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

Supplementary Material

1. Related Work

Correspondence Across Visual Sparsity. Establishing correspondence across the visual sparsity spectrum between sparse line representations and richly textured imagery poses challenges beyond conventional photo–sketch matching. Early approaches with hand-crafted descriptors like SIFT [9] and HOG [10] lack semantic robustness for extreme representation gaps. While deep models [1, 5] and advanced correlation operators [12] improve spatial matching on appearance-consistent pairs, they remain vulnerable when modalities differ fundamentally in structure and frequency. This vulnerability arises because visual encoders pre-trained on texture-rich distributions yield suboptimal representations for sparse inputs and generalize poorly across modalities [7]. Even with recent efforts [6] to inject high-level semantics, which partly mitigate frequency differences, fundamental inconsistencies in the spatial and frequency domains are not resolved, causing persistent cross-modal drift. This motivates spatial–frequency joint alignment: learning unified, clean semantic features that are invariant to modality-specific noise while selectively aggregating low-frequency components to overcome the texture bias.

Diffusion Features for Visual Understanding. Diffusion models [3] excel at image synthesis, with their U-Net architectures learning rich semantic representations. Though originally for denoising, these intermediate features encode valuable content and structure, making them promising for general visual understanding.

Recent work has applied diffusion features to semantic correspondence [2, 8, 17, 20, 21], segmentation [18], and virtual Try-on [4]. Features from models like Stable Diffusion [14] show exceptional correspondence performance. However, when matching sparse drawings against richly textured photos, pre-trained diffusion features are biased toward textures and generalize poorly to sketch-like inputs. This domain and frequency mismatch limits robust cross-modal alignment. Inspired by feature distillation approaches like CleanDIFT [16], we propose fine-tuning diffusion models with sketch data to enhance cross-modal semantic feature extraction.

2. Experimental Details

This section outlines the datasets utilized for our experiments, along with our detailed implementation setup. We describe the standard benchmarks, the generation process of our multi-style dataset for robustness testing, and other key experimental settings such as training configurations, feature dimensions, and evaluation metrics.

Sketchy [15]. The full Sketchy dataset contains 12,500 photographs spanning 125 object categories, with each photograph associated with at least five paired sketches. We leverage sketch data from 104 categories within this dataset to finetune CleanDIFT, resulting in our proposed Unified CleanDIFT model.

PSC6K [11]. This photo-sketch correspondence benchmark is constructed from the Sketchy dataset's test split through data augmentation techniques. The dataset encompasses 1,250 photographs, 150K manually annotated keypoints, and five paired sketches per photograph distributed across 125 object categories. Following the experimental protocol established in [6], we employ a 60:40 training/testing data split to train and evaluate our feature aggregation module, respectively.

MS-PSC6K. To rigorously investigate if a model achieves effective spatial-frequency decoupling and structural-semantic alignment between photos and sketches, we introduce a stylization-based data generation and evaluation protocol. This protocol is designed to test the model's robustness under conditions of "structure-preserved, appearance-altered" transformations. Examples shown in Fig. 1.

Our method builds upon existing photo-sketch datasets by applying the CAST [22] method, a series of artistic stylization transforms, to the photo modality to generate new test samples. This approach is predicated on a core assumption: the stylization process primarily perturbs the mid-to-high frequency spectral characteristics of an image, such as texture, brushstrokes, and color distribution, while leaving the low-frequency geometric contours and spatial structures highly intact. To quantitatively validate this assumption, radial power spectrum analysis (Fig. 4(a)) confirms that MS-PSC6K effectively reshapes frequency distributions. Specifically, it introduces bidirectional perturbations: for instance, the Abstract style injects high frequencies, while the Realism style attenuates them. This property allows the structural annotations of the original photograph (e.g., keypoint coordinates) to be transferred to its stylized variants at zero cost, enabling the construction of a large-scale, diverse test dataset without any manual re-annotation.

To ensure the protocol's validity and geometric fidelity, we implemented a rigorous filtering process. Using features from a pre-trained DINOv3 known for robustness to style variation, we quantitatively assessed key-

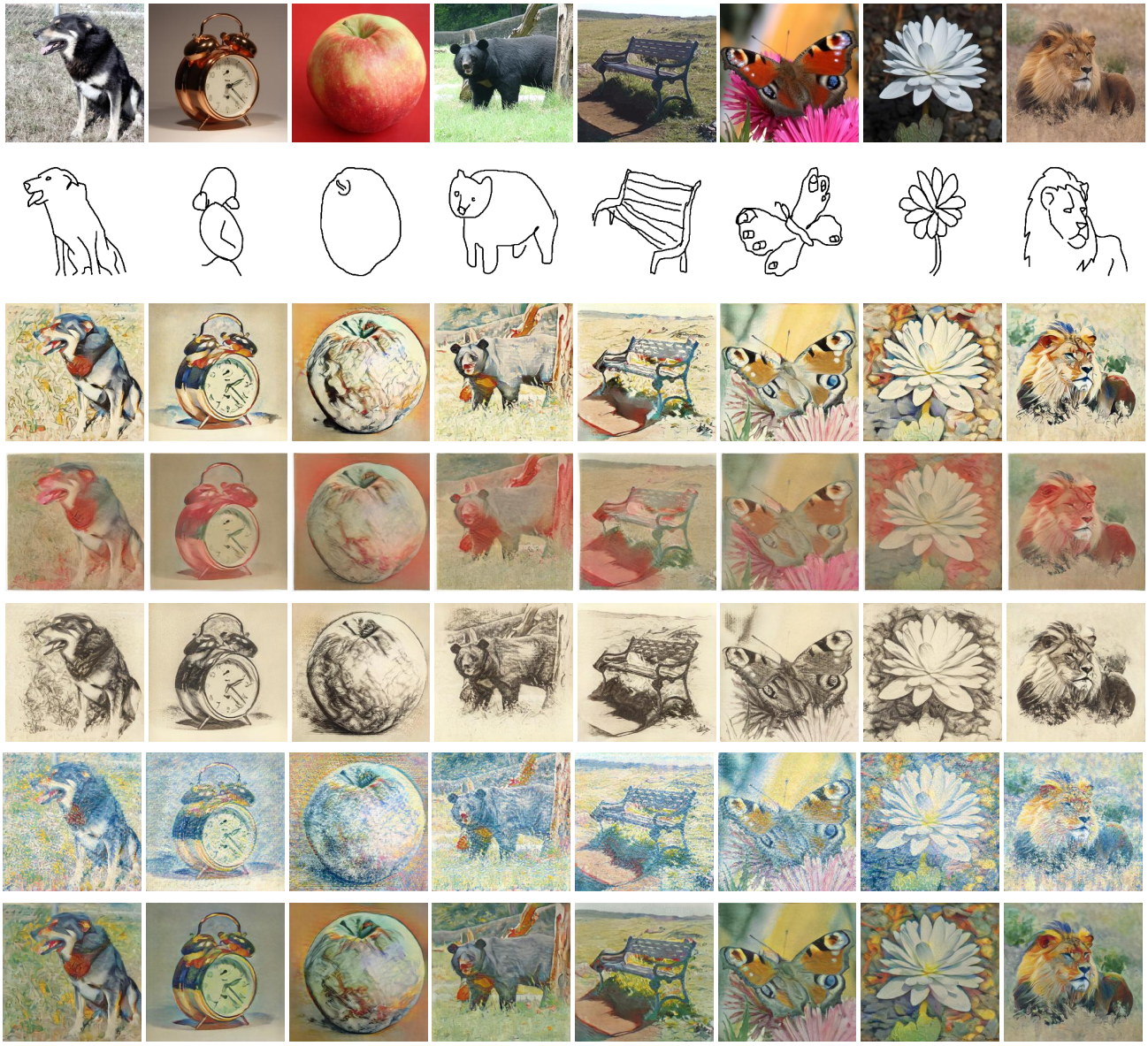


Figure 1. **Our MS-PSC6K dataset examples.** From top to bottom, each row shows the original photo, its paired sketch, and five stylized photo variants (Abstract, Realism, Baroque, Post-Impressionism, Neo-Impressionism). These examples illustrate our "structure-preserved, appearance-altered" protocol.

point spatial consistency before and after stylization, and excluded transformations causing significant geometric distortions. We then selected five representative artistic styles reflecting controlled yet rich perturbations: Abstract (low-frequency structural rearrangement), Baroque (moderately complex textures), Realism (high-fidelity detail retention), Post-Impressionism (boundary relaxation with directional strokes), and Neo-Impressionism (strong anisotropic high-frequency noise). Semantic correspondence between these stylized photos and original sketches measures the model's structural invariance under substantial appearance shifts.

We finetune CleanDIFT using Adam optimizer, with a learning rate of 1×10^{-5}, weight decay of 5×10^{-4}, and a batch size of 2, to obtain Unified CleanDIFT For LoRA with rank r of 16, scaling factor α of 64, weight matrix $\mathbf{B}$ is zero-initialized while $\mathbf{A}$ uses Gaussian initialization. The LoFFA module is trained for 40 epochs, using a learning rate of 1.25×10^{-3}, a batch size of 1, weight decay of 1×10^{-3}, and Gaussian augmentation with a standard deviation of 0.1. Images are resized to 512×512 for Unified CleanDIFT and 960×840 for DINOv2-ViT-B/14. All experiments use a single NVIDIA RTX 3090 GPU.

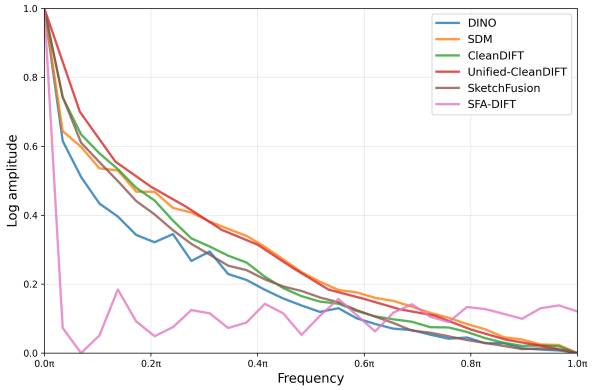


Figure 2. **Log-amplitude spectrum of PCA-reduced feature maps, averaged over 2500 samples.** This spectral analysis shows that while other models struggle with either low-frequency (global structure) or high-frequency (detail) information, our Unified CleanDIFT demonstrates a robust, balanced response across the entire spectrum.

Feature Dimension. Our method leverages features from four distinct layers: three layers extracted from the 2nd, 5th, and 8th upsampling layers of our fine-tuned Unified CleanDIFT U-Net with dimensions $640 \times 32 \times 32$, $1280 \times 16 \times 16$, and $1280 \times 8 \times 8$, respectively, along with features from the final layer of DINOv2-ViT-B/14 with dimension $768 \times 60 \times 60$. The input image size for Unified CleanDIFT is 512×512, while DINOv2 processes images at 840×840 resolution. To ensure spatial consistency, all features are resampled to a unified spatial dimension of 32×32 before being fed into our LoFFA module for feature aggregation. After processing through the LoFFA module, the aggregated features are transformed to the final dimension of $768 \times 32 \times 32$.

3. Frequency Analysis

Evaluation Metric. Given a set of N keypoint pairs, where each pair consists of a predicted keypoint location $\mathbf{p}_i$ at (x_i, y_i) and its corresponding ground truth location $\mathbf{g}_i$ at (x_i^{gt}, y_i^{gt}), the percentage of Correct Keypoints (PCK) [19] is computed as:

$$\text{PCK}(\alpha) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}[d(\mathbf{p}_i, \mathbf{g}_i) \leq \alpha \cdot \max(h, w)], \quad (1)$$

where $d(\mathbf{p}_i, \mathbf{g}_i) = \sqrt{(x_i - x_i^{gt})^2 + (y_i - y_i^{gt})^2}$ is the Euclidean distance between predicted and ground truth keypoints, $\mathbb{I}[\cdot]$ is the indicator function, α is the tolerance factor, and h, w are the height and width of the object's bounding box. The tolerance threshold is T is defined as $\alpha \cdot \max(h, w)$. For our experiments, we use three tolerance factors: PCK@10 with T_{10} set to $0.1 \cdot \max(h, w)$, PCK@5 with T_5 set to $0.05 \cdot \max(h, w)$, and PCK@1 with T_1 set to $0.01 \cdot \max(h, w)$.

Based on the spectral analysis shown in Fig. 2, we analyze the log-amplitude spectrum of PCA-reduced features from 2,500 sketch samples to evaluate frequency response characteristics across six methods: DINO [13], SDM [14], CleanDIFT [16], SketchFusion [6], our Unified CleanDIFT, and our complete SFA-DIFT pipeline.

DINOv2 (blue curve) demonstrates consistently poor performance with log-amplitude values below 0.4 throughout the spectrum, indicating fundamental limitations in capturing sketch-specific features due to its natural image training focus without sketch-domain adaptation.

SDM (orange curve) exhibits decay from approximately 0.6 at low frequencies to 0.1 at high frequencies. Its steep attenuation in mid-to-high frequency bands (0.2π-1.0π) reflects fundamental sketch processing limitations. SDM struggles with the sparsity-induced void problem where sparse line strokes create substantial empty regions, and cannot effectively infer semantic information in these void areas, resulting in incomplete feature representations consistent with its dense photorealistic image synthesis training.

CleanDIFT (green curve) shows strong low-frequency response ($\sim$0.7 at 0π) but rapid degradation beyond 0.3π, reaching below 0.2 in high-frequency regions. This indicates effective global structure preservation but limited detail retention, yielding overly smooth representations.

SketchFusion (red curve) maintains 0.6-0.7 amplitude in the 0-0.2π range but nevertheless exhibits similar high-frequency degradation, dropping to 0.15 beyond 0.6π, suggesting inherently suboptimal detail preservation despite sketch-domain optimization.

Unified CleanDIFT (brown curve) achieves balanced spectral response with consistently high amplitude values ($>$0.5) across low frequencies and superior high-frequency preservation, effectively retaining both global structures and fine-grained details.

SFA-DIFT (pink curve) presents unique spectral characteristics with initial sharp attenuation followed by stable mid-to-high frequency response (amplitude $\sim$0.15 beyond 0.2π). This results from our task-specific optimization prioritizing cross-modal semantic alignment over frequency domain fidelity. The optimization suppresses domain-specific frequency components that hinder correspondence matching while retaining semantically relevant information for robust image-sketch alignment, representing a deliberate trade-off enhancing cross-modal correspondence accuracy.

4. Layer-wise Comparison: DINOv2 and CLIP

As shown in Fig. 3, to investigate how pre-trained vision models encode cross-modal information, we conduct a layer-wise analysis of features from DINOv2 and CLIP on a zero-shot photo–sketch correspondence task, evaluating

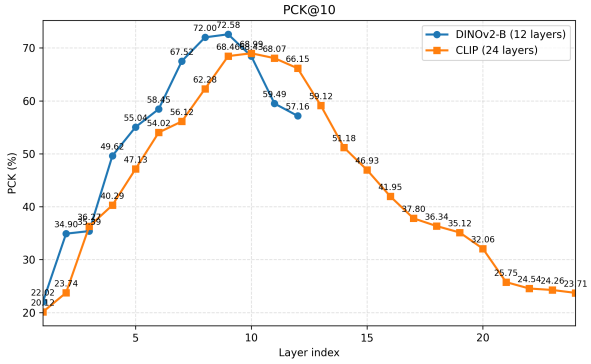


Figure 3. **PCK@10 across layers for zero-shot photo–sketch correspondence on the PSC6K dataset.** DINOv2 has 12 layers, while CLIP has 24 layers.

matching performance using PCK@10. Our results reveal a clear trend: under the zero-shot setting, intermediate-layer features significantly outperform last-layer features. We attribute this to the intermediate layers' ability to capture abstract structural and semantic representations shared between photos and sketches, whereas last-layer features tend to focus on image-specific texture and appearance details, making them less effective for bridging the domain gap between the two modalities. To ensure comparability and reproducibility, Considering these factors, we select DINOv2 as a complementary model.

5. Ablation Study on LoRA Rank Parameter

The selection of LoRA rank r represents a critical trade-off between model expressiveness and computational efficiency. The rank parameter determines the dimensionality of the low-rank decomposition matrices $\mathbf{A} \in \mathbb{R}^{r\times d}$ and $\mathbf{B} \in \mathbb{R}^{d\times r}$, directly influencing the model's capacity to capture task-specific adaptations. We employ LoRA to fine-tune CleanDIFT for enhancing sketch feature extraction capabilities and systematically evaluate different rank configurations with $r \in \{4, 8, 16, 32\}$ using Percentage of Correct Keypoints (PCK) as our evaluation metric.

As shown in Table 2, we evaluate the performance using PCK metrics at different tolerance levels (PCK@10, PCK@5, PCK@1) while monitoring the parameter overhead. The experimental results reveal that rank r of 16 achieves the best overall performance with PCK@10 of 85.71%, PCK@5 of 59.09%, and PCK@1 of 6.00%, representing the optimal balance between correspondence accuracy and computational efficiency. While r of 32 shows comparable performance in some metrics, it significantly increases the parameter count to 13.6M, making it computationally expensive without substantial performance gains. Conversely, lower ranks (specifically r of 4 and 8) result in insufficient model capacity, leading to suboptimal performance across all evaluation metrics due to limited expressiveness of the low-rank adaptation. Therefore, we adopt r of 16 as our default configuration, which provides 6.82M parameters while maintaining superior image-sketch correspondence quality and computational tractability.

6. Generalization Ability

Tab. 1 proves our robust generalization. On our newly collected Real-MS and Unpaired-PSC6K benchmarks, supervised methods fail due to domain bias. In contrast, our Unified CleanDIFT* maintains SOTA performance, showing significant advantages in zero-shot and unpaired settings.

Table 1. Quantitative Results on Unpaired-PSC6K and Real-MS.

Method	Unpaired-PSC6K			Real-MS		
	@1	@5	@10	@1	@5	@10
DINOv2-ViT-B/14*	6.50	47.00	72.38	1.50	42.75	65.88
CleanDIFT*	1.50	28.13	55.63	7.38	51.25	69.00
Fuse2Match*	**7.88**	57.25	81.38	**16.75**	**62.50**	74.88
Unified CleanDIFT*	6.88	**58.50**	**84.00**	9.50	57.00	**77.25**
Self-Sup‡	0.13	5.75	25.37	0.13	5.12	16.00
SketchFusion‡	4.63	47.25	75.63	2.22	22.63	38.75
SFA-DIFT‡ (Ours-full)	**6.75**	**62.00**	**86.63**	**8.75**	**53.88**	**71.88**

7. Computational Efficiency

We evaluate the computational efficiency of our SFA-DIFT approach on a single NVIDIA RTX 3090 GPU using the PyTorch framework. Our method demonstrates reasonable computational requirements across different training stages and inference phases.

Training Efficiency. LoRA fine-tuning for Unified CleanDIFT uses 20GB GPU memory and 9 hours for 6 epochs, with 6.82M adaptation parameters at rank 16, due to the large diffusion backbone and gradient updates. In contrast, LoFFA training is far more efficient: requiring only around 2GB of GPU memory with 6.96M parameters, and completes 40 epochs in 11 hours, demonstrating markedly lower resource demands.

Inference Performance. During inference, our method occupies approximately 14GB of GPU memory, which is manageable for modern GPU configurations. For establishing correspondences between a image-sketch pair, our approach requires 0.8 seconds per pair. The inference efficiency benefits from our strategic design choices: the LoRA adaptation introduces minimal computational overhead while the LoFFA module's focus on low-frequency components reduces processing complexity compared to full-spectrum feature manipulation.

This computational profile demonstrates that our SFA-DIFT achieves a favorable balance between performance

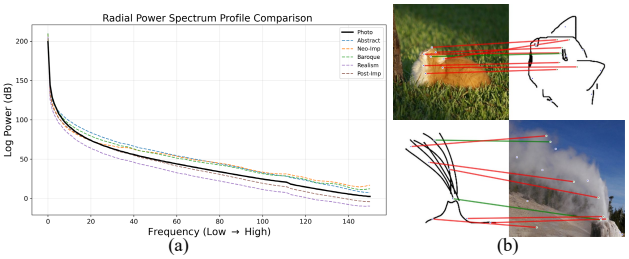
Rank r	PCK@10	PCK@5	PCK@1	Params (M)
4	85.26	<u>58.66</u>	<u>6.01</u>	1.71
8	85.28	58.62	5.96	3.41
16	**85.71**	**59.09**	6.00	6.82
32	<u>85.31</u>	58.61	**6.04**	13.65

Table 2. **Performance comparison across different LoRA ranks for CleanDIFT fine-tuning.** We evaluate PCK metrics at different tolerance levels while monitoring parameter overhead. Rank r of 16 achieves optimal balance between accuracy and efficiency. **Bold** indicates best performance, <u>underline</u> indicates second-best.

gains and resource requirements, making it suitable for both research applications and practical deployment scenarios.

8. Additional Qualitative Results

To further underscore the effectiveness and robustness of our SFA-DIFT approach, we present additional compelling qualitative results across diverse object categories and challenging scenarios in Fig. 5 and Fig. 6. These supplementary visualizations not only reinforce our quantitative findings but also provide deeper insights into the model's strong performance. Specifically, they highlight its resilience to significant consistently perturbations arising from variations in artistic style and texture, confirming its powerful generalization capabilities.



(a) (b)

Figure 4. (a) Power Spectrum Comparison. (b) Failure Cases.

9. Failure Modes.

As shown in Fig. 4(b), correspondence performance degrades on certain cases with extremely abstract sketches. Will discuss in final version.

References

[1] 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. 1

[2] Eric Hedlin, Gopal Sharma, Shweta Mahajan, Hossam Isack, Abhishek Kar, Andrea Tagliasacchi, and Kwang Moo Yi. Unsupervised semantic correspondence using stable diffusion. *Advances in Neural Information Processing Systems (NeurIPS)*, 36:8266–8279, 2023.

[3] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:6840–6851, 2020.

[4] Jeongho Kim, Guojung Gu, Minho Park, Sunghyun Park, and Jaegul Choo. Stableviton: Learning semantic correspondence with latent diffusion model for virtual try-on. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8176–8185, 2024.

[5] Seungwook Kim, Juhong Min, and Minsu Cho. Transformatcher: 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. 1

[6] 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, 3

[7] Ke Li, Kaiyue Pang, and Yi-Zhe Song. Photo pre-training, but for sketch. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2754–2764, 2023.

[8] Xinghui Li, Jingyi Lu, Kai Han, and Victor Adrian Prisacariu. Sd4match: Learning to prompt stable diffusion model for semantic matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 27558–27568, 2024.

[9] 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. 1

[10] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision (IJCV)*, page 91–110, 2004. 1

[11] 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

[12] 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. 1

[13] 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. 3

[14] 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. 1, 3

[15] 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. 1

Figure 5. **Qualitative cross-sparsity correspondence comparison.** Each row demonstrates cross-sparsity correspondence results between sketches and various textured images. For each method, we present correspondence results using both sketch-to-texture and texture-to-sketch configurations to evaluate bidirectional correspondence performance.

Figure 6. **Qualitative cross-sparsity correspondence comparison.** Each row demonstrates cross-sparsity correspondence results between sketches and various textured images. For each method, we present correspondence results using both sketch-to-texture and texture-to-sketch configurations to evaluate bidirectional correspondence performance.

[16] 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. 1, 3

[17] 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. 1

[18] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2955–2966, 2023.

[19] 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. 3

[20] 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. 1

[21] 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. 1

[22] Yuxin Zhang, Fan Tang, Weiming Dong, Haibin Huang, Chongyang Ma, Tong-Yee Lee, and Changsheng Xu. Domain enhanced arbitrary image style transfer via contrastive learning. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Proceedings*, page 1–8. ACM, 2022.