Gromov Wasserstein Optimal Transport for Semantic Correspondences

Francis Snelgar[1,2]
francis.snelgar@anu.edu.au
Ming Xu[1]
mingda.xu@anu.edu.au
Stephen Gould[1]
stephen.gould@anu.edu.au
Liang Zheng[1]
liang.zheng@anu.edu.au
Akshay Asthana[2]
akshay.asthana@seeingmachines.com

[1] School of Computing
Australian National University
Canberra, Australia

[2] Seeing Machines
Canberra, Australia

Abstract

Establishing correspondences between image pairs is a long studied problem in computer vision. With recent large-scale foundation models showing strong zero-shot performance on downstream tasks including classification and segmentation, there has been interest in using the internal feature maps of these models for the semantic correspondence task. Recent works observe that features from DINOv2 and Stable Diffusion (SD) are complementary, the former producing accurate but sparse correspondences, while the latter produces spatially consistent correspondences. As a result, current state-of-the-art methods for semantic correspondence involve combining features from both models in an ensemble. While the performance of these methods is impressive, they are computationally expensive, requiring evaluating feature maps from large-scale foundation models. In this work we take a different approach, instead replacing SD features with a superior matching algorithm which is imbued with the desirable spatial consistency property. Specifically, we replace the standard nearest neighbours matching with an optimal transport algorithm that includes a Gromov Wasserstein spatial smoothness prior. We show that we can significantly boost the performance of the DINOv2 baseline, and be competitive and sometimes surpassing state-of-the-art methods using Stable Diffusion features, while being 5–10x more efficient. We make code available at https://github.com/fsnelgar/semantic_matching_gwot.

1 Introduction

In contrast to the traditional counterpart, the *semantic* matching task requires finding correspondences between different instances of the same class (*e.g.*, the ears of two different cats) across images [6, 7, 19, 36, 37, 42]. It is an inherently challenging problem in computer vision due to large visual differences between classes, non-rigid objects and change in

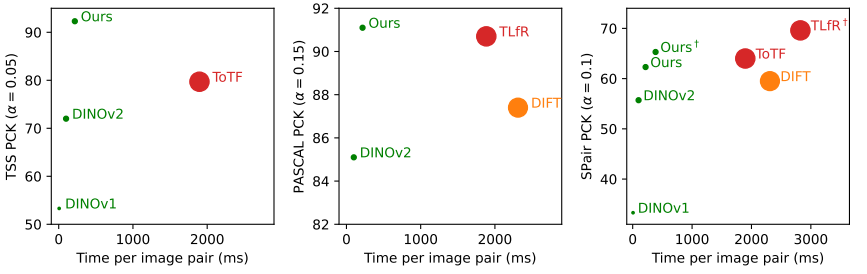

Figure 1: Accuracy and latency trade off for various methods. Our method achieves comparable accuracy while being much faster and requiring less memory. The size of the markers indicates the relative number of parameters in each method, and the colour indicates the model family. Green methods use DINO models, orange methods use Stable Diffusion, and red methods combine features from both. Methods marked with † use ground truth labels to flip keypoints at test time. See Tab. 3 for a detailed breakdown.

appearance due to camera pose. The task has numerous interesting applications including visual localisation [20, 42] semantic and few-shot segmentation [17, 35] and image transfer [50, 51]. There are several properties that are desirable in a semantic correspondence method. First, correspondences should be unique — each pixel in the target image should match at most one pixel in the source image. Second, the correspondence map should be locally spatially smooth, a pair of pixels that are sufficiently close in the source image should match to a pair that are similarly close in the target image, but global non rigid deformations mean this property is only desirable locally. Third, a matching algorithm must accommodate objects of different scale as well as background regions which have no semantic correspondences. Recent methods [3, 12, 26, 43, 54, 55] using pretrained foundation models [3, 38] have identified that the spatial smoothness property can be provided by Stable Diffusion features due to its strong spatial awareness in contrast to DINO features. We instead take a different approach and design a novel matcher based on optimal transport, encoding all properties — including spatial smoothness — into the matching algorithm, yielding competitive results with a fraction of the compute and memory. We summarise the speed and performance of zero-shot methods in Fig. 1, with our method achieving competitive performance at a fraction of the cost.

2 Related Works

2.1 Semantic Correspondences

Early semantic correspondence methods followed keypoint matching pipelines with handcrafted descriptors [8, 25, 40] combined with matching algorithms using geometric models [6, 22]. Cho *et al*. [6] uses hierarchical object proposals and the Probablistic Hough Transform (PHM) as a geometric model. ProposalFlow [15] extends this work, addressing the limitations of the global matching consensus. SiftFlow [22] combines coarse-to-dense SIFT descriptors with a belief propagation layer to encourage spatial smoothness.

Later methods [19, 23, 24, 36, 37] use convolutional neural networks in replacement of traditional descriptors. Long *et al*. [24] showed features from convolutional networks

worked as well as SIFT [25] descriptors for alignment and classification tasks. Rocco *et al*. [36] uses a parametric geometric layer to train CNNs end-to-end, later replacing it with a RANSAC [10] inspired layer to allow weakly supervised training. SCOT [23] uses a network pretrained on ImageNet [10] with optimal transport [45] and PHM [8] post processing.

In recent years a body of work [12, 26, 27, 43, 54, 55] has focused on using large foundation models including DINO [3, 9, 31] and Stable Diffusion [38] due to their powerful zero shot performance on downstream tasks. DIFT [43] use features from Stable Diffusion prompted with the class label, while Zhang *et al*. [54, 55] uses the ODISE [52] pipeline, including an implicit prompt from CLIP [34]. Several works combine features from several foundation models, either using PCA [55] or distillation [12] to reduce the computational burden. Several methods attempt to integrate the global object pose to resolve visual ambiguity, Zhang *et al*. [12, 54] uses labelled symmetric keypoints while Mariotti *et al*. [27] trains a spherical geometric prior to map DINO features to the unit sphere in object coordinates.

Different to these works, our contribution aims to improve the matching algorithm using features from a single model. We show that we can imbue desirable properties of foundation model features into the matching algorithm itself, while using less compute and memory. Our method is related to SCOT [23], which also uses optimal transport for matching. However, SCOT does not include properties such as spatial smoothness and symmetry within the OT problem. We show including these properties significantly improves matching accuracy.

2.2 Optimal Transport in Computer Vision

As our method is based on optimal transport (OT), we provide a brief review of related applications where OT is used for matching or alignment. We also discuss related works where Gromov Wasserstein OT has been used to imbue structure into the matching algorithm.

Optimal transport is a long studied problem [18, 29] and there are many variants including partial [9, 41] and unbalanced [1] transport and the Gromov Wasserstein [30, 32] formulation. For an excellent introduction into the problem we refer the reader to Thorpe's notes [45]. Applications of optimal transport include keypoint matching [41], positive-unlabelled learning [9] and object detection [33]. Others have exploited structure in the problem with Gromov Wasserstein optimal transport for temporal action segmentation [53], graph classification [47] alignment of fMRI data [46] and domain adaptation [13]. In this work we apply Gromov Wasserstein optimal transport to the new application of semantic correspondences.

3 Background

In this section we provide background on optimal transport which forms the basis of our method. First, we introduce the OT problem and how it can be used for matching problems, *e.g*., [41]. Next we introduce Gromov Wasserstein OT, which allows incorporating structure such as spatial structure, leveraging recent works [32, 53].

3.1 Kantorovich Optimal Transport

Optimal transport (OT) [29] is the problem of comparing two distributions and moving one to the other in the most efficient way possible. For the classic Kantorovich [18] OT with discrete measures $p \in \Delta_n$, $q \in \Delta_m$, where Δ_k is the $k-1$ dimensional probability simplex,

and a (symmetric) ground cost $C \in \mathbb{R}_+^{n\times m}$ (*i.e.*, the 'cost' of transporting mass from element p_i to q_i) the optimal transport map $T^\star$ is the solution to the linear program:

$$T^\star = \underset{T\in\mathcal{T}}{\operatorname{argmin}} \langle C, T\rangle, \tag{1}$$

where $\mathcal{T}$ is the transport polytope $\mathcal{T} = \{T \in \mathbb{R}_+^{n\times m} : T\mathbf{1}_m = p, T^\top \mathbf{1}_n = q\}$. For the applications in this work we consider T as a soft assignment between elements of p and q.

3.2 Gromov Wasserstein Optimal Transport

The optimal transport formulation introduced so far require that there exists a ground cost C between p and q. Gromov Wasserstein (GW) [30, 32] instead defines the transport cost between pairwise elements within the same measure. Given (symmetric) $C^p \in \mathbb{R}^{n\times n}$, $C^q \in \mathbb{R}^{m\times m}$ which defines a distance metric between pairs of elements in p (resp. q), the GW optimisation objective is given by

$$\mathcal{F}_{\mathrm{GW}}(C^p, C^q, T) = \sum_{\substack{(i,k)\in[n]\times[n]\\(j,l)\in[m]\times[m]}} L(C^p_{i,k}, C^q_{j,l}) T_{i,j} T_{k,l}, \tag{2}$$

where L defines a distance metric between elements of cost matrices C^p and C^q. As noted by Peyré [32] for $L\left(C^p_{i,k}, C^q_{j,l}\right) = C^p_{i,k} C^q_{j,l}$, $\mathcal{F}_{\mathrm{GW}}$ can be computed efficiently using only matrix operations as $\langle C^p T C^{q\top}, T\rangle$.

Kantorovich and GW optimal transport can be combined when there is a structural prior as well as a defined ground cost to form the Fused-GW problem [46, 47],

$$T^\star = \underset{T\in\mathcal{T}}{\operatorname{argmin}} \langle C, T\rangle + \mathcal{F}_{\mathrm{GW}}(C^p, C^q, T). \tag{3}$$

4 Optimal Transport for Semantic Correspondences

In this section we describe our method for semantic correspondences. We briefly introduce the task before detailing how we design an optimal transport based matching algorithm that encodes the properties of the semantic correspondence task.

4.1 Problem Statement

Given RGB image pairs $I, \hat{I} \in \mathbb{R}^{H\times W\times 3}$, we wish to find correspondences between pixels in $I, \hat{I}$ which represent semantically 'similar' parts. For each image, we assume we have access to the (normalised) feature maps from a foundation model $y = f_\theta(I) \in \mathbb{R}^{N\times D}$, where $N = \frac{H\times W}{P^2}$ is the spatial size of the feature maps with corresponding patch coordinates $\mathcal{X} \in [1, \frac{W}{P}] \times [1, \frac{H}{P}]$ and size P. To establish correspondences, we solve the optimisation problem

$$T^\star = \underset{T}{\operatorname{argmin}}\ \lambda \langle C, T\rangle + \lambda_{\mathrm{gw}} \mathcal{F}_{\mathrm{GW}}(C^p, C^q, T) + \lambda_{\mathrm{sym}} \mathcal{F}_{\mathrm{sym}}(T) + \lambda_{\mathrm{ub}} \mathrm{D}_{\mathrm{KL}}(T^\top \mathbf{1}_N, q), \tag{4}$$

where $p = \frac{1}{N}\mathbf{1}_N$, $q = \frac{1}{N}\mathbf{1}_N$ are measures over the respective sets of coordinates $\mathcal{X}$, $\hat{\mathcal{X}}$ and $\{T \in \mathbb{R}^{N\times N} : T\mathbf{1}_N = p\}$. The correspondence $\hat{x}_j$ for patch x_i is given by $j = \operatorname{argmax}(T_i^\star)$. We explain the exact form and reasoning for the terms in Eq. (4) in the following sections.

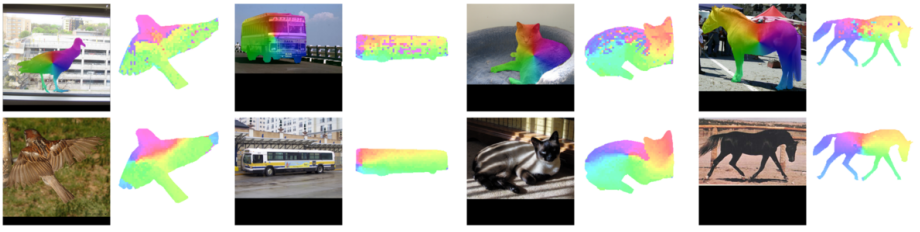

Figure 2: Impact of Gromov Wasserstein optimal transport. Each two-by-two grid contains an example image pair on the left, and dense correspondences using nearest neighbours and our GW method in the top right and bottom right respectively. Note that correspondences for our method are more spatially consistent.

4.2 Objective Functions

Feature Similarity: The cosine similarity between normalised features $y, \hat{y}$ captures the semantic similarity between patches, we therefore use $C = 1 - y\hat{y}^\top$ as the ground cost. However feature similarity is spatially noisy, and by itself does not encourage spatial smoothness.
Spatial Smoothness: To encourage spatial smoothness we include a Gromov Wasserstein objective $\mathcal{F}_{\mathrm{GW}}(C^p, C^q, T) = \langle C^p T C^{q\top}, T \rangle$, with cost matrices

$$C^p_{i,k} = \begin{cases} 1, & ||x_i - x_k||_2 < \delta_{\min} \\ 0, & \text{otherwise} \end{cases} \qquad C^q_{j,l} = \begin{cases} 1, & ||\hat{x}_j - \hat{x}_l||_2 > \delta_{\max} \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

Consider two pairs of patches with coordinates x_i, x_k and $\hat{x}_j, \hat{x}_l$ in images I and $\hat{I}$, respectively. The GW objective $\mathcal{F}_{\mathrm{GW}}$ is non-zero for pairs that are within radius $\delta_{\min}$ in I and further than radius $\delta_{\max}$ in $\hat{I}$. This encourages neighbouring patches to match to patches that are similarly close, while allowing for small amounts of local non rigid deformation. This results in matches that are more spatially consistent, as shown in Fig. 2. Note that since C^p, C^q are sparse, we can implement it using 2D convolutions, avoiding expensive matrix multiplication operations to further reduce compute and memory requirements.
Object Symmetry: Similar to Zhang *et al.* [54] we note that there often exists local ambiguity in matching between objects, where the global geometry of the object is required to correctly resolve local regions. For example, a 'wheel' may be visually similar to multiple other wheels, and the correct match can only be determined by considering the global pose of the vehicle. To address this we introduce a symmetry aware objective. We observe that many objects have a vertical axis of symmetry, and that under moderate changes in pose, the ordering of keypoints around this axis should be consistent across images. Let $\mathcal{G} = \{u_0, u_1, \ldots, u_M\}$ be the set of M pairs of symmetric keypoints for a given object where $u_m = (x_i, x_j)$ are the patch coordinates for the pair in I. The symmetry object function is

$$\mathcal{F}_{\mathrm{sym}}(S, T) = \sum_{\substack{(i,k)\in\mathcal{G} \\ (j,l)\in[N]}} -S_{i,j} S_{k,l} T_{i,j} T_{k,l}, \tag{6}$$

where $S \in \mathbb{R}^{N\times N}$ and $S_{i,j} = \mathrm{sign}(x_i^{(0)} - x_j^{(0)})$ indicating the relative ordering of patch coordinates horizontally (along the x-axis). While inspired by Gromov Wasserstein, $\mathcal{F}_{\mathrm{sym}}$ is not a strictly GW as S is not a valid metric space on $\mathcal{X}$ due to $S_{i,j} \neq S_{j,i}$. Nevertheless, it is still a

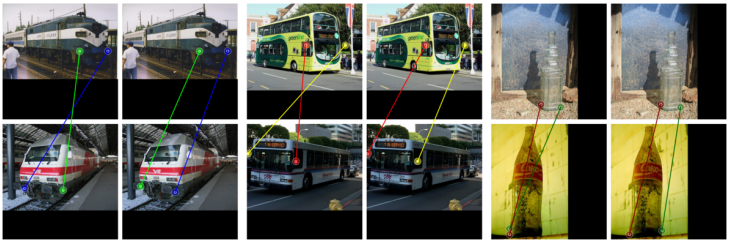

Figure 3: Effect of the symmetry loss. Each two-by-two grid contains matches without the symmetry loss on the left, and with the symmetry loss on the right. Note that with the symmetry loss ordering of the keypoint pair is maintained and matches are more plausible.

Figure 4: Impact of unbalanced optimal transport. Each two-by-two grid shows the image pair on the left, with results for balanced and partially balanced optimal transport on the top right and bottom right respectively. For objects with large scale differences (first and fourth examples) or occlusion (second and third examples), pixels have unequal importance and the balanced mass assumption doesn't hold. The unbalanced OT results are more plausible.

valid objective function and can be minimized in the same problem. We show examples of this in Fig. 3. Without the symmetry loss keypoints are matched incorrectly to opposite sides of the object. With the loss, ordering is maintained and keypoints are matched correctly.

4.3 Unbalanced Formulation

A balanced optimal transport problem would create one-to-one correspondences between all patches. However in practice this is not always desirable as object scale may vary significantly, and there may be regions which are not covisible. Therefore we relax the balanced assignment constraint to form a partially unbalanced problem. The first marginal constraint $T\mathbf{1}_N = p$ remains, insuring that a match is found for all patches in I, however the second constraint $T^\top \mathbf{1}_N = q$ is replaced with a KL divergence regularisation penalty $\mathrm{D}_{\mathrm{KL}}(T^\top \mathbf{1}, q)$. We show examples of the impact of the unbalanced formulation in Fig. 4.

5 Experiments

5.1 Implementation Details

We follow previous zero-shot methods and use a pre-trained and frozen foundation model as a feature extractor. We use the DINOv2 (ViT-B/16) model with an image resolution of

	TSS PCK $\alpha_{\text{image}} = 0.05$				PASCAL PCK $\alpha_{\text{image}} = k$		
Method	FG3DCar	JODS	Pascal	Avg.	0.05	0.1	0.15
CATs [8]	92.1	78.9	64.2	78.4	76.8	92.7	96.5
CATs++ [9]	-	-	-	-	84.9	93.8	96.8
PWarpC-CATs [49]	95.5	85.0	85.5	88.7	79.8	92.6	96.4
CNNGeo [54]	90.1	76.4	56.3	74.4	41.0	69.5	80.4
Semantic-GLU-Net [58]	95.3	82.2	78.2	85.2	48.3	72.5	85.1
SCOT [24]	95.3	81.3	57.7	78.1	63.1	85.4	**92.7**
DINOv1 [63]	64.7	51.2	36.7	53.3	-	-	-
DINOv2* [63]	82.8	73.9	53.9	72.0	63.0	79.2	85.1
SD [63]	93.9	69.4	57.7	77.7	-	-	-
DIFT* [63]	-	-	-	-	66.0	81.1	87.2
ToTF Fuse* [63]	94.3	73.2	60.9	79.7	71.5	85.8	90.6
TLfR [62]	-	-	-	-	**74.0**	**86.2**	90.7
DINOv2	83.9	75.1	55.1	71.4	59.4	76.8	82.4
Ours	**98.2**	**89.8**	**88.8**	**92.3**	70.6	**86.2**	91.1

Table 1: Performance on the TSS and PF-PASCAL datasets. We report `per-image` PCK results. Results in the top half of the table are for supervised methods, results in the bottom half of the table are for zero-shot methods. Results for our method are coloured in grey. Methods marked with * are taken from [54].

840-by-840 and extract all tokens except the class token from the last layer as features for matching. For the symmetric aware objective, we use the coordinates of the keypoints in the source image provided in the datasets for the symmetric pairs, defined by Zhang *et al.* [54]. We solve the OT problem in (4) using projected gradient descent with 50 steps. We provide a full breakdown of hyperparameters in the supplementary material.

5.2 Datasets

TSS: The TSS dataset [44] contains 400 image pairs primarily of vehicles curated from existing FG3D [21], PASCAL [16] and JODS [39] datasets. As vehicles are non deformable, methods benefit from having a strong spatial smoothness prior. Dense correspondence labels are provided for the objects of interest.

PF-PASCAL: PF-PASCAL [15] contains has 1,351 image pairs across 20 object categories. Keypoints are provided from the 2011 PASCAL [2] annotations. The dataset is more challenging than TSS, with greater variation in pose and scale.

SPair-71k: SPair [28] is a highly challenging semantic correspondence dataset with large variation in object pose and a higher degree of scene clutter. It is significantly larger than previous datasets, containing 70,958 image pairs of 18 categories with keypoint annotations.

5.3 Metrics

We use the standard Percentage of Correct Keypoints (PCK) metric for evaluation. For each keypoint, the predicted correspondence is considered correct if it is within $\alpha \cdot \max(w,h)$ radius of the ground truth match, where $0 \leq \alpha \leq 1$ and w,h are the width and height of the image for TSS and PF-PASCAL, or the bounding box for SPair-71k. We report results using the total number of correct keypoints in a category (or dataset) normalized by the total number of keypoints for the SPair-71k dataset, and normalised per image for TSS and PF-PASCAL datasets following recent prior works [12, 43, 54, 55].

Method	Aero	Bike	Bird	Boat	Bottle	Bus	Car	Cat	Chair	Cow	Dog	Horse	Motor	Person	Plant	Sheep	Train	TV	All
PER-IMAGE PCK $\alpha_{\text{bbox}} = 0.1$																			
PwarpC [49]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	35.3
CATs [8]	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
CATs++ [9]	60.6	46.9	82.5	41.6	56.8	64.9	50.4	72.8	29.2	75.8	65.4	62.5	50.9	56.1	54.8	48.2	80.9	74.9	59.9
SCOT [23]	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
PER-KEYPOINT PCK $\alpha_{\text{bbox}} = 0.1$																			
ASIC [24]	57.9	25.2	68.1	24.7	35.4	28.4	30.9	54.8	21.6	45.0	47.2	39.9	26.2	48.8	14.5	24.5	49.0	24.6	36.9
Dis.Diff† [2]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	70.6
S.Maps [27]	74.8	64.5	87.1	45.6	52.7	77.8	71.4	82.4	47.7	82.0	67.3	73.9	67.6	60.0	49.9	69.8	78.5	59.1	67.3
ToTF [53]	81.2	66.9	91.6	61.4	57.4	85.3	83.1	90.8	54.5	88.5	75.1	80.2	71.9	77.9	60.7	68.9	92.4	65.8	74.6
DIFT [43]	63.5	54.5	80.8	34.5	46.2	52.7	48.3	77.7	39.0	76.0	54.9	61.3	53.3	46.0	57.8	57.1	71.1	**63.4**	59.5
DINOv2 [53]	72.7	62.0	85.2	41.3	40.4	52.3	51.5	71.1	36.2	67.1	64.6	67.6	61.0	68.2	30.7	62.0	54.3	24.2	55.6
SD [53]	63.1	55.6	80.2	33.8	44.9	49.3	47.8	74.4	38.4	70.8	53.7	61.1	54.4	55.0	54.8	53.5	65.0	53.3	57.2
ToTF [53]	73.0	64.1	86.4	40.7	52.9	55.0	53.8	78.6	45.5	77.3	64.7	69.7	63.3	69.2	58.4	67.6	66.2	53.5	64.0
TLfR† [54]	**78.0**	**66.4**	**90.2**	44.5	**60.1**	**66.6**	**60.8**	**82.7**	**53.2**	**82.3**	69.5	**75.1**	**66.1**	**71.7**	**58.9**	**71.6**	**83.8**	55.5	**69.6**
DINOv2	72.5	63.3	85.7	39.9	41.4	51.6	51.7	71.0	35.8	67.9	65.5	67.9	59.2	67.2	30.1	60.8	54.7	25.0	55.7
Ours	73.6	65.1	87.6	45.5	50.5	55.1	56.5	79.3	45.0	76.1	67.7	70.1	61.6	69.6	41.8	64.0	66.8	41.3	62.3
Ours†	77.9	66.2	88.1	**45.8**	50.6	63.1	59.6	81.4	48.1	79.1	**69.8**	71.8	62.2	70.2	42.5	65.2	79.3	41.8	65.3

Table 2: Performance on the Spair-71k dataset. Due to inconsistencies in prior works we separate methods based on if they report `per-keypoint` or `per-image` results. Results in the top half of tables are for supervised methods, results in the bottom half of tables are for zero-shot methods. Results marked with † use ground truth annotations to reorder keypoints at test time for the Pose-Align method from [54]. Results for our method are coloured in grey.

5.4 Results

Comparison to State-of-the-art: We compare our method with state-of-the-art methods for the TSS and PASCAL datasets in Tab. 1. Our method improves on state-of-the-art for the TSS dataset, which we attribute to the fact that the dataset contains image pairs with low pose variation and non-deformable objects (*e.g.*, cars, trains, buses) and as such, a strong spatial smoothness prior is particularly beneficial. Our performance on the PASCAL dataset is comparable to existing zero-shot methods but still improves significantly over the nearest neighbour baseline at all thresholds. We also present results for the much more challenging Spair-71k dataset in Tab. 2. Our method lags slightly behind state-of-the art methods that use Stable Diffusion features, particularly in the `Plant` and `TV` categories. We attribute this to the strength of the underlying features, the Stable Diffusion (SD) features outperform DINOv2 features by over 20 percent in these categories. While our method significantly closes the gap, it still relies on strong features. We also include results using the Pose-Align technique from Zhang *et al.* [54], showing that it is orthogonal to our method and can further boost performance at the expensive of additional compute.

Computational Performance: We compare computational requirements for our method and other zero-shot methods in Tab. 3 and Fig. 1. All results are produced from the official repositories and are the average of 100 runs on an Nvidia RTX-4090 GPU. It is clear that other methods are dominated by feature extraction with the matching latency relatively inconsequential. In comparison our matching latency is higher, but still significantly less than Stable Diffusion. Overall our method requires less computational resources while still being competitive with more expensive state-of-the-art methods.

Ablation Study: We perform an ablation study of the different elements of our method in

Method	SD	DINO	Params	Feat. Ext. (ms)	Match (ms)	Total (ms)
DIFT[13]	✓		1.2B	2310	1	2311
TOTF[52]	✓	✓	1.3B	1880	17	1917
TlFR[54]	✓	✓	1.3B	1880	4	1904
DINOv2		✓	87M	97	2	99
Ours (GW)		✓	87M	97	120	217

Table 3: Computational requirements for zero-shot methods. For Stable Diffusion methods we include the VAE encoder and UNet in the parameter count. For DIFT, we include the text encoder as the method uses captions, for ToTF and TlFR we include the visual encoder from CLIP used for implicit captions. Feat. Ext. is the time for extracting features from an image pair, Match is the time for matching between features.

OT	GW	UB	Sym	Pose-Align	PCK($\alpha = 0.1$)	
					55.7	
✓					55.9	+0.3
✓	✓	✓			59.2	+3.3
✓	✓	✓	✓		62.3	+3.1
✓	✓	✓	✓	✓	65.3	+2.0

Table 4: Ablation study on the SPair-71k dataset.

Tab. 4. The GW, unbalanced (UB), and symmetry objective (Sym) all provide meaningful improvements over the baseline. The Pose-Align method [24] can be applied in addition to our method for further improvement.

Limitations: The GW spatial smoothness term is based on the premise that correspondences are spatially consistent — at least within a local neighbourhood. However in cases of extreme scale or pose change this assumption does not hold. Recall Eq. (5) penalises pairs of patches within a radius of δ_{min} that match to patches outside the radius δ_{max}. The first example in Fig. 5 shows a case where our choice of $\delta_{\text{min}}, \delta_{\text{max}}$ is inconsistent with the extreme scale difference. An interesting direction for future work would be to develop methods for estimation of adaptive filter sizes. Our method excels at removing noisy outliers, however can create smooth regions of incorrect correspondences as the proportion of outliers increases as shown in the second example. The one-to-one correspondence assumption is not always valid as shown in the third example. As we are using patch features at a lower spatial resolution, the small scale of the horse means several keypoints in the horse's head belong to the same patch. Lastly our symmetry assumption only holds for moderate pose variations, the fourth example shows an example of opposite viewpoints where this assumption is not valid.

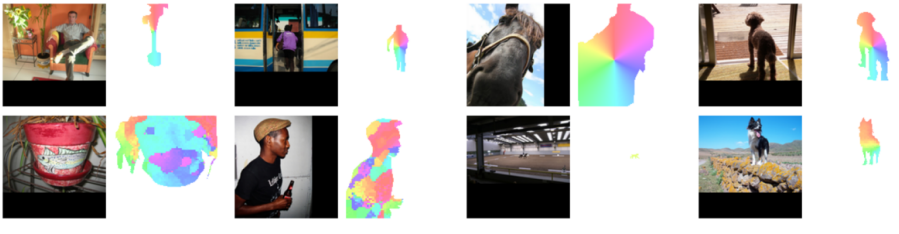

Figure 5: Failure cases of our method. Each two-by-two grid shows dense correspondences for a failure case. First and third examples show failure due to scale, second shows failure due to incorrect GW smoothing, and fourth failure due to invalid symmetry assumption.

6 Conclusion

In this work we present a novel optimal transport based matching algorithm for semantic correspondences. Unlike prior works that rely on Stable Diffusion features to encourage spatial smoothness, our method directly integrates this property into the matching process achieving competitive performance while requiring significantly less compute and memory. The proposed algorithm attains state-of-the-art results on the TSS dataset and competitive results on the PASCAL and SPair datasets. Promising directions for future works include exploring adaptive filter sizes for the Gromov–Wasserstein formulation, extending pairwise penalties to higher-order interactions, and applying our framework to video label propagation tasks.

Acknowledgments. This work was supported by an Australian Research Council (ARC) Linkage grant (LP21020093).

References

[1] A. Khamis, R. Tsuchida, M. Tarek, V. Rolland, and L. Petersson. Scalable Optimal Transport Methods in Machine Learning: A Contemporary Survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, pages 1–20, 2024. ISSN 1939-3539.

[2] Lubomir Bourdev and Jitendra Malik. Poselets: Body Part Detectors Trained Using 3D Human Pose Annotations. pages 1365–1372, September 2009.

[3] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Int. Conf. Comput. Vis.*, pages 9650–9660, 2021.

[4] Laetitia Chapel, Mokhtar Z. Alaya, and Gilles Gasso. Partial Optimal Tranport with applications on Positive-Unlabeled Learning. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Adv. Neural Inform. Process. Syst.*, volume 33, pages 2903–2913. Curran Associates, Inc., 2020.

[5] Minsu Cho, Suha Kwak, Cordelia Schmid, and Jean Ponce. Unsupervised object discovery and localization in the wild: Part-based matching with bottom-up region proposals. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1201–1210, 2015.

[6] Seokju Cho, Sunghwan Hong, Sangryul Jeon, Yunsung Lee, Kwanghoon Sohn, and Seungryong Kim. CATs: Cost Aggregation Transformers for Visual Correspondence. In *Adv. Neural Inform. Process. Syst.*, 2021.

[7] Seokju Cho, Sunghwan Hong, and Seungryong Kim. CATs++: Boosting Cost Aggregation With Convolutions and Transformers. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(6):7174–7194, 2023.

[8] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, volume 1, pages 886–893 vol. 1, June 2005.

[9] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision Transformers Need Registers, 2023. Publication Title: arXiv:2309.16588.

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 248–255, 2009.

[11] Martin A. Fischler and Robert C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6):381–395, June 1981. ISSN 0001-0782.

[12] Frank Fundel, Johannes Schusterbauer, Vincent Tao Hu, and Björn Ommer. Distillation of Diffusion Features for Semantic Correspondence. *IEEE Winter Conf. on Applications of Comput. Vis.*, 2025.

[13] Xiang Gu, Yucheng Yang, Wei Zeng, Jian Sun, and Zongben Xu. Keypoint-Guided Optimal Transport with Applications in Heterogeneous Domain Adaptation. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Adv. Neural Inform. Process. Syst.*, 2022.

[14] Kamal Gupta, Varun Jampani, Carlos Esteves, Abhinav Shrivastava, Abhinav Makadia, Noah Snavely, and Abhishek Kar. ASIC: Aligning Sparse Image Collections. In *Int. Conf. Comput. Vis.*, 2023.

[15] Bumsub Ham, Minsu Cho, Cordelia Schmid, and Jean Ponce. Proposal Flow. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3475–3484, 2016.

[16] Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In *Int. Conf. Comput. Vis.*, pages 991–998, 2011.

[17] Dahyun Kang and Minsu Cho. Integrative Few-Shot Learning for Classification and Segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022.

[18] L. V. Kantorovich. On the Translocation of Masses. *Journal of Mathematical Sciences*, 133(4):1381–1382, March 2006. ISSN 1573-8795.

[19] Zakaria Laskar, Hamed Rezazadegan Tavakoli, and Juho Kannala. Semantic Matching by Weakly Supervised 2D Point Set Registration. In *IEEE Winter Conf. on Applications of Comput. Vis.*, pages 1061–1069, 2019.

[20] Tunan Li, Zhaohuan Zhan, and Guang Tan. Accurate visual localization with semantic masking and attention. *EURASIP Journal on Advances in Signal Processing*, 2022(1): 42, May 2022. ISSN 1687-6180.

[21] Yen-Liang Lin, Vlad I. Morariu, Winston H. Hsu, and Larry S. Davis. Jointly Optimizing 3D Model Fitting and Fine-Grained Classification. In *Eur. Conf. Comput. Vis.*, 2014.

[22] Ce Liu, Jenny Yuen, and Antonio Torralba. SIFT Flow: Dense Correspondence across Scenes and Its Applications. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(5):978–994, 2011.

[23] Yanbin Liu, Linchao Zhu, Makoto Yamada, and Yi Yang. Semantic Correspondence as an Optimal Transport Problem. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4463–4472, 2020.

[24] Jonathan Long, Ning Zhang, and Trevor Darrell. Do convnets learn correspondence? In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, pages 1601–1609, 2014.

[25] David G. Lowe. Object Recognition from Local Scale-Invariant Features. In *Int. Conf. Comput. Vis.*, pages 1150–1157. IEEE Computer Society, 1999.

[26] Grace Luo, Lisa Dunlap, Dong Huk Park, Aleksander Holynski, and Trevor Darrell. Diffusion Hyperfeatures: Searching Through Time and Space for Semantic Correspondence. In *Adv. Neural Inform. Process. Syst.*, 2023.

[27] Octave Mariotti, Oisin Mac Aodha, and Hakan Bilen. Improving Semantic Correspondence with Viewpoint-Guided Spherical Maps. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 19521–19530, June 2024.

[28] Juhong Min, Jongmin Lee, Jean Ponce, and Minsu Cho. SPair-71k: A Large-scale Benchmark for Semantic Correspondence. *CoRR*, abs/1908.10543, 2019. arXiv: 1908.10543.

[29] G. Monge. *Mémoire sur la théorie des déblais et des remblais*. Imprimerie royale, 1781.

[30] Facundo Mémoli. Gromov–Wasserstein Distances and the Metric Approach to Object Matching. *Foundations of Computational Mathematics*, 11(4):417–487, August 2011. ISSN 1615-3383.

[31] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning Robust Visual Features without Supervision, 2023. Publication Title: arXiv:2304.07193.

[32] Gabriel Peyré, Marco Cuturi, and Justin Solomon. Gromov-Wasserstein Averaging of Kernel and Distance Matrices. volume 48 of *JMLR Workshop and Conference Proceedings*, pages 2664–2672. JMLR.org, 2016.

[33] Henri De Plaen, Pierre-François De Plaen, Johan A. K. Suykens, Marc Proesmans, Tinne Tuytelaars, and Luc Van Gool. Unbalanced Optimal Transport: A Unified Framework for Object Detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3198–3207. IEEE, 2023.

[34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. 2021.

[35] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. SAM 2: Segment Anything in Images and Videos. *arXiv preprint*, 2024.

[36] Ignacio Rocco, Relja Arandjelović, and Josef Sivic. End-to-End Weakly-Supervised Semantic Alignment. *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 6917–6925, 2017.

[37] Ignacio Rocco, Relja Arandjelovic, and Josef Sivic. Convolutional Neural Network Architecture for Geometric Matching. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(11): 2553–2567, November 2019. ISSN 1939-3539.

[38] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 10674–10685. IEEE, 2022.

[39] Michael Rubinstein, Armand Joulin, Johannes Kopf, and Ce Liu. Unsupervised Joint Object Discovery and Segmentation in Internet Images. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1939–1946, 2013.

[40] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary R. Bradski. ORB: An efficient alternative to SIFT or SURF. In *Int. Conf. Comput. Vis.*, pages 2564–2571. IEEE Computer Society, 2011.

[41] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperGlue: Learning Feature Matching With Graph Neural Networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4937–4946. Computer Vision Foundation / IEEE, 2020.

[42] Johannes L Schönberger, Marc Pollefeys, Andreas Geiger, and Torsten Sattler. Semantic visual localization. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 6896–6906, 2018.

[43] Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent Correspondence from Image Diffusion. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[44] Tatsunori Taniai, Sudipta N. Sinha, and Yoichi Sato. Joint Recovery of Dense Correspondence and Cosegmentation in Two Images. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4246–4255, 2016.

[45] Matthew Thorpe. Introduction to Optimal Transport, 2018. URL https://www.damtp.cam.ac.uk/research/cia/files/teaching/Optimal_Transport_Notes.pdf.

[46] Alexis Thual, Quang Huy Tran, Tatiana Zemskova, Nicolas Courty, Rémi Flamary, Stanislas Dehaene, and Bertrand Thirion. Aligning individual brains with fused unbalanced Gromov Wasserstein. *Adv. Neural Inform. Process. Syst.*, 35:21792–21804, 2022.

[47] Vayer Titouan, Nicolas Courty, Romain Tavenard, and Rémi Flamary. Optimal transport for structured data with application on graphs. pages 6275–6284. PMLR, 2019.

[48] Prune Truong, Martin Danelljan, Fisher Yu, and Luc Van Gool. Warp Consistency for Unsupervised Learning of Dense Correspondences. *Int. Conf. Comput. Vis.*, pages 10326–10336, 2021.

[49] Prune Truong, Martin Danelljan, Fisher Yu, and Luc Van Gool. Probabilistic Warp Consistency for Weakly-Supervised Semantic Correspondences. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 8698–8708, Los Alamitos, CA, USA, June 2022. IEEE Computer Society.

[50] Narek Tumanyan, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Splicing ViT Features for Semantic Appearance Transfer. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 10748–10757, 2022.

[51] Narek Tumanyan, Omer Bar-Tal, Shir Amir, Shai Bagon, and Tali Dekel. Disentangling Structure and Appearance in ViT Feature Space. *ACM Trans. Graph.*, November 2023. ISSN 0730-0301.

[52] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-Vocabulary Panoptic Segmentation With Text-to-Image Diffusion Models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2955–2966, June 2023.

[53] Ming Xu and Stephen Gould. Temporally Consistent Unbalanced Optimal Transport for Unsupervised Action Segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2024.

[54] Junyi Zhang, Charles Herrmann, Junhwa Hur, Eric Chen, Varun Jampani, Deqing Sun, and Ming-Hsuan Yang. Telling Left from Right: Identifying Geometry-Aware Semantic Correspondence. *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3076–3085, 2023.

[55] Junyi Zhang, Charles Herrmann, Junhwa Hur, Luisa F. Polanía, Varun Jampani, Deqing Sun, and Ming-Hsuan Yang. A tale of two features: stable diffusion complements DINO for zero-shot semantic correspondence. In *Adv. Neural Inform. Process. Syst.*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc. event-place: New Orleans, LA, USA.