Self-supervised Learning of Semantic Correspondence Using Web Videos

Donghyeon Kwon[1] Minsu Cho[1,2] Suha Kwak[1,2]
Dept. of CSE, POSTECH[1] Graduate School of AI, POSTECH[2]

Abstract

Existing datasets for semantic correspondence are often limited in terms of both the amount of labeled data and diversity of labeled keypoints due to the tremendous cost of manual correspondence labeling. To address this issue, we propose the first self-supervised learning framework that utilizes a large amount of web videos collected and annotated fully automatically. Our main motivation is that smooth changes between consecutive video frames allow to build accurate space-time correspondences with no human intervention. Hence, we establish space-time correspondences within each web video and leverage them for deriving pseudo correspondence labels between two distant frames of the video. In addition, we present a dedicated training strategy that facilitates stable training using web videos with such pseudo labels. Our experiments on public benchmarks demonstrated that the proposed method surpasses existing self-supervised learning models and that our self-supervised learning as pretraining for supervised learning improves performance substantially. Our codebase for web video crawling and pseudo label generation will be released public to promote future research.

1. Introduction

The task of semantic correspondence [9, 10, 40], *i.e.*, finding correspondences between images with intra-class variations as well as viewpoint changes, is a challenging problem in computer vision, and has interesting applications such as object discovery [3, 4, 51], few-shot segmentation [13, 21, 39], and visual localization [53, 56, 58]. Recently, learning-based methods have driven remarkable advances in this field [6, 11, 15, 17, 19, 24, 25, 34, 38, 42, 46, 47, 49, 60]. Despite their success, however, there is a critical obstacle to learning more robust matching: *the lack of large-scale densely annotated data for training*. Manual annotation of point-to-point correspondences is prohibitively costly since the number of image pairs and that of keypoints to match are both huge. Hence, existing datasets [2, 9, 10, 22, 41, 43, 57, 63, 69] are limited in terms of both the amount of labeled data and the diversity of annotated keypoints. This issue may be critical in particular for recent learning-based methods, which commonly rely on data-hungry models like deep neural networks.

In this paper, we tackle the issue of limited training data by learning with web-crawled videos. The reason for leveraging web-crawled videos is three-fold. First, videos are abundant and readily available on web repositories like YouTube, and thus allow us to construct a large-scale training dataset. Second, two temporally distant frames sampled from the same clip often capture non-trivial geometric/photometric variations of objects, and thus when augmented differently, they well simulate common inputs of the semantic correspondence task, *i.e.*, different images of the same class [41]. Third, while it is not straightforward to obtain correspondences between two distant frames, their intermediate frames in the clip bridge the gap with smooth changes, facilitating to compute reliable space-time correspondences over the frames without any supervision [6, 20, 26, 27, 31, 55, 61, 64, 65, 68]. Our strategy is thus to learn with random pairs of distant frames in automatically-crawled video clips while leveraging their space-time correspondences as pseudo labels.

However, using web videos for learning semantic correspondence introduces new challenges, such as failures of video retrieval and unreliable pseudo correspondence labels. Indeed, we need specialized methods for both dataset construction and learning using web videos. First, we develop an algorithm for collecting and annotating videos in a fully automatic manner, which is illustrated in Fig. 1. Our algorithm examines the retrieved videos by classifying their thumbnail images so that only videos with correct thumbnails are downloaded. Then each downloaded video is divided into clips with no abrupt transition that guarantee stable space-time correspondences. Also, our pseudo labeling pipeline detects outliers from the established correspondences to remove potentially incorrect pseudo labels.

Moreover, we present the first framework for learning semantic correspondence using such web videos as training data. The framework, depicted in Fig.2, leverages random pairs of distant frames along with their pseudo correspondence labels to train a network. In addition, a domain adversarial learning strategy [8] is employed to close the domain

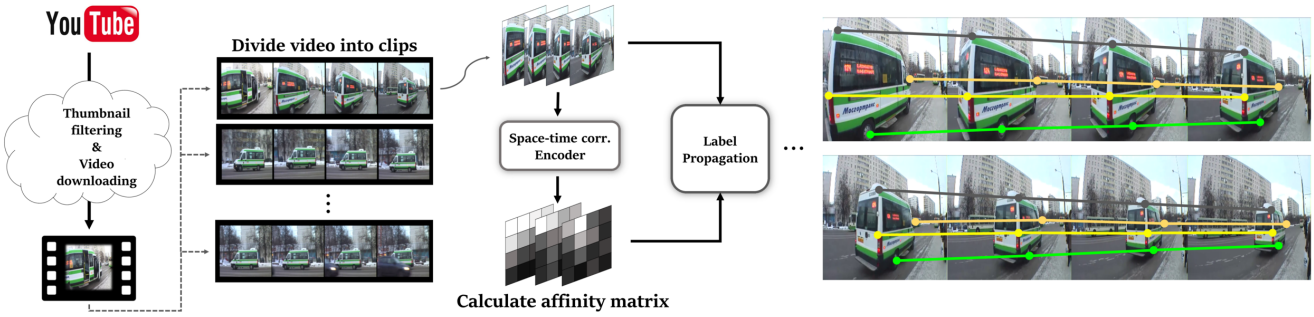


Figure 1. Our algorithm for collecting web videos (Section 3.2) and annotating them with pseudo correspondence labels (Section 3.3) fully automatically. The algorithm first downloads only thumbnail images and classify them to identify relevant videos. The relevant videos are then downloaded and divided into multiple clips with no abrupt transition. The algorithm trains a space-time correspondence model with the clips to generate dense pseudo correspondence labels for arbitrary pairs of frames of the clips.

gap between the web videos and common image datasets for the task.

The proposed framework was evaluated and compared with previous work on three public datasets, PF-Willow [9], PF-PASCAL [10], and SPair-71K [41], where it substantially outperformed existing self-supervised models, and improves performance of the state-of-the-art supervised learning model through transfer learning. In consequence, our work achieved the best on all the datasets in both self-supervised and strongly-supervised learning settings. Our major contribution is four-fold:

- We present the first attempt to utilize web videos for learning semantic correspondence in a self-supervised learning manner.
- We provide a fully automatic process for dataset construction and labeling using web videos. Our strategy exploits the exclusive advantages of videos over images for generating accurate pseudo correspondence labels.
- Our method outperformed existing self-supervised learning models and even substantially improved supervised learning performance through transfer learning.
- Our codebase for crawling and pseudo-labeling web videos will be open to public to promote future research.

2. Related Work

Semantic Correspondence. The task of semantic correspondence [9, 10, 40] has the objective of establishing correspondences between images with intra-class variations and viewpoint changes. Early approaches [5, 36, 57] utilized hand-crafted features to describe keypoints to be matched. Recent methods [6, 11, 15, 17, 19, 24, 25, 34, 38, 42, 46, 47, 49, 60] have shown remarkable progress by learning features and matching pipelines through deep neural networks using manually annotated datasets [2, 9, 10, 22, 41, 43, 57, 63, 69]. However, existing datasets are often limited in terms of the amount of image pairs and the number/diversity of keypoints annotated due to the tremendous annotation cost of point-to-point correspondences; this issue could be crucial in particular for the recent learning based methods. To address this issue, self-supervised learning strategies have been incorporated into semantic correspondence models [30, 37, 47, 59, 60]. One example is to synthesize image pairs with ground-truth correspondences by warping a real image for training. However, such synthetic geometric transformations usually fail to simulate intra-class variations exhibited in the real world. Another approach to dealing with the issue of limited training data is weakly supervised learning [15, 24, 46, 50], which utilizes only image-level class labels that indicate if a pair of images are of the same class. Also, a semi-supervised learning method [23] has been introduced to generate a large amount of pseudo correspondence labels of existing training data and exploit them to further improve performance.

Space-Time Correspondence. Space-time correspondence aims at finding correspondences across frames of a video clip [16, 20, 26, 27, 32, 55, 62, 66, 68]. The task has been addressed by self-supervised learning on unlabeled videos. Examples of pretext tasks for such self-supervised learning include color reconstruction [26, 27, 61], image reconstruction using auto-encoder [32], and cycle-consistency in time [16, 65, 68]. In particular, Jabri *et al*. [16] proposed a probabilistic framework using a contrastive random walk. Our pseudo label generation technique is built upon the method of Jabri *et al*. [16] to find dense space-time correspondences within each web video. The web videos and their pseudo labels are in turn used to learn a semantic correspondence model working on image pairs.

3. Method

Our framework consists of three stages: (1) retrieving videos from a web repository, (2) building space-time corre-

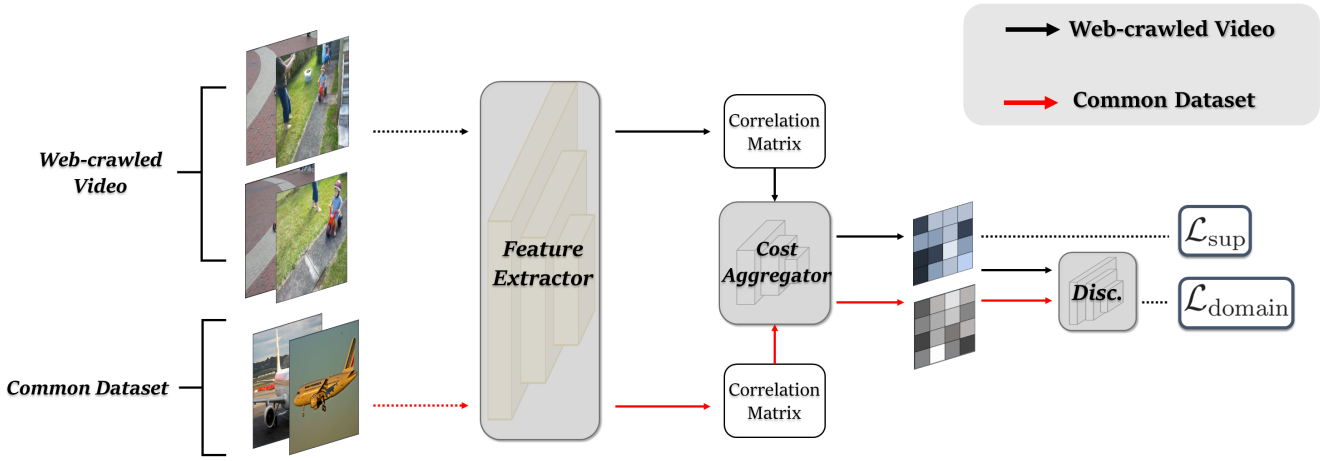


Figure 2. An overview of our framework using web videos. The web videos are used for conventional supervised learning of the base model (CATs [6] in this figure). Additionally, a common dataset is employed for domain adaptation learning, aiming to bridge the domain gap between the web videos and the common dataset without any form of supervision from the dataset.

spondences for each web video, and (3) training a semantic correspondence model with random pairs of distant frames sampled from the web videos while leveraging their space-time correspondences as pseudo labels. The first two stages are depicted in Fig. 1 and the last stage is illustrated in Fig. 2. The remainder of this section first provides preliminaries to our work and then elaborates on the three stages.

3.1. Preliminaries

Let I_s and I_t be source and target images that exhibit semantically similar objects. The goal of semantic correspondence is to match keypoints between I_s and I_t. To build correspondence between two images, a feature extraction network firstly extracts feature maps $F \in \mathbb{R}^{h \times w \times c}$, where $h \times w$ is the spatial resolution and c is the number of channels. Then we compute a correlation map $\mathcal{C} \in \mathbb{R}^{hw \times hw}$, where $\mathcal{C}(i,j) = {F_t^i}^\top F_s^j$ represents similarity between pixel i of F_t and pixel j of F_s. Unfortunately, this initial correlation map is often ambiguous and vulnerable to repetitive or textureless correspondence. Recent methods [6, 19, 29, 38, 40, 48, 49] remedy this by employing the cost aggregator to obtain a refined correlation map $\mathcal{C}'$ from the initial correlation map $\mathcal{C}$. We adopt CATs [6] as our base model for semantic correspondence. We also utilize common datasets [9, 10, 41] without their labels for closing the domain gap between images and videos.

3.2. Crawling Videos from Web Repository

Suppose that we have access to common semantic correspondence datasets for a set of pre-defined object classes. The first step to video crawling is to retrieve videos relevant to the object classes from YouTube using the class labels as search queries. However, the retrieved videos could be irrelevant to the classes of our interest due to errors of the search engine or the semantic ambiguity of the class labels as search queries (*e.g.*, when the class labels have multiple meanings). To mitigate this issue, following Hong *et al.* [14], we first download only the thumbnail images of the retrieved videos and then classify the images using a simple classifier trained for the classes of our interest using the labeled datasets; a video is downloaded only when its thumbnail classification score for the class label used as search query exceeds a predefined threshold.

Even though possibly irrelevant videos are filtered out in this manner, downloaded videos could be still inappropriate for estimating pseudo correspondence labels through space-time correspondence due to abrupt transitions between adjacent frames (*i.e.*, shot changes). Hence, each downloaded video is divided into multiple clips, each of which has no abrupt transition, through a shot detection method [1].

3.3. Generating Pseudo Labels for Web Videos

Our next step is to generate pseudo correspondence labels for the video clips obtained in Section 3.2. To this end, we first establish dense space-time correspondences for every pair of consecutive frames in every clip by training and deploying a self-supervised space-time correspondence model with no human intervention.[1] The established space-time correspondences are then used to derive pseudo correspondence labels between any arbitrary pairs of frame images. Since consecutive frames of each clip show smooth motions with no abrupt transition thanks to our video collection strategy in Section 3.2, the model can be trained stably and accordingly build accurate correspondences be-

[1] We adopt CRW [16] for this purpose, but any other recent space-time correspondence model can be incorporated.

tween consecutive frames.

Given the trained space-time correspondence model, pseudo correspondence labels are generated by propagating keypoints of the first frame to the remaining frames in the same clip; every pixel location of the first frame is considered as a keypoint as there is no manual keypoint annotation for the web videos. Let ϕ be an encoder of the space-time correspondence model and $F = \phi(I)$ be a feature map of an input image I. Then we compute the inter-frame affinity matrix $\mathcal{A}_{k,\,k+1} \in \mathbb{R}^{m \times m}$, where m is the number of pixels in each frame, as follows:

$$\mathcal{A}^{i,\,j}_{k,\,k+1} = \cos\big(F^i_k, F^j_{k+1}\big), \tag{1}$$

where cos denotes the cosine similarity function and F^i_k denotes the feature vector of pixel i of frame k. The affinity matrix $\mathcal{A}_{k,\,k+1}$ represents inter-pixel semantic affinity between two consecutive frames k and $k+1$.

Given the affinity matrices between all pairs of consecutive frames, the keypoints of the first frame are propagated sequentially to the other frames to build dense space-time correspondences within the clip. Let $\mathbf{y}_k \in \mathbb{N}^m$ indicate keypoint IDs of pixels in frame k (*i.e.*, $\mathbf{y}^i_k$ indicates the ID of the keypoint held by pixel i of frame k). Then $\mathbf{y}_k$ is propagated to frame $k+1$ for estimating $\mathbf{y}_{k+1}$ through the inter-frame affinity matrix as follows:

$$\mathbf{y}^j_{k+1} = \mathbf{y}^{\ell}_k, \quad \text{where } \ell = \underset{i}{\operatorname{argmax}}\,\big(\mathcal{A}^{i,\,j}_{k,\,k+1}\big). \tag{2}$$

To improve the accuracy of the pseudo correspondence labels, we detect and eliminate false label propagation. This is achieved by applying Isolation Forest [33], an algorithm devoted to outlier detection, to the pseudo labels generated by Eq. (2). We assume that a majority of the labels propagated between successive frames will exhibit similar degrees of change in their positions. Through Isolation Forest, we measure the degree of change in the x and y axes for each propagated label, and then identify and remove any labels with an unusual amount of change. This process leads to error-resistant pseudo correspondence labels.

Given these results, pseudo correspondence labels between two arbitrary frames k and k' of each clip are obtained simply by identifying keypoints co-occuring at both of the frames, *i.e.*, the intersection between $\mathbf{y}_k$ and $\mathbf{y}_{k'}$.

3.4. Learning Correspondence with Web Videos

Our learning framework mainly utilizes the web video dataset constructed in the previous section along with a common dataset for the task. Details of our training strategy using are illustrated in the following sections.

3.4.1 Supervised Learning with Web Videos

For supervised learning using web videos, we follow the training objective used in [6, 38, 40, 42]. Given generated pseudo keypoints for each randomly picked pair of images, we first generate the pseudo ground-truth flow field $\mathbb{F}_{\text{gt}}$ as in [40], and transform the refined correlation map $\mathcal{C}'$ into an estimated flow field $\mathbb{F}_{\text{est}}$. Then the supervised loss, called average end-point error [37], is applied to the two flow fields as follows:

$$\mathcal{L}_{\text{sup}} = \|\mathbb{F}_{\text{gt}} - \mathbb{F}_{\text{est}}\|_2. \tag{3}$$

3.4.2 Domain Adaptive Learning

We adopt a domain adaptive learning technique to mitigate possible negative effect of the domain gap between images of common correspondence datasets and web videos in a feature space. The objective is to train the correspondence model in such a way that its features for web videos and a common dataset are indistinguishable to the discriminator. For this purpose, we employ a gradient reversal layer [8]. Let d be the discriminator and D_w and D_c represent the set of frame images from web videos and a common image dataset, respectively. Then the domain adaptation loss is given by

$$\begin{aligned}\mathcal{L}_{\text{domain}} = &\frac{1}{|D_c|}\sum_{X_c \in D_c} \mathcal{L}_{\text{ce}}\Big(d\Big(r(G_c)\Big), c\Big) + \\ &\frac{1}{|D_w|}\sum_{X_w \in D_w} \mathcal{L}_{ce}\Big(d\Big(r(G_w)\Big), w\Big),\end{aligned} \tag{4}$$

where r indicates the gradient reversal layer and $\mathcal{L}_{\text{ce}}$ denotes the cross-entropy loss. w and c are domain labels indicating web videos and common image data, G_w and G_c are feature maps of given images X_w and X_c from the semantic correspondence model, respectively.

The total loss for our framework is then given by

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{sup}} + \lambda\mathcal{L}_{\text{domain}}, \tag{5}$$

where λ is a loss re-scaling factor.

4. Experiments

4.1. Implementation Details

Network architecture. We adopt CRW [16] for pseudo labeling of web videos in Section 3.3, and CATs [6] as our base semantic correspondence model for the training in Section 3.4. Their backbones are both ResNet [12]: ResNet-18 for the encoder of CRW, and ResNet-101 for the feature extraction network of CATs. Note that we use the same hyperparameters for the training of CRW and CATs, as they reported in the paper. ResNet-18 is used for the thumbnail classifier in Section 3.2. For the domain adaptation learning using our web videos ($\mathcal{L}_{\text{domain}}$ in Section 3.4), we use the last feature maps before its channel-wise averaged in

Method	Supervision type	Supervision Signal	PF-PASCAL			PF-Willow			SPair-71k
			0.05	0.1	0.15	0.05	0.1	0.15	0.1
SF-Net ResNet-101 [28]	Weak-sup.	bounding box	53.6	81.9	90.6	46.3	74.0	84.2	-
Weakalign ResNet-101 [46]		image-level label	49.0	74.8	84.0	37.0	70.2	79.9	20.9
RTNs ResNet-101 [24]			55.2	75.9	85.2	41.3	71.9	86.2	25.7
NC-Net ResNet101 [50]			54.3	78.9	86.0	33.8	67.0	83.7	20.1
PCSNet-SE ResNet101 [18]			59.8	80.3	88.5	42.6	75.1	88.0	26.5
PF HOG [10]	None	-	31.4	62.5	79.5	28.4	56.8	68.2	-
CNNGeo ResNet-101 [47]	Self-sup.	synthetic image pairs	41.0	69.5	80.4	36.9	69.2	77.8	20.6
A2Net ResNet-101 [54]			42.8	70.8	83.3	36.3	68.8	84.4	22.3
PMD ResNet-101 [30]			-	80.5	-	-	73.4	-	-
Ours ResNet-101	Self-sup.	synthetic keypoints	**50.7**	**80.6**	**90.0**	**44.5**	**74.7**	**87.9**	**27.3**

Table 1. Comparisons with self/weakly-supervised methods in PCK (%) on PF-PASCAL, PF-Willow, and SPair-71k. The backbone network of each method is indicated in the subscript. The best and the second best results are marked in **bold** and underline, respectively.

Method	Supervision type	Supervision Signal	PF-PASCAL			PF-Willow			SPair-71k
			0.05	0.1	0.15	0.05	0.1	0.15	0.1
SCNet VGG-16 [11]	Strong-sup.	keypoints	36.2	72.2	82.0	38.6	70.4	85.3	-
ANC-Net ResNet-101-FCN [29]			-	86.1	-	-	-	-	28.7
HPF ResNet-101 [40]			60.1	84.8	92.7	45.9	74.4	85.6	28.2
DHPF ResNet-101 [42]			75.7	90.7	95.0	49.5	77.6	89.1	37.3
CHMNet ResNet-101 [38]			80.1	91.6	94.9	52.7	79.4	87.5	46.3
MMNet ResNet-101 [67]			77.6	89.1	94.3	-	-	-	40.9
TransforMatcher ResNet-101 [25]			**80.8**	91.8	-	-	76.0	-	53.7
CATs ResNet-101 [6]			75.4	92.6	96.4	50.3	79.2	90.3	49.9
Ours ResNet-101	Strong-sup. (transferred)	keypoints	80.4	**93.6**	**96.8**	**54.8**	**80.9**	**91.0**	**54.0**

Table 2. Comparisons with supervised methods in PCK (%) on PF-PASCAL, PF-Willow and SPair-71k. The backbone network of each method is indicated in the subscripts. The best and the second best results are marked in **bold** and underline, respectively.

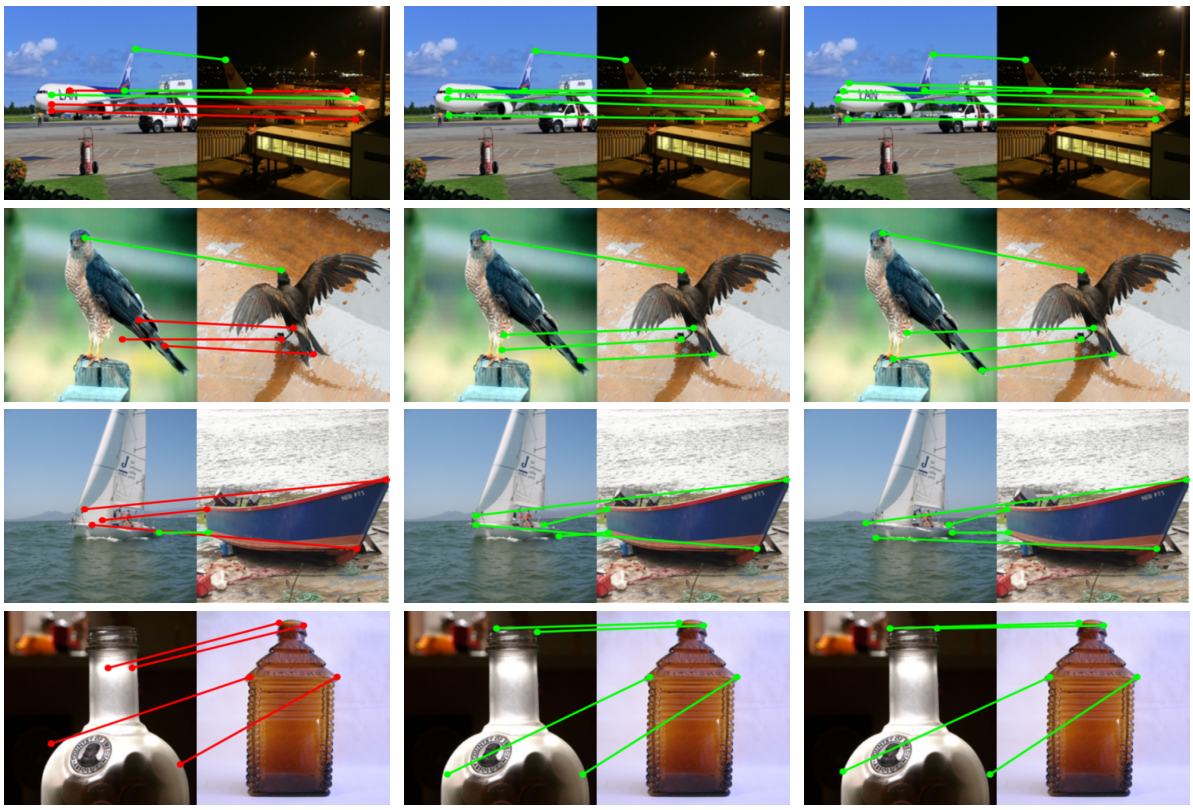
Methods	aero.	bicy.	bird	boat	bott.	bus	car	cat	chai.	cow	dog	hors.	mbik.	pers.	plan.	shee.	trai.	tv	all
CNNGeo [47]	23.4	16.7	40.2	14.3	36.4	27.7	26.0	32.7	12.7	27.4	22.8	13.7	20.9	21.0	17.5	10.2	30.8	34.1	20.6
A2Net [54]	22.6	18.5	42.0	16.4	37.9	30.8	26.5	35.6	13.3	29.6	24.3	16.0	21.6	22.8	20.5	13.5	31.4	36.5	22.3
WeakAlign [46]	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 [50]	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 [40]	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 [34]	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 [42]	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 [38]	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 [67]	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
TransforMatcher [25]	**59.2**	**39.3**	73.0	**41.2**	52.5	**66.3**	**55.4**	67.1	**26.1**	**67.1**	56.6	53.2	**45.0**	39.9	42.1	35.3	75.2	68.6	53.7
CATs [6]	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 [23]	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
Ours	55.8	38.8	**77.3**	38.2	**52.7**	57.5	50.0	**67.8**	25.4	65.4	**63.8**	**59.3**	44.2	**49.7**	40.4	**41.7**	**75.3**	**70.5**	**54.0**

Table 3. Per-class quantitative results in PCK (%) ($\alpha_k = 0.1$) on SPair-71K. The best and the second best results are marked in **bold** and underline, respectively.

the cost aggregator. The discriminator used in $\mathcal{L}_{\text{domain}}$ consists of two 3×3 convolutional layers followed by two fully-connected layers.

Web video dataset. We collect 2,124 videos retrieved in Section 3.2 and get 36,879 clips with 2,806,736 frame pairs; object classes of PF-PASCAL, PF-WILLOW, and SPair-71K are used as search queries for the video retrieval. The thumbnail classifier used in Section 3.2 is trained on the PASCAL VOC dataset. We use the same pseudo correspondence labels for all conducted experiments.

Datasets for evaluation. We conduct experiments on the three standard benchmarks for semantic correspondence, SPair-71K [41], PF-PASCAL [10], and PF-Willow [9]. SPair-71k contains 70,958 image pairs with diverse view-



(a) CATs [6] (b) Ours (transferred) (c) Ground-truth

Figure 3. Qualitative results on SPair-71K [41] *test* set.

point and scale variations. PF-PASCAL provides 1,351 image pairs from 20 categories of the PASCAL VOC [7] dataset, and PF-Willow provides 900 image pairs from 4 categories. Note that all previous weakly, self-, and strongly-supervised methods evaluated in Table 1 and Table 2 follow the standard evaluation protocol [6, 15, 40, 42]: For SPair-71K, they are trained and evaluated on the training and test splits, respectively, while for PF-PASCAL and PF-Willow, they are trained on the training split of PF-PASCAL and evaluated on the test split of each dataset.

Data augmentation. We use the same photometric augmentation lists as in [6] for data augmentation. Random horizontal flip, random perspective transform, color jittering, and random grayscale are applied differently to different frames of the same clip.

Optimizer. We adopt AdamW [35] optimization with learning rate of 3e-5 and weight decay of 0.05.

Hyper-parameters. The threshold for thumbnail classification scores in Section 3.2 is set to 0.8, following [14]. The loss re-scaling parameter λ of $\mathcal{L}_{\text{domain}}$ is set to 0.025. The number of estimators used in Isolation Forest is set to 100 following its default setting.

Software. We mainly use Pytorch [44] for a deep learning framework and Scikit-learn [45] for Isolation Forest.

Evaluation metric. We adopt the percentage of correct keypoints (PCK) as the performance metric. Given a set of estimated and ground-truth keypoint pairs $\mathcal{K} = \{(k_{\text{est}}(m), k_{\text{GT}}(m)\}$, PCK is computed by

$$\text{PCK}(\mathcal{K}) = \frac{1}{M}\sum_{m=1}^{M} \mathbb{1}[\|k_{\text{GT}}(m) - k_{\text{est}}(m)\| \leq \alpha_k \cdot \max(H, W)], \tag{6}$$

where M is the number of keypoint pairs, H and W are the width and height of the entire image or object bounding box, and α_k is a scale factor.

4.2. Results

Comparison with self/weakly-supervised methods. We first compare our approach to self-supervised methods [30, 47, 54] and weakly-supervised methods [18, 24, 28, 46, 50]. In this setup, we train our model using pseudo labels generated from web videos for matching supervision, and the common dataset [9, 10, 41] for domain adaptation learning. Our method achieves state-of-the-art results, as shown in Table 1, with 27.3% PCK@0.1 for the challenging SPair-71K dataset, surpassing all other self-supervised and even

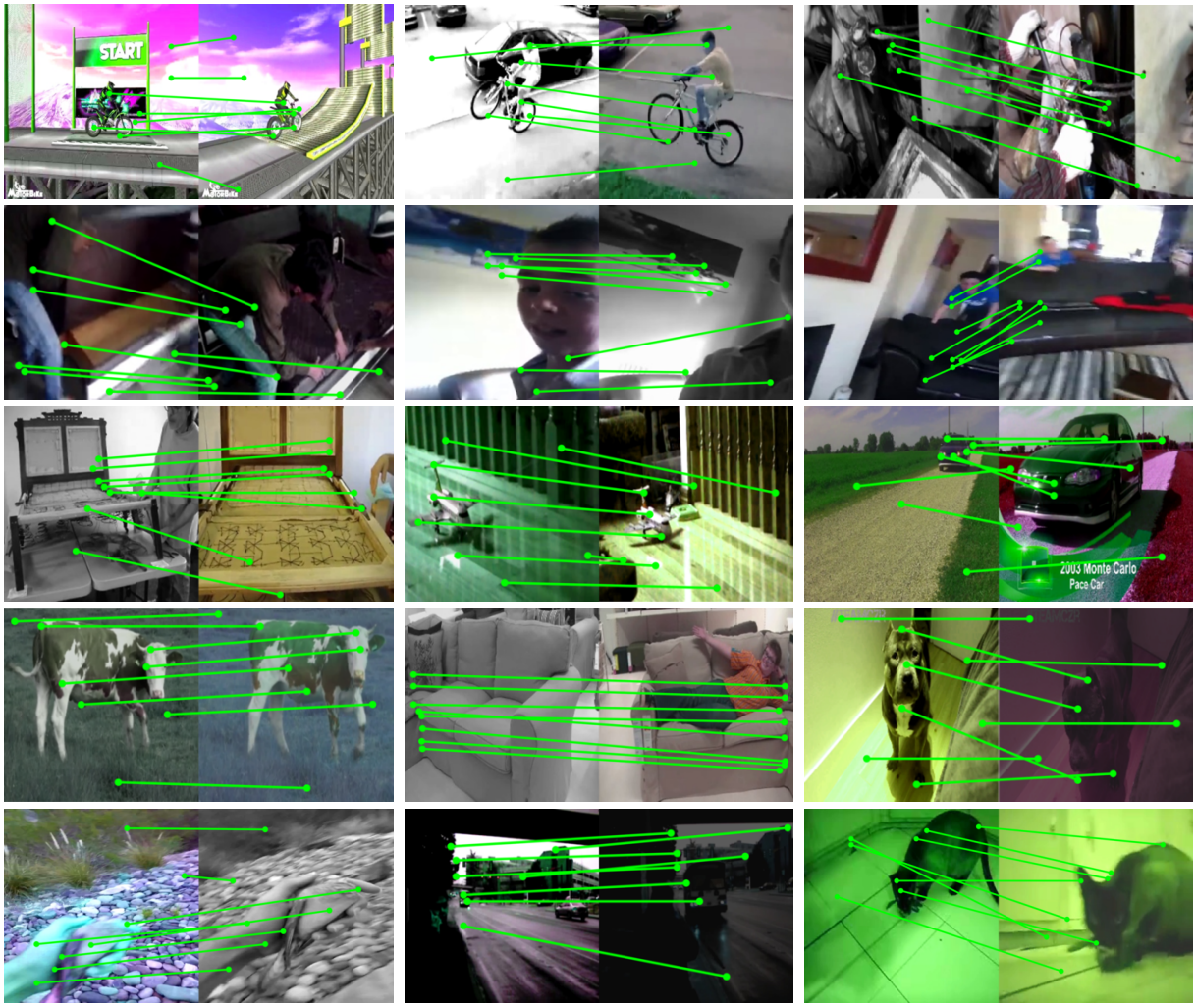


Figure 4. Qualitative results of pseudo correspondence labels on web-crawled videos. We randomly sampled eight keypoints for pseudo correspondence labels of each frame pair. Note that source and target images are augmented differently.

weakly-supervised methods. For PF-PASCAL, our method consistently outperforms previous self-supervised methods, achieving 80.6% PCK in the PCK@0.1 setting. Similarly, for PF-Willow, our method demonstrates its effectiveness, outperforming all previous self-supervised methods. It is important to note that our model primarily trains on pseudo labels generated from web videos, and the other common datasets are only used for domain adaptation learning without any supervision.

Comparison with strongly-supervised methods. We conducted a comparison between our method and recent state-of-the-art strongly-supervised models to demonstrate the effectiveness of our approach. We employed transfer learning, where the model was initially trained on pseudo labels from web videos and then fine-tuned on the target common dataset only with the supervised learning loss in Eq. (3). The results, presented in Table 2, show that our method achieves state-of-the-art performance. For PF-PASCAL, we obtained 96.8% PCK@0.15 and 93.6% PCK@0.1, and for PF-Willow, we achieved 54.8% PCK@0.05, 80.9% PCK@0.1, and 91.0% PCK@0.15. Additionally, our method demonstrates significant performance improvement on SPair-71K, with a PCK@0.1 score of 54.0%. We further compare our method with the others in terms of per-class accuracy on SPair-71K, showcasing their performance in detail. As shown in Table 3, our method achieves the best or second-best for almost all classes of SPair-71K, outperforming other methods in 9 out of the 18 classes. Its average PCK performance is 54.0%.

Qualitative analysis. Fig. 3 displays our results on SPair-71K, while Fig.4 exhibits example frame pairs with pseudo correspondence labels obtained from web-crawled videos.

4.3. Ablation Studies

We conduct ablation studies to investigate the effectiveness of each component of the proposed method. The ex-

λ	0.0025	0.005	0.025	0.05	0.25	0.5
PCK@0.1	27.0	26.9	27.3	27.0	26.5	27.1

Table 4. Impact of λ in Eq. (5) on SPair-71K test set.

$\mathcal{L}_{\text{sup}}$	$\mathcal{L}_{\text{domain}}$	PCK@0.1
✓	✓	**27.3**
✓	✗	25.7

Table 5. Ablation studies of different loss combinations in Eq. (5) in PCK@0.1 on SPair -71K [41] test set.

ORB	ISF	PCK@0.1
✓	✗	15.2
✗	✗	22.5
✗	✓	**27.3**

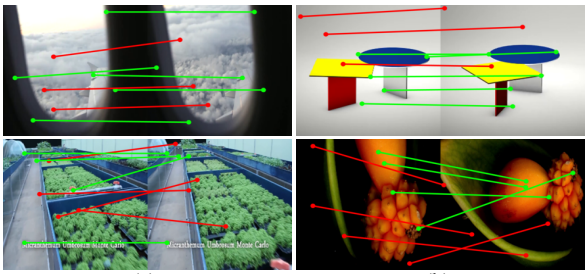
Table 6. Ablation studies of our pseudo label generation method in PCK@0.1 on SPair-71K [41] test set. ORB denotes using ORB algorithm [52] to directly generate pseudo labels (find correspondence between two images) instead of using space-time correspondence encoder, and ISF denotes using the error filtering technique based on Isolation Forest [33].

periments were conducted on the test set of SPair-71K [41].

Impact of λ in Eq. (5). As shown in Table 4, our method is not sensitive to the value of λ, the best performance was achieved when $\lambda = 0.025$ though.

Loss component analysis in Eq. (5) . In the learning phase described in Section 3.4, $\mathcal{L}_{\text{sup}}$ and $\mathcal{L}_{\text{domain}}$ are jointly optimized. The domain adaptation loss, $\mathcal{L}_{\text{domain}}$, is used to minimize the negative impact of the domain gap between the common dataset and web videos in the feature space of the correspondence model. We conducted a comparison to investigate the effectiveness of each loss term in Eq. (5). The results, presented in Table 5, indicate that each term contributes to the overall performance, and the best results are achieved when both losses are utilized.

Analysis on pseudo label generation method. In the pseudo label generation stage described in Section 3.3, we first utilize a space-time correspondence encoder to create pseudo correspondence labels from web videos. These labels are then filtered to remove any erroneous ones using the Isolation Forest algorithm [33]. In this section, we make a comparison to investigate effect of pseudo label generation strategy. First, to demonstrate the effectiveness of our approach using the space-time correspondence encoder, we train our model with pseudo labels generated by off-the-shelf image feature matching algorithm, ORB [52], instead of using the encoder, under the same conditions (*e.g.* augmentation) with our method. The results, as shown in the first row of Table 6, demonstrate poor performance with only 15.2% PCK@0.1 compared to our approach, which achieved 27.3% PCK@0.1. Second, we train our model using pseudo labels that were not filtered by the Isolation Forest algorithm [33]. As demonstrated in the second and third rows of Table 6, applying the filtering algorithm significantly improved the performance of the model. Based on these findings, we may conclude that utilizing the space-time encoder to find correspondences in web videos is highly effective, and applying the error filtering algorithm enhances performance effectively.



(a) (b)

Figure 5. Qualitative results of some failure cases of pseudo correspondence labels on web-crawled videos, where the pattern of texture is iterative (a) or featureless (b).

5. Limitation and discussion

Although our method could capture dense and reliable correspondence on consecutive frames, there are some failure cases where textures are iterative (Fig. 5(a)) or featureless (Fig. 5(b)). We believe developing an algorithm that detects or corrects possible errors on space-time correspondences of web videos by using readily available labeled data could be one of the promising future directions.

6. Conclusion

We have presented a self-supervised learning framework for semantic correspondence with the process for dataset construction and pseudo labeling using web videos. To mitigate the lack of large-scale dense annotation data for the training of semantic correspondence, we retrieve a large amount of videos from the web repository and generate pseudo correspondence labels by utilizing the space-time correspondence model. We then train a model with the pseudo correspondence labels and a common dataset. Our framework substantially improved performance over existing self-supervised methods for all three benchmarks.

Acknowledgement. This work was supported by Samsung Electronics Co., Ltd (IO201210-07948-01).

References

[1] Brandon Castellano. Pyscenedetect. 3

[2] Xianjie Chen, Roozbeh Mottaghi, Xiaobai Liu, Sanja Fidler, Raquel Urtasun, and Alan Yuille. Detect what you can: Detecting and representing objects using holistic models and body parts. In *CVPR*, 2014. 1, 2

[3] Yun-Chun Chen, Po-Hsiang Huang, Li-Yu Yu, Jia-Bin Huang, Ming-Hsuan Yang, and Yen-Yu Lin. Deep semantic matching with foreground detection and cycle-consistency. In *Asian Conference on Computer Vision (ACCV)*, 2018. 1

[4] Yun-Chun Chen, Yen-Yu Lin, Ming-Hsuan Yang, and Jia-Bin Huang. Show, match and segment: Joint weakly supervised learning of semantic matching and object co-segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2020. 1

[5] Minsu Cho, Suha Kwak, Cordelia Schmid, and Jean Ponce. Unsupervised object discovery and localization in the wild: Part-based matching with bottom-up region proposals. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 2

[6] Seokju Cho, Sunghwan Hong, Sangryul Jeon, Yunsung Lee, Kwanghoon Sohn, and Seungryong Kim. Cats: Cost aggregation transformers for visual correspondence. In *Proc. Neural Information Processing Systems (NeurIPS)*, 2021. 1, 2, 3, 4, 5, 6

[7] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The Pascal Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision (IJCV)*, 2010. 6

[8] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016. 1, 4

[9] Bumsub Ham, Minsu Cho, Cordelia Schmid, and Jean Ponce. Proposal flow. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 1, 2, 3, 5, 6

[10] Bumsub Ham, Minsu Cho, Cordelia Schmid, and Jean Ponce. Proposal flow: Semantic correspondences from object proposals. In *arXiv:1703.07144*, 2017. 1, 2, 3, 5, 6

[11] Kai Han, Rafael S. Rezende, Bumsub Ham, Kwan-Yee K. Wong, Minsu Cho, Cordelia Schmid, and Jean Ponce. Scnet: Learning semantic correspondence. In *International Conference on Computer Vision (ICCV)*, 2017. 1, 2, 5

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 4

[13] Sunghwan Hong, Seokju Cho, Jisu Nam, Stephen Lin, and Seungryong Kim. Cost aggregation with 4d convolutional swin transformer for few-shot segmentation. In *European Conference on Computer Vision*, pages 108–126. Springer, 2022. 1

[14] S. Hong, D. Yeo, S. Kwak, H. Lee, and B. Han. Weakly supervised semantic segmentation using web-crawled videos. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2224–2232, 2017. 3, 6

[15] Shuaiyi Huang, Qiuyue Wang, Songyang Zhang, Shipeng Yan, and Xuming He. Dynamic context correspondence network for semantic alignment. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, pages 2010–2019, 2019. 1, 2, 6

[16] Allan Jabri, Andrew Owens, and Alexei A. Efros. Space-time correspondence as a contrastive random walk. In *Proc. Neural Information Processing Systems (NeurIPS)*, 2020. 2, 3, 4

[17] Sangryul Jeon, Seungryong Kim, Dongbo Min, and Kwanghoon Sohn. Parn: Pyramidal affine regression networks for dense semantic correspondence. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Proc. European Conference on Computer Vision (ECCV)*, 2018. 1, 2

[18] Sangryul Jeon, Seungryong Kim, Dongbo Min, and Kwanghoon Sohn. Pyramidal semantic correspondence networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):9102–9118, 2022. 5, 6

[19] Sangryul Jeon, Dongbo Min, Seungryong Kim, Jihwan Choe, and Kwanghoon Sohn. Guided semantic flow. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Proc. European Conference on Computer Vision (ECCV)*, 2020. 1, 2, 3

[20] Sangryul Jeon, Dongbo Min, Seungryong Kim, and Kwanghoon Sohn. Mining better samples for contrastive learning of temporal correspondence. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1034–1044, 2021. 1, 2

[21] Dahyun Kang and Minsu Cho. Integrative few-shot learning for classification and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1

[22] Jaechul Kim, Ce Liu, Fei Sha, and Kristen Grauman. Deformable spatial pyramid matching for fast dense correspondences. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2307–2314, 2013. 1, 2

[23] Jiwon Kim, Kwangrok Ryoo, Junyoung Seo, Gyuseong Lee, Daehwan Kim, Hansang Cho, and Seungryong Kim. Semi-supervised learning of semantic correspondence with pseudo-labels. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19699–19709, June 2022. 2, 5

[24] Seungryong Kim, Stephen Lin, Sang Ryul Jeon, Dongbo Min, and Kwanghoon Sohn. Recurrent transformer networks for semantic correspondence. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. 1, 2, 5, 6

[25] Seungwook Kim, Juhong Min, and Minsu Cho. Transformatcher:match-to-match attention for semantic correspondence. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 2, 5

[26] Zihang Lai, Erika Lu, and Weidi Xie. MAST: A memory-augmented self-supervised tracker. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1, 2

[27] Z. Lai and W. Xie. Self-supervised learning for video correspondence flow. In *BMVC*, 2019. 1, 2

[28] J. Lee, D. Kim, J. Ponce, and B. Ham. Sfnet: Learning

object-aware semantic flow. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 5, 6

[29] S. Li, K. Han, T. W. Costain, H. Howard-Jenkins, and V. Prisacariu. Correspondence networks with adaptive neighbourhood consensus. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3, 5

[30] Xin Li, Deng-Ping Fan, Fan Yang, Ao Luo, Hong Cheng, and Zicheng Liu. Probabilistic model distillation for semantic correspondence. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2, 5, 6

[31] Xueting Li, Sifei Liu, Shalini De Mello, Xiaolong Wang, Jan Kautz, and Ming-Hsuan Yang. Joint-task self-supervised learning for temporal correspondence. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. 1

[32] Xueting Li, Sifei Liu, Shalini De Mello, Xiaolong Wang, Jan Kautz, and Ming-Hsuan Yang. Joint-task self-supervised learning for temporal correspondence. In *NeurIPS*, 2019. 2

[33] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining*. IEEE, 2008. 4, 8

[34] 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, 2, 5

[35] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *Proc. International Conference on Learning Representations (ICLR)*, 2019. 6

[36] David G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91–110, nov 2004. 2

[37] 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. 2, 4

[38] Juhong Min and Minsu Cho. Convolutional hough matching networks. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2940–2950, June 2021. 1, 2, 3, 4, 5

[39] Juhong Min, Dahyun Kang, and Minsu Cho. Hypercorrelation squeeze for few-shot segmentation. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2021. 1

[40] Juhong Min, Jongmin Lee, Jean Ponce, and Minsu Cho. Hyperpixel flow: Semantic correspondence with multi-layer neural features. In *ICCV*, 2019. 1, 2, 3, 4, 5, 6

[41] Juhong Min, Jongmin Lee, Jean Ponce, and Minsu Cho. Spair-71k: A large-scale benchmark for semantic correspondence. *arXiv prepreint arXiv:1908.10543*, 2019. 1, 2, 3, 5, 6, 8

[42] Juhong Min, Jongmin Lee, Jean Ponce, and Minsu Cho. Learning to compose hypercolumns for visual correspondence. In *ECCV*, 2020. 1, 2, 4, 5, 6

[43] David Novotny, Diane Larlus, and Andrea Vedaldi. I have seen enough: Transferring parts across categories. In *British Machine Vision Conference*, 2016. 1, 2

[44] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *AutoDiff, NIPS Workshop*, 2017. 6

[45] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. 6

[46] I. Rocco, R. Arandjelović, and J. Sivic. End-to-end weakly-supervised semantic alignment. *arXiv preprint arXiv:1712.06861*. 1, 2, 5, 6

[47] I. Rocco, R. Arandjelović, and J. Sivic. Convolutional neural network architecture for geometric matching. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1, 2, 5, 6

[48] I. Rocco, R. Arandjelović, and J. Sivic. Efficient neighbourhood consensus networks via submanifold sparse convolutions. In *Proc. European Conference on Computer Vision (ECCV)*, 2020. 3

[49] I. Rocco, M. Cimpoi, R. Arandjelović, A. Torii, T. Pajdla, and J. Sivic. Neighbourhood consensus networks. In *Proc. Neural Information Processing Systems (NeurIPS)*, 2018. 1, 2, 3

[50] Ignacio Rocco, Mircea Cimpoi, Relja Arandjelović, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Ncnet: Neighbourhood consensus networks for estimating image correspondences. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 44(2):1020–1034, 2022. 2, 5, 6

[51] Michael Rubinstein, Armand Joulin, Johannes Kopf, and Ce Liu. Unsupervised joint object discovery and segmentation in internet images. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1939–1946, 2013. 1

[52] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. Orb: An efficient alternative to sift or surf. In *2011 International Conference on Computer Vision*, pages 2564–2571, 2011. 8

[53] Johannes Schönberger, Marc Pollefeys, Andreas Geiger, and Torsten Sattler. Semantic visual localization. 06 2018. 1

[54] Paul Hongsuck Seo, Jongmin Lee, Deunsol Jung, Bohyung Han, and Minsu Cho. Attentive semantic alignment with offset-aware correlation kernels. In *ECCV*, 2018. 5, 6

[55] Jeany Son. Contrastive learning for space-time correspondence via self-cycle consistency. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14679–14688, June 2022. 1, 2

[56] Hajime Taira, Masatoshi Okutomi, Torsten Sattler, Mircea Cimpoi, Marc Pollefeys, Josef Sivic, Tomas Pajdla, and Akihiko Torii. InLoc: Indoor visual localization with dense matching and view synthesis. In *CVPR*, 2018. 1

[57] Tatsunori Taniai, Sudipta N. Sinha, and Yoichi Sato. Joint recovery of dense correspondence and cosegmentation in two images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 1, 2

[58] Carl Toft, Erik Stenborg, Lars Hammarstrand, Lucas Brynte, Marc Pollefeys, Torsten Sattler, and Fredrik Kahl. Semantic match consistency for long-term visual localization. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision – ECCV 2018*, pages 391–408, 2018. 1

[59] Prune Truong, Martin Danelljan, Luc V Gool, and Radu Timofte. Gocor: Bringing globally optimized correspondence volumes into your neural network. In *Proc. Neural Information Processing Systems (NeurIPS)*, volume 33, pages 14278–14290, 2020. 2

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

[61] Carl Vondrick, Abhinav Shrivastava, Alireza Fathi, Sergio Guadarrama, and Kevin Murphy. Tracking emerges by colorizing videos. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision – ECCV 2018*, 2018. 1, 2

[62] Carl Vondrick, Abhinav Shrivastava, Alireza Fathi, Sergio Guadarrama, and Kevin Murphy. Tracking emerges by colorizing videos. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Proc. European Conference on Computer Vision (ECCV)*, pages 402–419, 2018. 2

[63] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. "". Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. 1, 2

[64] Ning Wang, Wengang Zhou, and Houqiang Li. Contrastive transformation for self-supervised correspondence learning. In *AAAI*, 2021. 1

[65] Xiaolong Wang, Allan Jabri, and Alexei A. Efros. Learning correspondence from the cycle-consistency of time. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 2

[66] Jiarui Xu and Xiaolong Wang. Rethinking self-supervised correspondence learning: A video frame-level similarity perspective. *arXiv preprint arXiv:2103.17263*, 2021. 2

[67] Dongyang Zhao, Ziyang Song, Zhenghao Ji, Gangming Zhao, Weifeng Ge, and Yizhou Yu. Multi-scale matching networks for semantic correspondence. 2021. 5

[68] Zixu Zhao, Yueming Jin, and Pheng-Ann Heng. Modelling neighbor relation in joint space-time graph for video correspondence learning. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, pages 9960–9969, October 2021. 1, 2

[69] Tinghui Zhou, Yong Jae Lee, Stella X. Yu, and Alexei A. Efros. Flowweb: Joint image set alignment by weaving consistent, pixel-wise correspondences. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 1, 2