GAN-Supervised Dense Visual Alignment

William Peebles[1] Jun-Yan Zhu[2] Richard Zhang[3] Antonio Torralba[4] Alexei A. Efros[1] Eli Shechtman[3]
[1]UC Berkeley [2]Carnegie Mellon University [3]Adobe Research [4]MIT CSAIL

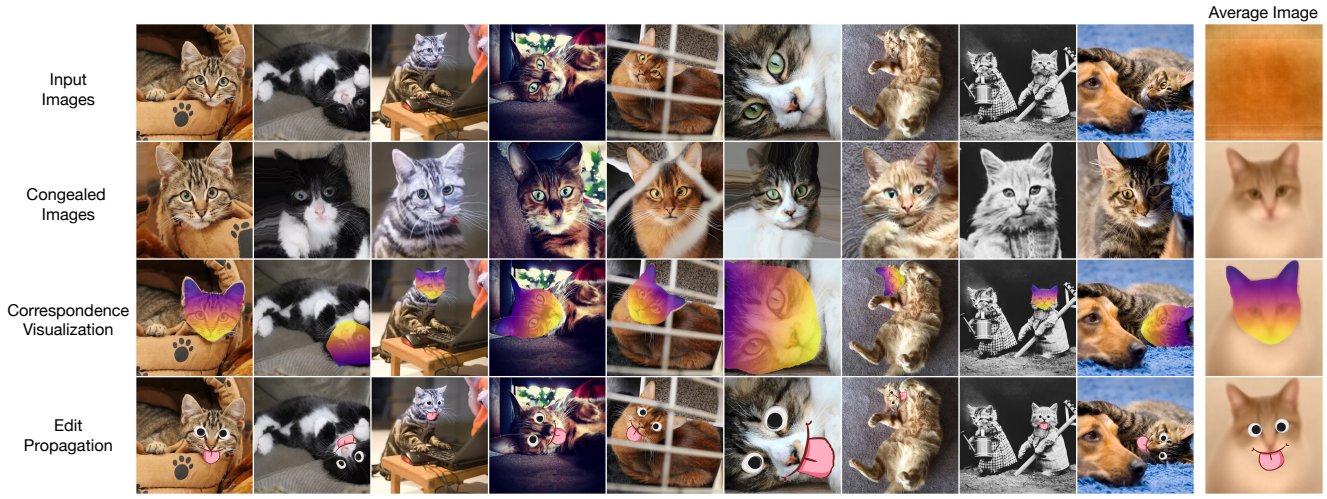

Figure 1. Given an input dataset of unaligned images, our GANgealing algorithm discovers dense correspondences between all images. **Top row**: Images from LSUN Cats and the dataset's average image. **Second row**: Our learned transformations of the input images. **Third row**: Dense correspondences learned by GANgealing. **Bottom row**: By annotating the average transformed image, we can propagate user edits to images and videos. **Please see our project page for detailed video results:** www.wpeebles.com/gangealing.

Abstract

We propose GAN-Supervised Learning, a framework for learning discriminative models and their GAN-generated training data jointly end-to-end. We apply our framework to the dense visual alignment problem. Inspired by the classic Congealing method, our GANgealing algorithm trains a Spatial Transformer to map random samples from a GAN trained on unaligned data to a common, jointly-learned target mode. We show results on eight datasets, all of which demonstrate our method successfully aligns complex data and discovers dense correspondences. GANgealing significantly outperforms past self-supervised correspondence algorithms and performs on-par with (and sometimes exceeds) state-of-the-art supervised correspondence algorithms on several datasets—without making use of any correspondence supervision or data augmentation and despite being trained exclusively on GAN-generated data. For precise correspondence, we improve upon state-of-the-art supervised methods by as much as 3×. We show applications of our method for augmented reality, image editing and automated preprocessing of image datasets for downstream GAN training.

Code and models: www.github.com/wpeebles/gangealing

1. Introduction

Visual alignment, also known as the correspondence or registration problem, is a critical element in much of computer vision, including optical flow, 3D matching, medical imaging, tracking and augmented reality. While much recent progress has been made on pairwise alignment (aligning image A to image B) [2, 14, 22, 34, 51, 57, 58, 60, 68–71, 75], the problem of global joint alignment (aligning *all* images across a dataset) has not received as much attention. Yet, joint alignment is crucial for tasks requiring a common reference frame, such as automatic keypoint annotation, augmented reality or edit propagation (see Figure 1 bottom row). There is also evidence that training on jointly aligned datasets (such as FFHQ [42], AFHQ [15], CelebA-HQ [40]) can produce higher quality generative models than training on unaligned data.

In this paper, we take inspiration from a series of classic works on automatic joint image set alignment. In particular, we are motivated by the seminal unsupervised *Congealing* method of Learned-Miller [48] which showed that a set of images could be brought into alignment by continually warping them toward a common, updating mode. While Congeal-

ing can work surprisingly well on simple binary images, such as MNIST digits, the direct pixel-level alignment is not powerful enough to handle most datasets with significant appearance and pose variation.

To address these limitations, we propose GANgealing: a *GAN-Supervised* algorithm that learns transformations of input images to bring them into better joint alignment. The key is in employing the latent space of a GAN (trained on the unaligned data) to automatically generate paired training data for a Spatial Transformer [35]. Crucially, in our proposed GAN-Supervised Learning framework, *both* the Spatial Transformer and the target images are learned jointly. Our Spatial Transformer is trained *exclusively* with GAN images and generalizes to real images at test time.

We show results spanning eight datasets—LSUN Bicycles, Cats, Cars, Dogs, Horses and TVs [87], In-The-Wild CelebA [52] and CUB [83]—that demonstrate our GANgealing algorithm is able to discover accurate, dense correspondences across datasets. We show our Spatial Transformers are useful in image editing and augmented reality tasks. Quantitatively, GANgealing significantly outperforms past self-supervised dense correspondence methods, nearly doubling key point transfer accuracy (PCK [4]) on many SPair-71K [59] categories. Moreover, GANgealing sometimes matches and even exceeds state-of-the-art correspondence-*supervised* methods.

2. Related Work

Pre-Trained GANs for Vision. Prior work has explored the use of GANs [27, 67] in vision tasks such as classification [10, 12, 55, 74, 84], segmentation [56, 79, 82, 90] and representation learning [7, 20, 21, 23, 36], as well as 3D vision and graphics tasks [28, 64, 72, 89]. Likewise, we share the goal of leveraging the power of pre-trained deep generative models for vision tasks. However, the relevant past methods follow a common two-stage paradigm of (1) synthesizing a GAN-generated dataset and (2) training a discriminative model on the fixed dataset. In contrast, our GAN-Supervised Learning approach learns *both* the discriminative model as well as the GAN-generated data jointly end-to-end. We do not rely on hand-crafted pixel space augmentations [12, 36], human-labeled data [28, 72, 79, 89, 90] or post-processing of GAN-generated datasets using domain knowledge [10, 56, 82, 89].

Joint Image Set Alignment. Average images have long been used to visualize joint alignment of image sets of the same semantic content (e.g., [78, 95]), with the seminal work of Congealing [32, 48] establishing unsupervised joint alignment as a research problem. Congealing uses sequential optimization to gradually minimize the entropy of the intensity distribution of a set of images by continuously warping each image via a parametric transformation (e.g., affine). It produces impressive results on well-structured datasets, such as digits, but struggles with more complex data. Subsequent work in this area assumes the data lies on a low-rank subspace [44, 66] or factorizes images as a composition of color, appearance and shape [62] to establish dense correspondences between instances of the same object category. FlowWeb [92] uses cycle consistency constraints to estimate a fully-connected correspondence flow graph. Every method above assumes that it is possible to align all images to a single central mode in the data. Joint visual alignment and clustering was proposed in AverageExplorer [95] but as a user-driven data interaction tool. Bounding box supervision has been used to align and cluster multiple modes within object categories [19]. Automated transformation-invariant clustering methods [24, 25] can align images in a collection before comparing them but work only in limited domains. Recently, Monnier et al. [63] showed that warps could be predicted with a network instead, removing the need for per-image optimization; this opened the door for simultaneous alignment and clustering of large-scale collections. Unlike our approach, these methods assume images can be aligned with simple (e.g., affine) color transformations; this assumption breaks down for complex datasets like LSUN.

Spatial Transformer Networks (STNs). A Spatial Transformer module [35] is one way to incorporate learnable geometric transformations in a deep learning framework. It regresses a set of warp parameters, where the warp and grid sampling functions are differentiable to enable backpropagation. STNs have seen success in discriminative tasks (e.g., classification) and applications such as robust filter learning [16, 37], view synthesis [26, 65, 93] and 3D representation learning [39, 86, 91]. Inverse Compositional STNs (IC-STNs) [49] advocate an iterative image alignment framework in the spirit of the classical Lukas-Kanade algorithm [6, 54]. Prior work has incorporated STNs in generative models for geometry-texture disentanglement [85] and image compositing [50]. In contrast, we use a generative model to directly produce training data *for* STNs.

3. GAN-Supervised Learning

In this section, we present GAN-Supervised Learning. Under this framework, $(\boldsymbol{x}, \boldsymbol{y})$ pairs are sampled from a pre-trained GAN generator, where $\boldsymbol{x}$ is a random sample from the GAN and $\boldsymbol{y}$ is the sample obtained by applying a *learned* latent manipulation to $\boldsymbol{x}$'s latent code. These pairs are used to train a network $f_\theta : \boldsymbol{x} \rightarrow \boldsymbol{y}$. This framework minimizes the following loss:

$$\mathcal{L}(f_\theta, \boldsymbol{y}) = \ell(f_\theta(\boldsymbol{x}), \boldsymbol{y}), \tag{1}$$

where ℓ is a reconstruction loss. In vanilla supervised learning, f_θ is learned on *fixed* $(\boldsymbol{x}, \boldsymbol{y})$ pairs. In contrast, in GAN-Supervised Learning, *both* f_θ and the targets $\boldsymbol{y}$ are learned jointly end-to-end. At test time, we are free to evaluate f_θ on *real* inputs.

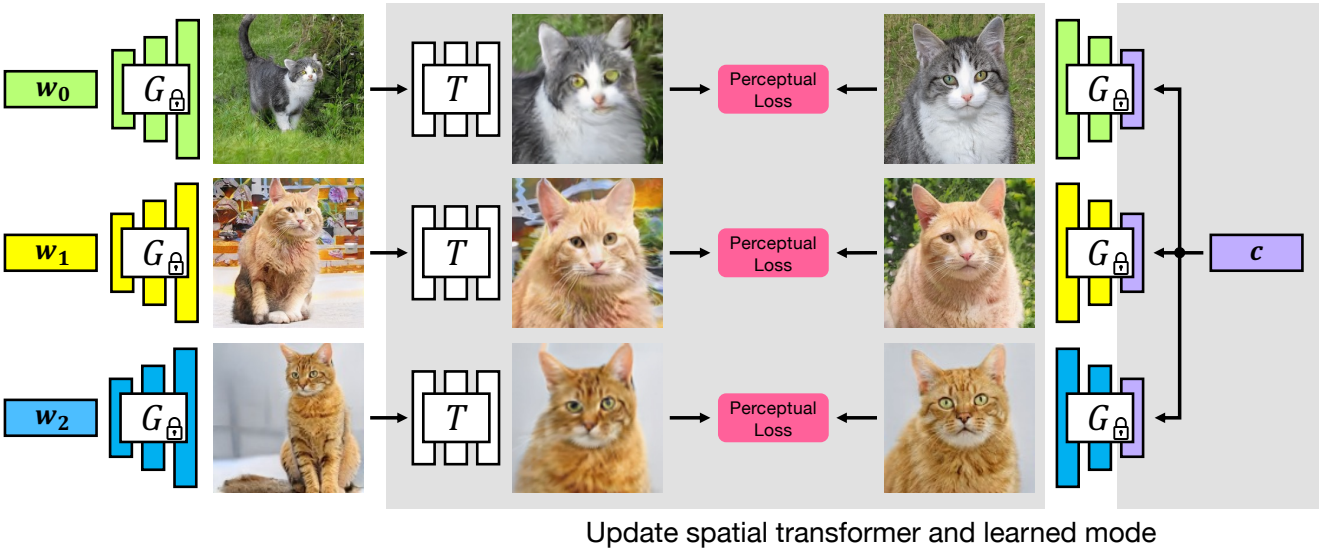

Figure 2. **GANgealing Overview**. We first train a generator G on unaligned data. We create a *synthetically-generated* dataset for alignment by learning a mode $\mathbf{c}$ in the generator's latent space. We use this dataset to train a Spatial Transformer Network T to map from unaligned to corresponding aligned images using a perceptual loss [38]. The Spatial Transformer generalizes to align *real* images automatically.

3.1. Dense Visual Alignment

Here, we show how GAN-Supervised Learning can be applied to Congealing [48]—a classic unsupervised alignment algorithm. In this instantiation, f_θ is a Spatial Transformer Network [35] T, and we describe our parameterization of inputs $\boldsymbol{x}$ and learned targets $\boldsymbol{y}$ below. We call our algorithm *GANgealing*. We present an overview in Figure 2.

GANgealing begins by training a latent variable generative model G on an unaligned input dataset. We refer to the input latent vector to G as $\mathbf{w} \in \mathbb{R}^{512}$. With G trained, we are free to draw samples from the unaligned distribution by computing $\boldsymbol{x} = G(\mathbf{w})$ for randomly sampled $\mathbf{w} \sim \mathcal{W}$, where $\mathcal{W}$ denotes the distribution over latents. Now, consider a fixed latent vector $\mathbf{c} \in \mathbb{R}^{512}$. This vector corresponds to a fixed synthetic image $G(\mathbf{c})$ from the original unaligned distribution. A simple idea in the vein of traditional Congealing is to use $G(\mathbf{c})$ as the target mode $\boldsymbol{y}$—i.e., we learn a Spatial Transformer T that is trained to warp every random unaligned image $\boldsymbol{x} = G(\mathbf{w})$ to the same target image $\boldsymbol{y} = G(\mathbf{c})$. Since G is differentiable in its input, we can optimize $\mathbf{c}$ and hence learn the target we wish to congeal towards. Specifically, we can optimize the following loss with respect to both T's parameters and the target image's latent vector $\mathbf{c}$ jointly:

$$\mathcal{L}_{\text{align}}(T, \mathbf{c}) = \ell(T(G(\mathbf{w})), G(\mathbf{c})), \tag{2}$$

where ℓ is some distance function between two images. By minimizing $\mathcal{L}$ with respect to the target latent vector $\mathbf{c}$, GANgealing encourages $\mathbf{c}$ to find a pose that makes T's job as easy as possible. If the current value of $\mathbf{c}$ corresponds to a pose that cannot be reached from most images via the transformations predicted by T, then it can be adjusted via gradient descent to a different vector that is "reachable" by more images.

This simple approach is reasonable for datasets with limited diversity; however, in the presence of significant appearance and pose variation, it is not reasonable to expect that every unaligned sample can be aligned to the exact same target image. Hence, optimizing the above loss does not produce good results in general (see Table 3). Instead of using the same target $G(\mathbf{c})$ for every randomly sampled image $G(\mathbf{w})$, it would be ideal if we could construct a *per-sample target* that retains the appearance of $G(\mathbf{w})$ but where the pose and orientation of the object in the target image is roughly identical across targets. To accomplish this, given $G(\mathbf{w})$, we produce the corresponding target by setting just a portion of the $\mathbf{w}$ vector equal to the target vector $\mathbf{c}$. Specifically, let $\text{mix}(\mathbf{c}, \mathbf{w}) \in \mathbb{R}^{512}$ refer to the latent vector whose first entries are taken from $\mathbf{c}$ and remaining entries are taken from $\mathbf{w}$. By sampling new $\mathbf{w}$ vectors, we can create an infinite pool of paired data where the input is the unaligned image $\boldsymbol{x} = G(\mathbf{w})$ and the target $\boldsymbol{y} = G(\text{mix}(\mathbf{c}, \mathbf{w}))$ shares the appearance of $G(\mathbf{w})$ but is in a learned, fixed pose. This gives rise to the GANgealing loss function:

$$\mathcal{L}_{\text{align}}(T, \mathbf{c}) = \ell(T(\underbrace{G(\mathbf{w})}_{\boldsymbol{x}}), \underbrace{G(\text{mix}(\mathbf{c}, \mathbf{w}))}_{\boldsymbol{y}}), \tag{3}$$

where ℓ is a perceptual loss function [38]. In this paper, we

opt to use StyleGAN2 [43] as our choice of G, but in principle other GAN architectures could be used with our method. An advantage of using StyleGAN2 is that it possesses some innate style-pose disentanglement that we can leverage to construct the per-image target described above. Specifically, we can construct the per-sample targets $G(\text{mix}(\mathbf{c}, \mathbf{w}))$ by using style mixing [42]—$\mathbf{c}$ is supplied to the first few inputs to the synthesis generator that roughly control pose and $\mathbf{w}$ is fed into the later layers that roughly control texture. See Table 3 for a quantitative ablation of the mixing "cutoff point" where we begin to feed in $\mathbf{w}$ (i.e., the cutoff point is chosen as a layer index in $\mathcal{W}^+$ space [1]).

Spatial Transformer Parameterization. Recall that a Spatial Transformer T takes as input an image and regresses and applies a (reverse) sampling grid $\mathbf{g} \in \mathbb{R}^{H \times W \times 2}$ to the input image. Hence, one must choose how to constrain the $\mathbf{g}$ regressed by T. In this paper, we explore a T that performs similarity transformations (rotation, uniform scale, horizontal shift and vertical shift). We also explore an arbitrarily expressive T that directly regresses unconstrained per-pixel flow fields $\mathbf{g}$. Our final T is a composition of the similarity Spatial Transformer into the unconstrained Spatial Transformer, which we found worked best. In contrast to prior work [50, 63], we do not find multi-stage training necessary and train our composed T end-to-end. Finally, our Spatial Transformer is also capable of performing horizontal flips at test time—please refer to Supplement B.4 for details.

When using the unconstrained T, it can be beneficial to add a total variation regularizer that encourages the predicted flow to be smooth to mitigate degenerate solutions: $\mathcal{L}_{\text{TV}}(T) = \mathcal{L}_{\text{Huber}}(\Delta_x \mathbf{g}) + \mathcal{L}_{\text{Huber}}(\Delta_y \mathbf{g})$, where $\mathcal{L}_{\text{Huber}}$ denotes the Huber loss and Δ_x and Δ_y denote the partial derivative w.r.t. x and y coordinates under finite differences. We also use a regularizer that encourages the flow to not deviate from the identity transformation: $\mathcal{L}_{\text{I}}(T) = ||\mathbf{g}||_2^2$.

Parameterization of c. In practice, we do not backpropagate gradients directly into $\mathbf{c}$. Instead, we parameterize $\mathbf{c}$ as a linear combination of the top-N principal directions of $\mathcal{W}$ space [29, 77]:

$$\mathbf{c} = \bar{\mathbf{w}} + \sum_{i=1}^{N} \alpha_i \mathbf{d}_i, \tag{4}$$

where $\bar{\mathbf{w}}$ is the empirical mean $\mathbf{w}$ vector, $\mathbf{d}_i$ is the i-th principal direction and α_i is the learned scalar coefficient of the direction. Instead of optimizing $\mathcal{L}$ w.r.t. $\mathbf{c}$ directly, we optimize it w.r.t. the coefficients $\{\alpha_i\}_{i=1}^N$. The motivation for this reparameterization is that StyleGAN's $\mathcal{W}$ space is highly expressive. Hence, in the absence of additional constraints, naive optimization of $\mathbf{c}$ can yield poor target images off the manifold of natural images. Decreasing N keeps $\mathbf{c}$ on the manifold and prevents degenerate solutions. See Table 3 for an ablation of N.

Our final GANgealing objective is given by:

$$\mathcal{L}(T, \mathbf{c}) = \mathbb{E}_{\mathbf{w} \sim \mathcal{W}}[\mathcal{L}_{\text{align}}(T, \mathbf{c}) + \lambda_{\text{TV}} \mathcal{L}_{\text{TV}}(T) + \lambda_I \mathcal{L}_{\text{I}}(T)]. \tag{5}$$

We set the loss weighting λ_{TV} at either 1000 or 2500 (depending on choice of ℓ) and the loss weighting λ_I at 1. See Supplement B for additional details and hyperparameters.

3.2. Joint Alignment and Clustering

GANgealing as described so far can handle highly-multimodal data (e.g., LSUN Bicycles, Cats, etc.). Some datasets, such as LSUN Horses, feature extremely diverse poses that cannot be represented well by a single mode in the data. To handle this situation, GANgealing can be adapted into a clustering algorithm by simply learning more than one target latent $\mathbf{c}$. Let K refer to the number of $\mathbf{c}$ vectors (clusters) we wish to learn. Since each $\mathbf{c}$ captures a specific mode in the data, learning multiple $\{\mathbf{c}_k\}_{k=1}^K$ would enable us to learn multiple modes. Now, each $\mathbf{c}_k$ will learn its own set of α coefficients. Similarly, we will now have K Spatial Transformers, one for each mode being learned. This variant of GANgealing amounts to simultaneously clustering the data and learning dense correspondence between all images within each cluster. To encourage each $\mathbf{c}_k$ and T_k pair to specialize in a particular mode, we include a hard-assignment step to assign unaligned synthetic images to modes:

$$\mathcal{L}_{\text{align}}^K(T, \mathbf{c}) = \min_k \mathcal{L}_{\text{align}}(T_k, \mathbf{c}_k) \tag{6}$$

Note that the $K = 1$ case is equivalent to the previously described unimodal case. At test time, we can assign an input fake image $G(\mathbf{w})$ to its corresponding cluster index $k^* = \arg\min_k \mathcal{L}_{\text{align}}(T_k, \mathbf{c}_k)$. Then, we can warp it with the Spatial Transformer T_{k^*}. However, a problem arises in that we cannot compute this cluster assignment for input *real* images—the assignment step requires computing $\mathcal{L}_{\text{align}}$, which itself requires knowledge of the input image's corresponding $\mathbf{w}$ vector. The most obvious solution to this problem is to perform GAN inversion [8, 11, 94] on input real images $\boldsymbol{x}$ to obtain a latent vector $\mathbf{w}$ such that $G(\mathbf{w}) \approx \boldsymbol{x}$. However, accurate GAN inversion for non-face datasets remains somewhat challenging and slow, despite recent progress [3, 33]. Instead, we opt to train a classifier that directly predicts the cluster assignment of an input image. We train the classifier using a standard cross-entropy loss on (input fake image, target cluster) pairs $(G(\mathbf{w}), k^*)$, where k^* is obtained using the above assignment step. We initialize the classifier with the weights of T (replacing the warp head with a randomly-initialized classification head). As with the Spatial Transformer, the classifier generalizes well to real images despite being trained exclusively on fake samples.

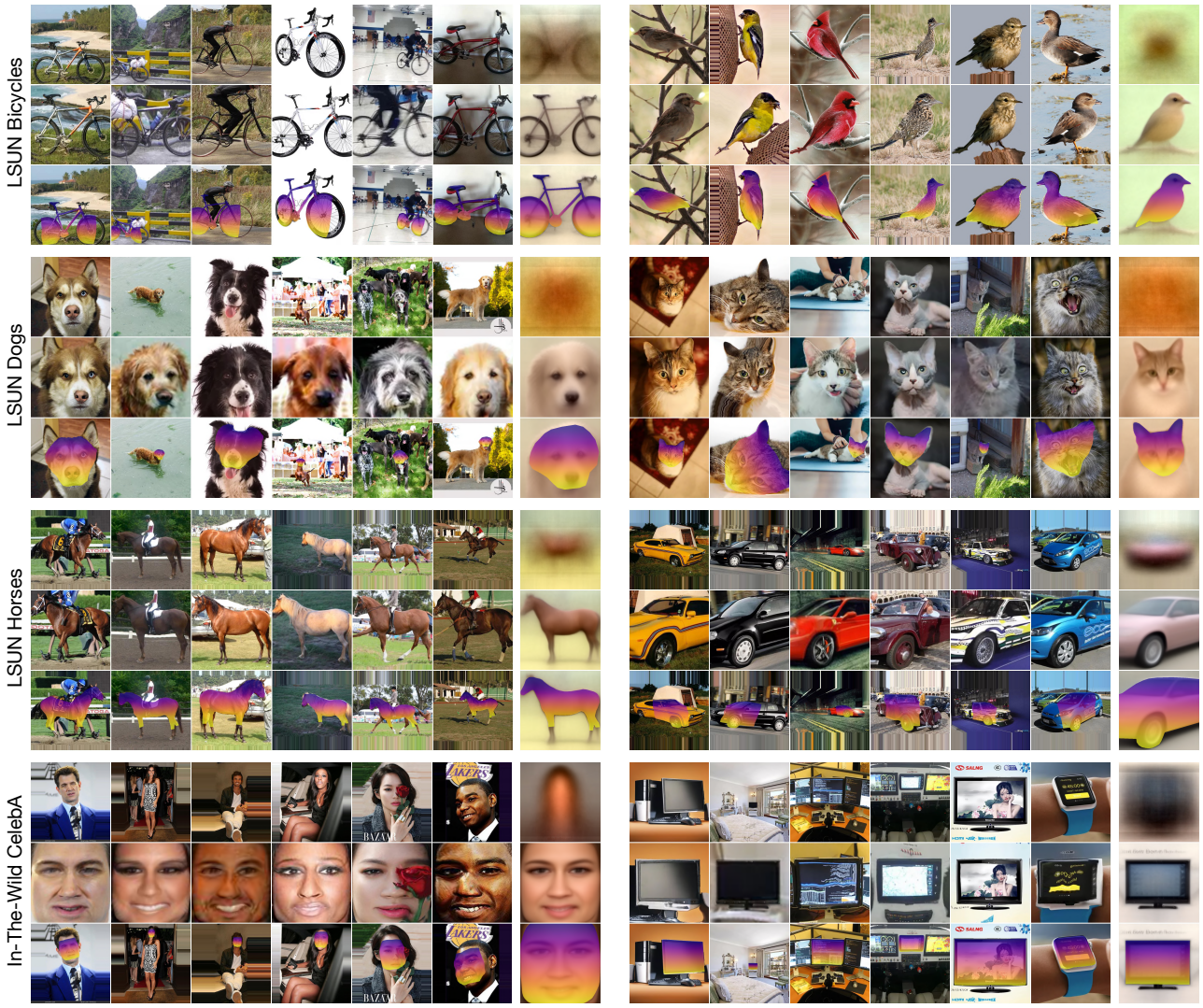

Figure 3. **Dense correspondence results on eight datasets**. For each dataset, the top row shows unaligned images and the dataset average image. The middle row shows our learned alignment of the input images. The bottom row shows dense correspondences between the images. For our clustering models (LSUN Horses and Cars), we show results for one selected cluster. See Supplement F for uncurated results.

4. Experiments

In this section, we present quantitative and qualitative results of GANgealing on eight datasets: LSUN Bicycles, Cats, Cars, Dogs, Horses and TVs [87], In-The-Wild CelebA [52] and CUB-200-2011 [83]. These datasets feature significant diversity in appearance, pose and occlusion of objects. Only LSUN Cars and Horses use clustering ($K = 4$)[1]; for all other datasets we use unimodal GANgealing ($K = 1$). Note that all figures except Figure 2 show our method applied to real images—not GAN samples. Please see www.wpeebles.com/gangealing for full results.

[1] K is a hyperparameter that can be set by the user. We found $K = 4$ to be a good default choice for our clustering models.

4.1. Propagation from Congealed Space

With the Spatial Transformer T trained, it is trivial to identify dense correspondences between real input images $\boldsymbol{x}$. A particularly convenient way to find dense correspondences between a set of images is by propagating from our *congealed coordinate space*. As described earlier, T both regresses and applies a sampling grid $\mathbf{g}$ to an input image. Because we use reverse sampling, this grid tells us where each point in the congealed image $T(\boldsymbol{x})$ maps to in the original image $\boldsymbol{x}$. This enables us to propagate *anything* from the congealed coordinate space—dense labels, sparse keypoints, etc. If a user annotates a *single* congealed image (or the average congealed image) they can then propagate those labels to an entire dataset by simply predicting the grid $\mathbf{g}$ for each im-

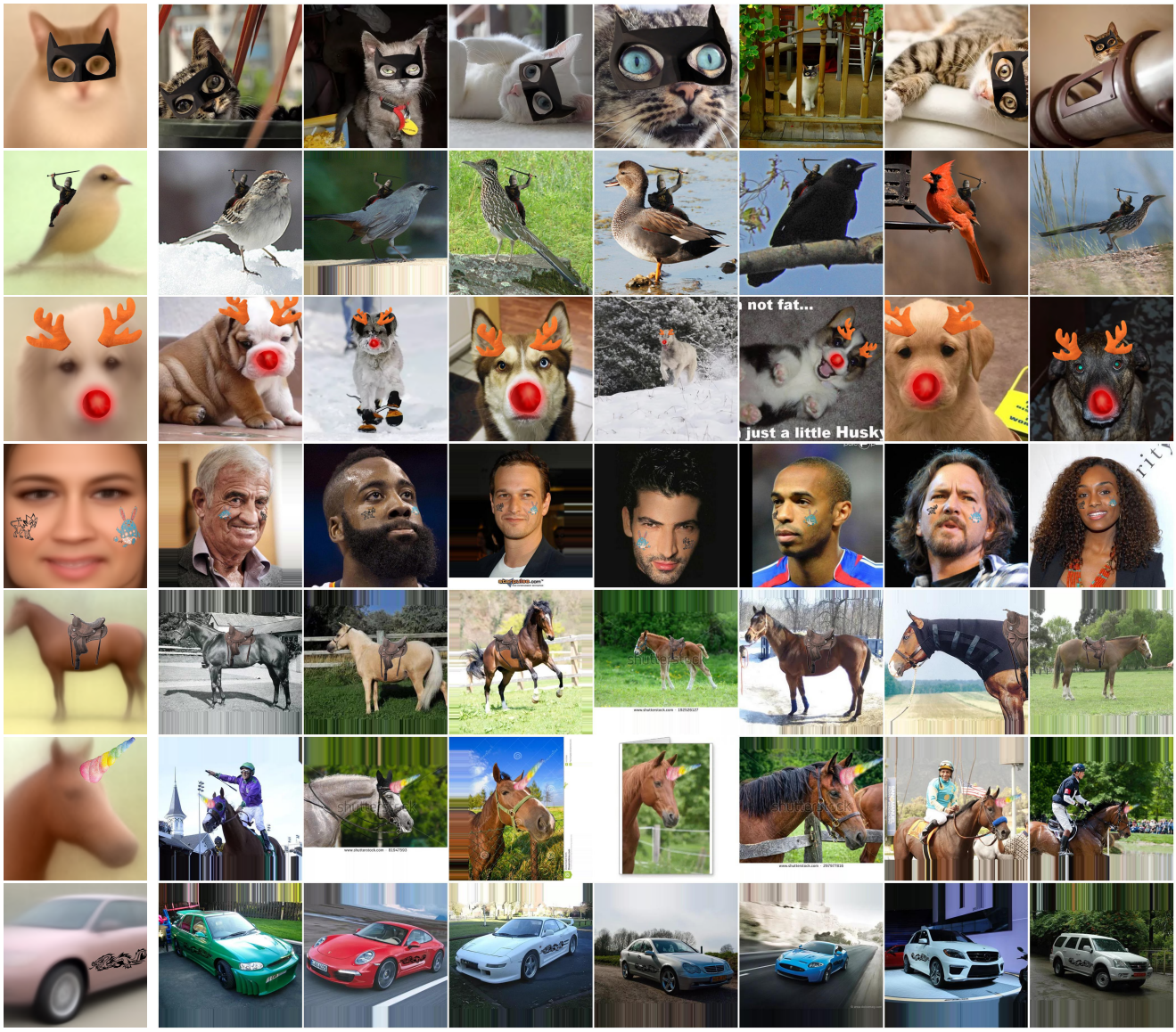

Figure 4. **Image editing with GANgealing**. By annotating just a single image per-category (our average transformed image), a user can propagate their edits to any image or video in the same category.

age $\boldsymbol{x}$ in their dataset via a forward pass through T. Figures 1 and 3 show visual results for all eight datasets—our method can find accurate dense correspondences in the presence of significant appearance and pose diversity. GANgealing accurately handles diverse morphologies of birds, cats with varying facial expressions and bikes in different orientations.

Image Editing. Our average congealed image is a template that can propagate any user edit to images of the same category. For example, by drawing cartoon eyes or overlaying a Batman mask on our average congealed cat, a user can effortlessly propagate their edits to massive numbers of cat images with forward passes of T. We show editing results on several datasets in Figures 4 and 1.

Augmented Reality. Just as we can propagate dense correspondences to images, we can also propagate to individual video frames. Surprisingly, we find that GANgealing yields remarkably smooth and consistent results when applied *out-of-the-box* to videos per-frame without leveraging any temporal information. This enables mixed reality applications like dense tracking and filters. GANgealing can outperform supervised methods like RAFT [75]—please see `www.wpeebles.com/gangealing` for results.

4.2. Direct Image-to-Image Correspondence

In addition to propagating correspondences from congealed space to unaligned images, we can also find dense correspondences directly between any pair of images $\boldsymbol{x}_A$

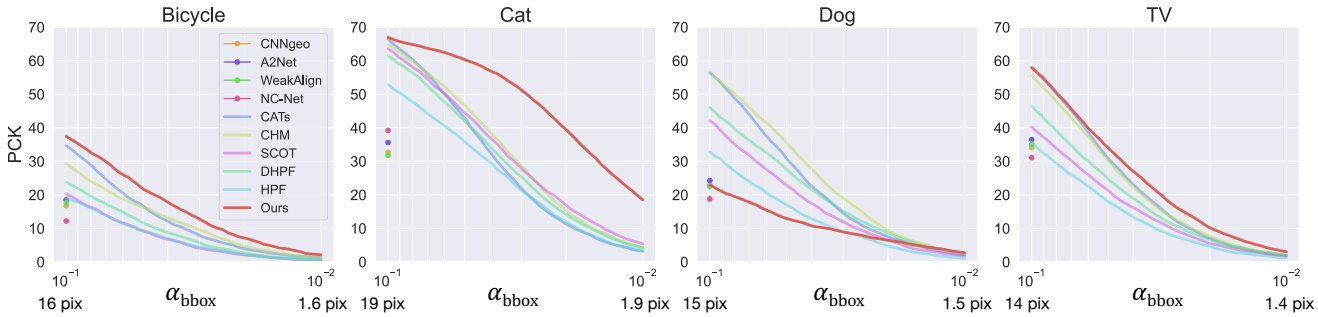

Figure 5. **PCK@α_{bbox} on various SPair-71K categories for α_{bbox} between 10^{-1} and 10^{-2}.** We report the average threshold (maximum distance for a correspondence to be deemed correct) in pixels for 256×256 images beneath each plot. GANgealing outperforms state-of-the-art supervised methods for very precise thresholds ($<$ 2 pixel error tolerance), sometimes by substantial margins.

and $\boldsymbol{x}_B$. At a high level, this merely involves applying the forward warp that maps points in $\boldsymbol{x}_A$ to points in $T(\boldsymbol{x}_A)$ and composing it with the reverse warp that maps points in the congealed coordinate space back to $\boldsymbol{x}_B$. Please refer to Supplement B.7 for details.

Quantitative Results. We evaluate GANgealing with PCK-Transfer. Given a source image $\boldsymbol{x}_A$, target image $\boldsymbol{x}_B$ and ground-truth keypoints for both images, PCK-Transfer measures the percentage of keypoints transferred from $\boldsymbol{x}_A$ to $\boldsymbol{x}_B$ that lie within a certain radius of the ground-truth keypoints in $\boldsymbol{x}_B$.

We evaluate PCK on SPair-71K [59] and CUB. For SPair, we use the α_{bbox} threshold in keeping with prior works. Under this threshold, a predicted keypoint is deemed to be correctly transferred if it is within a radius $\alpha_{\text{bbox}} \max(H_{\text{bbox}}, W_{\text{bbox}})$ of the ground truth, where H_{bbox} and W_{bbox} are the height and width of the object bounding box in the target image. For each SPair category, we train a StyleGAN2 on the corresponding LSUN category[2]—the GANs are trained on 256×256 center-cropped images. We then train a Spatial Transformer using GANgealing and directly evaluate on SPair. For CUB, we first pre-train a StyleGAN2 with ADA [41] on the NABirds dataset [81] and fine-tune it with FreezeD [61] on the training split of CUB, using the same image pre-processing and dataset splits as ACSM [46] for a fair comparison. When T performs a horizontal flip for one image in a pair, we permute our model's predictions for keypoints with a left versus right distinction.

SPair-71K Results. We compare against several self-supervised and state-of-the-art supervised methods on the challenging SPair-71K dataset in Table 1, using the standard $\alpha_{\text{bbox}} = 0.1$ threshold. Our method significantly outperforms prior self-supervised methods on several categories, nearly doubling the best prior self-supervised method's PCK on SPair Bicycles and Cats. *GANgealing performs on par with and even outperforms state-of-the-art correspondence-supervised methods on several categories.* We increase the previous best PCK on Bicycles achieved by Cost Aggregation Transformers [14] from 34.7% to 37.5% and perform comparably on Cats and TVs.

Method	Correspondence Supervision	SPair-71K Category			
		Bicycle	Cat	Dog	TV
HPF [58]	matching pairs + keypoints	18.9	52.9	32.8	35.6
DHPF [60]	matching pairs + keypoints	23.8	61.6	46.1	46.5
SCOT [51]	matching pairs + keypoints*	20.7	63.1	42.5	40.8
CHM [57]	matching pairs + keypoints	29.3	64.9	56.1	55.6
CATs [14]	matching pairs + keypoints	34.7	66.5	56.5	58.0
WeakAlign [69]	matching image pairs	17.6	31.8	22.6	35.1
NC-Net [70]	matching image pairs	12.2	39.2	18.8	31.1
CNNgeo [68]	self-supervised	16.7	32.7	22.8	34.1
A2Net [71]	self-supervised	18.5	35.6	**24.3**	36.5
GANgealing	**GAN-supervised**	**37.5**	**67.0**	23.1	**57.9**

Table 1. **PCK-Transfer@$\alpha_{\text{bbox}} = 0.1$ results on SPair-71K categories** (test split).

High-Precision SPair-71K Results. The usual $\alpha_{\text{bbox}} = 0.1$ threshold reported by most papers using SPair deems a correspondence correct if it is localized within roughly 10 to 20 pixels of the ground truth for 256×256 images (depending on the SPair category). In Figure 5, we evaluate performance over a range of thresholds between 0.1 and 0.01 (the latter of which affords a roughly 1 to 2 pixel error tolerance, again depending on category). GANgealing outperforms all supervised methods at these high-precision thresholds across all four categories tested. Notably, our LSUN Cats model improves the previous best SPair Cats PCK@$\alpha_{\text{bbox}} = 0.01$ achieved by SCOT [51] from 5.4% to 18.5%. On SPair TVs, we improve the best supervised PCK achieved by Dynamic Hyperpixel Flow [60] from 2.1% to 3.0%. Even on SPair Dogs, where GANgealing is outperformed by every supervised method at low-precision thresholds, we marginally outperform all baselines at the 0.01 threshold.

[2]We use off-the-shelf StyleGAN2 models for LSUN Cats, Dogs and Horses. Note that we do not evaluate PCK on our clustering models (LSUN Cars and Horses) as these models can only transfer points between images in the same cluster.

Figure 6. **GANgealing alignment improves downstream GAN training**. We show random, untruncated samples from StyleGAN2 trained on LSUN Cats versus our aligned LSUN Cats (both models trained from scratch). Our method improves visual fidelity.

Figure 7. **Various failure modes**: significant out-of-plane rotation and complex poses poorly modeled by GANs.

CUB Results. Table 2 shows PCK results on CUB, comparing against several 2D and 3D correspondence methods that use varying amounts of supervision. GANgealing achieves 57.5% PCK, outperforming all past methods that require instance mask supervision and performing comparably with the best correspondence-supervised baseline (58.5%).

Ablations. We ablate several components of GANgealing in Table 3. We find that learning the target mode $\mathbf{c}$ is critical for complex datasets; fixing $\mathbf{c} = \bar{\mathbf{w}}$ dramatically degrades PCK from 67% to 10.6% for our LSUN Cats model. This highlights the value of our GAN-Supervised Learning framework where *both the discriminative model and targets are learned jointly*. We additionally find that our baseline inspired by traditional Congealing (using a single learned target $G(\mathbf{c})$ for all inputs) is highly unstable and degrades PCK to as little as 7.7%. This result demonstrates the importance of our *per-input* alignment targets. We also ablate two choices of the perceptual loss ℓ: an off-the-shelf supervised option (LPIPS [88]) and a fully-unsupervised VGG-16 [73] pre-trained with SimCLR [13] on ImageNet-1K [17] (SSL)—there is no significant difference in performance between the two ($\pm 0.2\%$). Please see Table 3 for more ablations.

4.3. Automated GAN Dataset Pre-Processing

An exciting application of GANgealing is automated dataset pre-processing. Dataset alignment is an important yet costly step for many machine learning methods. GAN training in particular benefits from carefully-aligned and filtered datasets, such as FFHQ [42], AFHQ [15] and CelebA-HQ [40]. We can align input datasets using our similarity

Method	Supervision Required		PCK@0.1
	Inst. Mask	Keypoints	
Rigid-CSM (with keypoints) [47]	✓	✓	45.8
ACSM (with keypoints) [46]	✓	✓	51.0
IMR (with keypoints) [80]	✓	✓	58.5
Dense Equivariance [76]	✓		33.5
Rigid-CSM [47]	✓		36.4
ACSM [46]	✓		42.6
IMR [80]	✓		53.4
Neural Best Buddies [2]			35.1
Neural Best Buddies (with flip heuristic)			37.8
GANgealing			**57.5**

Table 2. **PCK-Transfer@0.1 on CUB**. Numbers for the 3D methods are reported from [46]. We sample 10,000 random pairs from the CUB validation split as in [46].

Ablation Description	Loss (ℓ)	$\mathcal{W}^+$ cutoff	λ_{TV}	N	PCK
Don't learn $\mathbf{c}$ (fix $\mathbf{c} = \bar{\mathbf{w}}$)	SSL	5	1000	0	10.6
Unconstrained $\mathbf{c}$ optimization	SSL	5	1000	512	0.34
Early style mixing cutoff	SSL	4	1000	1	60.5
Late style mixing cutoff	SSL	6	1000	1	65.0
No style mixing	SSL	14	1000	1	25.9
No style mixing (LPIPS)	LPIPS	14	1000	1	7.74
No $\mathcal{L}_{TV}$ regularizer	SSL	5	0	1	59.0
Lower λ_{TV} (LPIPS)	LPIPS	5	1000	1	66.7
Complete model (SSL)	SSL	5	1000	1	**67.2**
Complete model (LPIPS)	LPIPS	5	2500	1	**67.0**

Table 3. **GANgealing ablations for LSUN Cats**. We evaluate on SPair-71K Cats using $\alpha_{\text{bbox}} = 0.1$. SSL refers to using a self-supervised VGG-16 as the perceptual loss ℓ. N refers to the number of $\mathcal{W}$ space PCA coefficients learned when optimizing $\mathbf{c}$. Note that the LSUN Cats StyleGAN2 generator has 14 layers.

Spatial Transformer T to train generators with higher visual fidelity. We show results in Figure 6: training StyleGAN2 from scratch with our learned pre-processing of LSUN Cats yields high-quality samples reminiscent of AFHQ. As we show in Supplement E, our pre-processing accelerates GAN training significantly.

5. Limitations and Discussion

Our Spatial Transformer has a few notable failure modes as demonstrated in Figure 7. One limitation with GANgealing is that we can only reliably propagate correspondences that are visible in our learned target mode. For example, the learned mode of our LSUN Dogs model is the upper-body of a dog—this particular model is thus incapable of finding correspondences between, e.g., paws. A potential solution to this problem is to initialize the learned mode with a user-chosen image via GAN inversion that covers all points of interest. Despite this limitation, we obtain competitive results on SPair for some categories where many keypoints are not visible in the learned mode (e.g., cats).

In this paper, we showed that GANs can be used to train highly competitive dense correspondence algorithms from scratch with our proposed GAN-Supervised Learning framework. We hope this paper will lead to increased adoption of GAN-Supervision for other challenging tasks.

References

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4432–4441, 2019. 4

[2] 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):69, 2018. 1, 8

[3] Yuval Alaluf, Or Patashnik, and Daniel Cohen-Or. Restyle: A residual-based stylegan encoder via iterative refinement. *arXiv preprint arXiv:2104.02699*, 2021. 4

[4] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*, pages 3686–3693, 2014. 2

[5] David Arthur and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. Technical report, Stanford, 2006. 13

[6] Simon Baker and Iain Matthews. Lucas-kanade 20 years on: A unifying framework. *International journal of computer vision*, 56(3):221–255, 2004. 2

[7] Manel Baradad, Jonas Wulff, Tongzhou Wang, Phillip Isola, and Antonio Torralba. Learning to see by looking at noise. *arXiv preprint arXiv:2106.05963*, 2021. 2

[8] David Bau, Hendrik Strobelt, William Peebles, Jonas Wulff, Bolei Zhou, Jun-Yan Zhu, and Antonio Torralba. Semantic photo manipulation with a generative image prior. *ACM Transactions on Graphics (Proceedings of ACM SIGGRAPH)*, 38(4), 2019. 4

[9] David Bau, Jun-Yan Zhu, Jonas Wulff, William Peebles, Hendrik Strobelt, Bolei Zhou, and Antonio Torralba. Seeing what a gan cannot generate. In *IEEE International Conference on Computer Vision (ICCV)*, 2019. 21

[10] Victor Besnier, Himalaya Jain, Andrei Bursuc, Matthieu Cord, and Patrick Pérez. This dataset does not exist: training models from generated images. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2020. 2

[11] Andrew Brock, Theodore Lim, James M Ritchie, and Nick Weston. Neural photo editing with introspective adversarial networks. In *International Conference on Learning Representations (ICLR)*, 2017. 4

[12] Lucy Chai, Jun-Yan Zhu, Eli Shechtman, Phillip Isola, and Richard Zhang. Ensembling with deep generative views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14997–15007, 2021. 2

[13] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 8

[14] Seokju Cho, Sunghwan Hong, Sangryul Jeon, Yunsung Lee, Kwanghoon Sohn, and Seungryong Kim. Cats: Cost aggregation transformers for visual correspondence. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021. 1, 7, 21

[15] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1, 8, 17

[16] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. *arXiv preprint arXiv:1703.06211*, 2017. 2

[17] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009. 8

[18] Terrance DeVries, Michal Drozdzal, and Graham W Taylor. Instance selection for gans. *Advances in Neural Information Processing Systems*, 2020. 16

[19] Santosh Divvala, Alexei Efros, and Martial Hebert. Object instance sharing by enhanced bounding box correspondence. In *BMVC*, 2012. 2

[20] Jeff Donahue, Philipp Krähenbühl, and Trevor Darrell. Adversarial feature learning. In *International Conference on Learning Representations (ICLR)*, 2017. 2

[21] Jeff Donahue and Karen Simonyan. Large scale adversarial representation learning. *arXiv preprint arXiv:1907.02544*, 2019. 2

[22] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2758–2766, 2015. 1

[23] Vincent Dumoulin, Ishmael Belghazi, Ben Poole, Alex Lamb, Martin Arjovsky, Olivier Mastropietro, and Aaron Courville. Adversarially learned inference. In *International Conference on Learning Representations (ICLR)*, 2017. 2

[24] Brendan J. Frey and Nebojsa Jojic. Transformation-invariant clustering using the EM algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 25(1):1–17, 2003. 2

[25] J. Brendan Frey and Nebojsa Jojic. Estimating mixture models of images and inferring spatial transformations using the em algorithm. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1999. 2

[26] Yaroslav Ganin, Daniil Kononenko, Diana Sungatullina, and Victor Lempitsky. Deepwarp: Photorealistic image resynthesis for gaze manipulation. In *European Conference on Computer Vision*, pages 311–326. Springer, 2016. 2

[27] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2014. 2

[28] Zekun Hao, Arun Mallya, Serge Belongie, and Ming-Yu Liu. GANcraft: Unsupervised 3D Neural Rendering of Minecraft Worlds. In *ICCV*, 2021. 2

[29] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable gan controls. *arXiv preprint arXiv:2004.02546*, 2020. 4

[30] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Con-*

ference on Computer Vision and Pattern Recognition (CVPR), 2016. 13

[31] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, 2017. 16

[32] G. B. Huang, V. Jain, and E. Learned-Miller. Unsupervised joint alignment of complex images. In *2007 IEEE 11th International Conference on Computer Vision*, 2007. 2

[33] Minyoung Huh, Richard Zhang, Jun-Yan Zhu, Sylvain Paris, and Aaron Hertzmann. Transforming and projecting images to class-conditional generative networks. In *European Conference on Computer Vision (ECCV)*, 2020. 4

[34] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2462–2470, 2017. 1

[35] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *Advances in Neural Information Processing Systems*, pages 2017–2025, 2015. 2, 3

[36] Ali Jahanian, Xavier Puig, Yonglong Tian, and Phillip Isola. Generative models as a data source for multiview representation learning. *arXiv preprint arXiv:2106.05258*, 2021. 2

[37] Xu Jia, Bert De Brabandere, Tinne Tuytelaars, and Luc V Gool. Dynamic filter networks. In *Advances in Neural Information Processing Systems*, pages 667–675, 2016. 2

[38] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision (ECCV)*, 2016. 3

[39] Angjoo Kanazawa, David W Jacobs, and Manmohan Chandraker. Warpnet: Weakly supervised matching for single-view reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3253–3261, 2016. 2

[40] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *International Conference on Learning Representations (ICLR)*, 2018. 1, 8, 17

[41] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *arXiv*, 2020. 7

[42] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 4, 8, 16, 17

[43] 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. 4, 13

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

[45] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015. 14

[46] Nilesh Kulkarni, Abhinav Gupta, David F Fouhey, and Shubham Tulsiani. Articulation-aware canonical surface mapping. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 7, 8

[47] Nilesh Kulkarni, Abhinav Gupta, and Shubham Tulsiani. Canonical surface mapping via geometric cycle consistency. *IEEE International Conference on Computer Vision (ICCV)*, 2019. 8

[48] Erik G. Learned-Miller. Data driven image models through continuous joint alignment. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(2):236–250, 2006. 1, 2, 3

[49] Chen-Hsuan Lin and Simon Lucey. Inverse compositional spatial transformer networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[50] Chen-Hsuan Lin, Ersin Yumer, Oliver Wang, Eli Shechtman, and Simon Lucey. St-gan: Spatial transformer generative adversarial networks for image compositing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9455–9464, 2018. 2, 4

[51] 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 (CVPR)*, June 2020. 1, 7

[52] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *IEEE International Conference on Computer Vision (ICCV)*, December 2015. 2, 5

[53] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 14

[54] Bruce D. Lucas and Takeo Kanade. An iterative image registration technique with an application to stereo vision. In *Proceedings of the 7th International Joint Conference on Artificial Intelligence - Volume 2*, IJCAI'81, pages 674–679, 1981. 2

[55] Chengzhi Mao, Augustine Cha, Amogh Gupta, Hao Wang, Junfeng Yang, and Carl Vondrick. Generative interventions for causal learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3947–3956, 2021. 2

[56] Luke Melas-Kyriazi, Christian Rupprecht, Iro Laina, and Andrea Vedaldi. Finding an unsupervised image segmenter in each of your deep generative models. *arXiv preprint arXiv:2105.08127*, 2021. 2

[57] Juhong Min and Minsu Cho. Convolutional hough matching networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2940–2950, June 2021. 1, 7

[58] 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*, pages 3395–3404, 2019. 1, 7

[59] 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, 7

[60] Juhong Min, Jongmin Lee, Jean Ponce, and Minsu Cho. Learning to compose hypercolumns for visual correspondence. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*, pages 346–363. Springer, 2020. 1, 7

[61] Sangwoo Mo, Minsu Cho, and Jinwoo Shin. Freeze the discriminator: a simple baseline for fine-tuning gans. *arXiv preprint arXiv:2002.10964*, 2020. 7

[62] Hossein Mobahi, Ce Liu, and William T. Freeman. A compositional model for low-dimensional image set representation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, pages 1322–1329. IEEE Computer Society, 2014. 2

[63] Tom Monnier, Thibault Groueix, and Mathieu Aubry. Deep transformation-invariant clustering. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 7945–7955. Curran Associates, Inc., 2020. 2, 4

[64] Xingang Pan, Bo Dai, Ziwei Liu, Chen Change Loy, and Ping Luo. Do 2d gans know 3d shape? unsupervised 3d shape reconstruction from 2d image gans. In *International Conference on Learning Representations*, 2021. 2

[65] Eunbyung Park, Jimei Yang, Ersin Yumer, Duygu Ceylan, and Alexander C Berg. Transformation-grounded image generation network for novel 3d view synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[66] 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 Trans. Pattern Anal. Mach. Intell.*, 34(11):2233–2246, 2012. 2

[67] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *International Conference on Learning Representations (ICLR)*, 2016. 2

[68] Ignacio Rocco, Relja Arandjelovic, and Josef Sivic. Convolutional neural network architecture for geometric matching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6148–6157, 2017. 1, 7

[69] 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. 1, 7

[70] Ignacio Rocco, Mircea Cimpoi, Relja Arandjelović, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Neighbourhood consensus networks. In *Advances in Neural Information Processing Systems*, volume 31, 2018. 1, 7

[71] Paul Hongsuck Seo, Jongmin Lee, Deunsol Jung, Bohyung Han, and Minsu Cho. Attentive semantic alignment with offset-aware correlation kernels. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 349–364, 2018. 1, 7

[72] Yichun Shi, Divyansh Aggarwal, and Anil K Jain. Lifting 2d stylegan for 3d-aware face generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6258–6266, 2021. 2

[73] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *ICLR Workshop*, 2014. 8

[74] Fabio Henrique Kiyoiti dos Santos Tanaka and Claus Aranha. Data augmentation using gans. *arXiv preprint arXiv:1904.09135*, 2019. 2

[75] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European Conference on Computer Vision*, pages 402–419. Springer, 2020. 1, 6, 13, 21

[76] James Thewlis, Andrea Vedaldi, and Hakan Bilen. Unsupervised object learning from dense equivariant image labelling. In *Advances in Neural Information Processing Systems*, 2017. 8

[77] Yu Tian, Jian Ren, Menglei Chai, Kyle Olszewski, Xi Peng, Dimitris N. Metaxas, and Sergey Tulyakov. A good image generator is what you need for high-resolution video synthesis. In *International Conference on Learning Representations*, 2021. 4

[78] Antonio Torralba. http://people.csail.mit.edu/torralba/gallery/, 2001. 2

[79] Nontawat Tritrong, Pitchaporn Rewatbowornwong, and Supasorn Suwajanakorn. Repurposing gans for one-shot semantic part segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4475–4485, 2021. 2

[80] Shubham Tulsiani, Nilesh Kulkarni, and Abhinav Gupta. Implicit mesh reconstruction from unannotated image collections. 2020. 8

[81] Grant Van Horn, Steve Branson, Ryan Farrell, Scott Haber, Jessie Barry, Panos Ipeirotis, Pietro Perona, and Serge Belongie. Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 595–604, 2015. 7

[82] Andrey Voynov, Stanislav Morozov, and Artem Babenko. Big gans are watching you: Towards unsupervised object segmentation with off-the-shelf generative models. *arXiv preprint arXiv:2006.04988*, 2020. 2

[83] 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

[84] Eric Wu, Kevin Wu, David Cox, and William Lotter. Conditional infilling gans for data augmentation in mammogram classification. In *Image analysis for moving organ, breast, and thoracic images*, pages 98–106. Springer, 2018. 2

[85] Xianglei Xing, Ruiqi Gao, Tian Han, Song-Chun Zhu, and Ying Nian Wu. Deformable generator network: Unsupervised disentanglement of appearance and geometry. *arXiv preprint arXiv:1806.06298*, 2018. 2

[86] Xinchen Yan, Jimei Yang, Ersin Yumer, Yijie Guo, and Honglak Lee. Perspective transformer nets: Learning single-view 3d object reconstruction without 3d supervision. In *Advances in Neural Information Processing Systems*, pages 1696–1704, 2016. 2

[87] 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, 5

[88] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 8

[89] Yuxuan Zhang, Wenzheng Chen, Huan Ling, Jun Gao, Yinan Zhang, Antonio Torralba, and Sanja Fidler. Image gans meet differentiable rendering for inverse graphics and interpretable 3d neural rendering. *arXiv preprint arXiv:2010.09125*, 2020. 2

[90] Yuxuan Zhang, Huan Ling, Jun Gao, Kangxue Yin, Jean-Francois Lafleche, Adela Barriuso, Antonio Torralba, and Sanja Fidler. Datasetgan: Efficient labeled data factory with minimal human effort. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10145–10155, 2021. 2

[91] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. *arXiv preprint arXiv:1704.07813*, 2017. 2

[92] Tinghui Zhou, Yong Jae Lee, Stella Yu, and Alexei A Efros. Flowweb: Joint image set alignment by weaving consistent, pixel-wise correspondences. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 2

[93] Tinghui Zhou, Shubham Tulsiani, Weilun Sun, Jitendra Malik, and Alexei A Efros. View synthesis by appearance flow. *ECCV*, 2016. 2

[94] Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei A. Efros. Generative visual manipulation on the natural image manifold. In *ECCV*, 2016. 4

[95] 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. 2