Towards Geometry-Grounded Dense Semantic Matching with VGGT Priors

Songlin Yang[1], Tianyi Wei[2], Yushi Lan[3], Zeqi Xiao[4], Anyi Rao[1], and Xingang Pan[4]

[1] MMLab@HKUST, The Hong Kong University of Science and Technology, HK
[2] Microsoft, USA
[3] Visual Geometry Group, University of Oxford, UK
[4] MMLab@NTU, Nanyang Technological University, Singapore

Abstract. Semantic matching aims to establish pixel-level correspondences between instances of the same category, and previous approaches based on 2D foundation models have achieved promising results on this task. However, they fail to achieve geometry-grounded dense semantic matching, which encompasses geometric awareness, manifold preservation, and cross-image invisibility reasoning. This new matching task simultaneously needs geometry-aware descriptors and holistic dense matching mechanisms, but existing approaches rarely provide both. To bridge this gap, we leverage VGGT, a 3D geometric foundation model that inherently provides strong geometry-grounded priors to support both capabilities in a unified framework. However, directly transferring faces two challenges: task heterogeneity, where VGGT originally matches cross-view images of the same instance, not cross-instance variations in shape or appearance; and scarce dense matching annotations for task adaptation. To address these challenges, we propose an approach that (i) retains VGGT's intrinsic capabilities by reusing early feature stages, fine-tuning later ones, and adding a semantic head for bidirectional correspondences; and (ii) adapts VGGT for semantic matching under data scarcity through cycle-consistent training strategy, synthetic data augmentation, and progressive training recipe. Extensive experiments demonstrate that our approach achieves superior geometric awareness, manifold preservation, and matching reliability, outperforming previous baselines.

Keywords: Geometry-Grounded Dense Semantic Matching · Dense Semantic Correspondence · VGGT · 3D Geometric Foundation Models

1 Introduction

Semantic matching aims to establish pixel-level correspondences between semantically equivalent regions across two images with the same-category instances. Recent zero-shot works [7, 59] show promising results. They leverage features from 2D foundation models, such as Stable Diffusion [42, 46] and DINO [34], as pixel-level descriptors, and compute correspondences via nearest-neighbor

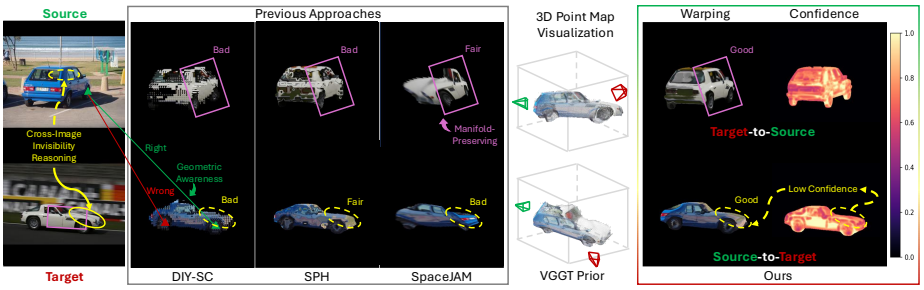

Fig. 1: Illustration of geometry-grounded dense semantic matching, limitations of previous approaches, and our improvements via leveraging VGGT priors. Previous approaches using 3D prior with nearest-neighbor matching (*e.g.*, DIY-SC [7], SPH [27]) or global alignment (*e.g.*, SpaceJAM [2]) each address only partially geometry-grounded requirements: geometric awareness, manifold preservation, and cross-image invisibility reasoning. By visualizing the predicted 3D point maps, we find VGGT [49], designed for geometry matching between views for one instance, can coarsely align same-category instances (more cases can be found in Fig. 8). VGGT also has inherent manifold-preserving matching capabilities. Built on VGGT, our approach not only achieves better geometry-aware and manifold-preserving matching, but also outputs prediction confidence for improving cross-image invisibility reasoning.

matching (*i.e.*, each pixel in the source image is matched to the pixel in the target image with the most similar feature).

However, practical downstream tasks (*e.g.*, morphing [54] and robotic affordance [19]) have raised the more demanding requirements of geometry-grounded dense matching, as shown in Fig. 1, which include: (i) Geometric Awareness: disambiguating symmetric or repetitive structures such as front and rear wheels; (ii) Manifold Preservation: maintaining consistent relative positions among pixels when mapped from source to target (*e.g.*, ensuring a source car window remains spatially coherent in the target image region, retaining its shape and internal layout rather than becoming distorted or fragmented); and (iii) Cross-Image Invisibility Reasoning: recognizing when a source pixel has no valid counterpart in the target image, typically due to occlusion or viewpoint-induced invisibility, which requires reasoning about 3D geometry rather than semantic similarity alone.

Previous approaches only partially achieve geometry-grounded matching. While some methods [2, 8, 27, 28, 36, 58] recognize the importance of geometric awareness and incorporate 3D priors to fine-tune 2D foundation model features, such as DIY-SC [7], SPH [27] in Fig. 1, they remain constrained by nearest-neighbor matching, failing to ensure manifold preservation and step towards cross-image invisibility reasoning (*e.g.*, occlusions). Conversely, methods such as SpaceJAM [2] in Fig. 1, that model global transformations to enforce manifold-preserving correspondences, often sacrifice the capacity for local non-rigid matching. These complementary shortcomings reveal that geometry-grounded seman-

tic matching requires a unified approach that simultaneously provides geometry-aware pixel descriptors and holistic dense matching mechanisms.

Considering these geometry-grounded matching requirements, we observe that they naturally align with recent advances in 3D geometric foundation models such as DUSt3R [50] and VGGT [49]. Although originally developed for 3D reconstruction from images, they provide strong geometry-aware features, and inherently learn image matching capabilities as a subtask of reconstruction [22]. As shown in Fig. 1, our preliminary evaluation of VGGT, a state-of-the-art reconstruction model, shows that it can coarsely align instances of the same category. Such capabilities make VGGT a highly promising basis for geometry-grounded dense semantic matching.

However, direct transferring faces two challenges: (i) Task Heterogeneity: VGGT was designed for geometric matching across-view images of the *same instance*, which differs fundamentally from semantic matching across *different instances* that exhibit variations in appearance and shape; (ii) Data Scarcity: The lack of dense correspondence annotations further complicate its adaptation.

To address these challenges, we design an approach that combines *capability retention* and *task-specific adaptation*. (i) For retaining geometry-aware features and holistic matching, we reuse VGGT's early feature stages, fine-tune its later ones, and append a semantic matching head that predicts dense bidirectional correspondence maps between source and target. (ii) For adaptation under limited annotations, we introduce a cycle-consistent training strategy that couples matching–reconstruction consistency with error–confidence correlation; curate a synthetic data pipeline generating diverse correspondences across categories and viewpoints; and adopt a progressive training recipe with aliasing artifact mitigation that gradually transfers dense correspondence ability from synthetic domains to real-world data. Extensive experiments and ablation studies demonstrate the superiority of our performance and the effectiveness of each key design.

Our contributions can be summarized as follows: (i) ***Beyond "Point Search" (Paradigm Shift)***: Traditional matching is bottlenecked by independent pixel-level nearest-neighbor search, which inherently fails at manifold preservation. We propose the first feed-forward framework unifying 3D-aware descriptors with a holistic dense matching rule. This enables demanding downstream tasks for geometry-grounded dense semantic matching that previous sparse matching simply cannot support. (ii) ***Beyond "Pseudo-3D"***: Semantic matching is inherently a 3D problem involving symmetry and occlusion. Rather than adapting 2D features with pseudo-3D labels, we leverage a native 3D foundation model to bridge geometric and semantic variations, providing a more fundamental and innovative solution. (iii) ***Beyond "Ad-Hoc Engineering"***: We resolve the long-standing scarcity of real-world dense correspondence labels via synthetic-to-real transfer. This may inspire the related tasks (*e.g.*, 3D registration) in overcoming the same bottleneck: transitioning from sparse, instance-specific to dense, category-level correspondence.

2 Related Works

Learning-Based Semantic Matching. Semantic matching is difficult due to appearance variations and scarce, ambiguous annotations [47,61]. Early methods relied on handcrafted feature descriptors [25], while the advent of deep learning facilitated the development of more effective feature extractors [17, 31, 41, 56] and direct semantic correspondence detection networks [10, 16, 40]. To address the limitations of scarce annotations, existing approaches have evolved three strategies: (i) leveraging weak supervision signals [5,20,48,55,60], (ii) employing warping and cycle consistency constraints [21, 47, 48, 63], and (iii) augmenting pseudo-label supervision [13, 15, 23].

2D-Foundation-Model-Based Semantic Matching. Recent research has demonstrated the potential of leveraging features from 2D foundation models for zero-shot semantic correspondence prediction [1, 6, 11, 44, 46, 59]. Among these, DINO features [4, 34] have been particularly effective [1, 8, 45, 58, 59]. Diffusion model features [42, 44] offer complementary strengths [8, 11, 24, 27, 46, 53, 58, 59]. However, simple nearest-neighbor search exhibits systematic limitations in disambiguating symmetric object parts [24, 26, 27, 43, 51, 58]. To address this, several studies have proposed appending adapter modules fine-tuned with pseudo or ground-truth labels [7, 53, 58]. Alternative approaches construct joint atlases for objects across multiple images [9,32]. [58] through keypoint-specific information, but this cannot resolve all symmetries via simple image transformations (*e.g.*, flipping) and requires keypoint-specific information that is generally unavailable. Approaches using 3D priors, such as DistillDIFT [8], SPH [27], and Jamais Vu [28], show promise but remain limited for cross-instance and complex cases.

3D Geometric Foundation Models. Recent works [62] such as DUSt3R [50] and VGGT [49] replace traditional multi-stage geometry pipelines with feed-forward transformers that directly predict 3D reconstruction signals (*e.g.*, camera poses and 3D point maps). Building on these efforts, MASt3R [22] demonstrates strong image matching performance within the same scene. In contrast, we extend this line of research from purely geometric to semantic matching, aiming to establish dense correspondences across different object instances within the same category.

3 Methodology

We introduce our approach from: (i) Architecture Adaptation: We review VGGT and present our motivation and extension (Sec. 3.1). (ii) Training: To adapt it to semantic matching with limited data, we propose a cycle-consistent training strategy (Sec. 3.2), curate a synthetic data pipeline (Sec. 3.3), and adopt a progressive training recipe (Sec. 3.4) with aliasing artifact mitigation (Sec. 3.5).

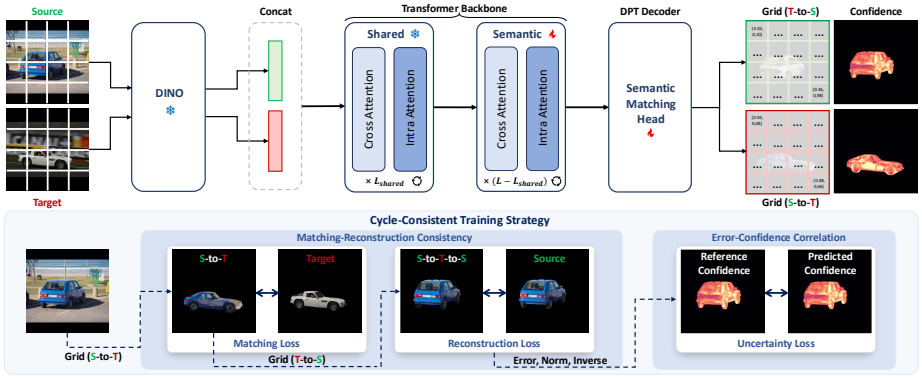

Fig. 2: Approach overview. (i) Pipeline: The source and target images are first processed through a DINO-based feature extractor, followed by a VGGT-adapted backbone for feature refinement, and finally output pixel-wise correspondences in grid form via a semantic matching head. (ii) Cycle-Consistent Training Strategy: Matching–reconstruction consistency, ensuring that mapping an image to its counterpart and back reconstructs the original, and error–confidence correlation, aligning reconstruction error with predicted confidence to encourage reliable predictions.

3.1 Semantic Correspondence Prediction with Geometry-Grounded Priors from VGGT

VGGT Preliminary. Given a set of N RGB images $\{\boldsymbol{I}_i\}_{i=1}^N \in \mathbb{R}^{H\times W}$, VGGT applies a single feed-forward transformer backbone with L blocks and DPT-based decoders [37], to predict camera parameters $\hat{\boldsymbol{g}}_i \in \mathbb{R}^9$ and pixel-level functional maps, such as depth $\hat{\boldsymbol{D}}_i \in \mathbb{R}^{H\times W}$ and 3D points $\hat{\boldsymbol{P}}_i \in \mathbb{R}^{3\times H\times W}$:

$$f:\ \{\boldsymbol{I}_i\}_{i=1}^N \longmapsto \{(\hat{\boldsymbol{g}}_i,\ \hat{\boldsymbol{D}}_i,\ \hat{\boldsymbol{P}}_i)\}_{i=1}^N. \tag{1}$$

Concretely, as shown in Fig. 2, each image is first patchified by a DINO encoder into tokens, which are processed by alternating frame-wise inter attention (within each image) and global cross attention (across all images) in the transformer backbone. Tokens are then reshaped into low-resolution feature maps $\hat{\boldsymbol{F}} \in \mathbb{R}^{C\times H'\times W'}$ and upsampled with DPT decoders to produce $\hat{\boldsymbol{D}}_i$ and $\hat{\boldsymbol{P}}_i$.

Motivation of Leveraging and Extending VGGT. Leveraging VGGT's priors and extending VGGT to semantic correspondence are well-motivated because: (i) Both cross-view geometric matching (same instance) and cross-instance semantic matching (same category) implicitly project images onto a canonical space, where instance-specific coordinates for geometry, category-level semantic landmarks for semantics. (ii) VGGT, trained for 3D reconstruction, extracts geometry-grounded representations from DINO's semantic-geometric features while implicitly yielding a semantic branch. (iii) VGGT's manifold-preserving matching capability is essential to current semantic correspondence approaches.

Semantic Correspondence Prediction. Building on VGGT pipeline, we aim to predict dense semantic correspondences between a source image $\boldsymbol{I}_s$ and a target image $\boldsymbol{I}_t$. As shown in Fig. 2, for the transformer backbone, we reuse the

early L_{shared} blocks (kept frozen) to extract geometry-grounded tokens, while fine-tuning the latter ($L - L_{shared}$) blocks to obtain semantic tokens. These are reshaped into feature maps $\hat{\boldsymbol{F}}_s$, $\hat{\boldsymbol{F}}_t \in \mathbb{R}^{C\times H'\times W'}$, which are fed into a new DPT-based semantic matching head ϕ_{match} to predict bidirectional sampling grids (*i.e.*, grids denote the coordinates used to sample color values from one image when generating the warping image): $\hat{\boldsymbol{G}}_{s\to t} \in [-1,1]^{2\times H\times W}$, $\hat{\boldsymbol{G}}_{t\to s} \in [-1,1]^{2\times H\times W}$, where coordinates are normalized to $[-1,1]$, following the convention of grid_sample function [35]. In addition, ϕ_{match} predicts pixel-wise confidence maps $\hat{\boldsymbol{C}}_s, \hat{\boldsymbol{C}}_t \in [0,1]^{H\times W}$ by adding one dimension, which provides reliability estimation for correspondences.

Training Objectives. The model is optimized through a progressive training recipe (Sec. 3.4) with: (i) *Supervised Loss (L_2 Loss)* from real data with sparse keypoints and synthetic data (Sec. 3.3) with dense correspondence labels (*i.e.*, grids); (ii) *Cycle-Consistency Loss* (Sec. 3.2) to refine matching and learn prediction uncertainty; and (iii) *Smoothness Loss* to mitigate aliasing artifacts (Sec. 3.5).

3.2 Cycle-Consistent Training Strategy

Given scarce dense annotations and the need to handle cross-image invisibility, we train the model to predict grids $\hat{\boldsymbol{G}}_{s\to t}, \hat{\boldsymbol{G}}_{t\to s}$ and confidences $\hat{\boldsymbol{C}}_s, \hat{\boldsymbol{C}}_t$ with a cycle-consistent strategy, including matching-reconstruction consistency and error-confidence correlation.

Matching-Reconstruction Consistency. Let $\mathcal{W}(\cdot, \boldsymbol{G})$ be grid_sample function, we can obtain matching and reconstruction images:

$$\hat{\boldsymbol{I}}_{s\to t} = \mathcal{W}(\boldsymbol{I}_s, \hat{\boldsymbol{G}}_{s\to t}), \qquad \hat{\boldsymbol{I}}_{t\to s} = \mathcal{W}(\boldsymbol{I}_t, \hat{\boldsymbol{G}}_{t\to s}), \tag{2}$$

$$\hat{\boldsymbol{I}}_{s\circlearrowleft} = \mathcal{W}(\hat{\boldsymbol{I}}_{s\to t}, \hat{\boldsymbol{G}}_{t\to s}), \qquad \hat{\boldsymbol{I}}_{t\circlearrowleft} = \mathcal{W}(\hat{\boldsymbol{I}}_{t\to s}, \hat{\boldsymbol{G}}_{s\to t}). \tag{3}$$

We then use $\boldsymbol{M}_s, \boldsymbol{M}_t \in \{0,1\}^{H\times W}$ as object masks [29, 38], and obtain the matching loss $\mathcal{L}_{\text{matching}}$ and reconstruction loss $\mathcal{L}_{\text{reconstruction}}$:

$$\mathcal{L}_{\text{matching}} = \boldsymbol{M}_t \odot \left\| E(\boldsymbol{I}_t) - E(\hat{\boldsymbol{I}}_{s\to t}) \right\|_2 \odot \hat{\boldsymbol{C}}_t + \boldsymbol{M}_s \odot \left\| E(\boldsymbol{I}_s) - E(\hat{\boldsymbol{I}}_{t\to s}) \right\|_2 \odot \hat{\boldsymbol{C}}_s, \tag{4}$$

$$\mathcal{L}_{\text{reconstruction}} = \boldsymbol{M}_s \odot \left\| \boldsymbol{I}_s - \hat{\boldsymbol{I}}_{s\circlearrowleft} \right\|_2 \odot \hat{\boldsymbol{C}}_s + \boldsymbol{M}_t \odot \left\| \boldsymbol{I}_t - \hat{\boldsymbol{I}}_{t\circlearrowleft} \right\|_2 \odot \hat{\boldsymbol{C}}_t. \tag{5}$$

where E is the DINO feature extractor. We weight predictions using the confidence maps $\hat{\boldsymbol{C}}_s, \hat{\boldsymbol{C}}_t$ to account for prediction uncertainty during matching.

Error-Confidence Correlation. We require the predicted confidence to be inversely correlated with the reconstruction error. Define per-pixel errors

$$\boldsymbol{e}_s = \left\| \boldsymbol{I}_s - \hat{\boldsymbol{I}}_{s\circlearrowleft} \right\|_2, \qquad \boldsymbol{e}_t = \left\| \boldsymbol{I}_t - \hat{\boldsymbol{I}}_{t\circlearrowleft} \right\|_2, \tag{6}$$

$$\boldsymbol{e}^*_s = \frac{\boldsymbol{e}_s - \min(\boldsymbol{M}_s \odot \boldsymbol{e}_s)}{\max(\boldsymbol{M}_s \odot \boldsymbol{e}_s) - \min(\boldsymbol{M}_s \odot \boldsymbol{e}_s)}, \qquad \boldsymbol{e}^*_t = \frac{\boldsymbol{e}_t - \min(\boldsymbol{M}_t \odot \boldsymbol{e}_t)}{\max(\boldsymbol{M}_t \odot \boldsymbol{e}_t) - \min(\boldsymbol{M}_t \odot \boldsymbol{e}_t)}. \tag{7}$$

and the reference confidences $\boldsymbol{C}_s = 1 - \boldsymbol{e^*}_s$, $\boldsymbol{C}_t = 1 - \boldsymbol{e^*}_t$. We correlate the confidence and error as:

$$\mathcal{L}_{\text{uncertainty}} = \left\| \boldsymbol{M}_s \odot (\boldsymbol{C}_s - \hat{\boldsymbol{C}}_s) \right\|_1 + \left\| \boldsymbol{M}_t \odot (\boldsymbol{C}_t - \hat{\boldsymbol{C}}_t) \right\|_1 - \lambda_{\text{conf}} \cdot \left(\sum \boldsymbol{M}_s \odot \hat{\boldsymbol{C}}_s + \sum \boldsymbol{M}_t \odot \hat{\boldsymbol{C}}_t \right). \tag{8}$$

3.3 Synthetic Data with Dense Annotations

We generate paired data as follows: (i) 3D Asset Generation: We select 18 categories from SPair-71k [30], expand textual descriptions for each instance via ChatGPT [33], and input these descriptions into Trellis [52]'s text-to-3D model to produce 3D assets. (ii) Rendering [39]: Each 3D asset is rendered from multiple viewpoints to generate RGB images and depth maps. The index of the corresponding 3D point for each pixel is recorded as an index map (serving as pixel-wise ground-truth labels). (iii) Multi-Condition Image Generation: For each rendered image, we extract a Canny [3] map and combine it with the depth map and different textual descriptions to condition FLUX [18], synthesizing RGB images with diverse textures. (iv) Paired Data Construction: For view-aligned pairs, transformations (*e.g.*, rotation, scaling) are applied to a canonical grid to generate warping grids (*i.e.*, $\boldsymbol{G}_{s\to t}, \boldsymbol{G}_{t\to s}$), followed by steps (i)-(iii) to produce paired images. For view-unaligned pairs, warping grids are derived by matching index maps between different 3D assets, then steps (i)-(iii) are used to generate paired images.

3.4 Progressive Training Recipe

During training, we optimize the model through a four-stage progressive strategy: (i) Synthetic Data Pretraining (Dense Supervision): Train exclusively on synthetic data with dense ground-truth annotations, employing only the L_2 loss and smoothness loss (Sec. 3.5). This stage aims to equip the model with VGGT's manifold-preserving mapping capabilities that generalize across different instances. (ii) Real Data Adaptation (Sparse Keypoint Supervision): Introduce real data with sparse ground-truth keypoints by adding a keypoint L_2 loss to the existing losses. This facilitates the transfer of dense mapping capabilities from synthetic to real-world domains. (iii) Matching Refinement (Matching and Reconstruction): Further optimize by incorporating the matching loss $\mathcal{L}_{\text{matching}}$ and reconstruction loss $\mathcal{L}_{\text{reconstruction}}$ to enhance matching precision. (iv) Uncertainty Learning (Confidence Prediction): Finally, integrate the uncertainty loss $\mathcal{L}_{\text{uncertainty}}$ to learn error-dependent confidence, enabling the model to predict prediction reliability.

3.5 Aliasing Artifacts and Mitigation

The matching head predicts continuous coordinates for discrete pixels, causing aliasing artifacts, visible as checkerboard patterns (Fig. 6, 6th column), due to the inherent ambiguity (*i.e.*, slight coordinate shifts in either direction remain plausible). We mitigate this with a smoothness loss that enforces spatial coherence by reshaping the grid into a vector and penalizing differences between adjacent coordinates.

4 Experimental Settings

4.1 Implementation Details

We use a VGGT-based Transformer with L=24 blocks, where the first 4 blocks are fixed, and the remaining 20 blocks are duplicated from VGGT to initialize for a new semantic branch. Additionally, a DPT decoder is added as the semantic matching head, extracting features from blocks [4, 11, 17, 23]. The model is trained for 5 days using a single A6000 GPU. The evaluation is performed on SPair-71k [30], AP-10k [57] (intra-species (I.S.), cross-species (C.S.), and cross-family (C.F.)) and our synthetic dataset.

4.2 Evaluation Metrics

(i) Matching Accuracy (From Sparse to Dense). For *sparse matching*, we adhere to standard protocols [9,12,27,30,53,58] and evaluate performance using the **(PCK)**. PCK@α is defined as the ratio of correctly predicted matched keypoints that lie within a radius of $R = \alpha \cdot \max(H, W)$ around their ground-truth positions, where H, W refer to the height and width of the object's bounding box, respectively. For *dense matching*, we evaluate the accuracy on the synthetic test set by measuring the Mean Squared Error (MSE) between the ground-truth images and the source images warped via predicted correspondences (*i.e.*, **Synthetic Dense**). Given the lack of dense annotations in real-world data, we utilize **DINO Accuracy** and **SD Accuracy** (defined as the MSE between the feature maps of the target and the warped source) as direct proxies to measure.
(ii) Geometric Awareness. To assess geometric awareness, we specifically compile results on symmetric keypoints (marked with an asterisk * in Tab. 1) on SPair-71k, which directly tests the model's ability to disambiguate repetitive or left-right symmetric structures.
(iii) Manifold Preservation. To holistically evaluate manifold preservation on real data, we compute the **FID** between target and warped source images. Assuming ideal cycle consistency ($\hat{I}_{s\circlearrowleft} \approx I_s$) implies geometric consistency and local invertibility, we further measure reconstruction fidelity via **PSNR** (signal-level), **SSIM** (structural), and **LPIPS** (perceptual). Together, these metrics quantify the model's manifold-preserving capabilities.
(iv) Cross-Image Invisibility Reasoning. We perform two specialized quantitative analyses on the synthetic dataset: (iv-a) AUC with Different Invisibilities (**AUC - Horizontal Rotation**): We calculate the Area Under the Curve (AUC) for occlusion detection across varying horizontal rotation angles, where increasing rotation serves as a direct proxy for greater cross-image invisibility due to occlusions. (iv-b) Confidence Evaluation (**MSE - Confidence**): We utilize our predicted confidence map as a universal reference to bin the matching errors of all evaluated methods into corresponding confidence intervals, verifying whether the confidence scores effectively reflect matching reliability and identify non-corresponding pixels.

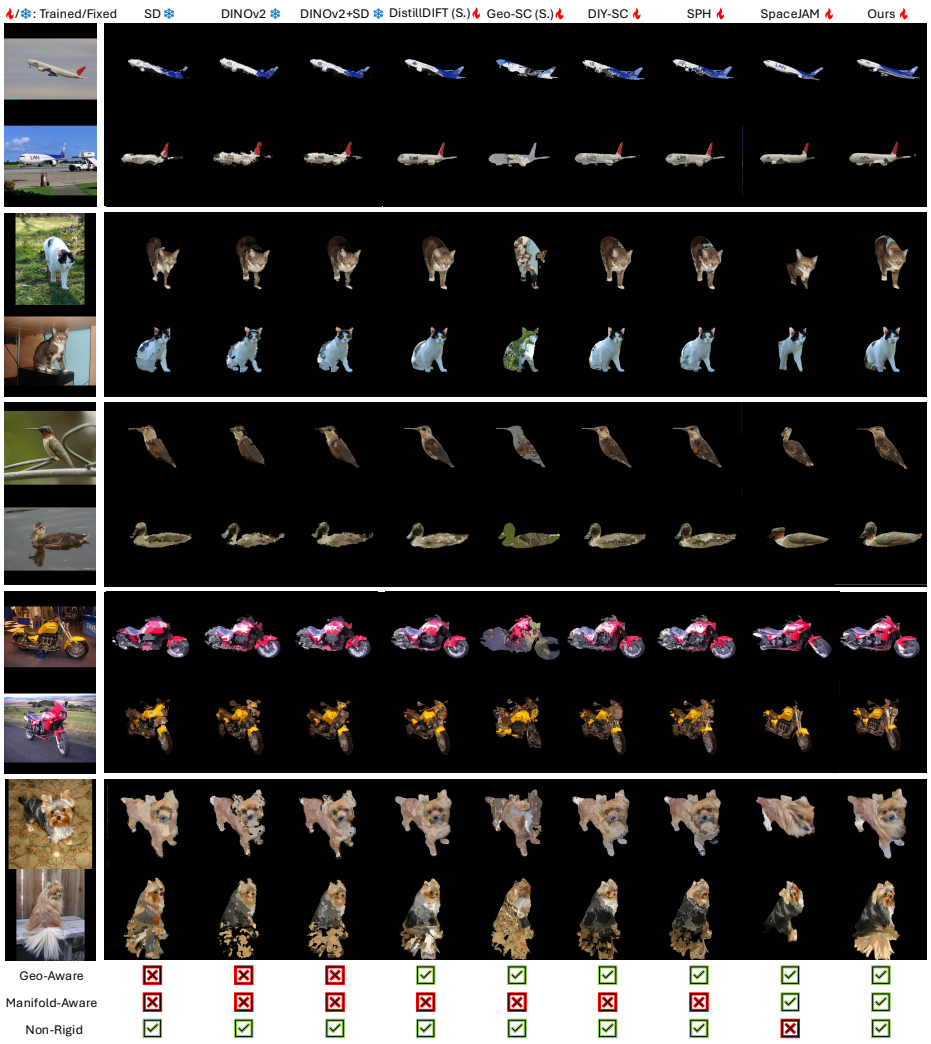

Fig. 3: Qualitative results of dense semantic matching compared with previous approaches.

4.3 Baseline Selection: Comprehensiveness and Fairness

We select baselines from: (i) Training Requirements: SD [46], DINO [34], and SD+DINO [59] are training-free, whereas Geo-SC [58] (Flip = No sparse keypoint tuning; S. = Sparse keypoint tuning applied), DIY-SC [7], SPH [27], ASIC [9], and SpaceJAM [2] involve training. Notably, DistillDIFT [8] (U.S. = No sparse keypoint tuning; S. = Sparse keypoint tuning applied) utilizes 3D synthetic data for training. Geo-SC, DIY-SC, SPH, ASIC, and SpaceJAM are trained with the same dataset as ours, while DistillDIFT uses more 3D synthetic data. (ii) Canon-

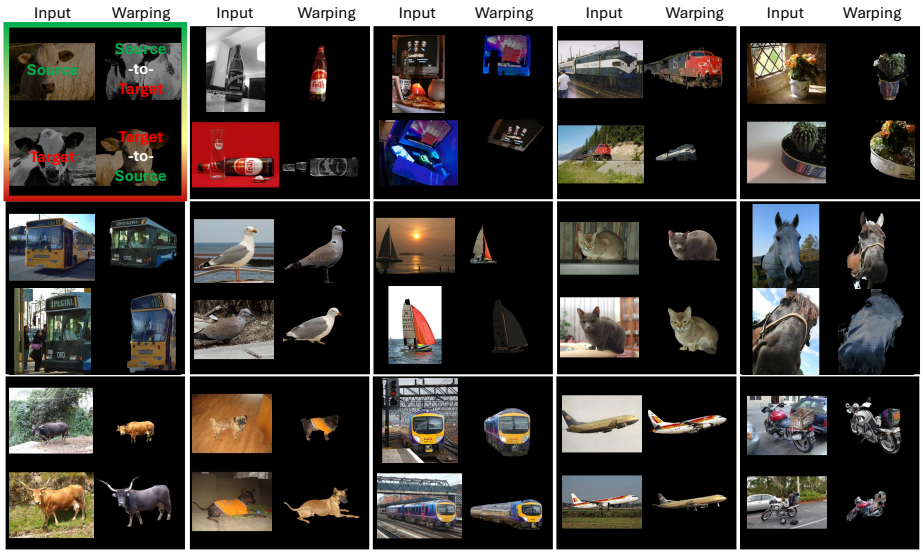

Fig. 4: More cases of dense semantic matching by our approach. These results highlight our superior generalization across diverse object categories, viewpoints, and non-rigid deformations.

ical Space Assumptions: SPH, ASIC, and SpaceJAM all adopt canonical space hypotheses. SPH explicitly requires same-category instances to share a spherical canonical space, while ASIC and SpaceJAM learn warp and transformations projected onto a canonical space, respectively.

5 Matching Evaluation

5.1 Dense Matching

Unlike sparse keypoint matching, which focuses only on a few salient points, dense semantic matching requires pixel-wise correspondences across the entire image, demanding geometric awareness, manifold preservation, and robustness to local non-rigidity. We present qualitative comparisons with baseline approaches in Fig. 3, and analyze across four critical dimensions: (i) Training Requirements: While fine-tuned approaches consistently outperform zero-shot approaches, GeoSC (S.) demonstrates limited effectiveness for dense matching despite improvements in keypoint matching. This stems from its keypoint-centric training objective that neglects dense correspondence. (ii) Geometric Awareness: Approaches incorporating geometric regularization, learning canonical spaces, or data augmentation exhibit superior performance in distinguishing symmetric and repetitive regions, a critical challenge for semantic matching. (iii) Manifold Preservation: With the exception of SpaceJAM and our approach, existing approaches

Table 1: Quantitative results on SPair-71k [30] and AP-10k [57]. 0.05* indicates additional aggregation on symmetric keypoints (*e.g.*, front/rear wheels).

Models		SPair-71k (PCK ↑)				AP-10k (PCK@0.1 ↑)			Synthetic ↓	DINO ↓	SD ↓	FID ↓	PSNR ↑	SSIM ↑	LPIPS ↓
		0.1	0.05	0.01	0.05*	I.S	C.S.	C.F.	Dense	Accuracy	Accuracy				
SD + DINO	[59]	59.9	44.7	7.9	40.3	62.9	59.3	48.3	0.20	0.35	0.28	28.73	17.88	0.59	0.249
DistillDIFT (U.S.)	[8]	60.8	45.4	8.0	44.7	65.8	64.2	56.1	0.16	0.21	0.18	32.04	18.08	0.62	0.242
Geo-SC (Flip)	[58]	65.4	49.1	9.9	51.0	68.7	64.6	52.7	0.14	0.23	0.16	29.45	18.02	0.63	0.245
DistillDIFT (S.)	[8]	65.3	49.9	8.9	49.8	66.9	64.7	58.0	0.15	0.17	0.16	18.98	18.95	0.63	0.221
Geo-SC (S.)	[58]	82.9	72.6	21.6	73.5	70.1	68.3	58.4	0.26	0.51	0.28	128.70	15.72	0.49	0.318
DIY-SC	[7]	71.6	54.2	10.1	53.8	70.6	69.1	57.8	0.11	0.15	0.14	25.46	17.22	0.56	0.268
SPH	[27]	64.4	48.2	8.4	49.3	65.4	63.1	51.0	0.10	0.17	0.15	23.89	16.51	0.53	0.289
ASIC	[9]	36.9	34.2	5.8	32.3	32.5	29.1	23.0	0.09	0.11	0.10	29.21	14.99	0.50	0.287
SpaceJAM	[2]	44.5	34.6	6.7	33.8	42.7	39.9	35.2	0.08	0.10	0.07	23.06	21.13	0.71	0.178
Ours		**84.2**	**76.3**	**25.9**	**77.8**	**73.2**	**71.4**	**60.9**	**0.07**	**0.06**	**0.06**	**3.58**	**23.84**	**0.78**	**0.142**

fail to maintain underlying manifold structures during matching. This limitation compromises their utility for downstream tasks (*e.g.*, affordance learning) that require precise preservation of original surface geometries. (iv) Local Non-Rigidity: While SpaceJAM's global transformation learning preserves manifolds, its inability to handle non-rigid deformations restricts performance on objects with complex topologies. Only our approach satisfies all the requirements for dense semantic correspondence while addressing these fundamental limitations. Additional quantitative evaluations appear in Tab. 1, with extended qualitative results presented in Fig. 4.

5.2 Sparse Keypoint Matching

Sparse keypoint matching evaluates the ability to match semantically salient regions with geometric awareness. Although previous approaches [8,27] attempted 3D synthetic data augmentation, they still struggle with geometry-ambiguous regions. In contrast, our approach surpasses previous approaches on this task, as shown in Tab. 1.

"Keypoint Overfitting" Trap. We observe an interesting phenomenon that solely overfitting sparse salient points neglects global manifold preservation. Despite high PCK, Geo-SC (S.) suffers in dense matching quality; in contrast, our method ensures superior fidelity. The 3D prior's robustness is confirmed on AP-10k, where both DIY-SC and ours outperform Geo-SC (S.), despite the latter's higher sparse PCK on SPair-71k. This illustrates the importance of our proposed holistic semantic matching paradigm and the introduction of 3D priors.

5.3 Matching Reliability: Invisibility & Confidence

We identify cross-image invisibility, where source pixels lack valid counterparts in the target image, as a critical yet overlooked factor that degrades matching reliability. Motivated by this, as shown in Fig. 5, we propose confidence map prediction, which not only introduces uncertainty calibration during matching learning to enhance optimization but also provides a reliability reference for downstream applications to selectively adopt high-confidence correspondences.

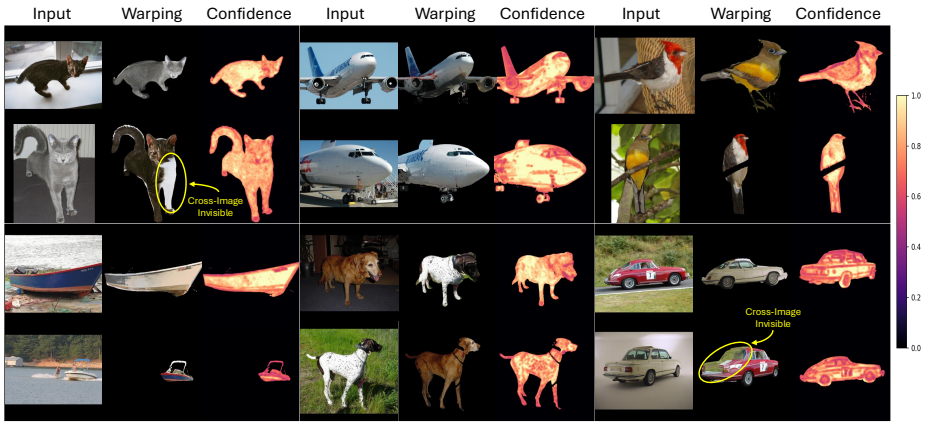

Fig. 5: More qualitative results of our approach on prediction reliability. Confidence map prediction is promising to addresse the overlooked issue of cross-image invisibility, enhancing optimization via uncertainty calibration and offering reliability cues for downstream tasks.

Table 2: Quantitative evaluation for invisibility & confidence on synthetic dataset with dense ground truth. Horizontal rotation angles denote the object's viewpoint deviation in the target image relative to the source. Confidence intervals represent predicted matching reliability, with higher values indicating greater certainty.

Models		AUC - Horizontal Rotation ↑					MSE - Confidence (%) ↓			
		5°	30°	45°	90°	Mean	90-100	80-90	60-80	0-60
SD + DINO	[59]	0.55	0.54	0.48	0.37	0.52	0.140	0.162	0.189	0.230
DistillDIFT (U.S.)	[8]	0.62	0.61	0.57	0.54	0.60	0.124	0.128	0.131	0.183
Geo-SC (Flip)	[58]	0.64	0.62	0.53	0.50	0.61	0.123	0.126	0.130	0.164
DistillDIFT (S.)	[8]	0.72	0.70	0.69	0.66	0.69	0.122	0.129	0.134	0.161
Geo-SC (S.)	[58]	0.59	0.52	0.48	0.42	0.51	0.182	0.189	0.195	0.296
DIY-SC	[7]	0.75	0.71	0.70	0.68	0.71	0.097	0.100	0.108	0.125
SPH	[27]	0.70	0.65	0.62	0.59	0.64	0.081	0.092	0.101	0.126
ASIC	[9]	0.82	0.80	0.75	0.72	0.78	0.082	0.089	0.095	0.112
SpaceJAM	[2]	0.85	0.81	0.73	0.70	0.81	0.078	0.083	0.089	0.108
Ours		**0.95**	**0.93**	**0.91**	**0.89**	**0.92**	**0.064**	**0.069**	**0.070**	**0.091**

To further evaluate invisibility quantitatively, we use AUC with different invisibilities. As shown in Tab. 2, ours consistently maintains the highest AUC across all angles. We then use our confidence map as a universal reference for all methods, and calculate MSE for the corresponding confidence interval based on the confidence value. All methods show an increase in MSE as confidence decreases. This consistent correlation proves that our confidence map truly identifies occlusions and non-corresponding pixels.

6 Ablation Study

Effectiveness of Each Key Design. As shown in Fig. 6 and Tab. 3, we assess the effectiveness of our proposed approach through two key aspects: (i)

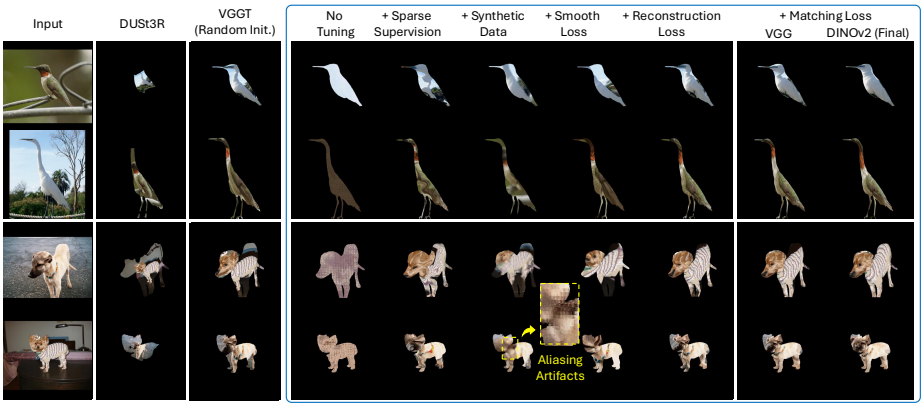

Fig. 6: Qualitative ablation results of the proposed approach. Ablation study on two key aspects: (i) 3D reconstruction pretraining, showing that DUSt3R and randomly initialized VGGT are inferior to our approach in semantic matching accuracy and training efficiency; and (ii) training strategy design, where stage-wise incorporation of synthetic data, sparse supervision, smoothness, reconstruction, and matching losses highlights the effectiveness of our final training recipe.

Table 3: Ablation of the proposed training recipe. We evaluate the contribution of each component step-by-step.

Settings	SPair-71k (PCK@0.1) ↑	Synthetic Dense ↓
DUSt3R	58.7	0.15
VGGT (Random Init.)	69.2	0.13
No Tuning	9.0	0.49
+ Sparse Supervision	79.4	0.18
+ Synthetic Data	79.3	0.11
+ Smoothness Loss	79.1	0.10
+ Reconstruction Loss	80.5	0.10
+ Matching Loss (VGG)	81.8	0.09
+ Matching Loss (DINO)	**84.2**	**0.07**

Table 4: Ablation of VGGT adaptation for shared blocks and DPT layers.

Models		SPair-71k	Synthetic ↓
Shared	DPT Layers	PCK@0.1 ↑	Dense
18	[4,11,17,23]	59.3	0.22
12	[4,11,17,23]	65.0	0.14
6	[4,11,17,23]	77.2	0.09
4	[4,11,17,23]	**84.2**	**0.07**
4	[3,10,16,22]	80.1	0.09
4	[2,9,15,21]	78.9	0.11
4	[1,8,14,20]	76.4	0.10

3D Reconstruction Pretraining Priors: We compare with DUSt3R [50] and randomly initialized VGGT. VGGT extends DUSt3R by using DINO as input and richer training data. DUSt3R shows manifold-preserving ability but lacks semantic richness due to the absence of 2D foundation model features (*e.g.*, DINO), while random VGGT achieves basic matching but requires far more refinement than our approach. (ii) Training Strategy Analysis: Our analysis reveals that directly using VGGT's geometric correspondences without tuning fails for semantic matching. The introduction of sparse supervision enables basic matching but compromises manifold preservation. Incorporating synthetic data restores manifold structure yet introduces aliasing artifacts. While smoothness loss mitigates these artifacts, it reduces matching accuracy. Further addition of reconstruction loss improves performance but remains suboptimal. We compare VGG loss [14] and DINO loss, ultimately selecting the latter as the optimal matching loss due to its superior fine-grained semantic matching accuracy.

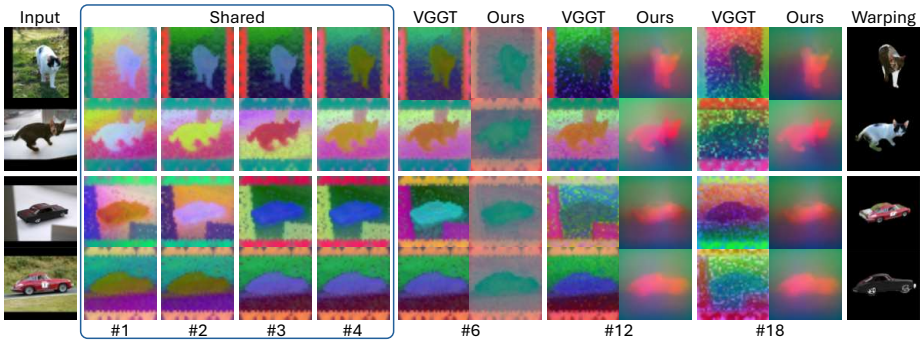

Fig. 7: Feature visualization of original VGGT backbone and our adapted semantic branch. Unlike the original VGGT branch, which yields noisy cross-instance activations, our fine-tuned semantic branch produces coherent, semantically aligned correspondences.

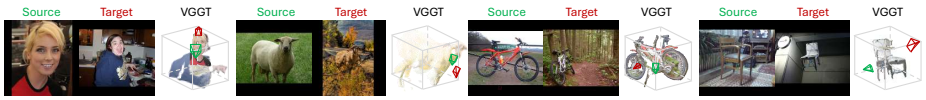

Fig. 8: More VGGT-predicted cases.

Backbone Block Selection and Feature Visualization. We propose an effective architectural adaptation that augments VGGT with semantic matching while preserving its extensibility for broader downstream tasks. We conduct in-depth analyses about this adaptation on two key aspects: (i) Backbone Adaptation Analysis: Testing the blocks of transformer backbone, we ultimately adopt the first 4 blocks as shared components and retain features from blocks [4, 11, 17, 23] for DPT input. Quantitative validations are provided in Tab. 4. (ii) Feature Visualization: We employ Principal Component Analysis (PCA) to project features from different blocks into three RGB channels for visualization. Visual comparisons reveal that VGGT's original branch struggles with cross-instance image pairs, exhibiting weak feature correlations and noisy activations. However, our fine-tuned semantic branch demonstrates clear matching coherence and smooth feature distribution, as shown in Fig. 7.

More VGGT Cases: Coarse Cross-Instance Alignment. To further validate our motivation regarding VGGT's capability for coarse alignment, we supplement additional examples in Fig. 8.

7 Conclusion

In this work, we have pioneered a framework for feed-forward dense semantic matching, fundamentally shifting the task from independent point-wise retrieval to holistic structural inference. By moving beyond the "Nearest Neighbor", our approach ensures global manifold consistency and structural coherence, enabling

the model to handle complex non-rigid deformations that traditional sparse matching paradigms fail to resolve. This transition from "searching for similarity" to "inferring geometric flow" provides a more promising foundation for dense semantic correspondence. We introduce a novel paradigm by repurposing 3D geometric foundation models for semantic correspondence. Our approach suggests that the native coupling of 3D structural priors and semantic descriptors offers a promising pathway to tackle symmetry and invisibility. By moving beyond conventional 2D-centric tuning, we leverage intrinsic 3D priors to effectively bridge the gap between geometry and semantics.

Limitations, Scalability and Future Work. (i) More Views: Although we propose a novel VGGT-based semantic matching approach, multi-view consistent semantic matching across scenes remains an open challenge. (ii) More Foundation Models: For compatibility with VGGT and fair baseline comparisons, we limit feature extraction to DINOv2. Future work could integrate additional features from other foundation models (*e.g.*, DINOv3 and Stable Diffusion). (iii) More Categories: To overcome the persistent scarcity of dense annotations, we have established a principled data-centric roadmap that synergizes synthetic-to-real transfer with cycle-consistent self-supervision. This strategy could be further explored on datasets with more diverse object categories. (iv) More Extreme: Beyond its current auxiliary role, our uncertainty modeling offers a path toward robust matching under extreme viewpoints and occlusions.

8 Acknowledgment

This work was supported by NTU SUG-NAP and S-Lab. This work was also partially supported by a grant from the NSFC/RGC Collaborative Research Scheme Project No. CRS_HKUST605/25.

References

1. Amir, S., Gandelsman, Y., Bagon, S., Dekel, T.: Deep ViT features as dense visual descriptors. ECCVW What is Motion For? (2022)
2. Barel, N., Weber, R.S., Mualem, N., Finder, S.E., Freifeld, O.: Spacejam: a lightweight and regularization-free method for fast joint alignment of images. In: European Conference on Computer Vision. pp. 180–197. Springer (2024)
3. Canny, J.: A computational approach to edge detection. IEEE Transactions on Pattern Analysis and Machine Intelligence **PAMI-8**(6), 679–698 (1986)
4. Caron, M., Touvron, H., Misra, I., Jegou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE (10 2021). https://doi.org/10.1109/iccv48922.2021.00951, http://dx.doi.org/10.1109/iccv48922.2021.00951
5. Chen, Y.C., Lin, Y.Y., Yang, M.H., Huang, J.B.: Show, match and segment: Joint weakly supervised learning of semantic matching and object co-segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence **43**(10), 3632–3647 (10 2021). https://doi.org/10.1109/tpami.2020.2985395, http://dx.doi.org/10.1109/tpami.2020.2985395

6. Cheng, X., Deng, C., Harley, A.W., Zhu, Y., Guibas, L.: Zero-shot image feature consensus with deep functional maps. In: European Conference on Computer Vision. pp. 277–293. Springer (2024)
7. Dünkel, O., Wimmer, T., Theobalt, C., Rupprecht, C., Kortylewski, A.: Do it yourself: Learning semantic correspondence from pseudo-labels. IEEE/CVF International Conference on Computer Vision (ICCV) (2025)
8. Fundel, F., Schusterbauer, J., Hu, V.T., Ommer, B.: Distillation of diffusion features for semantic correspondence. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). p. 6762–6774. IEEE (2 2025). https://doi.org/10.1109/wacv61041.2025.00658, http://dx.doi.org/10.1109/wacv61041.2025.00658
9. Gupta, K., Jampani, V., Esteves, C., Shrivastava, A., Makadia, A., Snavely, N., Kar, A.: ASIC: Aligning sparse in-the-wild image collections. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV). p. 4111–4122. IEEE (10 2023). https://doi.org/10.1109/iccv51070.2023.00382, http://dx.doi.org/10.1109/iccv51070.2023.00382
10. Han, K., Rezende, R.S., Ham, B., Wong, K.Y.K., Cho, M., Schmid, C., Ponce, J.: SCNet: Learning semantic correspondence. In: 2017 IEEE International Conference on Computer Vision (ICCV). p. 1849–1858. IEEE (10 2017). https://doi.org/10.1109/iccv.2017.203, http://dx.doi.org/10.1109/iccv.2017.203
11. Hedlin, E., Sharma, G., Mahajan, S., Isack, H., Kar, A., Tagliasacchi, A., Yi, K.M.: Unsupervised semantic correspondence using stable diffusion. Advances in Neural Information Processing Systems **36**, 8266–8279 (2023)
12. Huang, S., Yang, L., He, B., Zhang, S., He, X., Shrivastava, A.: Learning semantic correspondence with sparse annotations. In: European Conference on Computer Vision. pp. 267–284. Springer (2022)
13. Huang, Y., Sun, Y., Lai, C., Xu, Q., Wang, X., Shen, X., Ge, W.: Weakly supervised learning of semantic correspondence through cascaded online correspondence refinement. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV). p. 16208–16217. IEEE (10 2023). https://doi.org/10.1109/iccv51070.2023.01489, http://dx.doi.org/10.1109/iccv51070.2023.01489
14. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: European conference on computer vision. pp. 694–711. Springer (2016)
15. Kim, J., Ryoo, K., Seo, J., Lee, G., Kim, D., Cho, H., Kim, S.: Semi-supervised learning of semantic correspondence with pseudo-labels. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). p. 19667–19677. IEEE (6 2022). https://doi.org/10.1109/cvpr52688.2022.01908, http://dx.doi.org/10.1109/cvpr52688.2022.01908
16. Kim, S., Min, D., Jeong, S., Kim, S., Jeon, S., Sohn, K.: Semantic attribute matching networks. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). p. 12331–12340. IEEE (6 2019). https://doi.org/10.1109/cvpr.2019.01262, http://dx.doi.org/10.1109/cvpr.2019.01262
17. Kim, S., Min, D., Lin, S., Sohn, K.: DCTM: Discrete-continuous transformation matching for semantic flow. In: 2017 IEEE International Conference on Computer Vision (ICCV). p. 4539–4548. IEEE (10 2017). https://doi.org/10.1109/iccv.2017.485, http://dx.doi.org/10.1109/iccv.2017.485
18. Labs, B.F.: Flux. https://github.com/black-forest-labs/flux (2024)
19. Lai, Z., Purushwalkam, S., Gupta, A.: The functional correspondence problem. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). p.

15752–15761. IEEE (10 2021). https://doi.org/10.1109/iccv48922.2021.01548, http://dx.doi.org/10.1109/iccv48922.2021.01548
20. Lan, S., Yu, Z., Choy, C., Radhakrishnan, S., Liu, G., Zhu, Y., Davis, L.S., Anandkumar, A.: DiscoBox: Weakly supervised instance segmentation and semantic correspondence from box supervision. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE (10 2021). https://doi.org/10.1109/iccv48922.2021.00339, http://dx.doi.org/10.1109/iccv48922.2021.00339
21. Lan, Y., Loy, C.C., Dai, B.: DDF: Correspondence distillation from nerf-based gan. IJCV (2022)
22. Leroy, V., Cabon, Y., Revaud, J.: Grounding image matching in 3d with mast3r. In: European Conference on Computer Vision (ECCV) 2024. pp. 71–91. Springer Nature Switzerland, Cham, Switzerland (2024). https://doi.org/10.1007/978-3-031-73220-1_5
23. Li, X., Fan, D.P., Yang, F., Luo, A., Cheng, H., Liu, Z.: Probabilistic model distillation for semantic correspondence. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (6 2021). https://doi.org/10.1109/cvpr46437.2021.00742, http://dx.doi.org/10.1109/cvpr46437.2021.00742
24. Li, X., Lu, J., Han, K., Prisacariu, V.A.: Sd4match: Learning to prompt stable diffusion model for semantic matching. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). p. 27548–27558. IEEE (6 2024). https://doi.org/10.1109/cvpr52733.2024.02602, http://dx.doi.org/10.1109/cvpr52733.2024.02602
25. Liu, C., Yuen, J., Torralba, A.: SIFT Flow: Dense Correspondence Across Scenes and Its Applications, p. 15–49. Springer International Publishing (2016). https://doi.org/10.1007/978-3-319-23048-1_2, http://dx.doi.org/10.1007/978-3-319-23048-1_2
26. Luo, G., Dunlap, L., Park, D.H., Holynski, A., Darrell, T.: Diffusion hyperfeatures: Searching through time and space for semantic correspondence. Advances in Neural Information Processing Systems **36**, 47500–47510 (2023)
27. Mariotti, O., Aodha, O.M., Bilen, H.: Improving semantic correspondence with viewpoint-guided spherical maps. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). p. 19521–19530. IEEE (6 2024). https://doi.org/10.1109/cvpr52733.2024.01846, http://dx.doi.org/10.1109/cvpr52733.2024.01846
28. Mariotti, O., Du, Z., Bhalgat, Y., Mac Aodha, O., Bilen, H.: Jamais vu: Exposing the generalization gap in supervised semantic correspondence. arXiv preprint arXiv:2506.08220 (2025)
29. Medeiros, L.: Language segment-anything: Sam with text prompt. https://github.com/luca-medeiros/lang-segment-anything (2024), https://github.com/luca-medeiros/lang-segment-anything, accessed: 2025-08-31
30. Min, J., Lee, J., Ponce, J., Cho, M.: Spair-71k: A large-scale benchmark for semantic correspondence. arXiv preprint arXiv:1908.10543 (2019)
31. Novotny, D., Larlus, D., Vedaldi, A.: AnchorNet: A weakly supervised network to learn geometry-sensitive features for semantic matching. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). p. 2867–2876. IEEE (7 2017). https://doi.org/10.1109/cvpr.2017.306, http://dx.doi.org/10.1109/cvpr.2017.306
32. Ofri-Amar, D., Geyer, M., Kasten, Y., Dekel, T.: Neural congealing: Aligning images to a joint semantic atlas. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). p. 19403–19412. IEEE (6 2023).

https://doi.org/10.1109/cvpr52729.2023.01859, http://dx.doi.org/10.1109/cvpr52729.2023.01859
33. OpenAI: Chatgpt. https://openai.com/blog/chatgpt (2025), accessed: 2025-08-26
34. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., HAZIZA, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. TMLR (2023)
35. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32** (2019)
36. Qian, Q., Chen, H., Tomizuka, M., Keutzer, K., Wang, Q., Xu, C.: Bridging viewpoint gaps: Geometric reasoning boosts semantic correspondence. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11579–11589 (2025)
37. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 12179–12188 (2021)
38. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024)
39. Ravi, N., Reizenstein, J., Novotny, D., Gordon, T., Lo, W.Y., Johnson, J., Gkioxari, G.: Pytorch3d: An open-source library for 3d deep learning. arXiv preprint arXiv:2007.08501 (2020)
40. Rocco, I., Arandjelovic, R., Sivic, J.: Convolutional neural network architecture for geometric matching. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (7 2017). https://doi.org/10.1109/cvpr.2017.12, http://dx.doi.org/10.1109/cvpr.2017.12
41. Rocco, I., Arandjelovic, R., Sivic, J.: End-to-end weakly-supervised semantic alignment. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. p. 6917–6925. IEEE (6 2018). https://doi.org/10.1109/cvpr.2018.00723, http://dx.doi.org/10.1109/cvpr.2018.00723
42. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). p. 10674–10685. IEEE (6 2022). https://doi.org/10.1109/cvpr52688.2022.01042, http://dx.doi.org/10.1109/cvpr52688.2022.01042
43. Sommer, L., Dünkel, O., Theobalt, C., Kortylewski, A.: Common3d: Self-supervised learning of 3d morphable models for common objects in neural feature space. arXiv preprint arXiv:2504.21749 (2025)
44. Stracke, N., Baumann, S.A., Bauer, K., Fundel, F., Ommer, B.: Cleandift: Diffusion features without noise. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 117–127 (2025)
45. Suri, S., Walmer, M., Gupta, K., Shrivastava, A.: Lift: A surprisingly simple lightweight feature transform for dense vit descriptors. In: European Conference on Computer Vision. pp. 110–128. Springer (2024)
46. Tang, L., Jia, M., Wang, Q., Phoo, C.P., Hariharan, B.: Emergent correspondence from image diffusion. Advances in Neural Information Processing Systems **36**, 1363–1389 (2023)

47. Truong, P., Danelljan, M., Yu, F., Van Gool, L.: Warp consistency for unsupervised learning of dense correspondences. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). p. 10326–10336. IEEE (10 2021). https://doi.org/10.1109/iccv48922.2021.01018, http://dx.doi.org/10.1109/iccv48922.2021.01018
48. Truong, P., Danelljan, M., Yu, F., Van Gool, L.: Probabilistic warp consistency for weakly-supervised semantic correspondences. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). p. 8698–8708. IEEE (6 2022). https://doi.org/10.1109/cvpr52688.2022.00851, http://dx.doi.org/10.1109/cvpr52688.2022.00851
49. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: Vggt: Visual geometry grounded transformer. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 5294–5306 (2025)
50. Wang, S., Leroy, V., Cabon, Y., Chidlovskii, B., Revaud, J.: Dust3r: Geometric 3d vision made easy. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20697–20709 (2024)
51. Wimmer, T., Wonka, P., Ovsjanikov, M.: Back to 3d: Few-shot 3d keypoint detection with back-projected 2d features. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). p. 4154–4164. IEEE (6 2024). https://doi.org/10.1109/cvpr52733.2024.00398, http://dx.doi.org/10.1109/cvpr52733.2024.00398
52. Xiang, J., Lv, Z., Xu, S., Deng, Y., Wang, R., Zhang, B., Chen, D., Tong, X., Yang, J.: Structured 3d latents for scalable and versatile 3d generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 21469–21480 (2025)
53. Xue, F., Elflein, S., Leal-Taixé, L., Zhou, Q.: MATCHA: Towards matching anything. arXiv preprint arXiv:2501.14945 (2025)
54. Yang, S., Lan, Y., Chen, H., Pan, X.: Textured 3d regenerative morphing with 3d diffusion prior. IEEE/CVF International Conference on Computer Vision (ICCV) (2025)
55. Yang, S., Wang, W., Lan, Y., Fan, X., Peng, B., Yang, L., Dong, J.: Learning dense correspondence for nerf-based face reenactment. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 6522–6530 (2024)
56. Yi, K.M., Trulls, E., Lepetit, V., Fua, P.: LIFT: Learned Invariant Feature Transform, p. 467–483. Springer International Publishing (2016). https://doi.org/10.1007/978-3-319-46466-4_28, http://dx.doi.org/10.1007/978-3-319-46466-4_28
57. Yu, H., Xu, Y., Zhang, J., Zhao, W., Guan, Z., Tao, D.: AP-10k: A benchmark for animal pose estimation in the wild. In: Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2) (2021)
58. Zhang, J., Herrmann, C., Hur, J., Chen, E., Jampani, V., Sun, D., Yang, M.H.: Telling left from right: Identifying geometry-aware semantic correspondence. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). p. 3076–3085. IEEE (6 2024). https://doi.org/10.1109/cvpr52733.2024.00297, http://dx.doi.org/10.1109/cvpr52733.2024.00297
59. Zhang, J., Herrmann, C., Hur, J., Polania Cabrera, L., Jampani, V., Sun, D., Yang, M.H.: A tale of two features: Stable diffusion complements dino for zero-shot semantic correspondence. Advances in Neural Information Processing Systems **36**, 45533–45547 (2023)
60. Zhang, K., Fu, Y., Borse, S., Cai, H., Porikli, F., Wang, X.: Self-supervised geometric correspondence for category-level 6d object pose estimation in the wild. In: The Eleventh International Conference on Learning Representations (2023)

61. Zhang, K., Li, X., Lu, J., Han, K.: Semantic correspondence: Unified benchmarking and a strong baseline. arXiv preprint arXiv:2505.18060 (2025)
62. Zhang, W., Wu, Y., Li, S., Ma, W., Ma, X., Li, Q., Wang, Q.: Review of feed-forward 3d reconstruction: From dust3r to vggt (2025), https://arxiv.org/abs/2507.08448
63. Zhou, T., Krahenbuhl, P., Aubry, M., Huang, Q., Efros, A.A.: Learning dense correspondence via 3d-guided cycle consistency. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). p. 117–126. IEEE (6 2016). https://doi.org/10.1109/cvpr.2016.20, http://dx.doi.org/10.1109/cvpr.2016.20