Neural Congealing: Aligning Images to a Joint Semantic Atlas

Dolev Ofri-Amar[1] Michal Geyer[1] Yoni Kasten[2] Tali Dekel[1]

[1]Weizmann Institute of Science [2]NVIDIA Research

Figure 1. Given a set of input images, our method automatically detects and jointly aligns semantically-common content across the images. This is achieved through a test-time training approach that estimates a unified 2D atlas that represents the common semantic content, and dense mappings from the joint atlas to each of the input images. Our atlas and mappings are optimized per input set in a self-supervised manner by leveraging a pre-trained DINO-ViT model. Our method can be applied to diverse image sets, without requiring any additional training data, and allows us to automatically propagate an edit applied to a single image across the entire set.

Abstract

We present Neural Congealing – a zero-shot self-supervised framework for detecting and jointly aligning semantically-common content across a given set of images. Our approach harnesses the power of pre-trained DINO-ViT features to learn: (i) a joint semantic atlas – a 2D grid that captures the mode of DINO-ViT features in the input set, and (ii) dense mappings from the unified atlas to each of the input images. We derive a new robust self-supervised framework that optimizes the atlas representation and mappings per image set, requiring only a few real-world images as input without any additional input information (e.g., segmentation masks). Notably, we design our losses and training paradigm to account only for the shared content under severe variations in appearance, pose, background clutter or other distracting objects. We demonstrate results on a plethora of challenging image sets including sets of mixed domains (e.g., aligning images depicting sculpture and artwork of cats), sets depicting related yet different object categories (e.g., dogs and tigers), or domains for which large-scale training data is scarce (e.g., coffee mugs). We thoroughly evaluate our method and show that our test-time optimization approach performs favorably compared to a state-of-the-art method that requires extensive training on large-scale datasets. Project webpage: https://neural-congealing.github.io/

1. Introduction

Humans can easily associate and match semantically-related objects across images, even under severe variations in appearance, pose and background content. For example, by observing the images in Fig. 1, we can immediately focus and visually compare the different butterflies, while ignoring the rest of the irrelevant content. While computational methods for establishing *semantic* correspondences have seen a significant progress in recent years, research efforts are largely focused on either estimating *sparse* matching across *multiple* images (e.g., keypoint detection), or establishing *dense* correspondences between a *pair of images*. In this paper, we consider the task of joint *dense semantic alignment of multiple images*. Solving this long-standing task is useful for a variety of applications, ranging from editing image collections [36,56], browsing images through canonical primitives, and 3D reconstruction (e.g., [8,47]).

The task of joint image alignment dates back to the seminal congealing [13, 14, 20, 28], which aligns a set of images into a common 2D space. Recently, GANgealing [36] has modernized this approach for congealing an entire *domain* of images. This is achieved by leveraging a pre-trained GAN to generate images that serve as self-supervisory signal. Specifically, their method jointly learns both the mode of the generated images in the latent space of the GAN, and a network that predicts the mappings of the images into the joint mode. GANgealing demonstrated impressive results on in-the-wild image sets. Nevertheless, their method requires a StyleGAN model pre-trained on the domain of the test images, e.g., aligning cat images requires training StyleGAN on a large-scale cat dataset. This is a challenging task by itself, especially for unstructured image domains or uncurated datasets [34]. Moreover, they require additional extensive training for learning the mode and their mapping network (e.g., training on millions of generated images). In this work, we take a different route and tackle the joint alignment task in the challenging setting where only a test image set is available, without any additional training data. More specifically, given only a few images as input (e.g., $<$25 images), our method estimates the mode of the test set and their joint dense alignment, in a self-supervised manner.

We assume the input images share a common semantic content, yet may depict various factors of variations, such as pose, appearance, background content or other distracting objects (e.g., *Mugs* in Fig. 3). We take inspiration from the tremendous progress in representation learning, and leverage a pre-trained DINO-ViT – a Vision Transformer model trained in a self-supervised manner [4]. DINO-ViT features have been shown to serve as an effective visual descriptor, capturing localized and semantic information (e.g., [2, 49]). Here, we propose a *new self-supervised framework that jointly and densely aligns the images in DINO-ViT feature space*. To the best of our knowledge, we are the first to harness the power of DINO-ViT for *dense* correspondences between in-the-wild images. More specifically, given an image set, our framework estimates, at test-time: (i) a joint latent 2D atlas that represents the mode of DINO-ViT features across the images, and (ii) dense mappings from the atlas to each of the images. Our training objective is driven by a matching loss encouraging each image features to match the canonical learned features in the joint atlas. We further incorporate additional loss terms that allow our framework to robustly represent and align only the shared content in the presence of background clutter or other distracting objects.

Since our atlas and mappings are optimized per set, our method works in a zero-shot manner and can be applied to a plethora of image sets, including sets of mixed domains (e.g., aligning images depicting sculpture and artwork of cats), sets depicting related yet different object categories (e.g., dogs and tigers), or domains for which a dedicated generator is not available (e.g., coffee mugs). We thoroughly evaluate our method, and demonstrate that our test-time optimization framework performs favorably compared to [36] and on-par with state-of-the-art self-supervised methods. We further demonstrate how our atlas and mappings can be used for editing the image set with minimal effort by automatically propagating edits that are applied to a single image to the entire image set.

2. Related Work

Joint Image Set Alignment. Congealing [14, 20, 28] introduced the task of jointly aligning images into a common 2D image, representing the geometric mode of the set. This was done by minimizing the entropy of intensity values in a pixel stack after the alignment. They demonstrated the use of their method for several applications, including image classification given a few labels per class. These seminal works were extended by incorporating deep learning, combining unsupervised alignment with unsupervised feature learning [13], which showed improvement in face verification accuracy. Further, various methods were proposed to generalize congealing to several modes, e.g., through clustering [10, 11, 21, 26, 35, 51], and make it more robust to occlusions [37]. Other methods are based on pairwise optical flow, requiring either consistent matching across image pairs in a collection [42, 55], or factorizing the collection into simpler subspaces, simplifying the matching task [19, 33]. AverageExplorer [56] presented a user interactive framework for browsing Internet photo collections through average images representing modes in the collections.

Recently, GANgealing [36] used a Spatial Transformer Network (STN) [16] to predict a transformation from any image sampled from a predefined domain (e.g. cat images) into a shared aligned space. They leverage the style-pose disentanglement in a pre-trained StyleGAN2 model [17] to provide training supervision. Specifically, during training, their method simultaneously learns the mode of object pose across a large collection of images, and trains the STN to map each image to this mode. Their method demonstrates impressive results on complex data such as LSUN [54], yet requires extensive compute and large-scale training data. In contrast, we take the congealing task to the realm of test-time optimization, where only a small test image set is available (e.g., $<$25 images). Thus, our method can be applied on diverse domains, or sets that comprise of images from mixed yet related domains, as shown in Fig. 1, 3, 6 and 7.

Semantic Correspondences. Prior to the deep-learning era, various methods were proposed for the task of establishing sparse point correspondences between an image pair [23, 48]. However, due to their local nature and lack of global context, they cannot handle significant color and shape variations. Later, data-driven based descriptors opened the door to establishing correspondences based on higher level information by learning from data representations that encode semantic global information; such methods either work by extracting features from a pre-trained classification model (e.g., [1, 9, 46]), or by training a model end-to-end for establishing semantic correspondences (e.g., [40, 41]). Many of them use groundtruth supervision for training [5], while others aim to tackle the task in a weakly-supervised [40,41] or unsupervised fashion [2]. Our method also aims to learn semantic correspondences, however, we focus on aligning multiple images jointly by leveraging descriptors extracted from a pre-trained DINO-ViT model [4].

DINO-ViT Features as Local Semantic Descriptors. Recent works showed the power of ViT (Vision Transformer) features as local and global semantic descriptors, specifically features of a pre-trained DINO-ViT [4]. Several works [4, 45, 53] showed the use of these features for various applications such as instance segmentation, object discovery or transfer learning on downstream tasks. GCD [50] categorize all unlabeled images given only a partially labeled dataset. Other works used the pre-trained features for both object localization and segmentation [27], and even object part discovery and segmentation [7]. STEGO [12] showed a method for unsupervised semantic segmentation, which unlike most works that learn a feature per pixel, they use contrastive loss that produces low rank representation for the DINO features.

It was shown by [2] that the deep features of a pre-trained DINO-ViT encode semantic information at fine spatial granularity and capture semantic object parts. Furthermore, they showed that this information is shared across domains of semantically-related categories, e.g. cats and dogs, allowing them to design dense descriptors which are used for various applications of co-segmentation, part co-segmentation and point correspondences between images. Splice [49] established a method for semantic appearance transfer using DINO-ViT features, and presented additional powerful properties such as keys inversion, which shows the amount of details they hold of the original image.

We build upon the findings from [2, 49] and use DINO-ViT's spatial features as dense descriptors for aligning images from semantically-related categories.

3. Neural Atlas Congealing

The input to our method is a collection of images with common semantic content, e.g., a set of images depicting different types of guitars in natural scenes. The shared content may significantly differ across the images in appearance, structure, pose, and appear in cluttered scenes containing complex backgrounds or other distracting objects (e.g., Fig. 3). Our goal is to automatically detect the common content across the images and estimate a geometric transformation that maps each image into a joint 2D space. Our key idea is to harness the power of deep features extracted from a pre-trained (and fixed) DINO-ViT model, which have been shown to capture localized semantic information under significant appearance and pose variations [2, 4, 49].

Specifically, our framework, illustrated in Fig. 2, jointly aligns the images in DINO-ViT feature space using two learnable components: (i) a unified latent 2D atlas that represents the common semantic mode in DINO-ViT space across the images, and (ii) a Spatial Transformer Network (STN) that aligns each of the input images to the joint latent atlas. Our method is fully automatic and self-supervised: each of the input images is first fed into DINO-ViT and features (keys) are extracted from the last layer, and serve as our spatial semantic descriptors. The atlas representations, in which each pixel stores a latent feature, and the STN parameters are then optimized such that the transformed features of each image are aligned with the joint atlas features.

We further define a saliency value in each atlas pixel that takes a continuous value between zero and one. We optimize the atlas saliency using a voting-based loss w.r.t. rough initial image saliency masks estimated from DINO-ViT features in a pre-processing step [2]. This allows us to robustly align only the common regions in highly cluttered scenes.

3.1. Semantic Joint Image Alignment

Given an input image set $\{I_i\}_{i=1}^{N}$ and a pre-trained (and fixed) DINO-ViT model, we extract for each image its keys features from the last layer, denoted by $K_i \in \mathbb{R}^{H \times W \times D}$; an initial per-image saliency mask S_i is estimated by applying a simple clustering-based method directly to the extracted features (see [2]). These maps roughly capture salient foreground regions, but are often noisy and contain uncommon regions across the images, as seen in Fig. 4.

We define a learnable atlas $\mathcal{A}$ as a 2D grid of latent features $K_{\mathcal{A}} \in \mathbb{R}^{H_{\mathcal{A}} \times W_{\mathcal{A}} \times D}$, and a saliency mask $S_{\mathcal{A}} \in \mathbb{R}^{H_{\mathcal{A}} \times W_{\mathcal{A}}}$. We define the 2D mapping of each atlas point $\mathbf{x}_{\mathcal{A}}$ to each of the input images I_i as follows:

$$\mathbf{x}_i = \mathcal{M}(I_i, \mathbf{x}_{\mathcal{A}}) \tag{1}$$

where $\mathbf{x}_i$ is the estimated corresponding point of $\mathbf{x}_{\mathcal{A}}$ in image I_i. Note that applying $\mathcal{M}$ on all atlas coordinates allows us to backward warp I_i, S_i or K_i into the atlas space.

Similarly to [36], $\mathcal{M}$ is modeled as a composition of rigid and non-rigid transformations $\mathcal{M}_r \circ \mathcal{M}_f$, each is estimated by a separate STN. That is, $\mathcal{M}_r$ is a global 2D similarity transformation defined by:

$$\mathcal{M}_r(I, \mathbf{x}) = sR\mathbf{x} + \mathbf{t} \tag{2}$$

where $R \in SO(2)$, $\mathbf{t} \in \mathbb{R}^2$ and $s \in \mathbb{R}^+$ are 2D rotation, translation and global scale, respectively. Our method also supports horizontal flips, see the Supplementary Materials (SM) for details.

Given an image I, the non-rigid transformation $\mathcal{M}_f$ is defined by a dense flow field:

$$\mathcal{M}_f(I, \mathbf{x}) = \mathbf{x} + \mathbf{w} \tag{3}$$

where $\mathbf{w}$ is a per-pixel 2D offset.

In practice, each transformation is obtained by an STN model: one takes I_i as input and predicts the parameters of the similarity transformation $\mathcal{M}_r$; the image I_i is then backward warped using $\mathcal{M}_r$ and fed to the second STN that predicts $\mathcal{M}_f$. See SM for network architecture details.

3.2. Training

Given $\{I_i, S_i, K_i\}_{i=1}^{N}$, we now turn to the task of learning the joint atlas $\mathcal{A} = (K_{\mathcal{A}}, S_{\mathcal{A}})$ and 2D mappings $\mathcal{M} = (\mathcal{M}_r, \mathcal{M}_f)$. Our objective function incorporates four main loss terms and takes the following form:

$$\mathcal{L} = \mathcal{L}_{keys} + \lambda_s \mathcal{L}_{saliency} + \lambda_r \mathcal{L}_{reg_{\mathcal{M}}} + \lambda_a \mathcal{L}_{reg_{\mathcal{A}}} \tag{4}$$

where λ_s, λ_a and λ_r control the relative weights between the terms, and are fixed throughout training.

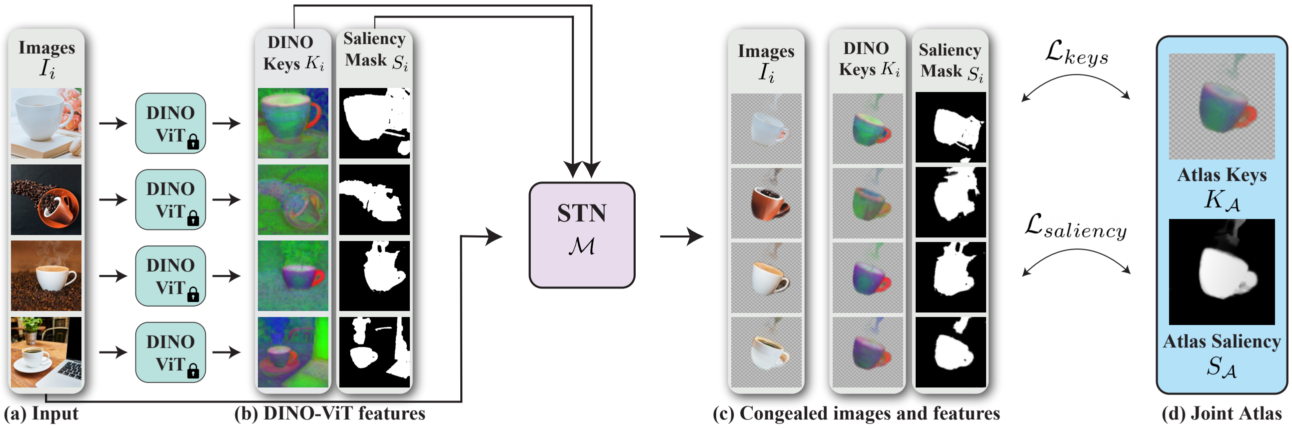

Figure 2. *Neural Congealing.* Each of the input images I_i in the set (a) is fed into a pre-trained and fixed DINO-ViT model. (b) Spatial features, K_i, are extracted and processed to estimate initial rough saliency masks S_i (see [2]), both of which are used to train: (i) a mapping network (STN), which (c) transforms each of the input images into a joint 2D space, and (ii) (d) a joint learnable 2D feature atlas, $K_{\mathcal{A}}$, and an atlas saliency map $S_{\mathcal{A}}$. Note that while the input saliency masks are noisy and contain objects from the background, our method is able to refine them and focus on the most common object in the set. See SM for the full set of images.

Semantic loss $\mathcal{L}_{\mathbf{keys}}$. The semantic loss is our driving loss and it encourages an alignment in DINO-ViT feature space between the atlas and each of the congealed images:

$$\mathcal{L}_{keys} = \sum_{i=1}^{N} \frac{1}{N \cdot \Sigma_{S_{\mathcal{A}}}} \sum_{\mathbf{x}_{\mathcal{A}}} S_{\mathcal{A}}(\mathbf{x}_{\mathcal{A}}) \cdot \mathcal{D}\left(K_i(\mathbf{x}_i), K_{\mathcal{A}}(\mathbf{x}_{\mathcal{A}})\right) \tag{5}$$

where $\Sigma_{S_{\mathcal{A}}} = \sum_{\mathbf{x}_{\mathcal{A}}} S_{\mathcal{A}}(\mathbf{x}_{\mathcal{A}})$, $K_{\mathcal{A}}(\mathbf{x}_{\mathcal{A}})$ is the atlas feature at location $\mathbf{x}_{\mathcal{A}}$, and $K_i(\mathbf{x}_i)$ is the corresponding feature in I_i (see Eq. 1). The distance metric $\mathcal{D}$ is defined by: $\lambda_l \left\| K_i(\mathbf{x}_i) - K_{\mathcal{A}}(\mathbf{x}_{\mathcal{A}}) \right\|_2^2 + \mathcal{D}_{cos}\left(K_i(\mathbf{x}_i), K_{\mathcal{A}}(\mathbf{x}_{\mathcal{A}})\right)$, where $\mathcal{D}_{cos}$ is the cosine distance. Note that since we are interested in aligning only the semantically common content across the images, our loss is weighted according to the atlas saliency $S_{\mathcal{A}}$.

Saliency loss $\mathcal{L}_{\mathbf{saliency}}$. This term serves as a saliency-voting loss that allows us to capture the common content across the images by the atlas saliency $S_{\mathcal{A}}$. Formally, the atlas saliency is optimized to match the initial congealed images' saliency masks. Since the image saliency masks are often rough and contain clutter or irrelevant salient objects of the scene (Fig. 2), we use a robust loss:

$$\mathcal{L}_{saliency} = \frac{1}{N \cdot N_{\mathcal{A}}} \sum_{i=1}^{N} \sum_{\mathbf{x}_{\mathcal{A}}} \rho_{0.7}\left(S_i(\mathbf{x}_i), S_{\mathcal{A}}(\mathbf{x}_{\mathcal{A}})\right) \tag{6}$$

where $N_{\mathcal{A}}$ is the number of pixels in the atlas, and $\rho_\delta(a, b)$ denotes the Huber loss [15] with parameter δ:

$$\rho_\delta(a, b) = \begin{cases} \frac{1}{2}(a-b)^2, & \text{if } |a-b| < \delta \\ \delta \cdot \left(|a-b| - \frac{1}{2}\delta\right) & \text{otherwise} \end{cases} \tag{7}$$

Intuitively, each image "votes" for the regions that should be salient in the atlas, and the aggregated common salient content is estimated.

Transformation regularization $\mathcal{L}_{\mathbf{reg}_{\mathcal{M}}}$. For obtaining a shared representation that is as undistorted as possible, while containing some distortions for aligning objects with different proportions of semantic parts, we apply regularization on both mapping networks:

$$\mathcal{L}_{reg_{\mathcal{M}}} = \lambda_{s_1}\mathcal{L}_{scale} + \lambda_{s_2}\mathcal{L}_{mag} + \mathcal{L}_{smooth} \tag{8}$$

where λ_{s_1} and λ_{s_2} are the relative weights.

$\mathcal{L}_{scale}$ regularizes $\mathcal{M}_r$ from changing the scale of the original images in the atlas space:

$$\mathcal{L}_{scale} = \frac{1}{N} \sum_{i=1}^{N} |1 - s_i|^2 \tag{9}$$

where s_i is the scale parameter of the learned rigid transformation for image I_i.

The non-rigid transformation is encouraged to be as small as possible:

$$\mathcal{L}_{mag} = \frac{1}{N \cdot N_{\mathcal{A}}} \sum_{i=1}^{N} \sum_{x_{\mathcal{A}}} \|\mathbf{w}_i\|_2^2 \tag{10}$$

where $\mathbf{w}$ is the per-pixel flow vector defined in Eq. (3). The term $\mathcal{L}_{smooth}$, defined as in [18], is used to prevent the non-rigid mapping from distorting the shared content by encouraging as rigid as possible mapping. Formally, this term is defined by:

$$\mathcal{L}_{smooth} = \frac{1}{N \cdot N_{\mathcal{A}}} \sum_{i=1}^{N} \sum_{x_{\mathcal{A}}} \left(\left\| J^T J \right\|_F + \left\| \left(J^T J\right)^{-1} \right\|_F \right) \tag{11}$$

where J is the Jacobian matrix of $\mathcal{M}$ at $\mathbf{x}_{\mathcal{A}}$. See SM for more details.

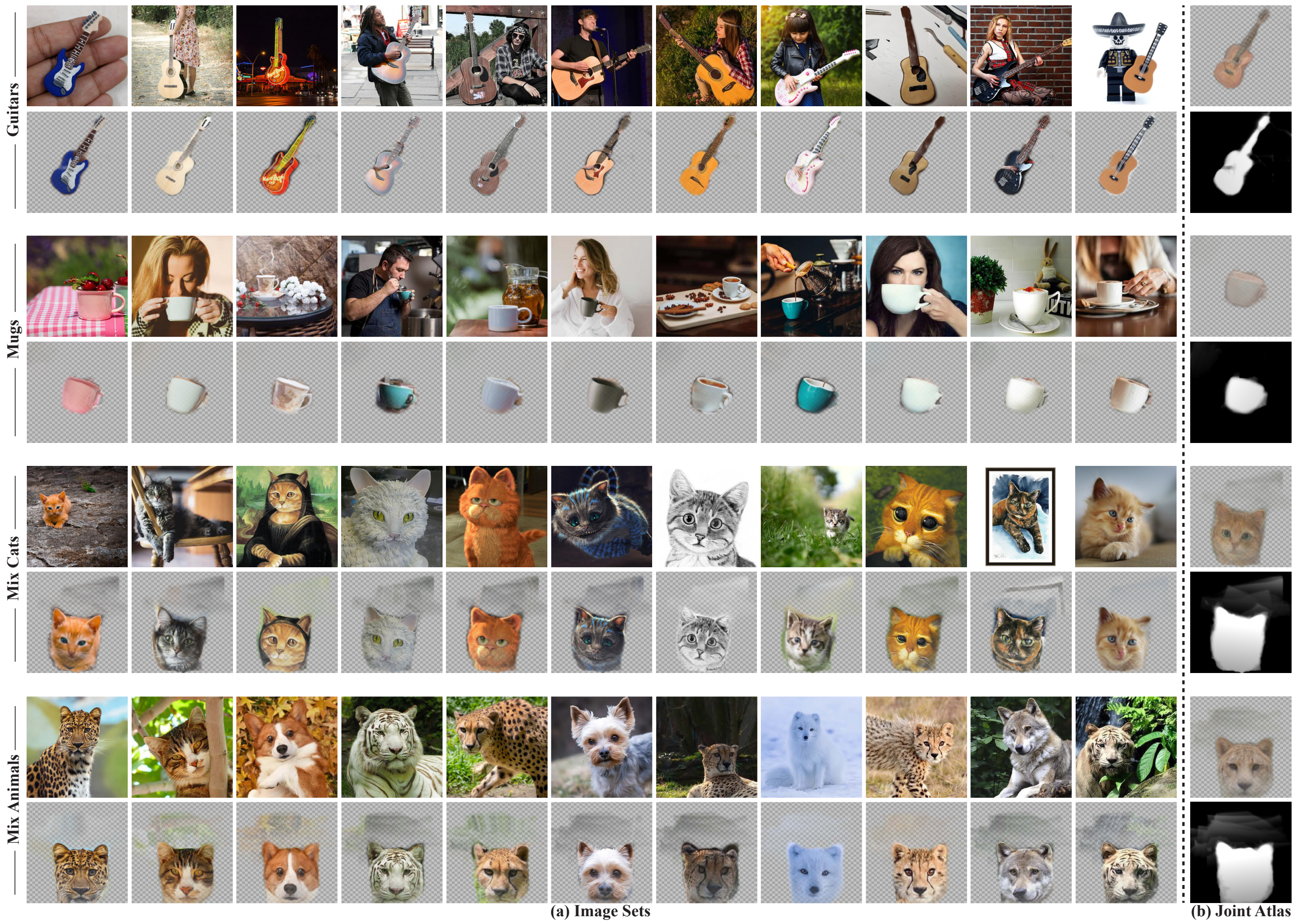

Figure 3. *Neural Congealing results.* We show results on a variety of image sets, each containing images with large variations in pose, appearance, scale, different domains, and cluttered backgrounds. We show (a) the original images together with the congealed images, and (b) the average image in atlas space together with the joint atlas saliency mask. See SM for results on all full sets.

Atlas regularization $\mathcal{L}_{\mathbf{reg}_\mathcal{A}}$. This term controls the localization and sparsity of the atlas:

$$\mathcal{L}_{reg_\mathcal{A}} = \mathcal{L}_{center} + \lambda_p \mathcal{L}_{sparsity} \tag{12}$$

where λ_p is the relative weight.

Since the position of the shared content in the atlas space is arbitrary, we define $\mathcal{L}_{center}$ to encourage the common object to be mapped to the center of the atlas space:

$$\mathcal{L}_{center} = \left\| \frac{1}{\Sigma_{S_\mathcal{A}}} \cdot \sum_{\mathbf{x}_\mathcal{A}} S_\mathcal{A}(\mathbf{x}_\mathcal{A}) \cdot \mathbf{x}_\mathcal{A} \right\|_2^2 \tag{13}$$

where $\mathbf{x}_\mathcal{A}$ coordinates are normalized to be in the range $(-1,1)\times(-1,1)$. By minimizing the norm of the saliency's center of mass, we encourage it to be as close as possible to the atlas center that is located at the origin.

We further observe that without any sparsity regularization, the atlas often contains non-common information. $\mathcal{L}_{sparsity}$ encourages both $S_\mathcal{A}$ and $K_\mathcal{A}$ to be sparse:

$$\mathcal{L}_{sparsity} = \mathcal{L}^{S_\mathcal{A}}_{sparsity} + \lambda_k \mathcal{L}^{K_\mathcal{A}}_{sparsity} \tag{14}$$

where λ_k is the relative weight. We follow [3, 24, 25] and define the sparsity loss term for the atlas saliency as a combination of L1- and L0-approximation regularization terms

$$\mathcal{L}^{S_\mathcal{A}}_{sparsity} = \gamma \left\| S_\mathcal{A} \right\|_1 + \Psi_0(S_\mathcal{A}) \tag{15}$$

where $\Psi_0(x) \equiv 2Sigmoid(5x) - 1$ is a smooth L0 approximation that penalizes non zero elements, and γ is the relative weight between the terms.

For the atlas features, we apply L1 sparsity loss on non-salient parts only:

$$\mathcal{L}^{K_\mathcal{A}}_{sparsity} = (1 - S_\mathcal{A}) \cdot \left\| K_\mathcal{A} \right\|_1 \tag{16}$$

3.3. Editing

Once we have the atlas representation, we can use the average image of all congealed images as a template for editing. Then, the edit in the atlas space is automatically propagated back to all original images. As in [36], given an image I_i and an RGBA edit image in atlas space we apply forward warping using $\mathcal{M}(I_i, \mathbf{x}_\mathcal{A})$ to the image space and then apply alpha blending of the warped edit image with I_i. One can also apply an edit on one of the images and propagate it to all the rest by passing through the atlas space.

4. Results

We tested our method on a variety of image sets, containing 5-25 images, from LSUN [54] and AFHQ [6] datasets, Pixabay [38], Shutterstock [44], and the general Internet. Each set contains objects that share the same semantic parts from different categories including painted/animated/real animals, mugs, guitars, etc. The domain, pose/orientation and appearance, as well as the amount of irrelevant salient objects in the background, may significantly change between the images in each set. We pre-process the images to 256×256 px, with border padding in case of an unsuited aspect-ratio. See SM for full implementation details.

4.1. Qualitative Results.

Sample input sets along with our joint alignment result can be seen in Fig. 1, 2, 3, 6 and 7. Fig. 3 also shows visualizations of the average image in the atlas space and the atlas saliency. The full set of results is included in the SM. As can be seen, our method successfully aligns diverse in-the-wild sets under significant differences in object scale (e.g. *Mugs* in Fig. 3), proportions between semantic parts (e.g., Garfield's ears in *Mix Cats*, or the Corgi's ears in *Mix Animals* in Fig. 3), slight differences in out-of-plane rotation (e.g., white tiger in *Mix Animals* in Fig. 3), and under non-rigid deformations (e.g., the butterflies in Fig. 1).

In addition, Fig. 6 shows an example where our method aligns images across different domains (paintings, food, etc.). This demonstrates the flexibility of our test-time training approach compared to GANgealing [36] which was trained on naturally-looking images, and thus struggles to generalize to other domains.

Edit results. Fig. 1 and 5 show sample edits automatically applied to the input set (Sec. 3.3). As seen, the edits are mapped correctly and accurately to the same semantic regions in all images, under significant variation in scale, pose and appearance. More edit results are included in SM.

Refined masks. As discussed in Sec. 3.1, the initial saliency masks extracted from DINO-ViT features [2] are typically very coarse and may contain irrelevant content such as other objects. Fig. 2 and 4 show examples of how our method manages to congeal these rough estimates into an accurate and refined mask that captures the shared content, while robustly filtering out cluttered background content or non-shared objects. The full set of saliency masks is included in the SM.

4.2. Quantitative Evaluation and Comparison.

We evaluate our framework on the task of semantic point correspondences on SPair-71K [31] and CUB-200-2011 [52]. Specifically, given a source image I_A and a target image I_B together with their ground truth point correspondences, we transfer the points from image I_A to the atlas space and map them back to image I_B to obtain its predicted points (see SM for technical details). We then measure for each set the PCK-Transfer, i.e., the percentage of keypoints that are mapped within the threshold of $\alpha \cdot \max(h, w)$ from the ground truth. We follow previous works and set $\alpha = 0.1$ for both benchmarks, and h, w to be

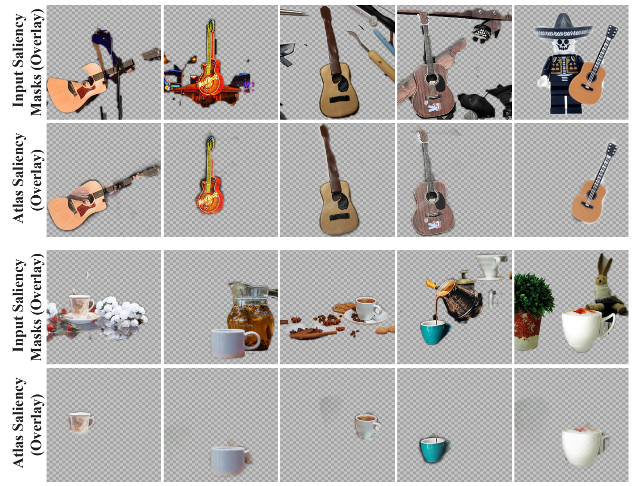

Figure 4. *Refined saliency masks (overlay).* Initial saliency masks extracted using [2], together with our final learned atlas saliency mask. Our method is able to capture the common object, while ignoring cluttered background content or irrelevant objects.

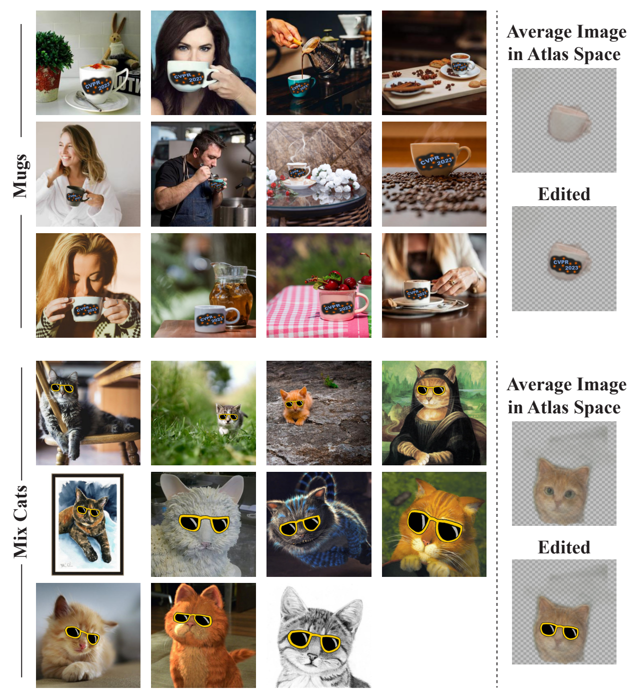

Figure 5. *Sample editing results* for two different sets. Edits are propagated to the same semantic parts in all images. See SM for more results.

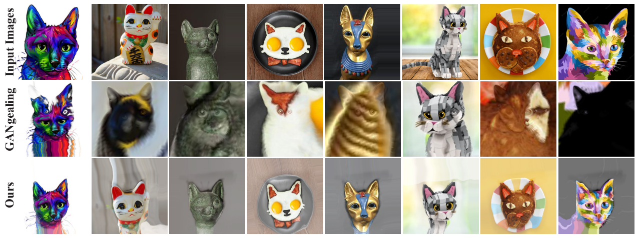

Figure 6. *Across-domain set.* While GANgealing [36] struggle with such images, our method is able to semantically align them.

the dimensions of the object's bounding box for SPair-71K. For CUB, we follow [36] and set h, w to be the image size.

Results on SPair-71K. We use the same pre-processing as in [36], applying border padding for non-square images and resizing to 256×256. We apply our method on each test set separately, each includes 25-26 images. See SM for further technical details.

Table 1 reports the results for our method, GANgealing, and a number of leading methods for semantic correspondences. As seen, our method outperforms GANgealing on most sets, and outperforms other self-supervised methods on all sets. Our method performs very well on the *Cat* and *Dog* sets, yet in the *Bicycle* set, due to the large deformations and the symmetric shape of the object, our performance decreases. Nevertheless, even in this challenging case, our performance is on-par with most supervised methods on this set. Note that all supervised methods have been directly trained or fine-tuned using ground-truth supervision on the SPair-71K training set.

Figure 7 shows qualitative comparisons to [36]; GANgealing converges to a global mode across a large dataset, which allows them to achieve good alignment for *Bicycle* and *Birds* (Fig. 7(c),(d)). However, in the *Dog* set this restricts their alignment to capturing only the head, while our method detects the common mode in a given set and can align the full body of the dogs (Fig. 7(a)).

There is an inherent tradeoff between aligning highly articulated content (e.g., the bodies of the non-rigid animals) and maintaining undistorted atlas representation. We demonstrate this tradeoff for the *Cat* set, by controlling the effective relative weight of our rigidity loss. As seen, with our default parameters, the method aligns the bodies of the cats yet their faces are not accurately aligned. By increasing the relative weight of the rigidity, we can encourage the model to focus on the most rigid part across the set, while disregarding the cats' body (even though it is a shared salient part across the set). In this setting, fine facial details are accurately aligned and our method outperforms all previous methods, including supervised methods.

Results on CUB-200-2011. Since our method works with small image sets, we randomly sample 14 sets of 25 images each, and train separately on each of them. For fair comparison, we apply the same pre-processing as in [36]. As

Method	Supervision	Cat	Dog	Bicycle
HPF [30]	supervised	52.9	32.8	18.9
DHPF [32]	supervised	61.6	46.1	23.8
SCOT [22]	supervised	63.1	42.5	20.7
CHM [29]	supervised	64.9	56.1	29.3
CATs [5]	supervised	66.5	56.5	34.7
WeakAlign [40]	weakly-supervised	31.8	22.6	17.6
NC-Net [41]	weakly-supervised	39.2	18.8	12.2
CNNgeo [39]	self-supervised	32.7	22.8	16.7
A2Net [43]	self-supervised	35.6	24.3	18.5
GANgealing [36]	GAN-supervised	67.0	23.1	**37.5**
Ours	self-supervised	54.5/**70.7***	**35.8**	29.1

Table 1. *PCK-Transfer on categories from SPair-71K with $\alpha = 0.1$.* *Results for relaxed setting, aligning only the cats' heads.

Method	Supervision	PCK@$\alpha = 0.1$
Ours w/o $\mathcal{L}_{keys}$	self-supervised	30.7
Ours w/o saliency masks	self-supervised	53.4
Ours w/o $\mathcal{L}_{reg_\mathcal{M}}$	self-supervised	36.9
Ours w/o $\mathcal{L}_{reg_\mathcal{A}}$	self-supervised	**64.9**
GANgealing [36]	GAN-supervised	56.8
Ours	self-supervised	**63.6**

Table 2. *PCK-Transfer on subsets of CUB-200-2011.* Comparison to [36], together with ablation for our different loss terms.

seen in Table 2, our method achieves better results compared to GANgealing. As seen in Fig. 7(c), GANgealing tends to hallucinate object parts, and struggles with aligning the heads, especially when the object pose differs significantly from the canonical pose learned from the entire domain. Our method, by optimizing the representation and mappings per set, manages to align the heads of the birds even under unusual poses, e.g., distant bird spreading wings (second column from the left).

4.3. Ablation Study

We ablate the different loss terms of our objective function (Eq. 4), both quantitatively in Table 2 and qualitatively in Fig. 8. Without our driving loss $\mathcal{L}_{keys}$, we notice a significant drop in performance. Fig. 8 shows that even though the saliency masks help a great deal in bringing the birds one on top of the other, there is no semantic alignment between them. Without saliency masks (no atlas saliency), our framework attempts to align all observed content, and thus struggles to converge, or converges only to small parts of the object, in cases of significant background clutter.

Removing $\mathcal{L}_{reg_\mathcal{M}}$ provides too much freedom to the non-rigid mapping, which converges either to a single point in the atlas, or spreads in disorder. Thus, the performance drops dramatically and the visual results are not appealing. Our method performs on-par without $\mathcal{L}_{reg_\mathcal{A}}$, however, we note that the initial saliencies for the CUB subsets are quite accurate, thus there is no much need for regularizing the atlas in this case. More generally, this loss allows us to obtain cleaner atlases (see SM). Finally, we see in Fig. 8 that $\mathcal{L}_{center}$ of $\mathcal{L}_{reg_\mathcal{A}}$ encourages the shared content to be centered, allowing us to keep the birds within atlas borders.

Figure 7. *Comparison to GANgealing [36].* We compare results from sets used in evaluation (Sec. 4.2). Since our method is based on alignment according to semantic DINO-ViT features, even in hard examples, our method focuses on aligning the most common salient object part, e.g., the birds in (c). Note that as in [36], our method supports horizontal flips (see SM for technical details).

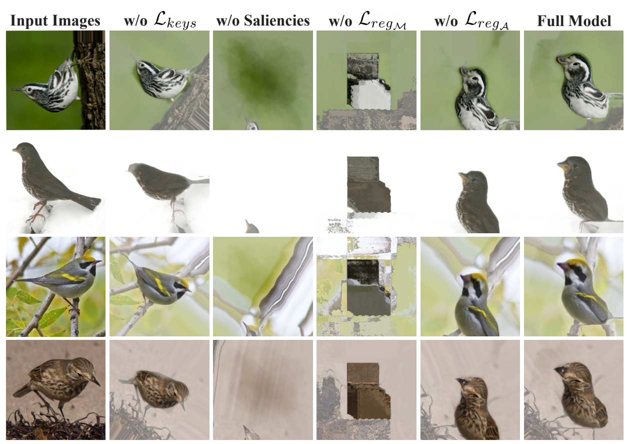

Figure 8. Ablation of our objective function's loss terms (Sec. 4.3).

5. Limitations

Our method relies on semantic similarity in the space of DINO-ViT feature space. Hence, in cases where these features do not capture the semantic association across the images, our method would not work well (e.g., image domains that are not well-represented in DINO's training data). In addition, in cases of extreme topological changes in the common object across the set, our method struggles to converge to a good alignment due to the strong rigidity constraints, e.g., Fig. 7(b) and Fig. 7(c). Furthermore, we notice that in sets containing symmetric objects with large rotation differences, the relative position between parts may affect the convergence, and may lead to partial alignment (e.g., matching the left ear of one image with the right ear of the other). In general, our framework is not designed to align images depicting more than one instance of the shared mode. In this case, our method may align arbitrarily one of the objects, and in other cases may fail to converge. See SM for some examples of failure cases.

6. Conclusions

We tackled the congealing task in a particularity challenging setting – jointly aligning a small set of in-the-wild images, without any additional training data other than the test set itself. We showed how to leverage the power of pre-trained DINO-ViT features for this task in a new test-time training framework. We demonstrated the key advantages of our approach w.r.t. existing state-of-the-art methods in its applicability to diverse image domains, lightweight training and overall performance. We further showed that our method can be used for automatically propagating edits to the entire set by simply editing a single image. We believe that our approach – combining test-time optimization with semantic information learned by external large-scale models – holds great promise for dense alignment tasks, and can motivate future research in this direction.

Acknowledgements: We thank Shai Bagon and Shir Amir for their insightful comments. This project received funding from the Israeli Science Foundation (grant 2303/20).

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)*, 2018. 2

[2] Shir Amir, Yossi Gandelsman, Shai Bagon, and Tali Dekel. Deep vit features as dense visual descriptors. *ECCVW What is Motion For?*, 2022. 2, 3, 4, 6

[3] Omer Bar-Tal, Dolev Ofri-Amar, Rafail Fridman, Yoni Kasten, and Tali Dekel. Text2live: Text-driven layered image and video editing. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XV*, pages 707–723. Springer, 2022. 5

[4] 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 IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021. 2, 3

[5] 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*, 2021. 2, 7

[6] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020. 6

[7] Subhabrata Choudhury, Iro Laina, Christian Rupprecht, and Andrea Vedaldi. Unsupervised part discovery from contrastive reconstruction. *Advances in Neural Information Processing Systems*, 2021. 3

[8] Gabriele Facciolo, Carlo De Franchis, and Enric Meinhardt-Llopis. Automatic 3d reconstruction from multi-date satellite images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 57–66, 2017. 1

[9] Philipp Fischer, Alexey Dosovitskiy, and Thomas Brox. Descriptor matching with convolutional neural networks: a comparison to sift. *arXiv preprint arXiv:1405.5769*, 2014. 2

[10] Brendan J Frey and Nebojsa Jojic. Estimating mixture models of images and inferring spatial transformations using the em algorithm. In *Proceedings. 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No PR00149)*. IEEE, 1999. 2

[11] Brendan J. Frey and Nebojsa Jojic. Transformation-invariant clustering using the em algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2003. 2

[12] Mark Hamilton, Zhoutong Zhang, Bharath Hariharan, Noah Snavely, and William T. Freeman. Unsupervised semantic segmentation by distilling feature correspondences. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*, 2022. 3

[13] Gary Huang, Marwan Mattar, Honglak Lee, and Erik Learned-Miller. Learning to align from scratch. *Advances in neural information processing systems*, 25, 2012. 1, 2

[14] 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. 1, 2

[15] Peter J. Huber. Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics*, 35(1):73 – 101, 1964. 4

[16] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. *Advances in neural information processing systems*, 28, 2015. 2

[17] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 2

[18] 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. 4

[19] Ira Kemelmacher-Shlizerman and Steven M Seitz. Collection flow. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1792–1799. IEEE, 2012. 2

[20] 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. 1, 2

[21] Xiaoming Liu, Yan Tong, and Frederick W Wheeler. Simultaneous alignment and clustering for an image ensemble. In *2009 IEEE 12th International Conference on Computer Vision*. IEEE, 2009. 2

[22] 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*, 2020. 7

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

[24] Erika Lu, Forrester Cole, Tali Dekel, Weidi Xie, Andrew Zisserman, David Salesin, William T. Freeman, and Michael Rubinstein. Layered neural rendering for retiming people in video. *ACM Trans. Graph.*, 2020. 5

[25] Erika Lu, Forrester Cole, Tali Dekel, Andrew Zisserman, William T. Freeman, and Michael Rubinstein. Omnimatte: Associating objects and their effects in video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 5

[26] Marwan A. Mattar, Allen R. Hanson, and Erik G. Learned-Miller. Unsupervised joint alignment and clustering using bayesian nonparametrics. In *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence, Catalina Island, CA, USA, August 14-18, 2012*. AUAI Press, 2012. 2

[27] Luke Melas-Kyriazi, Christian Rupprecht, Iro Laina, and Andrea Vedaldi. Deep spectral methods: A surprisingly strong baseline for unsupervised semantic segmentation and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 3

[28] Erik G Miller, Nicholas E Matsakis, and Paul A Viola. Learning from one example through shared densities on transforms. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No. PR00662)*. IEEE, 2000. 1, 2

[29] Juhong Min and Minsu Cho. Convolutional hough matching networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 7

[30] Juhong Min, Jongmin Lee, Jean Ponce, and Minsu Cho. Hyperpixel flow: Semantic correspondence with multi-layer neural features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019. 7

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

[32] Juhong Min, Jongmin Lee, Jean Ponce, and Minsu Cho. Learning to compose hypercolumns for visual correspondence. In *European Conference on Computer Vision*. Springer, 2020. 7

[33] Hossein Mobahi, Ce Liu, and William T Freeman. A compositional model for low-dimensional image set representation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1322–1329, 2014. 2

[34] Ron Mokady, Omer Tov, Michal Yarom, Oran Lang, Inbar Mosseri, Tali Dekel, Daniel Cohen-Or, and Michal Irani. Self-distilled stylegan: Towards generation from internet photos. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–9, 2022. 2

[35] Tom Monnier, Thibault Groueix, and Mathieu Aubry. Deep transformation-invariant clustering. *Advances in Neural Information Processing Systems*, 2020. 2

[36] 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, 2, 3, 5, 6, 7, 8

[37] Yigang Peng, Arvind Ganesh, John Wright, Wenli Xu, and Yi Ma. Rasl: Robust alignment by sparse and low-rank decomposition for linearly correlated images. *IEEE transactions on pattern analysis and machine intelligence*, 34(11):2233–2246, 2012. 2

[38] Pixabay. https://pixabay.com/. 6

[39] 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*, 2017. 7

[40] 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, 7

[41] Ignacio Rocco, Mircea Cimpoi, Relja Arandjelović, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Neighbourhood consensus networks. *Advances in neural information processing systems*, 2018. 2, 7

[42] Michael Rubinstein, Armand Joulin, Johannes Kopf, and Ce Liu. Unsupervised joint object discovery and segmentation in internet images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1939–1946, 2013. 2

[43] 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)*, 2018. 7

[44] Shutterstock. https://www.shutterstock.com/. 6

[45] Oriane Siméoni, Gilles Puy, Huy V. Vo, Simon Roburin, Spyros Gidaris, Andrei Bursuc, Patrick Pérez, Renaud Marlet, and Jean Ponce. Localizing objects with self-supervised transformers and no labels. In *32nd British Machine Vision Conference 2021, BMVC 2021, Online, November 22-25, 2021*. BMVA Press, 2021. 3

[46] Edgar Simo-Serra, Eduard Trulls, Luis Ferraz, Iasonas Kokkinos, Pascal Fua, and Francesc Moreno-Noguer. Discriminative learning of deep convolutional feature point descriptors. In *Proceedings of the IEEE international conference on computer vision*, pages 118–126, 2015. 2

[47] Xiao Tang, Tianyu Wang, and Chi-Wing Fu. Towards accurate alignment in real-time 3d hand-mesh reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11698–11707, October 2021. 1

[48] Engin Tola, Vincent Lepetit, and Pascal Fua. Daisy: An efficient dense descriptor applied to wide-baseline stereo. *IEEE transactions on pattern analysis and machine intelligence*, 32(5):815–830, 2009. 2

[49] 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. 2, 3

[50] Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Generalized category discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 3

[51] Andrea Vedaldi, Gregorio Guidi, and Stefano Soatto. Joint data alignment up to (lossy) transformations. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2008. 2

[52] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. . Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. 6

[53] 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*, 2022. 3

[54] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015. 2, 6

[55] Tinghui Zhou, Yong Jae Lee, Stella X Yu, and Alyosha A Efros. Flowweb: Joint image set alignment by weaving consistent, pixel-wise correspondences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1191–1200, 2015. 2

[56] Jun-Yan Zhu, Yong Jae Lee, and Alexei A Efros. Averageexplorer: Interactive exploration and alignment of visual data collections. *ACM Transactions on Graphics (TOG)*, 33(4):1–11, 2014. 1, 2