Learning Universal Semantic Correspondences with No Supervision and Automatic Data Curation

Aleksandar Shtedritski Andrea Vedaldi Christian Rupprecth
Visual Geometry Group, University of Oxford
{suny, vedaldi, rupprecht}@robots.ox.ac.uk

Abstract

We study the problem of learning semantic image correspondences without manual supervision. Previous works that tackled this problem rely on manually curated image pairs and learn benchmark-specific correspondences. Instead, we present a new method that learns universal correspondences once, from a large image dataset, and without using any manual curation. Despite their generality and despite using less supervision, our universal correspondences still outperform prior works, unsupervised and weakly supervised, in most benchmarks. Our approach starts from local features extracted by an unsupervised vision transformer, which obtain good semantic but poor geometric matching accuracy. It then learns a Transformer Adapter which improves the geometric accuracy of the features, as well as their compatibility between pairs of different images. The method combines semantic similarity with geometric stability obtained via cycle consistency and supervision via synthetic transformations. We use these features to also select pairs of matching images for training the unsupervised correspondences.

1. Introduction

Establishing image correspondences is a necessary step in numerous image analysis tasks, from image understanding to pose estimation and neural rendering. In this work, we tackle the problem of discovering *semantic* correspondences between pairs of images containing different objects or object types (Fig. 1), which is crucial for high-level visual reasoning in diverse environments. Due to the cost of collecting manual annotations for this task, and the absence of large-scale datasets for it, we aim at learning a correspondence predictor without manual supervision.

Most unsupervised approaches for learning correspondences are based on enforcing a form of cycle consistency between images [44, 46, 57, 58]. However, cycle consistency can only tell when correspondences are poor, but cannot suggest or induce meaningful correspondences by itself [57]. Hence, cycle consistency needs to be combined with additional unsupervised learning signals, such as synthetic transformations of the images, from which correspondences can be induced.

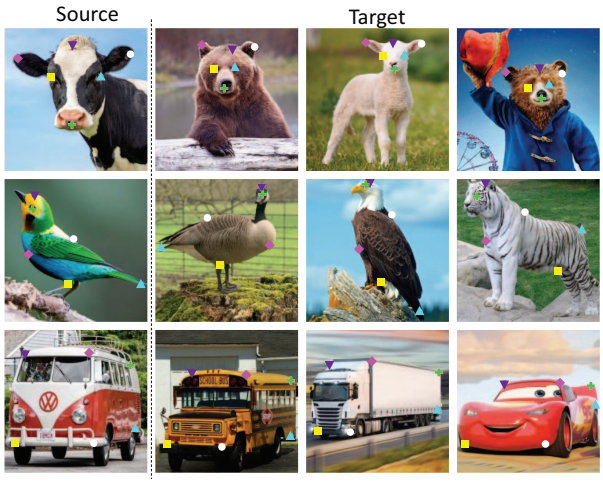

Figure 1. **Unsupervised Universal Semantic Correspondences.** Correspondences found by our method on images across different classes. The method is able to find meaningful correspondences across object categories and poses without any supervision, and with no manual curation of the training data. Best viewed digitally and in color.

Despite progress, it remains difficult to discover correspondences that generalize well to new categories and datasets. A clear sign of this limitation is that prior methods rely on training separate models for each downstream task or benchmark, which often requires domain-specific data curation [34, 42, 44, 46, 57, 58]. In this paper, we develop an unsupervised method that (1) **learns from a single, non-curated dataset** like ImageNet (2) **universal correspondences** that generalize to multiple benchmark datasets. Because of its generality, our correspondence predictor is more versatile than previous ones; furthermore, because training does not require data curation, it can be trained much more easily and cheaply on large datasets.

The first limitation that we address is the one of generalization. Approaches based on synthetic transformations of single images generalize poorly, as they can only learn 'semantically local' correspondences. For example, both a bird and a tiger have a head and a neck, but it is difficult to discover cross-category correspondences by applying random synthetic transformations to either (Fig. 1).

In order to address this limitation, we consider the 'dual' nature of correspondences and concepts. A coherent system of correspondences partitions 2D image points in equivalence classes, each corresponding to a certain concept or abstraction. For example, in multi-view geometry we often put all 2D projections of the same 3D physical point in correspondence, so the physical point can be thought of as the grouping of all its image projections. Likewise, the concept of 'neck' can be thought of as the grouping of all 2D points that correspond to the neck of all animals in all images.

Because of this duality, discovering semantic correspondences is closely related to discovering semantic parts. The latter has been boosted significantly by recent discoveries in self-supervised learning. Models such as DINO [6], that combine transformers (ViT) with deep invariant clustering, have been found to extract features that are well correlated with semantic object parts, all without using any manual supervision. Follow-up works [1, 60] have readily noted that such features can also establish semantic correspondences between images, and these *generalize* particularly well. Even so, methods such as ViT-DINO are *not explicitly designed* for this task. For example, we show empirically that the induced correspondences tend to confuse multiple occurrences of the same semantic part, e.g., by grouping all legs of all animals in a single cluster (Fig. 2). Hence, our goal is to improve the geometric accuracy and robustness of these correspondences. To this end, we make three technical contributions.

Our first contribution is to combine the information that is obtained from ViT-DINO (poor spatial locality but good generalization) and traditional cycle-consistency (good spatial locality but limited generalization). We derive from these a pair of complementary learning signals and distill a high-quality point matcher from them. Experiments clearly show the benefit of this combination, which significantly reduces the spatial uncertainty of the matches.

Our second contribution is a powerful formulation to predict correspondences. We cast this as retrieving a point in the target image given a point in the source, based on comparing descriptors derived from the ViT-DINO features. We learn a transformer network that takes the pre-trained ViT feature of a point in the source image and adapts it to generate a query descriptor to be matched in the second image. We also pass to the transformer a global descriptor of the target image. In this manner, the transformer can improve the locality of the source descriptor and adapt it to

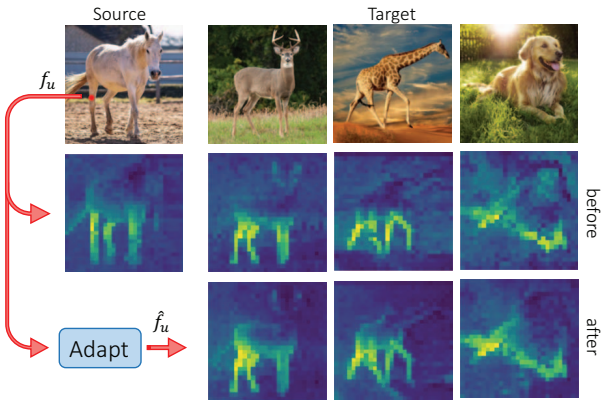

Figure 2. **Feature Similarity.** The heatmap is generated by computing the cosine similarity between a query feature in the leftmost image within the same image and with different ones, using ViT-DINO features. We see that the features capture the image semantics well (the heatmap highlights the legs), but do not provide precise localization (all legs are highlighted). We adapt the query feature f_u to an improved one, $\hat{f}_u$, that is more precise.

bridge appearance differences (e.g., matching a white dog to a black one), further boosting quality.

Our third contribution is an approach to learning unsupervised correspondences using a non-curated dataset such as ImageNet, which addresses the limitation that prior works require manually-curated image pairs. The challenge here is that not all pairs of source and target images can be matched. For example, it makes little sense to match points between a broom and an elephant. Prior work often uses curated pairs of images for training, ensuring their semantic compatibility. Here, we show that it is possible to fully *automatize pair selection* by finding images with sufficiently close global ViT descriptors. Our is the first method to learn point correspondences across different images from a general-purpose dataset without any curation of the matching image pairs.

We test our method against prior work on unsupervised image matching and show large gains in accuracy compared to the baseline ViT-DINO features in numerous benchmark datasets without the need to train separate models. Furthermore, for the most difficult datasets, we also obtain state-of-the-art performance when compared to all prior works on unsupervised keypoint matching, even when compared to weakly supervised methods.

2. Related work

Supervised Semantic Correspondences. Most approaches to semantic correspondences are based on cost volumes over image features, which are subsequently used to estimate displacements represented as geometric transformations [8, 27, 45, 47] or flow fields [28, 56, 57] from source to the target image. The aggregation of the

cost volumes has received particular attention, where 4D convolutions are traditionally used [32, 33, 48]. More recently, transformers have been shown to be very effective for cost aggregation as well [10, 11, 16, 23]. Differently from these works, we aggregate features using *global* rather than *local* descriptors, which is necessary when there are large appearance or pose changes between the pair of images. Other approaches focus on combining multilevel features [37, 39, 62, 67] or using optimal transport [35, 51]. Using pseudo labels and consistency across geometric transformations, [26] show that unsupervised objectives can improve the performance of supervised semantic correspondence methods.

Weakly Supervised and Unsupervised Semantic Correspondences. Early unsupervised approaches artificially warp images and predict the transformation [8, 36, 45, 55, 56], but that usually results in poor generalization. Recently, GANgealing [42] proposed to use GAN supervision for dense visual alignment. It achieves good performance, but it is limited to classes where a pre-trained GAN is available. Similarly, [40] learn a 2D atlas for a given class, but their method only works for a limited number of classes and needs careful manual curation of the images used. [68] learn semantic correspondences using 3D cycle consistency, requiring 3D CAD models during training. In the weakly supervised case, where pairs of positive and negative matches are given, various auxiliary losses can be used [17, 39, 49]. Another form of supervision is using forward-backward consistency [21] and cycle consistency [57, 58]. Truong *et al*. [57] show that some cycles, such as a simple forward-backward cycle, can lead to a degenerate solution where the model simply predicts the identity transformation, and suggest using a particular form of a triplet cycle instead. Some methods use pseudo matches [26] or discard inconsistent matches [46] using somewhat arbitrary thresholds. Instead, we propose a contrastive loss that deals with this in a principled manner. Finally, while all these methods use human-annotated pairs of images, our method is able to train with automatically discovered image pairs. A closely related line of work to unsupervised semantic correspondences is keypoint discovery, where the goal is to discover semantically meaningful keypoints without supervision [19, 20, 30, 50, 66].

Self-Supervised ViT Features. Self-supervised ViT features [6] have been shown to be surprisingly good visual descriptors [1], performing very well in the task of semantic correspondences, and others such as semantic segmentation, co-segmentation, etc. Recently, fine-tuning ViT-DINO features in a task-specific unsupervised way has shown great potential. [15] proposes STEGO, distilling ViT-DINO features for semantic segmentation. While its unsupervised objective clusters semantically similar features, here we do the opposite and learn features that lead to unique correspondences. Aygun *et al*. [2] use an equivariant representation learning approach, inspired by [24, 53, 54]. Similar to our work, the method learns to project self-supervised ViT features to a semantic correspondence space. However, they rely on task-specific datasets with image pairs, whereas we train a single model on unpaired data and evaluate on all correspondence datasets.

Feature Distillation. Feature distillation aims to compress the knowledge of a large and potentially complex model, usually called a teacher, into a smaller one, a student, while preserving performance [5]. Distillation can even improve the performance of the student compared to the teacher [3, 41, 65]. Li *et al*. [34] propose a student-teacher method for semantic correspondences, distilling a probabilistic teacher trained on synthetic data. Distillation of self-supervised ViT features has recently been used for semantic segmentation [15] and neural fields [29, 59].

3. Method

Given a source image $I \in \mathbb{R}^{3\times H\times W}$, a target image $J \in \mathbb{R}^{3\times H'\times W'}$, and a point $u \in [0,1]^2$ representing a 2D location in the source image I, we wish to find the corresponding point $v \in [0,1]^2$ in the target image J. We formulate this task as learning a predictor $\hat{v}_\theta(u; I, J)$ with parameters θ that maps any point u in the source image to the corresponding point v in the target.

To train the predictor $\hat{v}_\theta$ in an unsupervised manner, we score correspondences using ViT-DINO descriptors (Sec. 3.1) which we further improve via a transformer network (Sec. 3.2) that also incorporates geometric constraints (Sec. 3.3). We also automatically curate the training data (Sec. 3.4), a step which is usually done manually.

3.1. Scoring matches

Given the source image I, we use a self-supervised feature network to obtain a tensor $\Phi(I) \in \mathbb{R}^{D\times \frac{H}{K}\times \frac{W}{K}}$ of local feature descriptors (at a reduced image resolution). We use the symbol $f_u = \Phi_u(I) \in \mathbb{R}^D$ to denote the feature vector extracted from the source image I at location u via bilinear interpolation. We also use the symbol $g_v = \Phi_v(J) \in \mathbb{R}^D$ to denote the analogous feature vector for the target image J at location v.

We assess the 'compatibility' of points u and v in the two images by measuring the similarity of their feature vectors [51] using the function

$$s(f,g) = \exp\left(-\frac{\langle f,g\rangle}{\tau\|f\|\|g\|}\right), \tag{1}$$

which approaches 1 when the two descriptors are similar and 0 when they differ. $\tau > 0$ is a temperature parameter that controls the sharpness of the similarity function.

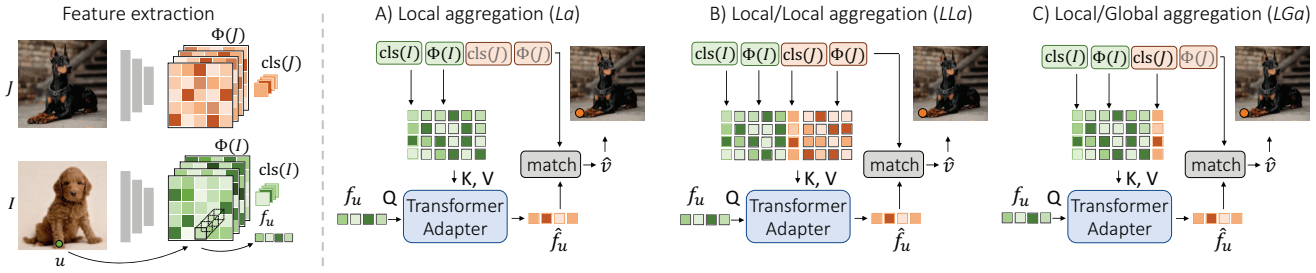

Figure 3. **Matching with a Transformer Adapter.** *Left:* Source and target images are encoded by a self-supervised ViT. We train a cross-transformer A that improves the query feature f_u for semantic matching. *Right:* Architecture variants.

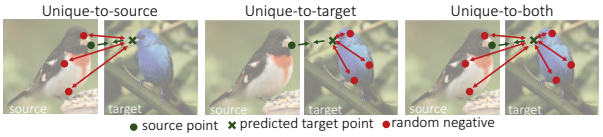

Figure 4. **Variants of the Contrastive Distillation Loss:** Depending on which images the negatives are sampled from, we define unique-to-source, unique-to-target, and unique-to-both losses. We find that the unique-to-both performs best.

We measure the quality of a match (u, v) via the *contrastive distillation loss* [12]:

$$\mathcal{L}_d(v|u, I, J) = -\log \frac{s(f_u, g_v)}{\sum_{h \in \Phi(I) \cup \Phi(J)} s(h, g_v)}. \quad (2)$$

This loss states that the match (u, v) should receive a score larger than most matches (t, v), where t is any location in the source or target images — in Eq. (2) the symbol h denotes the feature vector of this arbitrary point t.

Discussion. As shown in Fig. 4, we choose Eq. (2) as the best among a few alternatives. These losses verify that the target location v matches back to other points t in a unique manner. This works in the reverse order of the predictor $\hat{v}_\theta$, which estimates v from the source u. This is on purpose: next, we construct $\hat{v}_\theta$ by approximately maximizing a similar score in the forward direction, so it would not make sense to also use it for verifying the match.

3.2. Learning a matcher by adapting features

Here we define the function $\hat{v}_\theta$ that maps the source location u to the corresponding target location $\hat{v}_\theta(u; I, J)$. The scoring function s of Eq. (1) provides a natural basis to establish such correspondences: we can define $\hat{v}$ to be the location in the target image J that maximizes the compatibility score $s(f_u, g_{\hat{v}})$ with the location u in the source image I. However, the features used to compute this score can be improved: Fig. 2 shows that matching ViT-DINO feature matches semantic *parts* well, but confuses keypoints. Hence, we learn a function that improves f_u, turning it in a good keypoint descriptor rather than a part descriptor.

We do so by replacing the descriptor f_u of the source point with an improved version $\hat{f}_u = A_\theta(f_u; I, J)$ computed by a learnable *adapter* function A_θ, detailed below. Then, we define the probability $p(v|u) \propto s(A_\theta(f_u; I, J), g_v)$ of matching the source point u to the target point v be proportional to the updated compatibility score. Finally, we define the match

$$\hat{v}_\theta(u; I, J) = \sum_v v \cdot p(v|u) \quad (3)$$

to be the expected value of the target v given the source u. Equation (3) reduces to maximising the compatibility between descriptors $\hat{f}_u$ and g_v when the score temperature τ tends to zero.

Adapter architecture. As discussed in Sec. 1, the goal of the adapter A_θ is to improve the descriptor f_u to disambiguate keypoints and to better match to the style of the target image J. To implement the adapter A, we use the fact that features are extracted by a ViT. This means that in addition to the spatial features $\Phi(I)$, the network extracts also a class token $\text{cls} \in \mathbb{R}^D$ which encodes the overall appearance of the image [61]. By using this token, the network A can adapt the appearance of the source features to the overall style of the target. In practice, A is a cross-attention transformer that takes in the spatial features $\Phi(I)$, the class tokens $\text{cls}(I)$ from the source image and $\text{cls}(J)$ from the target image. This results in a sequence of length $|\Phi(I)|+2$ as key and value inputs to the transformer. Finally, the query to the transformer is the source descriptor f_u (see Fig. 3).

Discussion. Most related work [31, 34, 45, 56, 58] predicts an entire flow field at discrete pixel locations. In contrast, our design allows to evaluate a correspondence $\hat{v}_\theta(u; I, J)$ at individual source locations u with two advantages: (1) we only compute correspondences where they matter (typically the salient part of the image as seen in Sec. 3.4) and (2) u is not limited to exact pixel locations.

We also explored different designs for the transformer, shown in Fig. 3 and Sec. 4. Local aggregation (*La*) is the simplest, aggregating only source features. Local/Local aggregation (*LLa*) aggregates both source and target local features, as in *e.g.* COTR [23]. Our Local/Global aggregation

(*LGa*) combines local and global features from the source but only global features from the target. This is a bottleneck, forcing the adapter to predict what the query feature should look like in the target image.

Empirically, we found *LGa* (Fig. 3) to perform best. The *La* architecture, similar to [2], cannot account for the appearance of the target features by design. The *LLA* architecture, can do so but learns a degenerate solution: the model can satisfy Eq. (2) by *choosing* a feature from $\Phi(J)$ that is similar to f_u but different enough from the other features in $\Phi(I)$. With *LGa*, however, the model needs to *predict* what the query feature should look like in the target image.

3.3. Incorporating geometric constraints

While ViT-DINO features generalize well, they lack spatial precision. The latter can be improved by training $\hat{v}_\theta$ to fit synthetic transformations and to be cycle-consistent. We follow the recommendations of [57] and implement a cycle involving three images: the source image I, the target image J, and a random synthetic transformation $G(I)$ (rotation, flipping, *etc*.) of the source image. Formally, given a point u in the source image, we define the loss:

$$\begin{aligned}\mathcal{L}_c(\hat{v}_\theta|u,I,J) &= a+\beta(b+c), \quad \text{where} \\ a &= \|G(u)-\hat{v}_\theta(\hat{v}_\theta(u;I,J);J,I)\|^2, \\ b &= \|G(u)-\hat{v}_\theta(u;I,G(I))\|^2, \\ c &= \|u-\hat{v}_\theta(G(u);G(I),I)\|^2.\end{aligned} \tag{4}$$

The first term a closes the cycle $I \to J \to I$ from source to the target image and back, whereas terms b and c fit the (known) synthetic correspondences in the direction I to $G(I)$ and $G(I)$ to I. We automatically balance the terms by setting $\beta = \mathrm{sg}(a/(b+c))$ where sg is the stop-gradient operator.

3.4. Choosing images and points to match

Given a dataset of images $\mathcal{D}$ for training, we must choose a subset $\mathcal{P} \subset \mathcal{D}\times\mathcal{D}$ of image pairs $(I,J) \in \mathcal{P}$ that are meaningful to match. Randomly selecting pairs is unlikely to succeed as there are no meaningful correspondences between the vast majority of image pairs (*e.g*. car and fork, hamster and envelope, dog and plant, etc.). Prior work addressed this problem by only using the same image $(I, G(I))$ with augmentations, or by manually selecting meaningful pairs (I,J), usually based on matching the categories of the objects contained in the images. However, manual selection undermines the unsupervised nature of these approaches and also limits correspondences to only those that are manually deemed possible.

Instead, we propose to construct the pairs $\mathcal{P}$ from the images $\mathcal{D}$ automatically, in a process that we call *auto-curation*, based on global image similarity. We use the cls token as a global image descriptor and, given an image I, we construct pairs by randomly sampling from the top-k nearest neighbors images J in the dataset, based on this descriptor. This means that each image is potentially paired with k other images during training, generally yielding many more matching pairs than in the pre-defined case.

Finally, our training scheme also requires to sample points u in the source images. Semantic correspondences are likely only meaningful for foreground objects. Thus, during training, we sample u within the salient region $\Omega_I \subset [0,1]^2$ of the source image I, which we extract using the unsupervised segmentation method of [63].

With this, we can define the overall training objective for the matcher $\hat{v}_\theta$:

$$E(\theta) = \frac{1}{|\mathcal{P}|}\sum_{(I,J)\in\mathcal{P}}\frac{1}{|\Omega_I|}\sum_{u\in\Omega_I}\mathcal{L}(\theta|u,I,J)$$

which combines losses (4) and (2) as:

$$\mathcal{L}(\theta|u,I,J) = \mathcal{L}_d(\hat{v}_\theta(u;I,J)|u,I,J) + \lambda\mathcal{L}_c(\hat{v}_\theta|u,I,J).$$

4. Experiments

To evaluate the effectiveness of our method, we experiment on several benchmarks for semantic correspondences. We compare our automated results with methods trained in an unsupervised and weakly-supervised manner using manually-defined pairs of images of the target classes.

4.1. Datasets and Metrics

We perform experiments on four different datasets: PF-Pascal [14], Spair-71k [38], CUB-200-2011 (CUB) [64] and Stanford Dogs Extra (SDogs) [4, 25].

These datasets cover a wide range of objects and have a different level of difficulty. PF-Pascal contains 1351 image pairs, of which 300 are in the test set, selected from all 20 categories in PASCAL-VOC. The CUB and SDogs datasets contain images for fine-grained recognition. They have sparse keypoint labels, and contain images with varying appearances, poses, and backgrounds, making them particularly challenging. We test on 10,000 random pairs (the same as in [2]). Spair-71K is the most challenging dataset of the four, and the only one collected for the task of semantic correspondences. It contains 1800 images, from which 71k image pairs are created, and contains multiple classes, from animals to man-made objects. What makes it particularly challenging are the different poses and appearance of the objects, their small size in some of the images, and the additional objects present in the scenes.

For all datasets, we evaluate using the standard metric for this task: percentage of correct keypoints (PCK). Given a set of ground-truth points $\mathcal{P} = \{p_m\}_{m=1}^M$ and predictions

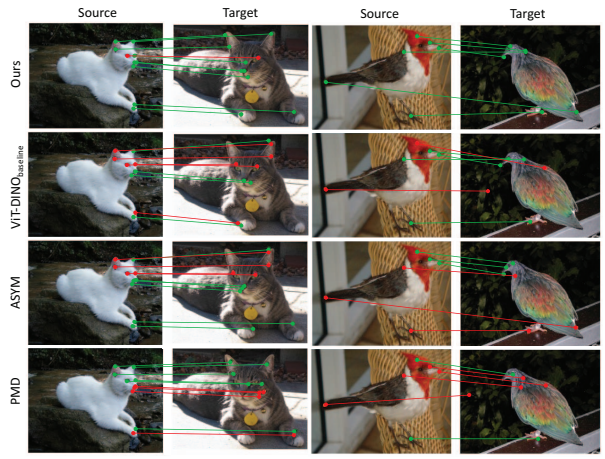

Figure 5. **Qualitative Comparison on SPair-71k**. Our method can find good semantic correspondences under strong appearance differences. PMD achieves state-of-the-art performance on the well-aligned PF-Pascal dataset, but struggles for complex image pairs. We provide more qualitative evaluating in the sup. matt.

	Method	DS	Spair-71k	PF-Pascal	CUB	SDogs
W	GANgealing [42]	✓	—	—	57.5	—
	NACongealing$_{\text{ViT-DINO}}$ [40]	✓	—	—	63.2	—
	WeakAlign $_{\text{res101}}$ [47]	✓	21.1	75.8	—	—
	CL$_{\text{ViT-DINO}}$ [2, 9]	✓	25.8	—	54.1	32.3
	NC-Net$_{\text{res101}}$ [49]	✓	26.4	78.9	—	—
	PMD$_{\text{res101}}$ [34]	✓	26.5	81.2	39.6†	32.2†
	DHPF$_{\text{res101}}$ [39]	✓	28.5	82.1	—	—
	ASYM$_{\text{ViT-DINO}}$ [2]	✓	32.9	—	65.2	45.2
	GSF$_{\text{res101}}$ [22]	✓	33.5	84.5	—	—
	LEAD$_{\text{ViT-DINO}}$ [2, 24]	✓	33.6	—	60.8	42.5
	PWarpC-NC-Net$_{\text{res101}}$ [58]	✓	35.3	84.4	—	—
U	PMD$_{\text{res101}}$ [34]	✓	—	**80.5**	—	—
	CNNGeo$_{\text{res101}}$ [45]	✓	18.1	69.5	—	—
	Sem-GLU-Net$_{\text{vgg16}}$ [56, 58]	✓	16.5	72.5	—	—
	A2Net$_{\text{res101}}$ [52]	✓	20.1	70.8	—	—
	ViT-DINO $_{\text{baseline}}$	✗	34.1	68.3	61.0	42.7
	Ours	✗	**35.9**	73.5	**66.2**	**43.7**

Table 1. **Evaluation.** W: weakly supervised. U: unsupervised. DS: dataset specific models that train on images from the dataset used for evaluation. †: evaluation of [43] using published weights and code of SPair-71k pre-trained model. Our method improves over prior unsupervised *and* weakly supervised work on three out of four challenging benchmarks. PF-Pascal contains only aligned image-pairs, favoring models with strong geometric regularization that often do not generalize well to the other datasets.

$\hat{\mathcal{P}} = \{p_m\}_{m=1}^{M}$, PCK is given by:

$$\text{PCK}(\mathcal{P}, \hat{\mathcal{P}}) = \frac{1}{M} \sum_{m=1}^{M} \mathbb{1}[\|\hat{p}_m - p_m\| \leq \delta].$$

Here, δ is a distance threshold given by $\delta = \alpha \max(H, W)$, where $0 < \alpha < 1$ is a chosen ratio and (H, W) is the image or bounding box size, depending on the dataset convention. Following standard practice, we report α_{img} for PF-Pascal and α_{bbox} for SPair-71K, CUB and SDogs. In all evaluations we use the standard $\alpha = 0.1$ unless stated otherwise.

4.2. Implementation Details

We train a single model on 6M images (to keep NN selection tractable) from ImageNet [13] and evaluate it on all benchmarks. Our experiments use DINO-pretrained ViT-B [7] as feature encoder. All images are resized to 224×224 and the ViT uses 8×8 patches with stride 8 are used as input. We use the 384-dimensional key features from the 9th layer, similarly to prior work [1, 2]. The temperature for the contrastive distillation loss $\mathcal{L}_d$ in Eq. (2) is set to $\tau_d = 0.1$. During training, we set the temperature for the scoring function in Eq. (1) to $\tau_H = 0.05$, and, during testing, we effectively set the temperature to the theoretical limit $\tau_H \to 0$ by using the $\arg\max$ operator in Eq. (3). We set the weight in the expression for the loss $\mathcal{L}$ to $\lambda = 5$.

For the descriptor adapter, we use the transformer decoder architecture of [18] with 4 transformer layers, setting the number of self-attention heads to 8 and the feed-forward dimensionality to 1024. To all spatial descriptors we concatenate sinusoidal positional embeddings, and to the `cls` tokens we concatenate learnable embeddings. The individual queries do not interact with each other, and during test-time only the locations of the query points are computed. During training, we randomly sample 200 points from the salient region of the source image.

In the cycle consistency loop of Eq. (4) we construct the transformation G as follows: (1) we sample a thin-plate spline transformation by randomly jittering a 10×10 grid, (2) we perform a random rescaled crop, and (3) apply color jittering. All further implementation details can be found in the supplementary material. Code and models will be released upon acceptance of the paper.

We use a schedule for the size of the image-pair neighborhood k, gradually increasing k and thus the diversity of pairs during training. In particular, we train with $k = 5$ for the first 10k iterations, $k = 100$ for the next 20k iterations, and $k = 500$ thereafter.

4.3. Results

Table 1 compares our results against the state of the art (weakly and unsupervised methods) on four datasets. We note that all other methods in Tab. 1 are *dataset specific* – they train on images from the same dataset they use for evaluation, whereas we train on one dataset (ImageNet-21k) and evaluate on different datasets at test time.

Our method improves the baseline ViT-DINO features on all datasets. This is in contrast to another recent method [2], whose objective for unsupervised correspondences leads to lower performance on some benchmarks when using ViT-DINO.

We report state-of-the-art results in unsupervised seman-

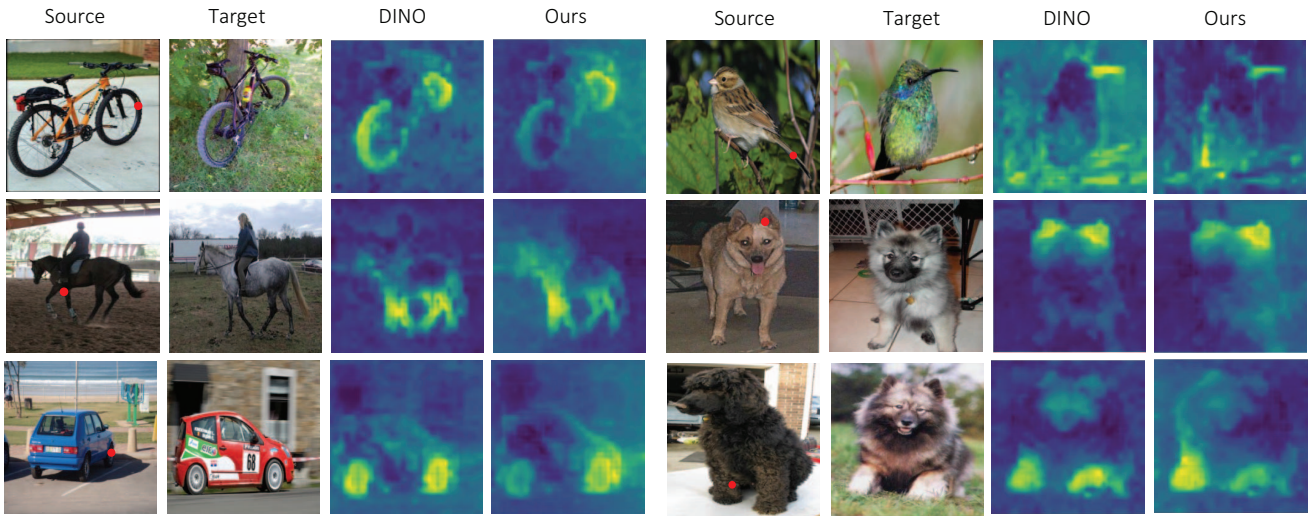

Figure 6. **Similarity Heatmaps.** We show the similarity between the target image and (1) ViT-DINO feature sampled at the query location marked in red (2) an adapted feature by our transformer adapter, resulting in less ambiguous matches. The pairs are from PF-Pascal, SPair71k, CUB and SDogs.

tic correspondences on three out of four datasets, SPair-71K dataset, CUB and SDogs. In SPair-71K and CUB we also outperform weakly-supervised methods which use manually-defined image pairs for all classes.

PMD [34] works better than our approach on the PF-Pascal dataset. PF-Pascal is different from the other benchmarks as it only contains image pairs with lower complexity correspondences: all image pairs are of aligned objects that have the same pose and their relative transformation can be approximated with a simple warp. Models that predict a flow field between the two images (such as PMD) have a clear advantage in this case, but fail when there is a complicated non-rigid transformation of the object between the source and the target. To support this hypothesis, we evaluate PMD's released *weakly* supervised model on CUB and SDogs and indeed see low performance outside of PF-Pascal (see also Fig. 5).

Per-class evaluation on PCK. To better understand the semanticity of the learned correspondences, we evaluate per-class PCK on SPair-71k in Tab. 2. We note that our model mainly achieves state-of-the-art on animal classes. We hypothesize a reason for this lies in the way we sample query points during training: randomly from the salient region of the image. For animals, most keypoints of interest are *within* the body of the animal, while for other objects, such as car, bus, TV, the points of interest usually lie at the boundary. Due to imperfections in the unsupervised saliency masks, our method less often samples boundary points and thus does not perform as well when the query points are at the boundary of the object at test time.

Efffect of weak supervision. To explore the competing effects of using (1) less training data, and (2) manually-defined image pairs (*i.e.*, weak supervision), we train our model on SPair71k, CUB and SDogs (as in [2]), using the provided manually-defined pairs and the same hyperparameters as before. Table 3 shows that using manually supervised image pairs does improve performance, but only slightly (∼1% relative improvement). Note that, by using manually defined image pairs in these smaller datasets, we observe the compound effect of using fewer, cleaner and more specific pairs. However, given that the difference in performance is small, this validates the robustness and *generalization* of our model, which is also *fully unsupervised*.

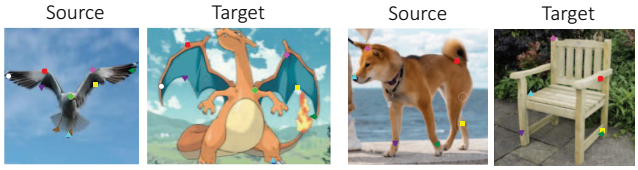

Figure 7. **Qualitative evaluation on out-of-distribution data.** Left: there are still semantic similarities that our model manages to find. Right: failure—there are no semantic correspondences.

Qualitative results. In Fig. 1 we show that our model can find good semantic correspondences across different scales (face, full body), and between categories (bird-tiger). We qualitatively compare our method to prior work in Fig. 5. We see that our method is robust across pose, scale, and appearance changes. Finally, in Fig. 7 we show results on out-of-distribution examples (animation and unrelated class). Our model can find good correspondences when there are semantically similar object parts in the two images.

In Fig. 6 we show heatmaps from cosine similarities between the target images and the ViT-DINO query features or the query features we get from our transformer adapter. We

Sup.	Method																			all
W	WeakAlign [47]	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
	$\text{SFNet}_{\text{res101}}$[31]	26.9	17.2	45.5	14.7	38.0	22.2	16.4	55.3	13.5	33.4	27.5	17.7	20.8	21.1	16.6	15.6	32.2	35.9	26.3
	$\text{PMD}_{\text{res101}}$ [34]	26.2	18.5	48.6	15.3	38.0	21.7	17.3	51.6	13.7	34.3	25.4	18.0	20.0	24.9	15.7	16.3	31.4	**38.1**	26.5
	$\text{ASYM}_{\text{ViT-DINO}}$ [2]	38.7	22.6	68.5	21.2	30.7	27.6	25.3	55.9	18.7	44.5	43.9	33.3	20.7	38.6	16.7	31.7	35.2	25.3	32.9
	$\text{LEAD}_{\text{ViT-DINO}}$ [2, 24]	43.0	24.2	68.6	23.4	31.5	26.2	25.5	57.1	19.5	43.2	42.4	32.0	22.9	41.1	19.4	29.9	34.2	26.2	33.6
U	CNNGeo [45]	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 [52]	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
	$\text{ViT-DINO}_{\text{baseline}}$	**44.5**	24.7	68.6	**24.2**	31.7	26.6	26.0	57.3	19.6	44.3	42.9	33.0	**23.4**	40.6	19.7	30.8	34.3	26.7	34.1
	Ours	40.6	**24.9**	**74.0**	18.9	37.2	26.7	22.3	**59.5**	**20.3**	**49.3**	**46.3**	**37.2**	18.5	**41.5**	**21.3**	**35.2**	**43.8**	32.2	**35.9**

Table 2. **Per-class evaluation on SPair-71k.** W: weakly supervised. U: unsupervised. Over all classes our method performs best. Additionally, in most cases our method improved over its teacher ViT-DINO. Mechanical objects (trains, cars, and motorbikes) are challenging for our method, potentially due to keypoints often appearing at the boundaries of objects which are underrepresented in our training.

Sup.	Methods	SPair71k	CUB	SDogs
W	$\text{CL}_{\text{ViT-DINO}}$ [2, 9]	25.8	54.1	32.3
	$\text{LEAD}_{\text{ViT-DINO}}$ [2, 24]	33.6	60.8	42.5
	$\text{ASYM}_{\text{ViT-DINO}}$ [2]	32.9	65.2	**45.2**
	Ours	**36.4**	**66.8**	44.2
U	$\text{DINO}_{\text{baseline}}$	34.1	61.0	42.7
	Ours	35.9	66.2	43.7

Table 3. **Using Weak Supervision.** Weak supervision trades cleanliness of training pairs with limited amount of training data resulting in marginal improvements.

Loss $\mathcal{L}_c$	Loss $\mathcal{L}_{dT}$	Loss $\mathcal{L}_{dS}$	Arch.	Mask	Spair-71k	CUB	SDogs
✓	✗	✗	*LGa*	✓	17.5	29.0	23.9
✓	✓	✗	*LGa*	✓	31.7	60.6	42.2
✓	✗	✓	*LGa*	✓	33.0	64.2	42.4
✗	✓	✓	*LGa*	✓	33.9	64.4	42.2
✓	✓	✓	*LLa*	✓	32.3	62.9	41.9
✓	✓	✓	*La*	✓	32.8	63.6	42.7
✓	✓	✓	*LGa*	✗	27.6	57.9	39.0
✓	✓	✓	*LGa*	✓	**35.9**	**66.2**	**43.7**

Table 4. **Ablation studies.** Losses: $\mathcal{L}_c$, $\mathcal{L}_{dT}$, and $\mathcal{L}_{dS}$ are the cycle consistency, distillation-from-target, and distillation-from-source losses, respectively. Adaptation: *La*, *LLa*, and *LGa* are Local, Local/Local, and Local/Global adaptation techniques, respectively.

Schedule	Spair-71k	CUB	SDogs
Random	27.9	53.3	39.6
Top_5 NNs	29.5	58.9	40.3
Top_{100} NNs	29.5	58.6	40.6
Top_{500} NNs	28.6	57.1	40.6
$\text{Top}_{5\to100\to500}$ NNs	**35.9**	**66.2**	**43.7**

Table 5. **Pair Selection Strategy.** A curriculum of gradually increasing the sampling neighborhood improves results.

see that the adapter successfully disambiguates between left and right, and front and back. They are also more precise, leading to better localization.

Ablation Study. First, we explore the performance using different losses and architectures in Tab. 4. The losses $\mathcal{L}_{dT}$ and $\mathcal{L}_{dS}$ form the contrastive loss in Eq. (2), and they represent sampling negatives from the target and source image, respectively (as in Fig. 4). We find that sourcing from both images is important, and the geometric prior of the cycle consistency improves the performance of the imperfect ViT-DINO features. Furthermore, adapting using *LGa* is superior to the other methods. Interestingly, *La* performs better than *LLa* — this confirms our hypothesis that, although this architecture is successful in the supervised regime, it is prone to shortcut learning and poor performance in the unsupervised setting. Finally, sampling points from pseudo masks further boosts results. This is expected, as, without the mask, we force the model to find correspondences in the background, which often do not exist.

Next, we look at the importance of using a schedule when selecting nearest neighbor pairs during training (Tab. 5). Using a random pair results in the lowest performance, as often pairs do not contain meaningful correspondences (*e.g.* cup and zebra, etc.). Even when we only sample from a small neighborhood (top five NNs), performance already greatly improves. When we increase the neighborhood size without using a schedule, we see no further gains, and in fact, it drops when we increase to 500. We explain this with the fact that while correspondences across large appearance changes are important for the model to pick up, they are difficult to discover without supervision. We solve this with the schedule that gradually increases the difficulty of the pairs.

5. Conclusions

We have presented a method to learn image correspondences without supervision. The method starts from a self-supervised ViT feature extractor and improves it by learning an adapter network that, given a set of features form a source image and the target image, steers the source features to better match the target. The method can be trained on a general-purpose image dataset without requiring manually-paired images for supervision. It obtains high-quality matches which outperform previous unsupervised methods, particularly on the hardest datasets, where matching requires a good degree of semantic abstraction.

Ethics. We use the PF-Pascal, SPair-71k, CUB and Stanford Dogs Extra datasets in a manner compatible with their terms. Some of these images may accidentally contain personal data (faces), but there is no extraction of personal or biometric information in this research. For further details on ethics, data protection, and copyright please see https://www.robots.ox.ac.uk/~vedaldi/research/union/ethics.html.

Acknowledgements. We thank Tim Franzmeyer, Bruno Korbar and Yash Bhalgat for proofreading. A. Shtedritski is supported by EPSRC EP/S024050/1. A. Vedaldi and C. Rupprecht are supported by ERC-CoG UNION 101001212. C. Rupprecht is also partially supported by VisualAI EP/T028572/1.

References

[1] Shir Amir, Yossi Gandelsman, Shai Bagon, and Tali Dekel. Deep ViT features as dense visual descriptors. *CoRR*, abs/2112.05814, 2021. 2, 3, 6

[2] Mehmet Aygun and Oisin Mac Aodha. Demystifying unsupervised semantic correspondence estimation. In *European Conference on Computer Vision (ECCV)*, 2022. 3, 5, 6, 7, 8

[3] Hessam Bagherinezhad, Maxwell Horton, Mohammad Rastegari, and Ali Farhadi. Label refinery: Improving imagenet classification through label progression. *arXiv preprint arXiv:1805.02641*, 2018. 3

[4] Benjamin Biggs, Oliver Boyne, James Charles, Andrew Fitzgibbon, and Roberto Cipolla. Who left the dogs out?: 3D animal reconstruction with expectation maximization in the loop. In *ECCV*, 2020. 5

[5] Cristian Buciluǎ, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 535–541, 2006. 3

[6] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021. 2, 3

[7] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proc. ICCV*, 2021. 6

[8] Jianchun Chen, Lingjing Wang, Xiang Li, and Yi Fang. Arbicon-net: Arbitrary continuous geometric transformation networks for image registration. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. 2, 3

[9] Zezhou Cheng, Jong-Chyi Su, and Subhransu Maji. On equivariant and invariant learning of object landmark representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9897–9906, 2021. 6, 8

[10] 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. 3

[11] Seokju Cho, Sunghwan Hong, and Seungryong Kim. Cats++: Boosting cost aggregation with convolutions and transformers. *arXiv preprint arXiv:2202.06817*, 2022. 3

[12] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *Proc. CVPR*. IEEE, 2005. 4

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

[14] 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. 5

[15] Mark Hamilton, Zhoutong Zhang, Bharath Hariharan, Noah Snavely, and William T Freeman. Unsupervised semantic segmentation by distilling feature correspondences. *arXiv preprint arXiv:2203.08414*, 2022. 3

[16] Sunghwan Hong, Seokju Cho, Seungryong Kim, and Stephen Lin. Integrative feature and cost aggregation with transformers for dense correspondence. *arXiv preprint arXiv:2209.08742*, 2022. 3

[17] Shuaiyi Huang, Qiuyue Wang, Songyang Zhang, Shipeng Yan, and Xuming He. Dynamic context correspondence network for semantic alignment. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2010–2019, 2019. 3

[18] Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, et al. Perceiver io: A general architecture for structured inputs & outputs. *arXiv preprint arXiv:2107.14795*, 2021. 6

[19] Tomas Jakab, Ankush Gupta, Hakan Bilen, and Andrea Vedaldi. Unsupervised learning of object landmarks through conditional image generation. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2018. 3

[20] Tomas Jakab, Ankush Gupta, Hakan Bilen, and Andrea Vedaldi. Self-supervised learning of interpretable keypoints from unlabelled videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[21] Sangryul Jeon, Seungryong Kim, Dongbo Min, and Kwanghoon Sohn. Parn: Pyramidal affine regression networks for dense semantic correspondence. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 351–366, 2018. 3

[22] Sangryul Jeon, Dongbo Min, Seungryong Kim, Jihwan Choe, and Kwanghoon Sohn. Guided semantic flow. In *European Conference on Computer Vision*, pages 631–648. Springer, 2020. 6

[23] Wei Jiang, Eduard Trulls, Jan Hosang, Andrea Tagliasacchi, and Kwang Moo Yi. Cotr: Correspondence transformer for matching across images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6207–6217, 2021. 3, 4

[24] Tejan Karmali, Abhinav Atrishi, Sai Sree Harsha, Susmit Agrawal, Varun Jampani, and R Venkatesh Babu. Lead: Self-

supervised landmark estimation by aligning distributions of feature similarity. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 623–632, 2022. 3, 6, 8

[25] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Fei-Fei Li. Novel dataset for fine-grained image categorization: Stanford dogs. In *Proc. CVPR workshop on fine-grained visual categorization (FGVC)*, volume 2. Citeseer, 2011. 5

[26] 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

[27] Seungryong Kim, Stephen Lin, Sang Ryul Jeon, Dongbo Min, and Kwanghoon Sohn. Recurrent transformer networks for semantic correspondence. *Advances in neural information processing systems*, 31, 2018. 2

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

[29] Sosuke Kobayashi, Eiichi Matsumoto, and Vincent Sitzmann. Decomposing nerf for editing via feature field distillation. *arXiv preprint arXiv:2205.15585*, 2022. 3

[30] Tejas D Kulkarni, Ankush Gupta, Catalin Ionescu, Sebastian Borgeaud, Malcolm Reynolds, Andrew Zisserman, and Volodymyr Mnih. Unsupervised learning of object keypoints for perception and control. *Advances in neural information processing systems*, 32, 2019. 3

[31] Junghyup Lee, Dohyung Kim, Jean Ponce, and Bumsub Ham. Sfnet: Learning object-aware semantic correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2278–2287, 2019. 4, 8

[32] Jae Yong Lee, Joseph DeGol, Victor Fragoso, and Sudipta N Sinha. Patchmatch-based neighborhood consensus for semantic correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13153–13163, 2021. 3

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

[34] Xin Li, Deng-Ping Fan, Fan Yang, Ao Luo, Hong Cheng, and Zicheng Liu. Probabilistic model distillation for semantic correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7505–7514, 2021. 1, 3, 4, 6, 7, 8

[35] 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

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

[37] 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

[38] Juhong Min, Jongmin Lee, Jean Ponce, and Minsu Cho. Spair-71k: A large-scale benchmark for semantic correspondence. *arXiv preprint arXiv:1908.10543*, 2019. 5

[39] 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. 3, 6

[40] Dolev Ofri-Amar, Michal Geyer, Yoni Kasten, and Tali Dekel. Neural congealing: Aligning images to a joint semantic atlas. *arXiv preprint arXiv:2302.03956*, 2023. 3, 6

[41] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3967–3976, 2019. 3

[42] 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, 3, 6

[43] Simon Prince, Peng Li, Yun Fu, Umar Mohammed, and James H. Elder. Probabilistic models for inference about identity. *PAMI*, 34(1), 2012. 6

[44] Ignacio Rocco, Relja Arandjelovic, and Josef Sivic. Convolutional neural network architecture for geometric matching. In *Proc. CVPR*, 2017. 1

[45] 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. 2, 3, 4, 6, 8

[46] Ignacio Rocco, Relja Arandjelovic, and Josef Sivic. End-to-end weakly-supervised semantic alignment. In *Proc. CVPR*, 2018. 1, 3

[47] 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. 2, 6, 8

[48] Ignacio Rocco, Relja Arandjelović, and Josef Sivic. Efficient neighbourhood consensus networks via submanifold sparse convolutions. In *European conference on computer vision*, pages 605–621. Springer, 2020. 3

[49] I. Rocco, M. Cimpoi, R. Arandjelović, A. Torii, T. Pajdla, and J. Sivic. Neighbourhood consensus networks. In *Proceedings of the 32nd Conference on Neural Information Processing Systems*, 2018. 3, 6

[50] Serim Ryou and Pietro Perona. Weakly supervised keypoint discovery. *arXiv preprint arXiv:2109.13423*, 2021. 3

[51] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4938–4947, 2020. 3

[52] 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. 6, 8

[53] James Thewlis, Samuel Albanie, Hakan Bilen, and Andrea Vedaldi. Unsupervised learning of landmarks by descriptor vector exchange. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6361–6371, 2019. 3

[54] James Thewlis, Hakan Bilen, and Andrea Vedaldi. Unsupervised learning of object frames by dense equivariant image labelling. *Advances in neural information processing systems*, 30, 2017. 3

[55] Prune Truong, Martin Danelljan, Luc V Gool, and Radu Timofte. Gocor: Bringing globally optimized correspondence volumes into your neural network. *Advances in Neural Information Processing Systems*, 33:14278–14290, 2020. 3

[56] Prune Truong, Martin Danelljan, and Radu Timofte. Glunet: Global-local universal network for dense flow and correspondences. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6258–6268, 2020. 2, 3, 4, 6

[57] Prune Truong, Martin Danelljan, Fisher Yu, and Luc Van Gool. Warp consistency for unsupervised learning of dense correspondences. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10346–10356, 2021. 1, 2, 3, 5

[58] Prune Truong, Martin Danelljan, Fisher Yu, and Luc Van Gool. Probabilistic warp consistency for weakly-supervised semantic correspondences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8708–8718, 2022. 1, 3, 4, 6

[59] Vadim Tschernezki, Iro Laina, Diane Larlus, and Andrea Vedaldi. Neural feature fusion fields: 3d distillation of self-supervised 2d image representations. *arXiv preprint arXiv:2209.03494*, 2022. 3

[60] Narek Tumanyan, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Splicing ViT features for semantic appearance transfer. In *Proc. CVPR*, 2022. 2

[61] Narek Tumanyan, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Splicing vit features for semantic appearance transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10748–10757, 2022. 4

[62] Nikolai Ufer and Björn Ommer. Deep semantic feature matching. In *Proc. CVPR*, 2017. 3

[63] Yangtao Wang, Xi Shen, Shell Xu Hu, Yuan Yuan, James L Crowley, and Dominique Vaufreydaz. Self-supervised transformers for unsupervised object discovery using normalized cut. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14543–14553, 2022. 5

[64] Peter Welinder, Steve Branson, Takeshi Mita, Catherine Wah, Florian Schroff, Serge Belongie, and Pietro Perona. Caltech-ucsd birds 200. 2010. 5

[65] Junho Yim, Donggyu Joo, Jihoon Bae, and Junmo Kim. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4133–4141, 2017. 3

[66] Yuting Zhang, Yijie Guo, Yixin Jin, Yijun Luo, Zhiyuan He, and Honglak Lee. Unsupervised discovery of object landmarks as structural representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2694–2703, 2018. 3

[67] 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. 3

[68] Tinghui Zhou, Philipp Krahenbuhl, Mathieu Aubry, Qixing Huang, and Alexei A Efros. Learning dense correspondence via 3d-guided cycle consistency. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 117–126, 2016. 3