BRIDGEVIEW: ENHANCING CROSS-VIEW CONSISTENCY OF VISION FOUNDATION MODELS VIA FROZEN OPPOSITE-VIEW TARGETS

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision foundation models provide powerful frozen features, yet image-based pretraining does not require two observations of the same surface to share a descriptor. Existing approaches enforce this relation with an intermediate scene representation and a rendered teacher. Reconstruction is costly, and the rendered teacher is a visibility-weighted mixture in the reconstructed field, so matching it moves student features away from pretrained coordinates. Leaving that coordinate system raises scene-level agreement while frozen semantic transfer falls. Instead, we establish cross-view consistency directly inside the pretrained coordinate system. To this end, we propose BridgeView, which uses the frozen opposite-view descriptor as the teacher on a same-surface pair formed by calibrated geometry. When the encoder cell at that pixel mixes neighboring instances, a bounded residual refines this readout with instance context while remaining inside a cone of the frozen vector. When many pairs observe one instance relation, their backward updates are coordinated inside that relation. We validate BridgeView on indoor RGB-D benchmarks. Across DINOv3 backbones, it raises cross-view correspondence, semantic segmentation, visual localization, and TSDF map rendering, and trains over $10\times$ faster than SnD. Scaling the pair corpus to 3M further improves correspondence. These results suggest a scalable route for large vision foundation models to obtain cross-view consistency inside the pretrained coordinate system. Code is provided in the supplementary material.

1 INTRODUCTION

Vision Foundation Models (VFMs) in the DINO family provide strong and robust features that transfer well to a wide range of downstream perception tasks, and they are widely used as frozen backbones (Caron et al., 2021; Oquab et al., 2024; Siméoni et al., 2025). Their image-based pretraining, however, does not enforce cross-view consistency: two observations of the same physical surface need not receive the same descriptor. This is not a minor limitation. Correspondence, visual localization, and reconstruction all depend on that relation, and Probe3D identifies the missing cross-view consistency of current foundation models as a distinct and measurable weakness (El Banani et al., 2024).

Existing approaches add this consistency through an intermediate scene representation and a rendered teacher. FiT3D fits a Gaussian scene per environment, aggregates descriptors from calibrated views, renders feature maps, and distills them into the encoder (Yue et al., 2024). SnD follows the same lift–render–distill path with feed-forward reconstruction and renders the teacher from novel viewpoints during training (Shavin & Benaim, 2026). In both cases, the training target is a feature rendered from a reconstructed scene. This target is a visibility-weighted mixture in the reconstructed field, not the frozen descriptor at the matched pixel. Matching this target raises shared-instance agreement across views, but it does so by moving student features away from pretrained coordinates. On frozen DINOv3-B correspondences, a Gaussian-rendered teacher reduces pretrained-coordinate cosine to 0.8905 and semantic probe agreement to 0.6109. Reconstruction also makes this route expensive.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

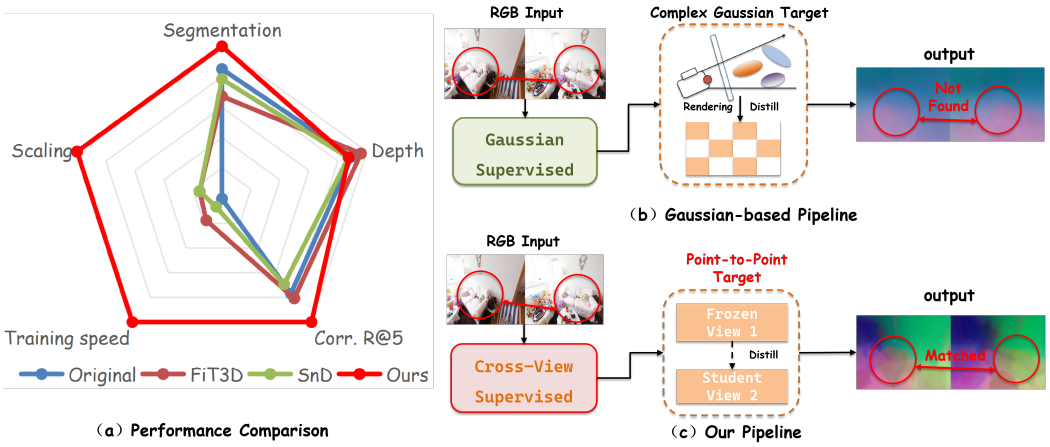

Figure 1: BridgeView enhances dense foundation representations with cross-view consistency from frozen opposite-view targets.

Instead, we establish cross-view consistency directly inside the pretrained coordinate system. To this end, we propose BridgeView. On a same-surface pair formed by calibrated geometry, the frozen opposite-view descriptor is the teacher. Matching this readout places two observations of one surface in pretrained coordinates. To make this readout reliable at scale, we address three challenges. First, a one-way projection can land on a different surface, so a pair is retained only when both views agree they observe the same surface. Second, a certified pixel can fall in an encoder cell that mixes neighboring instances. Pooling that cell into a prototype pulls the target toward the neighboring instance and away from the frozen readout, so a bounded residual refines this readout with instance context while remaining inside a cone of the frozen vector. Third, one instance relation can be observed by many pairs at once. Independent backward steps then pull that correspondence in disagreeing directions, so those updates are coordinated inside the relation. The overall information flow of BridgeView is summarized in Fig. 1.

We validate BridgeView on indoor RGB-D benchmarks. Across DINOv3 backbones, it raises cross-view correspondence, semantic segmentation, visual localization, and TSDF map rendering, and trains over $10\times$ faster than SnD. Scaling the pair corpus to 3M further improves correspondence. These results suggest a scalable route for large vision foundation models to obtain cross-view consistency inside the pretrained coordinate system.

The contributions of this paper are as follows.

1. BridgeView reads the frozen opposite-view descriptor as the teacher on a calibrated same-surface pair. Two observations of one surface then share **pretrained coordinates** rather than a rendered mixture.
2. A bounded residual and a coordinated backward step keep this target **reliable at scale**. The residual stays inside a cone of the frozen vector when an encoder cell is mixed; the coordination prevents independent updates from pulling the same correspondence apart.
3. Indoor RGB-D experiments show that BridgeView raises correspondence and semantic segmentation, including on DINOv3-7B, and improves localization and TSDF rendering. This pair-based training is over $10\times$ faster than SnD and further improves correspondence as the corpus scales to 3M.

2 Related Work

Dense visual foundation models. Recent advances in ViT-based vision foundation models have produced highly transferable visual representations that excel on a variety of 2D tasks (Dosovitskiy et al., 2021; Radford et al., 2021; Zhou et al., 2022; Touvron et al., 2022; He et al., 2022). Our

work builds upon encoders from the DINO family (Caron et al., 2021; Oquab et al., 2024; Siméoni et al., 2025). Despite their success, image-based pretraining does not require two observations of the same surface to receive consistent descriptors, and Probe3D measures this missing cross-view consistency (El Banani et al., 2024). This missing consistency limits correspondence, visual localization, and reconstruction.

Cross-view correspondence. That missing consistency is a relation between two views of the same surface. LoFTR, DKM, RoMa, and LightGlue recover it as pixel pairs that correspond to the same 3D point, and use those pairs for camera pose, SfM, and visual localization (Sun et al., 2021; Edstedt et al., 2023; 2024; Lindenberger et al., 2023). CroCo and CroCo v2 reconstruct a masked view with another view of the same scene (Weinzaepfel et al., 2022; 2023). DUSt3R, MASt3R, and VGGT make cameras, pointmaps, and dense matches explicit (Wang et al., 2024; Leroy et al., 2024; Wang et al., 2025). The common objective of this line is to train a matcher or a geometric model and to emit correspondences, cameras, or pointmaps. That output leaves the underlying encoder unchanged, so the gain stays inside the matching or reconstruction task. Inspired by this two-view relation, we form same-surface pairs from calibrated geometry. BridgeView avoids training a matcher. Instead, it uses the pair to enhance an existing DINO-family encoder, and the exported backbone reads a single RGB image.

Scene-mediated VFM enhancement. Existing methods use an intermediate scene representation to add that missing consistency to a pretrained 2D encoder. Neural Radiance Fields have become a cornerstone for photorealistic novel-view synthesis (Mildenhall et al., 2020). 3D Gaussian Splatting provides an explicit scene representation with faster rendering (Kerbl et al., 2023). Building on these representations, feature-field methods lift image descriptors into 3D for querying and novel-view synthesis (Tschernezki et al., 2022; Kobayashi et al., 2022; Kerr et al., 2023; Zhou et al., 2024). These methods distill 2D features into a 3D representation. FiT3D and SnD supervise a 2D encoder with a rendered teacher (Yue et al., 2024; Shavin & Benaim, 2026). That teacher is a visibility-weighted mixture in the reconstructed field. BridgeView uses the frozen opposite-view descriptor as the teacher, so the training target remains in pretrained coordinates, and a bounded instance residual with relation gradient coordination keeps that target reliable at scale.

3 METHOD

3.1 PROBLEM DEFINITION

The input is a calibrated RGB-D pair and a frozen teacher that maps each view v to a token grid $\bar{Z}^v$; the student produces a grid Z^v of the same shape. For a pixel $\boldsymbol{u}$ in view v and its match $\boldsymbol{u}^{\bar{v}}$ in the opposite view, which observe the same surface, the learning problem is

$$R(Z^v, \boldsymbol{u}) \;\approx\; R(\bar{Z}^{\bar{v}}, \boldsymbol{u}^{\bar{v}}), \tag{1}$$

where $R(G, \boldsymbol{u})$ is bilinear readout of a token grid G at pixel $\boldsymbol{u}$, and $\bar{v}$ is the opposite view. The right-hand side is a readout of the frozen grid. Existing methods instead replace it with a teacher rendered from a reconstructed scene, a visibility-weighted mixture $\hat{Z}^v(\boldsymbol{u}) = \sum_k w_k(\boldsymbol{u})\, \bar{Z}^v_k$ in that field, where k indexes reconstructed primitives and $w_k(\boldsymbol{u})$ is the visibility weight at $\boldsymbol{u}$. Matching this mixture relocates the student in the reconstructed field. Keeping the frozen readout as the target avoids this relocation, but it raises three obstacles. A one-way projection can land on a different surface. The encoder cell at a matched pixel can mix neighboring instances. One instance relation can be observed by many certified pixels whose updates disagree. Sections 3.2–3.4 address these obstacles in turn, also shown in Fig. 2. Training and implementation details are in Appendix A.3.

3.2 CERTIFIED FROZEN PAIR READOUT

A one-way projection can land on a different surface. When the two cameras do not both observe the same point, the projected pixel belongs to an occluder or a background region, and the opposite-view descriptor then describes the wrong observation. Using such a pair as a target pulls the student toward a descriptor of a different surface.

To address this, CFPR retains a pixel pair only when both views agree that they observe the same surface, and then reads the frozen opposite-view descriptor at the opposite endpoint. Let $\boldsymbol{u}^a$ be a

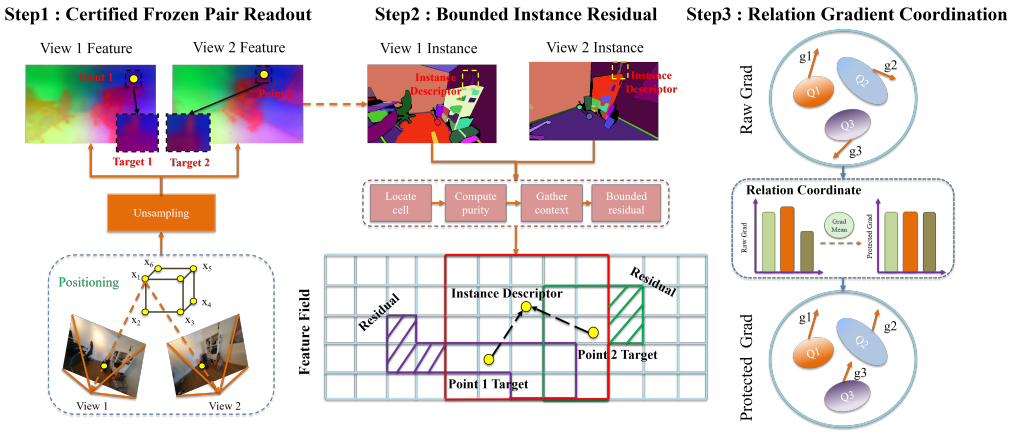

Figure 2: Architecture overview. A frozen opposite-view target on a matched pair is refined by a bounded instance residual and relation gradient coordination.

pixel in view a. Calibrated cameras lift it to a world point and project it into view b. Write $\boldsymbol{u}^b$ for the landing pixel in view b. The same-surface test uses four gates, a validity indicator $\mathbf{1}_{\text{valid}}$, 3D agreement e_{3D}, round-trip reprojection e_a, e_b, and depth agreement e_d^a, e_d^b. The scalars τ_{3D}, τ_{rep}, and τ_d are the thresholds of those residuals. CFPR sets

$$\chi_{ab}(\boldsymbol{u}^a) = \mathbf{1}_{\text{valid}}\,\mathbf{1}[e_{3D} \le \tau_{3D}]\,\mathbf{1}[e_a, e_b \le \tau_{\text{rep}}]\,\mathbf{1}[e_d^a, e_d^b \le \tau_d]. \tag{2}$$

Thus $\chi_{ab}(\boldsymbol{u}^a) = 1$ if and only if the landing is valid and every residual lies inside its threshold. The reverse indicator $\chi_{ba}(\boldsymbol{u}^b)$ is built from b to a in the same way. The certified pair set is $\mathcal{P}_{ab} = \{(\boldsymbol{u}^a, \boldsymbol{u}^b) \mid \chi_{ab}(\boldsymbol{u}^a)\,\chi_{ba}(\boldsymbol{u}^b) = 1\}$. This formulation ensures that every retained pair observes the same surface in both cameras. The transport maps and residual definitions are in Appendix C.1.

For a certified pair i with endpoints $(\boldsymbol{u}_i^a, \boldsymbol{u}_i^b) \in \mathcal{P}_{ab}$, CFPR reads the opposite frozen descriptor and stops its gradient,

$$p_i^v = \text{sg}\big(R(\bar{Z}^{\bar{v}}, \boldsymbol{u}_i^{\bar{v}})\big), \tag{3}$$

where R is the bilinear readout of Eq. equation 1, $v \in \{a, b\}$ is the student's view, $\bar{v}$ is the other view, sg is the stop-gradient, and $p_i^v \in \mathbb{R}^C$ is the point target in view v. Matching $R(Z^v, \boldsymbol{u}_i^v)$ to this target places two observations of one surface in pretrained coordinates.

3.3 Bounded Instance Residual

A certified pixel can still fall in an encoder cell that mixes neighboring instances. The encoder emits one descriptor per cell, so the pixel footprint of that cell can cover two identities, and the cell mean is then a mixture. Pooling the cell into a prototype pulls this mean toward the neighbor. Moving the frozen opposite-view readout toward that mean without a bound rotates the target away from the descriptor that CFPR just read.

To address this, BIR adds instance and boundary context as a residual around that frozen readout, and keeps the result inside a cone of that readout. We first test whether the cell is mixed. Let M be the training-time instance map, $\Omega(\boldsymbol{u})$ the pixel footprint of the encoder cell that contains the certified pixel $\boldsymbol{u}$, and $m(\boldsymbol{u})$ the identity occupying the most pixels in that cell. Occupancy is the fraction of the cell held by $m(\boldsymbol{u})$,

$$\rho(\boldsymbol{u}) = \frac{1}{|\Omega(\boldsymbol{u})|} \sum_{x \in \Omega(\boldsymbol{u})} \mathbf{1}[M(x) = m(\boldsymbol{u})]. \tag{4}$$

Thus $\rho(\boldsymbol{u}) = 1$ if and only if $\Omega(\boldsymbol{u})$ is one instance. BIR admits the cell when $\rho(\boldsymbol{u}) \ge \tau_\rho$, with $\tau_\rho = 0.75$.

When the cell is admitted, two frozen residuals enter around p_i^v. Let $r = m(\boldsymbol{u}_i^a)$ and $s = m(\boldsymbol{u}_i^b)$ be the endpoint identities. Write μ_r^a for the occupancy-gated mean frozen readout of identity r

in view a, and μ_s^b for identity s in view b. The interior residual is the normalized mean $\mu_i = \mathrm{norm}(\mu_r^a + \mu_s^b)$, with $\mathrm{norm}(x) = x/\|x\|_2$. Let b_i^v be the mean frozen readout on neighboring cells whose dominant identity differs from $m(\boldsymbol{u}_i^v)$, minus p_i^v. The boundary residual is the component of this offset orthogonal to the frozen readout,

$$b_{i,\perp}^v = b_i^v - \langle b_i^v, p_i^v \rangle p_i^v. \tag{5}$$

The refined target adds both residuals and projects onto a cone of p_i^v,

$$t_i^v = \mathrm{cap}_\phi\big(\mathrm{norm}(p_i^v + \Delta_i^v), p_i^v\big), \tag{6}$$

with $\Delta_i^v = \alpha_i(\mu_i - p_i^v) + \beta_i b_{i,\perp}^v$ and $\phi = 12°$. Here α_i and β_i are nonnegative scales on the interior residual and the boundary residual. Each scale is positive only when that residual is admitted, and is otherwise zero. The map $\mathrm{cap}_\phi(x, p)$ returns x if $\angle(x, p) \le \phi$, and otherwise the unit vector in $\mathrm{span}\{p, x\}$ at angle ϕ from p. If both scales are zero, then $t_i^v = p_i^v$. This formulation keeps instance and boundary context inside the cone of the frozen opposite-view readout.

3.4 Relation Gradient Coordination

One instance relation (r, s) is observed by many certified pixels in the same image pair and student view. Each pixel has its own cosine gap to its BIR target and its own pointwise gradient. Independent updates then pull the correspondence in disagreeing directions, which appears as variance in the shared update. RGC therefore mixes each pointwise gradient with the group mean to reduce this variance, and leaves the scalar loss unchanged.

Formally, write i for a certified endpoint in student view v. Let $z_i = R(Z^v, \boldsymbol{u}_i^v)$ be the bilinear readout of the student grid at that pixel, and let t_i^v be the BIR target in the same view. The per-endpoint cosine gap and its gradient are

$$\ell_i = 1 - \langle z_i, t_i^v \rangle, \qquad \boldsymbol{g}_i = \frac{\partial \ell_i}{\partial z_i}. \tag{7}$$

where $z_i, t_i^v \in \mathbb{R}^C$ are unit-length, $\ell_i \in [0, 2]$ is the cosine gap, and $\boldsymbol{g}_i \in \mathbb{R}^C$ is the gradient of endpoint i.

RGC groups endpoints that share one image pair, one student view, and one identity pair (r, s). Write G_i for that set. Thus $j \in G_i$ if and only if j and i share that pair, view, and identity pair, and $|G_i| \ge 2$ if and only if the relation is observed more than once.

Treat each gradient in G_i as a noisy observation of one shared relation gradient $\bar{\boldsymbol{g}}^*$,

$$\boldsymbol{g}_j = \bar{\boldsymbol{g}}^* + \boldsymbol{\varepsilon}_j, \qquad \mathbb{E}[\boldsymbol{\varepsilon}_j] = \mathbf{0}, \qquad \mathrm{Cov}(\boldsymbol{\varepsilon}_j) = \sigma^2 I. \tag{8}$$

where $\bar{\boldsymbol{g}}^* \in \mathbb{R}^C$ is the expected gradient of the relation, $\boldsymbol{\varepsilon}_j$ is zero-mean endpoint noise, and σ^2 is the noise variance. The group mean $\bar{\boldsymbol{g}}_{G_i} = \frac{1}{|G_i|} \sum_{j \in G_i} \boldsymbol{g}_j$ then has variance $\sigma^2/|G_i|$, so averaging reduces variance by the group size. This model treats the noise terms as uncorrelated. Nearby pixels can share noise.

When $|G_i| \ge 2$, RGC mixes the pointwise gradient with that mean,

$$\widetilde{\boldsymbol{g}}_i = (1 - \lambda)\boldsymbol{g}_i + \lambda \bar{\boldsymbol{g}}_{G_i}. \tag{9}$$

where $\widetilde{\boldsymbol{g}}_i \in \mathbb{R}^C$ is the coordinated update of endpoint i. We set $\lambda = 0.50$, an equal mix of the pointwise gradient and the relation mean. As shown in Figure 3, larger $|G_i|$ lowers this noise under the model. If $|G_i| < 2$, then $\lambda = 0$ and $\widetilde{\boldsymbol{g}}_i = \boldsymbol{g}_i$. The forward pass keeps ℓ_i, and the backward pass writes $\widetilde{\boldsymbol{g}}_i$ in place of $\boldsymbol{g}_i$. This formulation coordinates updates inside one instance relation.

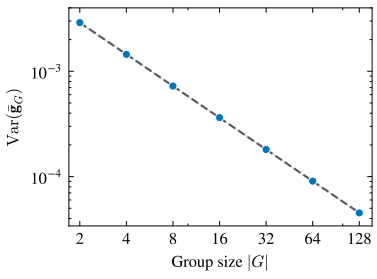

Figure 3: Larger $|G_i|$ lowers update noise.

3.5 Training Objective

The student is supervised by the BIR target t_i^v, formed symmetrically in both views of each image pair. With the per-endpoint gap ℓ_i from Eq. equation 7, the batch objective is

$$\mathcal{L}_{\mathrm{target}} = \sum_{i \in \mathcal{Q}} w_i \, \ell_i, \qquad w_i = \frac{1}{|\mathcal{Q}|}. \tag{10}$$

where $\mathcal{Q}$ is the set of certified endpoints in the current mini-batch and w_i is the uniform weight over that set.

Backpropagation uses the coordinated gradient $\widetilde{\boldsymbol{g}}_i$ from RGC. The teacher remains frozen. Training updates only the low-rank adapters $\Delta\theta$ on the student encoder. At inference, the encoder processes a single RGB image. Depth, camera poses, and instance maps are not used.

4 Experiments

4.1 Experimental Settings

4.1.1 Dataset and Metrics

We evaluate whether the enhanced features improve cross-view consistency while retaining frozen semantic segmentation. Correspondence and indoor localization are measured on ScanNet (Dai et al., 2017). Semantic segmentation and monocular depth are evaluated on ScanNet++, ScanNet, and NYUv2 (Yeshwanth et al., 2023; Silberman et al., 2012). TSDF map rendering is evaluated on withheld views of ScanNet++. Training pairs are overlapping RGB-D views sampled from the official ScanNet++ training split of 230 scenes. The 3M corpus covers these 230 scenes with 13,043 view pairs per scene on average. That split is disjoint from the official ScanNet++, ScanNet, and NYUv2 evaluation splits, and from the fixed ScanNet correspondence set. Correspondence reports median error and R@10. Localization reports median rotation and translation error and pose recall. TSDF rendering reports MV δ_1, Accuracy@10cm, and Completeness@10cm. Linear probes on the frozen encoder report mIoU and RMSE. Dataset splits and evaluation protocols are given in Appendix A.1. Pair construction is given in Appendix A.2.

4.1.2 Implementation Details

We compare the original frozen encoder against FiT3D and SnD (Yue et al., 2024; Shavin & Benaim, 2026). SnD and BridgeView update LoRA adapters on the student encoder. FiT3D fine-tunes the full encoder. The geometric-matcher comparison uses RoMa v2 (Edstedt et al., 2025), which reads both views at test time and does not update the image encoder. Unless otherwise stated, BridgeView is trained on the 3M pair corpus from the 230-scene ScanNet++ training split. Training uses calibrated color and depth pairs. Intrinsics, poses, and depth are used only to construct the training targets. The instance map M is produced by SAM (Kirillov et al., 2023) on each training view and is used only to construct BIR targets. The exported encoder reads an RGB image; depth, intrinsics, poses, and the instance map are not used at inference. Within each comparison, the methods share the backbone initialization, training split, image inputs, and preprocessing, use a matched update budget, and are measured on the same NVIDIA H20 GPUs. Further settings are given in Appendix A.3.

4.2 Evaluation Results

4.2.1 Overall Performance

Correspondence across views. As shown in Table 1, BridgeView achieves the lowest Median Error and the highest R@10 among all compared methods on DINOv3-S and DINOv3-B. Against SnD, Median Error falls from 29.62 to 17.87 pixels on DINOv3-S and from 26.54 to 12.88 pixels on DINOv3-B, while R@10 rises from 31.75% to 42.90% and from 34.03% to 46.34%. On DINOv3-7B, BridgeView raises R@10 from 45.43% to 58.21% over Original. As shown in Appendix Table 8, SnD with the same LoRA setting exceeds the memory budget at this scale. The improvement holds across all three backbone sizes.

A geometric matcher is the other existing answer to the missing cross-view relation, and it solves a different problem. RoMa v2 recovers that relation as two-view pixel pairs but leaves the image encoder unchanged, so its gain stays inside the matching model and both views remain required at test time. As shown in Table 2, BridgeView strongly outperforms RoMa v2 on the same 1,500 pairs and the same query pixels. Median Error falls from 55.71 to 12.88 pixels, and R@5, R@10, and R@20 rise from 2.77%, 4.60%, and 11.32% to 35.34%, 46.34%, and 56.09%. The matcher can emit correspondences without making two observations of one surface agree in the encoder, whereas BridgeView writes that agreement into the encoder and the exported backbone then reads a single

Table 1: Correspondence and frozen dense transfer on DINOv3.

		ScanNet Corr.		ScanNet++		ScanNet		NYUv2	
Backbone	Method	Med. Err. ↓	R@10 ↑	mIoU ↑	RMSE ↓	mIoU ↑	RMSE ↓	mIoU ↑	RMSE ↓
DINOv3-S	Original	26.80	37.50	32.37	0.3083	45.86	0.2939	70.65	**0.4137**
	FiT3D	18.55	41.00	32.43	0.2985	45.59	0.2964	69.55	0.4504
	SnD	29.62	31.75	31.67	0.3144	45.79	**0.2906**	70.16	0.4222
	BridgeView	**17.87**	**42.90**	**32.88**±0.11	**0.2952**±0.0007	**46.45**±0.08	0.2954±0.0004	**70.76**±0.06	0.4225±0.0005
DINOv3-B	Original	33.75	35.32	35.90	0.2864	51.20	0.2684	74.89	0.3765
	FiT3D	19.75	39.46	35.18	0.2674	49.70	0.2721	74.14	0.4116
	SnD	26.54	34.03	35.19	0.2941	50.67	**0.2593**	75.13	**0.3668**
	BridgeView	**12.88**	**46.34**	**36.70**±0.07	**0.2567**±0.0009	**51.53**±0.09	0.2636±0.0008	**75.28**±0.04	0.3694±0.0007
DINOv3-7B	Original	15.56	45.43	39.49	0.2876	57.54	0.2104	78.04	0.3070
	BridgeView	**5.68**	**58.21**	**40.02**±0.08	**0.2753**±0.0007	**58.43**±0.08	**0.2093**±0.0005	**78.69**±0.04	**0.2972**±0.0004

Table 2: DINOv3-B correspondence against RoMa v2.

Method	Med. Err. ↓	R@5 ↑	R@10 ↑	R@20 ↑
RoMa v2	55.71	2.77	4.60	11.32
BridgeView	**12.88**	**35.34**	**46.34**	**56.09**

RGB image. These results show that repairing the encoder, rather than adding a two-view matching model, is what raises correspondence.

In Appendix Table 12, we show our model performance when the pair corpus and the update budget are varied on DINOv3-B. The 640K set is a subset of the 3M pool. Because SnD and BridgeView consume different numbers of views per training unit, the comparison matches optimizer updates. Compared to SnD, BridgeView performs better under all configurations. At 10K updates, the 640K model already exceeds SnD, raising R@10 from 34.18% to 37.83% and lowering Median Error from 26.59 to 26.08 pixels. More significantly, with the 3M corpus at the same budget, R@10 rises to 44.53% and Median Error falls to 15.16 pixels. A longer SnD schedule does not improve correspondence. R@10 moves from 34.18% to 34.03%, and Median Error from 26.59 to 26.54 pixels. This behavior is unlike BridgeView, whose accuracy improves with longer training on both corpora, from 37.83% to 39.38% on 640K and from 44.53% to 46.34% on 3M. These observations show that SnD saturates under a longer schedule, while BridgeView continues to improve and surpasses SnD under every pair corpus and update budget.

Linear probe performance. BridgeView improves frozen semantic transfer on every DINOv3 backbone, and it does so without leaving the pretrained coordinate system. As shown in Table 1, BridgeView raises mIoU on ScanNet++, ScanNet, and NYUv2 to 32.88, 46.45, and 70.76 on DINOv3-S, and to 36.70, 51.53, and 75.28 on DINOv3-B, improving over Original, FiT3D, and SnD on every dataset. On DINOv3-7B, these scores rise from 39.49, 57.54, and 78.04 to 40.02, 58.43, and 78.69 over Original. On monocular depth, BridgeView reduces ScanNet++ RMSE from 0.3083 to 0.2952 on DINOv3-S and from 0.2864 to 0.2567 on DINOv3-B, and lowers RMSE on every dataset on DINOv3-7B over Original. SnD keeps the lower RMSE on ScanNet and NYUv2: 0.2906 and 0.4222 on DINOv3-S, and 0.2593 and 0.3668 on DINOv3-B. BridgeView remains competitive on those sets, at 0.2954 and 0.4225 on DINOv3-S and at 0.2636 and 0.3694 on DINOv3-B. These results show that BridgeView improves frozen semantic transfer, while remaining competitive with SnD on monocular depth.

4.2.2 Training Efficiency

As shown in Table 3, BridgeView uses 0.114 GPU seconds per view on matched NVIDIA H20 GPUs, giving $5.4\times$ and $10.3\times$ speedups over FiT3D and SnD, with peak memory 9.78 GiB. FiT3D fine-tunes the full encoder on one RGB image; SnD and BridgeView update LoRA, on a three-view context-to-target unit and a two-view pair respectively, so the table reports GPU-s/view. DINOv3-7B peak memory is in Appendix Table 8. These results show that the gap to SnD is from dropping scene reconstruction and rendering, not from a different adapter.

Table 3: Training cost on matched H20 GPUs.

Method	Adapter	Training unit	Views/unit	GPU-s/view ↓	Peak GiB ↓
FiT3D	Full encoder	Image	1	0.612	14.93
SnD	LoRA	Context to target	3	1.170	27.89
BridgeView	LoRA	Cross-view pair	2	**0.114**	**9.78**

Table 4: DINOv2-S ablation at two pair corpora.

Variant	Corpus	ScanNet++		ScanNet		NYUv2	
		mIoU ↑	RMSE ↓	mIoU ↑	RMSE ↓	mIoU ↑	RMSE ↓
CFPR	640K	32.32	0.3682	44.93	0.3150	69.64	0.4268
	3M	31.74	0.4301	42.79	0.3886	68.18	0.5219
+BIR	640K	32.45	0.3648	45.08	0.3135	69.90	0.4249
	3M	32.58	0.3649	44.85	0.3199	69.73	0.4319
+BIR+RGC	640K	**32.77**	0.3523	**45.42**	0.3109	**70.44**	0.4213
	3M	**32.90**	0.3546	**45.20**	0.3188	**70.24**	0.4312

Table 5: DINOv3-B ablation at two pair corpora.

Variant	Corpus	ScanNet++		ScanNet		NYUv2	
		mIoU ↑	RMSE ↓	mIoU ↑	RMSE ↓	mIoU ↑	RMSE ↓
CFPR	640K	36.19	0.2636	51.40	0.2615	75.28	0.3639
	3M	35.69	0.2884	50.36	0.2848	74.75	0.4012
+BIR	640K	36.28	0.2659	51.53	0.2609	75.39	0.3648
	3M	36.53	0.2603	51.43	0.2639	75.25	0.3685
+BIR+RGC	640K	**36.43**	0.2654	**51.56**	0.2599	**75.42**	0.3645
	3M	**36.70**	0.2567	**51.53**	0.2636	**75.28**	0.3694

4.2.3 Progressive Component Ablation

Component ablation. As shown in Tables 4 and 5, adding BIR and then RGC to CFPR raises frozen mIoU on ScanNet++, ScanNet, and NYUv2. At 3M, CFPR alone lowers frozen transfer relative to its 640K row. ScanNet mIoU falls from 44.93 to 42.79 on DINOv2-S and from 51.40 to 50.36 on DINOv3-B. Adding BIR reverses those drops, raising 3M ScanNet mIoU by 2.06 to 44.85 and by 1.07 to 51.43. Adding RGC keeps that recovery and adds a smaller increment, to 45.20 and 51.53. At 640K the same order holds with smaller steps. BIR raises ScanNet mIoU from 44.93 to 45.08 on DINOv2-S and from 51.40 to 51.53 on DINOv3-B. RGC then raises these scores to 45.42 and 51.56. The full method is best on every mIoU column at both budgets. These results show that BIR accounts for the 3M recovery, and RGC keeps the gain.

4.2.4 Consistency Across Views for Geometric Tasks

As shown in Table 6, BridgeView achieves the lowest rotation and translation error and the highest pose recall among Original, FiT3D, and SnD on withheld-view localization with DINOv3-B. Against SnD, rotation error falls from 4.40° to 2.66°, translation error from 0.1820 m to 0.0997 m, and pose recall rises from 59.80% to 76.13%. On TSDF map rendering, BridgeView also leads every compared method. Against FiT3D, the strongest TSDF baseline, MV δ_1, Accuracy@10cm, and Completeness@10cm rise from 55.22, 22.28, and 20.98 to 61.67, 28.40, and 26.67. These results show that the correspondence gain transfers to pose estimation and TSDF rendering on withheld views.

Table 6: DINOv3-B localization and TSDF rendering.

Method	Indoor localization			ScanNet++ TSDF map rendering		
	Med. Rot. Err. ↓	Med. Trans. Err. ↓	R@(1 m, 10°) ↑	MV δ_1 ↑	Accuracy@10cm ↑	Completeness@10cm ↑
Original	5.64°	0.2163 m	57.33	50.88	18.72	17.48
FiT3D	5.47°	0.2259 m	61.47	55.22	22.28	20.98
SnD	4.40°	0.1820 m	59.80	54.10	22.27	20.79
BridgeView	**2.66°**	**0.0997 m**	**76.13**	**61.67**	**28.40**	**26.67**

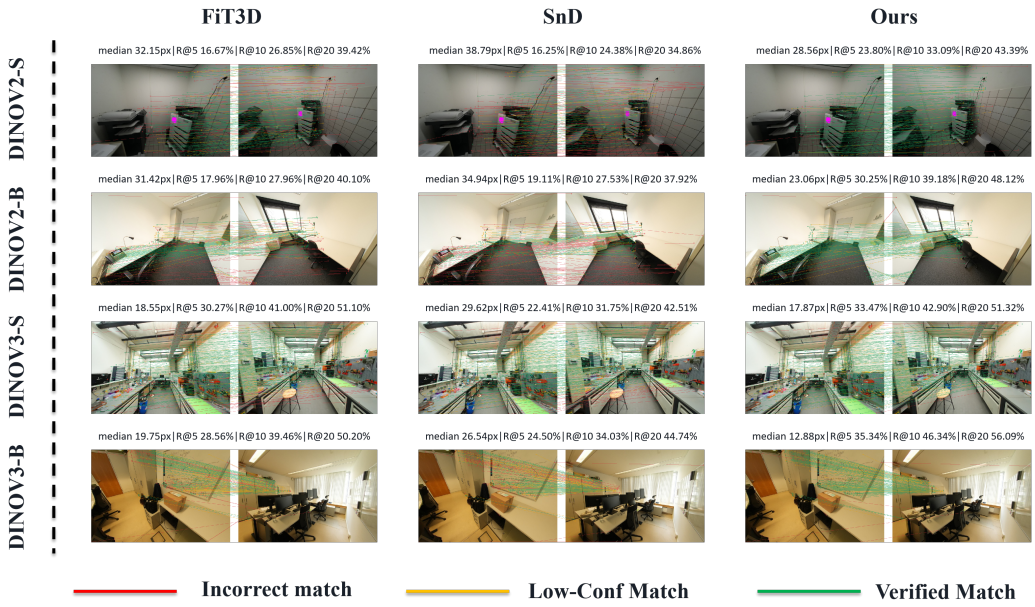

Figure 4: Correspondence matches across views for identical queries on the DINO family.

4.2.5 Qualitative Analysis

Figure 4 reveals qualitative differences, highlighting performance gaps between methods on identical source queries. Green lines denote matches within five pixels of the geometrically derived endpoint, and red lines denote larger errors. Across the displayed pairs, FiT3D and SnD retain more errors over long displacements. BridgeView produces denser correct matches under viewpoint and content changes. Appendix Figure 5 provides further qualitative results in a shared PCA space.

5 Conclusion

We presented BridgeView, which establishes cross-view consistency from frozen opposite-view targets. CFPR forms the teacher at the matched opposite pixel, BIR adds a bounded residual around this readout when neighboring instances occupy one encoder cell, and RGC coordinates the backward step inside one instance relation. This formulation ensures two observations of one surface share a descriptor in pretrained coordinates. Experiments on indoor RGB-D show that BridgeView raises cross-view correspondence and semantic segmentation on DINOv3, improves visual localization and TSDF map rendering over prior reconstruction-based methods, and trains over $10\times$ faster than SnD. Scaling the training corpus to 3M pairs yields additional correspondence gains. These results show that large vision foundation models can obtain cross-view consistency inside the pretrained coordinate system.

AI USE STATEMENT

OpenAI Codex was used solely for language polishing of this manuscript. The authors reviewed all edits and take responsibility for the final content.

REPRODUCIBILITY STATEMENT

Section 3 specifies the method and its optimization objective, while Section 4.1 defines the datasets, metrics, and evaluation setting. Appendix A.1 specifies the dataset splits and downstream evaluation protocols. Appendix A.2 specifies pair construction. Appendix A.3 records the training operating points. Appendix C writes the transport maps and residual definitions. Pair-corpus and update-budget controls are reported in Appendix B.3.

REFERENCES

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. ScanNet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

Johan Edstedt, Ioannis Athanasiadis, Mårten Wadenbäck, and Michael Felsberg. DKM: Dense kernelized feature matching for geometry estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Johan Edstedt, Qiyu Sun, Georg Bökman, Mårten Wadenbäck, and Michael Felsberg. RoMa: Robust dense feature matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Johan Edstedt, David Nordström, Yushan Zhang, Georg Bökman, Jonathan Astermark, Viktor Larsson, Anders Heyden, Fredrik Kahl, Mårten Wadenbäck, and Michael Felsberg. RoMa v2: Harder better faster denser feature matching. *arXiv preprint arXiv:2511.15706*, 2025.

Mohamed El Banani, Amit Raj, Kevis-Kokitsi Maninis, Abhishek Kar, Yuanzhen Li, Michael Rubinstein, Deqing Sun, Leonidas Guibas, Justin Johnson, and Varun Jampani. Probing the 3d awareness of visual foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), 2023.

Justin Kerr, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. LERF: Language embedded radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Sosuke Kobayashi, Eiichi Matsumoto, and Vincent Sitzmann. Decomposing NeRF for editing via feature field distillation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with MASt3R. In *European Conference on Computer Vision (ECCV)*, 2024.

Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. LightGlue: Local feature matching at light speed. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. In *European Conference on Computer Vision (ECCV)*, 2020.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021.

David Shavin and Sagie Benaim. Splat and distill: Augmenting teachers with feed-forward 3d reconstruction for 3d-aware distillation. In *International Conference on Learning Representations (ICLR)*, 2026.

Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from RGBD images. In *European Conference on Computer Vision (ECCV)*, 2012.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025.

Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. LoFTR: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Hugo Touvron, Matthieu Cord, and Hervé Jégou. DeiT III: Revenge of the ViT. In *European Conference on Computer Vision (ECCV)*, 2022.

Vadim Tschernezki, Iro Laina, Diane Larlus, and Andrea Vedaldi. Neural feature fusion fields: 3d distillation of self-supervised 2d image representations. In *International Conference on 3D Vision (3DV)*, 2022.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jérôme Revaud. DUSt3R: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Philippe Weinzaepfel, Vincent Leroy, Thomas Lucas, Romain Brégier, Yohann Cabon, Vaibhav Arora, Leonid Antsfeld, Boris Chidlovskii, Gabriela Csurka, and Jérôme Revaud. CroCo: Self-supervised pre-training for 3d vision tasks by cross-view completion. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Philippe Weinzaepfel, Thomas Lucas, Vincent Leroy, Yohann Cabon, Vaibhav Arora, Romain Brégier, Gabriela Csurka, Leonid Antsfeld, Boris Chidlovskii, and Jérôme Revaud. Croco v2: Improved cross-view completion pre-training for stereo matching and optical flow. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. ScanNet++: A high-fidelity dataset of 3d indoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Yuanwen Yue, Anurag Das, Francis Engelmann, Siyu Tang, and Jan Eric Lenssen. Improving 2d feature representations by 3d-aware fine-tuning. In *European Conference on Computer Vision (ECCV)*, 2024.

Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. Image BERT pre-training with online tokenizer. In *International Conference on Learning Representations (ICLR)*, 2022.

Shijie Zhou, Haoran Chang, Sicheng Jiang, Zhiwen Fan, Zehao Zhu, Dejia Xu, Pradyumna Chari, Suya You, Zhangyang Wang, and Achuta Kadambi. Feature 3DGS: Supercharging 3d gaussian splatting to enable distilled feature fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

A Implementation Details

This section records the evaluation protocols, the pair corpus, and the training operating points used throughout the paper. Additional tables and figures are collected in Section B. The transport maps, residuals, and residual scales are written in Section C.

A.1 Evaluation Protocols

All evaluations keep the enhanced image encoder frozen. Within each comparison, every feature variant uses the same image preprocessing, the same feature resolution, the same evaluation samples, and the same downstream head. Depth, camera poses, and instance maps are not encoder inputs at test time.

Semantic segmentation and monocular depth. Semantic segmentation trains a linear classifier on the frozen patch tokens and reports mIoU. Monocular depth trains the shared depth probe on the same tokens and reports RMSE. Each probe reads the patch grid of the corresponding backbone and does not update the encoder. We use the official training and validation or test splits of ScanNet++, ScanNet, and NYUv2. The split, the probe architecture, and the training budget of the probe are identical for Original, FiT3D, SnD, and BridgeView.

ScanNet correspondence. ScanNet correspondence retrieves, for each query patch, the patch in the paired view whose descriptor is nearest in cosine distance. It reports Median Error in pixels and Recall at 10 pixels (R@10). The evaluation set is fixed: $1{,}500$ view pairs and $1{,}000$ geometrically valid queries per pair. Every method uses quarter-resolution features and the same input-size rule, so differences come from the descriptor and not from the retrieval protocol.

Geometric matcher comparison. As shown in Table 2, the matcher comparison uses the same $1{,}500$ ScanNet pairs and the same query pixels as Table 1. Median error and R@10 of BridgeView are the DINOv3-B numbers of that table; R@5 and R@20 are the matching radii of the same retrieval. RoMa v2 (Edstedt et al., 2025) is scored with those same thresholds.

A.1.1 Downstream Geometry Protocol Details

Localization. Indoor localization forms 2D–3D associations from dense feature matches and evaluates every method with the same pose-estimation pipeline. Valid depth lifts a matched pixel to a 3D point. The estimator then recovers a camera pose and reports rotation error, translation error, and pose recall against the reference poses. An estimate that fails to return a pose counts as a failure, so recall is computed over the full query set.

TSDF map rendering. TSDF rendering uses DINOv3-B features on held-out ScanNet++ scenes. All methods share the same mapping frames, withheld views, camera calibration, depth-prediction protocol, and TSDF integration settings. Withheld RGB frames are excluded from mapping. Rendered depth is compared with reference depth using MV δ_1, Accuracy@10cm, and Completeness@10cm.

A.2 Pair Corpus Construction

This section specifies the source, size, sampling rule, and deduplication of the pair corpora used to train BridgeView.

We form pairs from calibrated RGB-D views in the official ScanNet++ training split (Yeshwanth et al., 2023). That split contains 230 scenes. It is disjoint from the official ScanNet++, ScanNet, and NYUv2 evaluation splits, and from the fixed ScanNet correspondence set.

Let s and t be two frames in one scene, ordered by capture sequence. A pair (s, t) enters the pool if and only if the frame index of t differs from that of s by at least $\mathrm{stride} = 1$, and the depth-projected overlap is at least 0.10. Overlap is the fraction of valid depth in one view that projects into the other. It is measured on a regular grid with spatial stride 32. Overlap is then assigned to one of three bins:

Table 7: Operating points for CFPR, BIR, and RGC.

Module	Quantity	Setting
CFPR	τ_{3D}, τ_{rep}, τ_d	0.05 m, 2 px, 0.01 m
CFPR	stride / pair quota	4 / 12 × 16 grid, 256 per cell, 8192 per pair
BIR	τ_ρ, $n_{\min}$, $s_{\min}$	0.75, 8, 0.50
BIR	α, β	0.2, 0.05
BIR	cone ϕ	12° ($\cos\phi \approx 0.978$)
BIR	neighborhood	two tokens, at least four outside the instance
RGC	λ	0.50 (equal mix)
RGC	mix rule	$\lvert G_i\rvert \geq 2$, else pointwise

low $[0.10, 0.30)$, mid $[0.30, 0.60)$, and high $[0.60, 1.00]$. The pool keeps these bins in balance across the 230 scenes, so a scene does not dominate one overlap band.

The identity of a pair is the undirected key $(\text{scene}, \min(s,t), \max(s,t))$. Self-pairs and reversed copies of the same two frames are removed. The resulting 3M corpus contains 3,000,000 unique pairs, or 13,043 pairs per scene on average. A scene contributes at most its eligible pair count. The sampler does not oversample a scene that has fewer eligible pairs.

The 640K corpus is a 640,000-pair subset of this 3M pool. It is not sampled independently from the raw frames. Both corpora use the same 230 scenes, the same overlap threshold, and the same undirected identity. Training reads the 3M corpus once under a uniform sampler, for 46,875 updates at a global batch of 64. Training reads the 640K subset once, for 10,000 updates at the same batch size.

A.3 Training Protocol

Within each matched enhancement comparison, the methods use the same backbone initialization, training split, image inputs, preprocessing, and optimizer update budget. Only the supervision route and its training unit differ. SnD and BridgeView update LoRA adapters on the student encoder. FiT3D fine-tunes the full encoder. The teacher encoder remains frozen.

Optimization and training scale. The student is adapted with LoRA of rank 8 and scaling 8 on every linear layer in the last eight transformer blocks. Optimization uses AdamW at 2×10^{-5} with weight decay 10^{-4}, BF16, and a global batch of 64 on eight GPUs. The learning rate is constant. The teacher remains frozen. All methods use the same optimizer family, the same numerical precision, and matched update budgets within a comparison. Section A.2 specifies the 640K and 3M pair corpora used under those budgets. As shown in Table 7, CFPR, BIR, and RGC use the listed operating points throughout.

Training and inference inputs. Camera intrinsics, poses, depths, and the SAM instance map are used only to construct training targets. The instance map is produced offline by SAM (Kirillov et al., 2023) on each training view. At inference, the enhanced encoder processes a single RGB image and requires none of these inputs.

Efficiency measurement. Training time and memory are measured on matched NVIDIA H20 hardware. GPU seconds per view divides the elapsed time of one training unit by the number of RGB views in that unit. The timed window includes the forward pass, the backward pass, and target construction for that unit. Peak memory is the maximum allocated GPU memory recorded under the same protocol for every method.

B Additional Experimental Results

This section reports DINOv3-7B training memory, the target audit, the consolidated dense metrics, the pair-budget controls, and the shared-space PCA figure.

Table 8: DINOv3-7B peak training memory.

Method	Backbone	Peak GiB ↓
SnD (same LoRA)	DINOv3-7B	OOM (> 95.22)
BridgeView (LoRA)	DINOv3-7B	**76.2**

Table 9: Mean frozen-probe mIoU on ScanNet++, ScanNet, and NYUv2.

Backbone	Original	FiT3D	SnD	BridgeView
DINOv3-S	49.63	49.19	49.21	**50.03**
DINOv3-B	54.00	53.01	53.66	**54.50**

As shown in Table 8, SnD with the same LoRA setting exceeds memory on DINOv3-7B, above 95.22 GiB, while BridgeView completes training at 76.2 GiB. These results show that the pair-based teacher remains trainable at the backbone scale where the reconstructed-scene teacher does not.

B.1 Target Audit

As shown in Table 9, BridgeView raises the mean frozen-probe mIoU over Original, FiT3D, and SnD on both DINOv3-S and DINOv3-B, using the same linear probes as the main tables. Against Original, the mean rises from 49.63 to 50.03 on DINOv3-S and from 54.00 to 54.50 on DINOv3-B. FiT3D and SnD lower this mean to 49.19 and 49.21 on DINOv3-S, and to 53.01 and 53.66 on DINOv3-B. These results show that the frozen opposite-view teacher raises the mean semantic score that a rendered-scene teacher reduces.

As shown in Table 10, a Gaussian teacher on paired ScanNet++ views with frozen DINOv3-B raises shared-instance cosine from 0.6869 to 0.7134, but lowers pretrained-coordinate cosine from 1.0000 to 0.8905 and semantic probe agreement from 1.0000 to 0.6109. The direct opposite-view descriptor keeps both pretrained-coordinate cosine and probe agreement at 1.0000. These results show that matching the rendered teacher leaves pretrained coordinates and probe agreement, so BridgeView reads the frozen opposite-view descriptor.

B.2 Dense Transfer and Cross-View Correspondence

As shown in Table 11, BridgeView has the lowest Median Error and the highest R@10 in every backbone group that reports correspondence. Against SnD, Median Error falls from 38.79 to 28.56 pixels on DINOv2-S and from 34.94 to 23.06 pixels on DINOv2-B, while R@10 rises from 24.38% to 33.09% and from 27.53% to 39.18%. The DINOv3 correspondence and mIoU columns match Table 1. On DINOv3-7B, Median Error falls from 15.56 to 5.68 pixels and R@10 rises from 45.43% to 58.21% over Original. As shown in Table 8, SnD with the same LoRA setting exceeds memory at that scale. On DINOv2-S, SnD keeps the higher ScanNet and NYUv2 mIoU, 46.24 and 70.82 versus 45.20 and 70.24. On DINOv2-B, SnD keeps the higher ScanNet mIoU, 51.15 versus 50.95, and it keeps the lower ScanNet and NYUv2 RMSE on both DINOv2 backbones. On DINOv3, BridgeView leads every mIoU column and remains competitive with SnD on monocular depth. These results show that a frozen opposite-view teacher helps most when the frozen space is already strong. On DINOv2, the weak frozen space limits the semantic gain while the correspondence gain remains large. On DINOv3, the same target improves both.

B.3 Pair Corpus and Exposure Scaling

As shown in Table 12, BridgeView outperforms SnD under every pair corpus and update budget on DINOv3-B. Section A.2 specifies that the 640K set is a subset of the 3M pool. The comparison matches optimizer updates, because the two methods consume different numbers of views per training unit. At 10K updates, the 640K model already exceeds SnD, raising R@10 from 34.18% to 37.83% and lowering Median Error from 26.59 to 26.08 pixels. At the same budget, the 3M corpus raises R@10 to 44.53% and lowers Median Error to 15.16 pixels. A longer SnD schedule

Table 10: Frozen opposite-view target versus a Gaussian teacher.

Metric	Direct descriptor	Gaussian target
Pretrained coordinate cosine	**1.0000**	0.8905
Shared instance cosine	0.6869	0.7134
Semantic probe prediction agreement	**1.0000**	0.6109

Table 11: Correspondence and frozen transfer across backbones.

		ScanNet Corr.		ScanNet++		ScanNet		NYUv2	
Backbone	Method	Med. Err. ↓	R@10 ↑	mIoU ↑	RMSE ↓	mIoU ↑	RMSE ↓	mIoU ↑	RMSE ↓
DINOv2-S	Original	55.52	14.07	30.34	0.4110	43.52	0.3425	67.76	0.4613
	FiT3D	32.15	26.85	30.31	0.3641	38.04	0.3811	64.32	0.5642
	SnD	38.79	24.38	32.46	0.3593	**46.24**	**0.3069**	**70.82**	**0.4143**
	BridgeView	**28.56**	**33.09**	**32.90**	**0.3546**	45.20	0.3188	70.24	0.4312
DINOv2-B	Original	53.69	15.90	32.79	0.3429	47.43	0.3048	71.47	0.4175
	FiT3D	31.42	27.96	34.24	0.3553	45.17	0.3368	70.59	0.5068
	SnD	34.94	27.53	35.38	0.3014	**51.15**	**0.2689**	74.68	**0.3691**
	BridgeView	**23.06**	**39.18**	**35.42**	**0.2996**	50.95	0.2737	**74.82**	0.3838
DINOv3-S	Original	26.80	37.50	32.37	0.3083	45.86	0.2939	70.65	**0.4137**
	FiT3D	18.55	41.00	32.43	0.2985	45.59	0.2964	69.55	0.4504
	SnD	29.62	31.75	31.67	0.3144	45.79	**0.2906**	70.16	0.4222
	BridgeView	**17.87**	**42.90**	**32.88**	**0.2952**	**46.45**	0.2954	**70.76**	0.4225
DINOv3-B	Original	33.75	35.32	35.90	0.2864	51.20	0.2684	74.89	0.3765
	FiT3D	19.75	39.46	35.18	0.2674	49.70	0.2721	74.14	0.4116
	SnD	26.54	34.03	35.19	0.2941	50.67	**0.2593**	75.13	**0.3668**
	BridgeView	**12.88**	**46.34**	**36.70**	**0.2567**	**51.53**	0.2636	**75.28**	0.3694
DINOv3-7B	Original	15.56	45.43	39.49	0.2876	57.54	0.2104	78.04	0.3070
	BridgeView	**5.68**	**58.21**	**40.02**	**0.2753**	**58.43**	**0.2093**	**78.69**	**0.2972**

does not improve correspondence: R@10 moves from 34.18% to 34.03%, and Median Error from 26.59 to 26.54 pixels. BridgeView improves with longer training on both corpora, from 37.83% to 39.38% on 640K and from 44.53% to 46.34% on 3M, with Median Error falling to 22.77 and 12.88 pixels. These results show that SnD saturates under a longer schedule, while BridgeView continues to improve.

B.4 Shared Feature Space PCA Visualization

As shown in Figure 5, descriptors from paired views are projected into one PCA space fitted jointly across the four methods and both views, so a color change is a change in the shared coordinates. The original features break corresponding surfaces into fragments and change color across views. A Gaussian-mediated teacher reduces part of that variation and keeps discontinuities near object boundaries. BridgeView produces more continuous colors on the same surface across views and keeps transitions between adjacent regions. These visualizations show that BridgeView improves cross-view consistency without smoothing away local structure, whereas a Gaussian-mediated teacher trades local structure for scene-level smoothness. This is consistent with the correspondence and semantic gains in the main tables.

C Method Details

This section writes the maps that Section 3 names and leaves to the appendix: the three projections, the five residuals, reciprocal support, and the cone map. The identity equations of CFPR, BIR, and RGC remain in the main text.

Table 12: Pair corpus and update budget on DINOv3-B.

Method	Configuration	Updates	Med. Err. ↓	R@10 ↑
SnD	Scene triplets	10K	26.59	34.18
SnD	Scene triplets	46,875	26.54	34.03
BridgeView	640K pairs	10K	26.08	37.83
BridgeView	640K pairs	46,875	22.77	39.38
BridgeView	3M pairs	10K	15.16	44.53
BridgeView	3M pairs	46,875	**12.88**	**46.34**

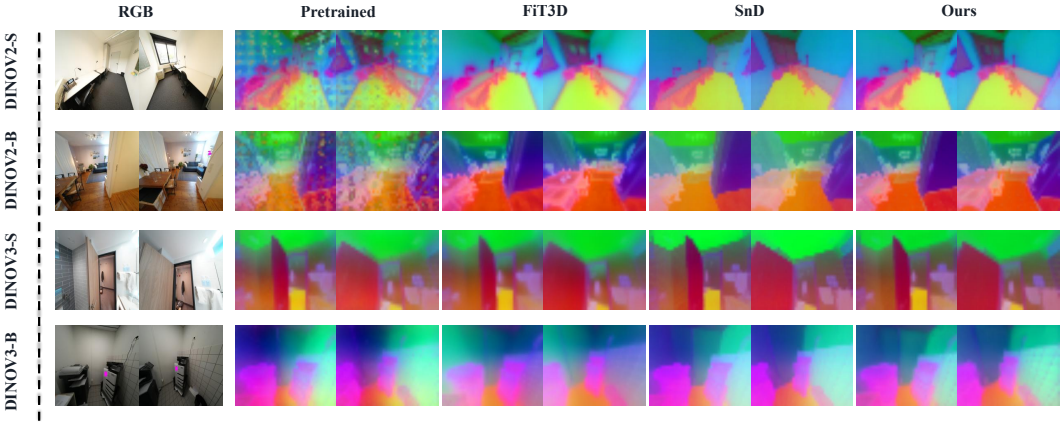

Figure 5: Shared space PCA visualization of dense features from paired views across four backbones and four feature variants.

C.1 Certified Frozen Pair Readout

Let $\boldsymbol{u} \in \mathbb{R}^2$ be a pixel in view a with measured depth $D^a(\boldsymbol{u}) \in \mathbb{R}_{>0}$, let $K^a, K^b \in \mathbb{R}^{3\times3}$ be the camera intrinsics, and let $T^a, T^b \in \mathrm{SE}(3)$ be the camera-to-world poses. Let $\widetilde{\boldsymbol{u}} = (u_x, u_y, 1)^\top$ denote the homogeneous pixel. Lifting $\boldsymbol{u}$ to a world point and projecting into view b gives

$$\boldsymbol{X}_a(\boldsymbol{u}) = T^a\big[D^a(\boldsymbol{u})\,(K^a)^{-1}\widetilde{\boldsymbol{u}};\,1\big], \qquad (\boldsymbol{v}, \widehat{d}_b) = \pi\big(K^b(T^b)^{-1}\boldsymbol{X}_a(\boldsymbol{u})\big). \tag{11}$$

Here $\boldsymbol{X}_a(\boldsymbol{u}) \in \mathbb{R}^3$ is the back-projected point in metres, π performs perspective division, $\widehat{d}_b \in \mathbb{R}_{>0}$ is the predicted depth of that landing in metres, and $\boldsymbol{v} \in \mathbb{R}^2$ is the landing pixel in view b. The reverse map lifts $\boldsymbol{v}$ with $D^b(\boldsymbol{v})$ and projects back into view a,

$$\boldsymbol{X}_b(\boldsymbol{v}) = T^b\big[D^b(\boldsymbol{v})\,(K^b)^{-1}\widetilde{\boldsymbol{v}};\,1\big], \qquad (\boldsymbol{u}', \widehat{d}_a) = \pi\big(K^a(T^a)^{-1}\boldsymbol{X}_b(\boldsymbol{v})\big). \tag{12}$$

Here $\widetilde{\boldsymbol{v}}$ is the homogeneous form of $\boldsymbol{v}$, $\boldsymbol{X}_b(\boldsymbol{v}) \in \mathbb{R}^3$ is the world point from view b, $\boldsymbol{u}' \in \mathbb{R}^2$ is the return pixel in view a, and $\widehat{d}_a \in \mathbb{R}_{>0}$ is the predicted depth of that return in metres. Closing the trip applies Eq. equation 11 to $\boldsymbol{u}'$,

$$(\boldsymbol{v}', \widehat{d}'_b) = \pi\big(K^b(T^b)^{-1}\boldsymbol{X}_a(\boldsymbol{u}')\big). \tag{13}$$

where $\boldsymbol{v}' \in \mathbb{R}^2$ is the return landing in view b.

The one-way candidate fails when the two cameras do not both observe the same point. The residuals used by the CFPR indicator χ_{ab} are

$$\begin{gathered} e_{3D} = \|\boldsymbol{X}_a(\boldsymbol{u}) - \boldsymbol{X}_b(\boldsymbol{v})\|_2, \quad e_a = \|\boldsymbol{u} - \boldsymbol{u}'\|_2, \quad e_b = \|\boldsymbol{v} - \boldsymbol{v}'\|_2, \\ e_d^a = |D^a(\boldsymbol{u}) - \widehat{d}_a|, \quad e_d^b = |D^b(\boldsymbol{v}) - \widehat{d}_b|. \end{gathered} \tag{14}$$

Here e_{3D} is 3D agreement in metres, e_a and e_b are round-trip reprojection in pixels, and e_d^a and e_d^b are depth agreement in metres. The gate $\chi_{ab}(\boldsymbol{u})$ is the product of validity and the three residual tests. As shown in Table 7, the scalars τ_{3D}, τ_{rep}, and τ_d take the listed values. This construction makes $\chi_{ab}(\boldsymbol{u}) = 1$ if and only if both views observe the same surface at the tested pixel.

C.2 Bounded Instance Residual

For endpoint identities $r = m(\boldsymbol{u}_i^a)$ and $s = m(\boldsymbol{u}_i^b)$, let n_{rs} be the number of pairs in $\mathcal{P}_{ab}$ with those identities, n_r^a the number of certified endpoints in view a with identity r, and n_s^b the corresponding count in view b. Reciprocal support is the smaller of the two directed pair-fractions,

$$s_{rs} = \min\Big(\frac{n_{rs}}{n_r^a}, \frac{n_{rs}}{n_s^b}\Big). \tag{15}$$

Thus $s_{rs} = 1$ if and only if every certified endpoint of r in a and of s in b is paired with the other. The pairing is used when $n_{rs} \geq n_{\min}$ and $s_{rs} \geq s_{\min}$.

Let μ_r^a be the ℓ_2-normalized mean of frozen readouts over cells in view a that satisfy $\rho \geq \tau_\rho$ and have dominant identity r, and define μ_s^b analogously in view b. The cross-view prototype is $\mu_i = \mathrm{norm}(\mu_r^a + \mu_s^b)$, with $\mathrm{norm}(x) = x/\|x\|_2$. Let $\mathcal{N}_i^v$ be the encoder cells around the certified endpoint in view v whose dominant identity is not $m(\boldsymbol{u}_i^v)$, and let b_i^v be the mean frozen readout on that neighborhood minus p_i^v. The neighborhood uses a two-token radius and requires at least four outside-instance cells.

Let $\angle(x, p) = \arccos(\langle x, p\rangle)$ for unit vectors x and p. The cone around a frozen readout p is the spherical cap $\{x : \angle(x, p) \leq \phi\}$ in feature space. Write $\widehat{x} = x/\|x\|_2$ and let $v = \widehat{x} - \langle \widehat{x}, p\rangle p$ be the component of $\widehat{x}$ orthogonal to p. The map cap_ϕ is

$$\mathrm{cap}_\phi(x, p) = \begin{cases} \widehat{x}, & \angle(\widehat{x}, p) \leq \phi, \\ \cos\phi\, p + \sin\phi\, \dfrac{v}{\|v\|_2}, & \text{otherwise.} \end{cases} \tag{16}$$

so the refined target stays on that cap. Semantic class labels are not used. As shown in Table 7, the scales α, β and the cutoffs $n_{\min}$, $s_{\min}$, and τ_ρ take the listed values.

C.3 Relation Gradient Coordination

Relation gradient coordination (RGC) groups endpoints that share one image pair, one student view, and one identity pair (r, s). Write G_i for that set. Thus $j \in G_i$ if and only if j and i share that pair, view, and identity pair, and $|G_i| \geq 2$ if and only if the relation is observed more than once. The forward pass keeps the pointwise gap ℓ_i. The backward pass writes the mixed gradient $\widetilde{\boldsymbol{g}}_i = (1 - \lambda)\boldsymbol{g}_i + \lambda \bar{\boldsymbol{g}}_{G_i}$ when $|G_i| \geq 2$, and writes $\boldsymbol{g}_i$ otherwise. The mix uses $\lambda = 0.50$.

C.4 Relation to Geometry-Oriented Models

Cross-view methods recover a relation between two views of one surface. CroCo reconstructs a masked view with another view of the same scene (Weinzaepfel et al., 2022). CroCo v2 scales that completion setting to large real image pairs and dense geometric transfer (Weinzaepfel et al., 2023). DUSt3R predicts dense pointmaps, MASt3R adds matching-oriented local features, and VGGT predicts cameras, depths, pointmaps, and tracks (Wang et al., 2024; Leroy et al., 2024; Wang et al., 2025). The product of this line is a matcher or a geometric model. The encoder that those methods train is the geometry predictor itself, and the gain is read out as correspondences, cameras, or pointmaps.

VFM enhancement starts from a DINO-family encoder that already organizes dense features around semantic and appearance cues. FiT3D and SnD add cross-view consistency by supervising that encoder with a teacher rendered from a reconstructed scene (Yue et al., 2024; Shavin & Benaim, 2026). BridgeView has the same enhancement objective. Calibrated RGB-D observations first verify that two pixels observe the same surface, and the frozen descriptor at one endpoint supervises the other. The view pair is a training source. The exported backbone reads a single RGB image.