Semantic Correspondence as an Optimal Transport Problem

Yanbin Liu[1], Linchao Zhu[1], Makoto Yamada[2,3], Yi Yang[1]
[1]ReLER, University of Technology Sydney, [2]RIKEN AIP, [3]Kyoto University
csyanbin@gmail.com, {linchao.zhu,yi.yang}@uts.edu.au, makoto.yamada@riken.jp

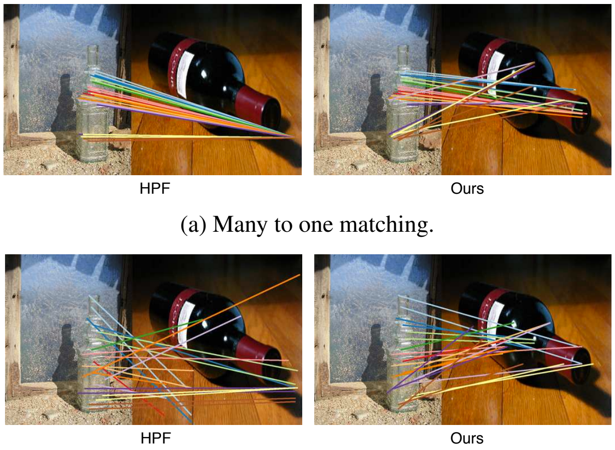
Abstract

Establishing dense correspondences across semantically similar images is a challenging task. Due to the large intra-class variation and background clutter, two common issues occur in current approaches. First, many pixels in a source image are assigned to one target pixel, i.e., ***many to one matching****. Second, some object pixels are assigned to the background pixels, i.e.,* ***background matching****. We solve the first issue by global feature matching, which maximizes the total matching correlations between images to obtain a global optimal matching matrix. The row sum and column sum constraints are enforced on the matching matrix to induce a balanced solution, thus suppressing many to one matching. We solve the second issue by applying a staircase function on the class activation maps to re-weight the importance of pixels into four levels from foreground to background. The whole procedure is combined into a unified optimal transport algorithm by converting the maximization problem to the optimal transport formulation and incorporating the staircase weights into optimal transport algorithm to act as empirical distributions. The proposed algorithm achieves state-of-the-art performance on four benchmark datasets. Notably, a 26% relative improvement is achieved on the large-scale SPair-71k dataset.*

1. Introduction

Establishing dense correspondences across semantically similar images is one of the fundamental tasks in computer vision that has potential applications such as semantic segmentation [35, 40], image registration [24], and image editing [7, 39]. This is a challenging task due to the large intra-class variation, viewpoint changes and background clutter.

Recent methods employ powerful image features from convolutional neural networks. Semantic flow approaches attempt to establish a flow field between images based on single [4, 33, 34] or multiple layers [30] feature maps. Semantic alignment methods cast semantic correspondence as a geometric alignment problem to regress the global transformation parameters using self-supervised [18, 32], weakly-supervised [21, 33] or keypoints [24] supervision.



(a) Many to one matching.

(b) Background matching.

Figure 1: We solve two problems caused by current approaches, such as HPF [30]. (a) Many pixels in a source image are assigned to one target pixel. (b) Some object pixels are assigned to the background pixels.

However, *many to one matching* problem and *background matching* problem hinder the development of semantic correspondence.

First, *many to one matching* occurs when many pixels in a source image are assigned to one target pixel. We solve this problem by global feature matching, which maximizes the total matching correlations between images. Most existing approaches [4, 19, 30, 33] for semantic correspondence rely on the correlation map which is computed by individual feature matching. The individual matching scheme does not care about the mutual relation between features, thus is sensitive to large intra-class variations and repetitive patterns. For example, in Figure 1a (Left), due to repetitive pattern in left bottle, the individual matching assigns many source pixels to one target pixel. Although, semantic alignment methods [21, 24, 32] try to suppress many to one matching by estimating the global transformation parameters, they are easily distracted by occlusion and non-rigid deformations.

In our method, maximizing the total matching correlations leads to a global optimal matching matrix, which is insensitive to repetitive patterns (e.g., Figure 1a (Right)). For the matching matrix, each row represents matching scores from a source pixel to all target pixels and each column represents scores from all source pixels to a target pixel. We enforce each row sum and column sum to be a fixed value according to the prior distributions of pixels. This avoids large values in a whole row or column, thus reducing the many to one matching.

Second, *background matching* happens when some object pixels are assigned to background pixels due to the intra-class appearance variation and background clutter, as shown in Figure 1b (Left). Recent methods deal with this by soft-inlier score [33] or attention [18], whereas they need special network design and rely on large amount of training data. In this paper, we reuse feature extraction network with neglected cost to obtain the class activation map (CAM), which is a good indicator for the foreground and background areas. However, the original CAM is not well calibrated for source and target images, e.g., same part of an object from two images may have different values. Therefore, we propose a staircase function to re-weight pixels of an activation map into four levels: hot spots, object, context and background with decreasing values. With staircase reweighting, background pixels are unlikely assigned to foreground, thus reducing the background matching.

We combine all proposed modules in a unified optimal transport framework. This is implemented by converting the correlation maximization to optimal transport formulation and incorporating the staircase weights to act as empirical distributions in optimal transport. We summarize the main contributions as follows:

- We model semantic correspondence as an optimal transport problem (SCOT) in a unified framework. The row sum and column sum constraints can be naturally incorporated to suppress many to one matching.
- We propose a staircase function applied on the class activation maps with neglected cost to suppress the background matching.
- The proposed algorithm achieves state-of-the-art performance on four benchmark datasets, especially a 26% relative improvement on the large-scale SPair-71k dataset.

2. Related Work

Semantic correspondence. Early works on semantic correspondence employ hand-crafted descriptors like SIFT [29] or HOG [8] together with geometric models [12, 23]. Cho *et al.* [3] use region proposals and HOG features in Probablistic Hough Matching (PHM) algorithm for semantic matching and object discovery. Ham *et al.* [13] extend this work with local-offset matching and introduce the PF-PASCAL benchmark with keypoint-level annotations.

Recent methods employ image features from convolutional neural networks. Many of them [4, 19, 25, 28, 30, 34] are semantic flow approaches that attempts to find correspondence for individual pixel or patches. Han *et al.* [19] develop a dynamic fusion strategy based on attention mechanism to obtain a context-aware semantic representation. Lee *et al.* [25] train a CNN for semantic correspondence by using images annotated with binary foreground masks. Min *et al.* [30] use beam search algorithm on validation split of the specific dataset to find the optimal subset of deep convolutional layers.

Other methods [15, 18, 32, 33] formulate semantic correspondence as a geometric alignment problem trained using different levels of supervision. Rocco *et al.* [32] propose a two-stage regression model that utilizes self-supervision from synthetically generated images. Rocco *et al.* [33] then develop a semantic alignment model that is end-to-end trainable from weakly supervised data. Laskar *et al.* [24] cast semantic correspondence as solving a 2D point set registration problem by using keypoint-level supervision. Different from these methods, the proposed algorithm does not rely on specific kind of supervision and is flexible to use either pre-trained or finetuned models.

Class activation map. The idea of generating class activation map (CAM) from a classification CNN model is first introduced by Zhou *et al.* [45]. They compute a weighted sum of the feature maps of the last convolutional layer to obtain the class activation maps. Zhang *et al.* [44] then provide a simple way by directly selecting the class-specific feature maps of the last convolutional layer and prove the equivalence to [45]. Gradient-weighted Class Activation Mapping (Grad-CAM) is proposed by Selvaraju *et al.* [36]. They utilize the gradients of any target concept to produce a coarse localization map highlighting the important regions in the image for predicting the concept.

Optimal transport. Optimal transport provides a way to infer the correspondence between two distributions. Recently, it has received great attention in various computer vision tasks. Courty *et al.* [5] solve domain adaptation problem by learning a transportation plan from source domain to target domain. Su *et al.* [38] employ optimal transport to deal with the 3D shape matching and surface registration problem. Other applications include generative model [1, 2, 10, 41], graph matching [42, 43], and etc. To the best of our knowledge, we are the first to model the semantic correspondence problem in optimal transport framework.

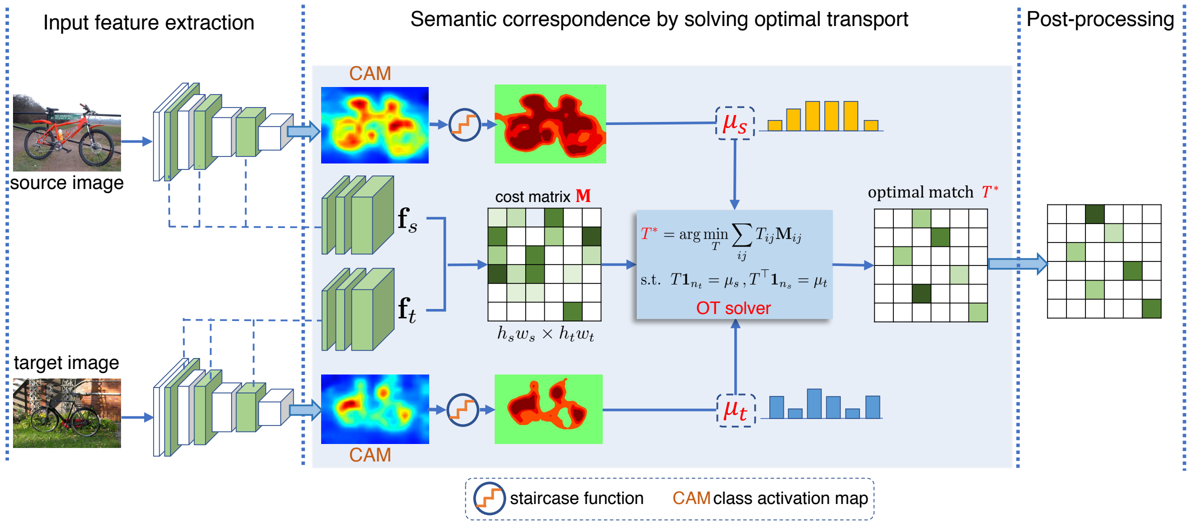


Figure 2: The proposed framework. A pair of images are input to the pre-trained CNN to get the multi-layer feature maps ($\mathbf{f}_s$,$\mathbf{f}_t$) and the class activation maps in a single forward pass. $\mathbf{f}_s$ and $\mathbf{f}_t$ are used to compute a cost matrix $\mathbf{M}$ representing the matching difference. Activation maps undergo a staircase function and are then normalized as the empirical probability distributions μ_s and μ_t. We deal with semantic correspondence by solving optimal transport problem to get the optimal match T^*, which is further post-processed to ensure geometric consistency.

3. Proposed Algorithm

In this section, we first introduce preliminary knowledge about optimal transport theory, then we describe how the semantic correspondence problem can be modeled in optimal transport framework, at last, we describe the implementation details about pre- and post-processing.

3.1. Preliminary

Optimal transport aims at computing a minimal cost transportation between a source distribution μ_s, and a target distribution μ_t. μ_s and μ_t are defined on probability space $X, Y \in \Omega$, respectively. When a meaningful cost function $c : X \times Y \mapsto \mathbb{R}^+$ is defined, the Kantorovich formulation [20] solves optimal transport by seeking for a probabilistic coupling $\boldsymbol{\pi} \in \mathcal{P}(X \times Y)$:

$$\boldsymbol{\pi}^* = \underset{\boldsymbol{\pi} \in \Pi(\mu_s, \mu_t)}{\arg\min} \int_{X \times Y} c(\boldsymbol{x}, \boldsymbol{y}) \boldsymbol{\pi}(\boldsymbol{x}, \boldsymbol{y}) d\boldsymbol{x} d\boldsymbol{y}\,, \tag{1}$$

where $\Pi(\mu_s, \mu_t) = \{\int_Y \boldsymbol{\pi}(x, y) d\boldsymbol{y} = \mu_s, \int_X \boldsymbol{\pi}(x, y) d\boldsymbol{x} = \mu_t, \boldsymbol{\pi} \geq \mathbf{0}\}$, i.e., $\boldsymbol{\pi}$ is the joint probability measure with marginals μ_s and μ_t.

Here we consider the case when those distributions are discrete empirical distributions, and can be written as

$$\mu_s = \sum_{i=1}^{n_s} p_i^s \delta(x_i)\,, \quad \mu_t = \sum_{i=1}^{n_t} p_i^t \delta(y_i)\,, \tag{2}$$

where $\delta(\cdot)$ denotes the Dirac function, n_s and n_t are the number of samples, p_i^s and p_i^t are the probability mass to the i-th sample, belonging to the probability simplex, i.e., $\sum_{i=1}^{n_s} p_i^s = \sum_{i=1}^{n_t} p_i^t = 1$. We define a cost matrix $\mathbf{M}$ with $\mathbf{M}_{ij}$ representing the distance between x_i and y_j. The optimal transport problem is:

$$\begin{aligned} T^* = & \underset{T \in \mathbb{R}_+^{n_s \times n_t}}{\arg\min} \sum_{ij} T_{ij} \mathbf{M}_{ij} \\ \text{s.t. } & T\mathbf{1}_{n_t} = \mu_s\,, T^\top \mathbf{1}_{n_s} = \mu_t\,. \end{aligned} \tag{3}$$

T^* is called the optimal transport plan or transport matrix. T_{ij} denotes the the optimal amount of mass to move from x_i to y_j in order to obtain an overall minimum cost.

3.2. Semantic correspondence as an OT problem

Given an input image pair $(\mathbf{I}_s, \mathbf{I_t})$ containing the same object, the goal of semantic correspondence is to estimate a matrix (e.g., T^* in Figure 2) representing the dense matching scores between pixels in two images. A key step in semantic correspondence is to compute the correlation map, which describes the matching similarities between any two locations from different images.

Correlation map A common strategy for computing correlation map is based on matching individual image features using cosine similarity. Given dense feature maps $\mathbf{f}_s \in \mathbb{R}^{h_s \times w_s \times D}$ and $\mathbf{f}_t \in \mathbb{R}^{h_t \times w_t \times D}$ of source and target images extracted from CNNs, the correlation map is computed as:

$$\mathbf{C} = \frac{\mathbf{f}_s \cdot \mathbf{f}_t^\top}{\|\mathbf{f}_s\| \|\mathbf{f}_t\|} \in \mathbb{R}^{h_s \times w_s \times h_t \times w_t}. \tag{4}$$

$\mathbf{C}_{ijkl}$ denotes the matching score between the (i, j)-th position in source feature map and (k, l)-th position in target feature map. The best match for (i, j) is computed as $\arg\max_{kl} \mathbf{C}_{ijkl}$.

In this process, each of the pairwise matching scores in position (i, j, k, l) is computed individually, without considering any mutual relation or additional constraints. However, since the large intra-class variation and background clutter are ubiquitous in semantic correspondence, this individual strategy often leads to two problems in matching (see Figure 1). Firstly, many source positions can be assigned to the same target position due to the individual $\arg\max$ assignment. This is an undesired property because for the same object it is more reasonable to match each part in source image to the corresponding part in target image, e.g., one-to-one matching. Secondly, foreground object may be assigned to the background due to high feature variation or illumination changes.

Optimal transport problem In this paper, instead of individual matching strategy, we model this problem from a global perspective. We first introduce a matrix $T \in \mathbb{R}^{h_s w_s \times h_t w_t}$ as the pairwise matching probability from source to target image. Then we resize correlation map $\mathbf{C}$ to the same shape as T and define the ***total correlation*** as $\sum_{ij} T_{ij} \mathbf{C}_{ij}$. Our goal is to maximize the total correlation to get a global optimal matching probability T^*. In order to avoid trivial solutions, we introduce the empirical distribution μ_s and μ_t as the probability of each point in source or target feature map. The values of μ_s and μ_t represent the importance of each point in feature map. Then the marginals of T are constrained to be μ_s and μ_t (i.e., the row sum of T is μ_s and column sum is μ_t). The problem is formulated as:

$$\begin{aligned} T^* &= \underset{T \in \mathbb{R}_+^{h_s w_s \times h_t w_t}}{\arg\max} \sum_{ij} T_{ij} \mathbf{C}_{ij} \\ \text{s.t. } & T \mathbf{1}_{h_t w_t} = \mu_s \, , T^\top \mathbf{1}_{h_s w_s} = \mu_t \, . \end{aligned} \tag{5}$$

If we define $\mathbf{M} = 1 - \mathbf{C}$ as the cost matrix denoting the matching difference, then Eq. 5 is equivalent to minimize the total matching difference:

$$\begin{aligned} T^* &= \underset{T \in \mathbb{R}_+^{h_s w_s \times h_t w_t}}{\arg\min} \sum_{ij} T_{ij} \mathbf{M}_{ij} \\ \text{s.t. } & T \mathbf{1}_{h_t w_t} = \mu_s \, , T^\top \mathbf{1}_{h_s w_s} = \mu_t \, , \end{aligned} \tag{6}$$

which is a standard optimal transport problem as in Eq. 3.

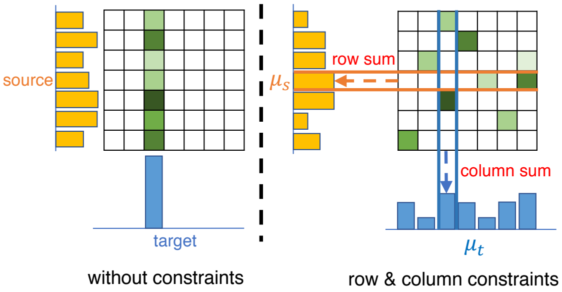


Figure 3: Left: Many source pixels are assigned to one target pixel. Right: The row sum and column sum constraints of the matching matrix suppress the many to one matching.

The intuition of modeling semantic correspondence as optimal transport is shown in Figure 3. Since the row sum and column sum of matching probability matrix is constrained to be μ_s and μ_t, the many to one matching problem occurring in other cosine similarity based methods [18, 30, 34], are significantly suppressed.

Computation of μ_s and μ_t with staircase re-weighting If we do not have any prior knowledge, then μ_s and μ_t can be set to uniform distributions, indicating same importance of each point in source and target feature maps. Since semantic correspondence suffers from background clutter issue, it is natural to recognize that the foreground object and background should be assigned different importance. Although some previous work [18, 33] showed similar idea, our method is flexible enough to incorporate any kind of prior into the unified optimal transport framework.

We generate the class activation maps [45] for source and target image as the prior information. Since we already have the feature extraction CNNs (detailed in next section), these maps are nearly zero cost due to the same forward pass with feature extraction. Let $\mathbf{f}_L \in \mathbb{R}^{h_L \times w_L \times d_L}$ denote the feature map of the last convolutional layer. It is fed into a Global Average Pooling layer (GAP) followed by a fully connected layer and a softmax layer for classification. The average value of the k_{th} feature map is $s_k = \frac{\sum_{i,j} \mathbf{f}_L(i,j,k)}{h_L \times w_L}$. $W^{fc} \in \mathbb{R}^{d_L \times C}$ denotes the fully connected layer weights, where C is the number of classes. Ignoring the bias term, the input to c_{th} softmax node can be defined as $y_c^{fc} = \sum_{k=0}^{d_L - 1} s_k W_{k,c}^{fc}$. The class activation map (CAM) of class c is obtained as follows,

$$A_c = \sum_{k=0}^{d_L - 1} \mathbf{f}_L(\cdot, \cdot, k) \cdot W_{k,c}^{fc} \, . \tag{7}$$

We choose the class with the highest classification probability and normalize A_c to the range of $[0, 1]$.

The original CAM is not well calibrated for source and

Algorithm 1 Optimal transport with sinkhorn algorithm.

```
**Input:** $\mu_s, \mu_t, \mathbf{M}, \epsilon, t_{max}$
Initialize $\mathbf{K} = e^{-\mathbf{M}/\epsilon}$ , $\mathbf{b} \leftarrow \mathbf{1}$ , $t \leftarrow 0$
**while** $t \leq t_{max}$ and not converge **do**
    $\mathbf{a} = \mu_s / (\mathbf{K}\mathbf{b})$
    $\mathbf{b} = \mu_t / (\mathbf{K}^\top \mathbf{a})$
**end while**
**Output:** $T = \text{diag}(\mathbf{a})\mathbf{K}\text{diag}(\mathbf{b})$
```

target images, e.g., same part of an object from two images may have different values. Therefore, we propose a staircase function to categorize the activation map into four levels according to their values: *hot spots*, *object*, *context* and *background*. Values of each category are adjusted as:

$$A_c(x, y) = \sum_{i=1}^{L} \gamma_i \mathbb{I}(A_c(x, y) > \beta_i) \,, \tag{8}$$

where $L = 4$ is the number of levels, β_i is the stair height denoting the threshold of i-th level, $\mathbb{I}(\cdot)$ is an indicator function whose value equals 1 only when the condition satisfies, γ_i is the stair width denoting the increased weight from previous level. γ_i and β_i are selected according to the validation set. Now, μ_s and μ_t can be computed as: $\mu_s(x, y) = A_c^s(x, y) / \sum A_c^s(x, y)$, $\mu_t(x, y) = A_c^t(x, y) / \sum A_c^t(x, y)$ and are then flattened to vectors. We call this strategy the *staircase re-weighting*.

Solving OT with Sinkhorn algorithm Exactly solving Eq. 6 with Network Flow solver requires the complexity of $O(n^3)$ (n proportional to $h_s w_s$ and $h_t w_t$). Following [6], we resort to the entropy-regularized optimal transport problem:

$$T^* = \underset{T \in \mathbb{R}_+^{h_s w_s \times h_t w_t}}{\arg\min} \sum_{ij} T_{ij} \mathbf{M}_{ij} + \epsilon H(T)$$
$$\text{s.t. } T\mathbf{1}_{h_t w_t} = \mu_s \,, T^\top \mathbf{1}_{h_s w_s} = \mu_t \,, \tag{9}$$

where $H(T) = \sum_{ij} T_{ij}(\log T_{ij} - 1)$ is the negative entropic regularization and $\epsilon > 0$ is the regularization parameter. Eq. 9 is a convex problem and can be solved using Sinkhorn-Knopp algorithm [37] with the complexity of $O(h_s w_s \times h_t w_t)$. Detailed solution is presented in Algorithm 1. Note that as Algorithm 1 only contains matrix multiplication and exponential operations, it is differentiable and can be computed efficiently. [1]

3.3. Pre- and post-processing

Input feature extraction Previous work [32, 33] usually extract feature from the last convolutional layer of deep neural network as matching primitives for semantic correspondence. However, this single layer feature cannot make full use of multi-level representations and fails to deal with ambiguous matching caused by intra-class variations. We follow the good practice of [30] to search and select multi-layer features from all candidate layers of a pre-trained CNN model.

A typical CNN takes an input image and produces a consecutive list of feature maps: $[\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_L]$ with $\mathbf{f}_i \in \mathbb{R}^{h_i \times w_i \times d_i}$. We use the percentage of correct keypoints (PCK) on validation set as a evaluation metric to compare different feature subsets. In order to search the optimal subset of feature maps, we run a variant of beam-search with a limited memory [30]. We maintain a memory containing at most N (beam size) subsets of layers. At each search step, we form the new subsets by adding the candidate layer to each of the current subsets in memory. With all old and new subsets, only the N top performing ones are kept in memory. This process goes on until we reach the maximum number of layers allowed.

After layer search, each image can be represented by a dense spatial feature grid $\mathbf{f} = [\mathbf{f}_{l_1}, \Phi(\mathbf{f}_{l_2}), \ldots, \Phi(\mathbf{f}_{l_k})] \in \mathbb{R}^{h \times w \times D}$, where l_i is the selected layer index and Φ denotes a function that upsamples feature map to the size of $\mathbf{f}_{l_1}$. We denote the features of source and target images as $\mathbf{f}_s$ and $\mathbf{f}_t$ respectively.

Post-processing In order to get the geometrically consistent matching, we employ the regularized Hough matching (RHM) [30] as the post-processing step. The key idea is to re-weight matching score by Hough space voting to enforce geometric consistency.

Let us assume $\mathbf{p}_s$ and $\mathbf{p}_t$ the position grids of feature maps $\mathbf{f}_s$ and $\mathbf{f}_t$. $R_s = (\mathbf{f}_s, \mathbf{p_s})$ and $R_t = (\mathbf{f}_t, \mathbf{p_t})$ are the coupled feature-position sets with r,r' being their elements. For the sake of simplicity, we denote $\mathcal{D}$ for two sets and m for a match: $\mathcal{D} = (R_s, R_t), m = (r, r')$ in $R_s \times R_t$. The matching confidence for m is denoted as $p(m|\mathcal{D})$. Since the source and target images contain the same object, we assume the common object can be located with *offset* x lying in a Hough space $\mathcal{X}$. The matching confidence is computed as:

$$p(m|\mathcal{D}) = p(m_a) \sum_{x \in \mathcal{X}} p(m_g|x) p(x|\mathcal{D}) \,, \tag{10}$$

$$p(x|\mathcal{D}) \propto \sum_{m} p(m_a) p(m_g|x) \,, \tag{11}$$

where $p(m_a)$ is the appearance matching probability, $p(m_g|x)$ is the geometric matching probability given an offset x, $p(x|\mathcal{D})$ is the geometry prior computed by aggregating individual votes into the Hough space scores.

In this work, we set $p(m_a) = T^*$. For $p(m_g|x)$, we estimate it by comparing $\mathbf{p}_s(i, j) - \mathbf{p}_t(i, j)$ to the given

[1] In this work, we only use pre-trained CNN model. We implement a GPU version of Algorithm 1 for fast computation.

offset x. The two-dimensional offset bins is constructed and a Gaussian mask is centered on offset x to re-weight the values.

After we obtain the matching confidence $p(m|D)$, it is easy to transfer any keypoint from a source image to the target image. Given a keypoint x_p in a source image, we first compute the neighborhood pixels $\mathcal{N}(x_p)$ whose feature map receptive fields cover this keypoint, and we compute the displacement between x_p and centers of the receptive fields, denoted by $\{d(x_q)\}_{x_q \in \mathcal{N}(x_p)}$. Let y_q denotes the target point for x_q computed from $p(m|D)$ by nearest neighbor assignment. The corresponding keypoint for x_p is the average of $\{y_q + d(x_q)\}_{x_q \in \mathcal{N}(x_p)}$.

4. Experiments

In this section we describe our benchmarks and evaluation metric, give implementation details, and compare our method to baselines and the state-of-the-art.

4.1. Benchmarks and evaluation metric

SPair-71k [31]. Due to the high expense of ground-truth annotations, previous datasets are relatively small and do not show much variability. The newly-released SPair-71k dataset contains 70,958 image pairs with diverse variations in viewpoint and scale, which is a reliable testbed for studying real problems of semantic correspondence.

TSS [40], **PF-PASCAL** [14], and **PF-WILLOW** [13]. TSS contains 400 image pairs divided into three groups: FG3DCar [26], JODS [35], and PASCAL [16]. PF-PASCAL contains 1,351 image pairs from the 20 object categories of the PASCAL VOC [11] dataset. PF-WILLOW contains 900 image pairs of 4 object categories. For a fair comparison, we follow the settings of previous work [15, 19, 25, 30, 33] to evaluate our model.

Evaluation metric. We employ the commonly-used metric of percentage of correct keypoints (PCK), which counts the number of correctly predicted keypoints given a fixed threshold. Given predicted keypoint $\mathbf{k}_{pr}$ and ground-truth keypoint $\mathbf{k}_{gt}$, the prediction is considered correct if the following condition satisfies:

$$d(\mathbf{k}_{pr}, \mathbf{k}_{gt}) \leq \alpha_\tau \cdot \max(w_\tau, h_\tau). \tag{12}$$

Here, $d(\cdot, \cdot)$ is the Euclidean distance, w_τ and h_τ are the width and height of either an entire image or object bounding box according to the criterion $\tau \in \{\text{img}, \text{bbox}\}$, α_τ is a fixed threshold (e.g., $\alpha_\tau = 0.1$). PCKs of all image pairs are averaged to get the final PCK. Following [30], we evaluate PF-PASCAL with α_{img}, PF-WILLOW and SPair-71k with the more stringent criterion α_{bbox}. For TSS, we follow [40, 19] to compute the PCK over a dense set of keypoints.

4.2. Implementation details

We use two CNNs as main backbone networks for feature and activation map extraction: ResNet50 and ResNet101 [17] pre-trained on ImageNet [9]. No fine-tuning is performed in any manner in our algorithm.

To select multiple-layer features, we run the search algorithm proposed in [30] with the proposed optimal transport matching on validation set. For SPair-71k, the best layer subsets are $(0, 11, 12, 13)$ with ResNet-50 and $(0, 19, 27, 28, 29, 30)$ with ResNet-101. For PF-PASCAL and PF-WILLOW, the best layer subsets are $(2, 22, 24, 25, 27, 28, 29)$ with ResNet-101. The optimal layers are different from [30] since we consider the total correlations rather than individual feature matching.

In Eq. 8, we set $\boldsymbol{\gamma} = [0.5, 0.3, 0.1, 0.1]$ and $\boldsymbol{\beta} = [0.0, 0.4, 0.5, 0.6]$ according to the PCK of the validation set. In Algorithm 1, we set $\epsilon = 0.05$ and $t_{max} = 50$.

4.3. Evaluation results on SPair-71k

Comparisons to state-of-the-art First, we compare per-class PCK on the SPair-71k dataset with state-of-the-art methods in Table 1. The overall PCK of the proposed algorithm outperforms the state-of-the-art [30] by 7.4 (relative 26%), which is a huge improvement. And for all classes, our algorithm surpasses [30] by a large margin. Among all candidate algorithms, our algorithm achieves the best PCK on 16 out of 18 classes. This proves the effectiveness and robustness of our optimal transport algorithm in finding global optimal matching.

To better understand the performance of our algorithm under complex conditions, we report the results according to different variation factors with various difficulty levels in Table 2. SPair-71k dataset contains diverse variations in view-point, scale, truncation and occlusion, which is a reliable testbed to study the problem of semantic correspondence. The results clearly show that the proposed algorithm outperforms all other methods by a large margin in all conditions, which demonstrates the high stability of the proposed algorithm.

Ablation studies on feature matching To verify the effect of global feature matching in optimal transport, we compare optimal transport with individual feature matching (i.e, cosine) on SPair-71k. We first introduce the baselines. Here, "**Cos-NN**" denotes the cosine matching scores (Eq. 4) followed by nearest neighbor assignment (NN). "**Cos-RHM**" denotes the cosine matching scores followed by regularized Hough matching (RHM), which is equivalent to the HPF algorithm in [30] [2]. Similarly, "**OT-NN**"

[2] For fair comparison, we use multiple-layer features selected in Sec 4.2 instead of [30].

Methods		aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	dog	horse	moto	person	plant	sheep	train	tv	all
Authors' original models	CNNGeo [32]	21.3	15.1	34.6	12.8	31.2	26.3	24.0	30.6	11.6	24.3	20.4	12.2	19.7	15.6	14.3	9.6	28.5	28.8	18.1
	A2Net [18]	20.8	17.1	37.4	13.9	33.6	29.4	26.5	34.9	12.0	26.5	22.5	13.3	21.3	20.0	16.9	11.5	28.9	31.6	20.1
	WeakAlign [33]	23.4	17.0	41.6	14.6	37.6	28.1	26.6	32.6	12.6	27.9	23.0	13.6	21.3	22.2	17.9	10.9	31.5	34.8	21.1
	NC-Net [34]	24.0	16.0	45.0	13.7	35.7	25.9	19.0	50.4	14.3	32.6	27.4	19.2	21.7	20.3	20.4	13.6	33.6	40.4	26.4
SPair-71k finetuned models	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 [18]	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
SPair-71k validation	HPF [30]	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
	Ours	**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**

Table 1: Per-class PCK (α_{bbox}=0.1) results on SPair-71k. All models in this table use ResNet101 as the backbone. For the authors' original models, the models of [32, 18] trained on PASCAL-VOC with self-supervision, [33, 34] trained on PF-PASCAL with weal-supervision are used for evaluation. For SPair-71k-finetuned models, the original models are further finetuned on SPair-71k dataset (results come from SPair-71k benchmark [31]). For SPair-71k validation models, [30] and our method are tuned using validation split of SPair-71k only. The best performances are shown in bold.

Methods		View-point			Scale			Truncation				Occlusion				All
		easy	medi	hard	easy	medi	hard	none	src	tgt	both	none	src	tgt	both	
Identity mapping		7.3	3.7	2.6	7.0	4.3	3.3	6.5	4.8	3.5	5.0	6.1	4.0	5.1	4.6	5.6
Authors' original models	CNNGeo [32]	25.2	10.7	5.9	22.3	16.1	8.5	21.1	12.7	15.6	13.9	20.0	14.9	14.3	12.4	18.1
	A2Net [18]	27.5	12.4	6.9	24.1	18.5	10.3	22.9	15.2	17.6	15.7	22.3	16.5	15.2	14.5	20.1
	WeakAlign [33]	29.4	12.2	6.9	25.4	19.4	10.3	24.1	16.0	18.5	15.7	23.4	16.7	16.7	14.8	21.1
	NC-Net [34]	34.0	18.6	12.8	31.7	23.8	14.2	29.1	22.9	23.4	21.0	29.0	21.1	21.8	19.6	26.4
SPair-71k finetuned models	CNNGeo [32]	28.8	12.0	6.4	24.8	18.7	10.6	23.7	15.5	17.9	15.3	22.9	16.1	16.4	14.4	20.6
	A2Net [18]	30.9	13.3	7.4	26.1	21.1	12.4	25.0	17.4	20.5	17.6	24.6	18.6	17.2	16.4	22.3
	WeakAlign [33]	29.3	11.9	7.0	25.1	19.1	11.0	24.0	15.8	18.4	15.6	23.3	16.1	16.4	15.7	20.9
	NC-Net [34]	26.1	13.5	10.1	24.7	17.5	9.9	22.2	17.1	17.5	16.8	22.0	16.3	16.3	15.2	20.1
SPair-71k validation	HPF [30]	35.6	20.3	15.5	33.0	26.1	15.8	31.0	24.6	24.0	23.7	30.8	23.5	22.8	21.8	28.2
	Ours	**42.7**	**28.0**	**23.9**	**41.1**	**33.7**	**21.4**	**39.0**	**32.4**	**30.0**	**30.0**	**39.0**	**30.3**	**28.1**	**26.0**	**35.6**

Table 2: PCK analysis by variation factors on SPair-71k ($\alpha_{\text{bbox}} = 0.1$). The variation factors include view-point, scale, truncation, and occlusion with various difficulty levels. All models in this table use ResNet101 as the backbone.

Backbone	Methods	src pts	trg matches	PCK
ResNet50	Cos-NN	4099	558	28.0
	OT-NN	4099	**1184**	**29.4**
	Cos-RHM	4099	783	26.6
	OT-RHM	4099	**1322**	**31.3**
ResNet101	Cos-NN	4099	445	30.6
	OT-NN	4099	**1062**	**33.7**
	Cos-RHM	4099	701	27.8
	OT-RHM	4099	**1261**	**34.8**

Table 3: PCK results ($\alpha_{\text{bbox}} = 0.1$) on SPair-71k dataset with feature matching computed by cosine (Cos) and optimal transport (OT). "src pts" denotes the average number of points from source feature maps. "trg matches" denotes the average number of unique points on target maps assigned to the source points.

and "**OT-RHM**" denote matching scores computed by optimal transport (Eq. 9) *without class activation maps* (i.e., μ_s and μ_t are set to uniform distribution) followed by corresponding post-processing. Finally, "**OT-RHM-CAM**" denotes the baseline using the original class activation maps, while "**OT-RHM-Stair**" denotes our model with staircase re-weighting on class activation maps, which is our ultimate model.

The results are shown in Table 3. It can be seen that in all settings (various backbones and geometric post-processing), the proposed optimal transport solution beats the corresponding baseline by a large margin. In order to study the *many to one matching* issue shown in Figure 1a, we calculated the average number of unique points on target maps assigned to source. As shown in Table 3, optimal transport has about twice individual matches as many as cosine methods, this agrees with our motivation that global feature matching can significantly suppress the *many to one matching*. We further observe that RHM can increase the number of target matches for both cosine and optimal transport. However, the PCK of Cos-RHM even drops compared with Cos-NN, while OT-RHM continues to increase over OT-NN. We conjecture that too many duplicated assignments of cosine matching hinders the effectiveness of RHM for better geometric adjustment. We also notice that ResNet101 has a smaller number of target matches than ResNet50. The reason is that the deeper network has a larger receptive field in deeper layers that makes features of these layers less distinguishable.

Methods	ResNet50	ResNet101
OT-NN	29.4	33.2
OT-NN-CAM	29.6	32.6
OT-NN-Stair	**30.2**	**34.2**
OT-RHM	31.3	34.8
OT-RHM-CAM	31.4	34.5
OT-RHM-Stair	**32.1**	**35.6**

Table 4: Ablation study of staircase re-weighting on SPair-71k dataset. PCK results with $\alpha_{\text{bbox}} = 0.1$ are reported.

Methods	FG3D.	JODS	PASC.	Avg.
CNNGeo$_{\text{res101}}$ [32]	90.1	76.4	56.3	74.3
DCTM$_{\text{CAT-FCSS}}$ [22]	89.1	72.1	61.0	74.0
Weakalign$_{\text{res101}}$ [33]	90.3	76.4	56.5	74.4
RTNs$_{\text{res101}}$ [21]	90.1	78.2	**63.3**	77.2
NC-Net$_{\text{res101}}$ [34]	94.5	81.4	57.1	77.7
DCCNet$_{\text{res101}}$ [19]	93.5	**82.6**	57.6	77.9
HPF$_{\text{res101}}$ [30]	93.6	79.7	57.3	76.9
Ours$_{\text{res101}}$	**95.3**	81.3	57.7	**78.1**

Table 5: Evaluation results on TSS dataset. Subscripts of the method names indicate backbone networks used. We report the PCK scores with $\alpha = 0.05$ and the best results are in bold.

Ablation studies on staircase re-weighting We then investigate the effect of staircase re-weighting under different backbones and geometric post-processing. From Table 4, we can see that original CAM has little or no improvement compared to the baselines while our staircase re-weighting enjoys at least 0.8% PCK increase. Considering the staircase re-weighting strategy shares the same CNN forward pass with feature extraction and the extra cost is nearly free, this is a promising improvement. We believe this strategy can perform better with more accurate class activation maps. We leave this for future study.

4.4. TSS, PF-PASCAL, and PF-WILLOW

Table 5 shows the evaluation results on TSS dataset. The proposed method outperforms previous methods on one of the three groups of the TSS dataset and the average performance over three groups on the TSS dataset sets up a new state of the art.

Table 6 summarizes comparisons to state-of-the-art methods on PF-PASCAL and PF-WILLOW. Following [30], we use the backbone of FCN [27] pre-trained with PASCAL VOC 2012 [11] [3]. Different levels of supervisory signals are used in the deep network models, such as self-supervision [32, 18], weak-supervision [22, 33, 34, 19, 21], keypoints [15, 24] and masks [25]. In the contrary, HPF [30] and our method only use the pre-trained ImageNet models and the validation set. Table 6 shows that the proposed method achieves the state-of-the-art results on both benchmarks with various thresholds α. It need to be further noticed that when α becomes smaller (stricter criterion), our method gains larger advantage over others. This indicates that our method generates more accurate keypoint predictions, so it can perform better with small threshold.

Methods	PF-PASCAL (α_{img})			PF-WILLOW (α_{bbox})		
	0.05	0.1	0.15	0.05	0.1	0.15
PF$_{\text{HOG}}$ [13]	31.4	62.5	79.5	28.4	56.8	68.2
CNNGeo$_{\text{res101}}$ [32]	41.0	69.5	80.4	36.9	69.2	77.8
A2Net$_{\text{res101}}$ [18]	42.8	70.8	83.3	36.3	68.8	84.4
DCTM$_{\text{CAT-FCSS}}$ [22]	34.2	69.6	80.2	38.1	61.0	72.1
Weakalign$_{\text{res101}}$ [33]	49.0	74.8	84.0	37.0	70.2	79.9
NC-Net$_{\text{res101}}$ [34]	54.3	78.9	86.0	33.8	67.0	83.7
DCCNet$_{\text{res101}}$ [19]	-	82.3	-	43.6	73.8	86.5
RTNs$_{\text{res101}}$ [21]	55.2	75.9	85.2	41.3	71.9	86.2
SCNet$_{\text{VGG16}}$ [15]	36.2	72.2	82.0	38.6	70.4	85.3
NN-Cyc$_{\text{res101}}$ [24]	55.1	85.7	94.7	40.5	72.5	86.9
SFNet$_{\text{res101}}$ [25]	-	78.7	-	-	74.0	-
HPF$_{\text{res101}}$ [30]	60.1	84.8	92.7	45.9	74.4	85.6
HPF$_{\text{res101-FCN}}$ [30]	63.5	88.3	**95.4**	48.6	76.3	88.2
Ours$_{\text{res101}}$	63.1	85.4	92.7	47.8	76.0	87.1
Ours$_{\text{res101-FCN}}$	**67.3**	**88.8**	**95.4**	**50.7**	**78.1**	**89.1**

Table 6: Evaluation results on PF-PASCAL and PF-WILLOW. Subscripts of the method names indicate backbone networks used. Different levels of supervision are used, such as self-supervision [32, 18], weak-supervision [22, 33, 34, 19, 21], keypoints [15, 24] and masks [25]. HPF [30] and our method only use pre-trained models and the validation set. The best performances are shown in bold. We borrow the results of [15, 13, 22, 32, 33] from [21].

5. Conclusion

We propose to model semantic correspondence as an optimal transport problem (SCOT). We solve semantic correspondence by maximizing the total correlations between pixels in two images, which is equivalent to the standard optimal transport problem. We then apply a staircase function on the class activation maps generated from feature extraction CNNs with neglected extra cost to re-weight the importance of foreground and background pixels. These reweighted maps are normalized to serve as prior information for empirical distributions in optimal transport. The ablation studies clearly demonstrate the effectiveness of each component. And SCOT outperforms state-of-the-art on standard benchmarks by a large margin.

Acknowledgments. Yanbin Liu, Linchao Zhu and Yi Yang were supported by ARC DP200100938. Makoto Yamada was supported by the JST PRESTO program JPMJPR165A.

[3] For this network, we directly extract the max-aggregated class masks as class activation map. Image-level annotations are not used.

References

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *ICML*, 2017. 2

[2] Charlotte Bunne, David Alvarez-Melis, Andreas Krause, and Stefanie Jegelka. Learning generative models across incomparable spaces. In *ICML*, 2019. 2

[3] 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 *CVPR*, 2015. 2

[4] Christopher B Choy, JunYoung Gwak, Silvio Savarese, and Manmohan Chandraker. Universal correspondence network. In *NIPS*, 2016. 1, 2

[5] Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. *IEEE TPAMI*, 39(9):1853–1865, 2016. 2

[6] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *NIPS*, 2013. 5

[7] Kevin Dale, Micah K Johnson, Kalyan Sunkavalli, Wojciech Matusik, and Hanspeter Pfister. Image restoration using online photo collections. In *ICCV*, 2009. 1

[8] N Danal. Histgram of oriented gradients for human detection. In *CVPR*, pages 886–893, 2005. 2

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 6

[10] Ishan Deshpande, Yuan-Ting Hu, Ruoyu Sun, Ayis Pyrros, Nasir Siddiqui, Sanmi Koyejo, Zhizhen Zhao, David Forsyth, and Alexander G Schwing. Max-sliced wasserstein distance and its use for gans. In *CVPR*, 2019. 2

[11] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *IJCV*, 111(1):98–136, 2015. 6, 8

[12] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 2

[13] Bumsub Ham, Minsu Cho, Cordelia Schmid, and Jean Ponce. Proposal flow. In *CVPR*, 2016. 2, 6, 8

[14] Bumsub Ham, Minsu Cho, Cordelia Schmid, and Jean Ponce. Proposal flow: Semantic correspondences from object proposals. *IEEE TPAMI*, 40(7):1711–1725, 2017. 6

[15] Kai Han, Rafael S Rezende, Bumsub Ham, Kwan-Yee K Wong, Minsu Cho, Cordelia Schmid, and Jean Ponce. Scnet: Learning semantic correspondence. In *ICCV*, 2017. 2, 6, 8

[16] Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In *ICCV*, 2011. 6

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 6

[18] Paul Hongsuck Seo, Jongmin Lee, Deunsol Jung, Bohyung Han, and Minsu Cho. Attentive semantic alignment with offset-aware correlation kernels. In *ECCV*, 2018. 1, 2, 4, 7, 8

[19] Shuaiyi Huang, Qiuyue Wang, Songyang Zhang, Shipeng Yan, and Xuming He. Dynamic context correspondence network for semantic alignment. In *ICCV*, 2019. 1, 2, 6, 8

[20] Leonid Kantorovitch. On the translocation of masses. *Management Science*, 5(1):1–4, 1958. 3

[21] Seungryong Kim, Stephen Lin, SANG RYUL JEON, Dongbo Min, and Kwanghoon Sohn. Recurrent transformer networks for semantic correspondence. In *NeurIPS*, 2018. 1, 8

[22] Seungryong Kim, Dongbo Min, Stephen Lin, and Kwanghoon Sohn. Dctm: Discrete-continuous transformation matching for semantic flow. In *ICCV*, 2017. 8

[23] Yehezkel Lamdan, Jacob T Schwartz, and Haim J Wolfson. Object recognition by affine invariant matching. In *CVPR*, 1988. 2

[24] Zakaria Laskar, Hamed Rezazadegan Tavakoli, and Juho Kannala. Semantic matching by weakly supervised 2d point set registration. In *WACV*, 2019. 1, 2, 8

[25] Junghyup Lee, Dohyung Kim, Jean Ponce, and Bumsub Ham. SFNet: Learning Object-aware Semantic Correspondence. In *CVPR*, 2019. 2, 6, 8

[26] Yen-Liang Lin, Vlad I Morariu, Winston Hsu, and Larry S Davis. Jointly optimizing 3d model fitting and fine-grained classification. In *ECCV*, 2014. 6

[27] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015. 8

[28] Jonathan L Long, Ning Zhang, and Trevor Darrell. Do convnets learn correspondence? In *NIPS*, 2014. 2

[29] David G Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004. 2

[30] Juhong Min, Jongmin Lee, Jean Ponce, and Minsu Cho. Hyperpixel flow: Semantic correspondence with multi-layer neural features. In *ICCV*, 2019. 1, 2, 4, 5, 6, 7, 8

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

[32] Ignacio Rocco, Relja Arandjelovic, and Josef Sivic. Convolutional neural network architecture for geometric matching. In *CVPR*, 2017. 1, 2, 5, 7, 8

[33] Ignacio Rocco, Relja Arandjelović, and Josef Sivic. End-to-end weakly-supervised semantic alignment. In *CVPR*, 2018. 1, 2, 4, 5, 6, 7, 8

[34] Ignacio Rocco, Mircea Cimpoi, Relja Arandjelović, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Neighbourhood consensus networks. In *NeurIPS*, 2018. 1, 2, 4, 7, 8

[35] Michael Rubinstein, Armand Joulin, Johannes Kopf, and Ce Liu. Unsupervised joint object discovery and segmentation in internet images. In *CVPR*, 2013. 1, 6

[36] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, 2017. 2

[37] Richard Sinkhorn. Diagonal equivalence to matrices with prescribed row and column sums. *The American Mathematical Monthly*, 74(4):402–405, 1967. 5

[38] Zhengyu Su, Yalin Wang, Rui Shi, Wei Zeng, Jian Sun, Feng Luo, and Xianfeng Gu. Optimal mass transport for shape matching and comparison. *IEEE TPAMI*, 37(11):2246–2259, 2015. 2

[39] Richard Szeliski et al. Image alignment and stitching: A tutorial. *Foundations and Trends® in Computer Graphics and Vision*, 2(1):1–104, 2007. 1

[40] Tatsunori Taniai, Sudipta N Sinha, and Yoichi Sato. Joint recovery of dense correspondence and cosegmentation in two images. In *CVPR*, 2016. 1, 6

[41] Jiqing Wu, Zhiwu Huang, Dinesh Acharya, Wen Li, Janine Thoma, Danda Pani Paudel, and Luc Van Gool. Sliced wasserstein generative models. In *CVPR*, 2019. 2

[42] Hongteng Xu, Dixin Luo, and Lawrence Carin. Scalable gromov-wasserstein learning for graph partitioning and matching. *arXiv preprint arXiv:1905.07645*, 2019. 2

[43] Hongteng Xu, Dixin Luo, Hongyuan Zha, and Lawrence Carin Duke. Gromov-wasserstein learning for graph matching and node embedding. In *ICML*, 2019. 2

[44] Xiaolin Zhang, Yunchao Wei, Jiashi Feng, Yi Yang, and Thomas S Huang. Adversarial complementary learning for weakly supervised object localization. In *CVPR*, 2018. 2

[45] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, 2016. 2, 4