35.92	**57.44**	**21.00**	**83.24**
VII	**WarpC** + SSIM	**54.92**	75.65	78.04	2.43	35.01	56.44	26.01	74.64
VIII	Supervised	38.83	72.42	77.34	2.15	32.52	53.88	37.91	56.15
IX	**WarpC** + Sup. ft.	**56.68**	**81.33**	**84.76**	**2.41**	**34.67**	**55.89**	**22.78**	**78.19**

Table 3. Analysis and comparison of learning objectives.

4.2. Geometric matching

Here, we train the recent GLU-Net [50] and RANSAC-Flow [41] architectures with our unsupervised learning approach and compare them against state-of-the-art dense geometric matching methods.

Experimental set-up for GLU-Net: We follow the training procedure explained in Sec. 4.1 and refer to the resulting model as WarpC-GLU-Net. The original GLU-Net [50] is trained using solely the warp-supervision (3) on a different training set. For fair comparison, we also report results of our altered GLU-Net architecture when trained on MegaDepth with our warp distribution. This corresponds to setting (IV) in Tab. 3, which we here call GLU-Net*.

Experimental set-up for RANSAC-Flow: We addi-

	MegaDepth [23]				RobotCar [29, 20]			
	PCK-1	PCK-3	PCK-5	PCK-10	PCK-1	PCK-3	PCK-5	PCK-10
SIFT-Flow [25]	8.70	12.19	13.30	-	1.12	8.13	16.45	-
NCNet [39]	1.98	14.47	32.80	-	0.81	7.13	16.93	-
DGC-Net [32]	3.55	20.33	32.28	-	1.19	9.35	20.17	-
GLU-Net [50, 49]	21.58	52.18	61.78	69.81	2.30	17.15	33.87	55.67
GLU-Net-GOCor [49]	37.28	61.18	68.08	74.39	2.31	17.62	35.18	57.26
GLU-Net*	38.50	59.60	60.33	66.21	2.36	17.18	33.28	54.47
WarpC-GLU-Net	50.61	73.80	78.61	82.94	**2.51**	**18.59**	**35.92**	**57.44**
RANSAC-Flow [41]	52.60	83.46	86.80	88.80	2.09	15.94	31.61	53.06
WarpC-RANSAC-Flow	**53.77**	**84.23**	**88.18**	**90.53**	2.29	17.23	34.42	56.12

Table 4. State-of-the-art comparison for geometric matching.

tionally use our unsupervised strategy to train RANSAC-Flow [41]. In the original work [41], the network is trained on MegaDepth [23] image pairs that are coarsely pre-aligned using feature matching and Ransac. Training is separated into three stages. First, the network is trained using the SSIM loss (1), which is further combined with the forward-backward consistency loss (2) in the second stage. In the last stage, a matchability mask is also trained, by weighting the previous losses with the predicted mask and including a mask regularization term. For our WarpC-RANSAC-Flow, we also follow a three-step training using the same training pairs. As for the WarpC-GLU-Net training, we add our visibility mask $\widehat{V}$ in the second training stage. In the third stage, we train the matchability mask by simply replacing $\widehat{V}$ in (8) with the predicted mask, and adding the same mask regularizer as in RANSAC-Flow.

Results: In Tab. 4, we report results on MegaDepth and RobotCar. Note that we only compare to methods that do not finetune on the test set. Our approach WarpC-GLU-Net outperforms the original GLU-Net and baseline GLU-Net* by a large margin at all PCK thresholds. Our proposed unsupervised objective enables the network to handle the large and complex 3D motions present in real image pairs, as evidenced in Fig. 5, top. Our unsupervised approach WarpC-RANSAC-Flow also achieves a substantial improvement compared to RANSAC-Flow. Importantly, WarpC-RANSAC-Flow shows much better generalization capabilities on RobotCar. The poorer generalization of photometric-based objectives, such as SSIM [54] here, further supports our findings in Sec. 4.1. Interestingly, training the matchability branch of RANSAC-Flow with our objective results in drastically more accurate mask predictions. This is visualized in Fig. 5, middle, where our approach WarpC-RANSAC-Flow effectively identifies unreliable matching regions such as the sky (in red), whereas RANSAC-Flow, trained with the SSIM loss, is incapable of discarding the sky and field as unreliable.

4.3. Semantic matching

Finally, we evaluate our approach for the task of semantic matching by training SemanticGLU-Net [50], a version of GLU-Net specifically designed for semantic images, which includes multi-resolution features and NC-Net [39].

Experimental set-up: Following [37, 3], we only fine-tune a pre-trained network on semantic correspondence data.

Methods	Features	TSS [46]				PF-Pascal [9]	
		FG3DCar	JODS	Pascal	Avg.	$\alpha=0.05$	$\alpha=0.1$
CNNGeo [37]	ResNet-101	90.1	76.4	56.3	74.4	41.0	69.5
WeakAlign [37]	ResNet-101	90.3	76.4	56.5	74.4	49.0	75.8
RTNs [18]	ResNet-101	90.1	78.2	63.3	77.2	55.2	75.9
PARN [17]	ResNet-101	89.5	75.9	71.2	78.8	-	-
NC-Net [39]	ResNet-101	94.5	81.4	57.1	77.7	-	78.9
DCCNet [12]	ResNet-101	93.5	82.6	57.6	77.9	55.6	**82.3**
DHPF [34]	ResNet-101	-	-	-	-	56.1	82.1
SAM-Net [19]	VGG-19	96.1	82.2	67.2	81.8	60.1	80.2
GLU-Net [50]	VGG-16	93.2	73.3	71.1	79.2	42.2	69.1
GLU-Net-GOCor [49]	VGG-16	94.6	77.9	77.7	83.4	36.6	56.8
SemanticGLU-Net [50]	VGG-16	94.4	75.5	78.3	82.8	46.0	70.6
WarpC-SemanticGLU-Net	VGG-16	**97.1**	**84.7**	**79.7**	**87.2**	**62.1**	81.7

Table 5. State-of-the-art comparison for semantic matching.

Specifically, we start from the SemanticGLU-Net weights provided by the authors, which are trained with warp-supervision without using any correspondences from flow annotations. We finetune this network on the PF-PASCAL training set [9], which consists of 20 object categories, using our unsupervised loss (Sec. 3.5).

Datasets and metrics: We first evaluate on the test set of **PF-Pascal** [9]. In line with [10], we report the PCK with a pixel threshold equal to $\alpha \cdot \max(h_q, w_q)$, where h_q and w_q are the dimensions of the query image and $\alpha = (0.05, 0.1)$. To demonstrate generalization capabilities, we also validate our trained model on **TSS** [46], which provides dense flow field annotations for the foreground object in each pair. We report the PCK for $\alpha = 0.05$. We also provide results on PF-Willow [8] and SPair-71K [33] in suppl. K.3.

Results: Results are reported in Tab. 5. Our approach WarpC-SemanticGLU-Net sets a new state-of-the-art on TSS by obtaining a remarkable improvement compared to previous works. On the PF-Pascal dataset, our method ranks first for the small threshold $\alpha = 0.05$ with a substantial 2% increase compared to second best method. It obtains marginally lower PCK (0.6%) than DCCNet [12] for $\alpha = 0.1$, but the later approach employs a much deeper feature backbone, beneficial on semantic images. Nevertheless, our unsupervised fine-tuning provides 16% and 11.1% gain, for each threshold respectively, over the baseline, demonstrating that our objective effectively copes with the radical appearance changes encountered in the semantic matching task. A visual example is shown in Fig. 5 bottom.

5. Conclusion

We propose an unsupervised learning objective for dense correspondences, particularly suitable for scenarios with large changes in appearance and geometry. From a real image pair, we construct an image triplet and design a regression loss based on the flow-constraints existing between the triplet. When integrated into three recent dense correspondence networks, our approach outperforms state-of-the-art for multiple geometric and semantic matching datasets.

Acknowledgements: This work was supported by the ETH Zürich Fund (OK), a Huawei Gift, the ETH Future Computing Laboratory (EFCL) financed by a gift from Huawei Technologies, Amazon AWS, and an Nvidia GPU grant.

References

[1] Vassileios Balntas, Karel Lenc, Andrea Vedaldi, and Krystian Mikolajczyk. Hpatches: A benchmark and evaluation of handcrafted and learned local descriptors. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 3852–3861, 2017. 6

[2] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. In *BMVC*, 2014. 6

[3] 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. 1, 2, 8

[4] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Häusser, Caner Hazirbas, Vladimir Golkov, Patrick van der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 2758–2766, 2015. 1, 2

[5] Mihai Dusmanu, Ignacio Rocco, Tomás Pajdla, Marc Pollefeys, Josef Sivic, Akihiko Torii, and Torsten Sattler. D2-net: A trainable CNN for joint description and detection of local features. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 8092–8101, 2019. 1

[6] 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

[7] 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

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

[9] 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. 2, 8

[10] 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. 1, 8

[11] Berthold K. P. Horn and Brian G. Schunck. "determining optical flow": A retrospective. *Artif. Intell.*, 59(1-2):81–87, 1993. 2

[12] 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. 2, 8

[13] Tak-Wai Hui, Xiaoou Tang, and Chen Change Loy. Liteflownet: A lightweight convolutional neural network for optical flow estimation. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 8981–8989, 2018. 1

[14] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1647–1655. IEEE Computer Society, 2017. 1

[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] Joel Janai, Fatma Güney, Anurag Ranjan, Michael J. Black, and Andreas Geiger. Unsupervised learning of multi-frame optical flow with occlusions. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XVI*, pages 713–731, 2018. 2, 5

[17] 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. 2, 8

[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. 8

[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. 1, 2, 8

[20] Måns Larsson, Erik Stenborg, Lars Hammarstrand, Marc Pollefeys, Torsten Sattler, and Fredrik Kahl. A cross-season correspondence dataset for robust semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 9532–9542, 2019. 2, 6, 8

[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

[22] Xinghui Li, Kai Han, Shuda Li, and Victor Prisacariu. Dual-resolution correspondence networks. 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

[23] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 2041–2050, 2018. 2, 6, 8

[24] Jing Liao, Yuan Yao, Lu Yuan, Gang Hua, and Sing Bing Kang. Visual attribute transfer through deep image analogy. *ACM Trans. Graph.*, 36(4), July 2017. 1

[25] Ce Liu, Jenny Yuen, and Antonio Torralba. SIFT flow: Dense correspondence across scenes and its applications. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(5):978–994, 2011. 1, 8

[26] Liang Liu, Jiangning Zhang, Ruifei He, Yong Liu, Yabiao Wang, Ying Tai, Donghao Luo, Chengjie Wang, Jilin Li, and Feiyue Huang. Learning by analogy: Reliable supervision from transformations for unsupervised optical flow estimation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 6488–6497, 2020. 2

[27] Pengpeng Liu, Irwin King, Michael R. Lyu, and Jia Xu. Ddflow: Learning optical flow with unlabeled data distillation. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 8770–8777, 2019. 2

[28] Pengpeng Liu, Michael R. Lyu, Irwin King, and Jia Xu. Selflow: Self-supervised learning of optical flow. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 4571–4580, 2019. 2

[29] Will Maddern, Geoff Pascoe, Chris Linegar, and Paul Newman. 1 Year, 1000km: The Oxford RobotCar Dataset. *The International Journal of Robotics Research (IJRR)*, 36(1):3–15, 2017. 2, 6, 8

[30] N. Mayer, Eddy Ilg, Philip Häusser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4040–4048, 2016. 2

[31] Simon Meister, Junhwa Hur, and Stefan Roth. UnFlow: Unsupervised learning of optical flow with a bidirectional census loss. In *AAAI*, New Orleans, Louisiana, Feb. 2018. 1, 2, 3, 5, 7

[32] 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, 2, 3, 8

[33] Juhong Min, Jongmin Lee, Jean Ponce, and Minsu Cho. Spair-71k: A large-scale benchmark for semantic correspondence. *CoRR*, abs/1908.10543, 2019. 8

[34] 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, 8

[35] Zhe Ren, Junchi Yan, Bingbing Ni, Bin Liu, Xiaokang Yang, and Hongyuan Zha. Unsupervised deep learning for optical flow estimation. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 1495–1501, 2017. 1, 2

[36] 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, 3, 6

[37] 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, 8

[38] Ignacio Rocco, Relja Arandjelovic, and Josef Sivic. Efficient neighbourhood consensus networks via submanifold sparse convolutions. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part IX*, pages 605–621, 2020. 2

[39] 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. 2, 8

[40] 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

[41] Xi Shen, François Darmon, Alexei A Efros, and Mathieu Aubry. Ransac-flow: generic two-stage image alignment. In *16th European Conference on Computer Vision*, 2020. 2, 3, 6, 7, 8

[42] Abhinav Shrivastava, Tomasz Malisiewicz, Abhinav Gupta, and Alexei A. Efros. Data-driven visual similarity for cross-domain image matching. *ACM Transaction of Graphics (TOG) (Proceedings of ACM SIGGRAPH ASIA)*, 30(6), 2011. 1

[43] Patrice Y. Simard, Dave Steinkraus, and John C. Platt. Best practices for convolutional neural networks applied to visual document analysis. In *Proceedings of the Seventh International Conference on Document Analysis and Recognition - Volume 2*, ICDAR '03, page 958, USA, 2003. IEEE Computer Society. 6

[44] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27, pages 568–576. Curran Associates, Inc., 2014. 1

[45] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and

cost volume. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 8934–8943, 2018. 1

[46] 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. 2, 8

[47] Georgi Tinchev, Shuda Li, Kai Han, David Mitchell, and Rigas Kouskouridas. Xresolution correspondence networks. *CoRR*, abs/2012.09842, 2020. 2

[48] Prune Truong, Stefanos Apostolopoulos, Agata Mosinska, Samuel Stucky, Carlos Ciller, and Sandro De Zanet. Glampoints: Greedily learned accurate match points. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10731–10740, 2019. 1

[49] 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, 8

[50] 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, 3, 6, 7, 8

[51] 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

[52] Yang Wang, Peng Wang, Zhenheng Yang, Chenxu Luo, Yi Yang, and Wei Xu. Unos: Unified unsupervised optical-flow and stereo-depth estimation by watching videos. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 8071–8081, 2019. 2

[53] Yang Wang, Yi Yang, Zhenheng Yang, Liang Zhao, and Wei Xu. Occlusion aware unsupervised learning of optical flow. *CoRR*, abs/1711.05890, 2017. 2, 3, 5

[54] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.*, 13(4):600–612, 2004. 2, 3, 7, 8

[55] Olivia Wiles, Sébastien Ehrhardt, and Andrew Zisserman. D2D: learning to find good correspondences for image matching and manipulation. *CoRR*, abs/2007.08480, 2020. 1, 2

[56] 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

[57] Jason J. Yu, Adam W. Harley, and Konstantinos G. Derpanis. Back to basics: Unsupervised learning of optical flow via brightness constancy and motion smoothness. In *Computer Vision - ECCV 2016 Workshops - Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III*, pages 3–10, 2016. 1, 2, 3

[58] 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

[59] Tinghui Zhou, Yong Jae Lee, Stella X. Yu, and Alexei A. Efros. Flowweb: Joint image set alignment by weaving consistent, pixel-wise correspondences. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 1191–1200, 2015. 2