

When Lines Meet Textures: Spatial-Frequency Aligned Diffusion Features for Cross-Sparsity Correspondence

Mingrui Zhu Fengzhi Wang Xin Wei* Jun Wang Nannan Wang* Xinbo Gao
 Xidian University

Abstract

Establishing accurate correspondence between sparse line representations and rich textured imagery remains a formidable challenge. While diffusion features excel in semantic correspondence, they struggle to bridge the fundamental gap between abstract sketches and texture-rich photographs. We identify two critical disparities: **spatial domain misalignment** from structural abstraction differences, and **frequency domain inconsistencies** from texture density variations. Based on this analysis, we propose **SFA-DIFT**, a novel approach that learns **spatial-frequency aligned diffusion features** for robust cross-modal correspondence. Unlike previous methods focusing solely on spatial alignment, our key innovation performs dual-domain alignment by learning unified clean diffusion features while strategically aggregating low-frequency components in the frequency domain. This comprehensive spatial-frequency alignment enables equitable understanding between sparse abstractions and rich textures. To validate our approach, we extend the existing sketch-photo correspondence dataset (PSC6K) by generating multi-style textured imagery, creating MS-PSC6K, a comprehensive correspondence benchmark. Extensive experiments demonstrate that SFA-DIFT achieves state-of-the-art performance, delivering substantial improvements with an average of **0.87%** on PCK@1, **2.20%** on PCK@5, and **0.95%** on PCK@10 over previous best methods, validating the effectiveness and robustness of our dual-domain alignment approach. Codes are available on <https://github.com/Mofr77/SFA-DIFT>.

1. Introduction

Establishing semantic correspondence between simple line drawings and complex textured images (as shown in Fig. 1(a)) presents a fundamental challenge. While humans effortlessly establish correspondences between sparse lines and rich photographs through shared semantic understanding, computer vision algorithms [15, 19] struggle to navi-

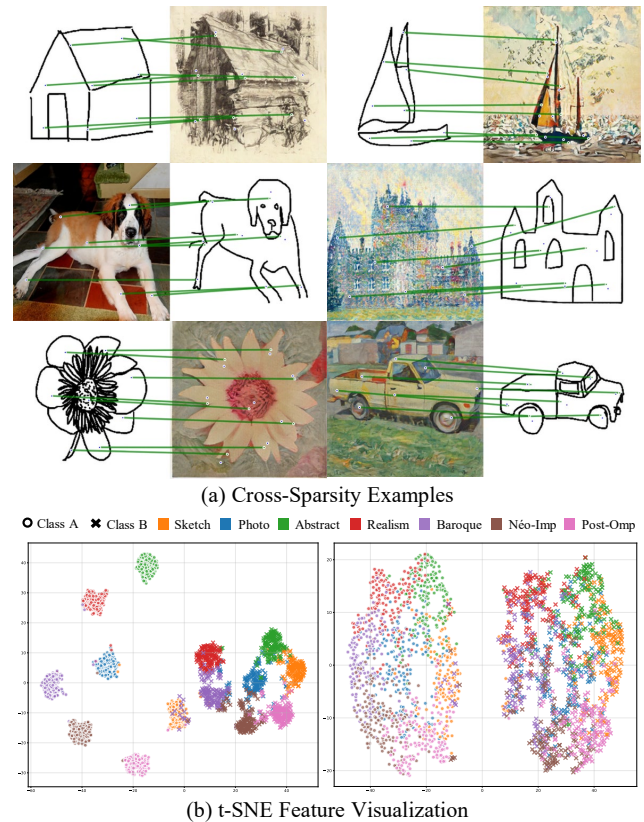


Figure 1. **Cross-sparsity correspondence examples and feature visualization.** (a) Representative cross-sparsity correspondence examples from extended MS-PSC6K. (b) t-SNE [20] visualization comparing cross-modal features from Stable Diffusion (SD) and our SFA-DIFT method. Best viewed when zoomed in.

*Corresponding authors.

Existing methods exhibit notable limitations in cross-modal correspondence between sparse sketches and texture-rich images: dedicated encoders [17, 19] suffer from poor generalization due to the scarcity of sketch data, while text-prompt-based learning [3, 14, 26] fails to address the underlying feature distribution discrepancies required for dense correspondence estimation. Although Stable Diffusion (SD) models [24] show promise [31, 36], its features remain heavily biased toward texture and appearance, leading to pronounced distribution shift on pure line drawings. Systematic analyses [11, 15] of these fundamental cross-modal disparities remain limited. Therefore, bridging this representational gap and aligning cross-modal semantics constitute the core challenge of this work.

Our analysis reveals two key challenges that limit diffusion features for accurate correspondence between sparse line drawings and texture-rich images. First, the standard forward noising process introduces noise, degrading feature purity, which is particularly detrimental to clean line representations. CleanDIFT [30] addresses this issue by aligning features across timesteps, allowing for direct feature extraction from clean images. Nevertheless, its effectiveness on sketches remains limited. Second, diffusion models struggle to capture low-frequency semantic information from sketches. While prior work [15] utilized CLIP [23] to modulate low-frequency semantic signals, the method is insufficient to fully bridge the modality gap for establishing dense correspondences.

Motivated by these observations, we propose **SFA-DIFT**, a novel approach that learns spatial-frequency aligned diffusion features for correspondence between sparse line representations and rich textured imagery. Unlike previous methods that exclusively focus on spatial domain feature alignment, our method addresses the fundamental disparities between sparse line and rich textured imagery across both spatial and frequency domains through a dual-domain alignment strategy. To address the spatial misalignment of lines, our approach builds upon a pre-trained CleanDIFT backbone and employs parameter-efficient Low-Rank Adaptation (LoRA) [8] to enhance sketch comprehension through unsupervised fine-tuning on sketch data. This process learns **unified clean diffusion features (Unified CleanDIFT)** that map both modalities into a shared semantic space while suppressing modality-specific noise. This guides the model to focus on the core geometric structure, making the representation robust to the visual differences between sparse lines and dense textures.

To achieve comprehensive spatial-frequency alignment, we further introduce a **low-frequency feature aggregation (LoFFA)** module that strategically processes low-frequency components. Specifically, the module decomposes multi-scale features via Discrete Wavelet Transform (DWT) and selectively applies convolution and aggregation to the low-

frequency components (LL sub-band), directing the model to focus on shared structural patterns that are crucial for cross-modal correspondence. This dual-domain alignment approach enables our method to bridge both the spatial domain gaps caused by visual appearance differences and frequency domain disparities arising from sketch sparsity. As visualized in Fig. 1(b), the t-SNE [20] visualization reveals that our method achieves superior cross-modal clustering by semantic classes rather than modality types, effectively mapping sparse line drawings and texture-rich images into a unified semantic space, while SD shows high modality discrimination with distinct modality-specific clustering. Main contributions are as follows:

- We conduct a **systematic analysis revealing fundamental discrepancies** between sparse line drawings and texture-rich images **across both spatial and frequency domains**, and demonstrate that existing diffusion features fail to bridge these multi-domain gaps, resulting in inaccurate correspondence estimation.
- We propose SFA-DIFT, which integrates **unified clean diffusion features** via parameter-efficient LoRA fine-tuning and **low-frequency feature aggregation**, achieving spatial-frequency alignment for robust correspondence estimation between sparse lines and rich textures.
- We introduce MS-PSC6K, an **extended multi-style version of PSC6K**, providing a comprehensive evaluation protocol for cross-modal correspondence. Extensive experiments on both PSC6K and MS-PSC6K demonstrate that our approach achieves state-of-the-art performance, with **substantial improvements over previous methods** in sparse line-rich texture correspondence.

2. Understanding Line-Texture Disparities

We decompose the fundamental correspondence challenge between sparse line drawings and texture-rich images into two critical perspectives: spatial disparities arising from line drawing abstraction, and frequency disparities stemming from texture density variations. Through both conceptual analysis and formal characterization, we show how these domain-specific gaps manifest in cross-sparsity correspondence tasks. We then critically examine the limitations of existing diffusion-based methods in bridging these multi-faceted disparities, establishing the foundation for our dual-domain alignment approach.

2.1. Spatial Perspective

The first fundamental gap stems from line drawings being sparse spatial abstractions of object structure. Unlike photographs that capture dense surface detail, sketches are cognitively selective representations [1] that preserve only principal contours while omitting textures and fine geometric features. This abstraction creates two primary chal-

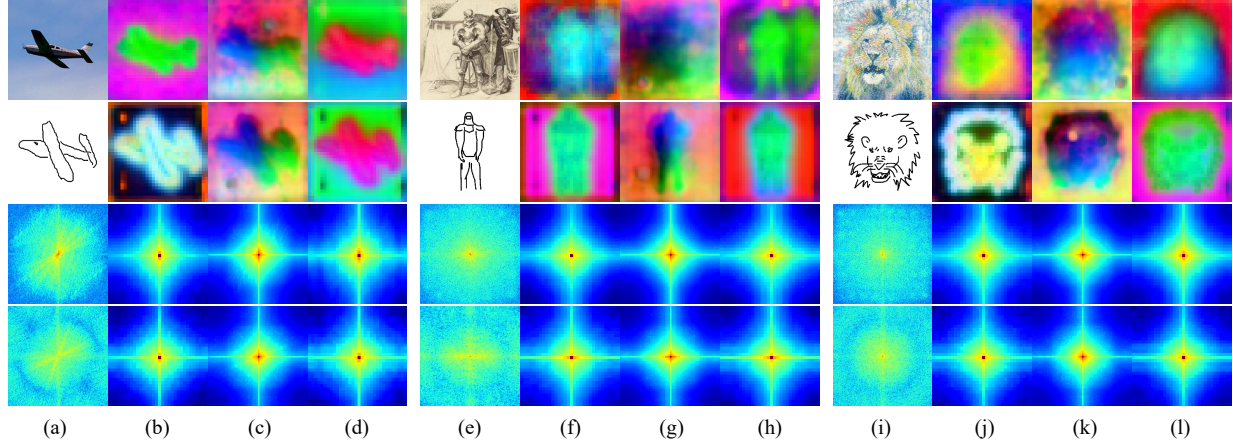


Figure 2. **Cross-sparsity feature analysis in spatial and frequency domains on MS-PSC6K dataset.** Rows 1-2: Image pairs and PCA-visualized features. Rows 3-4: Frequency spectra of extracted features. Columns show different methods: (a, e, i) original images, (b, f, j) SD, (c, g, k) SketchFusion, (d, h, l) CleanDIFT.

lenges: “sparsity”, where numerous textured regions lack corresponding sketch elements [34], and “misalignment”, where line positions may deviate from true object boundaries [29], introducing systematic spatial shifts. Effective correspondence thus requires higher-level understanding [10] of shared structural topology.

2.2. Frequency Perspective

The second gap emerges from fundamental frequency-domain disparities between texture-rich images and sparse line drawings. We quantify this difference through log-magnitude Fourier spectrum analysis (Fig. 2(a, e, i)). Textured images exhibit energy broadly distributed across frequencies, with low frequencies capturing macro shape, mid frequencies encoding surface patterns, and high frequencies preserving fine details [25]. Conversely, sketches display distinctly different spectral characteristics: dominant low-frequency plateaus from overall structure, broadband high-frequency spikes from sharp edges, and notably sparse mid-frequency content [7]. This spectral mismatch creates three critical challenges: cross-modal feature alignment becomes problematic in frequency space, texture-dependent representations fail on sketch inputs, and correspondence ambiguity is amplified across domains.

2.3. Limitations of Prior Efforts

Early hand-crafted descriptors like SIFT [16] and HOG [18] lack semantic robustness for extreme representation gaps. Recent deep models [4, 12] and advanced correlation operators [21] achieve better performance on appearance-consistent pairs, but struggle when modalities exhibit fundamental differences. The core issue lies in training bias: general models trained on texture-rich images develop inherent preferences for photometric features. Our PCA anal-

ysis of SD features [24] (Fig. 2(b, f, j)) confirms this limitation: when processing sparse sketches, SD produces pronounced “feature voids” and noise-like artifacts in feature maps, disrupting semantic localization and creating dispersed, entangled representations. This occurs because the model searches for textures in line-only inputs, causing spatial correspondence collapse. The t-SNE visualization (Fig. 1(b) left) further demonstrates this bias through distinct modality-specific clustering, indicating high modality discrimination that hinders cross-modal understanding.

SketchFusion [15] attempts to bridge modality differences by modulating signals with CLIP [23]. While this approach injects high-level semantics and reduces spatial discrepancies by aligning sketch and photo features semantically, it addresses symptoms rather than root causes. The method functions as a high-level patch that cannot resolve fundamental low-level frequency mismatches, leaving feature embeddings largely separated (Fig. 2(c, g, k)).

CleanDIFT [30] addresses SD’s limitations through denoising, achieving markedly improved feature smoothness and noise suppression (Fig. 2(d, h, l)). For sketch contours [2], CleanDIFT demonstrates “feature densification,” effectively inferring missing spatial structures from sparse line drawings, a clear improvement over SD’s feature discontinuity. This enhancement manifests as strengthened low-frequency components in both image and sketch features, visible as brighter, broader central regions in frequency spectra, reflecting preference for global shape information while suppressing high-frequency noise. However, this denoising mechanism introduces a critical trade-off: overly smoothed feature maps attenuate semantic expressiveness, particularly for sketches. The resulting incomplete cross-modal activation fusion reveals that the domain gap persists despite CleanDIFT’s improvements.

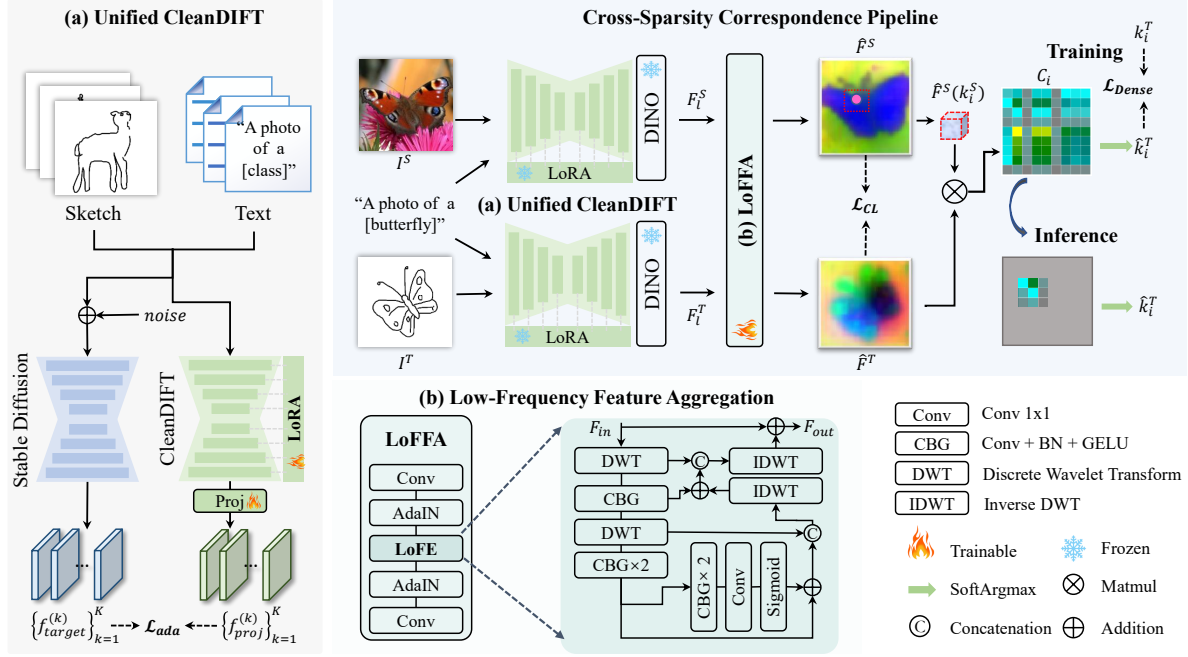


Figure 3. Our SFA-DIFT framework first creates a **Unified CleanDIFT** extractor via LoRA fine-tuning, and then uses its frozen features to train a **Low-Frequency Feature Aggregation (LoFFA)** module that amplifies low-frequency components for robust correspondence.

3. Method

Based on the dual-domain gap analysis in Sec. 2, we propose **Spatial-Frequency Aligned Diffusion Features (SFA-DIFT)** to establish effective correspondence between sparse lines and texture-rich imagery. Our approach addresses both gaps through two complementary stages: (1) **Unified Clean Diffusion Feature Learning** bridges spatial domain disparities via parameter-efficient fine-tuning, and (2) **Low-Frequency Feature Aggregation** resolves frequency domain mismatches through wavelet-based processing. Fig. 3 illustrates the complete framework.

3.1. Unified Clean Diffusion Feature Learning

Our analysis demonstrates that existing diffusion features exhibit significant cross-modal misalignment, with sparse sketches and texture-rich images forming separate clusters in the feature space. To address this fundamental limitation, we introduce a unified clean diffusion feature (Unified CleanDIFT) learning stage that maps both modalities into a shared semantic subspace while preserving their essential structural information. We employ Low-Rank Adaptation (LoRA) [8] to efficiently fine-tune the pre-trained CleanDIFT model on sketch data in an unsupervised manner. By integrating LoRA into all linear projection layers of the U-Net backbone, this approach preserves the model’s rich understanding of textured images while effectively adapting to sketch-specific characteristics. For each weight ma-

trix $W \in \mathbb{R}^{d \times d}$, the adaptation is formulated as:

$$W' = W + \Delta W, \quad \text{where} \quad \Delta W = \alpha B A, \quad (1)$$

where $A \in \mathbb{R}^{r \times d}$ and $B \in \mathbb{R}^{d \times r}$ are trainable low-rank matrices with rank $r \ll d$, and α is a learnable scaling factor. Only A and B are updated during training. Our training process employs a teacher-student framework with two parallel forward passes:

1. Teacher Path with Noisy Input: Given a clean sketch x_0 , we first apply forward diffusion to obtain noisy version x_t at randomly sampled timestep $t \in [1, T]$, using noise $\epsilon \sim \mathcal{N}(0, I)$. The original, frozen SD model then processes x_t to extract intermediate features from its U-Net layers. We denote these as the target features, $F_{\text{target}} = \{F_{\text{target}}^{(k)}\}_{k=1}^K$, where k indexes the selected layers.

2. Student Path with Clean Input: The clean sketch x_0 is fed into our LoRA-enhanced model using fixed timestep $t' = 261$, following the optimal setting for clean image feature extraction in [30]. The predicted features are denoted as $F_{\text{pred}} = \{F_{\text{pred}}^{(k)}\}_{k=1}^K$.

During training, to align with timestep-varying diffusion features, we introduce pointwise projection heads explicitly conditioned on the diffusion timestep. Given the predicted features F_{pred} as input, these heads produce a set of projected features $F_{\text{proj}} = \{F_{\text{proj}}^{(k)}\}_{k=1}^K$.

Both pathways use the unified text prompt “a photo of a [class]” to maintain consistency with the pre-trained model’s conditioning space.

Learning Objective. We seek to maximize the cosine similarity between predicted and target features. The loss function is formulated as the negative cosine similarity, computed as the expectation over the data distribution and sampled timesteps:

$$\mathcal{L}_{\text{ada}} = \mathbb{E}_{x_0, \epsilon, t} \left[- \sum_{k=1}^K \frac{F_{\text{proj}}^{(k)}(x_0, t') \cdot F_{\text{target}}^{(k)}(x_t, t)}{\|F_{\text{proj}}^{(k)}(x_0, t')\| \|F_{\text{target}}^{(k)}(x_t, t)\|} \right], \quad (2)$$

where K represents the number of aligned feature layers, as established in [30]. Through sampling t over the complete timestep range, the model learns to produce temporally-invariant, unified clean diffusion features.

3.2. Low-Frequency Feature Aggregation

Despite addressing spatial misalignment, a frequency disparity persists between low-frequency sketches and broad-spectrum images [11, 37]. To address this, we introduce the Low-Frequency Feature Aggregation (LoFFA) module that enhances shared low-frequency information while suppressing modality-specific high-frequency noise.

The LoFFA module processes a set of L multi-scale features $\mathbf{F}^S = \{F_l^S\}_{l=1}^L$ and $\mathbf{F}^T = \{F_l^T\}_{l=1}^L$, extracted from both modalities by Unified CleanDIFT and DINOv2 [22], where each feature map F_l has a dimension of $C_l \times H_l \times W_l$. For each layer l , the processing pipeline begins with channel dimension reduction via convolution, followed by Adaptive Instance Normalization (AdaIN) [9] to align the image feature's distribution to the sketch's distribution, leaving the sketch feature's distribution unchanged.

At its core is the Low-Frequency Enhancement (LoFE) module, which selectively amplifies low-frequency components to bridge the frequency domain gap between source and target features. Taking the source feature $F_l^{S, \text{in}}$ as an example, the LoFE process begins with a hierarchical two-level Discrete Wavelet Transform (DWT).

First, the input feature $F_l^{S, \text{in}}$ undergoes an initial DWT, yielding the first-level low-frequency component $F_{\text{LL}, l}^{S, (1)}$ and high-frequency components $F_{\text{H}, l}^{S, (1)}$. This low-frequency component is then passed through a Convolutional Block with Batch Normalization and GELU (CBG) before a second DWT decomposition, which results in the second-level components $F_{\text{LL}, l}^{S, (2)}$ and $F_{\text{H}, l}^{S, (2)}$, as shown below:

$$F_{\text{LL}, l}^{S, (1)}, F_{\text{H}, l}^{S, (1)} = \text{DWT}(F_l^{S, \text{in}}), \quad (3)$$

$$F_{\text{LL}, l}^{S, (2)}, F_{\text{H}, l}^{S, (2)} = \text{DWT}(\text{CBG}(F_{\text{LL}, l}^{S, (1)})). \quad (4)$$

The resulting lowest-frequency component $F_{\text{LL}, l}^{S, (2)}$, is then enhanced via a sigmoid-gated modulation mechanism, formulated as:

$$\tilde{F}_{\text{LL}, l}^{S, (2)} = F_{\text{LL}, l}^{S, (2)} \odot (1 + M), \quad (5)$$

where the learnable attention mask M is generated by a lightweight, sigmoid-activated convolutional network applied to $F_{\text{LL}, l}^{S, (2)}$. The resulting term $(1 + M)$, where 1 denotes a tensor of ones, forms an adaptive scaling factor. For reconstruction, a weighted sum combines the enhanced subband with its preceding low-frequency counterpart, and the result is then passed through a series of Inverse Discrete Wavelet Transforms (IDWTs) with the high-frequency components ($F_{\text{H}, l}^{S, (2)}$ and $F_{\text{H}, l}^{S, (1)}$), using learnable weights to balance structural preservation and cross-modal enhancement. The entire operation ultimately produces the low-frequency enhanced features, denoted as $\mathcal{H}(F_l^{S, \text{in}})$.

To ensure training stability, the LoFE process is embedded within a scaled residual connection:

$$F_l^{S, \text{out}} = F_l^{S, \text{in}} + \beta(\mathcal{H}(F_l^{S, \text{in}}) - F_l^{S, \text{in}}), \quad (6)$$

where β is a learnable scalar. The resulting enhanced source feature $F_l^{S, \text{out}}$ then passes through a second AdaIN layer and a final convolutional layer to produce the refined source feature $F_l^{S'}$. Similarly, the target feature $F_l^{T, \text{in}}$ follows the same LoFE processing pipeline to produce the refined target feature $F_l^{T'}$, ensuring both modalities receive equivalent low-frequency enhancement.

After processing all layers through LoFFA, we obtain refined features $\{F_l^{S'}\}_{l=1}^L$ and $\{F_l^{T'}\}_{l=1}^L$. The final comprehensive representation for each modality is computed through adaptive weighted summation:

$$\hat{\mathbf{F}}^S = \sum_{l=1}^L \omega_l F_l^{S'}, \hat{\mathbf{F}}^T = \sum_{l=1}^L \omega_l F_l^{T'}, \quad (7)$$

where ω_l is a learnable scalar weight that adaptively balances the contribution of each refined feature layer.

Training Objective. Given the final representations $\hat{\mathbf{F}}^S$ and $\hat{\mathbf{F}}^T$ and their corresponding sets of P keypoint, $\mathbf{K}^S = \{k_i^S\}_{i=1}^P$ and $\mathbf{K}^T = \{k_i^T\}_{i=1}^P$, we formulate a dual objective loss function to train the LoFFA module.

We apply CLIP-style symmetric contrastive loss [23] on features sampled at keypoint locations to pull corresponding pairs closer while pushing non-corresponding pairs apart:

$$\mathcal{L}_{\text{CL}} = \text{CL}(\hat{\mathbf{F}}^S(\mathbf{K}^S), \hat{\mathbf{F}}^T(\mathbf{K}^T)), \quad (8)$$

where $\hat{\mathbf{F}}(\mathbf{K})$ denotes sampling features from the map $\hat{\mathbf{F}}$ at all locations in the set $\mathbf{K}$.

To address the limitations of traditional End-Point Error (EPE) loss [6] on feature maps with potential discontinuities, we adopt the dense matching loss from [36]. For each source keypoint k_i^S , we compute a dense similarity map with the entire target feature map: $C_i = \hat{\mathbf{F}}^S(k_i^S)^\top \hat{\mathbf{F}}^T$. The predicted target location $\hat{k}_i^T$ obtained via differentiable Soft-Argmax: $\hat{k}_i^T = \text{SoftArgmax}(C_i)$, which enables backpropagating gradients to all spatial locations. The

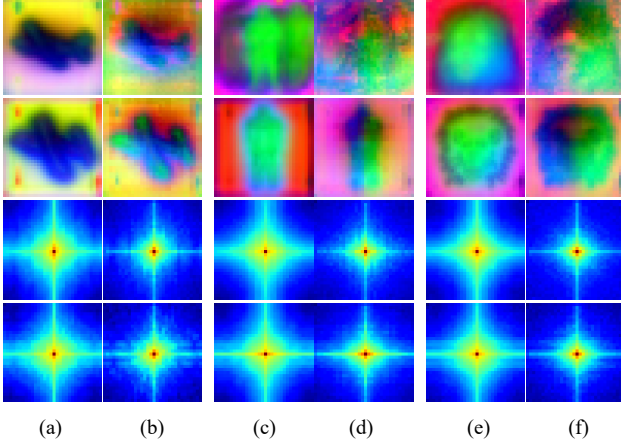


Figure 4. **Progressive feature refinement analysis demonstrating model improvements.** Rows 1-2: PCA-projected feature visualizations showing spatial domain characteristics. Rows 3-4: Frequency domain analysis of the corresponding extracted features. Method comparison: (a, c, e) Unified CleanDIFT, (b, d, f) SFA-DIFT with LoFFA enhancement.

dense loss is formulated as:

$$\mathcal{L}_{\text{Dense}} = \sum_i \|\hat{k}_i^T - (k_i^T + \epsilon)\|_2, \quad (9)$$

where ϵ is a small Gaussian noise added to the ground-truth target keypoint positions for regularization.

The final training objective combines both losses:

$$\mathcal{L} = \mathcal{L}_{\text{CL}} + \mathcal{L}_{\text{Dense}}. \quad (10)$$

3.3. Bridging the Spatial and Frequency Gap

Our two-stage approach achieves effective spatial-frequency alignment. As shown in Fig. 4, Unified CleanDIFT first establishes consistent spatial features across modalities. LoFFA then aggregates shared low-frequency components while suppressing modality-specific high-frequency noise, resulting in concentrated spectral energy in the low-to-mid frequency range and sharpened spatial boundaries. This dual-domain alignment enables robust correspondence between sparse sketches and texture-rich images.

4. Experiments

4.1. Experimental Setup

Datasets. To validate our method’s effectiveness on cross-sparsity correspondence, we introduce MS-PSC6K, an extension of the PSC6K [19] dataset. The original PSC6K provides 1,250 photos from 125 object categories, each with five corresponding sketches and 150,000 manually annotated keypoints. To create MS-PSC6K, we generate five

Table 1. **Quantitative comparison with state-of-the-art methods on PSC6K datasets.** * indicates zero-shot methods. ‡ indicates supervised methods.

Methods	PCK@1	PCK@5	PCK@10
SD*	3.02	38.66	69.77
CLIP*	3.18	40.31	68.99
DINOv2-ViT-B/14*	4.27	44.50	72.58
DINOv3-ViT-B/16*	5.98	56.11	83.05
SD+DINO*	3.55	39.36	66.67
CleanDIFT + DINO*	5.73	55.47	83.81
Fuse2Match*	5.92	48.87	72.25
CleanDIFT+DINO‡	<u>9.44</u>	70.25	<u>90.99</u>
GMRW-SC‡	-	53.61	82.20
Self-Sup‡	5.33	58.02	84.87
SketchFusion‡	-	<u>70.31</u>	89.86
SFA-DIFT‡ (Ours)	9.81 0.37%↑	72.94 2.63%↑	92.70 1.71%↑

stylized versions with distinct textures for each photo, introducing greater style diversity. This results in a new dataset of 7,500 texture-rich images. Additionally, we adopt the Sketchy [27] dataset, which contains 62,500 sketches across 125 categories, to fine-tune CleanDIFT into our Unified CleanDIFT model.

Evaluation Metric. We assess correspondence accuracy using the Percentage of Correct Keypoints (PCK) [33]. PCK is calculated within a threshold of $\alpha(h, w)$, where α represents the tolerance factor and (h, w) denote the height and width dimensions of the object’s bounding box. In our experiments, we evaluate performance using three tolerance values: $\alpha \in \{0.1, 0.05, 0.01\}$, corresponding to PCK@10, PCK@5, and PCK@1 metrics, respectively.

Complete details on method implementation, training protocols, dataset partitioning, and the creation and radial power spectrum analysis of the MS-PSC6K dataset are available in the supplementary material.

4.2. Results

Photo-Sketch Correspondence on PSC6K Dataset. As shown in Tab. 1, zero-shot baselines such as SD [35], Fuse2Match [38] and DINOv2 [22], which operate without target dataset exposure, demonstrate suboptimal performance due to significant representation gaps. We also compare against supervised methods like SketchFusion [15], Self-Sup [19] and GMRW-SC [28]. These methods, defined by parameter updates on the target dataset, also show limited effectiveness. In sharp contrast, our proposed SFA-DIFT model achieves a breakthrough, establishing a new state-of-the-art with 9.81% PCK@1, 72.94% PCK@5, and 92.70% PCK@10.

Multi-Style Photo-Sketch Correspondence on MS-PSC6K Dataset. Evaluating existing SOTA methods’ generalization [15, 19] on our stylistically diverse MS-

Table 2. **Quantitative comparison with state-of-the-art methods on MS-PSC6K datasets.** Performance is evaluated across five artistic styles (Abstract, Baroque, Realism, Post-Imp, Neo-Imp). * indicates zero-shot methods. † indicates supervised methods.

Method	Abstract			Realism			Baroque			Neo-Impressionism			Post-impressionism		
	@1	@5	@10	@1	@5	@10	@1	@5	@10	@1	@5	@10	@1	@5	@10
SD*	3.17	40.53	70.70	3.20	40.12	70.85	3.12	39.36	69.09	3.30	40.49	71.40	3.33	41.27	71.74
CLIP *	2.65	34.72	64.09	2.52	34.49	64.48	2.55	33.87	62.80	2.25	32.88	62.85	2.58	35.76	65.62
DINOv2-ViT-B/14*	3.32	39.26	68.16	3.46	39.44	68.25	3.23	37.98	66.73	3.01	37.48	67.57	3.63	40.87	69.58
DINOv3-ViT-B/16*	5.15	55.41	83.20	5.34	54.49	80.03	4.96	52.56	80.97	4.88	53.63	81.94	5.59	57.09	84.37
SD+DINO *	3.88	44.20	74.66	3.78	43.98	74.95	3.64	43.15	73.57	3.43	43.28	74.94	3.90	44.96	75.76
Fuse2Match*	5.05	45.53	70.43	4.93	45.17	70.17	4.72	43.66	68.16	4.71	45.83	71.43	5.32	47.50	71.97
CleanDIFT+DINO *	5.28	54.74	83.70	5.24	54.60	83.60	5.28	54.92	83.82	5.34	55.20	84.11	5.53	56.01	84.54
CleanDIFT+DINO †	<u>7.15</u>	<u>66.02</u>	<u>89.79</u>	<u>7.76</u>	<u>66.82</u>	<u>90.07</u>	<u>7.72</u>	<u>66.70</u>	<u>89.83</u>	<u>7.88</u>	<u>67.86</u>	<u>90.63</u>	<u>8.02</u>	<u>68.09</u>	<u>90.81</u>
Self-Sup*	4.19	51.96	81.34	3.41	46.14	76.26	3.54	47.54	78.10	3.31	45.22	76.54	4.45	54.00	82.83
SketchFusion *	4.43	48.85	77.24	3.66	44.11	73.88	3.91	45.05	74.09	3.74	44.16	73.81	4.93	50.45	78.41
Self-Sup†	3.27	45.23	77.33	2.76	40.71	73.10	3.01	42.54	74.85	2.46	38.17	72.63	3.48	46.66	78.42
SketchFusion†	5.05	51.83	78.98	4.50	48.43	77.39	4.66	49.11	76.92	4.29	48.18	77.03	4.95	52.15	79.31
SFA-DIFT †	8.98	70.43	91.49	8.84	68.45	90.68	8.47	68.25	90.30	8.59	68.89	90.88	8.60	70.05	91.74
	1.83%↑	4.41%↑	1.70%↑	1.08%↑	1.63%↑	0.61%↑	0.75%↑	1.55%↑	0.47%↑	0.71%↑	1.03%↑	0.25%↑	0.58%↑	1.96%↑	0.93%↑

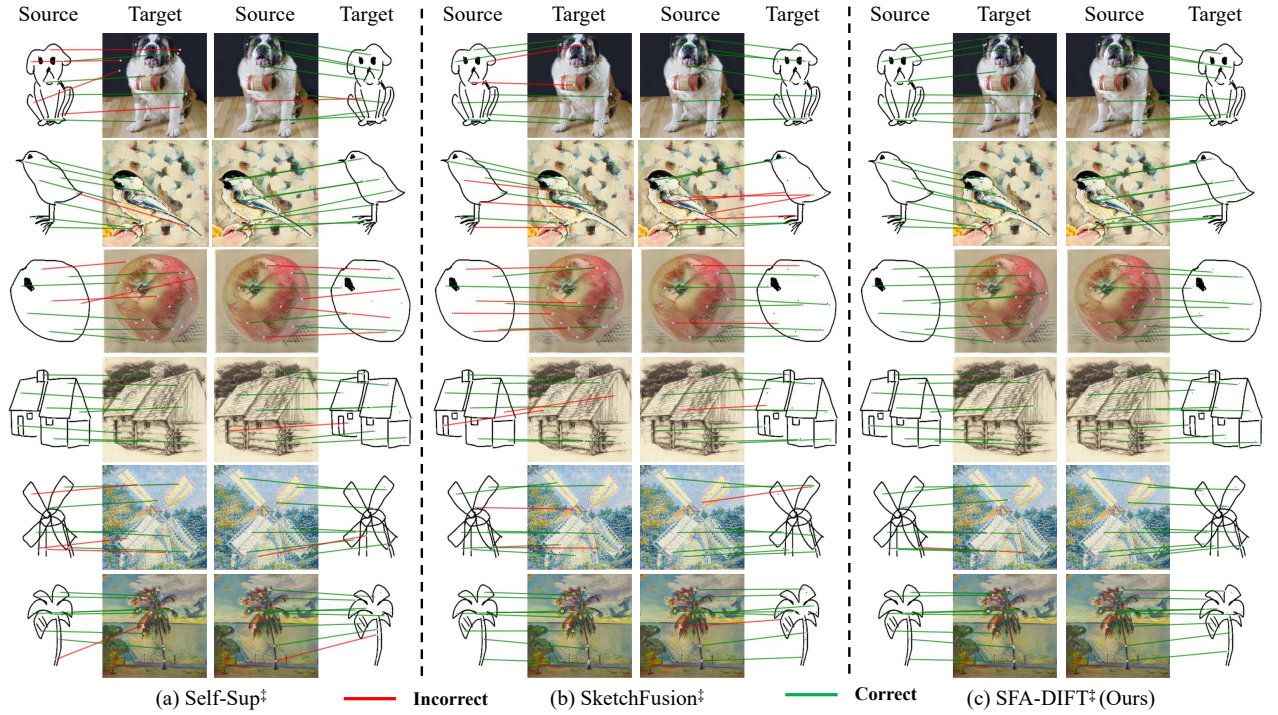


Figure 5. **Qualitative cross-sparsity correspondence comparison.** Each row demonstrates cross-sparsity correspondence results between sketches and various textured images. Methods (a), (b), and (c) represent different approaches. For each method, we present correspondence results using both sketch-to-texture and texture-to-sketch configurations to evaluate bidirectional correspondence performance. **SFA-DIFT** achieves superior accuracy and robustness.

PSC6K dataset, we found that direct transfer of pre-trained weights caused a marked performance degradation, indicating domain sensitivity. Critically, re-training on MS-PSC6K further degraded their performance, suggesting a failure to converge on such heterogeneous data. In contrast, our method demonstrated superior robustness. Averaging across the five style variations, it achieved new state-of-the-art results of 8.70% PCK@1, 69.21% PCK@5, and 91.02%

PCK@10, despite a slight performance dip, as detailed in Tab. 2. Qualitative results are presented in Fig. 5. More visualizations are in the supplementary material.

4.3. Robustness to Rich Textural Variations

A key limitation in evaluating model robustness is that absolute performance drops under style and texture perturbations remain unnormalized, making fair baseline compar-

Table 3. **Robustness Ratio (RR) of various approaches under texture variations.** * indicates zero-shot methods. ‡ indicates supervised methods.

Methods	RR@1	RR@5	RR@10
SD*	1.07	1.04	1.01
CLIP*	0.79	0.85	0.93
DINOv2-ViT-B/14*	0.78	0.81	0.94
DINOv3-ViT-B/16*	0.87	<u>0.97</u>	0.98
SD+DINO*	1.05	1.11	1.12
Fuse2Match*	0.80	0.93	0.98
CleanDIFT*	0.94	0.99	0.99
CleanDIFT+DINOv2*	0.93	0.99	1.00
Self-Sup*	0.71	0.84	0.93
SketchFusion*	-	0.66	0.88
Self-Sup‡	0.56	0.74	0.87
SketchFusion‡	-	0.71	0.87
CleanDIFT+DINOv2‡	0.82	0.96	<u>0.99</u>
Unified CleanDIFT* (Ours)	0.99	0.99	1.00
SFA-DIFT‡ (Ours-full)	0.87	0.95	0.98

Table 4. **Ablation study.** Quantitative analysis of method variants. * indicates zero-shot methods. ‡ indicates supervised methods.

Methods	PCK@1	PCK@5	PCK@10
CleanDIFT*	5.59	57.27	84.57
Unified CleanDIFT*	5.94	58.69	85.57
Unified CleanDIFT + DINOv2*	5.62	56.76	85.07
Unified CleanDIFT + Conv‡	<u>8.68</u>	68.78	90.17
LoFFA w/o AdaIN‡	8.57	68.64	<u>90.52</u>
LoFE → Conv‡	8.53	<u>69.01</u>	90.51
DWT & IDWT → Conv‡	7.96	66.98	89.77
Single DWT & IDWT‡	8.51	68.45	<u>90.52</u>
SFA-DIFT‡ (Ours-full)	8.89	69.83	91.32

isons difficult [5]. To address this limitation, we propose the **Robustness Ratio (RR)**, a standardized ratio-based metric that enables consistent comparisons across diverse style and texture variations:

$$\text{RR} := \frac{\frac{1}{N} \sum_{i=1}^N \text{PCK}(x_i^s, \hat{x}_i^{t,\text{pert}})}{\frac{1}{N} \sum_{i=1}^N \text{PCK}(x_i^s, \hat{x}_i^{t,\text{orig}})}. \quad (11)$$

Here, N denotes the number of samples, $\hat{x}_i^{t,\text{pert}}$ denotes the prediction on the texture-perturbed image, while $\hat{x}_i^{t,\text{orig}}$ denotes the prediction on the original image.

As reported in Tab. 3, baselines like CleanDIFT + DINOv2 [30] and Self-Sup [19] show RR degradation after fine-tuning on MS-PSC6K. Lacking explicit frequency alignment, they overfit to high-frequency styles in pixel and feature domains, impairing generalization across stylistic perturbations. Note that $\text{RR} > 1$ indicates a bias toward diverse styles rather than photos. In contrast, our Unified CleanDIFT maintains a stable RR near 1.0, demonstrating genuine invariance to textural shifts. The slight RR decrease

of our full SFA-DIFT is expected given task complexity, where high-variance distractors (textures, color shifts) compete with essential structural cues. This lowers performance improvement rates compared to simpler photo-sketch correspondence, reducing the RR metric. Importantly, SFA-DIFT achieves superior robustness through explicit spatial-frequency regularization. By suppressing high-frequency style noise while preserving mid-to-low frequency consistency, our method maintains substantially higher post-training RR than baselines, optimally balancing domain specialization and generalization robustness.

4.4. Ablation Study

We evaluate component contributions using averaged metrics across MS-PSC6K and PSC6K (Tab. 4). Zero-shot pre-trained models like CleanDIFT show limited effectiveness, highlighting the inadequacy of general-purpose features for fine-grained cross-modal matching.

To address this, we introduce the LoFFA module and validate its design through five ablation settings. Compared to a fairly supervised Unified CleanDIFT + Conv‡ baseline, LoFFA yields significant gains. Internally, variants including LoFFA without AdaIN‡, LoFE to Conv‡, DWT and IDWT to Conv‡, and Single DWT and IDWT‡ confirm the necessity of distribution alignment, low-frequency enhancement, explicit frequency transforms, and deep frequency decomposition, respectively. SFA-DIFT outperforms all variants, proving frequency-aware alignment effectively bridges the modality gap.

5. Conclusion

This paper addresses the challenging correspondence problem between sparse line drawings and texture-rich photographs by identifying fundamental cross-modal discrepancies. We reveal that correspondence failures stem from spatial abstraction gaps and frequency distribution inconsistencies. Our SFA-DIFT framework tackles these challenges through parameter-efficient LoRA-based fine-tuning for spatial alignment and lightweight low-frequency aggregation for frequency reconciliation. Extensive evaluation on our MS-PSC6K benchmark demonstrates state-of-the-art performance with significant improvements. The results validate that effective cross-sparsity correspondence requires simultaneous spatial and frequency domain alignment, providing a principled framework for understanding and addressing line-texture mismatches.

Limitation and Future Work. The use of Diffusion features in our approach presents inherent limitations in inference efficiency, with an average processing time of 0.8 seconds per correspondence estimation. We plan to explore more computationally efficient and semantically effective diffusion feature representations in future research.

6. Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grants 62576261, U22A2096, and 62506280, in part by the Scientific and Technological Innovation Teams in Shaanxi Province under grant 2025RS-CXTD-011, in part by the Shaanxi Province Core Technology Research and Development Project under grant 2024QY2-GJHX-11, and in part by the CCF-Kuaishou Large Model Explorer Fund (NO. CCF-KuaiShou 2025006).

References

- [1] Rudolf Arnheim. *Art and visual perception: A psychology of the creative eye*. Univ of California Press, 1954. 2
- [2] Wengling Chen and James Hays. Sketchygan: Towards diverse and realistic sketch to image synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9416–9425, 2018. 3
- [3] De Cheng, Zhipeng Xu, Xinyang Jiang, Dongsheng Li, Nannan Wang, and Xinbo Gao. Prompt disentanglement via language guidance and representation alignment for domain generalization. *IEEE Transactions on Pattern Analysis and Machine Intelligence (IEEE TPAMI)*, pages 1–18, 2026. 2
- [4] Seokju Cho, Sunghwan Hong, Sangryul Jeon, Yunsung Lee, Kwanghoon Sohn, and Seungryong Kim. Cats: Cost aggregation transformers for visual correspondence. *Advances in Neural Information Processing Systems (NeurIPS)*, 34: 9011–9023, 2021. 3
- [5] Wenye Chong. Towards robust semantic correspondence: A benchmark and insights. *arXiv preprint arXiv:2508.00272*, 2025. 8
- [6] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. FlowNet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2758–2766, 2015. 5
- [7] 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. 3
- [8] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *International Conference on Learning Representations (ICLR)*, 1(2):3, 2022. 2, 4
- [9] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1501–1510, 2017. 5
- [10] Seyed Mehdi Iranmanesh, Hadi Kazemi, Sobhan Soleymani, Ali Dabouei, and Nasser M Nasrabadi. Deep sketch-photo face recognition assisted by facial attributes. In *2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–10. IEEE, 2018. 3
- [11] Min Woo Kim and Nam Ik Cho. Whfl: Wavelet-domain high frequency loss for sketch-to-image translation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 744–754, 2023. 2, 5
- [12] Seungwook Kim, Juhong Min, and Minsu Cho. Transformer: Match-to-match attention for semantic correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8697–8707, 2022. 3
- [13] Subhadeep Koley, Ayan Kumar Bhunia, Aneeshan Sain, Pinaki Nath Chowdhury, Tao Xiang, and Yi-Zhe Song. Picture that sketch: Photorealistic image generation from abstract sketches. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6850–6861, 2023. 1
- [14] Subhadeep Koley, Ayan Kumar Bhunia, Aneeshan Sain, Pinaki Nath Chowdhury, Tao Xiang, and Yi-Zhe Song. Text-to-image diffusion models are great sketch-photo matchmakers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16826–16837, 2024. 2
- [15] Subhadeep Koley, Tapas Kumar Dutta, Aneeshan Sain, Pinaki Nath Chowdhury, Ayan Kumar Bhunia, and Yi-Zhe Song. Sketchfusion: Learning universal sketch features through fusing foundation models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2556–2567, 2025. 1, 2, 3, 6
- [16] Ce Liu, Jenny Yuen, and Antonio Torralba. Sift flow: Dense correspondence across scenes and its applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence (IEEE TPAMI)*, 33(5):978–994, 2010. 3
- [17] Qing Liu, Lingxi Xie, Huiyu Wang, and Alan L Yuille. Semantic-aware knowledge preservation for zero-shot sketch-based image retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3662–3671, 2019. 2
- [18] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision (IJCV)*, page 91–110, 2004. 3
- [19] Xuanchen Lu, Xiaolong Wang, and Judith E Fan. Learning dense correspondences between photos and sketches. In *International Conference on Machine Learning (ICML)*, pages 22899–22916. PMLR, 2023. 1, 2, 6, 8
- [20] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research (JMLR)*, 9(Nov):2579–2605, 2008. 1, 2
- [21] Juhong Min and Minsu Cho. Convolutional hough matching networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2940–2950, 2021. 3
- [22] 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. 5, 6
- [23] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry,

- Amanda Aspell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pages 8748–8763. PmLR, 2021. [2](#), [3](#), [5](#)
- [24] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. [2](#), [3](#)
- [25] Daniel Ruderman and William Bialek. Statistics of natural images: Scaling in the woods. *Advances in Neural Information Processing Systems (NeurIPS)*, 6, 1993. [3](#)
- [26] Aneeshan Sain, Ayan Kumar Bhunia, Pinaki Nath Chowdhury, Subhadeep Koley, Tao Xiang, and Yi-Zhe Song. Clip for all things zero-shot sketch-based image retrieval, fine-grained or not. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2765–2775, 2023. [1](#), [2](#)
- [27] 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. [6](#)
- [28] Ayush Shrivastava and Andrew Owens. Self-supervised spatial correspondence across modalities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6383–6393, 2025. [6](#)
- [29] Jifei Song, Qian Yu, Yi-Zhe Song, Tao Xiang, and Timothy M Hospedales. Deep spatial-semantic attention for fine-grained sketch-based image retrieval. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 5551–5560, 2017. [3](#)
- [30] Nick Stracke, Stefan Andreas Baumann, Kolja Bauer, Frank Fundel, and Björn Ommer. Cleandift: Diffusion features without noise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 117–127, 2025. [2](#), [3](#), [4](#), [5](#), [8](#)
- [31] Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. *Advances in Neural Information Processing Systems (NeurIPS)*, 36:1363–1389, 2023. [2](#)
- [32] 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. *ACM Transactions on Graphics (TOG)*, 41(4):1–11, 2022. [1](#)
- [33] Yi Yang and Deva Ramanan. Articulated human detection with flexible mixtures of parts. *IEEE Transactions on Pattern Analysis and Machine Intelligence (IEEE TPAMI)*, page 2878–2890, 2013. [6](#)
- [34] 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 (IJCV)*, 122(3):411–425, 2017. [3](#)
- [35] 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. *Advances in Neural Information Processing Systems (NeurIPS)*, 36:45533–45547, 2023. [6](#)
- [36] 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. [2](#), [5](#)
- [37] Tianpei Zhang, Jufeng Zhao, Yiming Zhu, and Guangmang Cui. Wife-fusion: Wavelet-aware intra-inter frequency enhancement for multi-model image fusion. *arXiv preprint arXiv:2506.03555*, 2025. [5](#)
- [38] Jing Zuo, Jiaqi Wang, Yonggang Qi, and Yi-Zhe Song. Fuse2match: Training-free fusion of flow, diffusion, and contrastive models for zero-shot semantic matching. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 1–10, 2025. [6](#)