UNIVERSITY OF CALIFORNIA SAN DIEGO

The photo-sketch correspondence problem: a new benchmark and a self-supervised approach

A thesis submitted in partial satisfaction of the
requirements for the degree Master of Science

in

Computer Science

by

Xuanchen Lu

Committee in charge:

Professor Xiaolong Wang, Chair
Professor Hao Su, Co-Chair
Professor Manmohan Chandraker

2022

The Thesis of Xuanchen Lu is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2022

TABLE OF CONTENTS

Thesis Approval Page iii

Table of Contents iv

List of Figures v

List of Tables vi

Acknowledgements vii

Abstract of the Thesis viii

Introduction 1

Chapter 1 Photo-Sketch Correspondence Benchmark 5
1.1 Sampling Photo-Sketch Pairs 5
1.2 Collecting Human Keypoint Annotations 5

Chapter 2 Weakly-supervised Photo-Sketch Correspondence 8
2.1 Feature Encoder ϕ 9
2.2 Warp Estimator T 10
2.3 Weighted Perceptual Similarity 11
2.4 Additional Objectives 12

Chapter 3 Experiments and Results 15
3.1 Implementation Details 15
3.2 Photo-sketch Correspondence Estimation 16
3.3 Ablation Study 17
3.4 Comparing model and human error patterns 18
3.5 Shape Bias in Learned Representation 20

Chapter 4 Related Work 21
4.1 Self-supervised Representation Learning 21
4.2 Weakly-supervised Semantic Correspondence Learning 21

Chapter 5 Conclusions 23

Bibliography 27

LIST OF FIGURES

Figure 1.1. Examples of human-annotated photo-sketch pairs from our new correspondence benchmark. 7

Figure 2.1. The self-supervised framework for learning photo-sketch correspondence by estimating a dense flow that warps one image to the other. 8

Figure 2.2. Example image pairs, feature maps, weight maps, and final results processed in our warp estimator. 13

Figure 3.1. Measuring human and model consistency. 19

Figure 3.2. Comparing the degree of shape vs. texture bias between models trained with different objectives. 20

Figure 5.1. More alignment examples on the PSC6K dataset. 24

Figure 5.2. More alignment examples on the PSC6K dataset. 25

Figure 5.3. More alignment examples on the PSC6K dataset. 26

LIST OF TABLES

Table 3.1. State-of-the-art comparison for photo-sketch correspondence learning. 16

Table 3.2. Ablation study on the feature encoder training. 18

Table 3.3. Ablation study on correspondence estimation. 18

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my advisors, Professor Judith Fan and Professor Xiaolong Wang. Their invaluable guidance shapes my academic curiosity and develops my general view of open problems in the intersection of human cognition and computer vision.

I would like to acknowledge Professor Hao Su and Professor Manmohan Chandraker for their help with my thesis review.

I also want to give thanks to Jiarui Xu, Justin Yang, Holly Huey, and other lab members for their brilliant insights and kind help during my study.

This thesis, in full, is currently being prepared for submission for publication of the material as it may appear in a conference, 2023, Xiaolong Wang, Judith Fan. The thesis author was the primary researcher and author of this material.

ABSTRACT OF THE THESIS

The photo-sketch correspondence problem:
a new benchmark and a self-supervised approach

by

Xuanchen Lu

Master of Science in Computer Science

University of California San Diego, 2022

Professor Xiaolong Wang, Chair
Professor Hao Su, Co-Chair

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 [67] 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-flow methods. However, our benchmark also revealed systematic biases shared by the suite of models we tested that are distinct from 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.

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 [30, 1]. 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 [17, 55, 85]. 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 [15, 67, 86, 87, 4]. 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 [60], including sketch generation [78]. 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) [2], 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 and

unsolved aspect of 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 [47, 18]. 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. Classic methods [3, 42, 50] determine the alignment with hand-crafted feature descriptors such as SIFT [52] or DOG [11]. More recently developed methods [24, 63, 76], 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 [16, 32]. 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 on large, labeled image datasets can achieve state-of-the-art performance [46, 70, 26]. Moreover, models trained in this way currently provide the most quantitatively accurate models of biological vision in non-human primates and humans [84, 41, 61, 5]. 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 [29, 40]. Recent advances in self-supervised representation learning have begun to approach the performance of supervised models without the need for such labels [82, 25], while also emulating key aspects of visual processing in biological systems [91, 45]. 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.

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, shown in Figure 1.1. 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 [67], 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 broadly consistent with the view that line drawings, as sparse as they are, are meant to accurately convey 3D shape [27], as opposed to the view that they are arbitrary arrangements of marks whose associations with objects are established purely by convention [22]. On the other hand, the nonlinear "warping" approach we propose diverges from the strongest version of the shape-based view, which is not well equipped to handle the kinds of visual distortions that human-generated sketches exhibit [15, 67, 17]. Our system consists of two main components: the first is a multimodal image encoder trained with a contrastive loss [82, 91], 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 [36] 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.

Chapter 1

Photo-Sketch Correspondence Benchmark

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 [67], which already contains 75,471 human sketches produced from 12,500 unique photographs spanning 125 object categories.

1.1 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 [67]. We then randomly sampled 5 sketches from among the remaining valid sketches produced of each photo, resulting in 6250 unique photo-sketch pairs.

1.2 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 for each pseudo-part. Next, we recruited 1384 participants using the Prolific crowdsourcing platform to provide annotations. Participants provided informed consent in accordance with the UC San Diego IRB. On each trial, participants were cued with a keypoint appearing on a sketch and indicated its corresponding location in a photo appearing next to it (Figure 1.1). 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; there were flagged as outliers and excluded from the determination of the centroid.

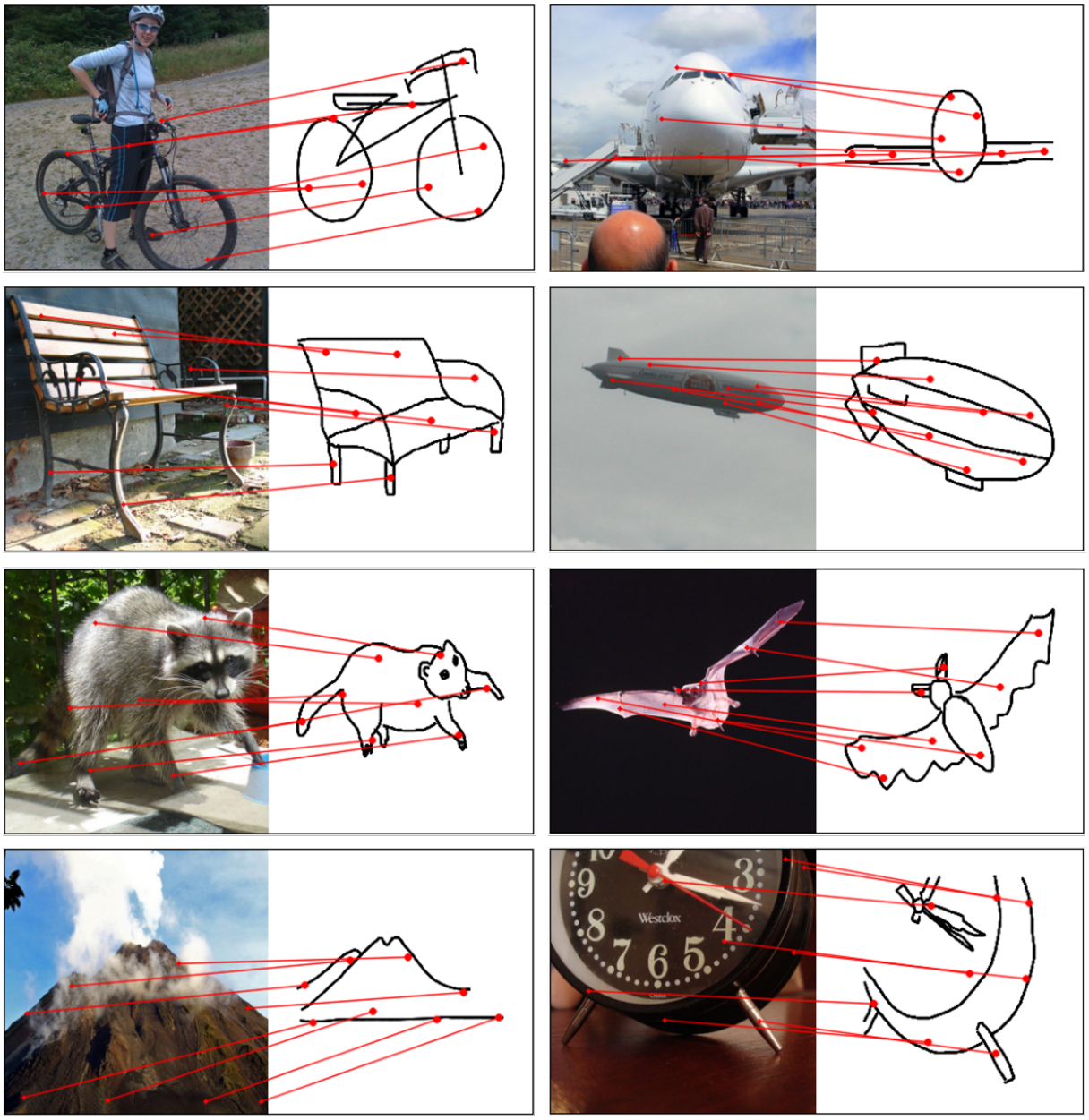

Figure 1.1. Examples of human-annotated photo-sketch pairs from our new correspondence benchmark.

Chapter 2

Weakly-supervised Photo-Sketch Correspondence

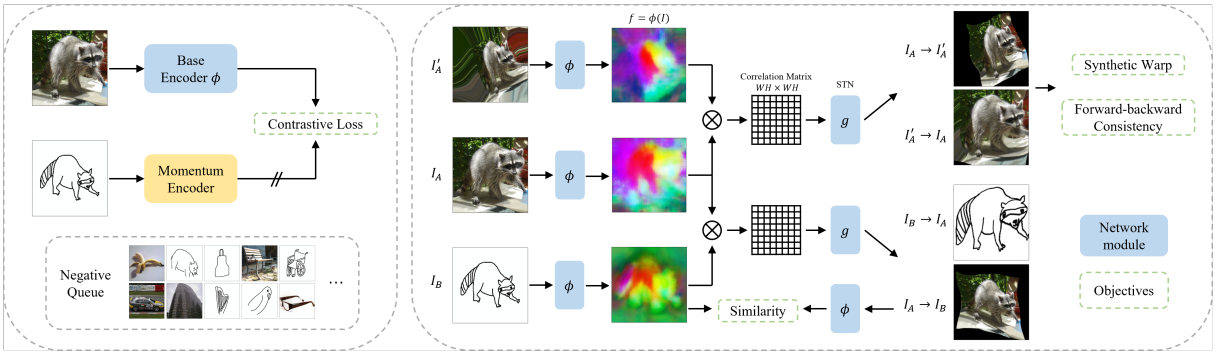

Figure 2.1. We propose a self-supervised framework for learning photo-sketch correspondence by estimating a dense flow that warps one image to the other. The framework consists of a multi-modal feature encoder that aligns the photo-sketch representation with a contrastive loss, and a STN-based warp estimator to predict transformation that maximizes the similarity between feature maps of the two images. The estimator learns to optimize a combination of similarity metric, forward-backward consistency, and the synthetic pseudo groundtruth flow.

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 an RGB photowea $I_p \in \mathbb{R}^{h \times w \times 3}$ that depict the same object (Figure 2.1). 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.

2.1 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 [82, 6, 25] 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 [25], 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[57] as follows:

$$\mathscr{L}_{nce} = -\log \frac{\exp(q \cdot k^+ / \tau)}{\exp(q \cdot k^+ / \tau) + \sum_{k^-} \exp(q \cdot k^- / \tau)}, \tag{2.1}$$

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 batch normalization (BN) [34] in the encoder with conditional batch normalization [12] for better domain alignment. Detailed implementation and experiment are reported in chapter 3.

2.2 Warp Estimator T

Given the representation of the source and target image from encoder, we estimate the displacement field with T as $F_{I_s \to I_t} = T(I_s, I_t)$. Inspired by [71], we propose a simplified pyramidal warp estimation module for ResNet backbone.

Affinity function f. We start by considering the affinity at a single layer. With the source and target image feature maps $x_s \in \mathbb{R}^{c \times h \times w}$ and $x_t \in \mathbb{R}^{c \times h \times w}$, we construct the pairwise affinity matrix as $A_{(s,t)} \in \mathbb{R}^{hw \times hw}$. 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, $A_{(s,t)}(i,j) = x_s(i)^T x_t(j)$.

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 $X_t = \{x_t^i\}_{i=0}^{n-1}$. Since the affinity matrix $A_{(s,t)}$ is irrelevant to the feature dimension, we compute the affinity on each of the selected layers, giving $\{A_{(s,t)}^i\}_{i=0}^{n-1}$. We bi-linearly upsample all selected feature maps to the same spatial resolution, so as to align the shape of affinity matrices. They are concatenated to obtain the final affinity matrix $A_{(s,t)}^* \in \mathbb{R}^{n \times hw \times hw}$.

Now, we formally define the affinity function we use f: $f(X_s, X_t) := A_{(s,t)}^* \in \mathbb{R}^{n \times hw \times hw}$.

Estimation Module g. Module g takes in the final 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 feature affinity warped by previous block and regresses displacement field at corresponding scale. 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

$$F^{(1)} = g_1(f(X_s, X_t)), \tag{2.2}$$

$$F^{(k)} = F^{(k-1)} + g_i(f(warp(X_s, F^{(k-1)}), X_t)), \tag{2.3}$$

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.

2.3 Weighted Perceptual Similarity

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

Perceptual similarity *s.* For an image pair (I_s, I_t), the model estimates 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 aligns semantically with the target.

In the ideal case, each pixel in the warped source $X_{s\to t}$ will have the largest 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 correctly matching pixel from all pixels in target X_t:

$$s(n,i) = -\log \frac{\exp\left(A^*_{(s\to t,t)}(n,i,i)/\tau\right)}{\sum_j \exp\left(A^*_{(s\to t,t)}(n,i,j)/\tau\right)}, \tag{2.4}$$

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

Weight function w**.** While it is possible to optimize flow estimation with the formula above, 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, 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:

$$w(n,i) = scale(\max_{j}[norm(A^{*}_{(s\to t,t)})(n,i)] \tag{2.5}$$

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

$$\mathscr{L}_{sim}(n,i) = w(n,i)s(n,i) \tag{2.6}$$

We visualize the image pairs, feature maps, weight maps, and final results of samples from the photo-sketch correspondence benchmark to exhibit the function of each part in Figure 2.2.

2.4 Additional Objectives

In addition to the perceptual similarity loss, we consider two other weakly-supervised losses to assist robust warp estimation and stabilize training.

Synthetic warp. Many approaches [63] use synthetically generated image pair as a direct supervision to the warp estimator, where the estimator is trained to optimize toward

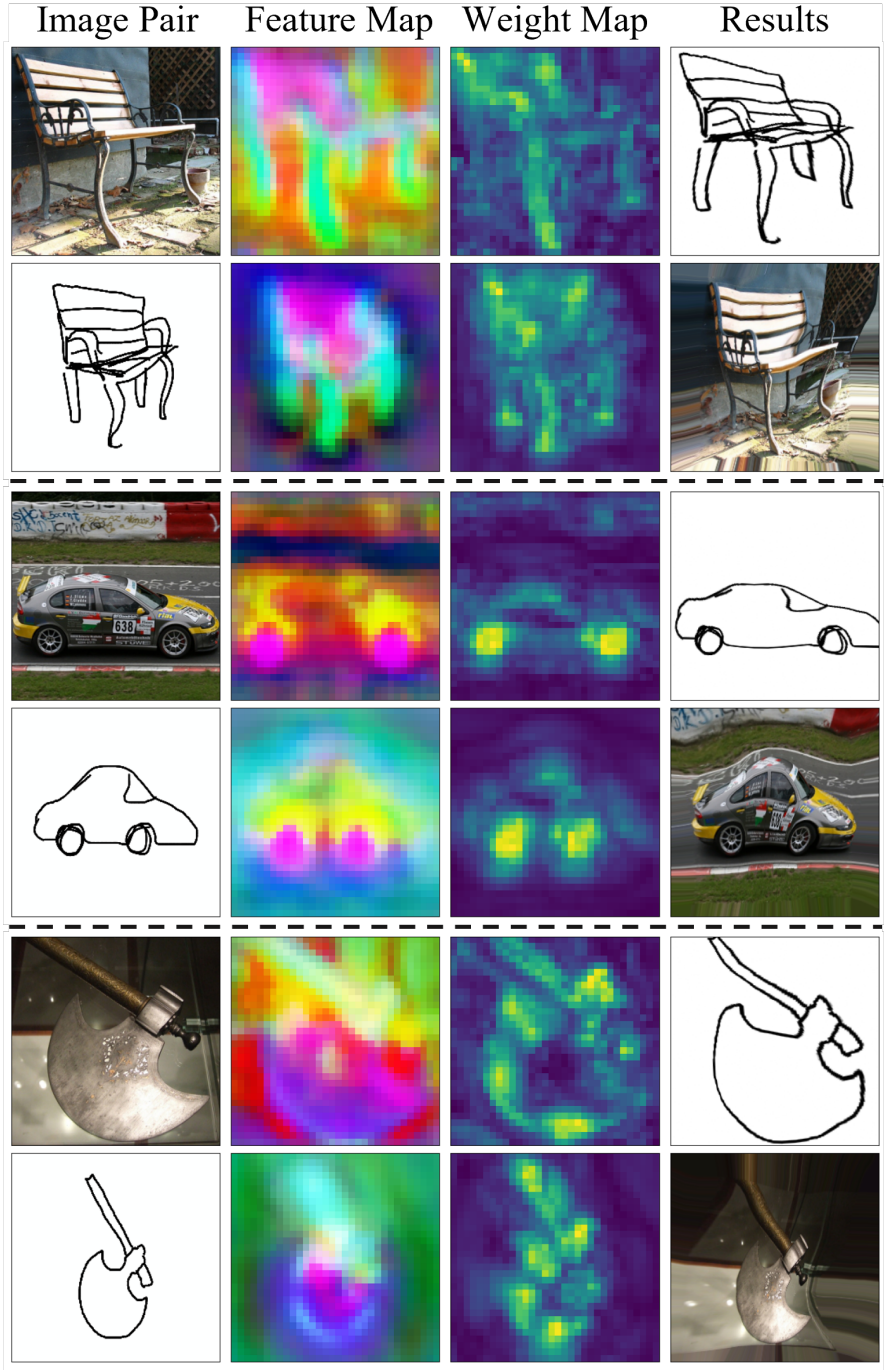

Figure 2.2. Example image pairs, feature maps, weight maps, and final results processed in our warp estimator. The weight maps highlight semnatic parts that have the largest correlation between the two images. We use PCA to project the feature dimentions to 3 principal components as RGB.

the constructed ground-truth flow. It not only accelerates convergence, but also prevents the network from converging to trivial solutions during initialization. With an arbitrary source image I_s, we randomly sample a displacement field $\hat{F}_{s\to s'}$ and generate the synthetic target image as $I_{s'} = warp(I_s, \hat{F}_{s\to s'})$. Following [71], each STN block has a different weight in the loss function. The goal is to minimize

$$\mathscr{L}_{syn} = \sum_k \alpha_k \| F^{(k)}_{s\to s'} - \hat{F}_{s\to s'} \|, \tag{2.7}$$

In practice, we sample the synthetic flow as a composition of affine and Thin-plate Spline (TPS) transformations.

Forward-backward consistency. Forward-backward consistency is a classic idea in tracking [79, 81, 35] and flow estimation [53, 62, 37, 76, 31] 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 apply this loss to both real photo-sketch pairs and synthetic image pairs, by minimizing the $L2$ norm between the identity flow and the composition of the forward flow and backward flow:

$$\mathscr{L}_{con} = \| warp(F_{s\to t}, F_{t\to s}) - F_{\mathbb{I}} \|, \tag{2.8}$$

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

Overall, our final objective is

$$\mathscr{L} = \lambda_{sim}\mathscr{L}_{sim} + \lambda_{syn}\mathscr{L}_{syn} + \lambda_{con}\mathscr{L}_{con}, \tag{2.9}$$

Chapter 3

Experiments and Results

In this section, 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.

3.1 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 [25] on ImageNet-2012 [13]. 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 [25, 8], with $dim = 128, m = 0.999, t = 0.07, lr = 0.03$ and a two-layer MLP head. Noticeably, we set the size of 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 0.003, leading to 2500 epochs of training in total. We set the weights of the objectives to $\lambda_{sim} = 0.1, \lambda_{syn} = 1.0, \lambda_{con} = 1.0$. We compute $\mathscr{L}_{sim}$ using the features after ResNet stages 2 and 3, and the temperature is set

to $\tau = 0.001$. The weights of the STN blocks in $\mathscr{L}_{syn}$ are $\alpha_1 = 1.0, \alpha_2 = 0.5, \alpha_3 = 0.25$.

We train the network with the SGD optimizer[66], weight decay of $1e-4$, the native mixed precision of Pytorch [58], and the batch size of 256. We adopt a cosine learning rate decay schedule [51].

3.2 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 3.1. For fair comparisons, we retrain existing open-sourced methods on the same photo-sketch dataset we used to develop our own model [67]. We report their PCK for $\alpha = (0.05, 0.1)$ in two settings: tranfer (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.

Our approach sets a new state-of-the-art in the field of photo-sketch correspondence. Although we only regress flow at the 16×16 scale which is less than the granularity of PCK-05, our ResNet-101 model gains a substantial increase of +0.77%/+4.38% compared to the second best method WarpC-SemanticGLU-Net[76]. This is surprising as the latter method

Table 3.1. State-of-the-art comparison for photo-sketch correspondence learning.

		Transfer		**Retrain**	
Methods	Encoder	PCK-5	PCK-10	PCK-5	PCK-10
CNNGeo [63]	ResNet-101	27.59	57.71	19.19	42.57
WeakAlign [63]	ResNet-101	35.65	68.76	43.55	78.60
NC-Net [65]	ResNet-101	40.60	63.50	–	–
DCCNet [31]	ResNet-101	42.43	66.53	–	–
PMD [74]	VGG-16	35.77	71.24	–	–
WarpC-SemanticGLUNet [76]	VGG-16	48.79	71.43	56.78	79.70
Ours	ResNet-18	–	–	54.32	**82.12**
Ours	ResNet-101	–	–	**57.55**	**84.08**

benefits from flow resolution four times as large as ours, and additional two-stage training on CityScape[10], DPED[33], and ADE[89]. 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 our model predicts in photo-to-sketch correspondence (Figure 5.1, Figure 5.2) and sketch-to-photo correspondence(Figure 5.3).

3.3 Ablation Study

In Table 3.2 we analyze different training schemes for our 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 classical contrastive learning on the photo-sketch dataset harms model performance, because the domains of photo and sketch are separated in the representation space. The best result comes from contrastive learning on photo-sketch pairs as it provides strongest supervision for learning discriminative features. In Table 3.3 we analyze the key components of our correspondence estimation on the ResNet-18 version of our model. We first show the importance of our perceptual similarity loss, which is essential in aligning the feature space from the two modalities. Using multiple feature layers, conditional BN, and the weight function further improves model performance.

Table 3.2. Ablation study on the feature encoder training.

Training Description	PCK-5	PCK-10
ImageNet only	41.67	76.80
CL on individual image	36.75	68.26
CL on image class	52.03	80.69
CL on image pair	54.32	82.12

Table 3.3. Ablation study on correspondence estimation.

Ablation Description	PCK-5	PCK-10
No $\mathscr{L}_{sim}$	22.62	57.68
No perceptual $\mathscr{L}_{sim}$	37.19	73.79
No multiple feature layers	52.11	82.03
No conditional BN	52.37	81.43
No weight function w	50.96	81.66
Complete model	54.32	82.12

3.4 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 3.1. 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.

Figure 3.1. 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[24] as PF.

3.5 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 [20]. 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 [20, 19]. 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 3.2).

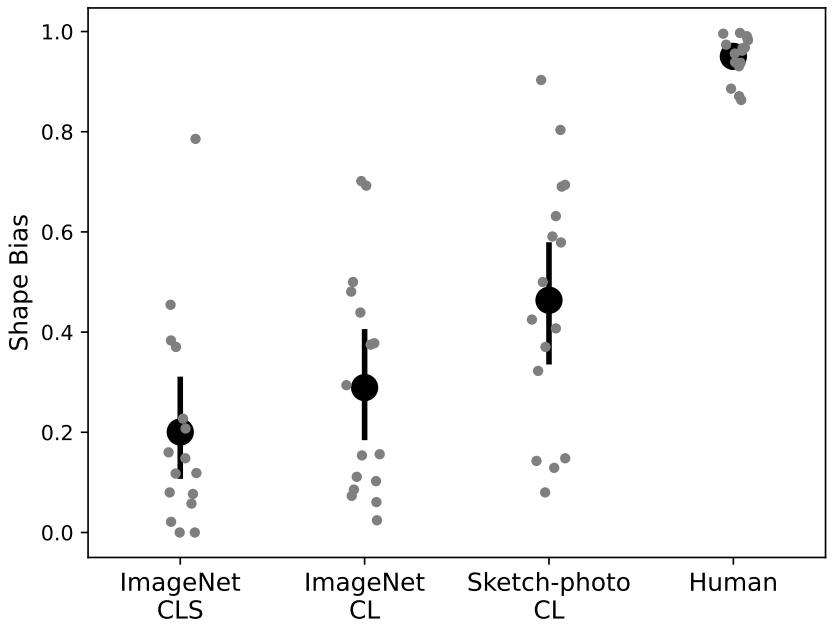

Figure 3.2. Comparing the degree of shape vs. texture bias between models trained with different objectives. Higher values suggest that the model recognition depends more on shape information. Our model exhibits more human-like performance. Each dot represents an object category from [20]. Error bars indicate 95% CI.

Chapter 4

Related Work

4.1 Self-supervised Representation Learning

Learning with self-supervision aims to obtain generic representations for diverse downstream tasks with minimal dependence on human labels [80, 14, 59, 56, 88, 21, 82]. 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 [91, 45, 61], including the ability to understand sketches [29, 40]. In particular, recently proposed *contrastive learning* techniques demonstrate competitive performance with supervised baselines not only on visual recognition [28, 57, 82, 6, 25, 23, 7, 9], but also on learning visual representations from inputs varying across sensory views [72, 73], across frames in video [35, 83, 90], and even between text and images [60, 39]. 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.

4.2 Weakly-supervised Semantic Correspondence Learning

Geometric matching [54, 49, 64, 69, 75] is perhaps the most basic form of correspondence prediction, which aims to align two views of the same scene. By contrast, *semantic matching*

[24, 63, 65, 31, 48, 76] 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 [63, 37, 68], image pairs [65, 44, 43, 38, 31, 48, 76], and class labels [77]. 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 [24, 63, 37, 68, 48, 76] or a cost volume [65, 31, 77]. In this work, we propose a photo-sketch correspondence learning framework that explicitly estimates the dense flow field with image pair-level supervision.

Chapter 5

Conclusions

What is needed to develop artificial systems that learn to perceive the visual world as robustly as humans do? While there have been tremendous recent advances in the performance of artificial vision systems on a variety of tasks, there are 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 [67]. 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 based on contrastive learning and STN is effective in capturing photo-sketch correspondence, but there remains a systematic gap with human performance. Taken together, we hope that these findings, together with a new challenging fine-grained multimodal image understanding benchmark will catalyze progress towards achieving more human-like vision systems.

Figure 5.1. More alignment examples on the PSC6K dataset.

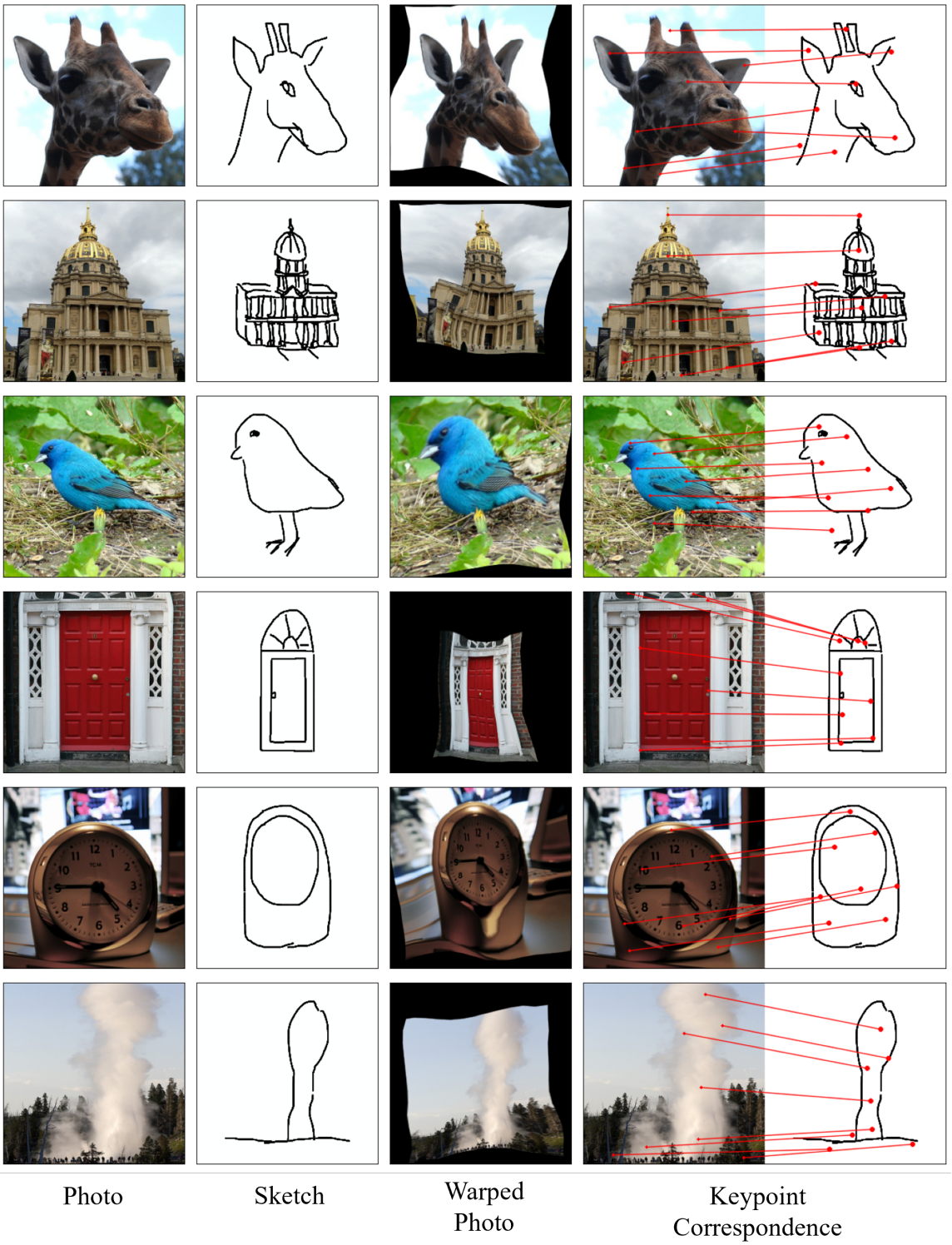

Figure 5.2. More alignment examples on the PSC6K dataset.

Figure 5.3. More alignment examples on the PSC6K dataset.

Bibliography

[1] Maxime Aubert, Adam Brumm, Muhammad Ramli, Thomas Sutikna, E Wahyu Saptomo, Budianto Hakim, Michael J Morwood, Gerrit D van den Bergh, Leslie Kinsley, and Anthony Dosseto. Pleistocene cave art from sulawesi, indonesia. *Nature*, 514(7521):223–227, 2014.

[2] Ayu Marliawaty I Gusti Bagus, Tiago Marques, Sachi Sanghavi, James J DiCarlo, and Martin Schrimpf. Primate inferotemporal cortex neurons generalize better to novel image distributions than analogous deep neural networks units. In *SVRHM 2022 Workshop@ NeurIPS*.

[3] Alexander C Berg, Tamara L Berg, and Jitendra Malik. 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, pages 26–33. IEEE, 2005.

[4] Ayan Kumar Bhunia, Yongxin Yang, Timothy M Hospedales, Tao Xiang, and Yi-Zhe Song. 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*, pages 9779–9788, 2020.

[5] Santiago A Cadena, George H Denfield, Edgar Y Walker, Leon A Gatys, Andreas S Tolias, Matthias Bethge, and Alexander S Ecker. Deep convolutional models improve predictions of macaque v1 responses to natural images. *PLoS computational biology*, 15(4):e1006897, 2019.

[6] 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.

[7] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big self-supervised models are strong semi-supervised learners. *Advances in neural information processing systems*, 33:22243–22255, 2020.

[8] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020.

[9] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9640–9649, 2021.

[10] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[11] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 886–893. Ieee, 2005.

[12] Harm De Vries, Florian Strub, Jérémie Mary, Hugo Larochelle, Olivier Pietquin, and Aaron C Courville. Modulating early visual processing by language. *Advances in Neural Information Processing Systems*, 30, 2017.

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

[14] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE international conference on computer vision*, pages 1422–1430, 2015.

[15] Mathias Eitz, Ronald Richter, Tamy Boubekeur, Kristian Hildebrand, and Marc Alexa. Sketch-based shape retrieval. *ACM Transactions on graphics (TOG)*, 31(4):1–10, 2012.

[16] Judith E Fan, Robert D Hawkins, Mike Wu, and Noah D Goodman. Pragmatic inference and visual abstraction enable contextual flexibility during visual communication. *Computational Brain & Behavior*, 3(1):86–101, 2020.

[17] Judith E Fan, Daniel LK Yamins, and Nicholas B Turk-Browne. Common object representations for visual production and recognition. *Cognitive science*, 42(8):2670–2698, 2018.

[18] Jerry Fodor. The revenge of the given. *Contemporary debates in philosophy of mind*, pages 105–116, 2007.

[19] Robert Geirhos, Kantharaju Narayanappa, Benjamin Mitzkus, Tizian Thieringer, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Partial success in closing the gap between human and machine vision. *Advances in Neural Information Processing Systems*, 34:23885–23899, 2021.

[20] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018.

[21] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018.

[22] Nelson Goodman. *Languages of art: An approach to a theory of symbols*. Hackett publishing, 1976.

[23] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

[24] Bumsub Ham, Minsu Cho, Cordelia Schmid, and Jean Ponce. Proposal flow. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3475–3484, 2016.

[25] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[26] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

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

[28] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018.

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

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

[31] Shuaiyi Huang, Qiuyue Wang, Songyang Zhang, Shipeng Yan, and Xuming He. Dynamic context correspondence network for semantic alignment. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2010–2019, 2019.

[32] Holly Huey, Xuanchen Lu, Caren Walker, and Judith Fan. Explanatory drawings prioritize functional properties at the expense of visual fidelity. 2021.

[33] Andrey Ignatov, Nikolay Kobyshev, Radu Timofte, Kenneth Vanhoey, and Luc Van Gool. Dslr-quality photos on mobile devices with deep convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3277–3285, 2017.

[34] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR, 2015.

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

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

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

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

[39] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021.

[40] John M Kennedy and Abraham S Ross. Outline picture perception by the songe of papua. *Perception*, 4(4):391–406, 1975.

[41] Seyed-Mahdi Khaligh-Razavi and Nikolaus Kriegeskorte. Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS computational biology*, 10(11):e1003915, 2014.

[42] Jaechul Kim, Ce Liu, Fei Sha, and Kristen Grauman. Deformable spatial pyramid matching for fast dense correspondences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2307–2314, 2013.

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

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

[45] Talia Konkle and George A Alvarez. Instance-level contrastive learning yields human brain-like representation without category-supervision. *BioRxiv*, 2020.

[46] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

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

[48] Xin Li, Deng-Ping Fan, F. Yang, Ao Luo, Hong Cheng, and Zicheng Liu. Probabilistic model distillation for semantic correspondence. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7501–7510, 2021.

[49] Xinghui Li, Kai Han, Shuda Li, and Victor Prisacariu. Dual-resolution correspondence networks. *Advances in Neural Information Processing Systems*, 33:17346–17357, 2020.

[50] Ce Liu, Jenny Yuen, and Antonio Torralba. Sift flow: Dense correspondence across scenes and its applications. *IEEE transactions on pattern analysis and machine intelligence*, 33(5):978–994, 2010.

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

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

[53] Simon Meister, Junhwa Hur, and Stefan Roth. Unflow: Unsupervised learning of optical flow with a bidirectional census loss. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

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

[55] Kushin Mukherjee, Robert XD Hawkins, and Judith W Fan. Communicating semantic part information in drawings. In *CogSci*, pages 2413–2419, 2019.

[56] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European conference on computer vision*, pages 69–84. Springer, 2016.

[57] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[58] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[59] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016.

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

[61] Rishi Rajalingham, Elias B Issa, Pouya Bashivan, Kohitij Kar, Kailyn Schmidt, and James J DiCarlo. 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.

[62] 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.

[63] 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.

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

[65] Ignacio Rocco, Mircea Cimpoi, Relja Arandjelović, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Neighbourhood consensus networks. *Advances in neural information processing systems*, 31, 2018.

[66] Sebastian Ruder. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*, 2016.

[67] Patsorn Sangkloy, Nathan Burnell, Cusuh Ham, and James Hays. The sketchy database: learning to retrieve badly drawn bunnies. *ACM Transactions on Graphics (TOG)*, 35(4):1–12, 2016.

[68] 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.

[69] Xi Shen, François Darmon, Alexei A Efros, and Mathieu Aubry. Ransac-flow: generic two-stage image alignment. In *European Conference on Computer Vision*, pages 618–637. Springer, 2020.

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

[71] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8934–8943, 2018.

[72] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *European conference on computer vision*, pages 776–794. Springer, 2020.

[73] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? *Advances in Neural Information Processing Systems*, 33:6827–6839, 2020.

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

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

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

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

[78] Yael Vinker, Ehsan Pajouheshgar, Jessica Y Bo, Roman Christian Bachmann, Amit Haim Bermano, Daniel Cohen-Or, Amir Zamir, and Ariel Shamir. Clipasso: Semantically-aware object sketching. *arXiv preprint arXiv:2202.05822*, 2022.

[79] Carl Vondrick, Abhinav Shrivastava, Alireza Fathi, Sergio Guadarrama, and Kevin Murphy. Tracking emerges by colorizing videos. In *Proceedings of the European conference on computer vision (ECCV)*, pages 391–408, 2018.

[80] Xiaolong Wang and Abhinav Gupta. Unsupervised learning of visual representations using videos. In *Proceedings of the IEEE international conference on computer vision*, pages 2794–2802, 2015.

[81] Xiaolong Wang, Allan Jabri, and Alexei A Efros. Learning correspondence from the cycle-consistency of time. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2566–2576, 2019.

[82] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3733–3742, 2018.

[83] Jiarui Xu and Xiaolong Wang. Rethinking self-supervised correspondence learning: A video frame-level similarity perspective. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10075–10085, 2021.

[84] Daniel LK Yamins, Ha Hong, Charles F Cadieu, Ethan A Solomon, Darren Seibert, and James J DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the national academy of sciences*, 111(23):8619–8624, 2014.

[85] Justin Yang and Judith E Fan. Visual communication of object concepts at different levels of abstraction. *arXiv preprint arXiv:2106.02775*, 2021.

[86] Qian Yu, Feng Liu, Yi-Zhe Song, Tao Xiang, Timothy M Hospedales, and Chen-Change Loy. Sketch me that shoe. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 799–807, 2016.

[87] Qian Yu, Yongxin Yang, Feng Liu, Yi-Zhe Song, Tao Xiang, and Timothy M Hospedales. Sketch-a-net: A deep neural network that beats humans. *International journal of computer vision*, 122(3):411–425, 2017.

[88] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European conference on computer vision*, pages 649–666. Springer, 2016.

[89] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127(3):302–321, 2019.

[90] Chengxu Zhuang, Tianwei She, Alex Andonian, Max Sobol Mark, and Daniel Yamins. Unsupervised learning from video with deep neural embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9563–9572, 2020.

[91] Chengxu Zhuang, Siming Yan, Aran Nayebi, Martin Schrimpf, Michael C Frank, James J DiCarlo, and Daniel LK Yamins. Unsupervised neural network models of the ventral visual stream. *Proceedings of the National Academy of Sciences*, 118(3):e2014196118, 2021.