Self-Supervised Spatial Correspondence Across Modalities

Ayush Shrivastava Andrew Owens
University of Michigan
https://ayshrv.com/cmrw

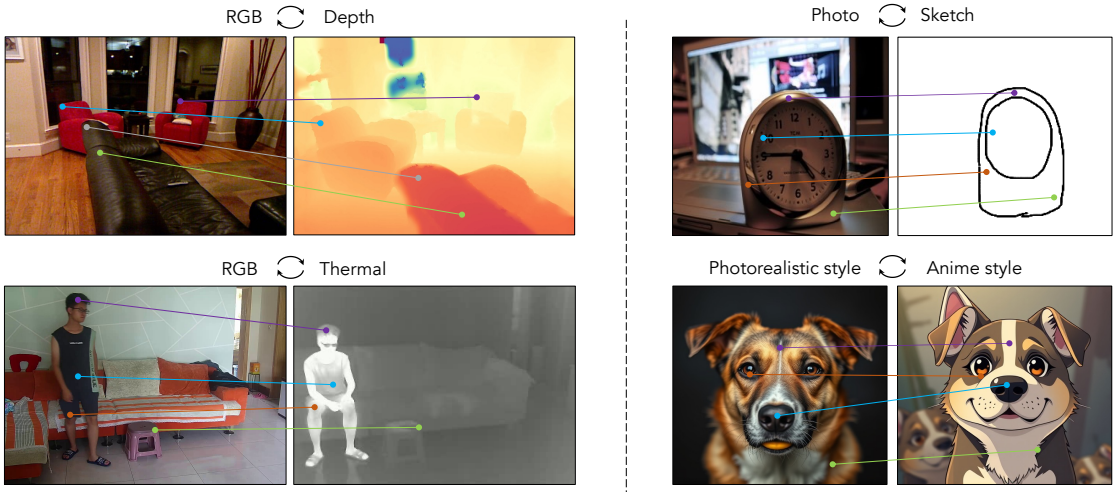

Figure 1. **Finding spatial correspondences across modalities.** We present a method for cross-modal matching, trained entirely through self-supervision using a simple formulation based on contrastive random walks [14]. (a) Given two images taken by different visual modalities and at different positions and times, we predict the pairs of image patches that physically correspond to the same points. (b) We also apply our method to semantic matching tasks, using a visual encoder initialized with pretrained DINOv2 [29] weights and fine-tuned during training. These include tasks such as photo-sketch alignment [28] and style-based matching between images of different styles generated with a text-to-image model [19].

Abstract

We present a method for finding cross-modal space-time correspondences. Given two images from different visual modalities, such as an RGB image and a depth map, our model identifies which pairs of pixels correspond to the same physical points in the scene. To solve this problem, we extend the contrastive random walk framework to simultaneously learn cycle-consistent feature representations for both cross-modal and intra-modal matching. The resulting model is simple and has no explicit photo-consistency assumptions. It can be trained entirely using unlabeled data, without the need for any spatially aligned multimodal image pairs. We evaluate our method on both geometric and semantic correspondence tasks. For geometric matching, we consider challenging tasks such as RGB-to-depth and RGB-to-thermal matching (and vice versa); for semantic matching, we evaluate on photo-sketch and cross-style image alignment. Our method achieves strong performance across all benchmarks.

1. Introduction

Cameras take a multitude of different forms. While RGB, thermal, and depth cameras each record images, what is stored within each pixel differs drastically. Consequently, when a scene is captured using cameras that are based on different sensory modalities, it is difficult to determine which pixels within them correspond to the same physical points. These cross-modal pixel correspondences have a number of applications, such as cross-modal registration, 3D reconstruction, and multimodal data fusion [1, 6, 22, 47]. Recent work on multimodal learning has provided effective ways of learning cross-modal correspondences through self-supervision, yet these methods largely rely on having paired data from multiple sensors, which is not available for pixel-level correspondence.

Traditional self-supervised methods for pixel correspondences often make strong assumptions about the visual appearance of pixels that should be matched together, such as by assuming that matching image patches should be photo-

consistent [16, 17, 38, 51, 57], or that cross-modal translation methods can estimate one modality from another as part of the matching process [1]. However, these assumptions are frequently violated in multimodal data, making it challenging to use them as general-purpose multimodal matching methods. Depth images and RGB images, for example, may look alike to a human, but their intensity values record information that is so different—depth *vs.* lightness—that they cannot be matched using local information. And cross-modal prediction requires being able to solve a notoriously difficult problem—monocular depth estimation—without explicit paired data.

We take inspiration from recent advances in self-supervised tracking [2, 14, 23, 48] that learn correspondences between video frames through *cycle consistency*. These methods assume that the content within a scene persists between frames, and that there should thus be a one-to-one correspondence between the pixels in them. Our approach extends the contrastive random walk [14] to the *cross-modal* pixel correspondence problem. We create a directed graph in which nodes correspond to image patches in each modality, and where edges connect patches across modalities. We train a network, based on the recently proposed global matching transformer [36, 49], to assign transition probabilities to pairs of patches for a random walk. A random walker uses these transition probabilities to step through the graph, moving from patches in one modality to another, and then back. To train the model, we maximize the walker's return probability, thus encouraging matches to be cycle-consistent.

A key challenge in our setting is that, unlike many correspondence tasks, the patches we match are visually very different. We therefore incorporate *intra-modal* random walks between images of the same modality, such that the model must learn a representation that simultaneously allows it to match intra- and inter-modally. To get these image pairs, we apply data augmentation to input images, simulating the challenges of matching frames of a video. We also show that, in contrast to image-based matching [36], spatial smoothness priors are important for obtaining strong performance. Moreover, in contrast to other cross-modal correspondence methods [1, 17, 38], our method does not rely on hand-crafted measures of photoconsistency. Instead, it learns visual similarity through cycle-consistent representation learning, defined by the learned embedding space, thereby making it possible to apply it to different pairs of modalities without any modifications.

We evaluate our model on both geometric and semantic cross-modal correspondence tasks. For geometric matching, we focus on two challenging tasks: RGB-to-depth and RGB-to-thermal matching (Figure 1a). In both cases, we match images captured at different times and positions, with different sensors. For semantic matching (Figure 1b), we use a visual encoder initialized with pretrained DINOv2 [29] weights and fine-tune it during training. We evaluate on photo-to-sketch alignment, where our method performs competitively with approaches specifically tailored to this task, and on a new cross-style image matching task, using images generated in different styles from the same diffusion model [19]. Our work makes several contributions:

- We show that space-time cross-modal pixel correspondence can be learned from unlabeled data through cycle consistency.
- We extend the contrastive random walk framework for cross-modal pixel correspondence.
- Through experiments, we show that our model successfully finds correspondences in geometric tasks such as RGB-to-depth and RGB-to-thermal, significantly outperforming prior methods. In semantic tasks like photo-to-sketch matching, our method is competitive with other self-supervised approaches.
- We propose benchmarks for evaluating RGB-to-depth, RGB-to-thermal, and cross-style image matching. For cross-style matching, we use generative models to synthesize images with similar content but different styles.

2. Related Work

Cycle consistency for correspondence. Zhou et al. [55] proposed learning correspondences across 2D views of 3D objects, by matching the viewpoints to a 3D model, then projecting back to other images. In contrast, we focus on matching two images that differ in *modality*, *space*, and *time*, rather than just 3D viewpoint. Later work extended cycle consistency to space-time tracking. Wang et al. [48] trained a model to track forward and backward in time, penalizing deviation from the original position, using a formulation based on the spatial transformer [15]. Jabri et al.[14] framed this as a random walk on a space-time graph, enabling self-supervised tracking, followed by improvements such as shortcut removal[41] and smoothness priors [2] for fine-grained correspondence. Recent works further integrated this framework with global matching transformers [36, 49] to address the tracking-any-point problem [4].

While these works focus on temporal matching, we extend contrastive random walks to cross-modal matching, introducing both intra- and inter-modal random walks. Although contrastive random walks have been explored in multimodal settings like image-audio alignment [11] or audio-to-audio matching [3], these approaches operate at the patch or node level and do not aim for dense pixel-to-pixel correspondence, which is the focus of our work.

Cross-modal pixel correspondence. Cross-modal correspondence has been explored through contrastive learning [10, 31, 44], but these approaches typically learn global or patch-level embeddings, not dense pixel-wise matches.

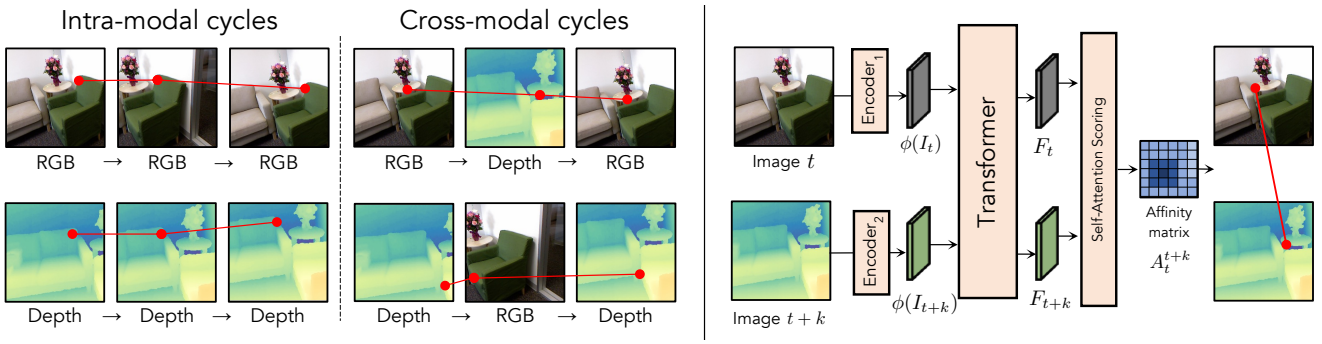

Figure 2. **Model Architecture.** We learn to find pixel-level correspondences between images that may differ in sensory modality, time, and scene position. Given images from two modalities (e.g., unpaired RGB and depth images from the same scene), we perform a contrastive random walk on a graph whose nodes come from patches within the two images using a global matching transformer architecture [36]. We simultaneously perform auxiliary intra-modal random walks within each modality's augmented crops of images to improve the model's ability to avoid local minima during optimization. Through this process, we learn to match in both directions (e.g., RGB-to-depth and depth-to-RGB).

Arar et al.[1] proposed a method for RGB-infrared matching by jointly training a GAN for modality translation and a spatial transformation network. In contrast, our model does not require explicit cross-modal translation; it learns from cycle-consistency alone. Other methods such as[28] address photo-sketch alignment through a two-stage process involving cross-modal embeddings, whereas we learn fine-grained correspondence directly through self-supervision. Some recent work has explored matching across 2D-3D modalities (e.g., RGB and point clouds [22, 47]), but these approaches often rely on explicit geometric priors and operate on sparse data like LiDAR. Our method, by contrast, is designed for dense 2D image domains and does not rely on any handcrafted similarity metric.

Self-supervised correspondence. Unsupervised learning of correspondence has a long history in optical flow [16, 17, 38, 51, 57], often relying on photometric loss as a training signal. Other approaches use cues like color similarity [20, 23, 46] or temporal coherence [7, 50] for self-supervised tracking. However, these techniques assume single-modality inputs and struggle to generalize to cross-modal settings. Simulated data has also been used for learning correspondence [4, 5, 8, 9, 39], but generating realistic, time-varying multimodal signals remains a significant challenge. In contrast, our method learns from real multimodal videos without requiring ground-truth annotations or simulation.

3. Method

Our goal is to learn pixel-wise correspondences between images from different modalities (e.g., RGB, thermal, depth, sketch, stylistically diverse images) solely through self-supervision. We build upon the contrastive random walk framework [14] and extend it to handle multiple modalities (Figure 2). The recently proposed Global Matching Random Walk (GMRW) model [36] allows for dense correspondences between RGB images by enforcing cycle consistency via contrastive random walks over videos [36, 49]. We leverage this framework to train on unlabeled videos from different modalities, learning cycle-consistent tracking across pairs such as RGB-depth, RGB-thermal, photo-sketch, and cross-style images.

However, we find that directly optimizing the model for cross-modal cycle consistency leads to poor convergence: training plateaus at a suboptimal solution with high loss and non-semantic matches. We hypothesize that this difficulty arises due to the substantial domain gap between modalities. For instance, unlike space-time matching—where initial random features provide reasonable gradients—cross-modal inputs may align arbitrarily (e.g., bright regions in RGB could align with distant areas in depth due to similarly high intensity values). To mitigate this, we introduce auxiliary intra-modality random walk supervision, encouraging the model to learn representations effective for both intra- and cross-modality matching.

3.1. Model architecture

To address the challenge of cross-modal dense correspondence, we require an architecture capable of producing high-resolution matches while remaining modality-agnostic. We adopt the Global Matching Random Walk (GMRW) model [36], which performs all-pairs pixel matching [43, 49] and is trained using a cycle consistency objective via contrastive random walks. Crucially, it does not rely on modality-specific assumptions like photo-consistency. Each modality uses a dedicated visual encoder to produce input tokens, while a shared Transformer backbone performs the global matching across modalities.

Image Features. For geometric tasks, we use a CNN-based encoder to extract visual features. For each image I_t^m at time t and modality $m \in \{\texttt{rgb}, \texttt{depth}, \texttt{thermal}\}$, we

compute d-dimensional features $\phi(I_t^m) \in \mathbb{R}^{\frac{H}{c} \times \frac{W}{c} \times d}$ with a downsampling factor $c = 4$. Each modality has its own encoder, and we append 2D positional encodings to the features. For depth inputs, the single-channel image is replicated across three channels to match the RGB architecture.

For semantic tasks (e.g., photo-sketch, cross-style), we use DINOv2 [29] as a shared visual encoder across modalities. We initialize it with pretrained weights and finetune it, leveraging the rich semantic priors learned during DINOv2 pretraining.

Cross-modal Matching Transformer. Given a pair of images $I_t^{m_1}$ and $I_{t+k}^{m_2}$, we feed their extracted features into a shared Transformer consisting of six blocks of self-attention, cross-attention, and feed-forward layers. The final layer produces correlation features $F_t^{m_1}$ and $F_{t+k}^{m_2}$. This module supports both intra-modality ($m_1 = m_2$) and cross-modality ($m_1 \neq m_2$) matching.

Cross-modal Transition Matrix. We compute the transition matrix from $F_t^{m_1}$ and $F_{t+k}^{m_2}$ as $A_{t,t+k}^{m_1,m_2} = \text{softmax}(F_t^{m_1}(F_{t+k}^{m_2})^\top/\tau)$, which represents the likelihood of a pixel in modality m_1 at time t corresponding to a pixel in modality m_2 at time $t+k$. This matrix defines the transition probabilities for contrastive random walks (CRW) in both intra- and cross-modality matching. Following [36, 49], we normalize features using $\tau = \sqrt{d}$ instead of the L2 normalization used in earlier work [2, 14, 41].

The expected change in pixel position is computed as:

$$\mathbf{f}_{t,t+k}^{m_1,m_2} = \mathbb{E}_{A_{t,t+k}}[A_{t,t+k}^{m_1,m_2} D - D] \tag{1}$$

where $\mathbf{f}_{t,t+k}$ is the predicted flow between modalities m_1 and m_2, D is the constant pixel coordinate grid and $A_{t,t+k}^{m_1,m_2}$ is the transition matrix from frame t to $t+k$.

3.2. Learning cross-modal correspondences

Cross-modal cycle-consistency. To learn cross-modal correspondences, we extend the cycle consistency objective from GMRW [36] to multi-modal data. Given a *palindrome* sequence $\{I_t^{m_1}, I_{t+k}^{m_2}, I_t^{m_1}\}$, we treat it as a space-time graph and train our model to perform random walks across it. The model estimates two transition matrices: $A_{t,t+k}^{m_1,m_2}$ which maps pixels from modality m_1 at time t to modality m_2 at time $t+k$, and $A_{t+k,t}^{m_2,m_1}$, which returns the walker back to modality m_1 at time t. We chain these two transitions to obtain a round-trip path from a point in $I_t^{m_1}$, through $I_{t+k}^{m_2}$, and back. To enforce cycle consistency, we maximize the probability that the walker returns to its original position. This is done using the label-warping loss from [36]:

$$\mathcal{L}_{\texttt{cross-crw}} = \mathcal{L}_{\texttt{CE}}(A_{t,t+k}^{m_1,m_2} A_{t+k,t}^{m_2,m_1}, T_f^b(I)), \tag{2}$$

where I is the identity matrix and $T_f^b(I)$ is the warped label target designed to prevent shortcut learning [36], and $\mathcal{L}_{\texttt{CE}}$ is the cross-entropy loss.

Intra-modal cycle-consistency. We further apply the same contrastive random walk objective within each modality to stabilize training. For a palindrome $\{I_{\texttt{ori}}^{m_i}, I_{\texttt{aug}}^{m_i}, I_{\texttt{ori}}^{m_i}\}$, where $I_{\texttt{aug}}^{m_i}$ is an augmented crop of $I_{\texttt{ori}}^{m_i}$, the intra-modal cycle loss is:

$$\mathcal{L}_{\texttt{intra-crw}} = \sum_{i=1}^{2} \mathcal{L}_{\texttt{CE}}(A_{\texttt{ori,aug}}^{m_i} A_{\texttt{aug,ori}}^{m_i}, T_f^b(I)) \tag{3}$$

Smoothness loss. To encourage spatial coherence, we apply an edge-aware smoothness loss [17] — used in prior contrastive random walk work [2, 36]. It penalizes large second-order derivatives in flow, weighted by image gradients. We apply this loss only when the source image is from the RGB modality, where visual similarity is a reliable cue for perceptual grouping. The loss is defined as:

$$\mathcal{L}_{\texttt{smooth}} = \mathbb{E}_p \sum_{d \in \{x,y\}} \exp(-\lambda_c I_d(p)) |\frac{\partial^2 \mathbf{f}_{s,t}(p)}{\partial d^2}| \tag{4}$$

where the p is a pixel in the RGB image, $I_d(p)$ is the average gradient magnitude across color channels in direction d. The coefficient λ_c controls edge sensitivity.

Overall loss. The final training objective combines all losses:

$$\mathcal{L}_{\texttt{cross-crw}} + \mathcal{L}_{\texttt{intra-crw}} + \lambda_s \mathcal{L}_{\texttt{smooth}} \tag{5}$$

4. Experiments

Our method estimates pixel-level correspondences between pairs of images across different modalities. We evaluate its performance on both geometric and semantic correspondence tasks. For geometric matching, we consider two settings: RGB-Depth and RGB-Thermal. For RGB-Depth, we use the NYU Depth V2 dataset [37] for both training and evaluation. For RGB-Thermal, we train and evaluate on two datasets: the indoor Thermal-IM dataset [42] and the outdoor KAIST dataset [13]. Notably, our method does not require spatially or temporally aligned image pairs during training. We sample frames at different time steps, allowing for scene changes and motion, while assuming partial scene overlap. For semantic correspondence, we consider two tasks. The first is photo-sketch alignment, using the PSC6K dataset [28], which includes annotated keypoints that match real-world photos with corresponding human-drawn sketches. The second task is cross-style image matching, where the goal is to match corresponding points across stylistic variants (e.g., photorealistic, anime, watercolor) of the same scene. We construct this benchmark using a text-to-image generation model conditioned on consistent prompts and varying style modifiers. For all benchmarks, evaluation sets are created by manually annotating keypoints across modalities or, where applicable, by propagating tracked points from the RGB domain using a point-tracking algorithm.

Table 1. **RGB-Depth and RGB-Thermal Matching**. We compare our method with supervised and several unsupervised baselines on NYU-Depth V2, Thermal-IM and KAIST datasets. Our method performs best when computing cross-modality matching, significantly outperforming the baselines. D: Depth, T: Thermal, SSL: Self-Supervised Learning

Method	SSL	NYU-Depth, $< \delta^x_{\text{avg}} \uparrow$				Thermal-IM, $< \delta^x_{\text{avg}} \uparrow$				KAIST, $< \delta^x_{\text{avg}} \uparrow$	
		Intra-modal		Cross-modal		Intra-modal		Cross-modal		Cross-modal	
		RGB → RGB	D → D	RGB → D	D → RGB	RGB → RGB	T → T	RGB → T	T → RGB	RGB → T	T → RGB
RAFT [43]		**91.5**	59.2	7.9	1.3	81.7	53.3	5.6	0.9	29.2	7.4
GMFlow [49]		79.7	58.2	12.7	12.5	**82.2**	**53.4**	3.8	2.6	23.1	22.3
Arar et al. [1]	✓	-	-	1.3	0.8	-	-	2.3	1.9	2.1	4.7
$\text{CycleGAN}_{\text{RGB→D/T}}$ + GMFlow	✓	-	-	8.5	7.8	-	-	7.9	7.1	8.3	8.2
$\text{CycleGAN}_{\text{D/T→RGB}}$ + GMFlow	✓	-	-	16.2	16.6	-	-	6.1	5.8	6.8	7.4
ARFlow [25]	✓	76.1	53.5	7.5	7.4	82.1	53.3	12.5	13.2	31.0	30.4
ARFlow (Retrained)	✓	-	-	9.3	8.1	-	-	13.4	13.1	29.1	27.7
DIFT [40]	✓	40.6	18.5	3.3	4.3	68.2	50.3	17.5	18.2	7.1	8.8
SD-DINO [52]	✓	25.2	13.9	7.8	6.4	44.5	49.9	29.3	34.9	19.6	20.6
Ours	✓	80.1	**62.3**	**33.5**	**34.3**	81.6	53.1	**41.8**	**47.9**	**35.2**	**34.1**

4.1. Geometric Correspondence

RGB-Depth Training. We train our model on the NYU Depth V2 dataset [37], which contains approximately 400K unlabeled RGB-D frames. Training is performed in three stages: 1) intra-modality CRW applied to RGB-RGB and Depth-Depth pairs, 2) additionally cross-modality CRW applied to RGB-Depth and Depth-RGB pairs, 3) adding the smoothness loss to encourage spatial coherence. During training, we randomly sample two frames ($I_t^{m_1}$, $I_{t+k}^{m_2}$) from a scene, separated by a few random timesteps, drawn from both modalities. We then apply different random resized crops to forward ($I_t^{m_1}$, $I_{t+k}^{m_2}$) and backward images ($I_{t,\text{aug}}^{m_1}$) [36] and train for cycle-consistency losses.

RGB-Thermal Training. We use the Thermal-IM and KAIST datasets for RGB-Thermal training. Thermal-IM consists of 783 video sequences with RGB and thermal modalities, though they are spatially unaligned. KAIST contains 320 video sequences with spatially aligned RGB and thermal views. As with RGB-Depth training, we randomly sample two frames per scene from both modalities and follow the same three-stage training pipeline with intra- and cross-modality CRW losses.

RGB-Depth Evaluation. Since RGB and depth frames in NYU are spatially aligned, we use PIP++ [54], a point tracking method, to generate ground-truth correspondences. We create 10-frame video clips and track points across them, retaining those visible in all frames. These tracks are used to evaluate RGB-RGB, Depth-Depth, and RGB-Depth correspondence. Our final dataset includes 250 video clips with an average of 688 annotated tracks per clip.

RGB-Thermal Evaluation. Thermal-IM videos are not spatially aligned, so we manually annotate 100 RGB-Thermal frame pairs across 5 timesteps, with 10 keypoints each, following the protocol from TAP-Vid [4], resulting in 1,000 evaluation points. For KAIST, where RGB and thermal frames are aligned, we use a strategy similar to NYU Depth, using CoTracker [18] to generate correspondences, as it performs better than other trackers on KAIST videos.

Evaluation Metrics. We adopt the positional accuracy metric from TAP-Vid [4], denoted as $< \delta^x_{\text{avg}}$. This metric reports the fraction of visible keypoints that are predicted within a pixel distance threshold from ground truth, averaged over thresholds $1, 2, 4, 8, 16$.

4.2. Semantic Correspondence

Photo-Sketch Matching. We use the PSC6K dataset [28] for both training and evaluation. The dataset contains 6,250 annotated photo-sketch pairs, with 8 keypoints per pair in the evaluation set, and over 130K photo-sketch pairs in the training set. Performance is measured using Percentage of Correct Keypoints (PCK) at thresholds $\alpha = 0.05$ and 0.1, denoted as PCK-5 and PCK-10.

Cross-style Image Matching. We introduce a new benchmark to evaluate correspondence across different visual styles of the same scene. We generate images using the Flux text-to-image model [19], conditioned on a base prompt and a style modifier (added to the prompt). We consider 10 distinct styles: anime, dark, light, photorealistic, pixel art, watercolor, comic book, neon, pastel, and sci-fi (examples in Figure 6). Prompts are generated using BLIP-2[21] by captioning images from the ImageNet dataset [35], and are then used to generate stylistic variants via Flux.

We generate 10K images per style, resulting in 100K training samples. For evaluation, we sample 50 unseen prompts and generate 10 images per prompt—one in each style—

yielding 500 test images. We manually annotate corresponding keypoints across styles for each prompt. All possible pairs of styles are compared within each prompt, resulting in 2,250 total image pair comparisons in the test set.

5. Results

We train our model for RGB-Depth, RGB-Thermal, photo-sketch and cross-style matching tasks and present both quantitative and qualitative results alongside comparisons with baseline methods.

5.1. Geometric correspondence

Table 1 compares our method against a range of supervised and unsupervised baselines for RGB-Depth and RGB-Thermal matching, evaluated in the direct setting where query frames are matched directly to target frames without chaining [36]. Among the supervised baselines, we include RAFT [43] and GMFlow [49], state-of-the-art optical flow models trained on RGB images, applied to thermal and depth data by treating them as RGB inputs. For unsupervised baselines, we evaluate Arar et al.[1], a CycleGAN[56]-based cross-modal registration method, which struggles to generalize due to the difficulty of translating between modalities without paired supervision. To address this, we introduce a variant (CycleGAN+GMFlow) that first translates depth or thermal inputs to RGB using CycleGAN, followed by matching in the RGB domain using GMFlow. We also test the reverse direction—translating RGB to depth or thermal—before applying GMFlow. We further evaluate ARFlow [25], an unsupervised optical flow model trained with photometric losses, and include a retrained version adapted to our setting with modality-specific encoders. Finally, we compare against recent diffusion-based correspondence methods, including DIFT [40], which leverages Stable Diffusion features, and SD-DINO [52], which combines Stable Diffusion and DINOv2 features for robust geometric and semantic matching. Across both RGB-Depth and RGB-Thermal tasks, our method significantly outperforms all baselines.

Qualitative results. We present geometric matching qualitative results showing the effectiveness of our method in cross-modal matching. As shown in Figures 3 and 4, our method consistently produces accurate and meaningful correspondences across RGB-Thermal and RGB-Depth pairs, significantly outperforming RAFT and GMFlow, which often yield noisy or incorrect matches.

Model Ablations. Table 2 presents ablations on the NYU-Depth and Thermal-IM datasets. Training with only intra-modal losses yields good performance within the same modality but fails on cross-modal tasks. In contrast, using only cross-modal losses leads to unstable training and poor convergence. Pretraining with intra-modal losses followed by joint intra- and cross-modal training significantly improves performance across both settings. Adding a smoothness loss further enhances cross-modal accuracy, highlighting its value as a regularization strategy.

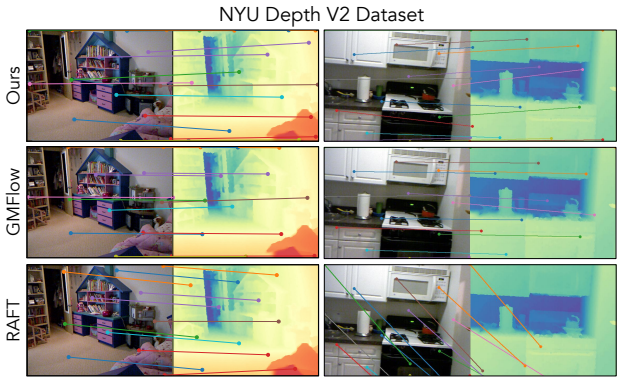

Figure 3. **RGB-Depth Matching.** We show qualitative comparisons on the NYU Depth V2 dataset. RAFT and GMFlow struggle to establish accurate correspondences in the depth domain, while our method successfully matches keypoints across RGB and depth images. The points shown are randomly sampled from the dataset annotations.

5.2. Semantic correspondence

Photo-Sketch Matching. In Table 3, we compare our method on PSC6K with several baselines from [28], including DINOv2+NN [29], which performs nearest-neighbor search on DINO features, and SD-DINO [52]. When using a CNN encoder, our method performs poorly due to the lack of semantic priors. Replacing it with DINOv2 features and fine-tuning the encoder leads to a significant performance boost, highlighting the importance of DINOv2's semantic pretraining. Our final model, with DINOv2 and full training, performs best among our variants and is competitive with state-of-the-art methods. Qualitative results are shown in Figure 5.

Cross-style Image Matching. We evaluate our model (using a DINOv2 encoder) alongside baselines including DINOv2+NN, DIFT, and SD-DINO in Table 4. We also compare with GeoAwareSC [53], a state-of-the-art supervised semantic correspondence method. Our model outperforms all baselines, including the supervised approach, albeit by a small margin. The performance gap between our method and DINOv2+NN demonstrates the effectiveness of our learned correspondence framework over simple feature similarity using pretrained DINO features. Qualitative results are presented in Figure 6.

6. Discussion

We have presented a modality-agnostic method for learning pixel-level correspondences between images taken using different visual sensors. Our method extends the contrastive

Table 2. **Model Ablations.** We evaluate the impact of different loss functions and show that pretraining with intra-modal losses followed by fine-tuning with all losses achieves the best performance for our method. X denotes D (Depth) or T (Thermal), based on the dataset.

Losses					NYU Depth, $< \delta^x_{\text{avg}} \uparrow$				Thermal-IM, $< \delta^x_{\text{avg}} \uparrow$	
					Intra-modal		Cross-modal		Cross-modal	
$\mathcal{L}^{\text{RGB}\to\text{RGB}}_{\texttt{intra-crw}}$	$\mathcal{L}^{\text{X}\to\text{X}}_{\texttt{intra-crw}}$	$\mathcal{L}^{\text{RGB}\to\text{X}}_{\texttt{inter-crw}}$	$\mathcal{L}^{\text{X}\to\text{RGB}}_{\texttt{inter-crw}}$	$\mathcal{L}_{\texttt{smooth}}$	RGB → RGB	D → D	RGB → D	D → RGB	RGB → T	T → RGB
✓	✓	-	-	-	78.8	49.2	2.5	2.2	4.9	6.2
-	-	✓	✓	-	18.5	4.4	5.6	4.5	6.2	8.3
✓	✓	✓	✓	-	80.4	61.1	19.1	21.1	30.2	38.5
✓	✓	✓	✓	✓	80.1	62.3	33.5	34.3	41.8	47.9

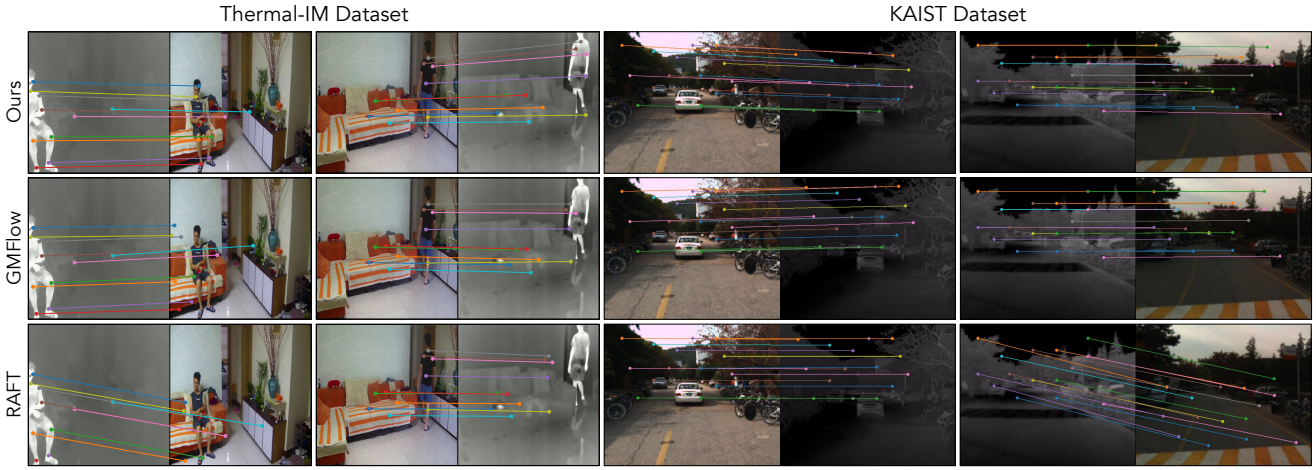

Figure 4. **RGB-Thermal Matching.** Qualitative comparisons on the Thermal-IM (left) and KAIST (right) datasets. Our method accurately tracks keypoints across RGB and thermal images, even in the presence of motion. In contrast, RAFT and GMFlow often fail to produce meaningful correspondences, frequently matching points to the same spatial location in the other modality. Here, we show all 10 annotated points from the dataset.

Table 3. **Photo-Sketch Matching**. Qualitative comparison on the PSC6K dataset [28], demonstrating the competitive performance of our method against several baselines.

Method	Trained on PSC6K	PCK-5	PCK-10
CNNGeo [33]		27.59	57.71
CNNGeo [33]	✓	19.19	42.57
DINOv2 + NN [29]		11.48	31.66
SD-DINO [52]		33.10	70.50
WeakAlign [33]		35.65	68.76
WeakAlign [33]	✓	43.55	78.60
NC-Net [34]		40.60	63.50
DCCNet [12]		42.43	66.53
PMD [24]		35.77	71.24
WarpC-Net [45]		48.79	71.43
WarpC-Net [45]	✓	56.78	79.70
PSCNet [28]	✓	**57.92**	**84.72**
Ours (CNN + Stage 1, 2, 3)	✓	26.22	60.89
Ours (DINO + Stage 2, 3)	✓	50.66	80.70
Ours (DINO + Stage 1, 2, 3)	✓	53.61	82.20

Table 4. **Cross-style Image Matching (Sec. 4.2).** Quantitative comparisons on the cross-style image matching task. We compare our method against a few self-supervised and one supervised baseline, and show that it outperforms both.

Method	SSL	PCK-5	PCK-10
DINOv2 + NN [29]	✓	23.13	56.52
DIFT [40]	✓	39.81	57.97
SD-DINO [52]	✓	60.14	80.72
GeoAwareSC [53]		68.30	79.67
Ours (DINO)	✓	**69.26**	**84.19**

random walk to perform cross-modal matching. A random walker transitions from patches in one modality to another, and then back, using transition probabilities that are produced by a global matching transformer [1, 36]. We avoid local minima in this process by including intra-modal random walks between frames of the same modality, as well as a spatial smoothness constraint. We evaluate our method on diverse tasks—including RGB-to-depth, RGB-to-thermal, photo-to-sketch, and cross-style image matching—and find

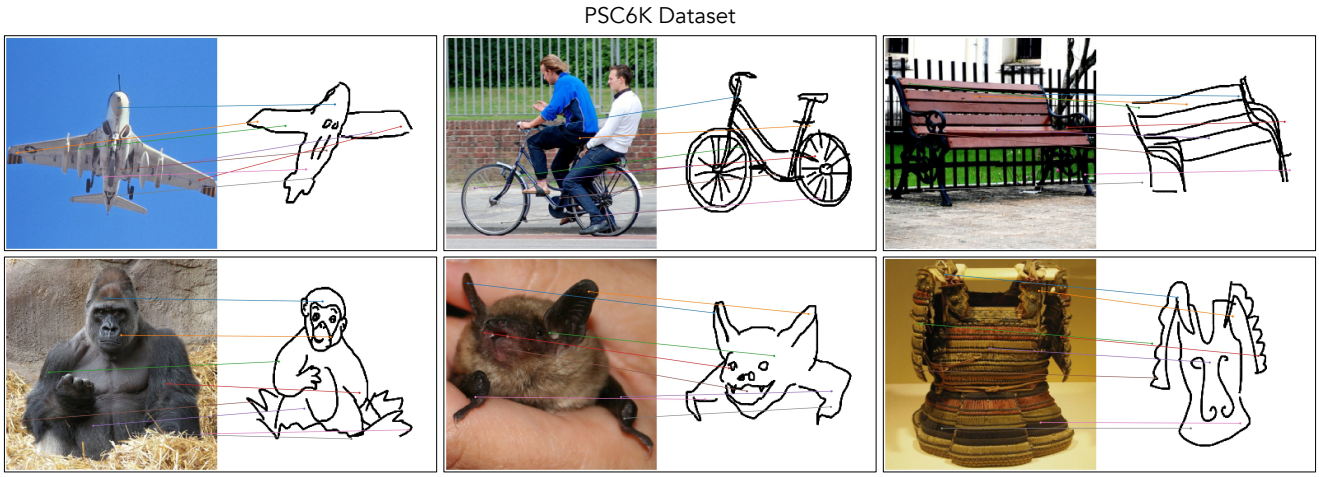

Figure 5. **Photo-Sketch Matching.** Qualitative results of our method on the PSC6K dataset [28]. For each image pair, all annotated points from the dataset are used as query keypoints.

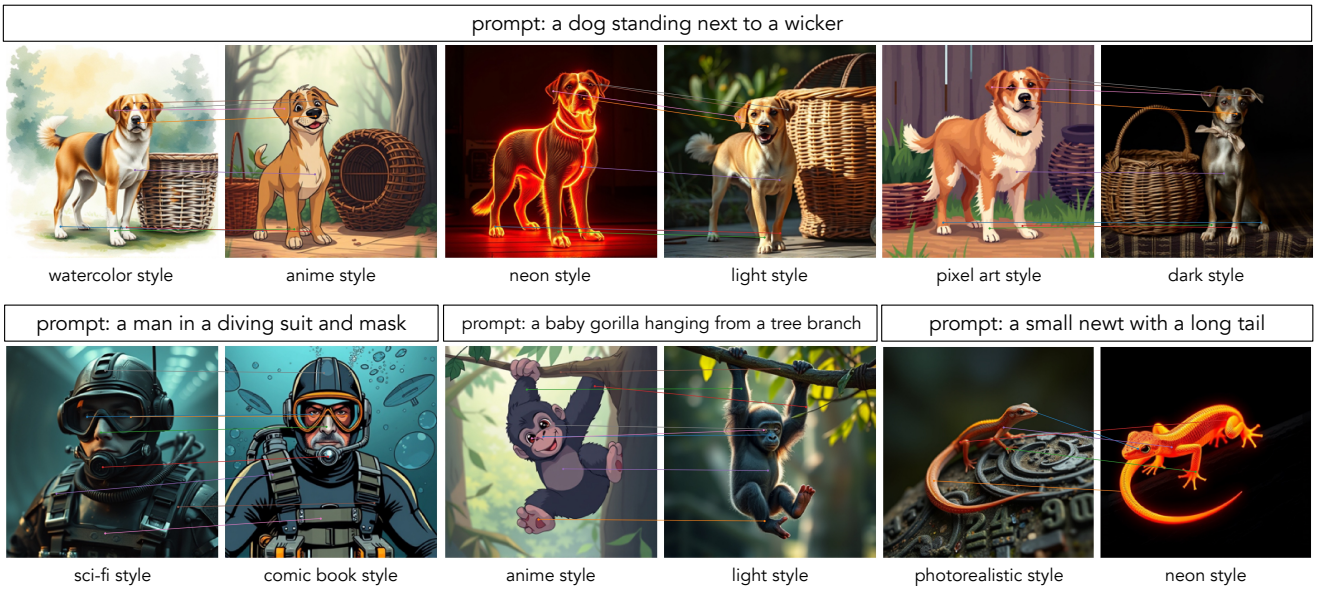

Figure 6. **Cross-style Image Matching.** We show qualitative results of our method on matching images across different styles. The images are generated using the Flux [19] model, with the prompts and styles used for generation shown above.

that it significantly outperforms existing methods on geometric correspondence and is competitive with state-of-the-art methods on semantic matching. We see our work as opening new directions in multimodal matching, making it significantly easier to match images without hand-crafted measures of visual similarity. Our approach can be trained entirely using multimodal video data, which can be acquired at scale. We also see our work as experimentally demonstrating the generality of self-supervised matching methods, which can be applied in scenarios that would be difficult to simulate or acquire labeled data for.

Limitations. We have demonstrated our model for four domains, RGB-depth, RGB-thermal, photo-sketch, cross-style matching. While the model successfully handles these cases, it is possible that other modalities would present additional challenges, especially if they lacked other distinctive visual structures that may be used in matching, such as occluding contours (which are visible in all 3 current modalities, RGB, Thermal, Depth). In cross-style image matching, we observe occasional failures, such as confusion between left and right limbs in animals (see Figure 6, top-right). While our method places few assumptions on the underlying signal (e.g., no hand-crafted photo-consistency assumption), both of our datasets include RGB images due to their ubiquity.

Acknowledgements. We thank Xuanchen Lu, Adam Harley, Qianqian Wang, Daniel Geng, Ziyang Chen, Jeongsoo Park, Yiming Dou and the reviewers for the valuable discussion and feedback. This work was supported by Toyota Research Institute and Cisco Systems.

References

[1] Moab Arar, Yiftach Ginger, Dov Danon, Amit H Bermano, and Daniel Cohen-Or. Unsupervised multi-modal image registration via geometry preserving image-to-image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020. 1, 2, 3, 5, 6, 7

[2] Zhangxing Bian, Allan Jabri, Alexei A Efros, and Andrew Owens. Learning pixel trajectories with multiscale contrastive random walks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6508–6519, 2022. 2, 4

[3] 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. 2

[4] Carl Doersch, Ankush Gupta, Larisa Markeeva, Adrià Recasens, Lucas Smaira, Yusuf Aytar, João Carreira, Andrew Zisserman, and Yi Yang. Tap-vid: A benchmark for tracking any point in a video. *arXiv preprint arXiv:2211.03726*, 2022. 2, 3, 5, 13

[5] Carl Doersch, Yi Yang, Mel Vecerik, Dilara Gokay, Ankush Gupta, Yusuf Aytar, Joao Carreira, and Andrew Zisserman. Tapir: Tracking any point with per-frame initialization and temporal refinement. *arXiv preprint arXiv:2306.08637*, 2023. 3

[6] Yiming Dou, Fengyu Yang, Yi Liu, Antonio Loquercio, and Andrew Owens. Tactile-augmented radiance fields. *Computer Vision and Pattern Recognition (CVPR)*, 2024. 1

[7] Daniel Gordon, Kiana Ehsani, Dieter Fox, and Ali Farhadi. Watching the world go by: Representation learning from unlabeled videos. *arXiv preprint arXiv:2003.07990*, 2020. 3

[8] Klaus Greff, Francois Belletti, Lucas Beyer, Carl Doersch, Yilun Du, Daniel Duckworth, David J Fleet, Dan Gnanapragasam, Florian Golemo, Charles Herrmann, et al. Kubric: A scalable dataset generator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3749–3761, 2022. 3

[9] Adam W Harley, Zhaoyuan Fang, and Katerina Fragkiadaki. Particle video revisited: Tracking through occlusions using point trajectories. In *European Conference on Computer Vision*, pages 59–75. Springer, 2022. 3

[10] 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. 2

[11] Xixi Hu, Ziyang Chen, and Andrew Owens. Mix and localize: Localizing sound sources in mixtures. *Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[12] Shuaiyi Huang, Qiuyue Wang, Songyang Zhang, Shipeng Yan, and Xuming He. Dynamic context correspondence network for semantic alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2010–2019, 2019. 7

[13] Soonmin Hwang, Jaesik Park, Namil Kim, Yukyung Choi, and In So Kweon. Multispectral pedestrian detection: Benchmark dataset and baselines. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 4, 15

[14] 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. 1, 2, 3, 4

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

[16] Joel Janai, Fatma Guney, Anurag Ranjan, Michael Black, and Andreas Geiger. Unsupervised learning of multi-frame optical flow with occlusions. In *Proceedings of the European conference on computer vision (ECCV)*, pages 690–706, 2018. 2, 3

[17] Rico Jonschkowski, Austin Stone, Jonathan T Barron, Ariel Gordon, Kurt Konolige, and Anelia Angelova. What matters in unsupervised optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 557–572. Springer, 2020. 2, 3, 4

[18] Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker: It is better to track together. In *Proc. ECCV*, 2024. 5, 14

[19] Black Forest Labs. Flux. `https://github.com/black-forest-labs/flux`, 2024. 1, 2, 5, 8

[20] Zihang Lai, Erika Lu, and Weidi Xie. Mast: A memory-augmented self-supervised tracker. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6479–6488, 2020. 3

[21] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 5

[22] Minhao Li, Zheng Qin, Zhirui Gao, Renjiao Yi, Chenyang Zhu, Yulan Guo, and Kai Xu. 2d3d-matr: 2d-3d matching transformer for detection-free registration between images and point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14128–14138, 2023. 1, 3

[23] Xueting Li, Sifei Liu, Shalini De Mello, Xiaolong Wang, Jan Kautz, and Ming-Hsuan Yang. Joint-task self-supervised learning for temporal correspondence. *Advances in Neural Information Processing Systems*, 32, 2019. 2, 3

[24] Xin Li, Deng-Ping Fan, Fan Yang, Ao Luo, Hong Cheng, and Zicheng Liu. Probabilistic model distillation for semantic correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7505–7514, 2021. 7

[25] Liang Liu, Jiangning Zhang, Ruifei He, Yong Liu, Yabiao Wang, Ying Tai, Donghao Luo, Chengjie Wang, Jilin Li, and

Feiyue Huang. Learning by analogy: Reliable supervision from transformations for unsupervised optical flow estimation. In *IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, 2020. 5, 6

[26] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 12

[27] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 12

[28] Xuanchen Lu, Xiaolong Wang, and Judith E Fan. Learning dense correspondences between photos and sketches. In *International Conference on Machine Learning*, pages 22899–22916. PMLR, 2023. 1, 3, 4, 5, 6, 7, 8, 15

[29] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 1, 2, 4, 6, 7

[30] 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. 12

[31] 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*. PMLR, 2021. 2

[32] E. Riba, D. Mishkin, D. Ponsa, E. Rublee, and G. Bradski. Kornia: an open source differentiable computer vision library for pytorch. In *Winter Conference on Applications of Computer Vision*, 2020. 12

[33] 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. 7

[34] Ignacio Rocco, Mircea Cimpoi, Relja Arandjelović, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Neighbourhood consensus networks. *Advances in neural information processing systems*, 31, 2018. 7

[35] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015. 5

[36] Ayush Shrivastava and Andrew Owens. Self-supervised any-point tracking by contrastive random walks. In *European Conference on Computer Vision (ECCV)*, 2024. 2, 3, 4, 5, 6, 7, 12

[37] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part V 12*, pages 746–760. Springer, 2012. 4, 5, 15

[38] Austin Stone, Daniel Maurer, Alper Ayvaci, Anelia Angelova, and Rico Jonschkowski. Smurf: Self-teaching multi-frame unsupervised raft with full-image warping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3887–3896, 2021. 2, 3

[39] Deqing Sun, Daniel Vlasic, Charles Herrmann, Varun Jampani, Michael Krainin, Huiwen Chang, Ramin Zabih, William T Freeman, and Ce Liu. Autoflow: Learning a better training set for optical flow. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10093–10102, 2021. 3

[40] Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 5, 6, 7

[41] Yansong Tang, Zhenyu Jiang, Zhenda Xie, Yue Cao, Zheng Zhang, Philip HS Torr, and Han Hu. Breaking shortcut: Exploring fully convolutional cycle-consistency for video correspondence learning. *arXiv preprint arXiv:2105.05838*, 2021. 2, 4

[42] Zitian Tang, Wenjie Ye, Wei-Chiu Ma, and Hang Zhao. What happened 3 seconds ago? inferring the past with thermal imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17111–17120, 2023. 4, 15

[43] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 402–419. Springer, 2020. 3, 5, 6

[44] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, 2020. 2

[45] 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. 7

[46] 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. 3

[47] Haiping Wang, Yuan Liu, Bing Wang, Yujing Sun, Zhen Dong, Wenping Wang, and Bisheng Yang. Freereg: Image-to-point cloud registration leveraging pretrained diffusion models and monocular depth estimators. *arXiv preprint arXiv:2310.03420*, 2023. 1, 3

[48] 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. 2

[49] Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, and Dacheng Tao. Gmflow: Learning optical flow via global matching. In *Proceedings of the IEEE/CVF conference on*

computer vision and pattern recognition, pages 8121–8130, 2022. 2, 3, 4, 5, 6

[50] 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. 3

[51] Jason J. Yu, Adam W. Harley, and Konstantinos G. Derpanis. Back to basics: Unsupervised learning of optical flow via brightness constancy and motion smoothness. In *Computer Vision - ECCV 2016 Workshops, Part 3*, 2016. 2, 3

[52] Junyi Zhang, Charles Herrmann, Junhwa Hur, Luisa Polania Cabrera, Varun Jampani, Deqing Sun, and Ming-Hsuan Yang. A Tale of Two Features: Stable Diffusion Complements DINO for Zero-Shot Semantic Correspondence. *arXiv preprint arxiv:2305.15347*, 2023. 5, 6, 7

[53] Junyi Zhang, Charles Herrmann, Junhwa Hur, Eric Chen, Varun Jampani, Deqing Sun, and Ming-Hsuan Yang. Telling left from right: Identifying geometry-aware semantic correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3076–3085, 2024. 6, 7

[54] Yang Zheng, Adam W. Harley, Bokui Shen, Gordon Wetzstein, and Leonidas J. Guibas. Pointodyssey: A large-scale synthetic dataset for long-term point tracking. In *ICCV*, 2023. 5, 13

[55] Tinghui Zhou, Philipp Krahenbuhl, Mathieu Aubry, Qixing Huang, and Alexei A Efros. Learning dense correspondence via 3d-guided cycle consistency. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 117–126, 2016. 2

[56] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017. 6

[57] Yuliang Zou, Zelun Luo, and Jia-Bin Huang. Df-net: Unsupervised joint learning of depth and flow using cross-task consistency. In *Proceedings of the European conference on computer vision (ECCV)*, pages 36–53, 2018. 2, 3

A.1. Implementation details

Here, we present the model architecture in detail and the hyperparameters used during training and evaluation.

A.1.1. Model Architecture

The input to our model is a pair of images of modalities m_1, m_2 $(I_1^{m_1}, I_2^{m_2})$ of size (H, W) and the output is an affinity matrix A of size (H, W, H, W) where $A[i, j, k, l]$ represents the probability of pixel (i, j) in $I_1^{m_1}$ transitioning to pixel (k, l) in $I_2^{m_2}$.

CNN Encoder. Our feature encoder is similar to GMRW [36], we extract the features at 1/4 scale. We train a different encoder for each modality. The CNN architecture is as follows:

Layer	Output	Details
`input`	$H, W, 3$	
`conv1`	$H/2, W/2, 64$	kernel 7×7, stride 2
`res1`	$H/2, W/2, 96$	kernel 3×3, stride 1
`res2`	$H/2, W/2, 96$	kernel 3×3, stride 1
`res3`	$H/4, W/4, 128$	kernel 3×3, stride 2
`res4`	$H/4, W/4, 128$	kernel 3×3, stride 1
`res5`	$H/4, W/4, 128$	kernel 3×3, stride 1
`res6`	$H/4, W/4, 128$	kernel 3×3, stride 1

Table 5. CNN Architecture

Transformer. In our transformer model, we stack 6 layers of transformer blocks. Each block consists of a self-attention, cross-attention and feed-forward network. The transformer feature dimension is 128, and the feed-forward network, consisting of 2 linear layers, expands the dimension by $4\times$. For efficiency, we use shifted local window attention [26] where the attention window is $1/16$th the size of the input image.

A.1.2. Training Details

We train our network in Pytorch [30], using the AdamW [27] optimizer with the constant learning rate of 1.6×10^{-4}. We train on 4 A40 GPUs with an effective batch size 24. For RGB-to-Depth matching, we train the model in three stages: Stage 1 for 50K iterations, Stage 2 for 100K iterations, and Stage 3 for an additional 20K iterations. This full schedule takes approximately 7 days. For RGB-to-Thermal matching, we train Stage 1 for 30K iterations, Stage 2 for 100K iterations, and Stage 3 for another 20K iterations. For Photo-to-Sketch matching, we initialize the visual encoder with a pretrained DINOv2 model, and train for 12K, 10K, and 28K iterations in Stages 1, 2, and 3 respectively with a learning rate of 1.6×10^{-7}. During Stage 3, we linearly increase the weight of the smoothness loss λ_s over the first 10K iterations.

Data augmentation. For data augmentation, we perform different resize-crop transformations T^f and T^b on the forward and backward cycle in the contrastive random walk. To implement this, we use the `kornia` [32] library. We use it to apply the same transformation (T^f/T^b) to the entire forward/backward cycle and also use T^f and T^b to compute T_f^b and warp the label i.e. $T_f^b(I)$. We use `RandomResizedCrop` with range of size ratio of cropped area set to $(0.08, 1.0)$ and the range of aspect ratio of cropped area set to $(0.7, 1.3)$. We set these augmentation hyperparameters by following GMRW [36].

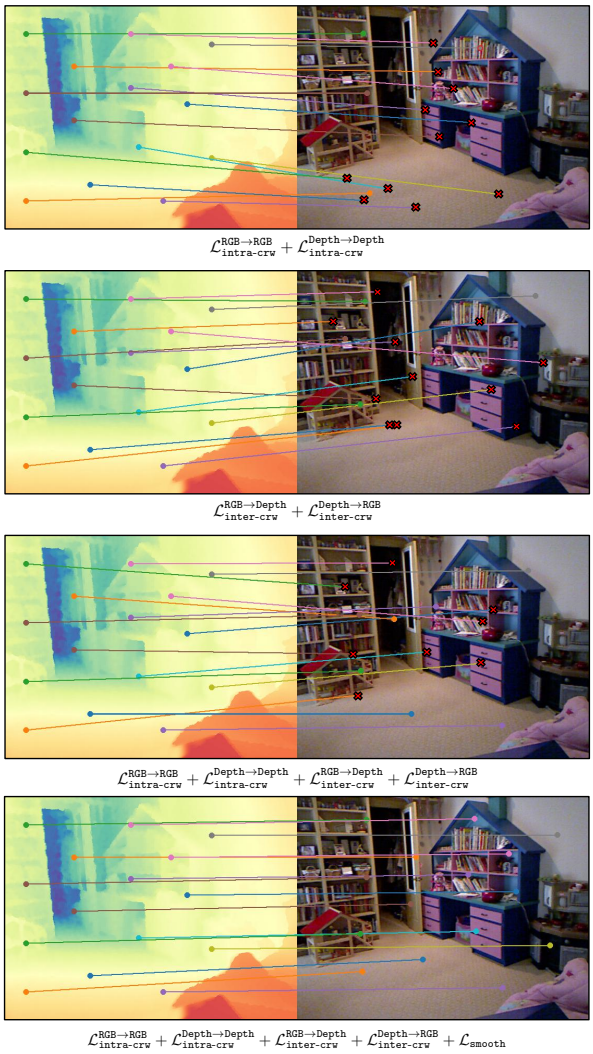

Figure 7. **Qualitative examples for RGB-Depth matching with different losses. Zoom in for details.** Points with red cross show the incorrect correspondences (not within 50 px distance of ground truth).

A.1.3. Hyperparameters

Here are the rest of the hyperparameters used in the model and in training.

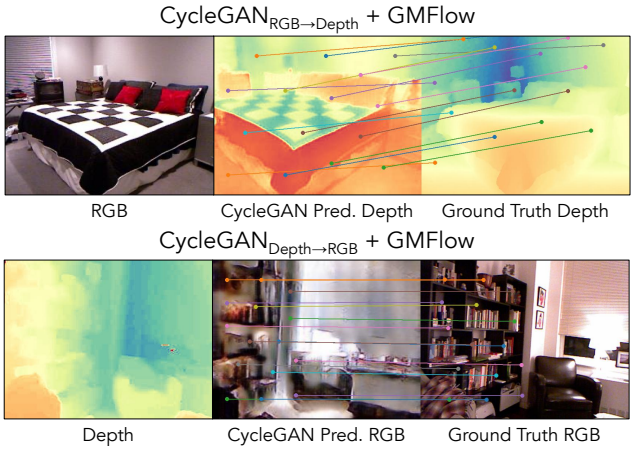

Figure 8. **CycleGAN+GMFlow Baseline.** We show results for a CycleGAN-based baseline that performs image translation from RGB to the depth domain, followed by matching in the depth space. We also present the reverse setup, where images are translated from depth to RGB, and matching is performed in the RGB space. In both cases, the presence of translation artifacts leads to inaccurate matching.

Hyperparameter	**Value**
Learning rate schedule	Constant
Learning rate	1.6×10^{-4}
Optimizer	AdamW
Weight Decay	10^{-4}
Temperature τ	$\sqrt{128}$
Effective Batch size	24
Time stride k	Randomly chosen from $[1, 10]$
Smoothness loss weight λ_s	linear increase of $[0, 1]$ over $[100\text{k}, 120\text{k}]$ steps for RGB-Depth

Table 6. **Hyperparameters for training the model.**

A.2. Dataset details

NYU Depth Evaluation Dataset. We leverage the fact that high-quality pseudo ground-truth correspondences can be generated by matching visual frames using off-the-shelf multi-frame tracking methods, in combination with the known calibration between RGB and depth sensors. Specifically, we use PIP++ [54] to track points across video clips of length 10. We retain only those tracks that are consistently visible across all frames, reducing the impact of occlusions and tracking failures. We further perform manual visual inspection to ensure the tracks are of reasonable quality and do not drift from their original trajectories. Since RGB and depth frames are spatially aligned via sensor calibration, these tracks can serve as ground truth for RGB-RGB, Depth-Depth, and RGB-Depth matching. The dataset comprises 250 video clips, each with 10 frames and an average of 688 annotated tracks per clip.

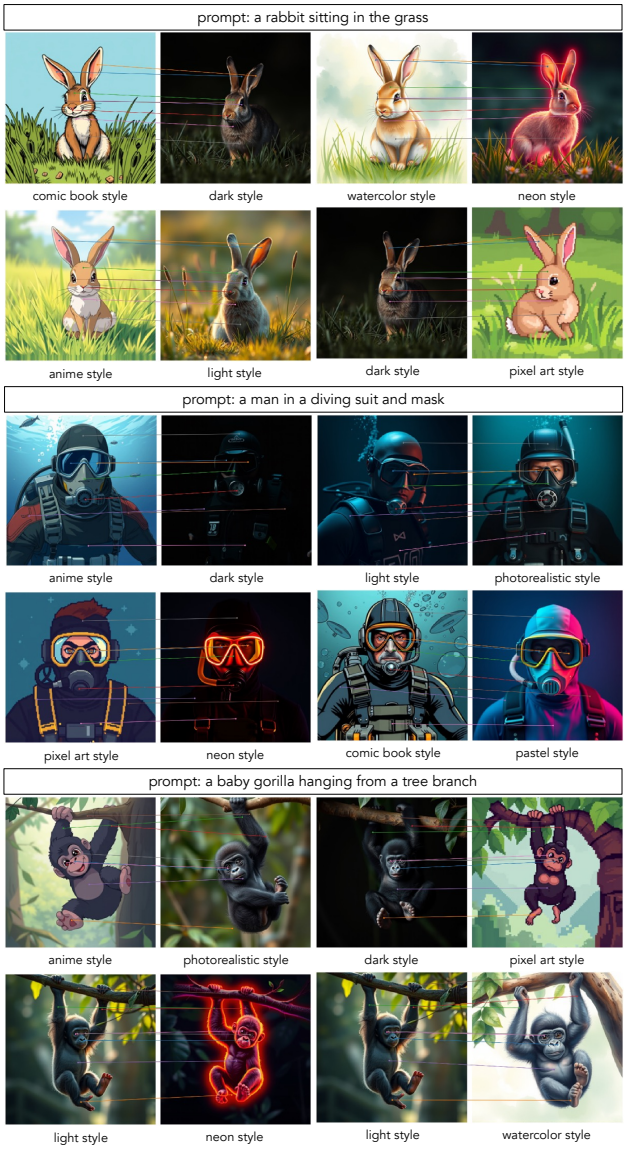

Figure 9. **Cross-style Matching.** We show qualitative results from our model on matching images generated in different styles by a image generation model.

Thermal-IM Evaluation Datasets. In the Thermal-IM dataset, RGB and thermal images are not spatially aligned. To establish ground-truth correspondences, we follow a protocol similar to TAP-Vid [4], manually annotating 100 RGB-Thermal image pairs across 5 timesteps with 10 keypoints per frame, resulting in a total of 1000 evaluation points. For the KAIST dataset, where RGB and thermal images are aligned, we adopt a strategy similar to that used for the NYU Depth dataset. Ground-truth correspondences for RGB-RGB, RGB-Thermal, and Thermal-Thermal matching are obtained using

CoTracker [18] on RGB sequences for videos of length 5.

A.3. Qualitative examples

We present additional qualitative results for our model trained on RGB-Depth matching in Figure 10, and RGB-Thermal matching in Figure 11. For semantic correspondence tasks, we include more examples for Photo-to-Sketch matching in Figure 12, and Cross-style matching in Figure 9.

We also provide qualitative results for the CycleGAN+GMFlow baseline from Table 1, shown in Figure 8. In this setup, a CycleGAN is trained to translate RGB images to depth, and the pretrained GMFlow model is applied to estimate correspondences between the generated and target depth images. Notably, the CycleGAN often assigns inconsistent depth values to pixels with different colors, even when they belong to the same physical depth plane, resulting in inaccurate matches. We also present results from the reverse setup, where depth images are translated to RGB, and matching is performed in the RGB space. In this case, the generated RGB images often contain significant artifacts and fail to accurately reconstruct the original appearance, degrading correspondence quality.

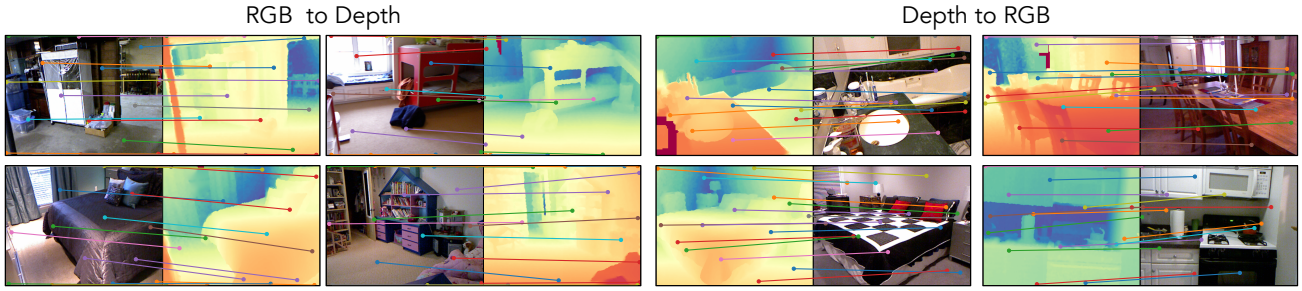

Figure 10. **RGB-Depth Matching.** We show qualitative results from our model on NYU-Depth V2 dataset [37].

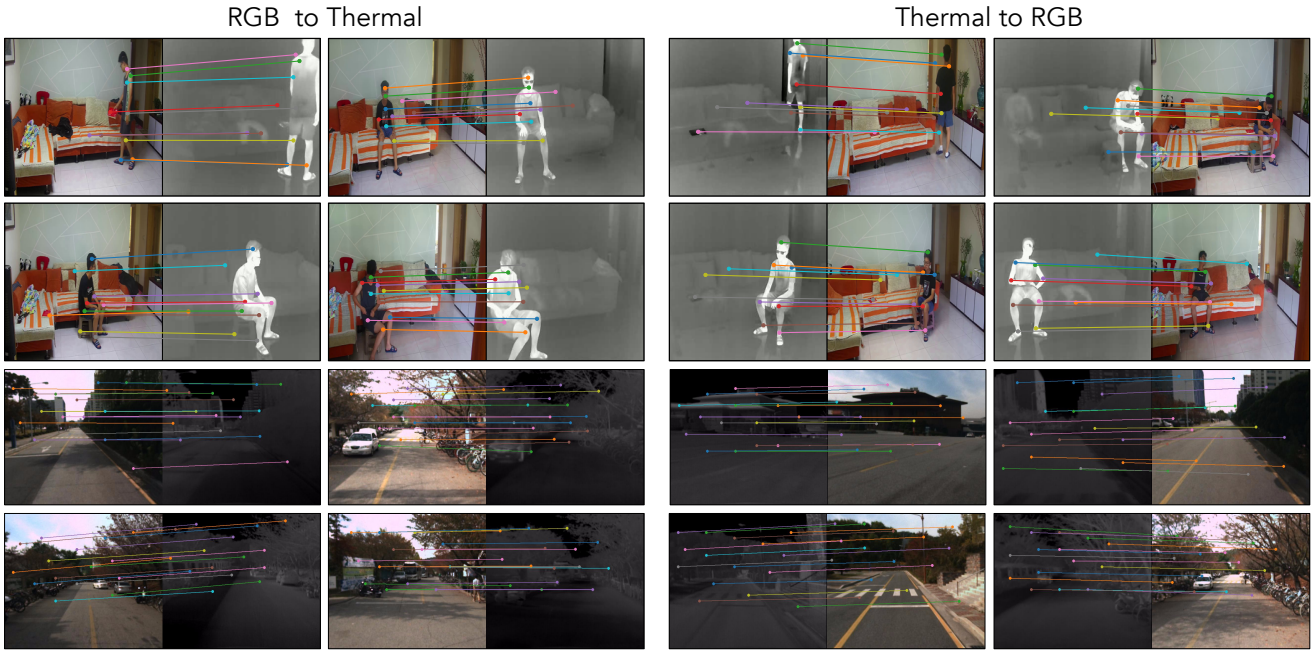

Figure 11. **Thermal-IM Matching.** We show qualitative results from our model on Thermal-IM [42] and KAIST datasets [13].

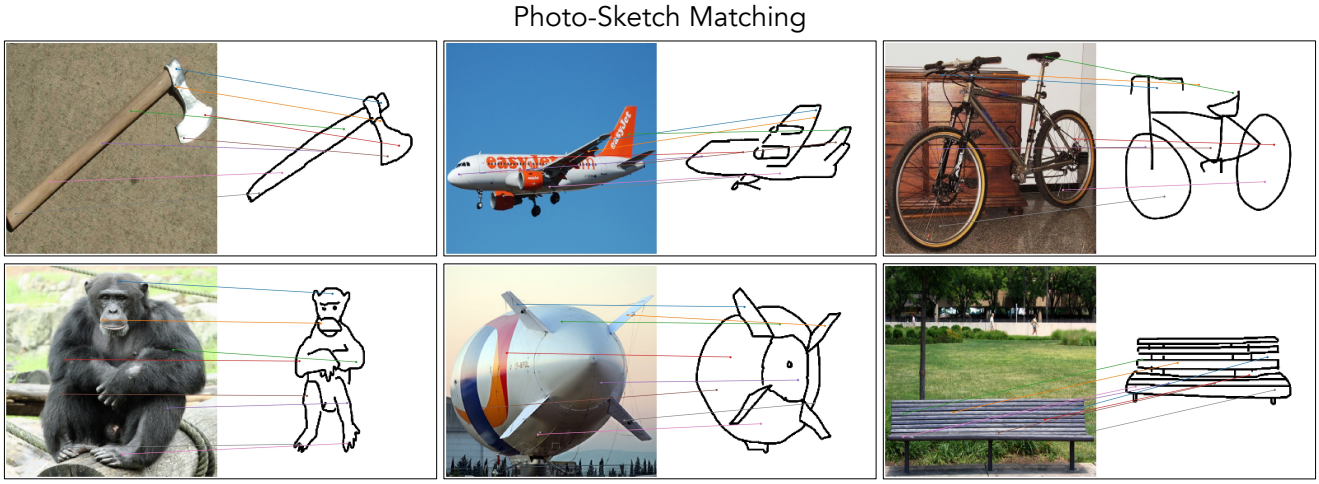

Figure 12. **Photo-Sketch Matching.** We show qualitative results from our model on PSC6K dataset [28].