Weakly Supervised Learning of Semantic Correspondence through Cascaded Online Correspondence Refinement

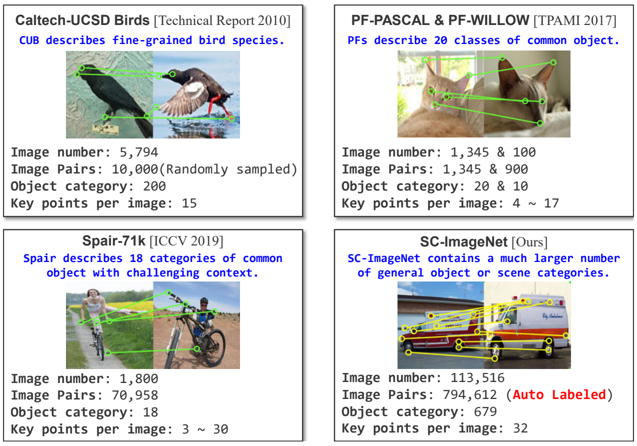
Yiwen Huang[1,*], Yixuan Sun[2,*], Chenghang Lai[1], Qing Xu[3],
Xiaomei Wang[3], Xuli Shen[1] and Weifeng Ge[1,†]
[1]School of Computer Science, Fudan University, Shanghai, China
[2]Academy of Engineering & Technology, Fudan University, Shanghai, China
[3]UniDT Technology, Shanghai, China
wfge@fudan.edu.cn

Abstract

In this paper, we develop a weakly supervised learning algorithm to learn robust semantic correspondences from large-scale datasets with only image-level labels. Following the spirit of multiple instance learning (MIL), we decompose the weakly supervised correspondence learning problem into three stages: image-level matching, region-level matching, and pixel-level matching. We propose a novel cascaded online correspondence refinement algorithm to integrate MIL and the correspondence filtering and refinement procedure into a single deep network and train this network end-to-end with only image-level supervision, i.e., without point-to-point matching information. During the correspondence learning process, pixel-to-pixel matching pairs inferred from weak supervision are propagated, filtered, and enhanced through masked correspondence voting and calibration. Besides, we design a correspondence consistency check algorithm to select images with discriminative key points to generate pseudo-labels for classical matching algorithms. Finally, we filter out about 110,000 images from the ImageNet ILSVRC training set to formulate a new dataset, called SC-ImageNet. Experiments on several popular benchmarks indicate that pre-training on SC-ImageNet can improve the performance of state-of-the-art algorithms efficiently. Our project is available on https://github.com/21210240056/SC-ImageNet.

1. Introduction

Learning semantic correspondence between object instances of the same category has become a fundamental problem in computer vision [9, 13]. Semantic matching methods have various applications in few shot learning [22, 14], multi-object tracking [25], image editing [19, 31] and etc. With the breakthrough of deep learning, state-of-the-art algorithms have achieved impressive achievements [9, 44, 14]. However, current popular semantic correspondence datasets, such as FG3DCar [37], PF-PASCAL [13], SPair71k [28] and Caltech-UCSD Birds [39], only contain limited annotated samples and thus weaken the generalization ability of existing algorithms.


Caltech-UCSD Birds [Technical Report 2010]
CUB describes fine-grained bird species.
PF-PASCAL & PF-WILLOW [TPAMI 2017]
PFs describe 20 classes of common object.
Spair-71k [ICCV 2019]
Spair describes 18 categories of common object with challenging context.
SC-ImageNet [Ours]
SC-ImageNet contains a much larger number of general object or scene categories.

Caltech-UCSD Birds:
Image number: 5,794
Image Pairs: 10,000(Randomly sampled)
Object category: 200
Key points per image: 15

PF-PASCAL & PF-WILLOW:
Image number: 1,345 & 100
Image Pairs: 1,345 & 900
Object category: 20 & 10
Key points per image: 4 ~ 17

Spair-71k:
Image number: 1,800
Image Pairs: 70,958
Object category: 18
Key points per image: 3 ~ 30

SC-ImageNet:
Image number: 113,516
Image Pairs: 794,612 (Auto Labeled)
Object category: 679
Key points per image: 32

Figure 1. Annotation statistics of three popular semantic matching datasets and our SC-ImageNet. Green circles and lines demonstrate key-point and semantic correspondence relationship annotated by humans. Yellow circles and lines are our automatically generated pseudo-labels.

Different from other visual recognition tasks [7, 45, 21, 46, 15], building dense semantic correspondence datasets needs to identify important object parts or salient feature points, which is much more complex and labor-expensive. Figure 1 illustrates the annotation statistics of three popular semantic matching datasets. As listed above, Caltech-UCSD Birds [39] has 200 fine-greind bird species and 5,794 annotated images, PF-PASCAL [13] contains 20 different object categories and 1,345 annotated images, and

$*$: Equal Contribution
†: Corresponding Author

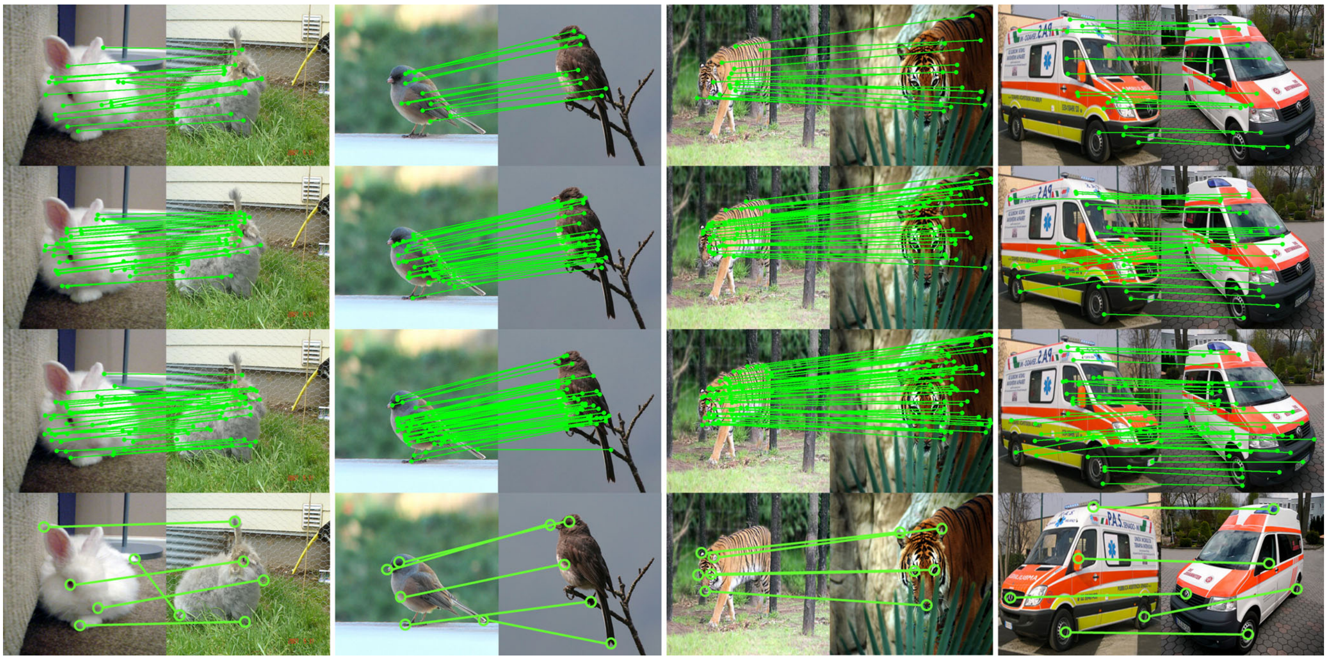


Figure 2. Visual comparison of semantic correspondences generated by different matching levels. All images are selected from ImageNet [7]. Rows from top to bottom demonstrate pseudo-labels generated by the image-level, region-level, and pixel-level matching stages and the last row stands for correspondences annotated by humans.

SPair71k [28] contains 18 different object categories and 1,800 annotated images. The number of key points in each image varies from 3 to 30, and the average number is 7. Compared with other large-scale datasets, such as ImageNet [7] and MS COCO [21], the annotated images in the semantic correspondence datasets are far from enough.

This paper aims to solve the problem of training deep neural networks for semantic matching tasks with insufficient training data. Our motivation is to investigate whether we can learn accurate semantic correspondence from large-scale datasets with only image-level annotations. We adopt the weakly supervised learning scheme [36, 11, 3, 12, 42] to learn semantic correspondences from the ImageNet ILSVRC training set since it has a massive amount of labeled data and various object categories. Given an image pair from ImageNet, our core idea is to follow multiple instance learning (MIL [10, 2, 5, 43, 40]) pipelines to treat images as bags and point-to-point matching pairs as instances to train binary matching classifiers. We select ImageNet ILSVRC training set as our database and incorporate the MIL pipeline in OICR [36] and MEFF [11] to decompose the problem into the image-level matching stage, region-level stage, and pixel-level stage. Now the problem becomes two folds: 1) How to gradually refine the correspondences during the learning process? 2) How to select and filter out images with salient feature points from ImageNet?

To learn reliable correspondences from the ImageNet dataset, we propose a novel cascaded online correspondence refinement algorithm to integrate the image-level, region-level, and pixel-level matching modules into a single network and train it in an end-to-end manner. As shown in Figure 2, in the image-level matching module, image labels are used to judge whether appropriate semantic correspondence exists in image pairs, and correspondences with high confidence in positive image pairs are identified as reliable correspondences. The region-level matching module accepts supervision from the previous stage and conducts robust region-matching to improve the matching accuracy further. Finally, the pixel-level matching module gets better supervision from the previous stage and trains the semantic matching head as previous methods [44, 18]. To ensure learning efficiency, different from state-of-the-art semantic matching algorithms [34, 29, 44, 6], we design a matching pipeline consisting of a transformer feature backbone, a gated cross attention module, and a correlation aggregation module, which is proven to be very powerful. Another critical point in our multiple-instance learning pipeline is to design a correspondence filtering and refinement module to improve correspondences in different stages. We incorporate a saliency detector, SelfReformer [41], to segment foreground objects and employ regularized Hough matching (RHM) [27] to ensure further matching consistency.

After learning correspondences from image-level annotations, we generated a new dataset with high-quality pseudo-labels for semantic correspondence and called it semantic correspondence ImageNet (SC-ImageNet). In ImageNet, some image categories are not object-centric, lack salient and unique feature points, or contain too complex background clutters, which are unsuitable for semantic correspondence learning. We remove these object categories and filter out 679 image categories manually. To further improve the dataset quality, we compute the matching quality of each image and select images that contain more high-quality salient feature points to formulate our new dataset. The matching quality is measured by the matching consistency index, which is introduced with details in subsequent sections. We select the top 30% images to formulate the SC-ImageNet and use it as a pre-training set for several popular semantic matching methods. Experiments demonstrate that our cascaded online correspondence refinement network trained in a weakly supervised manner can achieve competitive results with their supervised counterpart on PF-PASCAL and PF-WILLOW benchmarks. And if we pre-train state-of-the-art algorithms on SC-ImageNet and fine-tune them in the common fully supervised setting, we get impressive improvements on various datasets.

In summary, our contribution can be written as follows:

- We introduce a new weakly supervised semantic correspondence scheme, which can learn pixel-level matching relationships in image pairs. It provides a new perspective for semantic matching methods to learn reliable correspondences from large-scale, weakly annotated datasets.
- We propose a novel cascaded online correspondence refinement pipeline that integrates multiple instance learning and correspondence filtering and refinement into a single neural network that can be trained end-to-end.
- We build a new dataset based on ImageNet for semantic correspondence, called SC-ImageNet. It contains 679 object categories, 113,516 images, and 794,612 semantic correspondence pairs, and is much larger than existing benchmarks. Experiments with state-of-the-art algorithms indicate that SC-ImageNet pre-trained models show strong generalization ability and can benefit the subsequent fine-tuning process.

2. Related Work

Semantic Correspondence. Semantic Correspondence aims to find associations between different instances from the same object category. Unlike previous image-matching tasks, semantic matching methods focus on finding semantic associations between object instances. According to the usage of deep-level features, the semantic correspondence method can be first divided as handcrafted feature designing based methods [24, 4, 23, 27] and end-to-end learning methods [26, 44, 6, 18, 14]. Feature designing based methods usually use the frozen pretrained backbone to first extract features. Before calculating the similarity matrix and generating correspondence, several selection [27] or calculation [24, 4, 23] based preprocessing are often used to improve the discriminative of semantic representation among patches. While the end-to-end learning methods can enhance the deep feature in a dynamic and implicit way and the deep semantic descriptor can update its parameters as the task requires during the training process. Current state-of-the-art methods [26, 44, 6, 18, 14] are mostly designed in this way. However, the issue deviations from the goal of semantic correspondence is that, current benchmarks such as SPair-71K [27] and PF-PASCAL [13] only provide sparse keypoint annotation. This drawback leads to the performance limitation of most methods until now. As a result, it is necessary to propose a way to boost the pair annotation with low cost. Hence, we propose a new coarse-to-fine MIL method which can generate sufficient annotation with only image-level annotation (categories) required.

Weakly Supervised Learning. Currently fully supervised semantic correspondence methods [6, 44, 18] have achieved impressive performance. However, since it is difficult to identify the accurate matching between images, the datasets are always insufficient. To expand this work's domain, weakly-supervised algorithms requiring only object-level or image-level annotation are important. In which, work [34, 29, 33] based on image-level annotation tried to obtain the pixel-to-pixel matching relationship by comparing positive and negative samples. While DISCOBOX [20] proposed a collaborative training method which optimized the implicitly used matching similarity matrix by supervising object-level bounding boxes and masks. And work [17, 16, 1] tried to directly boost the matching pairs with further training procedure on them to obtain better performance. Following the idea of ”weakly-supervised” and label boosting, we propose a weakly supervised semantic correspondence pipeline, which utilizes different level of supervision.

3. Method

The proposed pseudo-label generation framework tried to build up a coarse-to-fine pipeline (shown in Figure 3), which used a three stage selection pipeline to find out accurate pixel-to-pixel (p2p) matching with only image-level instructions. During the whole process, we followed the idea of multiple instance learning (MIL). For the first stage, we tried to extract coarse regions $(\mathbf{R}_s, \mathbf{R}_t)$ which had potential pixel pairs to be matched on given image pair $(\mathbf{I}_s, \mathbf{I}_t)$. Afterwards, a region-level selection was used to find out a suspect pixel pair $(\mathbf{P}_s, \mathbf{P}_t)$ for each $(\mathbf{R}_s, \mathbf{R}_t)$. Finally, the output $(\mathbf{P}_s, \mathbf{P}_t)$ was used to supervise the generation of pixel-to-pixel correspondence. After training from this pipeline, we can obtain a pixel-to-pixel correspondence generator which can provide pseudo-labels for semantic correspondence method pretraining. That is, we only used the category of each image and can train a p2p pseudo-label gen-

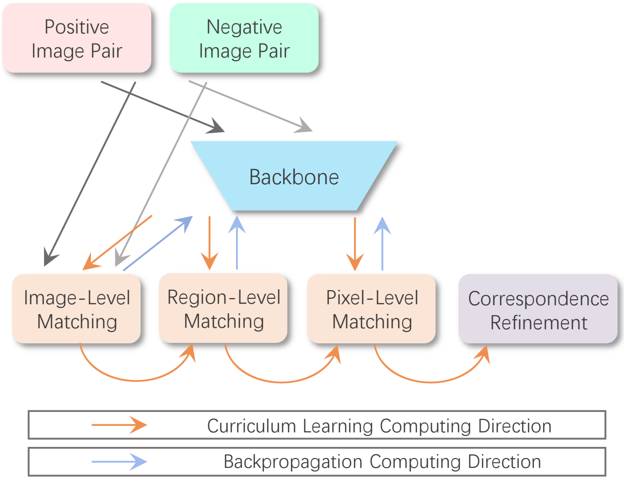


Figure 3. An overview of our three stage multi-instance learning pipeline with image-level, region-level and pixel-level refinement.

erator. In the following parts, we will first introduce the proposed multiple instance learning pipeline, from image-level, region-level, to pixel-level. Then, we will analyze the pseudo-label generation process, finally, an overview of proposed SC-ImageNet dataset will be given.

3.1. Design of Modules

Here we first briefly introduce the common modules (shown on Figure 4) used by all stages namely matching volume aggregation module (MVAM) and pseudo matching pair selection module (PMPS) to make it clear for following introduction of the coarse-to-fine refinement pipeline.

Matching Volume Aggregation Module. This module is designed to build up a refined similarity volume ($\mathbf{v}_f$) from a feature pair $(\mathbf{F}_s, \mathbf{F}_t)$ extracted by iBOT [47] backbone from $(\mathbf{I}_s, \mathbf{I}_t)$ or $(\mathbf{R}_s, \mathbf{R}_t)$. The module can be divided into three components namely the feature level aggregation $\Psi(\mathbf{F}_s, \mathbf{F}_t)$, matching volume calculation $\Phi(\widetilde{\mathbf{F}}_s, \widetilde{\mathbf{F}}_t)$ and volume level aggregation $\rho(\mathbf{v}_r)$ shown on Figure 4. The $\Psi(\cdot)$ for $s \rightarrow t$ direction can be represented as in Eq.1.

$$\begin{aligned} \mathbf{G}_s^l &= \sigma(\mathbf{F}_s \otimes \mathbf{T}_t), \\ \mathbf{G}_t^l &= \sigma(\mathbf{F}_t \otimes \mathbf{T}_s), \\ \widetilde{\mathbf{F}}_s &= \mathtt{MHA}(\mathbf{F}_s^l, \mathbf{F}_t^l \odot \mathbf{G}_t^l) \odot \mathbf{G}_s^l, \end{aligned} \tag{1}$$

in which $\sigma(\cdot)$ stands for the sigmoid activate function, $\mathbf{T}_t, \mathbf{T}_s$ stand for the extracted token for global semantic representation, $\widetilde{\mathbf{F}}_s, \widetilde{\mathbf{F}}_t$ stand for the refined features, $\otimes$ stands for vector inner product, $\odot$ stands for token-wised product and $\mathtt{MHA}(\cdot)$ stands for multi-head attention.

After calculating the cosine similarity of $\widetilde{\mathbf{F}}_s$ & $\widetilde{\mathbf{F}}_t$ (w/o cls tokens) and reshaping that into 4D format, we acquired the similarity volume $\mathbf{v}_r \in \mathbb{R}^{h_s \times w_s \times h_t \times w_t}$. Following the work [14], we introduced the 4D window attention structure to design the volume level aggregation block represented as $\rho(\cdot)$. In order to prevent high computational costs, we used 4D space downsampling for input similarity volume, and used feature channel redistribution to restore the initial shape of the similarity volume afterwards. With the design of our MVAM, we can acquire a similarity volume $\mathbf{v}_f \in \mathbb{R}^{h_s \times w_s \times h_t \times w_t}$ with significant match/mismatch similarity score differences for subsequent pseudo-label selection.

Pseudo Matching Pair Selection Module. With the accurate similarity volume acquired, we try to extract possible next-level (such as the region-level compared with image-level) correspondence for fine-grained supervision. In order to achieve such a goal, it is necessary to select the similarity extremum with neighborhood consistency on the obtained similarity volume. Inspired by previous works [27, 26], here we introduced the correspondence hough voting mechanism. In which, similarity volumes are converted into hough space and the neighborhood hough matching consistency scores are used to re-determine the similarity of a matching pair. Afterward, a one-to-one greedy selection method is used to generate final pseudo matchings.

3.2. C2F Multi-Instance Learning

Image-Level Refinement Pipeline. For image-level calculation, we design a positive-negative contrastive learning schema. We first build positive pairs (with the same classes) and negative pairs (with different classes) for further process. After the cascaded MVAM and PMPS module, we can acquire several rough estimations of suspected correspondence for each pair. However, these matching relationships are not accurate, which only indicate that there may be correlations within a certain range of source and target points. Thus, we firstly use a saliency detector to make sure that the selected suspects are all from foreground. And then we use a neighborhood window of 5×5 to represent the possible matching option regions for each suspected pair. Finally, such windows are the input for region-level refinement.

Region-Level Refinement Pipeline. For region-level refinement, our goal is to find out the pixel-to-pixel (p2p) pairs with the highest confidence. Importantly, only one pixel pair will be chosen for each input region. Same to the image-level refinement, we also generate positive and negative samples in the granularity of regions. After the same MVAM, PMPS, and foreground constraint, suspected p2p pairs are selected. Here we directly extract the keypoint pair with the highest similarity in output 4D volume. According to the recorded position, we assemble p2p matching pairs back to the whole image. With further non-maximum suppression, 16 highly confident pairs are selected.

Pixel-Level Pseudo Generator. With such 16 selected pairs, here we tried to train a pseudo-label generator. This stage inherits basic structure from previous levels. The pixel-level pseudo generator uses the image pairs as the input and outputs a trained similarity volume like previous

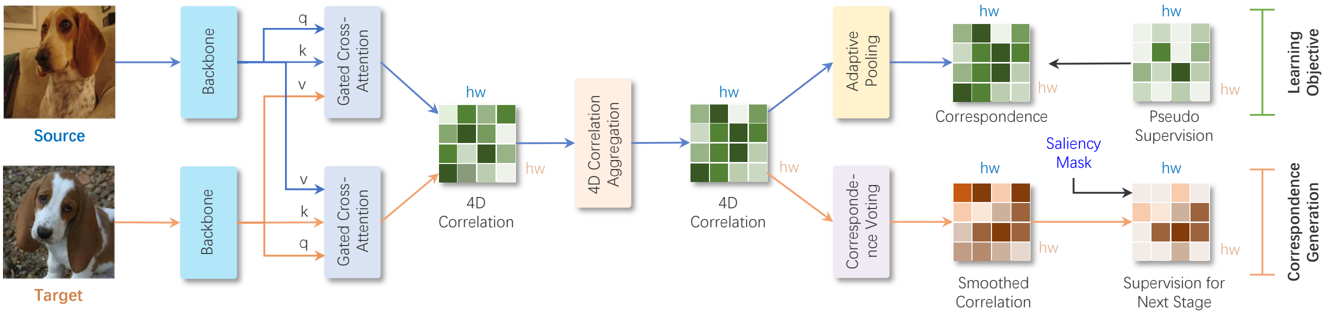


Figure 4. **An overview of pseudo label generation Pipeline.** The Pipeline contains two modules as Matching Volume Aggregation Module (MVAM) to generate accurate similarity volume and Pseudo Matching Pair Selection Module (PMPS) to select potential fine-grained correspondence for supervision in next satge.

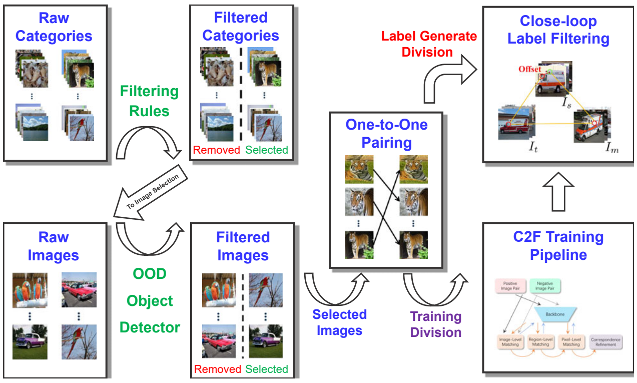


Figure 5. An overview of our dataset construction pipeline with the usage of proposed modules.

semantic correspondence methods [23, 44]. We note that, although the predicted similarity volume can generate pixel-to-pixel prediction, a further selection is still required.

3.3. Training

In our pipeline, only the categories of images are used as supervision. For the image-level and region-level supervision, we designed a contrastive learning task to maximize the differences between the refined similarity volumes with and without matching relationships (as in Eq. 2).

$$\mathcal{L} = \frac{1}{N} \sum_{i \in (1,N)} -\mathtt{AP}(\mathbf{v}_f^i[pos]) + \mathtt{AP}(\mathbf{v}_f^i[neg]), \tag{2}$$

in which $\mathtt{AP}(\cdot)$ stands for adaptive pooling on refined similarity volume $\mathbf{v}_f$. Importantly, the negative pair is generated by batch shifting as work [34]. And for pixel-level supervision, we use the binary cross entropy loss as in [44] to supervise the pixel-level pipeline with correspondences selected from region-level. Finally, the final loss is the weighted average of image-level loss ($\mathcal{L}_I$), region-level loss ($\mathcal{L}_R$), pixel-level loss ($\mathcal{L}_P$) as shown on Eq. 3.

$$\mathcal{L}_{all} = \lambda_1 \mathcal{L}_I + \lambda_2 \mathcal{L}_R + \lambda_3 \mathcal{L}_P. \tag{3}$$

3.4. Close-loop Consistency Restriction

After the training process, here we tried to use pixel-level pseudo generator to generate labels for ImageNet dataset. In this process, we needed to further select image pairs with accurate pseudo p2p pairs. To this end, we introduced a match closed-loop check method to select final image pairs and their pseudo-labels. During the inference time, we introduced the third picture $\mathbf{I}_m$ for the pair $(\mathbf{I}_s, \mathbf{I}_t)$. For a specific point $\mathbf{P}_s$ selected from $\mathbf{I}_s$, we can acquire the $\mathbf{P}'_s$ with the matching path of $\mathbf{I}_s \rightarrow \mathbf{I}_m \rightarrow \mathbf{I}_t \rightarrow \mathbf{I}_s$. We used the offset between $\mathbf{P}_s$ and $\mathbf{P}'_s$ as final score to evaluate whether image pairs (using the average scores) and single matching pair can be selected. Finally, we can acquire 794,612 image pairs with 32 pseudo keypoint pairs for each.

3.5. SC-ImageNet Dataset

With the listed modules designed, we build SC-ImageNet from a large scale classification dataset, ImageNet [7], which contains 1,281,167 images of 1,000 categories. Here we provide an overview of our dataset construction pipeline in Figure 5. Before the raw ImageNet becomes appropriate for semantic matching, three levels of selection are required. Firstly, some classes are not suitable for single-object semantic correspondence task. For example, there are usually crowded instances in images from category ‘conch’. At the same time, it is hard to locate salient and unique key-point in images which are labeled as ‘lakeside’. Therefore, we remove 321 categories from ImageNet. Afterwards, an out-of-distribution object detector [8] trained on MS COCO [21] is used to select images with only one high confidence (> 0.9) object. Such filtering results in 379,458 single-instance images from 679 classes. During training, images from the same category are randomly paired, and an image can only be used once. As for pseudo-labels generation, the method introduced in section 3.4 is performed 100 times for an image, and images with top-30% consistency score are selected. Eventually, 113,516 highly consistent single-instance images of 679 category compose 794,612 image pairs.

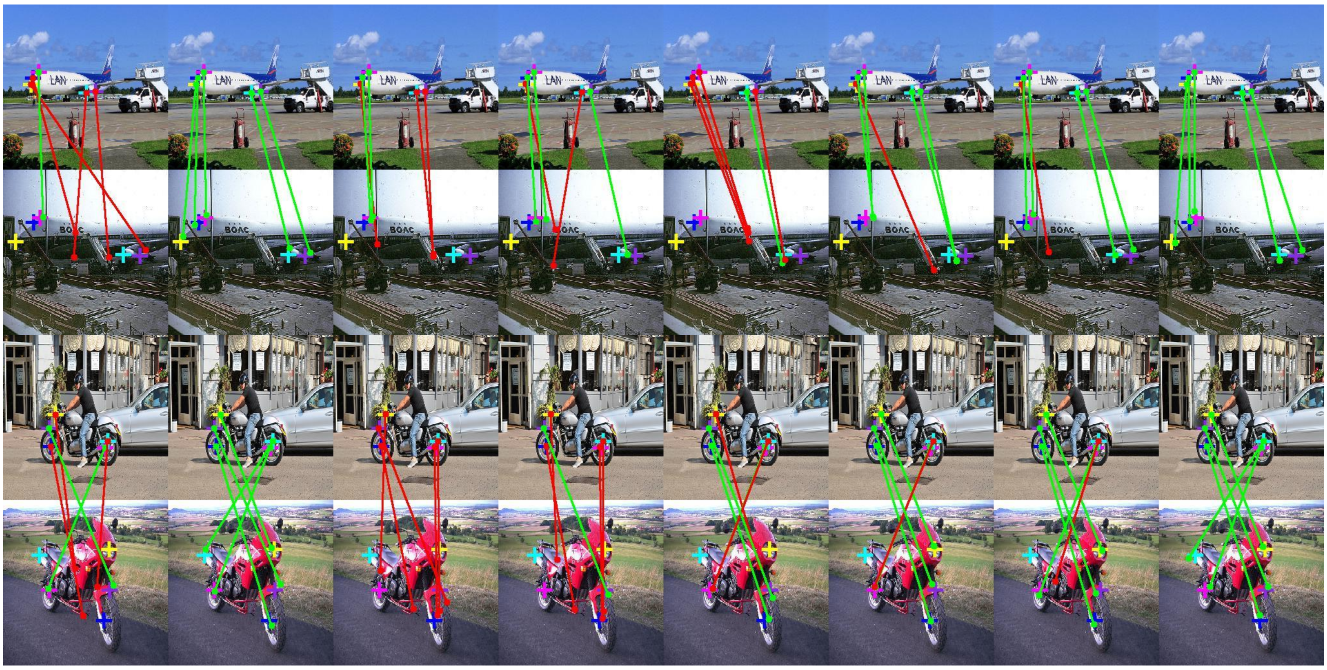


Figure 6. **Qualitative results on SPair-71k [27].** TM means TransforMatcher [18]. Other methods are proposed in CATs[6], DHPF[29] and VAT[14], respectively. +Ours demonstrates semantic correspondence models are firstly pre-trained with our SC-ImageNet, then fine-tuned on SPair-71k. Colored cross indicates corresponding key-point (ground-truth). Green dots and lines mean correct matches, while red dots and lines mean incorrect matches (measured by PCK@0.1 as in Table 2).

Method	Train Set	PF-PASCAL		PF-WILLOW	
		0.05	0.10	0.05	0.10
NCNet [34]	PF-P.	53.9	78.9	52.7	84.3
DHPF [29]	PF-P.	56.1	82.1	50.2	80.2
Ours	FImN.	56.8	81.5	57.6	85.3
Ours	PF-P.	58.3	86.2	60.7	88.7

Table 1. **Quantitative evaluation of pseudo-label generator on PF-PASCAL and PF-WILLOW dataset.** PF-P. for PF-PASCAL, FImN. for Filtered ImageNet [7] with inappropriate categories and multi-instance images removed. Note that DHPF [29] are trained with weak supervision.

4. Experiments

To prove the effectiveness of our pseudo-label generator as well as the proposed SC-ImageNet, we hold on several experiments. Firstly, we compared the pseudo-label generator directly with weakly supervised method, then we evaluated the performance of state-of-the-art (SOTA) methods, such as CATs [6] and VAT [14], on PF-PASCAL, PF-WILLOW [13], and SPair-71k [27] with SC-ImageNet pre-training, which were compared to models trained with their original settings. And to demonstrate the correctness of our pseudo-label generation method, we designed a series of ablation studies for crucial components in the workflow of pseudo-label generating.

4.1. Implementation Details

For the implementation of pseudo-label generator, our method is trained with an SGD optimizer, where learning rate is set to 3×10^{-5} with momentum of 0.9 for all layers except the frozen iBOT-B [47] backbone. In addition, the pseudo-label generator is built on PyTorch-GPU [30] with 8 NVidia RTX 3090 GPU, and the batch-size is 4 for each GPU, where each input image is resized to 512×512. Furthermore, SelfReformer [41] trained on DUTS-TR [38] is used for saliency detection in our coarse-to-fine pipeline. While for the pre-training and fine-tuning of SOTA, we followed the vanilla implementation. Note that, the hyper-parameters, data augmentation settings, and input resolution are all identical to original ones for fair comparison.

4.2. Experimental Settings

Datasets. We selected the widely used semantic correspondence datasets such as SPair-71k [27] and PF-PASCAL [13] to hold on experiments. To further evaluate the generalization ability of algorithms, experiments with PF-PASCAL training/fine-tuning and PF-WILLOW [13] testing were also introduced. PF-PASCAL is composed of 1,345 image pairs, which are selected from 20 object classes, split into 700 (train), 300 (val), 300 (test) pairs, respectively. As a complement, there are another 900 image pairs of 10 classes

Method	Supervision	Learning signal	PF-PASCAL			PF-WILLOW			Spair-71k
			0.05	0.10	0.15	0.05	0.10	0.15	0.10
ProposalFlow[13]	None	-	31.4	62.5	79.5	28.4	56.8	68.2	-
CNNGeo[32]	Self-sup.	Synthetic pairs	41.0	69.5	80.4	36.9	69.2	77.8	20.6
A2Net[35]			42.8	70.8	83.3	36.3	68.8	84.4	22.3
Weakalign[33]	Weak-sup.	Images	49.0	74.8	84.0	37.0	70.2	79.9	20.9
NC-Net[34]			53.9	78.9	86.0	52.7	84.3	92.2	20.1
DHPF[29]			56.1	82.1	91.1	50.2	80.2	91.1	27.7
DHPF[29]	Sup.	Keypoints	75.7	90.7	95.0	49.5	77.6	89.1	37.3
CATs[6]			75.4	92.6	96.4	50.3	79.2	90.3	49.9
TransforMatcher[18]			80.8	91.8	95.2	48.9	76.0	86.1	53.7
VAT[14]			78.2	92.3	96.2	52.8	81.6	91.4	55.5
CATs[6]+SemiMatch[17]	Semi-sup.	Keypoints	80.1	**93.5**	96.6	54.0	82.1	92.1	50.7
SCorrSAN[16]			81.5	93.3	96.6	54.1	80.0	89.8	55.3
DHPF[29]+Ours	Weak-sup. (GP.) Semi-sup. (PT.&FT.)	Images (GP.) Keypoints (PT.&FT.)	76.5	90.7	95.3	50.8	77.9	90.2	39.5
CATs[6]+Ours			78.4	92.9	96.6	53.2	84.0	**94.5**	54.6
TransforMatcher[18]+Ours			82.3	92.4	95.8	49.9	77.2	87.4	56.3
SCorrSAN[16]+Ours			**84.4**	93.4	**96.8**	**57.4**	82.3	91.6	58.9
VAT[14]+Ours			80.5	93.0	96.7	57.1	**85.1**	94.1	**60.3**

Table 2. **Quantitative evaluation on PF-PASCAL, PF-Willow [13] and SPair-71k [27].** The best results in bold, and the second best results are underlined. Note that the input resolution of original algorithms is not modified. GP. is short for the generation of pseudo-labels. PT. & FT. means pre-train and fine-tune.

Method	aero.	bike	bird	boat	bott.	bus	car	cat	chai.	cow	dog	hors.	mbik.	pers.	plan.	shee.	trai.	tv	all
CNNGeo[32]	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[35]	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[33]	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[34]	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
DHPF (Sup.)[29]	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
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
TransforMatcher[18]	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
VAT[14]	58.8	40.0	75.3	40.1	52.1	59.7	44.2	69.1	23.3	75.1	61.9	57.1	46.4	49.1	**51.8**	41.8	80.9	70.1	55.5
CATs[6]+SemiMatch[17]	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
SCorrSAN[16]	57.1	40.3	78.3	38.1	51.8	57.8	47.1	67.9	25.2	71.3	63.9	49.3	45.3	49.8	48.8	40.3	77.7	69.7	55.3
DHPF[29]+Ours	40.5	25.9	70.2	23.2	41.0	33.5	23.1	62.1	22.4	48.8	46.9	36.6	32.0	43.3	28.6	27.0	56.9	47.7	39.5
CATs[6]+Ours	58.8	44.6	71.7	41.0	49.2	71.2	46.4	73.0	23.3	69.8	58.3	59.8	55.8	39.2	33.0	41.6	75.8	73.9	54.5
TransforMatcher[18]+Ours	**66.5**	49.8	72.5	39.4	**53.7**	67.3	**55.4**	70.2	**31.3**	71.8	60.2	54.7	53.6	34.8	41.5	39.9	82.9	69.2	56.3
SCorrSAN[16]+Ours	64.5	47.5	78.0	39.9	49.2	65.1	49.0	74.0	29.4	75.4	64.3	60.9	54.8	52.0	48.1	**47.4**	**87.0**	**75.4**	58.9
VAT[14]+Ours	65.2	**50.1**	**81.0**	**43.3**	51.2	**76.1**	48.9	**79.1**	18.4	**80.8**	**67.4**	**63.8**	**57.6**	**57.6**	45.4	46.2	82.7	74.5	**60.3**

Table 3. **Per-class quantitative evaluation on SPair-71k dataset [27].** The best results are in bold, and the second best results are underlined. Note that the input resolution of original algorithms is not modified.

in PF-WILLOW for further evaluation. SPair-71k is a large-scale dataset containing 70,958 image pairs from 18 classes with obvious intraclass variations, scale difference, occlusion and truncation. In SPair-71k, 53,340 image pairs are processed during training, while 5,384 image pairs are designed for validation, and there are 12,234 image pairs in the test set. However, except for so many image pairs, these datasets contain limited number of images. For instance, 1,800 images make up SPair-71k.

Evaluation Metric. Following pervious works [6, 14, 16, 18, 29], PCK@α (percentage of correct keypoints with threshold α) is employed as our evaluation metric. During the calculation of PCK@α, a predicted key-point is considered correct when it falls into the circle of radius $\alpha \times d$ centering at its ground-truth counterpart, where d is the length of longer side of image (PF-PASCAL) or object bounding box (PF-WILLOW and SPair-71k), at the same time, α is a hyper-parameter standing for precision.

4.3. Matching Results

To begin with, since there are not sparse annotations in the ImageNet [7], we evaluate the performance of our pseudo-label generator on the test set of PF-PASCAL and PF-WILLOW [13]. As shown in Table 1, when training on PF-PASCAL with only image-level supervision, our generator outperforms current best weakly supervised semantic matching algorithm (weakly-supervised DHPF [34]) by 2.2%/10.5% PCK@0.05 and 4.1%/8.5% PCK@0.10 on PF-PASCAL and PF-WILLOW. In addition, our generator can still outperform the DHPF (weak-sup.) by 0.7%/7.4% PCK@0.05 on PF-PASCAL and PF-WILLOW when just training on filtered ImageNet without further fine-tuning.

Next, we compared the performance of state-of-the-art models with pre-training on SC-ImageNet to those with original training settings to prove that the high-quality pseudo-labels can help. Specifically, SOTA models are firstly pre-trained on SC-ImageNet. Afterwards, pre-trained

Component		PF-PASCAL		PF-WILLOW	
		0.05	0.10	0.05	0.10
(I)	Full Procedure	56.8	81.5	57.6	85.3
(II)	(I) w/o SD	55.9	80.7	56.4	84.7
(III)	(I) w/o RHM	55.6	79.1	56.1	84.0
(IV)	(I) L(RHM) = 3	56.0	80.8	56.7	84.9
(V)	(I) L(RHM) = 5*	56.8	81.5	57.6	85.3
(VI)	(I) L(RHM) = 7	56.2	81.1	56.8	85.0
(VII)	(I) Shared Attn.	52.6	74.7	53.6	78.5

Table 4. **Ablation study for network components of pseudo-labels generator.** All model are trained on ImageNet [7]. SD for Saliency Detection, L(RHM) for hyperpixel length in the RHM module [27], and Shared Attn. for shared weighted gated cross-attention blocks in three stages. *: Case (V) equals to Case (I).

models are fine-tuned with target dataset, such as SPair-71k and PF-PASCAL, for final evaluation on accordant test set. Compared with vanilla ones, models pre-trained with SC-ImageNet can predict more precise key-point, especially on the test set of SPair-71k, which is the most challenging for its obvious intraclass variations, scale difference, occlusion and truncation. Such results in Table 2 demonstrate that with proposed pre-training settings, the effectiveness and robustness of models are significantly improved. Furthermore, according to Table 3, the majority of outstanding accuracy is given by pre-trained models, providing more evidence for the robustness and effectiveness of our SC-ImageNet. Our qualitative results are shown in Figure 6.

In addition, our pre-training strategy with additional automatically labeled images could be seen as a semi-supervised approach. Nevertheless, SOTA semi-supervised semantic correspondence methods [17, 16] mainly propose strategies with additional key-point pairs. Hence, we conduct experiments to analyze their congruities and divergences. As shown in Table 2 and Table 3, our method is competitive in comparison to SemiMatch [17], especially in the challenging SPair-71k [27] dataset. At the same time, comparing to vanilla settings, SCorrSAN [16] pre-trained with our SC-ImageNet performs better on all target datasets, which indicates that cooperation between additional image pairs and additional key-point pairs is beneficial for models' effectiveness, robustness and generalization ability.

4.4. Ablation Study

Effects of Network Components. In Table 4, we evaluate network components in our pseudo-label generator (I) by removing each from the full pipeline. From (I) to (II), PCK decline demonstrates that with the help of saliency detection, our model can mainly focus on foreground. Besides, a larger performance drop in the comparison between (I) and (III) indicates that geometry information in the Hough space is critical to region-level correspondence. Furthermore, (IV) $\sim$ (VI) shows that appropriate hyperpixel size is helpful to RHM, where small hyperpixel lacks in geometry information and big hyperpixel leads to waste of computational resources. However, the most significant PCK drop occurs when gated cross-attention blocks share the same weight across all three stages, as a consequence, features in different stages should be processed independently.

Network Architecture		PF-PASCAL		PF-WILLOW	
		0.05	0.10	0.05	0.10
(I)	Full Procedure	56.8	81.5	57.6	85.3
(II)	(I) IIR=256 × 256	53.2	79.4	55.7	84.6
(III)	(I) IIR=384 × 384	54.3	80.2	54.7	84.4
(IV)	(I) IIR=512 × 512*	56.8	81.5	57.6	85.3
(V)	(I) w/o Stage 3	56.1	80.5	57.0	84.6
(VI)	(V) w/o Stage 2	51.0	77.4	55.2	83.0

Table 5. **Ablation study for stages and input resolution of pseudo-labels generator.** All model are trained on ImageNet [7]. IIR means input image resolution. *: Case (IV) equals to Case (I).

Effects of Network Architectures and Input Resolution. We also explore the importance of each network stage and input resolution to our pseudo-label generator. As shown in Table 5, small input resolution results in slight performance drop in PF-PASCAL, according to (II) $\sim$ (IV). However, when tested on PF-WILLOW, such slight drop even disappear in (II) and (III), which demonstrates powerful generalization ability of our model. As for network architecture, (V) shows that pixel-level matching plays a critical role in revising the location of predicted key-point. And (VI) demonstrates that region-level matching is fundamental to our method because of the essentiality of geometry information in weakly-supervised semantic correspondence.

5. Conclusion

In this paper, we have presented a novel weakly-supervised learning scheme, which investigates pixel-level matching relationships in image pairs from large-scale, weakly annotated datasets. We also introduce a novel cascaded online correspondence refinement pipeline to integrate semantic correspondence relationship from image-level, region-level, to pixel-level, with a single end-to-end neural network. On this basis, we build SC-ImageNet from ImageNet [7], the largest semantic correspondence dataset so far, containing 679 categories, 113,516 images as well as 794,612 pairs. Experiments on SOTA algorithms indicate that SC-ImageNet pre-trained models show strong generalization ability and can benefit the subsequent fine-tuning.

6. Acknowledgment

This work was supported by National Natural Science Foundation of China under Grant (No.62106051), Shanghai Pujiang Program (No.21PJ1400600), and National Key R&D Program of China (2022YFC3601405). This paper is also supported by UniDT's Cognitive Computing and Few Shot Learning Project.

References

[1] Shir Amir, Yossi Gandelsman, Shai Bagon, and Tali Dekel. Deep vit features as dense visual descriptors. *arXiv preprint arXiv:2112.05814*, 2(3):4, 2021. 3

[2] Stuart Andrews, Ioannis Tsochantaridis, and Thomas Hofmann. Support vector machines for multiple-instance learning. *Advances in neural information processing systems*, 15, 2002. 2

[3] Nikita Araslanov and Stefan Roth. Single-stage semantic segmentation from image labels. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4253–4262, June. 2

[4] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In *European conference on computer vision*, pages 404–417. Springer, 2006. 3

[5] Marc-André Carbonneau, Veronika Cheplygina, Eric Granger, and Ghyslain Gagnon. Multiple instance learning: A survey of problem characteristics and applications. *Pattern Recognition*, 77:329–353, 2018. 2

[6] Seokju Cho, Sunghwan Hong, Sangryul Jeon, Yunsung Lee, Kwanghoon Sohn, and Seungryong Kim. Cats: Cost aggregation transformers for visual correspondence. *Advances in Neural Information Processing Systems*, 34:9011–9023, 2021. 2, 3, 6, 7

[7] 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, 2009. 1, 2, 5, 6, 7, 8

[8] Xuefeng Du, Zhaoning Wang, Mu Cai, and Yixuan Li. Vos: Learning what you don't know by virtual outlier synthesis. *Proceedings of the International Conference on Learning Representations*, 2022. 5

[9] Mark Everingham, SM 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. 1

[10] James Foulds and Eibe Frank. A review of multi-instance learning assumptions. *The knowledge engineering review*, 25(1):1–25, 2010. 2

[11] Weifeng Ge, Sibei Yang, and Yizhou Yu. Multi-evidence filtering and fusion for multi-label classification, object detection and semantic segmentation based on weakly supervised learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1277–1286, 2018. 2

[12] Sheng Guo, Weilin Huang, Haozhi Zhang, Chenfan Zhuang, Dengke Dong, Matthew R Scott, and Dinglong Huang. Curriculumnet: Weakly supervised learning from large-scale web images. In *Proceedings of the European conference on computer vision (ECCV)*, pages 135–150, 2018. 2

[13] 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, 3, 6, 7

[14] Sunghwan Hong, Seokju Cho, Jisu Nam, Stephen Lin, and Seungryong Kim. Cost aggregation with 4d convolutional swin transformer for few-shot segmentation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIX*, pages 108–126. Springer, 2022. 1, 3, 4, 6, 7

[15] Lianghua Huang, Xin Zhao, and Kaiqi Huang. Got-10k: A large high-diversity benchmark for generic object tracking in the wild. *IEEE transactions on pattern analysis and machine intelligence*, 43(5):1562–1577, 2019. 1

[16] Shuaiyi Huang, Luyu Yang, Bo He, Songyang Zhang, Xuming He, and Abhinav Shrivastava. Learning semantic correspondence with sparse annotations. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XIV*, pages 267–284. Springer, 2022. 3, 7, 8

[17] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19699–19709, 2022. 3, 7, 8

[18] Seungwook Kim, Juhong Min, and Minsu Cho. Transformatcher: Match-to-match attention for semantic correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8697–8707, 2022. 2, 3, 6, 7

[19] Kazuma Kobayashi, Ryuichiro Hataya, Yusuke Kurose, Mototaka Miyake, Masamichi Takahashi, Akiko Nakagawa, Tatsuya Harada, and Ryuji Hamamoto. Decomposing normal and abnormal features of medical images for content-based image retrieval of glioma imaging. *Medical Image Analysis*, 74:102227, 2021. 1

[20] Shiyi Lan, Zhiding Yu, Christopher Choy, Subhashree Radhakrishnan, Guilin Liu, Yuke Zhu, Larry S Davis, and Anima Anandkumar. Discobox: Weakly supervised instance segmentation and semantic correspondence from box supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3406–3416, 2021. 3

[21] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 1, 2, 5

[22] Weide Liu, Chi Zhang, Guosheng Lin, and Fayao Liu. Crcnet: Few-shot segmentation with cross-reference and region–global conditional networks. *International Journal of Computer Vision*, pages 1–18, 2022. 1

[23] 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. 3, 5

[24] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004. 3

[25] Tim Meinhardt, Alexander Kirillov, Laura Leal-Taixe, and Christoph Feichtenhofer. Trackformer: Multi-object tracking with transformers. In *Proceedings of the IEEE/CVF Con-*

ference on Computer Vision and Pattern Recognition, pages 8844–8854, 2022. 1

[26] 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. 3, 4

[27] 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. 2, 3, 4, 6, 7, 8

[28] Juhong Min, Jongmin Lee, Jean Ponce, and Minsu Cho. Spair-71k: A large-scale benchmark for semantic correspondence, 2019. 1, 2

[29] Juhong Min, Jongmin Lee, Jean Ponce, and Minsu Cho. Learning to compose hypercolumns for visual correspondence. In *European Conference on Computer Vision*, pages 346–363. Springer, 2020. 2, 3, 6, 7

[30] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 6

[31] William Peebles, Jun-Yan Zhu, Richard Zhang, Antonio Torralba, Alexei A Efros, and Eli Shechtman. Gan-supervised dense visual alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13470–13481, 2022. 1

[32] 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. 7

[33] 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. 3, 7

[34] 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. 2, 3, 5, 6, 7

[35] 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 (ECCV)*, pages 349–364, 2018. 7

[36] Peng Tang, Xinggang Wang, Xiang Bai, and Wenyu Liu. Multiple instance detection network with online instance classifier refinement. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2843–2851, 2017. 2

[37] 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

[38] Lijun Wang, Huchuan Lu, Yifan Wang, Mengyang Feng, Dong Wang, Baocai Yin, and Xiang Ruan. Learning to detect salient objects with image-level supervision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 136–145, 2017. 6

[39] Peter Welinder, Steve Branson, Takeshi Mita, Catherine Wah, Florian Schroff, Serge Belongie, and Pietro Perona. Caltech-ucsd birds 200. 2010. 1

[40] Made Satria Wibawa, Kwok-Wai Lo, Lawrence S Young, and Nasir Rajpoot. Multi-scale attention-based multiple instance learning for classification of multi-gigapixel histology images. In *Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VII*, pages 635–647. Springer, 2023. 2

[41] Yi Ke Yun and Weisi Lin. Selfreformer: Self-refined network with transformer for salient object detection. *arXiv preprint arXiv:2205.11283*, 2022. 2, 6

[42] Dingwen Zhang, Junwei Han, Gong Cheng, and Ming-Hsuan Yang. Weakly supervised object localization and detection: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(9):5866–5885, 2021. 2

[43] Hongrun Zhang, Yanda Meng, Yitian Zhao, Yihong Qiao, Xiaoyun Yang, Sarah E Coupland, and Yalin Zheng. Dtfd-mil: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18802–18812, 2022. 2

[44] 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, 5

[45] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017. 1

[46] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017. 1

[47] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. *International Conference on Learning Representations (ICLR)*, 2022. 4, 6