ASIC: Aligning Sparse in-the-wild Image Collections

Kamal Gupta[1,2], Varun Jampani[1], Carlos Esteves[1],
Abhinav Shrivastava[2], Ameesh Makadia[1], Noah Snavely[1], Abhishek Kar[1]

[1]Google [2]University of Maryland, College Park

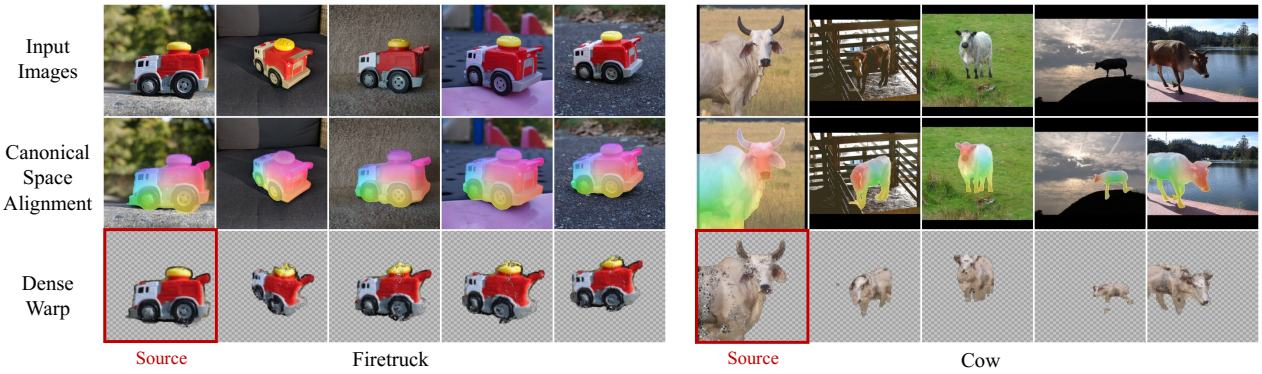

Figure 1: **Globally consistent and dense aligments with ASIC.** Given a small set (∼10-30) of images of an object or object category captured in-the-wild, our framework computes a dense and consistent mapping between all the images in a self-supervised manner. **First row:** Unaligned sets of images from the SAMURAI (Keywest) and SPair-71k (Cow) datasets. **Second row:** Dense correspondence maps produced by our method. **Third row:** Image in the first column warped to the images in columns 2-5.

Abstract

We present a method for joint alignment of sparse in-the-wild image collections of an object category. Most prior works assume either ground-truth keypoint annotations or a large dataset of images of a single object category. However, neither of the above assumptions hold true for the long-tail of the objects present in the world. We present a self-supervised technique that directly optimizes on a sparse collection of images of a particular object/object category to obtain consistent dense correspondences across the collection. We use pairwise nearest neighbors obtained from deep features of a pre-trained vision transformer (ViT) model as noisy and sparse keypoint matches and make them dense and accurate matches by optimizing a neural network that jointly maps the image collection into a learned canonical grid. Experiments on CUB, SPair-71k and PF-Willow benchmarks demonstrate that our method can produce globally consistent and higher quality correspondences across the image collection when compared to existing self-supervised methods. Code and other material will be made available at https://kampta.github.io/asic.

1. Introduction

Given an image of a car, we as humans, can easily map corresponding pixels between this car and an arbitrary collection of car images. Our visual system is able to achieve this (rather impressive) feat using a multitude of cues - low level photometric consistency, high level visual grouping and our priors on cars as an object category (shape, pose, materials, illumination etc.). The above is also true for an image of a "never-before-seen" object (as opposed to a common object category such as cars) where humans demonstrate surprisingly robust generalization despite lacking an object or category specific priors [7]. These correspondences in turn inform downstream inferences about the object such as shape, affordances, and more. In this work, we tackle this problem of "low-shot dense correspondence" – *i.e.* given only a small in-the-wild image collection (∼10–30 images) of an object or object category, we recover dense and consistent correspondences across the entire collection.

Prior works addressing this problem of dense alignment in "in-the-wild" image collections assume availability of annotated keypoint matches and image pairs [13, 55], a mesh of the object [42, 86], or a very large collection of images

of the object [58, 61]. These assumptions often do not hold for the long-tail of objects that exist in real world imagery. This long-tail is unavoidable; no matter how many new images we annotate, we will keep uncovering new and rare categories of objects. In our work, we show that it is possible to achieve dense correspondence of small in-the-wild image collections without any manual annotations by leveraging the power of large self-supervised vision models. Aligning these image sets can be useful for a wide range of applications such as edit propagation for images and videos, as well as downstream problems such as pose and shape estimation.

In the presence of a limited number of samples and a high-dimensional search space, dense correspondence and joint alignment of an image set is a challenging optimization problem. We draw inspiration from classical image alignment methods [44, 74] where images are warped (or congealed) to a consistent canonical pose before classification using simple transformations, as well as recent works on per-image-set optimization [39, 53, 80], where the inductive model biases coupled with additional regularization allows for learning a good solution with self-supervision. Our framework, dubbed ASIC, consists of a small image-to-image network which predicts a dense per-pixel mapping from the image to a two-dimensional canonical grid. This canonical grid is parameterized as a multi-channel learned embedding and stores an RGB color along with an alpha value indicating whether the location represents the object or the background. We devise a novel contrastive loss function to ensure that semantic keypoints from different images map to a consistent location in canonical space.

The key contribution of our work is to exploit noisy and sparse pseudo-correspondences between a pair of images and extend them to learn consistent dense correspondences across the image collection. These pseudo-correspondences can be obtained using any of the large self-supervised learning (SSL) models [10–12, 22, 24, 57, 62] which learn without explicit labels on large internet-scale data. In order to make them accurate, we enforce pair-wise consistency across the image collection with an alignment network and a self-supervised keypoint consistency loss. Further, we introduce additional regularization via equivariance and reconstruction terms to get dense correspondences across collection.

Fig. 1 demonstrates the dense and consistent mapping learned by our model for two image sets. We also evaluate our method on 18 image categories in SPair-71k [56] dataset, 4 categories in PF-Willow [23], 3 fine-grained categories in the CUB [81], as well as 5 collections in SAMURAI [9] datasets and show that ASIC is competitive against unsupervised keypoint correspondence approaches, and often outperforms them. An additional advantage of learning a joint canonical mapping is that our method suffers significantly less drift when propagating keypoints on a sequence of images (instead of just a single image pair). In order to evaluate the keypoint consistency over a sequence of k images, we propose a new metric k-CyPCK (or k-cycle PCK) in Sec. 4.4 and show that our method outperforms existing methods by over 20% at both the low and high precision settings. In summary, our contributions are as follows:

- We introduce a test-time optimization technique to recover consistent dense correspondence maps over a small collection of in-the-wild images.
- We design a novel loss function and several regularization terms to encourage mapping to be consistent across multiple images in a given collection.
- We perform extensive quantitative and qualitative evaluations on 4 different datasets (spanning 30 object categories) to show that our method is competitive with the unsupervised methods, often outperforming them.
- We propose a novel metric k-CyPCK to evaluate consistency of keypoint propagation over a sequence of images, which is not captured by traditional metrics such as PCK.

2. Related Work

Correspondence between image pairs. Keypoint matching or correspondence between images is one of the oldest tasks in computer vision. Some very early works focused on finding dense optical flow [6, 8, 26] between pairs of consecutive images in videos via a variational framework to optimize flow based on pixel intensities. Sparse keypoint matching, e.g., using SIFT descriptors [51, 84], also gained importance due to applications in tracking [52, 78] and structure from motion (SfM) [2, 19, 70]. SIFT Flow [49] proposed the idea of using SIFT descriptors for dense alignment between image pairs. Initial deep learning based correspondence works [15, 17, 40] replaced SIFT with deep features. With the availability of labeled datasets, a number of works have performed end-to-end matching with deep networks [29, 36, 46–48, 50, 54, 55, 65, 68, 73, 88]. However, a shortcoming of these aforementioned works is that they usually require large labeled datasets, and often fail to generalize on unseen objects or scenes.

Joint alignment of image sets. The concept of a canonical image has long been used for the task of object detection via template matching [20, 32]. Learned-Miller *et al*. [28, 44] formalized the task of jointly aligning a set of images (i.e., congealing them) by continuously warping each image (*e.g*. via affine transformations) to minimize the entropy distribution of the image set. [27] use deep features from multiple resolutions in place of hand-crafted features. GANgealing [61] extended this idea by constraining the canonical image to be the output of a pre-trained StyleGAN [21, 38]. In a similar vein, CoordGAN [58] trains a structure-texture disentangled GAN with a canonical coordinate frame as input. Both of these works attempt to solve a similar tasks as ours,

but are limited by data-hungry GAN training. Some works exploits 3D shape as a means for consistent dense correspondences across image collections [37, 42, 43, 83] but require access to additional signals such as category specific 3D templates, segmentation masks or keypoint correspondences. In contrast, our work attempts to learn dense correspondences in a low-shot setting where GAN training is infeasible and in the absence of additional training signals. As mentioned before, we do so primarily by leveraging large pre-trained SSL models as our source of semantic priors on general imagery.

Self-supervised correspondence discovery. To overcome the lack of large datasets with ground-truth correspondence, recent work seeks to combine the idea of distilling deep features from a network trained with self-supervision on large-scale image datasets. Some of these works optimize for proxy losses computed with known transformations [5, 34, 41, 59, 64, 71, 75, 76, 79]. Like these methods, we also train our network to be equivariant to synthetic geometric transformations. However, a key difference is that we also train with pseudo-correspondences 'across' real images, which allows the method to generalize better and build a consistent mapping across the given image collection.

Deep Matching Prior [25] and Neural Best Buddies [1] optimize for only a single pair of images to match deep features of one image to another. More recently PSCNet [35] and Neural Congealing [60] train large networks for simultaneously matching deep features for image pairs by learning a flow from image to image, and image to canonical space respectively. However, these methods have limited flexibility in the deformation space and do not generalize well to out of plane rotations present in datasets such as SPair-71k. We allow our model to map different image regions arbitrarily to different parts of the canonical space. In Tab. 1, we show that this allows us to generate more accurate correspondences and generalize to more object categories.

3. ASIC Framework

Given a collection of images of an object or an object category, our goal is to assign corresponding pixels in all the images to a unique location in a canonical space. By doing so, we can use this learned canonical space as an intermediary when mapping pixels from one image to any another image in the collection while guaranteeing global consistency. The absence of ground truth annotations, small size of the datasets we consider ($\sim$10 - 30 images) and the presence of occlusions and variations in shape, texture, viewpoint, and background lighting all serve to make this task highly challenging. We introduce a simple yet robust framework with a novel self-supervised contrastive loss function over image pairs, as well as auxiliary regularization losses on this learned canonical space to find consistent dense correspondences across the collection.

3.1. Obtaining Pseudo-correspondences

Prior and concurrent works have shown that deep features extracted from large pre-trained networks contain useful local semantic information [10, 14, 31, 85]. In this work, we use DINO [10] to extract these local semantic features. Note that these features are only extracted for obtaining pseudo-correspondences only once and are not used during the training. Given a pair of images $\mathbf{I}^{\mathrm{a}}$ and $\mathbf{I}^{\mathrm{b}}$, we obtain feature maps $\mathbf{F}^{\mathrm{a}}$ and $\mathbf{F}^{\mathrm{b}}$ using DINO. Here $\mathbf{F}^{\mathrm{a}} = \{\mathbf{f}_i^{\mathrm{a}}\}$ and $\mathbf{F}^{\mathrm{b}} = \{\mathbf{f}_i^{\mathrm{b}}\}$ represents the sets of feature vectors $\mathbf{f} \in \mathbb{R}^{\mathrm{d}}$ for all spatial locations $\mathbf{p}_i = (x, y) \in \mathbb{R}^2$. In practice, we obtain these feature maps at a coarser resolution, but for brevity, we do not introduce new notations for low-resolution feature maps. We define our pseudo keypoint correspondences, between the two images $\mathbf{I}^{\mathrm{a}}$ and $\mathbf{I}^{\mathrm{b}}$ as all pairs of locations of feature vectors that are mutual nearest neighbors, *i.e.*,

$$\{(\mathbf{p}_i^{\mathrm{a}}, \mathbf{p}_j^{\mathrm{b}}) \mid \left(\mathrm{NN}(\mathbf{f}_i^{\mathrm{a}}, \mathbf{F}^{\mathrm{b}}) = \mathbf{f}_j^{\mathrm{b}}\right) \wedge \left(\mathrm{NN}(\mathbf{f}_j^{\mathrm{b}}, \mathbf{F}^{\mathrm{a}}) = \mathbf{f}_i^{\mathrm{a}}\right)\}$$

where $\mathrm{NN}(\mathbf{f}_i^{\mathrm{a}}, \mathbf{F}^{\mathrm{b}})$ corresponds to the nearest neighbor of the normalized feature vector $\mathbf{f}_i^{\mathrm{a}}$ in the set of feature vectors $\mathbf{F}^{\mathrm{b}}$. The mutual nearest neighbors are usually noisy and sparse, and they serve as pseudo-correspondences for training our alignment network which we discuss next. Rather than using VGG-19 [72] features as proposed in [1], we use DINO features [4, 10]. In general, we observe that better or more discriminative features lead to a better performance in our downstream task as also shown in Section 4.3.

3.2. Architecture

Fig. 2 gives the high-level overview of the framework. Formally, we are given an image collection consisting of N images $\{\mathbf{I}^k\}_{k=1}^N$. We want to train an alignment network $\mathbf{\Phi}_{\text{align}}$ that takes a single image as input at a time and outputs $\mathbf{C} = \mathbf{\Phi}_{\text{align}}(\mathbf{I})$, the canonical space coordinate map, of the image. The canonical space coordinate map $\mathbf{C}$ has the same spatial dimensions as the input $H \times W$ and contains (u, v) coordinates in the shared canonical grid for that location. We parameterize this alignment network $\mathbf{\Phi}_{\text{align}} : \mathbb{R}^{H \times W \times 3} \rightarrow \mathbb{R}^{H \times W \times 2}$ with a fully convolutional U-Net [66] trained from scratch for the collection. Each pixel location $\mathbf{p} = (x, y)$ of this map consists of a 2-dimensional $\mathbf{c} = (u, v)$ coordinate.

These coordinates corresponds to a location in a learned canonical grid $\mathbf{G} \in \mathbb{R}^{H' \times W' \times 4}$. The canonical grid $\mathbf{G}$ is also two-dimensional but can have arbitrary height and width $H' \times W'$, and is shared by all the images in the collection. Each location in canonical grid $\mathbf{G}$ stores an (r, g, b, α) value which corresponds to colors (r, g, b) and a probability α that this location corresponds to a foreground pixel in the image. The original image, and a foreground visibility mask can now be reconstructed using this shared canonical grid $\mathbf{G}$, canonical space mapping $\mathbf{C}$, and a differentiable warp operator commonly used in spatial transformer networks [33]. For

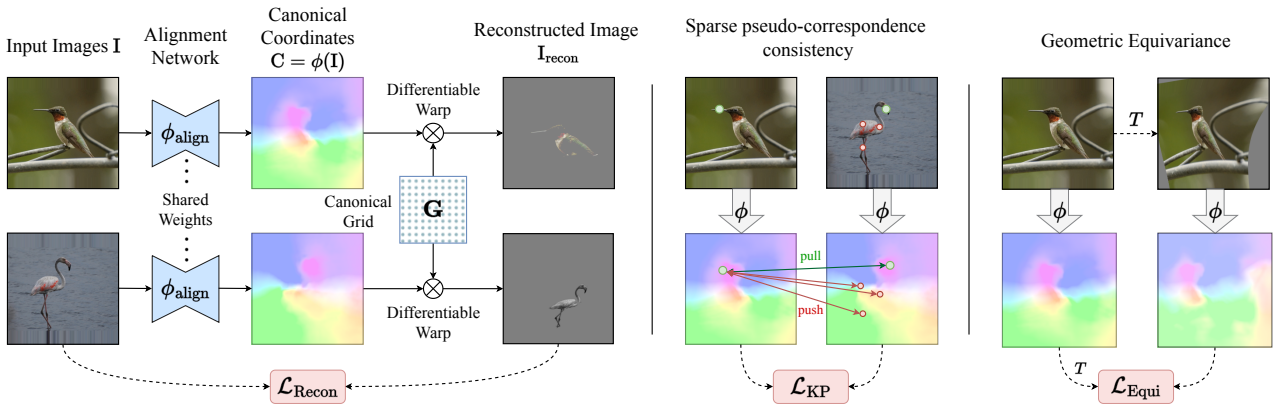

Figure 2: **ASIC Architecure.** The alignment network $\mathbf{\Phi}_{\text{align}}$ predicts canonical space coordinates for all pixels for all input images. Images can be reconstructed using a differentiable warp from the canonical space. In order to align semantically similar pixels from different images to the same location in the canonical space, we propose two primary loss functions $\mathcal{L}_{\text{KP}}$ and $\mathcal{L}_{\text{Recon}}$. Please refer to the Sec. 3.2 for more details.

the mapping to be meaningful, we want semantically similar points from different images to map to the same location in the canonical space. In the next section, we describe the training loss we devised to this end.

3.3. Training Objectives

Sparse pseudo-correspondence consistency. The central goal of our framework is to ensure that semantically similar points in the images are aligned in the canonical space. Recall from the Sec. 3.1, that we pre-compute the pseudo-correspondences between all pairs of images using mutual nearest neighbors in the SSL feature space. Since SSL models are not trained for the task of correspondence, the pseudo-correspondences are noisy and sparse. Our first loss term is targeted at improving the accuracy of the correspondences by jointly aligning them for all pairwise combinations of images in our collection. Formally, given an image pair $(\mathbf{I}^{\text{a}}, \mathbf{I}^{\text{b}})$, we denote all the K pseudo-correspondences in the pair by $\{\mathbf{p}_i^{\text{a}}, \mathbf{p}_i^{\text{b}}\}_{i=1}^{K}$. We apply the alignment network $\mathbf{\Phi}_{\text{align}}$ to $\mathbf{I}^{\text{a}}$ and $\mathbf{I}^{\text{b}}$ independently to obtain the canonical space coordinates for each pixel in the pair, which we denote by $\{\mathbf{c}_i^{\text{a}}, \mathbf{c}_i^{\text{b}}\}_{i=1}^{K}$. We want to map each keypoint location in $\mathbf{I}^{\text{a}}$ as close as possible to its counterpart in $\mathbf{I}^{\text{b}}$, while pushing it away from the mapping of other keypoints in $\mathbf{I}^{\text{b}}$. To achieve this, we define our first loss function $\mathcal{L}_{\text{KP}}$ as

$$\mathcal{L}_{\text{KP}} = -\sum_{i=1}^{K} \log \frac{\exp(-\parallel \mathbf{c}_i^{\text{a}} - \mathbf{c}_i^{\text{b}} \parallel^2 / \tau)}{\sum_{j=1}^{K} \exp(-\parallel \mathbf{c}_i^{\text{a}} - \mathbf{c}_j^{\text{b}} \parallel^2 / \tau)} \tag{1}$$

where τ is a hyperparameter Very small values of τ might lead the network to place too much emphasis on mismatched keypoints, and thus lead to training instability. Very high values, on the other hand, might not yield a end up not providing meaningful loss to train the alignment network. We use a fixed value of $\tau = 1.0$ in all our experiments. $\mathcal{L}_{\text{KP}}$ plays the key role in improving the accuracy of pseudo-correspondences jointly for all images in our collection. However, the number of pseudo-correspondences is still very small (typically 100-300 for an image pair) as compared to the number of pixel locations. Hence, this loss is sparse and we need to add extra regularization terms in order to learn dense alignment, that we will discuss next.

Geometric transformation equivariance. The loss $\mathcal{L}_{\text{KP}}$ is a sparse constraint on the canonical space, *i.e.* it is computed only at a few image locations and not densely. Furthermore - depending on the quality of the SSL features - the accuracy of keypoint matches, and hence the loss, can vary significantly. In order to make our learned mapping dense, we introduce a geometric equivariance regularization term in our loss function. We apply a random synthetic geometric transformation $\mathcal{T}$ to a given image $\mathbf{I}$. Since the output of the alignment network $\mathbf{\Phi}_{\text{align}}$ learns the canonical space coordinates for each location of input image, we can apply the same geometric transformation $\mathcal{T}$ to $\mathbf{\Phi}_{\text{align}}(\mathbf{I})$, and enforce an equivariance loss as follows

$$\mathcal{L}_{\text{Equi}} = \parallel \mathcal{T}(\mathbf{\Phi}_{\text{align}}(\mathbf{I})) - \mathbf{\Phi}_{\text{align}}(\mathcal{T}(\mathbf{I})) \parallel \tag{2}$$

where $\mathcal{T}$ is the geometric transformation. We choose thin plate spline (TPS) transformations [18] in our work, commonly used for image warps. $\mathcal{L}_{\text{KP}}$ and $\mathcal{L}_{\text{Equi}}$ serve as the two primary loss functions for the image set alignment problem, serving the purpose of making the pseudo-correspondences accurate and dense respectively. To further aid the training, we also propose the following auxiliary regularizations.

Total variation regularization. In order to encourage smooth mappings from from each image to the canonical space, we add a total variation (TV) regularization to the computed mapping $\mathbf{C}$. We found TV loss to be crucial to

mitigate degenerate solutions (see Sec. 4.5):

$$\mathcal{L}_{\text{TV}} = \mathcal{L}_{\text{Huber}}(\Delta_x(\mathbf{C} - \mathcal{I})) + \mathcal{L}_{\text{Huber}}(\Delta_y(\mathbf{C} - \mathcal{I})) \quad (3)$$

where $\mathcal{I}$ is the identity mapping (*i.e.* each pixel (x, y) in the image gets mapped to (x, y) in the canonical space), Δ_x and Δ_y denote the partial derivatives under finite differences w.r.t. x and y dimensions, and $\mathcal{L}_{\text{Huber}}$ denotes the Huber loss [30].

Reconstruction loss. All the loss terms so far are computed on the canonical coordinates given by $\mathbf{\Phi}_{\text{align}}$ and does not use the canonical grid $\mathbf{G}$. Recall that $\mathbf{G}$ is of the size $H' \times W' \times 4$, and allows us to reconstruct each image as well as a foreground visibility mask via a differentiable warp operator ($\mathcal{W}$) [33] such that $\mathbf{I}_{\text{Recon}}, \mathbf{M}_{\text{Recon}} = \mathcal{W}(\mathbf{G}, \mathbf{C})$. This allows us to compute a per image reconstruction loss using the original and reconstructed images. However a simple L_1 or L_2 loss will not suffice since $\mathbf{G}$ is shared for all the images in the collection and these images may come from wildly different backgrounds and lighting conditions. Furthermore, the two images might contain two different instances of the same object class, and may have different textures, shapes, and viewpoints. We instead minimize the perceptual (LPIPS) loss [87] which measures a perceptual patch level similarity between images. For the reconstructed mask, we compute pixel-wise binary cross entropy (BCE) loss using the image foregrounds obtained with co-segmentation [4].

$$\begin{aligned} \mathbf{I}_{\text{Recon}}, \mathbf{M}_{\text{Recon}} &= \mathcal{W}(\mathbf{G}, \mathbf{C}) \\ \mathcal{L}_{\text{Recon}} &= \text{LPIPS}(\mathbf{I}, \mathbf{I}_{\text{Recon}}) + \text{BCE}(\mathbf{M}, \mathbf{M}_{\text{Recon}}) \end{aligned} \quad (4)$$

Consistent part alignment. For our final auxiliary loss, we obtain part co-segmentation maps from the images [4, 14] by clustering deep ViT features into S semantic parts, then running GrabCut [67] to smoothen the part boundaries. Our hypothesis is that semantically similar parts in the images should get mapped to similar location in $\mathbf{G}$, Formally, for each image $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$, we obtain semantic part masks as a binary matrix, $\mathbf{P} \in \mathbb{R}^{H \times W \times S}$. Since we want a part across the image set to map to a compact location in the canonical space, we minimize the variance of the canonical space coordinates for all pixels belonging to a part:

$$\mathcal{L}_{\text{Parts}} = \sum_{s=1}^{S} \frac{1}{N_s} \sum_{i} \parallel \mathbf{c}_i^s - \mathbb{E}(\mathbf{c}^s) \parallel^2 \quad (5)$$

where N_s is the number of pixels belonging to the part, $\mathbf{c}_i^s$ is the canonical coordinates of i^{th} pixel location belonging to the s^{th} part, and $\mathbb{E}(\mathbf{c}_i^s)$ is the centroid of the s^{th} part. We fix the number of parts to 8 in all our experiments. Alternately, the number of parts can be computed using the elbow method [77] at the expense of additional compute.

4. Experiments

We evaluate our method on several real-world in-the-wild image collections of both rigid and non-rigid object categories. For all datasets, we use a fixed set of hyperparameters (provided in the appendix) unless specified otherwise.

Datasets. SPair-71k [56] consists of 1,800 images from 18 categories. We optimize over image collections derived from the SPair-71k test set for each category independently and report results on each individual category, as well as aggregate results over all 18 categories. In case of **PF-Willow** [23], we consider all 4 categories of the dataset containing $\sim$30 images. **CUB-200** [81] datasets consists of over 200 fine-grained categories. We optimized our model on the test sets of first 3 categories of the dataset, consisting of 15-20 images each. We also show qualitative results on 4 objects from **SAMURAI** dataset [9].

4.1. Canonical Space Alignment

One simple way to visualize the alignment of an image set when mapped to the canonical space $\mathbf{G}$ is to define a colormap over the canonical grid and color the image pixels according to their mapped location in the canonical grid. In Fig. 3, the first row for each collection contains sample input images. The second row shows discrete parts obtained via parts co-segmentation using [4]. While these parts are also consistent across the image set, our canonical space mapping (third row) can be seen as a dense and continuous co-segmentation. We show the results for six datasets: CUB-200 birds; Dogs, Cats, and Train from SPair-71k; and Robot and Shoe from SAMURAI. The colormap used for the canonical space is provided in the supplement. We observe that our method can find dense correspondences across highly varying poses, backgrounds, and lighting. It also maps common parts of objects in a dataset to nearby regions of the canonical space. This is evident in Fig. 3 where, for instance, the faces of different cats are colored similarly.

4.2. Visualizing Dense Correspondences

We can also find dense correspondences between a pair of images $\mathbf{I}^{\text{a}}$ and $\mathbf{I}^{\text{b}}$ using our framework. Recall that $\mathbf{\Phi}_{\text{align}}$ outputs canonical space coordinates $\mathbf{c}^{\text{a}}$ and $\mathbf{c}^{\text{b}}$ for each pixel location $\mathbf{p}^{\text{a}}$ and $\mathbf{p}^{\text{b}}$. In order to warp the source image $\mathbf{I}^{\text{a}}$ to a target image $\mathbf{I}^{\text{b}}$, for every foreground pixel $\mathbf{p}_i^{\text{a}}$ in $\mathbf{I}^{\text{a}}$, we need to find its nearest neighbor among the set of points in $\mathbf{p}^{\text{b}}$ in the canonical space: We perform this action for all the foreground pixels in the source image, and splat according to the nearest neighbor mapping to get our desired warped image. Fig. 4 shows qualitative results for 10 different datasets. The top row is the source image with foreground mask highlighted. The second row is the target image. We show results for two other pairwise image optimization approaches, and then our method in the last row. NBB [1] computes nearest neighbors using VGG-19 and applies a Moving Least Squares (MLS) optimization [69] to compute a dense flow from source to target. While the flow computed via MLS

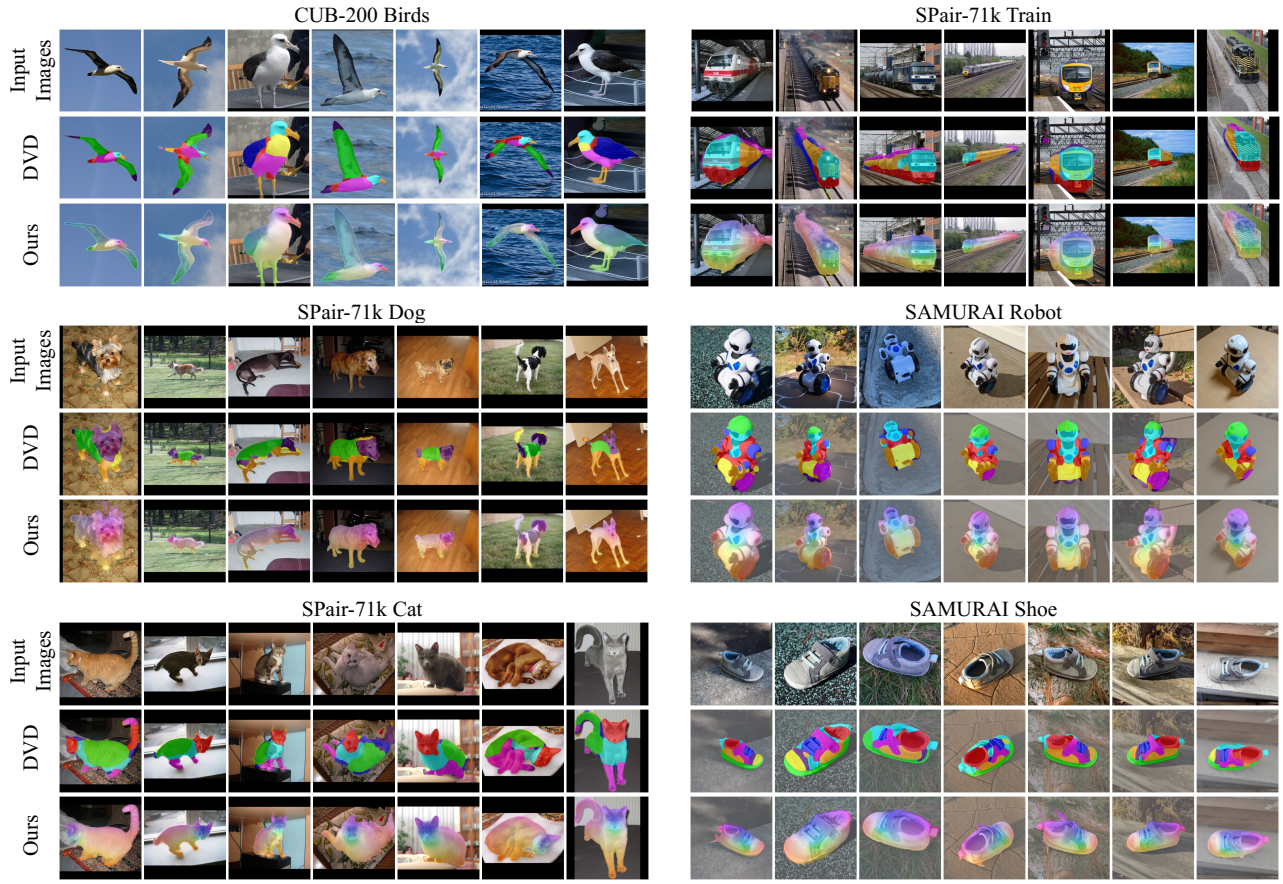

Figure 3: **Visualizing canonical space alignment.** For each dataset, the top row shows sample images from the dataset (composed of 10-30 images each). The middle row shows part co-segmentations computed by DVD [4]. DVD computes a coarse, discrete set of parts across the dataset. The bottom row shows the continuous canonical space mapping computed by our method. Our canonical space mapping is smooth and consistent across the images for each dataset/collection.

is smooth, it usually does not respect semantic correspondences, as evident in the figure. We extend their technique to use DINO features as well. With DINO and the nearest neighbor approach (DVD) [4], the semantic correspondences are arguably better, but since this approach relies on the output of a Vision Transformer (ViT), which has lower resolution than the image, it produces a sparse flow. Our method produces both dense and consistent flow between an image pair.

4.3. Pairwise Correspondence

Metric. For evaluating accuracy of pairwise correspondence, we use the PCK metric [82] (percentage of correct keypoints) on the SPair, CUB, and PF-Willow datasets.

Baselines. We categorize prior works based on the supervision used: (1) *Strong supervision* methods utilize human-annotated keypoints to learn pairwise image correspondence and achieve the best performance (on average). We include the numbers from a recent work [29] for reference purposes. (2) *GAN supervision* methods like [61] use a category-specific GAN pre-trained with large external datasets. While this method works well, it is restricted to only the categories for which large datasets are available and GAN training is feasible. (3) *Weak supervision* methods use category-level supervision (*i.e.* they assume that given pair/collection of images are from same category). They often resort to fine-tuning a large ImageNet [16] pre-trained network using a self-supervised loss function (*e.g.* with synthetic transformations) and optionally use additional information such as foreground masks or matching image pairs for training. Some of these works follow a *train/test* setting, where the network is fine-tuned on a separate set of training images. Note that in our work, we train a much smaller network from scratch instead of fine-tuning a large network. Some approaches (including ours) directly perform *test-time optimization* without additional training data or annotations.

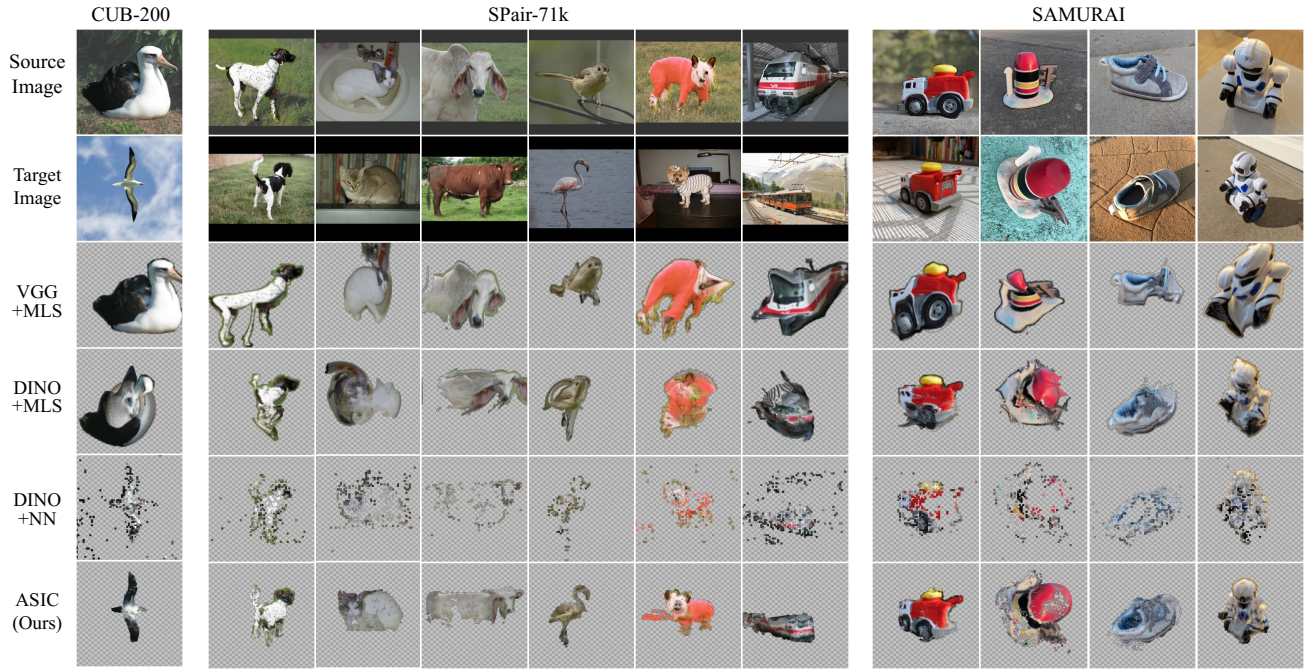

Figure 4: **Dense warping** from a source image (top row) to a target image (second row). We warp all foreground pixels (highlighted by a red overlay in the source image). Our methods produce dense and semantically more meaningful warps from the source to the target.

Table 1: **Evaluation on SPair-71k.** Per-class and average PCK@0.10 on test split. Highest PCK among *weakly supervised* methods in bold, second highest underlined. Scores marked with ($\star$) means the paper uses a fixed image from the test set as canonical image. Our method is competitive against other weak supervised approaches and often outperforms them.

Supervision	Method	Aero	Bike	Bird	Boat	Bottle	Bus	Car	Cat	Chair	Cow	Dog	Horse	Motor	Person	Plant	Sheep	Train	TV	All
Strong Supervision	CATs [13]	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
	PMNC [46]	54.1	35.9	74.9	36.5	42.1	48.8	40.0	72.6	21.1	67.6	58.1	50.5	40.1	54.1	43.3	35.7	74.5	59.9	50.4
	SCorrSAN [29]	57.1	40.3	78.3	38.1	51.8	57.8	47.1	67.9	25.2	71.3	63.9	49.3	45.3	49.8	48.8	40.3	77.7	69.7	55.3
GAN supervision	GANgealing [61]	-	37.5	-	-	-	-	-	67.0	-	-	23.1	-	-	-	-	-	-	57.9	-
Weak supervision (train/test)	CNNGeo [63]	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 [71]	22.6	18.5	42.0	16.4	37.9	**30.8**	26.5	35.6	13.3	29.6	24.3	16.0	21.6	22.8	**20.5**	13.5	31.4	36.5	22.3
	WeakAlign [64]	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
	NCNet [65]	17.9	12.2	32.1	11.7	29.0	19.9	16.1	39.2	9.9	23.9	18.8	15.7	17.4	15.9	14.8	9.6	24.2	31.1	20.1
	SFNet [45]	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
	PMD [48]	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
	PSCNet-SE [35]	28.3	17.7	45.1	15.1	37.5	30.1	27.5	47.4	14.6	32.5	26.4	17.7	24.9	24.5	19.9	16.9	34.2	37.9	27.0
Weak supervision (test-time optimization)	VGG+MLS [1]	29.5	22.7	61.9	**26.5**	20.6	25.4	14.1	23.7	14.2	27.6	30.0	29.1	24.7	27.4	19.1	19.3	24.4	22.6	27.4
	DINO+MLS [1, 10]	49.7	20.9	63.9	19.1	32.5	27.6	22.4	48.9	14.0	36.9	39.0	30.1	21.7	41.1	17.1	18.1	35.9	21.4	31.1
	DINO+NN [4]	57.2	24.1	67.4	24.5	26.8	29.0	27.1	52.1	15.7	42.4	43.3	30.1	23.2	40.7	16.6	24.1	31.0	24.9	33.3
	NeuCongeal [60]	-	**29.1***	-	-	-	-	-	53.3	-	-	35.2	-	-	-	-	-	-	-	-
	ASIC (Ours)	**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**

Table 2: **CUB-200 and PF-Willow.** PCK@0.10 with standard deviations (only for our method) for three CUB categories and four PF-Willow categories.

	CUB-200 (3 categories)		PF-Willow (4 categories)	
Method	PCK@0.1	PCK@0.05	PCK@0.1	PCK@0.05
PMD [48]	-	-	74.7	40.3
PSCNet-SE [35]	-	-	75.1	42.6
VGG+MLS [1]	25.8	18.3	63.2	41.2
DINO+MLS [1, 10]	67.0	52.0	66.5	45.0
DINO+NN [4]	68.3	52.8	60.1	40.1
ASIC (Ours)	**75.9±1.5**	**57.9±1.2**	**76.3±0.8**	**53.0±1.1**

NBB [1] optimizes a flow from one image to another using mutual nearest neighbors as control points [69]. While [1] shows the results by computing nearest neighbors from a VGG network, we further extend their work to utilize a DINO network. [4] simply computes nearest neighbors in DINO feature space. A concurrent work, Neural Congealing [60], is closest to our work, in that they also perform test-time training using a canonical atlas. However, for objects with large deformations (such as in SPair-71k), they need to apply category specific accommodations (for instance, fixing the atlas for bicycle category). Our canonical grid allows for large deformations and is learned in all cases with

a fixed set of hyperparameters. We obtain scores for other models from their respective papers (whenever available) or from [29]. Scores for [1, 4, 10] are computed using official code. The official code of [60] did not converge on several objects in our experiments, hence we report the quantitative results from the paper.

Discussion. Tab. 1 shows PCK@0.1 for all SPair-71k [56] categories. It is evident that having groundtruth keypoint annotations during training is highly beneficial; approaches that lack keypoints during training lag behind. We also observe that for categories with rigid objects (or less extreme deformations) such as 'Bottle' or 'Bus', weakly supervised approaches attain a similar performance as ours. However, in the objects with extreme variations such as animals/birds, our method outperforms other baselines. In our experiments, we observed that per-category hyperparameters can increase PCK performance further by $\sim 2\%$. This strategy is similar to Neural Congealing [60] where specific accommodations are made per category (*e.g.* tailored training regime for bicycle). However we report our numbers with a fixed set of hyperparameters for consistency.

Tab. 2 shows average results for the first 3 categories of the CUB dataset, and 4 categories of the PF-Willow dataset. Note that PF-Willow is an easier dataset compared to SPair-71k since it consists of rigid objects with little variation. Our method has performance similar to PSCNet-SE [35] when we compute PCK using threshold $\alpha_{\text{bbox}} = 0.1$ (which corresponds to a $\sim$20-pixels margin of error). However at higher precision ($\alpha_{\text{bbox}} = 0.05$), our method provides much larger gains compared to the baselines.

Choice of SSL for pseudo-correspondences. In our experiments, we obtain initial set of pseudo-correspondences by finding mutual nearest neighbors from frozen DINO (ViT-S/8) network. Note that DINO is not trained or fine-tuned in our experiments. Our alignment network $\mathbf{\Phi}_{\text{align}}$, which is much smaller than DINO, is trained from scratch. This is also in contrast with other weakly supervised techniques such as PMD which uses ResNet-101 / VGG-16 (> 40M params). We observe that performance of our framework can be improved further by using better pseudo-correspondences. We add the comparison using DINO (ViT-S/8) and DINOv2 (ViT-B/14) on CUB and PF-Pillow in Table 3. DINOv2 is able to improves the correspondence results of both DVD and ASIC in 4 PF-Willow categories, however, the difference is negligible in CUB.

Table 3: **DINO and DINOv2 for pseudo-correspondences.** Performance of our framework can be further improved by using stronger SSL networks for computing pseudo-correspondences.

	NN method	CUB-200 (3 categories)	PF-Willow (4 categories)
DVD [3]	DINO	68.3	54.0
ASIC (ours)	DINO	**75.4**	**76.3**
DVD [3]	DINOv2	67.9	61.8
ASIC (ours)	DINOv2	**75.4**	**78.4**

4.4. Image Set Correspondence

Our goal in this work is to recover dense and *consistent* correspondences. A shortcoming of the PCK metric is that it is only computed between image pairs. However, the errors in keypoint prediction tend to accumulate when transferring keypoints over a sequence of images. To address this limitation, we propose a new metric to measure consistency across multiple images, called k-CyPCK. Given a set of k images $\{\mathbf{I}^1, \mathbf{I}^2, \ldots, \mathbf{I}^k\}$ and an annotated keypoint in the first image $\mathbf{p}^1$ visible in **each** of the k images, we propagate $\mathbf{p}$ from $\mathbf{I}^1 \rightarrow \mathbf{I}^2, \mathbf{I}^2 \rightarrow \mathbf{I}^3, \ldots, \mathbf{I}^{k-1} \rightarrow \mathbf{I}^k, \mathbf{I}^k \rightarrow \mathbf{I}^1$ and get the corresponding predictions $\mathbf{p}^{1\rightarrow 2}, \mathbf{p}^{1\rightarrow 2\rightarrow 3}, \ldots$ and so on. As before, $\mathbf{p}^{1\rightarrow\cdots\rightarrow j}$ is considered to be predicted correctly if it is within a threshold $\alpha_{\text{bbox}} \cdot \max(H_{\text{bbox}}, W_{\text{bbox}})$ of the ground truth keypoints $\hat{\mathbf{p}}^{1\rightarrow\cdots\rightarrow j}$. We sum up all the correct predictions and plot scores at different values of α_{bbox} in Fig. 5. We choose $k = 4$ for all experiments (with additional results for other values of k provided in the supplement). In order to make the metric invariant to the order of images, we choose all possible permutations of k-length sequences in the given image set. Note that this number can become quite large if the image set is large. We recommend using a smaller random subsample in such cases.

Fig. 5 and Tab. 4 show that our method significantly outperforms the DINO+NN baseline for both small and large values of α_{bbox} across all datasets (complete results in the supplemental material). We attribute this result to having a consistent canonical space across the image collection that prevents errors in keypoint transfer from accumulating to large values.

Table 4: Consistent Correspondence Results measured using k-CyPCK for different values of (k, α_{bbox}). Our method produces correspondences which are far from consistent than other works, especially for large values of k.

Dataset	Method	$(\mathbf{2}, \mathbf{0.1})$	$(\mathbf{3}, \mathbf{0.1})$	$(\mathbf{4}, \mathbf{0.1})$	$(\mathbf{2}, \mathbf{0.01})$	$(\mathbf{3}, \mathbf{0.01})$	$(\mathbf{4}, \mathbf{0.01})$
Bicyle	DVD++ [4]	65.3	47.1	38.3	14.6	5.6	2.0
	ASIC (Ours)	92.8	88.9	92.0	59.3	43.3	40.9
Cat	DVD++ [4]	76.3	70.7	70.1	18.4	9.1	8.4
	ASIC (Ours)	84.7	84.3	84.2	54.2	42.7	33.9
Dog	DVD++ [4]	79.2	65.0	66.7	16.8	7.2	8.4
	ASIC (Ours)	91.5	90.0	87.9	47.3	31.5	27.4
TV	DVD++ [4]	52.7	31.6	23.4	8.8	4.8	2.2
	ASIC (Ours)	83.9	73.2	71.9	30.6	15.1	12.1
CUB1	DVD++ [4]	84.3	70.7	63.9	21.8	8.7	7.2
	ASIC (Ours)	94.1	94.4	94.1	58.3	51.5	48.8
CUB2	DVD++ [4]	79.1	67.4	62.4	13.9	8.3	5.6
	ASIC (Ours)	94.6	94.0	94.8	57.6	53.0	44.9
CUB3	DVD++ [4]	80.8	68.6	63.9	17.4	9.4	5.6
	ASIC (Ours)	94.8	94.8	92.9	62.9	56.2	49.7

4.5. Ablations

We perform an ablation study on our various proposed losses proposed, summarized in Tab. 5. We report average

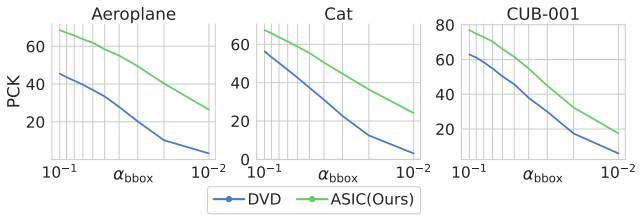

Figure 5: **Image Set Correspondence.** k-CyPCK at varying α_{bbox} (higher is better). Our method outperforms DVD baseline at both large and small values of α_{bbox}.

PCK@0.10 results for first 3 categories of the CUB-200 dataset and all 4 categories of the PF-Willow dataset. As expected, the keypoint loss $\mathcal{L}_{\text{KP}}$ plays the most important role in our overall framework. We also found the total variation regularization $\mathcal{L}_{\text{TV}}$ to be crucial for network training convergence. $\mathcal{L}_{\text{Equi}}$ is necessary for learning dense correspondence. Finally, $\mathcal{L}_{\text{Recon}}$ and $\mathcal{L}_{\text{Parts}}$ provide comparatively small improvements.

Table 5: **Ablation Study.** Average PCK@0.10 (for 3 and 4 categories respectively) in CUB and PF-Willow datasets.

Ablation	CUB-200 (3 categories)	PF-Willow (4 categories)
Complete objective	75.9	76.3
No $\mathcal{L}_{\text{KP}}$	22.8	36.2
No $\mathcal{L}_{\text{TV}}$	43.9	40.4
No $\mathcal{L}_{\text{Equi}}$	64.8	65.6
No $\mathcal{L}_{\text{Recon}}$	73.3	74.2
No $\mathcal{L}_{\text{Parts}}$	73.6	73.5

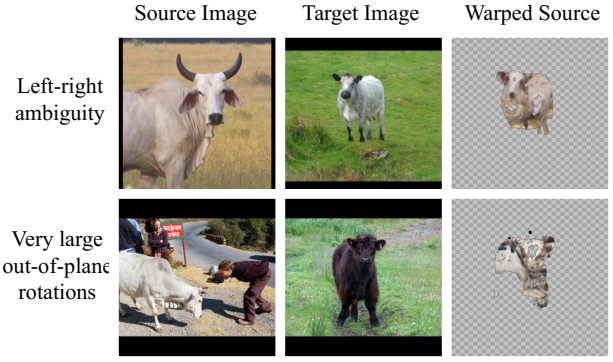

Figure 6: **Limitations.** Top row shows that our model can map left part of object in source to right part of object in target when object is symmetric. Bottom row shows that our model fails for very large out-of-plane rotations.

4.6. Limitations

Left-right ambiguity: One shortcoming of our approach is that it cannot differentiate well between left and right parts well for symmetric objects. We attribute this problem to the SSL models being invariant to left-right flips during their training. The top row of Fig. 6 shows that our model matches the left part of the cow torso in the source image to the right part of the torso in the target image (note left part of cow is not visible in the target image). Some heuristics used in prior works, such as flipping the source and target images and picking a combination that provides minimum total variation loss, could be used in our work as well. For brevity, we provide results without this heuristic.

Large shape changes: Our model doesn't handle large viewpoint changes well, especially when there are few intermediate viewpoints. In the bottom row of Fig. 6, we see that model is unable to warp the source cow image to target image even for the co-visible portions.

5. Conclusions

We propose ASIC, a method to address the challenging task of dense correspondences across images of an object or object category captured in-the-wild. ASIC utilizes noisy and sparse pseudo-correspondences in pre-trained ViT feature space to build an accurate and dense consistent mapping from image to a canonical space. Extensive qualitative and quantitative experiments show that ASIC works in low-shot settings and can deal with extreme variations in pose, background, occlusion, and object deformations. We also propose a new metric k-CyPCK to evaluate the consistency of keypoint predictions over a set of images beyond pair-wise consistency. ASIC can obtain consistent dense mappings competitive with supervised counterparts with just a few images. In future work, we will explore applications of ASIC in other few-shot downstream tasks such as reconstruction, pose estimation and tracking.

Acknowledgements. We thank the reviewers and Area Chairs for their time and valuable feedback. We would also like to thank Archana, Matt and Gowthami for reviewing early drafts of the paper. KG and AS were partially supported by DARPA SAIL-ON (W911NF2020009) and NSF CAREER Award (#2238769) to AS.

References

[1] Kfir Aberman, Jing Liao, Mingyi Shi, Dani Lischinski, Baoquan Chen, and Daniel Cohen-Or. Neural best-buddies: Sparse cross-domain correspondence. *ACM Transactions on Graphics (TOG)*, 37(4):1–14, 2018. 3, 5, 7, 8

[2] Sameer Agarwal, Yasutaka Furukawa, Noah Snavely, Ian Simon, Brian Curless, Steven M Seitz, and Richard Szeliski. Building rome in a day. *Communications of the ACM*, 54(10):105–112, 2011. 2

[3] Shir Amir et al. Deep vit features as dense visual descriptors. *ECCVW What is Motion For?*, 2021. 8

[4] Shir Amir, Yossi Gandelsman, Shai Bagon, and Tali Dekel.

Deep vit features as dense visual descriptors. *arXiv preprint arXiv:2112.05814*, 2(3):4, 2021. 3, 5, 6, 7, 8

[5] Mehmet Aygün and Oisin Mac Aodha. Demystifying unsupervised semantic correspondence estimation. In *European Conference on Computer Vision*, pages 125–142. Springer, 2022. 3

[6] Steven S. Beauchemin and John L. Barron. The computation of optical flow. *ACM computing surveys (CSUR)*, 27(3):433–466, 1995. 2

[7] Irving Biederman. Recognition-by-components: a theory of human image understanding. *Psychological review*, 94(2):115, 1987. 1

[8] Michael J Black and Padmanabhan Anandan. A framework for the robust estimation of optical flow. In *1993 (4th) International Conference on Computer Vision*, pages 231–236. IEEE, 1993. 2

[9] Mark Boss, Andreas Engelhardt, Abhishek Kar, Yuanzhen Li, Deqing Sun, Jonathan T. Barron, Hendrik P.A. Lensch, and Varun Jampani. SAMURAI: Shape And Material from Unconstrained Real-world Arbitrary Image collections. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 2, 5

[10] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *NeurIPS*, 2020. 2, 3, 7, 8

[11] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, pages 1597–1607. PMLR, 2020.

[12] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *CVPR*, 2021. 2

[13] Seokju Cho, Sunghwan Hong, Sangryul Jeon, Yunsung Lee, Kwanghoon Sohn, and Seungryong Kim. Cats: Cost aggregation transformers for visual correspondence. *NeurIPS*, 34:9011–9023, 2021. 1, 7

[14] Subhabrata Choudhury, Iro Laina, Christian Rupprecht, and Andrea Vedaldi. Unsupervised part discovery from contrastive reconstruction. *Advances in Neural Information Processing Systems*, 34:28104–28118, 2021. 3, 5

[15] Christopher B Choy, JunYoung Gwak, Silvio Savarese, and Manmohan Chandraker. Universal correspondence network. *Advances in neural information processing systems*, 29, 2016. 2

[16] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 6

[17] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 224–236, 2018. 2

[18] Jean Duchon. Splines minimizing rotation-invariant seminorms in sobolev spaces. In *Constructive theory of functions of several variables*, pages 85–100. Springer, 1977. 4

[19] Jan-Michael Frahm, Pierre Fite-Georgel, David Gallup, Tim Johnson, Rahul Raguram, Changchang Wu, Yi-Hung Jen, Enrique Dunn, Brian Clipp, Svetlana Lazebnik, et al. Building rome on a cloudless day. In *ECCV*, pages 368–381. Springer, 2010. 2

[20] Dariu M Gavrila. Multi-feature hierarchical template matching using distance transforms. In *Proceedings. Fourteenth international conference on pattern recognition (Cat. No. 98EX170)*, volume 1, pages 439–444. IEEE, 1998. 2

[21] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 2

[22] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. In *NeurIPS*, 2020. 2

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

[24] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, pages 9729–9738, 2020. 2

[25] Sunghwan Hong and Seungryong Kim. Deep matching prior: Test-time optimization for dense correspondence. In *ICCV*, pages 9907–9917, 2021. 3

[26] Berthold KP Horn and Brian G Schunck. Determining optical flow. *Artificial intelligence*, 17(1-3):185–203, 1981. 2

[27] Gary Huang, Marwan Mattar, Honglak Lee, and Erik Learned-Miller. Learning to align from scratch. *Advances in neural information processing systems*, 25, 2012. 2

[28] Gary B Huang, Vidit Jain, and Erik Learned-Miller. Unsupervised joint alignment of complex images. In *2007 IEEE 11th international conference on computer vision*, pages 1–8. IEEE, 2007. 2

[29] Shuaiyi Huang, Luyu Yang, Bo He, Songyang Zhang, Xuming He, and Abhinav Shrivastava. Learning semantic correspondence with sparse annotations. In *ECCV*, 2022. 2, 6, 7, 8

[30] Peter J Huber. Robust estimation of a location parameter. *Breakthroughs in statistics: Methodology and distribution*, pages 492–518, 1992. 5

[31] Wei-Chih Hung, Varun Jampani, Sifei Liu, Pavlo Molchanov, Ming-Hsuan Yang, and Jan Kautz. Scops: Self-supervised co-part segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 869–878, 2019. 3

[32] Sergey Ioffe and David A. Forsyth. Probabilistic methods for finding people. *International Journal of Computer Vision*, 43(1):45–68, 2001. 2

[33] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. *Advances in neural information processing systems*, 28, 2015. 3, 5

[34] 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

[35] Sangryul Jeon, Seungryong Kim, Dongbo Min, and Kwanghoon Sohn. Pyramidal semantic correspondence networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):9102–9118, 2021. 3, 7, 8

[36] 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. 2

[37] Angjoo Kanazawa, Shubham Tulsiani, Alexei A. Efros, and Jitendra Malik. Learning category-specific mesh reconstruction from image collections. In *ECCV*, 2018. 3

[38] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems*, 34:852–863, 2021. 2

[39] Yoni Kasten, Dolev Ofri, Oliver Wang, and Tali Dekel. Layered neural atlases for consistent video editing. *ACM Transactions on Graphics (TOG)*, 40(6):1–12, 2021. 2

[40] Seungryong Kim, Dongbo Min, Bumsub Ham, Sangryul Jeon, Stephen Lin, and Kwanghoon Sohn. Fcss: Fully convolutional self-similarity for dense semantic correspondence. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6560–6569, 2017. 2

[41] 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. 3

[42] Nilesh Kulkarni, Abhinav Gupta, David F Fouhey, and Shubham Tulsiani. Articulation-aware canonical surface mapping. In *CVPR*, pages 452–461, 2020. 1, 3

[43] Nilesh Kulkarni, Abhinav Gupta, and Shubham Tulsiani. Canonical surface mapping via geometric cycle consistency. *ICCV*, 2019. 3

[44] Erik G Learned-Miller. Data driven image models through continuous joint alignment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(2):236–250, 2005. 2

[45] 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. 7

[46] 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. 2, 7

[47] 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.

[48] 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. 2, 7

[49] Ce Liu, Jenny Yuen, and Antonio Torralba. Sift flow: Dense correspondence across scenes and its applications. *IEEE TPAMI*, 33(5):978–994, 2010. 2

[50] Yanbin Liu, Linchao Zhu, Makoto Yamada, and Yi Yang. Semantic correspondence as an optimal transport problem. In *CVPR*, pages 4463–4472, 2020. 2

[51] G Lowe. Sift-the scale invariant feature transform. *Int. J*, 2(91-110):2, 2004. 2

[52] Bruce D Lucas, Takeo Kanade, et al. *An iterative image registration technique with an application to stereo vision*, volume 81. Vancouver, 1981. 2

[53] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 2

[54] Juhong Min and Minsu Cho. Convolutional hough matching networks. In *CVPR*, pages 2940–2950, 2021. 2

[55] Juhong Min, Jongmin Lee, Jean Ponce, and Minsu Cho. Hyperpixel flow: Semantic correspondence with multi-layer neural features. In *ICCV*, pages 3395–3404, 2019. 1, 2

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

[57] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *CVPR*, pages 6707–6717, 2020. 2

[58] Jiteng Mu, Shalini De Mello, Zhiding Yu, Nuno Vasconcelos, Xiaolong Wang, Jan Kautz, and Sifei Liu. Coordgan: Self-supervised dense correspondences emerge from gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10011–10020, 2022. 2

[59] David Novotny, Samuel Albanie, Diane Larlus, and Andrea Vedaldi. Self-supervised learning of geometrically stable features through probabilistic introspection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3637–3645, 2018. 3

[60] Dolev Ofri-Amar, Michal Geyer, Yoni Kasten, and Tali Dekel. Neural congealing: Aligning images to a joint semantic atlas. In *CVPR*, 2023. 3, 7, 8

[61] William Peebles, Jun-Yan Zhu, Richard Zhang, Antonio Torralba, Alexei A Efros, and Eli Shechtman. Gan-supervised dense visual alignment. In *CVPR*, pages 13470–13481, 2022. 2, 6, 7

[62] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR, 2021. 2

[63] Ignacio Rocco, Relja Arandjelovic, and Josef Sivic. Convolutional neural network architecture for geometric matching. In *CVPR*, pages 6148–6157, 2017. 7

[64] Ignacio Rocco, Relja Arandjelović, and Josef Sivic. End-to-end weakly-supervised semantic alignment. In *CVPR*, pages 6917–6925, 2018. 3, 7

[65] Ignacio Rocco, Mircea Cimpoi, Relja Arandjelović, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Neighbourhood consensus networks. *NeurIPS*, 31, 2018. 2, 7

[66] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 3

[67] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. " grabcut" interactive foreground extraction using iterated graph cuts. *ACM transactions on graphics (TOG)*, 23(3):309–

314, 2004. 5

[68] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *CVPR*, pages 4938–4947, 2020. 2

[69] Scott Schaefer, Travis McPhail, and Joe Warren. Image deformation using moving least squares. In *ACM SIGGRAPH 2006 Papers*, pages 533–540, 2006. 5, 7

[70] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, pages 4104–4113, 2016. 2

[71] Paul Hongsuck Seo, Jongmin Lee, Deunsol Jung, Bohyung Han, and Minsu Cho. Attentive semantic alignment with offset-aware correlation kernels. In *ECCV*, pages 349–364, 2018. 3, 7

[72] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 3

[73] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8922–8931, 2021. 2

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

[75] 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

[76] 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

[77] Robert L Thorndike. Who belongs in the family. In *Psychometrika*. Citeseer, 1953. 5

[78] Carlo Tomasi and Takeo Kanade. Detection and tracking of point. *Int J Comput Vis*, 9:137–154, 1991. 2

[79] 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. 3

[80] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9446–9454, 2018. 2

[81] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. 2, 5

[82] Yi Yang and Deva Ramanan. Articulated human detection with flexible mixtures of parts. *IEEE transactions on pattern analysis and machine intelligence*, 35(12):2878–2890, 2012. 6

[83] Chun-Han Yao, Wei-Chih Hung, Yuanzhen Li, Michael Rubinstein, Ming-Hsuan Yang, and Varun Jampani. Lassie: Learning articulated shapes from sparse image ensemble via 3d part discovery. 2022. 3

[84] Zinan Zeng et al. Finding correspondence from multiple images via sparse and low-rank decomposition. In *ECCV 2012*. 2

[85] Junyi Zhang, Charles Herrmann, Junhwa Hur, Luisa Polania Cabrera, Varun Jampani, Deqing Sun, and Ming-Hsuan Yang. A tale of two features: Stable diffusion complements dino for zero-shot semantic correspondence. *arXiv preprint arXiv:2305.15347*, 2023. 3

[86] Jason Zhang, Gengshan Yang, Shubham Tulsiani, and Deva Ramanan. Ners: Neural reflectance surfaces for sparse-view 3d reconstruction in the wild. *Advances in Neural Information Processing Systems*, 34:29835–29847, 2021. 1

[87] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pages 586–595, 2018. 5

[88] 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. 2