Semi-Supervised Learning of Semantic Correspondence with Pseudo-Labels

Jiwon Kim[1*]   Kwangrok Ryoo[1*]   Junyoung Seo[1*]   Gyuseong Lee[1*]
Daehwan Kim[2]   Hansang Cho[2]   Seungryong Kim[1†]
[1]Korea University, Seoul, Korea [2]Samsung Electro-Mechanics, Suwon, Korea

{naancoco,kwangrok21,se780,jpl358,seungryong_kim}@korea.ac.kr

{daehwan85.kim,hansang.cho}@samsung.com

Abstract

Establishing dense correspondences across semantically similar images remains a challenging task due to the significant intra-class variations and background clutters. Traditionally, a supervised learning was used for training the models, which required tremendous manually-labeled data, while some methods suggested a self-supervised or weakly-supervised learning to mitigate the reliance on the labeled data, but with limited performance.

In this paper, we present a simple, but effective solution for semantic correspondence that learns the networks in a semi-supervised manner by supplementing few ground-truth correspondences via utilization of a large amount of confident correspondences as pseudo-labels, called SemiMatch. Specifically, our framework generates the pseudo-labels using the model's prediction itself between source and weakly-augmented target, and uses pseudo-labels to learn the model again between source and strongly-augmented target, which improves the robustness of the model. We also present a novel confidence measure for pseudo-labels and data augmentation tailored for semantic correspondence. In experiments, SemiMatch achieves state-of-the-art performance on various benchmarks.

1. Introduction

Establishing dense correspondences across semantically similar images, depicting different instances of the same object or scene category, can facilitate many Computer Vision applications such as semantic segmentation [42,52,61], object detection [34], or image editing [25,33]. Unlike classical dense correspondence problems such stereo matching or optical flow [21,58], semantic correspondence poses additional challenges from large intra-class appearance and ge-

(a) Supervised loss   (b) Self-supervised loss

(c) Unsupervised loss (Ours)   (d) Semi-supervised loss (Ours)
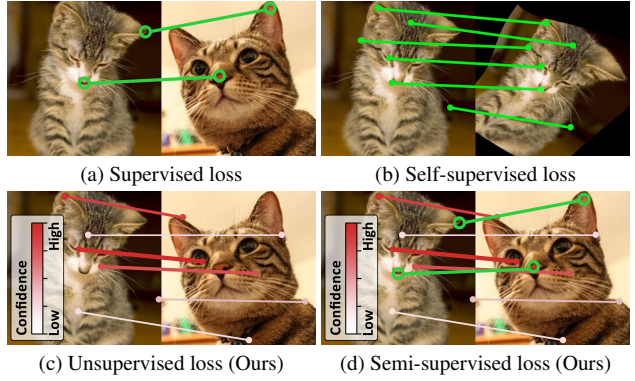
Figure 1. **Comparison of semantic correspondence methods in terms of different formulation of supervision.** Conventional methods leverage either (a) supervised loss using sparse ground-truth keypoint matches [16,43,45] or (b) self-supervised loss using synthetic flow field with random geometric parameters [40,48,63]. Unlike them, we present (c) unsupervised loss using pseudo-labels from a matching probability and (d) semi-supervised loss using both sparse ground-truth keypoints and confident pseudo-labels.

ometric variations [10,16,23].

Although formulated in various ways, most recent approaches [9,36,40,41,43,45,48,49,51,54,63,64] addressed these challenges by carefully designing deep neural networks, such as CNNs [36,40,43,45,48,49,51,54,63,64] or Transformers [9,59], based models. The most straightforward way to formulate a mapping function is to use ground-truth correspondences between the image pairs. Recent approaches [9,70], including conventional approaches, have been formulated in a *supervised* fashion (Fig. 1a). However, ground-truth keypoint pairs on the most standard benchmarks [15,44] can be the inherent bottleneck [10,16,43,45] because they are annotated subjectively and sparsely.

To alleviate the reliance on the ground-truth data, some methods [40,48,63,64] presented a *self-supervised* learning framework (Fig. 1b), using synthetic geometric warps of an image to generate a synthetic image pair. Although it turns out that it is an appealing alternative, using synthetic image pairs cannot account for extreme intra-class appear-

---

ance variations in semantic correspondence [48, 55]. On the other hand, some other methods [20, 23, 25, 45] presented a *weakly-supervised* learning framework that casts this task as a feature reconstruction between the images, but the loss function often fails to explain the correspondences across severely different instances among the same class.

On the other hand, most recent approaches in image classification task [3, 4, 19, 27, 57] have been popularly formulated in a *semi-supervised* learning framework, which enables learning the model on a large amount of *unlabeled* data with a few *labeled* data, and showed outstanding performance. Most recent trends of semi-supervised learning [3, 4, 19, 27, 57] integrate consistency regularization [2] and pseudo-labeling [29]. For instance, FixMatch [57] first generates a pseudo-label using the model's prediction on weakly-augmented unlabeled data and then encourages the prediction from strongly-augmented unlabeled data to follow the pseudo-label with confidence thresholding. This learning framework has become a promising solution to mitigate the reliance on large labeled data [18, 38, 57], but directly applying these techniques to semantic correspondence is challenging in that learning the matching networks requires pixel-level pseudo-labels.

In this paper, we present a novel *semi-supervised* learning framework, called SemiMatch, that generates pixel-level pseudo-labels using the model's prediction itself between source and *weakly-augmented* target and then encourages the model to predict the pseudo-label again between source and *strongly-augmented* target, as illustrated in Fig. 2. To account for the observation that all of the pseudo-labels may not help to boost performance, we introduce a novel confidence measure for pseudo-labels by considering object-centric foreground, forward-backward consistency, and uncertainty of probability itself. Tailored for semantic correspondence, we also present a matching-specialized augmentation that exploits a keypoint-aware cutout, helping the network to learn distinctive feature representations around the keypoints.

We evaluate our method on several benchmarks [15, 44]. Experimental results on various benchmarks show that using our novel loss function with sparsely-supervised loss function consistently improves the performance compared to the latest methods for semantic correspondence. We also provide an extensive ablation study to validate and analyze components in our learning framework.

2. Related works

Semantic Correspondence. The objective of semantic correspondence [10, 15, 24, 49] is to find correspondences across semantically similar images. Recent methods [9, 10, 16, 43, 45, 70] showed great progress by supervised loss, but they are limited by the availability of sparse ground-truth annotation. On the other hand, some methods [40, 48, 63, 64]

address the aforementioned limitations by training the network in a self-supervised manner, relying on synthetic warps of real images, without ground-truth flow and [32] presents a teacher-student model where the student model exploits the teacher model's generalized knowledge learned from synthetic data. Their improvement is mainly due to self-supervised training data, but synthetic geometric deformation cannot model intra-class appearance variations and realistic scene generation. Another alternative is to use a training loss that only requires a weak-level of supervision, given for each image pair as either positive (the same class) or negative (different class), as in [20, 45, 51]. While it [45] can achieve higher performance than strongly-supervised methods, it has fundamental limitations to supervise locating matches. Unlike the methods above, we present for the first time a semi-supervised learning framework which overcomes the lack of labeled ground-truth keypoints by utilizing a large amount of confident correspondences as pseudo-labels.

Semi-Supervised Learning. The most recent methods for semi-supervised learning have followed two trends; pseudo-labeling and consistency regularization. Pseudo-labeling [1, 29, 47, 56, 67, 71] encourages a model to follow the pseudo-label from the model's prediction itself closely related to entropy minimization [13] where the model's predictions are encouraged to be low-entropy (i.e., high-confidence) on unlabeled data. On the other hand, consistency regularization [2, 53, 62, 67] encourages the model to produce the same prediction when perturbations are applied to the input or the model. Consistency regularization-based methods, enforcing invariant representations across augmentations, rely heavily on the usage of strong data augmentation, so it is important which strong augmentations are used and how strong they are. Very recently, some state-of-the-art methods [3, 19, 27, 57, 67] combine pseudo-labeling and consistency regularization by a confidence-based strategy and separate weak and strong augmentations, but there still remains the problem of ignoring a large amount of unlabeled data due to reliance on fixed high thresholding to compute the unsupervised loss. Most aforementioned methods have focused on solving an image-level task, e.g., image classification, but methods for pixel-level tasks such as semantic correspondence are done limitedly and cannot be applied directly.

Uncertainty Estimation. Predicting uncertainty in the field of Computer Vision has been widely explored [7, 26, 28, 37], even before the expansion of deep learning. There are two approaches in uncertainty estimation: empirical and predictive. The former method [5, 14] approximates the uncertainty by sampling a finite number of weight configurations for a given network and computing the mean and variance of the predictions. In the latter method [46], a network
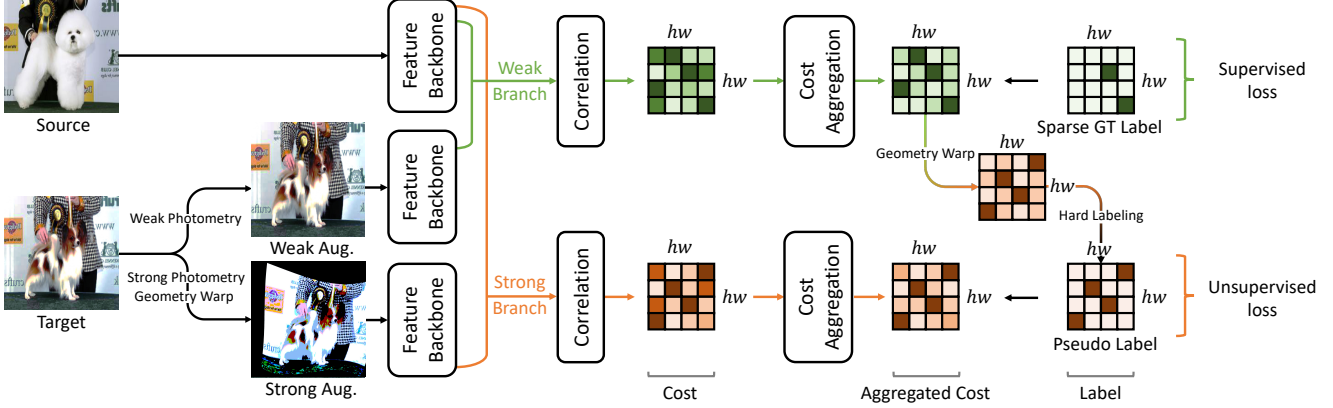
Figure 2. **Overview of our semi-supervised learning framework for semantic correspondence.** SemiMatch first augments the target image into a *weakly-augmented* one and a *strongly-augmented* one. The backbone matching networks extract features from these images, compute correlation cost, followed by aggregation. The aggregated cost between source and weakly-augmented target images is transformed by the same geometry warp used in strong augmentation and sharpened to generate a *pseudo-label*. The overall loss function is composed of supervised loss which uses the sparse ground-truth label and unsupervised loss which uses the pseudo-label.

is trained to infer the mean and variance of the distribution by design.

In semi-supervised learning in image classification, there were some approaches, such as Dash [68] and Flex-Match [69], to measure the uncertainty of pseudo-labels by applying a different threshold to each sample, unlike FixMatch [57] which assumes that the samples exceed the hand-crafted confidence threshold. In contrast to the existing semi-supervised method [57,68,69] which assumes that all pseudo-labels are certain without considering the uncertainty of the generated pseudo-label, we improve the performance by considering the uncertainty of the pseudo-label.

3. Methodology

3.1. Motivation

Given a pair of images, i.e., source I_s and target I_t, which represent semantically similar images, the goal of semantic correspondence is to establish matches between the two images at each pixel. To achieve this, most predominant methods with CNNs [36, 40, 43, 45, 48, 49, 51, 54, 63, 64] consist of two steps, including feature extraction and cost aggregation. First of all, feature extraction networks extract a feature $F \in \mathbb{R}^{h \times w \times d}$, where $h \times w$ is the spatial resolution and d is the channels. The similarities between feature maps, called cost volume, can be estimated by $\mathcal{C}(i,j) = F_t(i)^T F_s(j)$, where $i \in \{1,..,h_t w_t\}$ and $j \in \{1,..,h_s w_s\}$ However, initial cost volume itself is vulnerable to ambiguous, repetitive, or textureless matches. To disambiguate these, recent methods [9, 31, 36, 41] employ the cost aggregation networks for refining the initial matching similarities to achieve the aggregated cost $\mathcal{C}'(i,j)$.

Although existing methods [9, 31, 36, 41, 70] can be formulated in various ways, their outputs $\mathcal{C}$ or $\mathcal{C}'$ can be con-

sidered as a *matching probability* through a simple SoftMax function [30] such that $P(i) = p(I_s, I_t(i); \theta) \in \mathbb{R}^{h_s w_s \times 1}$ with the network parameters θ, defined across all the points in I_s for point i in I_t. Learning such networks in a *supervised* manner requires manually annotated ground-truth correspondences P_{GT}, which are extremely labor-intensive and involves subjectivity [9, 45, 70]. Thus, in semantic correspondence, only sparsely-annotated ground-truths are available, and the supervised loss function is defined such that

$$\mathcal{L}_{\mathrm{sup}} = \sum_i c(i)\mathcal{D}(p(I_s, I_t(i); \theta), P_{\mathrm{GT}}(i)), \qquad (1)$$

where $c(i)$ is a binary indicator for representing the existence of ground-truth $P_{\mathrm{GT}}(i)$, $P_{\mathrm{GT}}(i)$ is an one-hot vector form, and $\mathcal{D}(\cdot, \cdot)$ is the distance function, e.g., L2 distance [40] or cross-entropy [23, 51]. Due to the inherent nature of using sparse ground-truths, the distance function yields a limited performance [9, 70]. To alleviate the reliance on large ground-truth data, *self-supervised* learning methods [40, 48, 63, 64] have been popularly used, which generates synthetic matching pairs by applying a geometric warping on a single image. In specific, target image I_t (or source image I_s) is transformed by geometry warping operator $\mathcal{G}(\cdot; \phi)$ with a randomly-defined warping field ϕ, e.g., generated by affine [66] or thin-plate-spline (TPS) [6] transformation. The synthetic image pairs are then defined as I_t and $\mathcal{G}(I_t; \phi)$, and ϕ is used as a pseudo-label for them.

The self-supervised loss function is then defined as

$$\mathcal{L}_{\mathrm{self-sup}} = \sum_i \mathcal{D}(p(\mathcal{G}(I_t; \phi), I_t(i); \theta), P_\phi(i)). \qquad (2)$$

where P_ϕ is one-hot vector form of ϕ. Since this loss function does not require annotated ground-truths, and enables

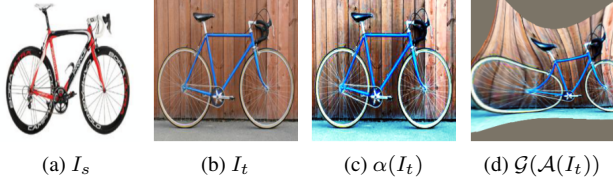| (a) I_s | (b) I_t | (c) $\alpha(I_t)$ | (d) $\mathcal{G}(\mathcal{A}(I_t))$ |

Figure 3. **Examples of augmented images in semi-supervised learning:** (a) source image I_s, (b) target image I_t (c) weakly-augmented target image $\alpha(I_t)$, and (d) strongly-augmented target image $\mathcal{G}(\mathcal{A}(I_t))$.

learning with dense labels, it can be an alternative to data hungry from sparse annotations [9, 70]. However, such synthetic image pairs cannot contain realistic appearance variations across the two images and model moving objects or occlusion, which limits the performance in semantic correspondence that often poses challenges by intra-class appearance and shape variations [30, 41].

3.2. Formulation

We present a novel *semi-supervised* learning framework, SemiMatch, for learning a matching model on a large amount of unlabeled pixels with few labeled pixels between source and target images. Following recent trends of semi-supervised learning in image classification [57, 67] that have not been applied in semantic correspondence yet, we present to extend the consistency regularization between two differently augmented instances from the same image [57] to semantic correspondence. They are commonly based on the assumption that when perturbations are applied to the input, the prediction should not change significantly. However, it is difficult to adapt the existing consistency regularization techniques [3, 4, 57] to semantic correspondence directly, predicting dense probabilities because they were designed for image classification task.

In this section, we study how to formulate the loss function for *unsupervised learning* within semantic correspondence framework that can be simultaneously used with sparsely supervised loss function for semi-supervised learning. First of all, as shown in Fig. 3 given source I_s and target I_t images, we build a triplet $\{I_s, \alpha(I_t), \mathcal{G}(\mathcal{A}(I_t))\}$, where $\alpha(\cdot)$ and $\mathcal{A}(\cdot)$ represent *weak* and *strong* photometric and geometric augmentations, respectively. Our key ingredient is to exploit the difference in difficulty levels of matching weak pairs and strong pairs, i.e., the weak pairs are easier to generate pseudo-labels for the strong pairs, inspired by [57]. However, without the geometric augmentation, the performance of trained models may be sub-optimal because the models may lack the robustness to geometric variations, which frequently occur in semantic correspondence. But if the geometric augmentation is applied simultaneously, direct consistency regularization cannot be formed,i.e., to perform consistency regularization on the geometrically warped image pairs, it is necessary to find the

Figure 4. **Comparison of matching difficulty between weak branch and strong branch.** The yellow and red lines show the result of the weak and strong branch, respectively. The green lines show when the weak branch and strong branch results are the same. Weak branch gives more accurate results.

pseudo labels in the misaligned image pair.

To overcome the aforementioned issues, we present a novel unsupervised loss function that jointly leverages photometric and geometric augmentations, defined such that

$$\mathcal{L}_{un-sup} = \sum_i m(i)\mathcal{D}(p(I_s, \mathcal{G}(\mathcal{A}(I_t(i)); \phi); \theta), Q(i)),$$

(3)

where $Q(i)$ is a pseudo-label defined as

$$Q(i) = \mathcal{G}(p(I_s, \alpha(I_t(i)); \theta); \phi),$$

(4)

which means the geometrically-warped matching probabilities between weak pairs with ϕ. And the same geometric warping applied to the target image, such that $\mathcal{G}(\mathcal{A}(I_t); \phi)$. Like previous consistency regularization methods [2], we generate pseudo-labels from better correspondences of the weak pairs than the strong pairs as shown in Fig. 4. By geometrically-warping the pseudo-labels, we can align matching probability for strong pairs, enabling the model to achieve robustness to strong augmentation as well.

In the following section, we will explain how to achieve the confidence $m(i)$ of pseudo-label.

3.3. Confidence of Pseudo-Label

Matching probabilities inferred from the weak branch are used to define pseudo-labels in the strong branch. However, incorrect pseudo-labels may hinder performance boosting, which is called confirmation bias [1] problem. To overcome this, some semi-supervised learning methods for image classification use simple thresholding of the probability value itself, e.g., FixMatch [57]. This is too a strong constraint in early iterations of training in that most false positives cannot be captured.

To alleviate this, we present three constraints to measure the confidence of pseudo-labels. First of all, we utilize a binary object mask of source and target images, which helps to limit the matching candidates within object-centric regions. In specific, since the pseudo-labels are defined at geometrically-warped regions, we define M_{mask} as a binary foreground object mask for target image I_t. Note that the object mask is synthesized using the min and max values of ground-truth keypoints.
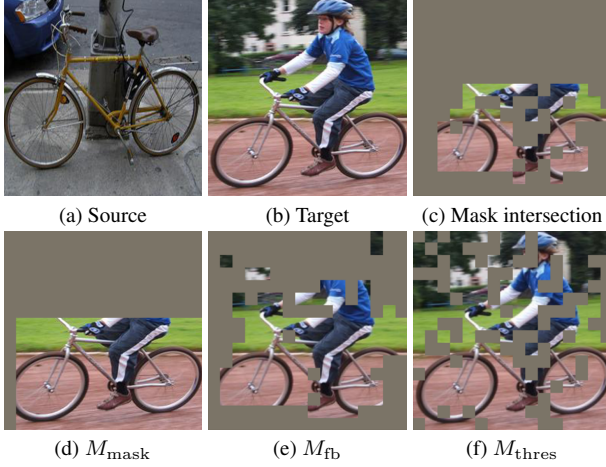
(a) Source   (b) Target   (c) Mask intersection

(d) M_{mask} (e) M_{fb} (f) M_{thres}

Figure 5. **Visualization of our confidence mask:** (a) source, (b) target, (c) mask intersection, (d) M_{mask}, (e) M_{fb}, and (f) M_{thres}. Sampled pixels by mask intersection belong to object-centric and discriminative image regions.

Secondly, we also utilize forward-backward consistency checking [35, 39, 60]. In specific, by estimating correspondences between $\{I_s, \alpha(I_t)\}$ in forward and backward directions simultaneously and checking the consistency between them, we determine the reliability of pseudo-labels based on the observation that if this consistency constraint is not satisfied, the points in $\alpha(I_t)$ are occluded at the matches in I_s (or vice-versa), and the estimated flow vector is incorrect. This constraint is denoted by M_{fb}.

Finally, unlike existing consistency regularization methods [57], using a few samples through high thresholding with a pre-defined scalar value τ, we present uncertainty-based weighting on the loss function itself. This lets a considerable amount of unlabeled data, especially at the early iteration of training phase, be used as pseudo-labels according to their different learning status. Specifically, we measure the uncertainty $u(i)$ of matching probabilities $P(i)$ as

$$u(i) = 1/\exp\left(\sum_j P(i,j)\log P(i,j)\right), \quad (5)$$

where $P(i,j)$ is j-th component of $P(i)$. $M_{\mathrm{thres}}(i)$ is then measured by $(1/u(i)) \odot 1\,(\max(P(i)) \geq \tau)$ to weight loss function depending on uncertainty where $\odot$ denotes Hadamard product.

Our final confidence m is determined by $\mathcal{G}(M_{\mathrm{mask}} \odot M_{\mathrm{fb}} \odot M_{\mathrm{thres}}; \phi)$. We visualize each mask component belonging to the intersection of the masks in Fig. 5.

3.4. Augmentation

Consistency regularization-based methods [2], enforcing invariant representations across augmentations, greatly depend on what kind of transformations are used for strongly-augmented images. Even though any kind of augmentation can be used in this framework, for semantic correspon-

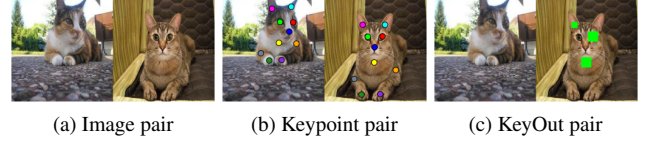(a) Image pair   (b) Keypoint pair   (c) KeyOut pair

Figure 6. **Visualization of KeyOut augmentation:** (a) image pair, (b) keypoints, and (c) keyOut applied to the target image.

dence, augmentation should be more focused on discriminative local parts of an object to infer the matches.

In this section, we propose a matching-specialized augmentation, called keypoint-guided CutOut (KeyOut), which cuts and removes boxes of a certain size around the keypoint location as shown in Fig. 6. It allows the model to learn to find the keypoint locations by integrating keypoint peripheral information.

3.5. Loss Functions

Finally, we propose two loss functions to train our model using different supervision, including *supervised* loss and *semi-supervised* regimes. As described above, any distance function $\mathcal{D}$ can be used for loss functions. In specific, following the common practice [40], L_{sup} is defined as the L2 distance such that

$$\mathcal{L}_{\mathrm{sup}} = \sum_i c(i)\|\xi(p(I_s, I_t(i); \theta)) - \xi(P_{\mathrm{GT}}(i))\|, \quad (6)$$

where $\xi(\cdot)$ is denoted as general max function including soft argmax and hard argmax. In addition, $L_{\mathrm{un-sup}}$ is formulated with the contrastive loss function [65] as

$$\mathcal{L}_{\mathrm{un-sup}} = -\sum_i m(i)\log\left(\frac{\exp\left(p(I_s(i'), G(i); \theta)/\gamma\right)}{\sum_j \exp(p(I_s(j), G(i); \theta)/\gamma)}\right), \quad (7)$$

where $G = \mathcal{G}(\mathcal{A}(I_t); \phi)$, i' and j represent the locations in source image, respectively, i' is determined as $\xi(Q(i))$, and γ is the temperature hyper-parameter.

Our total loss $\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{sup}} + \lambda\mathcal{L}_{\mathrm{un-sup}}$ where λ is a weight that is adaptively determined by the ratio between $\mathcal{L}_{\mathrm{sup}}$ and $\mathcal{L}_{\mathrm{un-sup}}$ such that $\lambda = \mathcal{L}_{\mathrm{sup}}^*/\mathcal{L}_{\mathrm{un-sup}}^*$, where $\mathcal{L}^*$ is the loss value itself and no back propagation happens.

3.6. Network Architecture

Our semi-supervised learning framework can be used in any deep networks for semantic correspondence [9, 22, 31, 36, 41, 43, 50, 51]. In this paper, we leverage the recent state-of-the-art network, especially focusing on cost aggregation stage, CATs [9] that explores global consensus among initial correlation map with the help of Transformer-based aggregator. By considering the outputs of the networks as matching probabilities, it directly leverages the proposed unsupervised loss, as well as supervised loss. Our overall architecture is shown in Fig. 2.

Method	Supervision	Learning signal	PF-PASCAL			PF-Willow			SPair-71k
			0.05	0.1	0.15	0.05	0.1	0.15	0.1
PF$_{HOG}$ [15]	None	-	31.4	62.5	79.5	28.4	56.8	68.2	-
CNNGeo$_{ResNet-101}$ [48]	Self-sup.	synthetic pairs	41.0	69.5	80.4	36.9	69.2	77.8	20.6
A2Net$_{ResNet-101}$ [55]			42.8	70.8	83.3	36.3	68.8	84.4	22.3
SF-Net$_{ResNet-101}$ [30]		bbox	53.6	81.9	90.6	46.3	74.0	84.2	-
Weakalign$_{ResNet-101}$ [49]	Weak-sup.	images	49.0	74.8	84.0	37.0	70.2	79.9	20.9
RTNs$_{ResNet-101}$ [23]			55.2	75.9	85.2	41.3	71.9	86.2	25.7
NC-Net$_{ResNet-101}$ [51]			54.3	78.9	86.0	33.8	67.0	83.7	20.1
DCC-Net$_{ResNet-101}$ [20]			55.6	82.3	90.5	43.6	73.8	86.5	-
DHPF$_{ResNet-101}$ [45]			56.1	82.1	91.1	50.2	_80.2_	_91.1_	37.3
SCNet$_{VGG-16}$ [16]	Sup.	keypoints	36.2	72.2	82.0	38.6	70.4	85.3	-
ANC-Net$_{ResNet-101-FCN}$ [31]			-	86.1	-	-	-	-	28.7
HPF$_{ResNet-101}$ [43]			60.1	84.8	92.7	45.9	74.4	85.6	28.2
DHPF$_{ResNet-101}$ [45]			75.7	90.7	95.0	49.5	77.6	89.1	37.3
CHMNet$_{ResNet-101}$ [41]			**80.1**	91.6	94.9	_52.7_	79.4	87.5	46.3
MMNet$_{ResNet-101}$ [70]			_77.6_	89.1	94.3	-	-	-	40.9
CATs$^{†}_{ResNet-101}$ [9]			67.5	89.1	94.9	46.6	75.6	87.5	42.4
CATs$_{ResNet-101}$ [9]			75.4	_92.6_	_96.4_	50.3	79.2	90.3	_49.9_
SemiMatch†	Semi-sup.	keypoints	75.0	91.7	95.6	47.4	76.3	88.2	43.0
SemiMatch			**80.1**	**93.5**	**96.6**	**54.0**	**82.1**	**92.1**	**50.7**

Table 1. **Quantitative evaluation on PF-PASCAL and PF-Willow [15] and SPair-71k [44].** Subscripts of each method's name indicate the feature backbone used. The best results in bold, and the second best results are underlined. CATs† means without fine-tuning feature.

Methods	aero.	bike	bird	boat	bott.	bus	car	cat	chai.	cow	dog	hors.	mbik.	pers.	plan.	shee.	trai.	tv	all
CNNGeo [48]	23.4	16.7	40.2	14.3	36.4	27.7	26.0	32.7	12.7	27.4	22.8	13.7	20.9	21.0	17.5	10.2	30.8	34.1	20.6
A2Net [55]	22.6	18.5	42.0	16.4	37.9	30.8	26.5	35.6	13.3	29.6	24.3	16.0	21.6	22.8	20.5	13.5	31.4	36.5	22.3
WeakAlign [49]	22.2	17.6	41.9	15.1	38.1	27.4	27.2	31.8	12.8	26.8	22.6	14.2	20.0	22.2	17.9	10.4	32.2	35.1	20.9
NC-Net [51]	17.9	12.2	32.1	11.7	29.0	19.9	16.1	39.2	9.9	23.9	18.8	15.7	17.4	15.9	14.8	9.6	24.2	31.1	20.1
HPF [43]	25.2	18.9	52.1	15.7	38.0	22.8	19.1	52.9	17.9	33.0	32.8	20.6	24.4	27.9	21.1	15.9	31.5	35.6	28.2
SCOT [36]	34.9	20.7	63.8	21.1	43.5	27.3	21.3	63.1	20.0	42.9	42.5	31.1	29.8	35.0	27.7	24.4	48.4	40.8	35.6
DHPF [45]	38.4	23.8	68.3	18.9	42.6	27.9	20.1	61.6	22.0	46.9	46.1	33.5	27.6	40.1	27.6	28.1	49.5	46.5	37.3
CHMNet [41]	49.1	33.6	64.5	_32.7_	44.6	47.5	43.5	57.8	21.0	61.3	54.6	43.8	35.1	**43.7**	38.1	_33.5_	70.6	55.9	46.3
MMNet [70]	43.5	27.0	62.4	27.3	40.1	50.1	37.5	60.0	21.0	56.3	50.3	41.3	30.9	19.2	30.1	33.2	64.2	43.6	40.9
CATs† [9]	46.5	26.9	69.1	24.3	44.3	38.5	30.2	62.7	15.9	53.7	52.2	46.7	32.7	35.2	32.2	31.2	68.0	49.1	42.4
CATs [9]	_52.0_	_34.7_	_72.2_	**34.3**	**49.9**	_57.5_	**43.6**	_66.5_	**24.4**	_63.2_	_56.5_	**52.0**	_42.6_	_41.7_	_43.0_	**33.6**	_72.6_	_58.0_	_49.9_
SemiMatch†	47.8	29.0	70.6	24.0	44.5	37.6	29.8	65.2	17.2	54.7	52.8	47.1	35.2	37.6	29.9	32.7	68.5	49.4	43.0
SemiMatch	**53.6**	**37.0**	**74.6**	32.3	_47.5_	**57.7**	42.4	**67.4**	_23.7_	**64.2**	**57.3**	_51.7_	**43.8**	40.4	**45.3**	33.1	**74.1**	**65.9**	**50.7**

Table 2. **Per-class quantitative evaluation on SPair-71k dataset [44].** The best results are in bold, and the second best results are underlined

4. Experiments

4.1. Implementation Details

In experiments, we evaluate our framework with the state-of-the-art network, CATs [9]. For a fair comparison, we use the same hyper-parameters and photometric augmentation lists following CATs [9]. For geometric transformation [30, 48, 63], we apply a combination of affine and thin-plate-spline with random transformation parameters in range [0,1] * t_{scale} (0.15 for affine, and 0.4 for tps). To generate matching-specialized augmentations, we add blur as in [8] and KeyOut on strong augmentation. We set $\gamma = 0.1$ and $\tau = 0.5$.

4.2. Experimental Settings

In this section, we conduct comprehensive experiments for semantic correspondence, by evaluating our framework through comparisons to state-of-the-art methods including HOG [15], CNNGeo [48], A2Net [55], SFNet [30], WeakAlign [49], SCNet [16], RTNs [23], NC-Net [51], DCC-Net [20], HPF [43], DHPF [45], SCOT [36], ANC-Net [31], CHM [41], CATs [9], MMNet [70].

Dataset. We extensively conduct experiments on three popular benchmarks for semantic correspondence: PF-PASCAL [15], PF-WILLOW [15], SPair-71k [44]. PF-PASCAL contains 1,351 semantically related image pairs from 20 categories of the PASCAL VOC dataset [12]. PF-Willow contains 900 image pairs from 4 categories. SPair-
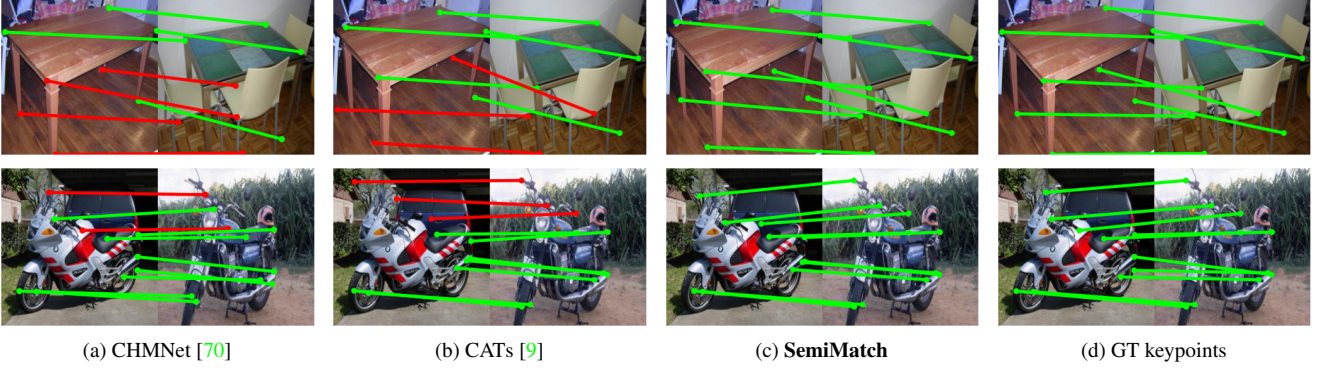
| | (a) CHMNet [70] | (b) CATs [9] | (c) **SemiMatch** | (d) GT keypoints |

Figure 7. **Qualitative results on PF-PASCAL [15].** (a) CHMNet [41] (b) CATs [9] (c) SemiMatch, and (d) sparse GT keypoints.

Component		PF-PASCAL		PF-Willow	
		0.05	0.1	0.05	0.1
(I)	SemiMatch	80.1	93.5	54.0	82.1
(II)	(I) w/o M_{mask}	<u>73.9</u>	<u>92.0</u>	**51.4**	**79.7**
(III)	(I) w/o M_{fb}	78.7	93.1	52.9	81.9
(IV)	(I) w/o M_{thres}	**72.3**	**91.3**	<u>52.1</u>	<u>81.5</u>

Table 3. **Ablation study of mask elements.**

71k consists of total 70,958 image pairs in 18 categories with diverse view-point and scale variations. We used the same split proposed in [44].

Evaluation Metric. Following the standard experimental protocol [15, 44], we use the percentage of correct keypoint (PCK@α), computed as the ratio of estimated keypoints within the threshold from ground-truths to the total number of keypoints. Given a set of predicted and ground-truth keypoint pairs $\mathcal{K} = \{(k_{\mathrm{pred}}(m), k_{\mathrm{GT}}(m))\}$, PCK can be defined as $\mathrm{PCK}(\mathcal{K}) = \frac{1}{M} \sum_m^M d\left(k_{\mathrm{pred}}(m), k_{\mathrm{GT}}(m)\right) \leq \alpha_k \cdot \max(H, W)$, where M is the number of keypoint pairs, $d(\cdot)$ is Euclidean distance; a threshold is scaled by $\alpha_k \cdot \max(H, W)$ in proportion to the larger portion of image for PF-PASCAL [15], and the object's bounding box for PF-Willow [15], and SPair-71k [44].

4.3. Matching Results

For fair comparisons with our baseline, CATs [9], and previous state-of-the-art methods, we exploit the same network architecture, ResNet-101 [17] pretrained on ImageNet [11]. As shown in Table 1, for PF-PASCAL [15], SemiMatch records state-of-the-art results with 80.1% PCK@0.05, 93.5% PCK@0.1 and 96.6% PCK@0.15. Without and with fine-tuning feature extraction backbone, SemiMatch outperforms CATs [9] by 7.5%/4.7% PCK@0.05, 2.6%/0.9% PCK@0.1, and 0.7%/0.2% PCK@0.15. It demonstrates the effectiveness of our semi-supervised framework with the confidence constraints of pseudo-labels and matching-specialized augmentation. Compared to CATs, it can operate more sensitively

at local regions through the significant performance improvement in PCK@0.05, which is the most strict matching criterion in PF-PASCAL. Experiments show that a large amount of pseudo-labels provide information on neighboring keypoints that cannot be provided by sparse keypoints. Generalization power of SemiMatch can be proven through the best performance in PF-willow for all PCKs by 1.3%, 1.9% and 1.0%, respectively compared to the previous state-of-the-art results. Finally, we also record the best performance with 50.7% PCK@0.1 even in SPair-71k [44] having large-scale variation. To show the effectiveness and robustness of our framework in detail, we compare per-class accuracy in Table 2 and our approach outperforms all state-of-the-art networks on 11 of the 18 classes. Our qualitative results are shown in Fig. 7.

4.4. Ablation Study

We conduct ablation analyses to investigate the effectiveness of components in our framework and also explore the effect of pseudo-labeling compared to CATs [9]. All experiments are conducted on PF-PASCAL dataset [15] and validated on PF-PASCAL and PF-Willow.

Effects of Confidence Mask. In Table 3, we evaluate each mask constraint in SemiMatch baseline (I) by removing each from the entire constraints. From (I) to (II), PCK decline shows that the network can learn more representative information of the object by foreground samples separated from the background. A large performance drop in (IV) compared to (III) shows that it is important to adjust a weighting on the loss function proportionally according to the uncertainty.

Effects of Keypoint-Based Augmentation. We also evaluate additional strong augmentation by CutOut, KeyOut with low probability and small box regions (KeyOut$_{\mathrm{weak}}$), and KeyOut with high probability and large box regions (KeyOut$_{\mathrm{strong}}$). As shown in Table 4, we can prove the effectiveness of our KeyOut, cutout regions based on keypoint locations by comparing (II) and (III)-(IV). Especially

(a) PCK@0.1 on PF-PASCAL   (b) PCK@0.05 PF-PASCAL   (c) PCK@0.1 on PF-Willow   (d) PCK@0.05 on PF-Willow
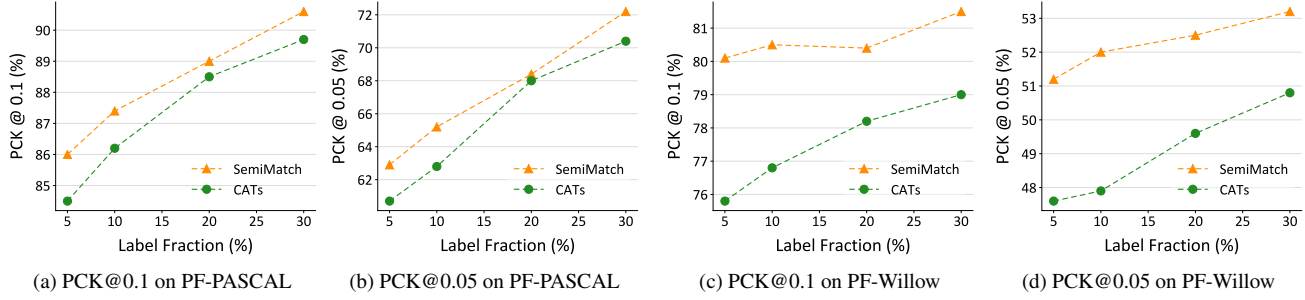
Figure 8. **PCK results of CATs [9] and SemiMatch on PF-PASCAL and PF-Willow with various label fractions:** (a) PCK@0.1 on PF-PASCAL, (b) PCK@0.05 on PF-PASCAL, (c) PCK@0.1 on PF-Willow, and (d) PCK@0.05 on PF-Willow.

Component		PF-PASCAL		PF-Willow	
		0.05	0.1	0.05	0.1
(I)	SemiMatch$_{Base}$	79.8	93.3	52.9	81.5
(II)	w/ CutOut$_{weak}$	79.8	93.2	51.8	81.1
(III)	w/ KeyOut$_{weak}$	**80.1**	**93.6**	**54.0**	**82.1**
(IV)	w/ KeyOut$_{strong}$	79.8	93.4	53.6	81.9

Table 4. **Ablation study of types of data augmentation.**

Warm-up	PF-PASCAL			PF-Willow		
	0.05	0.1	0.15	0.05	0.1	0.15
✓	**80.1**	**93.6**	96.6	**54.0**	82.1	92.1
✗	78.9	93.5	**96.8**	53.9	**82.3**	**92.7**

Table 5. **Ablation study of warm-up stage.**

Loss function	PF-PASCAL			PF-Willow		
	0.05	0.1	0.15	0.05	0.1	0.15
AEPE	78.9	92.7	96.4	53.6	**82.1**	**92.2**
Contrastive	**80.1**	**93.6**	**96.6**	**54.0**	**82.1**	92.1

Table 6. **Ablation study of loss function.**

on PF-Willow, (III) shows powerful generalization power compared to (I) and (II) by 1.1% and 2.2%.

Experiments with Label Fraction. We investigate the performance gap between CATs and SemiMatch according to the label fraction of PF-PASCAL dataset, referring to the percentage of data in entire image pairs. We conduct ablation experiments using labels of 5%, 10%, 20%, and 30% of the total dataset. Fig. 8 shows that SemiMatch is consistently ∼4.5% better than CATs in any label fraction setting on both PF-PASCAL and PF-Willow. Note that SemiMatch using 5%, 10%, 20%, and 30% of labels surpasses CATs using all labels by 0.9%, 1.3%, 1.2% and 2.3% PCK@0.1 for PF-Willow, respectively. Also, SemiMatch using 5%, 10%, 20%, and 30% of labels surpasses CATs using all labels by 0.9%, 1.7%, 2.2%, and 2.9% PCK@0.05 for PF-Willow.

Experiments with Warm-Up stage. The fundamental problem of pseudo-labeling is that in the early stages of the training, the model is hindered by incorrect pseudo-labels caused by parameter initialization, which is called confirmation bias [1]. Therefore, semi-supervised frameworks based on pseudo-labeling generally have a warm-up stage using only the labeled data. We evaluate the effect of warm-up stage results, which is 20 epochs of training with supervised loss. In experiments, there is no significant difference in performance with or without the warm-up, as shown in Table 5. It can be interpreted that our network explores the effectiveness of utilizing unlabeled data according to the model's learning status by uncertainty-based confidence measurement.

Comparison of Distance Function. A distance function for unsupervised loss in our semi-supervised framework can be defined as either AEPE or Contrastive loss. It should be noted that, SemiMatch shows better performance than CATs, our baseline, regardless of loss formulation. We experimentally observed that using Contrastive loss performs better than using AEPE as shown in Table 6.

5. Conclusion

In this paper, we have presented a novel semi-supervised learning framework, called SemiMatch, that exploits the pixel-level pseudo-label generated by source and weakly-augmented target to learn a model again by taking source and strongly-augmented target as input. We introduce a novel confidence measure for pseudo-labels to ignore incorrect pseudo-labels and augmentation tailored for semantic matching, exploiting keypoint locations, to learn the model to integrate keypoint peripheral information. We have shown that SemiMatch achieves state-of-the-art performance over the latest methods in several benchmarks.

References

[1] Eric Arazo, Diego Ortego, Paul Albert, Noel E O'Connor, and Kevin McGuinness. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *2020 International Joint Conference on Neural Networks*, pages 1–8. IEEE, 2020. 2, 4, 8

[2] Philip Bachman, Ouais Alsharif, and Doina Precup. Learning with pseudo-ensembles. *Advances in neural information processing systems*, 27:3365–3373, 2014. 2, 4, 5

[3] David Berthelot, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang, and Colin Raffel. Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring. *arXiv preprint arXiv:1911.09785*, 2019. 2, 4

[4] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin Raffel. Mixmatch: A holistic approach to semi-supervised learning. *arXiv preprint arXiv:1905.02249*, 2019. 2, 4

[5] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In *International Conference on Machine Learning*, pages 1613–1622. PMLR, 2015. 2

[6] FL Bookstein and WDK Green. A thin-plate spline and the decomposition of deformations. *Mathematical Methods in Medical Imaging*, 2:14–28, 1993. 3

[7] Andrés Bruhn and Joachim Weickert. A confidence measure for variational optic flow methods. In *Geometric Properties for Incomplete Data*, pages 283–298. Springer, 2006. 2

[8] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning, 2020. 6

[9] Seokju Cho, Sunghwan Hong, Sangryul Jeon, Yunsung Lee, Kwanghoon Sohn, and Seungryong Kim. Semantic correspondence with transformers. *arXiv preprint arXiv:2106.02520*, 2021. 1, 2, 3, 4, 5, 6, 7, 8

[10] Christopher B Choy, JunYoung Gwak, Silvio Savarese, and Manmohan Chandraker. Universal correspondence network. *arXiv preprint arXiv:1606.03558*, 2016. 1, 2

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 7

[12] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111(1):98–136, 2015. 6

[13] Yves Grandvalet, Yoshua Bengio, et al. Semi-supervised learning by entropy minimization. *CAP*, 367:281–296, 2005. 2

[14] Alex Graves. Practical variational inference for neural networks. *Advances in neural information processing systems*, 24, 2011. 2

[15] Bumsub Ham, Minsu Cho, Cordelia Schmid, and Jean Ponce. Proposal flow: Semantic correspondences from object proposals. *IEEE transactions on pattern analysis and machine intelligence*, 40(7):1711–1725, 2017. 1, 2, 6, 7

[16] Kai Han, Rafael S Rezende, Bumsub Ham, Kwan-Yee K Wong, Minsu Cho, Cordelia Schmid, and Jean Ponce. Scnet: Learning semantic correspondence. In *Proceedings of the IEEE international conference on computer vision*, pages 1831–1840, 2017. 1, 2, 6

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 7

[18] Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md Patwary, Mostofa Ali, Yang Yang, and Yanqi Zhou. Deep learning scaling is predictable, empirically. *arXiv preprint arXiv:1712.00409*, 2017. 2

[19] Zijian Hu, Zhengyu Yang, Xuefeng Hu, and Ram Nevatia. Simple: Similar pseudo label exploitation for semi-supervised classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15099–15108, 2021. 2

[20] Shuaiyi Huang, Qiuyue Wang, Songyang Zhang, Shipeng Yan, and Xuming He. Dynamic context correspondence network for semantic alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2010–2019, 2019. 2, 6

[21] Tak-Wai Hui, Xiaoou Tang, and Chen Change Loy. Liteflownet: A lightweight convolutional neural network for optical flow estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8981–8989, 2018. 1

[22] Sangryul Jeon, Dongbo Min, Seungryong Kim, Jihwan Choe, and Kwanghoon Sohn. Guided semantic flow. In *European Conference on Computer Vision*, pages 631–648. Springer, 2020. 5

[23] Seungryong Kim, Stephen Lin, Sangryul Jeon, Dongbo Min, and Kwanghoon Sohn. Recurrent transformer networks for semantic correspondence. *arXiv preprint arXiv:1810.12155*, 2018. 1, 2, 3, 6

[24] Seungryong Kim, Dongbo Min, Bumsub Ham, Sangryul Jeon, Stephen Lin, and Kwanghoon Sohn. Fcss: Fully convolutional self-similarity for dense semantic correspondence. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6560–6569, 2017. 2

[25] Seungryong Kim, Dongbo Min, Somi Jeong, Sunok Kim, Sangryul Jeon, and Kwanghoon Sohn. Semantic attribute matching networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12339–12348, 2019. 1, 2

[26] Claudia Kondermann, Daniel Kondermann, Bernd Jähne, and Christoph Garbe. An adaptive confidence measure for optical flows based on linear subspace projections. In *Joint Pattern Recognition Symposium*, pages 132–141. Springer, 2007. 2

[27] Chia-Wen Kuo, Chih-Yao Ma, Jia-Bin Huang, and Zsolt Kira. Featmatch: Feature-based augmentation for semi-supervised learning. In *European Conference on Computer Vision*, pages 479–495. Springer, 2020. 2

[28] Jan Kybic and Claudia Nieuwenhuis. Bootstrap optical flow confidence and uncertainty measure. *Computer Vision and Image Understanding*, 115(10):1449–1462, 2011. 2

[29] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, page 896, 2013. 2

[30] Junghyup Lee, Dohyung Kim, Jean Ponce, and Bumsub Ham. Sfnet: Learning object-aware semantic correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2278–2287, 2019. 3, 4, 6

[31] Shuda Li, Kai Han, Theo W Costain, Henry Howard-Jenkins, and Victor Prisacariu. Correspondence networks with adaptive neighbourhood consensus. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10196–10205, 2020. 3, 5, 6

[32] Xin Li, Deng-Ping Fan, Fan Yang, Ao Luo, Hong Cheng, and Zicheng Liu. Probabilistic model distillation for semantic correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7505–7514, 2021. 2

[33] Jing Liao, Yuan Yao, Lu Yuan, Gang Hua, and Sing Bing Kang. Visual attribute transfer through deep image analogy. *arXiv preprint arXiv:1705.01088*, 2017. 1

[34] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 1

[35] Pengpeng Liu, Irwin King, Michael R Lyu, and Jia Xu. Ddflow: Learning optical flow with unlabeled data distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8770–8777, 2019. 5

[36] Yanbin Liu, Linchao Zhu, Makoto Yamada, and Yi Yang. Semantic correspondence as an optimal transport problem. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4463–4472, 2020. 1, 3, 5, 6

[37] Oisin Mac Aodha, Ahmad Humayun, Marc Pollefeys, and Gabriel J Brostow. Learning a confidence measure for optical flow. *IEEE transactions on pattern analysis and machine intelligence*, 35(5):1107–1120, 2012. 2

[38] Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens Van Der Maaten. Exploring the limits of weakly supervised pretraining. In *Proceedings of the European conference on computer vision*, pages 181–196, 2018. 2

[39] Simon Meister, Junhwa Hur, and Stefan Roth. Unflow: Unsupervised learning of optical flow with a bidirectional census loss. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. 5

[40] Iaroslav Melekhov, Aleksei Tiulpin, Torsten Sattler, Marc Pollefeys, Esa Rahtu, and Juho Kannala. Dgc-net: Dense geometric correspondence network. In *2019 IEEE Winter Conference on Applications of Computer Vision*, pages 1034–1042. IEEE, 2019. 1, 2, 3, 5

[41] Juhong Min and Minsu Cho. Convolutional hough matching networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2940–2950, 2021. 1, 3, 4, 5, 6, 7

[42] Juhong Min, Dahyun Kang, and Minsu Cho. Hypercorrelation squeeze for few-shot segmentation. *arXiv preprint arXiv:2104.01538*, 2021. 1

[43] Juhong Min, Jongmin Lee, Jean Ponce, and Minsu Cho. Hyperpixel flow: Semantic correspondence with multi-layer neural features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3395–3404, 2019. 1, 2, 3, 5, 6

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

[45] 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 16*, pages 346–363. Springer, 2020. 1, 2, 3, 6

[46] David A Nix and Andreas S Weigend. Estimating the mean and variance of the target probability distribution. In *Proceedings of 1994 ieee international conference on neural networks*, volume 1, pages 55–60. IEEE, 1994. 2

[47] Hieu Pham, Zihang Dai, Qizhe Xie, and Quoc V Le. Meta pseudo labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11557–11568, 2021. 2

[48] Ignacio Rocco, Relja Arandjelovic, and Josef Sivic. Convolutional neural network architecture for geometric matching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6148–6157, 2017. 1, 2, 3, 6

[49] Ignacio Rocco, Relja Arandjelović, and Josef Sivic. End-to-end weakly-supervised semantic alignment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6917–6925, 2018. 1, 2, 3, 6

[50] Ignacio Rocco, Relja Arandjelović, and Josef Sivic. Efficient neighbourhood consensus networks via submanifold sparse convolutions. In *European Conference on Computer Vision*, pages 605–621. Springer, 2020. 5

[51] Ignacio Rocco, Mircea Cimpoi, Relja Arandjelovic, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Ncnet: Neighbourhood consensus networks for estimating image correspondences. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 1, 2, 3, 5, 6

[52] Michael Rubinstein, Armand Joulin, Johannes Kopf, and Ce Liu. Unsupervised joint object discovery and segmentation in internet images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1939–1946, 2013. 1

[53] Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. *Advances in neural information processing systems*, 29:1163–1171, 2016. 2

[54] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature

matching with graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4938–4947, 2020. 1, 3

[55] Paul Hongsuck Seo, Jongmin Lee, Deunsol Jung, Bohyung Han, and Minsu Cho. Attentive semantic alignment with offset-aware correlation kernels. In *Proceedings of the European Conference on Computer Vision*, pages 349–364, 2018. 2, 6

[56] Weiwei Shi, Yihong Gong, Chris Ding, Zhiheng MaXiaoyu Tao, and Nanning Zheng. Transductive semi-supervised deep learning using min-max features. In *Proceedings of the European Conference on Computer Vision*, pages 299–315, 2018. 2

[57] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *arXiv preprint arXiv:2001.07685*, 2020. 2, 3, 4, 5

[58] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8934–8943, 2018. 1

[59] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8922–8931, 2021. 1

[60] Narayanan Sundaram, Thomas Brox, and Kurt Keutzer. Dense point trajectories by gpu-accelerated large displacement optical flow. In *European conference on computer vision*, pages 438–451. Springer, 2010. 5

[61] Tatsunori Taniai, Sudipta N Sinha, and Yoichi Sato. Joint recovery of dense correspondence and cosegmentation in two images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4246–4255, 2016. 1

[62] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *arXiv preprint arXiv:1703.01780*, 2017. 2

[63] Prune Truong, Martin Danelljan, and Radu Timofte. Glunet: Global-local universal network for dense flow and correspondences. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6258–6268, 2020. 1, 2, 3, 6

[64] Prune Truong, Martin Danelljan, Luc Van Gool, and Radu Timofte. Gocor: Bringing globally optimized correspondence volumes into your neural network. *arXiv preprint arXiv:2009.07823*, 2020. 1, 2, 3

[65] Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3024–3033, 2021. 5

[66] Eric W Weisstein. Affine transformation. *https://mathworld. wolfram. com/*, 2004. 3

[67] Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V Le. Unsupervised data augmentation for consistency training. *arXiv preprint arXiv:1904.12848*, 2019. 2, 4

[68] Yi Xu, Lei Shang, Jinxing Ye, Qi Qian, Yu-Feng Li, Baigui Sun, Hao Li, and Rong Jin. Dash: Semi-supervised learning with dynamic thresholding. In *International Conference on Machine Learning*, pages 11525–11536. PMLR, 2021. 3

[69] Bowen Zhang, Yidong Wang, Wenxin Hou, Hao Wu, Jindong Wang, Manabu Okumura, and Takahiro Shinozaki. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. *arXiv preprint arXiv:2110.08263*, 2021. 3

[70] Dongyang Zhao, Ziyang Song, Zhenghao Ji, Gangming Zhao, Weifeng Ge, and Yizhou Yu. Multi-scale matching networks for semantic correspondence. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3354–3364, 2021. 1, 2, 3, 4, 6, 7

[71] Barret Zoph, Golnaz Ghiasi, Tsung-Yi Lin, Yin Cui, Hanxiao Liu, Ekin D Cubuk, and Quoc V Le. Rethinking pre-training and self-training. *arXiv preprint arXiv:2006.06882*, 2020. 2