Learning Dense Correspondences between Photos and Sketches

Xuanchen Lu [1] **Xiaolong Wang** [1] **Judith E. Fan** [1 2]

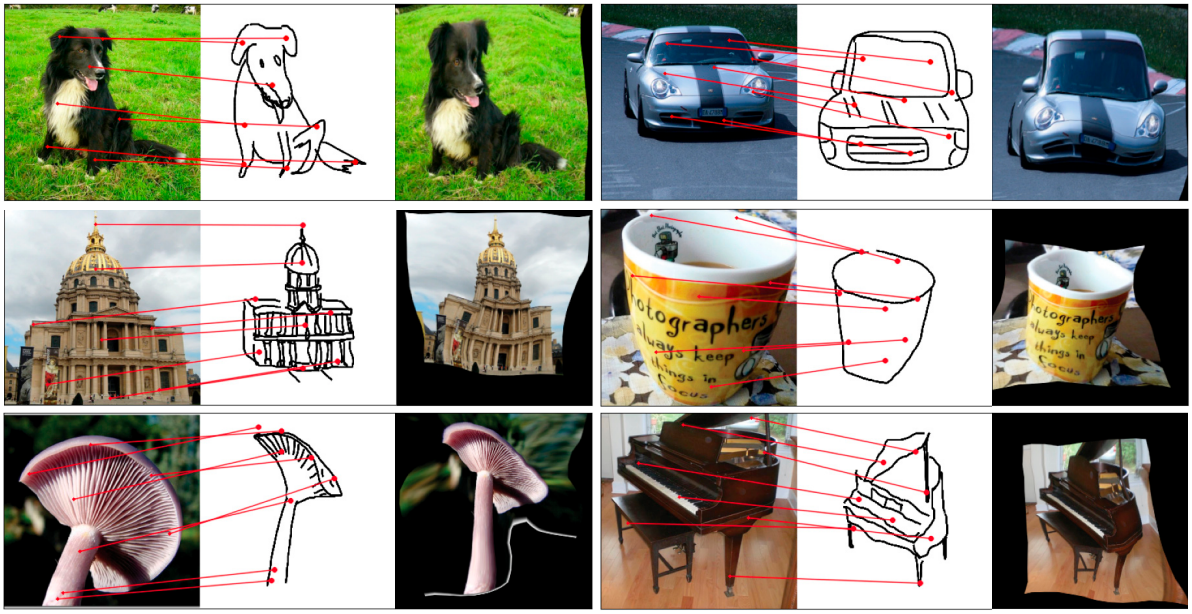

Figure 1: We propose a self-supervised method for learning the dense correspondence between sketches and photos. For each photo-sketch pair, we show the annotated keypoints from our benchmark dataset *PSC6K* (first column), the predicted correspondences (second column), and the result of warping the photo to the sketch (third column).

Abstract

Humans effortlessly grasp the connection between sketches and real-world objects, even when these sketches are far from realistic. Moreover, human sketch understanding goes beyond categorization — critically, it also entails understanding how individual elements within a sketch correspond to parts of the physical world it represents. What are the computational ingredients needed to support this ability? Towards answering this question, we make two contributions: first, we introduce a new sketch-photo correspondence benchmark, `PSC6k`, containing 150K annotations of 6250 sketch-photo pairs across 125 object categories, augmenting the existing Sketchy dataset (Sangkloy et al., 2016) with fine-grained correspondence metadata. Second, we propose a self-supervised method for learning dense correspondences between sketch-photo pairs, building upon recent advances in correspondence learning for pairs of photos. Our model uses a spatial transformer network to estimate the warp flow between latent representations of a sketch and photo extracted by a contrastive learning-based ConvNet backbone. We found that this approach outperformed several strong baselines and produced predictions that were quantitatively consistent with other warp-based methods. However, our benchmark also revealed systematic differences between predictions of the suite of models we tested and those of humans. Taken together, our work suggests a promising path towards developing artificial systems that achieve more human-like understanding of visual images at different levels of abstraction. Project page: `https://photo-sketch-correspondence.github.io`

[1]University of California, San Diego [2]Stanford University. Correspondence to: Judith Fan <jefan@stanford.edu>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

1. Introduction

Sketching is a powerful technique humans use to create images that capture key aspects of the visual world. It is also among the most enduring and versatile of image generation techniques, with the earliest known sketch-like images dating to at least 40,000-60,000 years ago (Hoffmann et al., 2018; Aubert et al., 2014). Although the retinal image cast by a sketch and a real-world object are highly distinct, humans are nevertheless able to grasp the meaning of that sketch at multiple levels of abstraction, including the category label that best applies to it, the specific object instance it represents, as well as detailed correspondences between elements in the sketch and the parts of the object (Fan et al., 2018; Mukherjee et al., 2019; Yang & Fan, 2021). What are the computational ingredients needed to achieve such robust image understanding across domains and at multiple levels of abstraction?

Generalizing across photorealistic and stylized image distributions. There has been substantial recent progress in the development of artificial vision systems that capture some key aspects of sketch understanding, especially sketch categorization and sketch-based image retrieval (Eitz et al., 2012; Sangkloy et al., 2016; Yu et al., 2016; 2017; Bhunia et al., 2020). In addition, the availability of larger models that have been trained on vast quantities of paired image and text data have led to encouraging results on tasks involving images exhibiting different visual styles (Radford et al., 2021), including sketch generation (Vinker et al., 2022). However, recent evidence suggests that even otherwise high-performing vision models trained on photorealistic image data do not generalize well to other image distributions as well as neurons in primate inferotemporal cortex (a key brain region supporting object categorization) (Bagus et al.), indicating that a large gap remains between the capabilities of current computer vision systems and those achieved by biological systems.

Perceiving semantic correspondences between images. In particular, a core open problem in human sketch understanding concerns the computational ingredients required to encode the internal structure of a sketch with sufficient fidelity to establish a detailed mapping between parts of a sketch with parts of the object it represents (Kulvicki, 2015; Fodor, 2007). The problem of discovering semantic correspondences between images is a well established problem in computer vision. In the typical setting, the goal is to establish dense correspondences between images containing objects belonging to the same class. Classical methods (Berg et al., 2005; Kim et al., 2013; Liu et al., 2010) determine the alignment with hand-crafted feature descriptors such as SIFT (Lowe, 1999) or DOG (Dalal & Triggs, 2005). More recently developed methods (Ham et al., 2016; Rocco et al., 2018a; Truong et al., 2021), which benefit from the robust feature representations learned by deep neural networks are more robust to variations in appearance and shape. However, finding correspondence between photos and sketches is particularly challenging as human-generated sketches are inherently selective, highlighting the most relevant aspects of an object's appearance at the expense of other aspects (Fan et al., 2020; Huey et al., 2021). Moreover, sketches typically lack the texture and color cues that can facilitate dense correspondence learning for color photos. As a consequence, the task of learning dense semantic correspondences between photos and sketches relies on a substantial degree of visual abstraction in order to establish strong semantic alignment between images from different modalities.

Self-supervised representation learning. A robust finding from the past decade is that deep neural networks trained with supervision at large, labeled image datasets can achieve state-of-the-art performance (Krizhevsky et al., 2017; Simonyan & Zisserman, 2014; He et al., 2016). Moreover, models trained in this way currently provide the most quantitatively accurate models of biological vision in non-human primates and humans (Yamins et al., 2014; Khaligh-Razavi & Kriegeskorte, 2014; Rajalingham et al., 2018; Cadena et al., 2019). Nevertheless, such models are unlikely to explain how humans are capable of achieving such robust image understanding across different modalities given the implausibility that such large, labeled datasets were available to or necessary for humans to learn to understand natural visual inputs, much less to interpret sketches (Hochberg & Brooks, 1962; Kennedy & Ross, 1975). Recent advances in self-supervised representation learning have begun to approach the performance of supervised models without the need for such labels (Wu et al., 2018; He et al., 2020), while also emulating key aspects of visual processing in biological systems (Zhuang et al., 2021; Konkle & Alvarez, 2020). However, it remains unclear to what degree these advances are sufficient to support challenging multi-domain image understanding tasks, including predicting dense photo-sketch correspondences.

Our contributions: Evaluating a self-supervised method for learning photo-sketch correspondences. Towards meeting these challenges, our paper makes two key contributions: first, we establish a new benchmark for photo-sketch dense correspondence learning: `PSC6k`. This benchmark consists of 150,000 pairs of keypoint annotations for 6250 photo-sketch pairs spanning 125 object categories. Each annotation consists of a keypoint marked by a human participant on an object in a color photo that they judged to correspond to a given keypoint appearing on a sketch of the same object. All photo-sketch pairs were sampled from the well established Sketchy dataset (Sangkloy et al., 2016), a collection of 75K sketches produced by humans to depict objects in 12.5K color photographs of objects spanning 125 categories.

Our second contribution is a self-supervised method for learning photo-sketch correspondences that leverages a learned nonlinear "warping" function to map one image to the other. This approach embodies the hypothesis that sketches preserve key information about spatial relations between an object's constituent parts, even if they also manifest distortions in the size and shape of these parts. This hypothesis is motivated by the view that representational line drawings, as sparse as they are, are meant to accurately convey 3D shape (Hertzmann, 2020), which stands in sharp contrast to the view that the relationship between drawings and objects are established purely by convention (Goodman, 1976). Nevertheless, the nonlinear "warping" approach we propose diverges from very strong versions of the 3D-shape-preservation account (Greenberg, 2021), which are not well equipped to handle the kinds of nonlinear visual distortions that human-generated sketches exhibit (Eitz et al., 2012; Sangkloy et al., 2016; Fan et al., 2018).

Our system consists of two main components: the first is a multimodal image encoder trained with a contrastive loss (Wu et al., 2018; Zhuang et al., 2021), with photos and sketches of the same object being treated as positive examples, and those depicting different objects as negative examples. The second component is a spatial transformer network (Jaderberg et al., 2015) that estimates the transformation between each photo and sketch and aims to maximize the similarity between the feature maps for both images. Using our newly developed `PSC6k` benchmark, we find that our system outperforms other existing self-supervised and weakly supervised correspondence learning methods, and thus establishes the new state-of-the-art for sketch-photo dense correspondence prediction. We will publicly release `PSC6k` with extensive documentation and code to enhance its usability to the research community.

2. Photo-Sketch Correspondence Benchmark (`PSC6k`)

Our first goal was to establish a novel photo-sketch correspondence benchmark satisfying two criteria: first, it should build directly upon existing benchmarks in sketch understanding and second, it should provide broad coverage of a wide variety of visual concepts. Towards that end, we developed `PSC6k` by directly augmenting the Sketchy dataset (Sangkloy et al., 2016), which already contains 75,471 human sketches produced from 12,500 unique photographs spanning 125 object categories.

2.1. Sampling Photo-Sketch Pairs

We sampled photo-sketch pairs from the original test split of the Sketchy dataset, which consisted of 1250 photos and their corresponding sketches. We manually filtered out sketches that were completely off-target or that depicted the photographed object from the wrong perspective (Sangkloy et al., 2016). We then randomly sampled 5 sketches from among the remaining valid sketches produced of each photo, resulting in 6250 unique photo-sketch pairs.

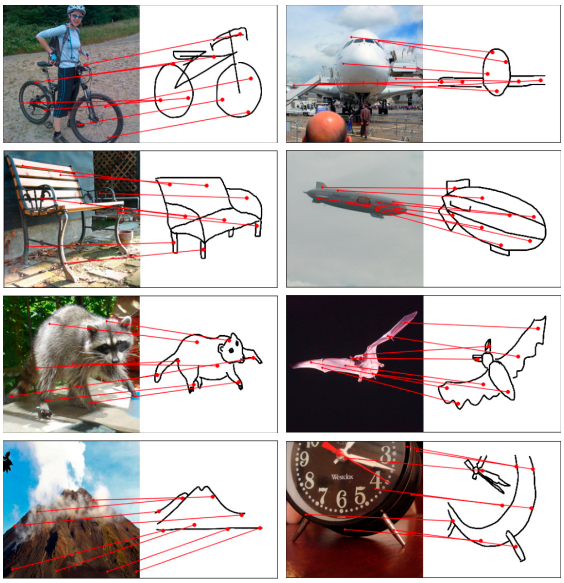

Figure 2: Examples of human-annotated photo-sketch pairs from our new photo-sketch correspondence benchmark `PSC6k`.

2.2. Collecting Human Keypoint Annotations

We formalize the problem of identifying photo-sketch correspondences as the ability to map a keypoint located on a sketch to the location in the source photograph that best corresponds to it. For example, a keypoint appearing on the left wing of a sketch of an airplane should be mapped to the "same" location on the left wing of the photograph of that same airplane. For each photo-sketch pair, we sampled 8 keypoints spanning as much of the object as possible. To determine these keypoints, we first computed segmentation masks for each sketch, relying upon the heuristic that outermost contour of the sketch naturally serves as the contour of the object in the sketch. The pixels covered by the segmentation mask were then clustered into 8 groups to estimate 8 "pseudo-part" regions. We employ nearest-neighbor-based spectral clustering to prioritize connectivity within each pseudo-part. A keypoint was then placed at the centroid of each pseudo-part.

This approach allowed us to automatically discover regions of the sketch that are likely to be semantically meaningful without the need for explicit part labels. However, this approach is also less sensitive to sketch regions that constitute only a small portion of object mask (e.g., a cat's whiskers). As such, future work could employ a combination of region-based and stroke-based keypoints to gain fuller coverage of semantically meaningful regions of sketches.

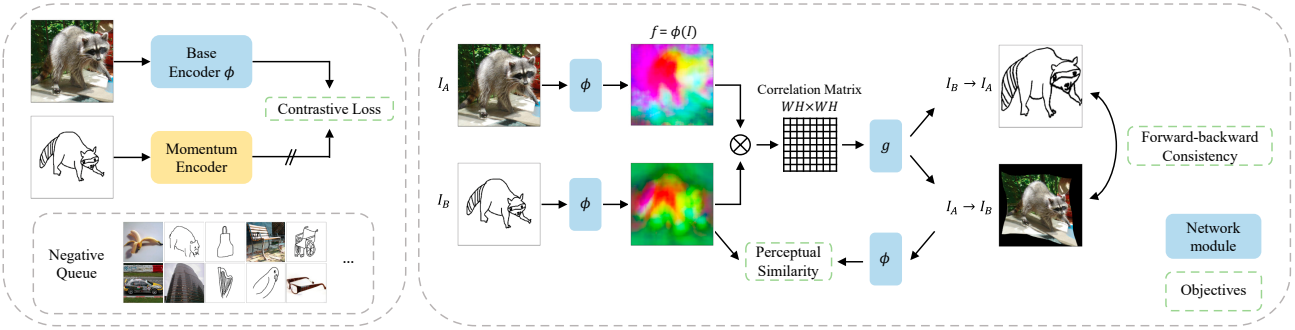

Figure 3: We propose a self-supervised framework for learning photo-sketch correspondence by estimating a dense flow that warps one image to the other. The framework consists of a multi-modal feature encoder that aligns the photo-sketch representation with a contrastive loss, and an STN-based warp estimator to predict transformation that maximizes the similarity between feature maps of the two images. The estimator learns to optimize a combination of weighted perceptual similarity and forward-backward consistency.

Next, we recruited 1,384 participants using the Prolific crowdsourcing platform to provide annotations. Participants provided informed consent in accordance with the UC San Diego Institutional Review Board (IRB). On each trial, participants were cued with a keypoint appearing on a sketch and asked to indicate its corresponding location in a photo appearing next to it. Each participant provided annotations for 125 photo-sketch pairs, one from each category. We collected three annotations from different participants for each keypoint in every sketch, resulting in 150,000 annotations across all 6250 photo-sketch pairs. We defined the centroid over these annotations as the ground-truth keypoint in the photo. In rare cases, there was one annotation out of three with an exceptionally large distance from the median location of all three annotations; these responses were flagged as outliers and excluded from the determination of the centroid. See Appendix A for additional details regarding the creation of this photo-sketch correspondence benchmark.

3. Weakly-supervised Photo-Sketch Correspondence

In this section, we present our weakly-supervised model for finding the pixel-level correspondence between photo-sketch pairs. We formulate the problem as estimating the displacement field across a sketch $I_s \in \mathbb{R}^{h \times w \times 3}$ and a photo $I_p \in \mathbb{R}^{h \times w \times 3}$ that depict the same object (Figure 3). Our goal is to find the cross-modal photo-sketch alignment in a weakly-supervised manner, by maximizing the perceptual similarity of an image in (I_p, I_s) and its warped counterpart. Our framework consists of a feature encoder ϕ that learns a shared feature space of photo and sketch, and a warp estimator T based on the spatial transformer network (STN) that directly predicts the displacement field $F \in \mathbb{R}^{h \times w \times 2}$, where we extract the dense correspondence.

3.1. Feature Encoder ϕ

Here we leverage advances in contrastive learning to develop a weakly-supervised feature encoder on photo-sketch data pairs. Contrastive learning obtains a feature representation by contrasting similar and dissimilar pairs. Here, the photo I_p and the sketch I_s depicting the same object become a natural choice to construct similar pairs. Unlike typical contrastive learning schemes (Wu et al., 2018; Chen et al., 2020a; He et al., 2020) that take augmented views of the same image I as positives, our model uses augmented views from the same photo-sketch pair (I_p, I_s). To minimize the contrastive loss over a set of photo-sketch pairs, the encoder must learn a feature space that attracts photo/sketch from the same pair and separates photo/sketch from distinct pairs.

Similar to (He et al., 2020), we formulate pair-level contrastive learning as a dictionary look-up problem. For a given photo-sketch pair (I_p, I_s), random data augmentation is applied to generate the view pair $(\widetilde{I}_p, \widetilde{I}_s)$. One view in the pair is randomly selected as the query and the other becomes the corresponding key. We denote their representations encoded by ϕ as q and k^+, respectively. The query token q should match its key k^+ over a set of negative keys k^- sampled from other photo-sketch pairs. To optimize this target, we minimize InfoNCE (Oord et al., 2018) as follows:

$$\mathcal{L}_{nce} = -\log \frac{\exp\left(q \cdot k^+ / \tau\right)}{\exp\left(q \cdot k^+ / \tau\right) + \sum_{k^-} \exp\left(q \cdot k^- / \tau\right)}, \quad (1)$$

where τ is a temperature hyperparameter scaling the data distribution in the metric space.

To explore the inherent similarity between photos and sketches, we use a shared encoder ϕ for images from both modalities. We replace the batch normalization (BN) (Ioffe & Szegedy, 2015) in the encoder with conditional batch normalization (De Vries et al., 2017) for better domain

alignment. Detailed implementation and experiment are reported in section 4.

3.2. Warp Estimator T

Given the source and target image I_s, I_t and their representation X_s, X_t, the warp estimator T predicts the displacement field $F_{I_s \to I_t} = T(X_s, X_t)$. Inspired by (Sun et al., 2018), we propose a simplified pyramidal warp estimation module for the ResNet backbone.

Affinity function f. While it is possible to estimate the correspondence based on the feature affinity at a specific layer of the encoder ϕ, e.g., the final convolutional layer, it is beneficial to evaluate affinities at multiple layers along the feature pyramid. We select a set of n feature layers of interest, denoted as $X_s = \{x_s^i\}_{i=0}^{n-1}$ and $X_t = \{x_t^i\}_{i=0}^{n-1}$. We bilinearly upsample all selected feature maps to the same spatial resolution, and concatenate them along the channel dimension for the multi-layer feature maps, $X_s \in \mathbb{R}^{c \times h \times w}$ and $X_t \in \mathbb{R}^{c \times h \times w}$.

With the source and target feature maps X_s and X_t, we compute affinity as the correlation between feature embeddings: with pixel i in feature map X_s and pixel j in feature map X_t, $A_{(s,t)}(i,j) = X_s(i)^T X_t(j)$. The pairwise affinity between every pixel in the source and target feature maps forms the affinity matrix $f(X_s, X_t) := A_{(s,t)} \in \mathbb{R}^{hw \times hw}$.

Estimation Module g. Module g takes the affinity matrix $A_{(s,t)}$ and directly estimates the displacement field F from the source image to the target image. Following the idea of coarse-to-fine refinement, it consists of three STN blocks at different scales with residual connections, denoted as g_1, g_2 and g_3. Each STN-block (except the first block) takes the affinity matrix warped by the previous block and regresses a new displacement field to refine the alignment. The first block g_1 regresses at the 4×4 scale, estimating displacement field $F^{(0)} \in \mathbb{R}^{4 \times 4 \times 2}$. g_2 and g_3 regress at the 8×8 and 16×16 scale, respectively. The displacement field at each block is computed as

$$F^{(1)} = g_1(f(X_s, X_t)), \tag{2}$$

$$F^{(k)} = F^{(k-1)} + g_i(f(warp(X_s, F^{(k-1)}), X_t)), \tag{3}$$

where $warp(I, F)$ operation warps image I to target according to the displacement field F. It is implemented with bilinear interpolation.

After g_3 generates the 16×16 displacement field, it is upsampled to full image resolution as the final estimation.

3.3. Weighted Perceptual Similarity

We propose using weighted perceptual similarity to evaluate the quality of estimated displacement field between the photo-sketch pair. Instead of directly evaluating similarity using the warped source feature map (direct similarity), we pass the warped source image into the feature encoder *again* and evaluate similarity using the new feature map, so that the feature encoder serves as a soft constraint that reduces warping artifacts and stabilizes training (perceptual similarity). We use subscripts to indicate the direction of warp; for example, the displacement field from I_s to I_t is denoted $F_{s \to t}$. We also denote the warped image as $I_{s \to t} = warp(I_s, F_{s \to t})$.

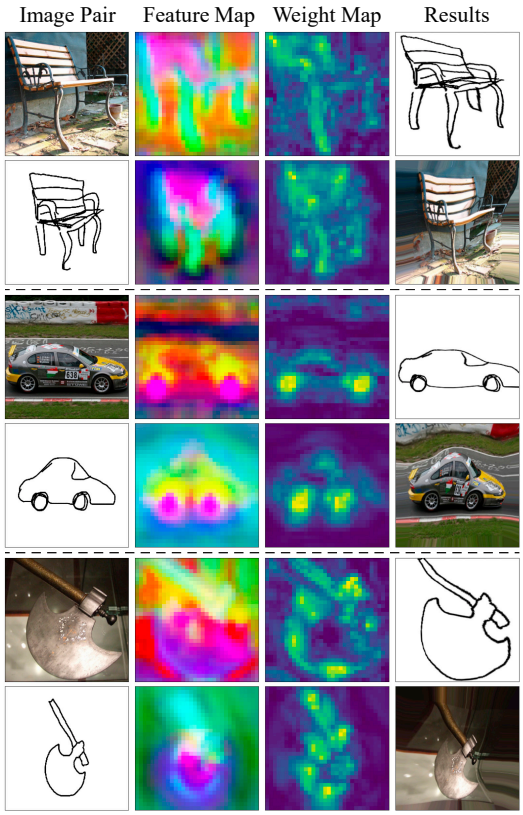

Figure 4: Example image pairs, feature maps, weight maps, and final results processed in our warp estimator. The weight maps highlight semantic parts that have the largest correlation between the two images. We use PCA to project the feature dimensions to 3 principal components as RGB.

Perceptual similarity s. For an image pair (I_s, I_t), the model estimates the flow $F_{s \to t}$ and renders the warped source image $I_{s \to t}$. The warped source image is passed through the encoder ϕ to generate its new set of feature maps $X_{s \to t}$, as well as its new affinity with the target $A_{(s \to t, t)}$. The new affinity matrix represents how well the warped source image semantically aligns with the target.

In the ideal case, each pixel in the warped source $X_{s \to t}$ will have the highest correlation with the pixel at the same location in the target X_t. This is reflected in the affinity space $A_{(s \to t, t)} \in \mathbb{R}^{n \times hw \times hw}$ as a maximized diagonal along the second and third axes. For a pixel in warped source $X_{s \to t}$,

we formulate the optimization as selecting the pixel that matches correctly from all pixels in target X_t:

$$s(n,i) = -\log \frac{\exp\left(A_{(s\to t,t)}(n,i,i)/\tau\right)}{\sum_j \exp\left(A_{(s\to t,t)}(n,i,j)/\tau\right)}, \quad (4)$$

where n is the index of the feature layer to evaluate on; i, j are indices of pixels in the source and target feature map.

Weight function w. While it is possible to optimize flow estimation with the above formula, there are two problems. First, sketches contain a large number of empty pixels, and photos often suffer from background clutter. Moreover, while the encoder activation generally lies over the entire object in the photo, the activation concentrates along the strokes in a sketch. As a result, optimizing the correspondence of every pixel is inefficient and biased toward the background. To focus optimization on important matches, we consider an intuitive rule: important pixels in one image should have greater affinities to the other image. It is formulated as a weight function:

$$w(n,i) = scale(\max_j [norm(A_{(s\to t,t)})(n,i)]) \quad (5)$$

where $norm$ is the normalization over the affinity matrix to penalize pixels that have multiple large affinities in the other image. $scale$ is an arbitrary operation to standardize the weight function. We use Min-Max to scale its distribution to $[0, 1]$.

Therefore, the final perceptual similarity loss is given by

$$\mathcal{L}_{sim}(n,i) = w(n,i)s(n,i) \quad (6)$$

In Figure 4, we visualize the image pairs, feature maps, weight maps, and the final alignment results of photo-sketch pairs from *PSC6k* to exhibit the function of each component in our estimator.

3.4. Additional Objectives

In addition to the perceptual similarity loss, we consider an additional self-supervised loss to assist robust warp estimation and stabilize training.

Forward-backward consistency. Forward-backward consistency is a classical idea in tracking (Vondrick et al., 2018; Wang et al., 2019; Jabri et al., 2020) and flow estimation (Meister et al., 2018; Rocco et al., 2017; Jeon et al., 2018; Truong et al., 2021; Huang et al., 2019) as constraints. Namely, we expect the estimated forward flow $F_{s\to t}$ to be the inverse of the estimated backward flow $F_{t\to s}$. It poses a strict constraint on the network for symmetric prediction. We minimize the $L2$ norm between the identity flow and the composition of the forward flow and backward flow:

$$\mathcal{L}_{con} = \|warp(F_{s\to t}, F_{t\to s}) - F_{\mathbb{I}}\|, \quad (7)$$

where $F_{\mathbb{I}}$ is the identity displacement that maps all locations to themselves.

Overall, our final objective is

$$\mathcal{L} = \lambda_{sim}\mathcal{L}_{sim} + \lambda_{con}\mathcal{L}_{con}, \quad (8)$$

4. Experiments

Here we empirically evaluate our method and compare it to existing approaches in dense correspondence learning on the photo-sketch correspondence benchmark. We analyze the difference between human annotations and predictions from existing methods. We show that our method establishes the state-of-the-art in the photo-sketch correspondence benchmark and learns a more human-like representation from the photo-sketch contrastive learning objectives. We conducted additional experiments to evaluate generalization to unseen categories in Appendix C.

4.1. Implementation Details

The input image size is set to 256 following our photo-sketch correspondence benchmark. We use ResNet-18 and ResNet-101 as our feature encoder. The encoder is initialized with pretrained weights from MoCo training (He et al., 2020) on ImageNet-2012 (Deng et al., 2009). We then train our encoder on the training split of Sketchy for 1300 epochs. Since there are multiple sketches for each photo in the dataset, at each epoch, we iterate through all photos and sample a corresponding sketch for each photo. We follow the recipe from MoCo (He et al., 2020; Chen et al., 2020c), with $dim = 128, m = 0.999, t = 0.07, lr = 0.03$ and a two-layer MLP head. Noticeably, we set the size of the memory queue to $K = 8192$ to prevent multiple positive pairs from appearing at the same time.

We then train the estimator for 1200 epochs with a learning rate of 0.003, leading to 2500 epochs of training in total. We set the weights of the objectives to $\lambda_{sim} = 0.1, \lambda_{con} = 1.0$. We compute $\mathcal{L}_{sim}$ using the features after ResNet stages 2 and 3, and the temperature is set to $\tau = 0.001$.

We apply the same set of augmentations to both feature encoder and the warp estimator, consisting of random color jitter, grayscale, and Gaussian blur, which are consistent with the settings in MoCo v2 (Chen et al., 2020c) and SimCLR (Chen et al., 2020a). However, we replace random cropping with a combination of affine and TPS transformations for a more complex spatial distortion.

We train the network with the SGD optimizer, a weight decay of $1e-4$, a batch size of 256, and the native mixed precision from PyTorch. We adopt a cosine learning rate decay schedule (Loshchilov & Hutter, 2016).

Methods	Encoder	Transfer		Retrain	
		PCK-5	PCK-10	PCK-5	PCK-10
CNNGeo (Rocco et al., 2018a)	ResNet-101	27.59	57.71	19.19	42.57
WeakAlign (Rocco et al., 2018a)	ResNet-101	35.65	68.76	43.55	78.60
NC-Net (Rocco et al., 2018b)	ResNet-101	40.60	63.50	–	–
DCCNet (Huang et al., 2019)	ResNet-101	42.43	66.53	–	–
PMD (Li et al., 2021)	VGG-16	35.77	71.24	–	–
WarpC-SemanticGLUNet (Truong et al., 2021)	VGG-16	48.79	71.43	56.78	79.70
Ours	ResNet-18	–	–	56.01	**82.89**
Ours	ResNet-101	–	–	**57.92**	**84.72**

Table 1: State-of-the-art comparison for photo-sketch correspondence learning.

4.2. Photo-sketch Correspondence Estimation

We evaluate our correspondence estimation results qualitatively and quantitatively. We compare our method with existing approaches in correspondence learning with image or pair-level supervision, and present a state-of-the-art comparison on photo-sketch correspondence in Table 1. For fair comparisons, we retrain existing open-sourced methods on the same photo-sketch dataset we used to develop our own model (Sangkloy et al., 2016). We report their PCK for $\alpha = (0.05, 0.1)$ in two settings: transfer (directly evaluate on photo-sketch correspondence with pretrained weights) and retrain (train from scratch on photo-sketch correspondence). Methods that fail to converge on photo-sketch dataset are left blank. In Appendix B, we include methods with stronger supervision to the table and detail the training/evaluation setting of each method.

Our approach sets a new state-of-the-art for photo-sketch correspondence. Although we only regress flow at 16×16, which is less than the granularity of PCK-05, our ResNet-101 model gains a substantial increase of +1.14%/+5.02% compared to the second-best method WarpC-SemanticGLUNet (Truong et al., 2021). This is surprising as the latter method benefits from flow resolution four times as large as ours, and additional two-stage training on CityScape (Cordts et al., 2016), DPED (Ignatov et al., 2017), and ADE (Zhou et al., 2019). Our smaller ResNet-18 model also outperforms most existing methods despite a significantly shallower feature encoder, demonstrating the effectiveness of our pair-based contrastive learning scheme in finding dense correspondences between images from different image modalities. We visualize more examples of the dense correspondence that our model predicts in Appendix D.

4.3. Ablation Study

We conduct two sets of ablation experiments on the ResNet-18 version of our framework. In Table 2 we analyze different training schemes for the feature encoder. In the first row, we directly use the pretrained weights from ImageNet contrastive learning. The following rows compare the performance of different ways of constructing positive pairs: 1) two augmented views from single images from the photo-sketch dataset, as in classical contrastive learning; 2) a photo and a sketch randomly sampled from the same class; and 3) a photo and a sketch from the same photo-sketch pair. We find that the pretrained model on ImageNet leads to the worst performance due to its failure to generalize to sketch data. Classical contrastive learning on the photo-sketch dataset also harms model estimation, because the domains of photo and sketch are not explicitly aligned in the representation space. The best result comes from contrastive learning on photo-sketch pairs, as it provides the strongest supervision for learning discriminative features. In Table 3, we analyze the key components of our correspondence estimation framework. We first show the importance of our objectives, by ablating the similarity loss, the perceptual version of the similarity loss, and the consistency loss. In addition, we show that the use of the weight function, multiple feature layers, and conditional BN further improves the model performance.

Training Description	PCK-5	PCK-10
ImageNet only	17.20	48.93
CL on individual image	44.41	75.67
CL on image class	54.81	81.72
CL on image pair	56.01	82.89

Table 2: Ablation study on training feature encoder.

Ablation Description	PCK-5	PCK-10
No $\mathcal{L}_{sim}$	17.46	49.43
No perceptual $\mathcal{L}_{sim}$	49.41	80.59
No $\mathcal{L}_{con}$	52.49	80.38
No weight function w	54.29	82.52
No multiple feature layers	55.19	83.14
No conditional BN	55.84	82.67
Complete model	56.01	82.89

Table 3: Ablation study on correspondence estimation.

4.4. Comparing model and human error patterns

To what degree do any of the models tested generate predictions that achieve the degree of consistency that we observe between individual human annotators? To evaluate this question, for each pair of systems (whether two models, two humans, or a model and a human), we computed the normalized mean pixel distance between the predictions they generated for a given photo-sketch pair, then normalized this distance by the image size. We find that while higher-performing models tend to produce predictions that are more similar to one another, all of the models taken together display systematic biases that are distinct from those of humans performing the photo-sketch correspondence task Figure 5. These results indicate the size of the current human-model gap and suggest that future progress on this benchmark will entail bringing human-model consistency values closer to that observed between individual humans.

Human1	0	0.06	0.06	0.12	0.13	0.13	0.21	0.14	0.12	0.2	0.18	0.14	0.15
Human2	0.06	0	0.06	0.12	0.13	0.13	0.21	0.14	0.12	0.2	0.18	0.14	0.15
Human3	0.06	0.06	0	0.12	0.13	0.13	0.21	0.14	0.12	0.2	0.18	0.14	0.15
Ours(PS)	0.12	0.12	0.12	0	0.07	0.06	0.15	0.1	0.08	0.17	0.13	0.07	0.09
WarpC(PS)	0.13	0.13	0.13	0.07	0	0.08	0.17	0.07	0.1	0.18	0.14	0.09	0.11
Weakalign(PS)	0.13	0.13	0.13	0.06	0.08	0	0.15	0.11	0.08	0.18	0.14	0.07	0.09
CNNGeo(PS)	0.21	0.21	0.21	0.15	0.17	0.15	0	0.2	0.14	0.24	0.18	0.12	0.1
WarpC(PF)	0.14	0.14	0.14	0.1	0.07	0.11	0.2	0	0.12	0.2	0.16	0.12	0.14
PMD(PF)	0.12	0.12	0.12	0.08	0.1	0.08	0.14	0.12	0	0.18	0.14	0.09	0.1
DCCNet(PF)	0.2	0.2	0.2	0.17	0.18	0.18	0.24	0.2	0.18	0	0.2	0.18	0.2
NCNet(PF)	0.18	0.18	0.18	0.13	0.14	0.14	0.18	0.16	0.14	0.2	0	0.13	0.14
Weakalign(PF)	0.14	0.14	0.14	0.07	0.09	0.07	0.12	0.12	0.09	0.18	0.13	0	0.06
CNNGeo(PF)	0.15	0.15	0.15	0.09	0.11	0.09	0.1	0.14	0.1	0.2	0.14	0.06	0

Figure 5: Measuring human and model consistency. Each cell represents the mean pixel distance between correspondence predictions generated by two systems (whether artificial or human), normalized by the image size. We denote models trained on Photo-sketch pairs with PS, and models trained on PF-Pascal (Ham et al., 2016) as PF.

4.5. Shape Bias in Learned Representation

Recent work has shown that ImageNet-trained CNNs are biased towards object texture compared to global object shape on image recognition tasks (Geirhos et al., 2018). Since sketch recognition requires relies on cues to object category apart from texture, we hypothesized that our photo-sketch contrastive learning pre-training procedure would mitigate this texture bias. To evaluate this hypothesis, we followed the same evaluation protocol as in (Geirhos et al., 2018; 2021). It devises a cue-conflict experiment in which a model aims to classify images with conflicting shape and texture. We report the shape bias of ResNet-18 models from several different training objectives: ImageNet classification (20.06%), ImageNet contrastive learning (28.93%), photo-sketch contrastive learning (46.36%), and the result of human participants (95.04%). The model trained on photo-sketch contrastive learning exhibits a reliably weaker texture bias (i.e., and thus stronger shape bias) than its photo-only counterparts (Figure 6).

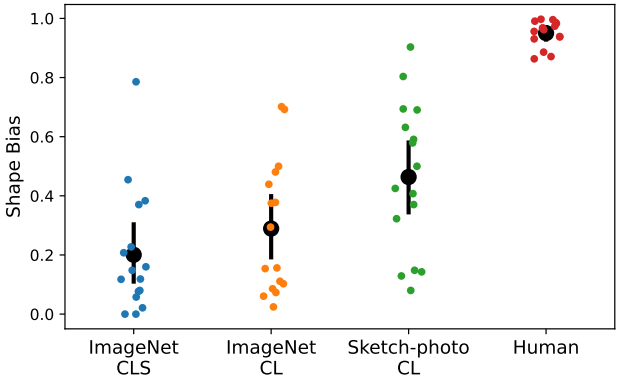

Figure 6: Comparing the degree of shape vs. texture bias between models trained with different objectives. Higher values suggest that the model recognition depends more on shape information. Our model exhibits more human-like performance. Each dot represents an object category from (Geirhos et al., 2018). Error bars indicate 95% CI.

5. Related Work

Self-supervised Representation Learning. Learning with self-supervision aims to obtain generic representations for diverse downstream tasks with minimal dependence on human labels (Wang & Gupta, 2015; Doersch et al., 2015; Pathak et al., 2016; Noroozi & Favaro, 2016; Zhang et al., 2016; Gidaris et al., 2018; Wu et al., 2018). Recent research on sketch understanding also benefits from such development (Pang et al., 2020; Xu et al., 2020; Bhunia et al., 2021). These approaches are especially important for making progress towards human-like image understanding, given that large numbers of labeled images are neither available to nor necessary for humans to develop robust perceptual abilities (Zhuang et al., 2021; Konkle & Alvarez, 2020; Rajalingham et al., 2018), including the ability to understand sketches (Hochberg & Brooks, 1962; Kennedy & Ross, 1975). In particular, recently proposed *contrastive learning* techniques demonstrate competitive performance with supervised baselines not only on visual recognition (Hjelm et al., 2018; Oord et al., 2018; Wu et al., 2018; Chen et al., 2020a; He et al., 2020; Grill et al., 2020; Chen et al., 2020b; 2021), but also on learning visual representations from inputs varying across sensory views (Tian et al., 2020a;b), across frames in video (Jabri et al., 2020; Xu & Wang, 2021; Zhuang et al., 2020), and even between text and images (Radford et al., 2021; Jia et al., 2021).

Here, we leverage contrastive learning-based pretraining to achieve strong performance on visual correspondence between images from highly distinct distributions (i.e., photos and sketches). To the best of our knowledge, ours is the first paper to successfully apply these approaches to the problem of photo-sketch dense correspondence prediction.

Weakly-supervised Semantic Correspondence Learning. Geometric matching (Melekhov et al., 2019; Li et al., 2020; Rocco et al., 2020; Shen et al., 2020; Truong et al., 2020) is perhaps the most basic form of correspondence prediction, which aims to align two views of the same scene. On the contrary, *semantic matching* (Ham et al., 2016; Rocco et al., 2018a;b; Huang et al., 2019; Li et al., 2021; Truong et al., 2021) aims to establish more abstract correspondences between the image of objects in the same class, in a way that is tolerant to greater variation in appearance and shape. Due to difficulties in collecting ground-truth data for dense correspondence learning, prior work has generally resorted to weak supervision, such as synthetic transformation on single images (Rocco et al., 2018a; Jeon et al., 2018; Seo et al., 2018) and image pairs (Rocco et al., 2018b; Kim et al., 2019; 2018; Jeon et al., 2020; Huang et al., 2019; Li et al., 2021; Truong et al., 2021; 2022). Various objectives have been proposed to explore the correspondence from weak supervision, including synthetic supervision, optimization of the cost volume, forward-backward consistency, or a combination of these objectives. Most work utilizes hierarchical features in deep models from supervised pretraining on ImageNet. The dense correspondence is then predicted with a dense flow field (Ham et al., 2016; Rocco et al., 2018a; Jeon et al., 2018; Seo et al., 2018; Li et al., 2021; Truong et al., 2021) or a cost volume (Rocco et al., 2018b; Huang et al., 2019; Truong et al., 2022). In this work, we propose a photo-sketch correspondence learning framework that explicitly estimates the dense flow field with image pair supervision.

6. Conclusions

What is needed to develop artificial systems that learn to perceive the visual world as keenly as humans do? While artificial vision systems have made dramatic improvements in a variety of tasks, there remain key aspects of human image understanding that continue to pose major challenges. Here we focused on one of these aspects: the ability to understand the semantic content of color photos and line drawings well enough to establish a detailed mapping between them. Our paper introduces a new photo-sketch correspondence benchmark containing 150K human annotations of 6250 sketch-photo pairs across 125 object categories, augmenting existing photo-sketch benchmark datasets (Sangkloy et al., 2016). In addition, we conduct several experiments to evaluate a self-supervised approach to learning to predict these correspondences and compare this approach to several strong correspondence learning baselines. Our results suggest that our approach combining contrastive learning and spatial transformer network is effective for capturing photo-sketch correspondences, but there remain systematic deviations from human judgments on the same task. Taken together, we hope that these findings, along with our new fine-grained multimodal image understanding benchmark, will catalyze progress towards achieving more human-like vision systems.

Acknowledgements

Many thanks to the members of the Cognitive Tools Lab and the Prof. Wang's Lab at UC San Diego for their helpful feedback and support. This work was supported by an NSF CAREER Award #2047191 to J.E.F.. J.E.F is additionally supported by an ONR Science of Autonomy award and a Stanford Hoffman-Yee grant. Prof. Wang's lab was supported, in part, by NSF CAREER Award IIS-2240014, DARPA LwLL, Amazon Research Award, and gifts from Qualcomm.

References

Aubert, M., Brumm, A., Ramli, M., Sutikna, T., Saptomo, E. W., Hakim, B., Morwood, M. J., van den Bergh, G. D., Kinsley, L., and Dosseto, A. Pleistocene cave art from sulawesi, indonesia. *Nature*, 514(7521):223–227, 2014.

Bagus, A. M. I. G., Marques, T., Sanghavi, S., DiCarlo, J. J., and Schrimpf, M. Primate inferotemporal cortex neurons generalize better to novel image distributions than analogous deep neural networks units. In *SVRHM 2022 Workshop@ NeurIPS*.

Berg, A. C., Berg, T. L., and Malik, J. Shape matching and object recognition using low distortion correspondences. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pp. 26–33. IEEE, 2005.

Bhunia, A. K., Yang, Y., Hospedales, T. M., Xiang, T., and Song, Y.-Z. Sketch less for more: On-the-fly fine-grained sketch-based image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9779–9788, 2020.

Bhunia, A. K., Chowdhury, P. N., Yang, Y., Hospedales, T. M., Xiang, T., and Song, Y.-Z. Vectorization and rasterization: Self-supervised learning for sketch and handwriting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5672–5681, 2021.

Cadena, S. A., Denfield, G. H., Walker, E. Y., Gatys, L. A., Tolias, A. S., Bethge, M., and Ecker, A. S. Deep con-

volutional models improve predictions of macaque v1 responses to natural images. *PLoS computational biology*, 15(4):e1006897, 2019.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020a.

Chen, T., Kornblith, S., Swersky, K., Norouzi, M., and Hinton, G. E. Big self-supervised models are strong semi-supervised learners. *Advances in neural information processing systems*, 33:22243–22255, 2020b.

Chen, X., Fan, H., Girshick, R., and He, K. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020c.

Chen, X., Xie, S., and He, K. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9640–9649, 2021.

Cho, S., Hong, S., Jeon, S., Lee, Y., Sohn, K., and Kim, S. Cats: Cost aggregation transformers for visual correspondence. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., and Schiele, B. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Dalal, N. and Triggs, B. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pp. 886–893. Ieee, 2005.

De Vries, H., Strub, F., Mary, J., Larochelle, H., Pietquin, O., and Courville, A. C. Modulating early visual processing by language. *Advances in Neural Information Processing Systems*, 30, 2017.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Doersch, C., Gupta, A., and Efros, A. A. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE international conference on computer vision*, pp. 1422–1430, 2015.

Eitz, M., Richter, R., Boubekeur, T., Hildebrand, K., and Alexa, M. Sketch-based shape retrieval. *ACM Transactions on graphics (TOG)*, 31(4):1–10, 2012.

Fan, J. E., Yamins, D. L., and Turk-Browne, N. B. Common object representations for visual production and recognition. *Cognitive science*, 42(8):2670–2698, 2018.

Fan, J. E., Hawkins, R. D., Wu, M., and Goodman, N. D. Pragmatic inference and visual abstraction enable contextual flexibility during visual communication. *Computational Brain & Behavior*, 3(1):86–101, 2020.

Fodor, J. The revenge of the given. *Contemporary debates in philosophy of mind*, pp. 105–116, 2007.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., and Brendel, W. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018.

Geirhos, R., Narayanappa, K., Mitzkus, B., Thieringer, T., Bethge, M., Wichmann, F. A., and Brendel, W. Partial success in closing the gap between human and machine vision. *Advances in Neural Information Processing Systems*, 34:23885–23899, 2021.

Gidaris, S., Singh, P., and Komodakis, N. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018.

Goodman, N. *Languages of art: An approach to a theory of symbols*. Hackett publishing, 1976.

Greenberg, G. Semantics of pictorial space. *Review of Philosophy and Psychology*, 12(4):847–887, 2021.

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

Ham, B., Cho, M., Schmid, C., and Ponce, J. Proposal flow. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3475–3484, 2016.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.

Hertzmann, A. Why do line drawings work? a realism hypothesis. *Perception*, 49(4):439–451, 2020.

Hjelm, R. D., Fedorov, A., Lavoie-Marchildon, S., Grewal, K., Bachman, P., Trischler, A., and Bengio, Y. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018.

Hochberg, J. and Brooks, V. Pictorial recognition as an unlearned ability: A study of one child's performance. *the american Journal of Psychology*, 75(4):624–628, 1962.

Hoffmann, D. L., Standish, C. D., García-Diez, M., Pettitt, P. B., Milton, J. A., Zilhão, J., Alcolea-González, J. J., Cantalejo-Duarte, P., Collado, H., de Balbín, R., Lorblanchet, M., Ramos-Muñoz, J., Weniger, G.-C., and Pike, A. W. G. U-th dating of carbonate crusts reveals neandertal origin of iberian cave art. *Science*, 359(6378):912–915, 2018. doi: 10.1126/science.aap7778. URL https://www.science.org/doi/abs/10.1126/science.aap7778.

Huang, S., Wang, Q., Zhang, S., Yan, S., and He, X. Dynamic context correspondence network for semantic alignment. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2010–2019, 2019.

Huey, H., Lu, X., Walker, C., and Fan, J. Explanatory drawings prioritize functional properties at the expense of visual fidelity. 2021.

Ignatov, A., Kobyshev, N., Timofte, R., Vanhoey, K., and Van Gool, L. Dslr-quality photos on mobile devices with deep convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3277–3285, 2017.

Ioffe, S. and Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pp. 448–456. PMLR, 2015.

Jabri, A., Owens, A., and Efros, A. Space-time correspondence as a contrastive random walk. *Advances in neural information processing systems*, 33:19545–19560, 2020.

Jaderberg, M., Simonyan, K., Zisserman, A., et al. Spatial transformer networks. *Advances in neural information processing systems*, 28, 2015.

Jeon, S., Kim, S., Min, D., and Sohn, K. Parn: Pyramidal affine regression networks for dense semantic correspondence. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 351–366, 2018.

Jeon, S., Min, D., Kim, S., Choe, J., and Sohn, K. Guided semantic flow. In *European Conference on Computer Vision*, pp. 631–648. Springer, 2020.

Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q., Sung, Y.-H., Li, Z., and Duerig, T. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pp. 4904–4916. PMLR, 2021.

Kennedy, J. M. and Ross, A. S. Outline picture perception by the songe of papua. *Perception*, 4(4):391–406, 1975.

Khaligh-Razavi, S.-M. and Kriegeskorte, N. Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS computational biology*, 10(11): e1003915, 2014.

Kim, J., Liu, C., Sha, F., and Grauman, K. Deformable spatial pyramid matching for fast dense correspondences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2307–2314, 2013.

Kim, S., Lin, S., Jeon, S. R., Min, D., and Sohn, K. Recurrent transformer networks for semantic correspondence. *Advances in neural information processing systems*, 31, 2018.

Kim, S., Min, D., Jeong, S., Kim, S., Jeon, S., and Sohn, K. Semantic attribute matching networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12339–12348, 2019.

Konkle, T. and Alvarez, G. A. Instance-level contrastive learning yields human brain-like representation without category-supervision. *BioRxiv*, 2020.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

Kulvicki, J. Analog representation and the parts principle. *Review of Philosophy and Psychology*, 6(1):165–180, 2015.

Li, X., Han, K., Li, S., and Prisacariu, V. Dual-resolution correspondence networks. *Advances in Neural Information Processing Systems*, 33:17346–17357, 2020.

Li, X., Fan, D.-P., Yang, F., Luo, A., Cheng, H., and Liu, Z. Probabilistic model distillation for semantic correspondence. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7501–7510, 2021.

Liu, C., Yuen, J., and Torralba, A. Sift flow: Dense correspondence across scenes and its applications. *IEEE transactions on pattern analysis and machine intelligence*, 33 (5):978–994, 2010.

Loshchilov, I. and Hutter, F. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

Lowe, D. G. Object recognition from local scale-invariant features. In *Proceedings of the seventh IEEE international conference on computer vision*, volume 2, pp. 1150–1157. Ieee, 1999.

Meister, S., Hur, J., and Roth, S. Unflow: Unsupervised learning of optical flow with a bidirectional census loss. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Melekhov, I., Tiulpin, A., Sattler, T., Pollefeys, M., Rahtu, E., and Kannala, J. Dgc-net: Dense geometric correspondence network. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1034–1042. IEEE, 2019.

Min, J. and Cho, M. Convolutional hough matching networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2940–2950, 2021.

Min, J., Lee, J., Ponce, J., and Cho, M. Hyperpixel flow: Semantic correspondence with multi-layer neural features. In *ICCV*, 2019.

Mukherjee, K., Hawkins, R. X., and Fan, J. W. Communicating semantic part information in drawings. In *CogSci*, pp. 2413–2419, 2019.

Noroozi, M. and Favaro, P. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European conference on computer vision*, pp. 69–84. Springer, 2016.

Oord, A. v. d., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Pang, K., Yang, Y., Hospedales, T. M., Xiang, T., and Song, Y.-Z. Solving mixed-modal jigsaw puzzle for fine-grained sketch-based image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10347–10355, 2020.

Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., and Efros, A. A. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2536–2544, 2016.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.

Rajalingham, R., Issa, E. B., Bashivan, P., Kar, K., Schmidt, K., and DiCarlo, J. J. Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks. *Journal of Neuroscience*, 38(33):7255–7269, 2018.

Rocco, I., Arandjelovic, R., and Sivic, J. Convolutional neural network architecture for geometric matching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6148–6157, 2017.

Rocco, I., Arandjelović, R., and Sivic, J. End-to-end weakly-supervised semantic alignment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6917–6925, 2018a.

Rocco, I., Cimpoi, M., Arandjelović, R., Torii, A., Pajdla, T., and Sivic, J. Neighbourhood consensus networks. *Advances in neural information processing systems*, 31, 2018b.

Rocco, I., Arandjelović, R., and Sivic, J. Efficient neighbourhood consensus networks via submanifold sparse convolutions. In *European conference on computer vision*, pp. 605–621. Springer, 2020.

Sangkloy, P., Burnell, N., Ham, C., and Hays, J. The sketchy database: learning to retrieve badly drawn bunnies. *ACM Transactions on Graphics (TOG)*, 35(4):1–12, 2016.

Seo, P. H., Lee, J., Jung, D., Han, B., and Cho, M. Attentive semantic alignment with offset-aware correlation kernels. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 349–364, 2018.

Shen, X., Darmon, F., Efros, A. A., and Aubry, M. Ransac-flow: generic two-stage image alignment. In *European Conference on Computer Vision*, pp. 618–637. Springer, 2020.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Sun, D., Yang, X., Liu, M.-Y., and Kautz, J. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8934–8943, 2018.

Tian, Y., Krishnan, D., and Isola, P. Contrastive multiview coding. In *European conference on computer vision*, pp. 776–794. Springer, 2020a.

Tian, Y., Sun, C., Poole, B., Krishnan, D., Schmid, C., and Isola, P. What makes for good views for contrastive learning? *Advances in Neural Information Processing Systems*, 33:6827–6839, 2020b.

Truong, P., Danelljan, M., and Timofte, R. Glu-net: Global-local universal network for dense flow and correspondences. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6258–6268, 2020.

Truong, P., Danelljan, M., Yu, F., and Van Gool, L. Warp consistency for unsupervised learning of dense correspondences. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10346–10356, 2021.

Truong, P., Danelljan, M., Yu, F., and Van Gool, L. Probabilistic warp consistency for weakly-supervised semantic correspondences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8708–8718, 2022.

Vinker, Y., Pajouheshgar, E., Bo, J. Y., Bachmann, R. C., Bermano, A. H., Cohen-Or, D., Zamir, A., and Shamir, A. Clipasso: Semantically-aware object sketching. *arXiv preprint arXiv:2202.05822*, 2022.

Vondrick, C., Shrivastava, A., Fathi, A., Guadarrama, S., and Murphy, K. Tracking emerges by colorizing videos. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 391–408, 2018.

Wang, X. and Gupta, A. Unsupervised learning of visual representations using videos. In *Proceedings of the IEEE international conference on computer vision*, pp. 2794–2802, 2015.

Wang, X., Jabri, A., and Efros, A. A. Learning correspondence from the cycle-consistency of time. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2566–2576, 2019.

Wu, Z., Xiong, Y., Yu, S. X., and Lin, D. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3733–3742, 2018.

Xu, J. and Wang, X. Rethinking self-supervised correspondence learning: A video frame-level similarity perspective. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10075–10085, 2021.

Xu, P., Song, Z., Yin, Q., Song, Y.-Z., and Wang, L. Deep self-supervised representation learning for free-hand sketch. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(4):1503–1513, 2020.

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., and DiCarlo, J. J. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the national academy of sciences*, 111(23):8619–8624, 2014.

Yang, J. and Fan, J. E. Visual communication of object concepts at different levels of abstraction. *arXiv preprint arXiv:2106.02775*, 2021.

Yu, Q., Liu, F., Song, Y.-Z., Xiang, T., Hospedales, T. M., and Loy, C.-C. Sketch me that shoe. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 799–807, 2016.

Yu, Q., Yang, Y., Liu, F., Song, Y.-Z., Xiang, T., and Hospedales, T. M. Sketch-a-net: A deep neural network that beats humans. *International journal of computer vision*, 122(3):411–425, 2017.

Zhang, R., Isola, P., and Efros, A. A. Colorful image colorization. In *European conference on computer vision*, pp. 649–666. Springer, 2016.

Zhou, B., Zhao, H., Puig, X., Xiao, T., Fidler, S., Barriuso, A., and Torralba, A. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127(3):302–321, 2019.

Zhuang, C., She, T., Andonian, A., Mark, M. S., and Yamins, D. Unsupervised learning from video with deep neural embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9563–9572, 2020.

Zhuang, C., Yan, S., Nayebi, A., Schrimpf, M., Frank, M. C., DiCarlo, J. J., and Yamins, D. L. Unsupervised neural network models of the ventral visual stream. *Proceedings of the National Academy of Sciences*, 118(3):e2014196118, 2021.

A. Details of the Photo-Sketch Correspondence Benchmark (PSC6k)

A.1. Keypoint Sampling

We visualize the steps we take to sample eight keypoints spanning the object in Figure 7. First, we fill in the outermost contour detected in the sketch to generate the segmentation of the object. In cases where multiple contours are detected due to unconnected strokes, we apply dilation and contour filling iteratively until all strokes are connected. We then cluster the pixels covered by the segmentation mask into 8 pseudo-parts, by building a nearest-neighbor-based affinity matrix over pixels and applying spectral clustering. Since the affinity between two pixels is defined by the shortest path instead of the L2 distance, it ensures a clustering that maintains the connectivity within each pseudo-part.

Figure 7: Example of the keypoint sampling process. We show the sketch, segmentation mask, pseudo-parts, and final keypoints.

A.2. Annotation Filtering

In rare cases, for a given keypoint, one of the three annotations has an exceptionally large distance from the median location $\tilde{x}$ of all three annotations, denoted as $d = \|x - \tilde{x}\|_2^2$. We gather the distance d for the 150,000 annotations that we collect and compute its mean and standard deviation. The annotations with d of three standard deviations away from the mean are then considered outliers and excluded from the final determination of the centroids. This rejects 0.74% of the annotations.

B. Additional Evaluation on PSC6k

B.1. Methods with Stronger Supervision

For a more comprehensive evaluation of existing correspondence learning methods on our PSC6k benchmark, we include methods with keypoint supervision in Table 4 and report their PCK for $\alpha = (0.05, 0.1)$. We report the performance of keypoint-supervised models in the transfer setting only (directly evaluate on the photo-sketch correspondence with pretrained weights), because they require supervision beyond what the sketchy training set provides. Interestingly, we observe that CATs (Cho et al., 2021) performs exceptionally well on the photo-sketch correspondence, even without retraining on photo-sketch pairs. This suggests its good ability of generalization.

B.2. Training and Evaluation Details

Evaluation setting. All methods are evaluated on our PSC6k benchmark using their original evaluation scripts. We make necessary edits to adapt the existing codes to PSC6k.

Training setting. In the transfer setting, we use the pretrained weights on PF-Pascal provided by each method. In the retrain setting, we train the methods on the training split of the Sketchy dataset (Sangkloy et al., 2016) using their codes and default hyperparameters. Since there is no validation split, we do not select the best checkpoint and evaluate with the last checkpoint after training.

Since the training set of Sketchy dataset is 88X larger than that of PF-Pascal, it is impossible to keep the original training epochs and learning rate schedule in large models such as WarpC. Therefore, we make the following changes to the series: instead of training 100 epochs as in the original settings, we find that training for 2 epochs on Sketchy has already guaranteed an optimal performance (since it leads to $1.77\times$ iterations compared to the original training scheme). We reduce the learning rate to $0.125\times$ in the second epoch to approximate the original LR schedule of the method.

Causes of blank entries. The retrain performance of several methods are left blank for the following reasons:

- The method requires stronger supervision than what the Sketchy training set provides. This applies to all methods with keypoint supervision.
- The method fails to converge on the photo-sketch correspondence task, which is observed in NC-Net and DCCNet. We hypothesize that since the sketch samples are out-of-domain, their cost volume optimization blocks fail to handle the large disparity between the representations of photos and sketches.
- The method does not provide codes for training: PMD.
- In addition, methods that did not release source code, failed to execute, or did not provide pre-trained weights are excluded from the table.

C. Generalization to Unseen Categories

To analyze the generalization capability of our proposed model, we evaluate its performance on categories that were not included during the training phase. Specifically, we randomly sample N categories from the full set of 125 categories in the Sketchy dataset, and hold them out during the training of both the feature encoder and warp estimator. Then we evaluate the model performance of correspondence estimation on these N held-out categories. We conduct experiments for N=10 and N=20. The mean performance and standard deviation were calculated based on three randomly

Sup	Methods	Transfer PCK-5	Transfer PCK-10	Retrain PCK-5	Retrain PCK-10
KP	HPF(Min et al., 2019)	50.55	78.18	–	–
	CHM (Min & Cho, 2021)	40.52	69.91	–	–
	PMD(Li et al., 2021)	28.62	63.95	–	–
	CATs(Cho et al., 2021)	**52.36**	**81.80**	–	–
Pair	CNNGeo (Rocco et al., 2018a)	27.59	57.71	19.19	42.57
	WeakAlign (Rocco et al., 2018a)	35.65	68.76	43.55	78.60
	NC-Net (Rocco et al., 2018b)	40.60	63.50	–	–
	DCCNet (Huang et al., 2019)	42.43	66.53	–	–
	PMD (Li et al., 2021)	35.77	71.24	–	–
	WarpC-SemanticGLUNet (Truong et al., 2021)	48.79	71.43	56.78	79.70
	Ours (ResNet-18)	–	–	56.01	**82.89**
	Ours (ResNet-101)	–	–	**57.92**	**84.72**

Table 4: Comprehensive evaluation for photo-sketch correspondence learning.

sampled held-out splits for each of the two conditions. The results are presented in Table 5.

As shown in the table, our method maintains a very decent performance on the 10/20 categories absent during training, with a decrease of -0.33%/-0.26% for 10 held-out categories and a decrease of -0.50%/-0.37% for 20 held-out categories. This shows that our method is robust in generalization to unseen categories.

# Categories (N)	PCK-5 ($\pm std$)	PCK-10 ($\pm std$)
0	56.01	82.89
10	55.68 (0.20)	82.63 (0.15)
20	55.51 (0.27)	82.52 (0.18)

Table 5: Model performance on unseen categories.

D. Additional Qualitative Results

We show typical failure patterns in Figure 8. Specifically, the model has degraded performance in 1) discriminating between commonly cooccurred objects; 2) aligning fine structures due to low resolution; and 3) handling non-continuous transformation caused by large changes in perspective and structure, which violates the continuity assumption in warp-based models that close points should correspond to close locations. We believe that they are the main problems that need to be addressed in future studies.

Lastly, we exhibit more examples of photo-sketch correspondence predicted by our model (Figure 9, Figure 10, Figure 11).

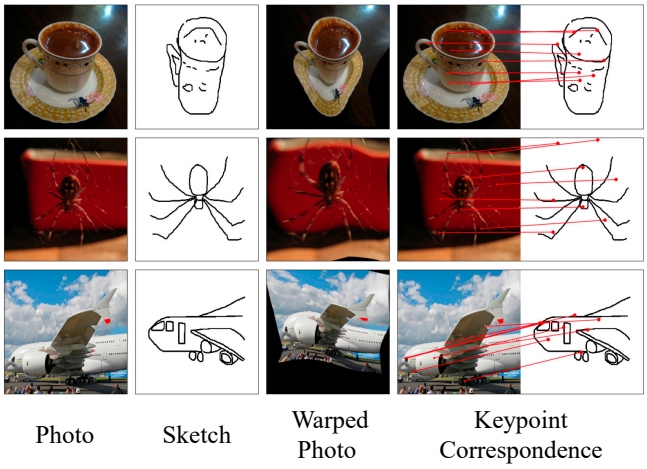

Figure 8: Examples of three typical failure patterns. The method has worse performance for: 1) commonly co-occurred objects, 2) fine structures, and 3) non-continuous transformations.

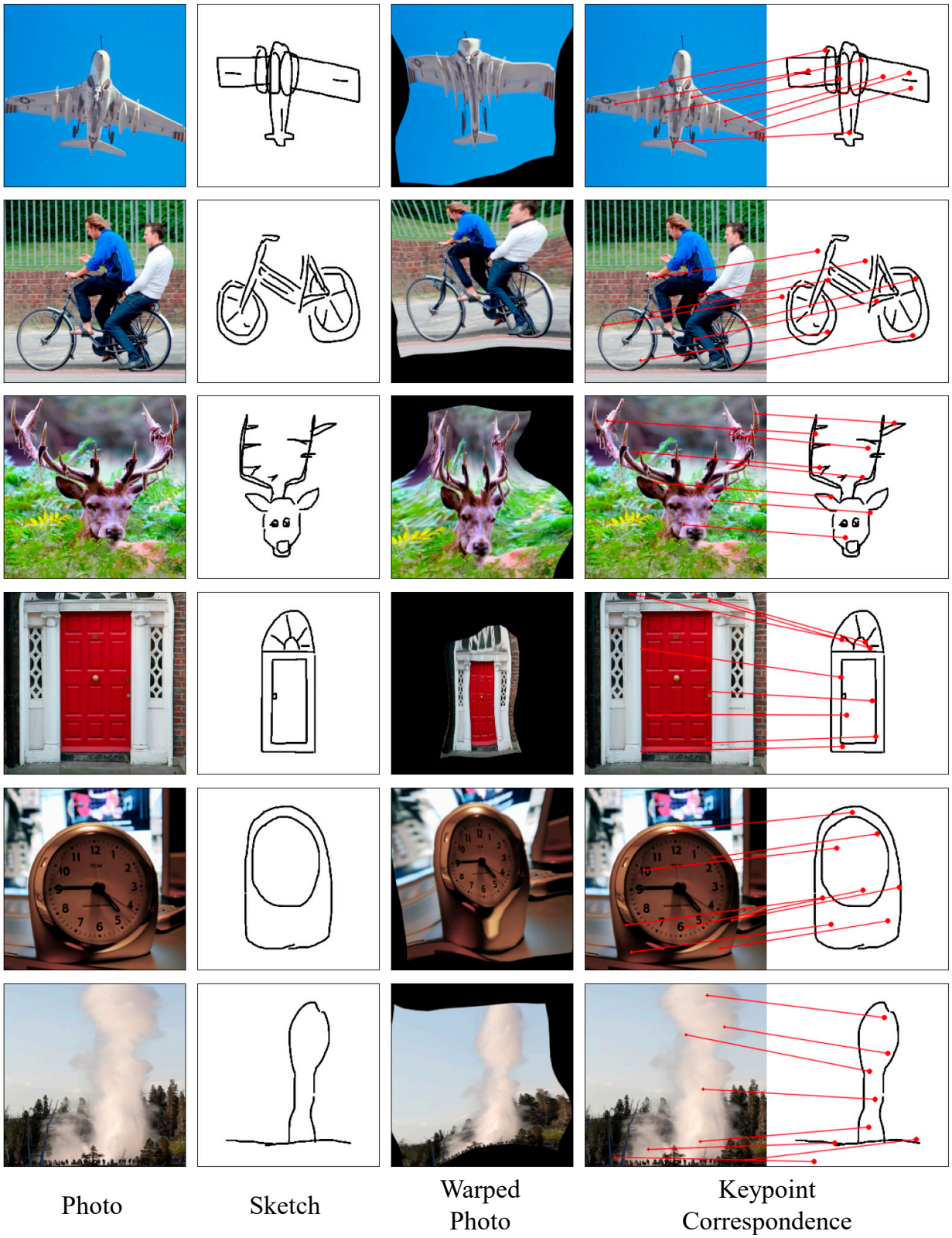

Figure 9: More alignment examples on the PSC6k dataset.

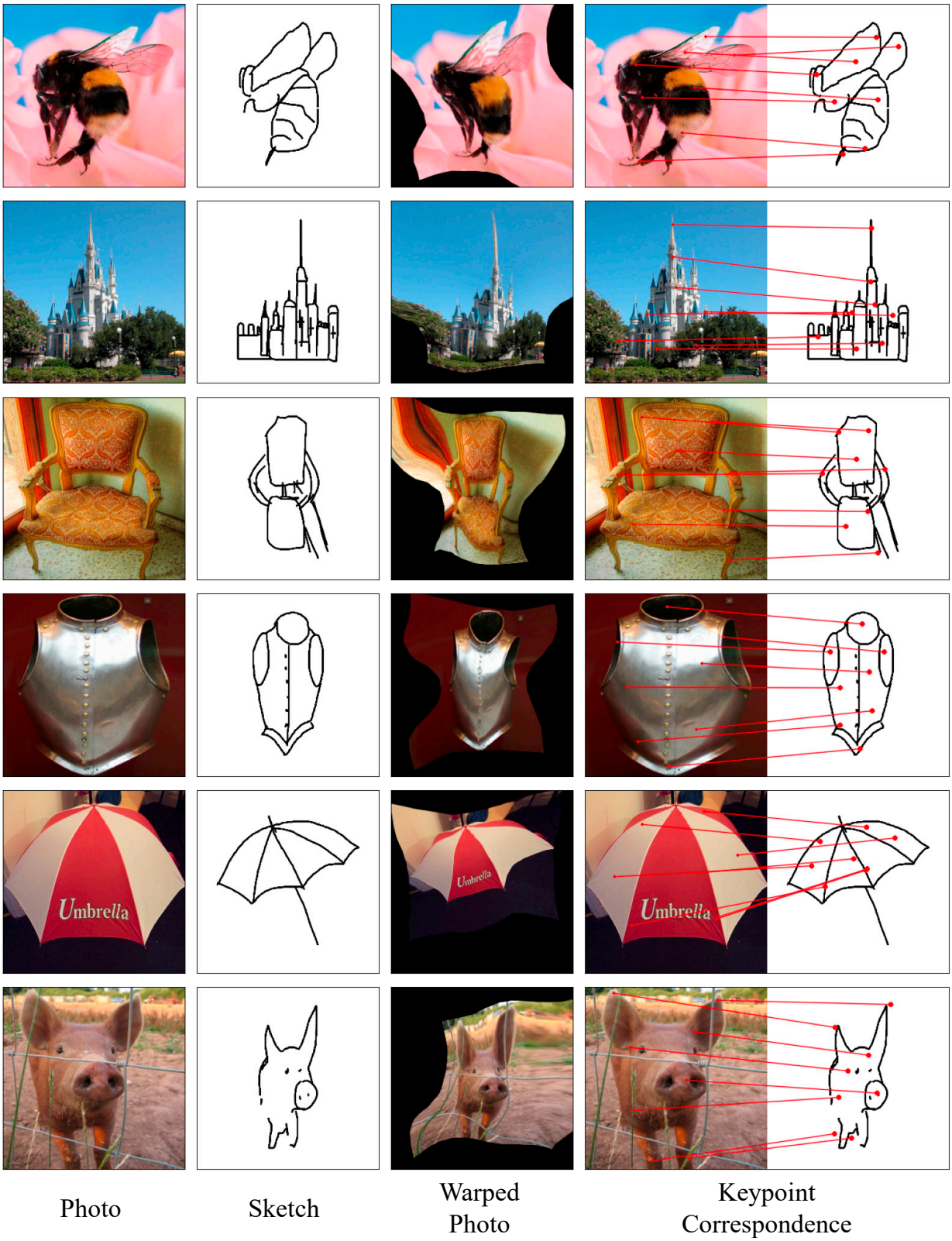

Figure 10: More alignment examples on the PSC6k dataset.

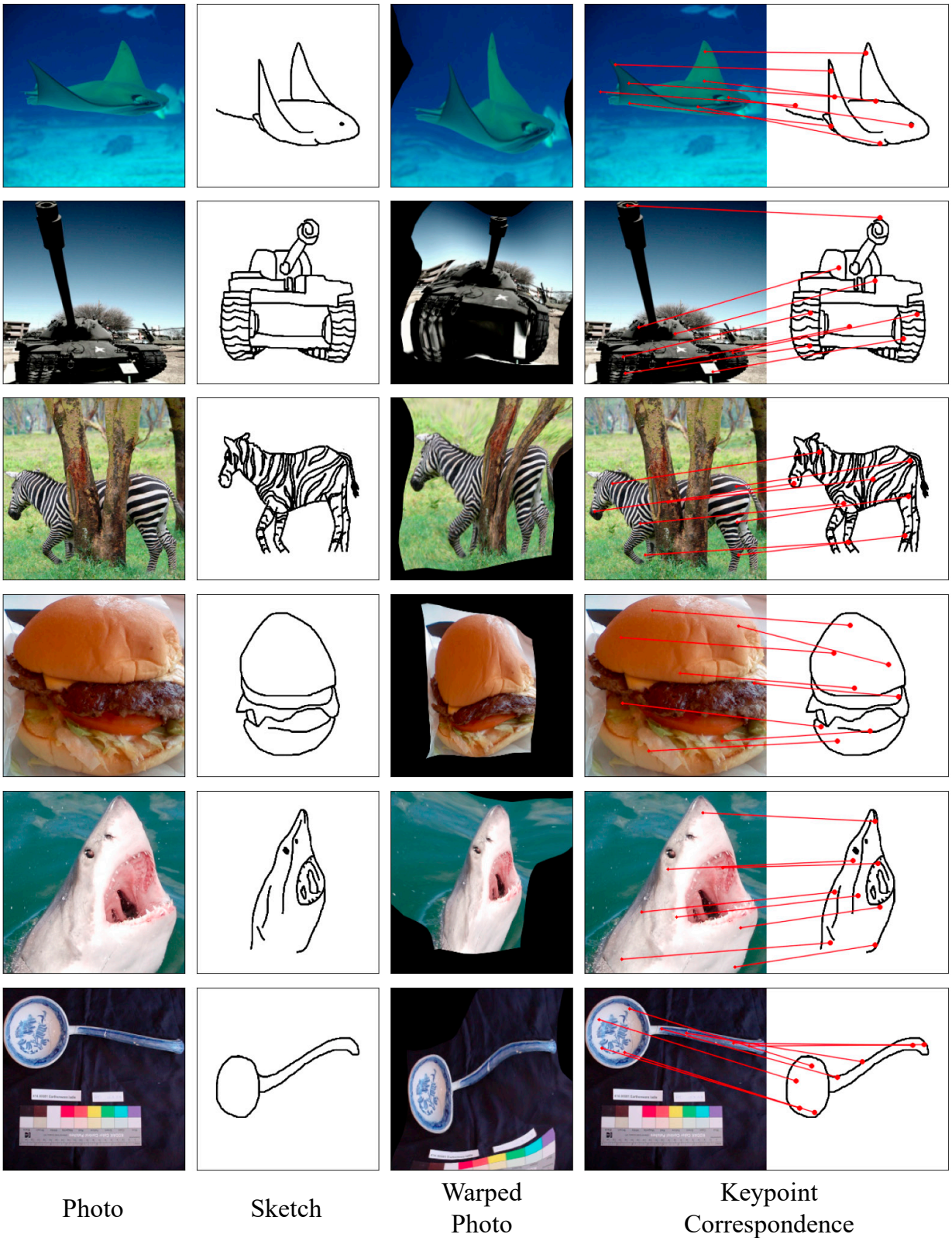

Figure 11: More alignment examples on the PSC6k dataset.