MARCO: Navigating the Unseen Space of Semantic Correspondence

Supplementary Material

In this appendix, we provide additional insights into MARCO, together with further experimental analyses. Specifically:

- **General applicability of our approach**. In Sec. A, we show that our dense self-distillation loss can produce coherent and generalizable representations when applied to previous state-of-the-art methods.
- **Results with pre-training**. In Sec. B, we present a variant of our model pre-trained on the AP-10K dataset, following recent approaches [32, 54].
- **Additional experiments**. Section C provides further analyses of MARCO, including studies on adapter placement and dimensionality, comparisons with full fine-tuning, a progressive ablation of the proposed components, and analyses of pseudo-label coverage, pseudo-label noise, and hyperparameter robustness.
- **Use of object masks**. In Sec. D, we discuss the role of object masks in our pipeline, show that they can be removed with negligible impact on results, and compare the supervision assumptions of different methods.
- **Compute details**. In Sec. E, we report model size and inference speed, and describe the protocol used to measure computational performance.
- **Pseudo-code of dense self-distillation**. In Sec. F, we summarize the proposed flow-anchoring procedure and self-distillation loss in algorithmic form.
- **Details on MP-100**. In Sec. G, we provide details of the proposed benchmark, including the curation protocol, evaluation splits, and extended quantitative results.
- **Qualitative examples**. In Sec. H, we present additional qualitative results for MARCO.

A. Broad Applicability of our Dense Self-Distillation via Flow Anchoring

A central motivation behind MARCO is to enhance the *geometric coherence* of learned feature representations, producing features that vary smoothly across the object surface to generalize to *unseen keypoints and categories*, despite the sparsity of landmark supervision. Our dense self-distillation via flow-anchoring is designed for this purpose: by mining reliable correspondences emerging in the feature space and propagating them across the object via piecewise-affine interpolation, the training objective encourages the backbone to maintain smooth, consistent geometry beyond the annotated regions. We raise the question of *whether our self-distillation strategy is tied to the MARCO architecture, or whether it constitutes a general training principle that can benefit other correspondence pipelines*. To assess its generality, we apply our dense self-distillation loss to the state-of-the-art Geo-SC model [54]. We compare three variants: *(i)* the original Geo-SC, *(ii)* Jamais Vu [32], which augments Geo-SC with a loss mapping object points to a learned canonical manifold, and *(iii)* Geo-SC augmented with our flow-anchoring loss. Because all variants share the same backbone, inference pipeline, and base loss, this experiment isolates the contribution of the auxiliary loss.

Table 5. **General applicability of our self-distillation loss.** By training the state-of-the-art Geo-SC model with our dense self-distillation objective, we markedly improve its generalization to unseen keypoints on SPair-U. In MARCO, we integrate this objective with a coarse-to-fine supervised loss, reaching state-of-the-art results while being smaller and faster. Per-image PCK (in %, ↑) on SPair-71k, and SPair-U (unseen keypoints); models trained on SPair-71k. Best results **bold**, 2nd best underlined.

		SPair-71k			SPair-U		
Method	**Encoders**	0.01	0.05	0.10	0.01	0.05	0.10
Jamais Vu [32]	SD+DINO	20.5	71.9	82.5	4.2	37.8	62.4
Geo-SC [54]	SD+DINO	21.7	72.8	83.2	3.9	35.4	56.9
↳ + our dense self-distillation loss	SD+DINO	20.8	73.0	83.6	4.2	38.1	63.4
MARCO *(ours)*	DINOv2	**27.0**	**77.6**	**87.2**	**4.7**	**41.7**	**67.5**

Effect of adding our loss to Geo-SC. As shown in Tab. 5, adding our flow-anchoring loss consistently improves Geo-SC. On SPair-U, which evaluates generalization to unseen keypoints, Geo-SC improves from 56.9 to 63.4 PCK@0.10, outperforming Jamais Vu, which requires monocular depth prediction from an external model for lifting keypoints in 3D. Importantly, our loss does not degrade in-domain accuracy: on SPair-71k, Geo-SC increases from 83.2 to 83.6 PCK@0.10, whereas Jamais Vu slightly impairs results (82.5). Despite these improvements, Geo-SC enhanced with our loss remains below the accuracy levels of MARCO, which achieves 87.2 PCK@0.10 on SPair-71k and 67.5 PCK@0.10 on SPair-U. This gap highlights the importance of coupling flow anchoring with the full MARCO framework: coarse-to-fine supervision, adapter-based feature enrichment, and efficient sub-patch refinement, which together strengthen spatial fidelity in DINOv2. Finally, MARCO attains these improvements with a *single* DINOv2 encoder (323M parameters), compared to the 950M-parameter SD+DINO dual-encoder used by Geo-SC and Jamais Vu, being approximately 10× faster.

B. Pre-training on AP-10k

Recent correspondence works have increasingly adopted an additional pre-training stage on the AP-10K dataset. This strategy was first introduced by Geo-SC [54], which also proposed the AP-10K benchmark itself. Subsequent methods, including Jamais Vu [32], followed this practice and incorporated AP-10K pre-training into their pipelines, making it a common component of recent evaluation protocols. To ensure a fair and direct comparison with these approaches, we therefore train a variant of MARCO that includes the same AP-10K pre-training stage. The corresponding results are reported in Tab. 6. Like prior works, our method benefits from pre-training. Notably, *(i)* simply adding more data does not aid generalization on SPair-U (*cf*. GECO, Geo-SC), and *(ii)* our SPair-only model outperforms previous baselines pre-trained on AP-10k.

Table 6. **Impact of pre-training.** Following recent works, we show a variant of our model pretrained on AP-10k and compare to other works in the same setting. † indicates a model jointly trained on SPair-71k and AP-10k, rather than pre-trained and then fine-tuned. Per-image PCK (in %, ↑) on SPair-71k, and SPair-U (unseen keypoints). Best results **bold**, 2nd best underlined

Method	Encoders	SPair-71k			SPair-U		
		0.01	0.05	0.10	0.01	0.05	0.10
Train on SPair-71k							
Jamais Vu [32]	SD+DINO	20.5	71.9	82.5	4.2	37.8	62.4
Geo-SC [54]	SD+DINO	21.7	72.8	83.2	3.9	35.4	56.9
MARCO *(ours)*	DINOv2	27.0	77.6	87.2	4.7	41.7	67.5
w/ Pre-train on AP-10k							
GECO [14] †	DINOv2	14.2	59.6	73.6	3.2	32.1	55.2
Jamais Vu [32]	SD+DINO	20.9	73.1	85.4	4.2	37.8	66.1
Geo-SC [54]	SD+DINO	22.0	75.3	85.6	4.1	35.5	57.1
MARCO *(ours)*	DINOv2	**28.5**	**78.3**	**87.3**	**5.0**	**44.2**	**69.7**

C. Additional Ablations

Adapter design study. We first analyze the architectural design choices underlying MARCO. Tab. 7 reports controlled ablations studying *(i)* different fine-tuning strategies, *(ii)* alternative parameter-efficient adaptation mechanisms, *(iii)* adapter placement across the transformer depth, and *(iv)* the dimensionality of the adapter bottleneck. Experiments are conducted using the same training schedule and evaluated on both SPair-71k and SPair-U. Three consistent trends emerge. First, fully fine-tuning the DINOv2 backbone harms generalization. While full fine-tuning increases SPair-71k accuracy to 67.0 PCK@0.10, the accuracy on SPair-U drops sharply to 43.9, compared to 54.9 when the backbone is kept frozen. This confirms that the pre-trained representation should remain largely frozen to preserve its semantic structure and generalization ability. A lighter strategy that fine-tunes only the QKV projections performs better (81.1 PCK@0.10 on SPair-71k and 59.6 on SPair-U), but still underperforms parameter-efficient adaptation. Second, among parameter-efficient strategies, AdaptFormer provides the best trade-off between in-domain accuracy and generalization. Classical Adapters [56] achieve 86.9 PCK@0.10 on SPair-71k and 65.7 on SPair-U, while LoRA [17] obtains slightly lower results (85.2 and 65.6 PCK@0.10, respectively). AdaptFormer [4] improves this balance, reaching 87.2 PCK@0.10 on SPair-71k and 67.5 on SPair-U. Third, the placement of adapters across the transformer depth plays an important role. Adaptation is most effective when applied to the upper transformer blocks (Layers 12–24), where high-level semantic features are formed. Applying adapters earlier in the network (*e.g.*, Layers 3–24) slightly reduces accuracy, while restricting adaptation to only the last few blocks also degrades accuracy. Finally, the dimensionality of the adapter bottleneck controls the balance between model capacity and regularization. A mid-sized bottleneck ($\times 0.5$, used in MARCO) achieves the best overall trade-off, while both larger ($\times 1$) and smaller ($\times 0.3$ or $\times 0.1$) bottlenecks slightly reduce accuracy.

Contribution of architectural and training components. Table 8 complements the ablation analysis in the main paper (Tab. 4). While the ablation in the main paper evaluates each component independently, this table presents a *progressive ablation* where the architectural and training components of MARCO are introduced sequentially. Starting from a frozen DINOv2 backbone, adding the proposed adapters and feature upsampling and training with the objective of Geo-SC [54], *i.e.* InfoNCE+ℓ_2, improves accuracy from 6.3 to 20.0 PCK@0.01 and from 53.9 to 78.9 PCK@0.10 on SPair-71k, while relying on a *single* DINOv2 encoder instead of the DINOv2+Stable Diffusion pipeline used by prior work. Replacing this objective with the proposed coarse-to-fine supervision further improves localization precision to 26.8 PCK@0.01 and 85.6 PCK@0.10. Finally, adding dense self-distillation substantially improves generalization to unseen keypoints, increasing SPair-U accuracy from 42.0 to 67.5 PCK@0.10, while also improving SPair-71k to 87.2 PCK@0.10.

Table 7. **Additional ablations** comparing fine-tuning *vs*. adaptation strategies, adapter placement, and bottleneck dimension. All results reported as PCK@0.10 (in %, ↑).

Setting	SPair-71k	SPair-U
DINOv2 *frozen*	53.9	54.9
Fine-tuning strategies		
Full FT	67.0	43.9
QKV-only FT	81.1	59.6
Adapter type		
Adapter [56]	86.9	65.7
LoRA [17]	85.2	65.6
AdaptFormer [4] (**ours**)	87.2	67.5
Adapter placement (AdaptFormer)		
Layers 3–24	85.5	64.5
Layers 6–24	85.6	65.3
Layers 9–24	86.2	66.0
Layers 12–24	86.7	66.7
Layers 15–24 (**ours**)	87.2	67.5
Layers 18–24	87.3	65.3
Layers 21–24	84.9	63.8
Adapter bottleneck size (AdaptFormer, Layers 12–24)		
Bottleneck $\times$ 1	86.2	67.1
Bottleneck $\times$ 0.5 (**ours**)	87.2	67.5
Bottleneck $\times$ 0.3	86.8	66.5
Bottleneck $\times$ 0.1	83.7	65.7

Pseudo-label coverage and noise. The dense self-distillation objective propagates correspondences across the object surface, extending supervision beyond the sparse annotated keypoints. As shown in Tab. 4 in the main paper, enabling dense self-distillation already increases accuracy on unseen keypoints from 41.8 to 64.7 PCK@0.10, producing pseudo-labels that densely cover the object surface, *i.e.* about 17k correspondences per object on average in SPair-71k. However, our goal is to maximize pseudo-label quality rather than raw coverage. Anchoring clusters using the GT keypoints improves accuracy to 67.5 PCK@0.10, while retaining an average coverage of about 13k correspondences. To estimate the quality of the pseudo-labels, we measure their sensitivity to noise by injecting Gaussian perturbations into the pseudo-label coordinates. As reported in Tab. 9, accuracy remains stable for perturbations up to $\sigma = 5$ pixels and only begins to degrade around $\sigma = 10$ pixels, suggesting that the intrinsic noise of the pseudo-labels re-

Table 8. **Ablation of architectural and training components.** Per-image PCK (in %, ↑) on SPair-71k (seen keypoints) and SPair-U (unseen keypoints). Adapters and feature upsampling are added to a frozen DINOv2 backbone, while training objectives progressively include standard supervision (InfoNCE + ℓ_2), the proposed coarse-to-fine loss, and dense self-distillation.

Architecture	Training objective			SPair-71k (seen keypoints)		SPair-U (unseen keypoints)	
	InfoNCE + ℓ_2 [54]	Coarse-to-fine	Dense Loss	PCK@0.01	PCK@0.10	PCK@0.01	PCK@0.10
DINOv2 (frozen)	✗	✗	✗	6.3	53.9	3.3	54.9
Adapter + Upsample	✓	✗	✗	20.0	78.9	1.9	39.7
	✗	✓	✗	26.8	85.6	2.1	42.0
	✗	✓	✓	**27.0**	**87.2**	**4.7**	**67.5**

mains below roughly 10 pixels.

Hyperparameters. Our solution is hyperparameter-free. In the flow-anchoring stage, flow vectors are grouped into k clusters to identify regions with coherent motion. While this step could require selecting the number of clusters, we instead initialize clustering with a large value (*e.g.*, $k = 15$) and use the Bayesian Information Criterion (BIC) to merge clusters with statistically consistent motion patterns. As shown in Tab. 10, this over-segmentation followed by BIC-based merging automatically determines a suitable number of clusters, avoiding the need to tune k.

D. Use of Object Masks

Recent works in semantic correspondence frequently leverage instance masks to concentrate feature matching on foreground regions. In unsupervised settings, masks are used to suppress background responses when mining feature matches [7, 10, 31]. Supervised methods also rely on masks, either to map object pixels to canonical 3D templates [32] or to restrict optimal-transport domains and sampling to the foreground region [14]. Similarly, Geo-SC performs mask-based pose alignment during inference through pose augmentation [54]. Masks are obtained by prompting SAM [22] with annotated keypoints. In MARCO, we follow this common practice: importantly, masks are used *only* as a weak spatial prior during pseudo-label generation, not as a direct supervision signal. Specifically, masks enter our pipeline in a single step:

> After extracting mutual nearest-neighbor (MNN) matches, we restrict candidate locations to pixels inside the object mask. The MNN matches are used to construct the Delaunay triangulation for flow estimation.

Their role is, therefore, identical to prior SOTA methods, acting purely as a spatial prior that filters out background regions. However, we observe that, in the absence of masks, we can simply derive a tight bounding box from the ground-truth keypoints and restrict MNN mining to this region. As reported in Tab. 11, the accuracy difference between the two variants is negligible:

- on SPair-71k, from 27.0 to 26.6 PCK@0.01 and from 87.2 to 86.7 PCK@0.10;
- on SPair-U, from 67.5 to 66.9 PCK@0.10.

In all cases, MARCO *without SAM-based masks* still exceeds all prior methods by a considerable margin. This shows that masks merely provide a mild foreground prior during pseudo-label mining and can be removed with almost no degradation. In other words, MARCO performs competitively even without any supervision beyond the sparse landmarks. For completeness, Tab. 11 summarizes the supervision used by each competing approach (keypoints, object masks, and depth) together with their accuracy on SPair-71k and SPair-U.

Table 9. **Pseudo-label noise estimation.** SPair-U PCK@0.10 (%, ↑) when Gaussian noise with standard deviation σ (px) is added to pseudo-label coordinates. Performance degrades near σ=10.

σ (px)	0	0.5	1	5	10
SPair-U	67.5 ± 0.2	67.7 ± 0.3	67.2 ± 0.2	67.1 ± 0.4	65.0 ± 0.5

Table 10. **Clustering sensitivity.** Performance on SPair-U (PCK@0.10, in %, ↑) for different initial numbers of clusters k. Initializing with a larger k and merging clusters using BIC yields the best result, avoiding the need to tune k.

k	3	5	10	15	15 + BIC
SPair-U	66.6	66.9	66.5	66.0	**67.5**

Table 11. **Comparison of supervision levels and acuracy.** Per-image PCK (in %, ↑) on SPair-71k and SPair-U. All recent methods use SAM masks during training. Geo-SC uses masks during inference (†). Jamais Vu additionally requires depth supervision. MARCO gives accurate results even with no supervision beyond the sparse keypoints.

	Supervision			SPair-71k		SPair-U	
Method	Keypoint	Mask	Depth	0.01	0.10	0.01	0.10
Unsupervised / Weakly Supervised							
DistillDIFT [10]	–	✓	–	8.9	65.1	–	–
DIY-SC [7]	–	✓	–	10.1	71.6	–	–
Supervised							
Geo-SC [54]	✓	†	–	21.7	83.2	3.9	56.9
GECO [14]	✓	✓	–	14.2	73.6	3.2	55.2
Jamais Vu [32]	✓	✓	✓	20.5	82.5	4.2	62.4
MARCO *(ours)*	✓	✓	–	**27.0**	**87.2**	**4.7**	**67.5**
MARCO *(ours)*	✓	–	–	26.6	86.7	4.3	66.9

E. Computational Cost

Table 12 compares the computational footprint of MARCO with respect to state-of-the-art solutions. Geo-SC [54] and Jamais Vu [32] use the same dual-encoder architecture combining Stable Diffusion and DINOv2, resulting in 950M parameters. In contrast, MARCO relies on a single DINOv2 backbone with adapters, to-

Table 12. **Compute comparison.** Model size and inference speed measured on an RTX4090 GPU. All methods use the same evaluation protocol: feature extraction at 840p, batched reference–target pairs, and the same soft-argmax keypoint prediction.

Method	Backbone	Params	FPS ↑
Geo-SC [54]	SD + DINOv2	950M	0.85
Jamais Vu [32]	SD + DINOv2	950M	0.85
MARCO	DINOv2	323M	**8.30**

taling 323M parameters. We measure inference speed on a single NVIDIA RTX4090 GPU. To ensure a fair comparison, all methods follow the same evaluation protocol: feature extraction at 840p resolution, batched reference–target image pairs, and the same soft-argmax prediction [54]. For Geo-SC and Jamais Vu we use the original Geo-SC implementation. Under these conditions, MARCO runs at 8.3 FPS, compared to 0.85 FPS for Geo-SC and Jamais Vu, corresponding to roughly a $10\times$ speedup.

F. Pseudocode of Dense Self-Distillation

For clarity, Algorithm 1 summarizes the procedure used to generate dense pseudo-correspondences and compute the dense self-distillation loss described in Sec. 3.3. The teacher network is maintained as an exponential moving average (EMA) of the student parameters and is used to generate stable pseudo-labels. Given a pair of images, the teacher features are used to mine reliable correspondences through mutual nearest-neighbor matches, which are combined with the available ground-truth keypoints. These correspondences are then *densified* by estimating a piecewise-affine warp obtained from a Delaunay triangulation of the seed points, producing a dense flow field between the two images. Flow vectors are subsequently clustered to identify regions with coherent motion, and clusters consistent with the ground-truth correspondences are retained to form pseudo-labels. These pseudo-correspondences supervise the student network through a regression loss, while the teacher parameters are updated via EMA during training.

Algorithm 1 Dense self-distillation via flow anchoring

Require: Source and target images $(\mathbf{I}^s, \mathbf{I}^t)$, sparse GT correspondences $\mathcal{E}$, student parameters θ_S, teacher parameters θ_T
Ensure: Self-distillation loss $\mathcal{L}_{\text{self}}$

Teacher / student feature extraction
$\mathbf{F}_T^s, \mathbf{F}_T^t \leftarrow \Phi_{\theta_T}(\mathbf{I}^s), \Phi_{\theta_T}(\mathbf{I}^t)$
$\mathbf{F}_S^s, \mathbf{F}_S^t \leftarrow \Phi_{\theta_S}(\mathbf{I}^s), \Phi_{\theta_S}(\mathbf{I}^t)$
Seed correspondence extraction
Compute mutual nearest neighbors:

$$\mathcal{P}_{\text{MNN}} = \{(\mathbf{u}, \mathbf{v}) \mid \text{NN}_{s\to t}(\mathbf{u}) = \mathbf{v} \wedge \text{NN}_{t\to s}(\mathbf{v}) = \mathbf{u}\}$$

Restrict $\mathcal{P}_{\text{MNN}}$ to pixels inside the object mask
Form seed correspondences: $\mathcal{P}_{\text{seed}} \leftarrow \mathcal{E} \cup \mathcal{P}_{\text{MNN}}$
Dense flow estimation
Construct Delaunay triangulation $\mathcal{T}$ over source points $\mathcal{P}_{\text{seed}}$
for each triangle $\tau \in \mathcal{T}$ **do**
    Define target triangle τ' from the matched vertices
    Estimate affine warp $\mathcal{W}_{\tau\to\tau'}$
end for
Compose all triangle warps into piecewise-affine mapping $\hat{\mathcal{W}}$
Compute dense flow field $\mathbf{D}(\mathbf{u}) = \hat{\mathcal{W}}(\mathbf{u}) - \mathbf{u}$
Flow clustering and GT anchoring
Cluster flow vectors $\mathbf{D}(\mathbf{u})$ using k-means
Merge clusters using BIC to obtain $\{\Omega_n\}$
for each cluster Ω_n **do**
    $C_n^s = \{\mathbf{u} \mid \mathbf{D}(\mathbf{u}) \in \Omega_n\}$
    $C_n^t = \{\mathbf{u} + \mathbf{D}(\mathbf{u}) \mid \mathbf{u} \in C_n^s\}$
end for
Retain clusters anchored by GT matches:

$$\mathcal{P}_{\text{self}} = \{(\mathbf{u}, \mathbf{u}+\mathbf{D}(\mathbf{u})) \mid \exists n, i : \mathbf{u} \in C_n^s, \mathbf{p}_i^s \in C_n^s, \mathbf{p}_i^t \in C_n^t\}$$

Self-distillation loss

$$\mathcal{L}_{\text{self}} = \frac{1}{|\mathcal{P}_{\text{self}}|} \sum_{(\hat{\mathbf{u}}, \hat{\mathbf{v}}) \in \mathcal{P}_{\text{self}}} \left\| \underset{\mathbf{u}\in\hat{\Lambda}}{\text{soft-argmax}}\, S(\hat{\mathbf{u}}, \mathbf{u}) - \hat{\mathbf{v}} \right\|_2^2$$

EMA teacher update $\theta_T \leftarrow \beta\theta_T + (1-\beta)\theta_S$
return $\mathcal{L}_{\text{self}}$

G. Details on the MP-100 Benchmark

Our goal is to establish an evaluation protocol to thoroughly assess the generalization ability of correspondence models beyond the specific landmarks and categories observed during training. Concurrently to our work, Jamais Vu [32] highlights a similar limitation and augments SPair-71k with only four additional keypoint definitions per category. While this extension is useful, SPair-U remains limited both in the number of new keypoints and in its restriction to the original SPair-71k taxonomy.

To broaden this perspective, we turn to the ecosystem of 2D pose estimation [55], where annotation schemes vary widely across object types. In particular, we repurpose the MP-100 dataset [48], a large-scale collection spanning 100 categories and 18k images. For context, SPair-71k contains only 1.8k images in total. Similarly to us, Geo-SC [54] adopts a pose estimation dataset, namely AP-10K, for semantic correspondence. However, their focus is primarily to provide a source of training, whereas our intent is to provide an evaluation benchmark to measure generalization under keypoint and category shift. We describe our curation protocol below.

Data curation. We first filter out images with fewer than three visible keypoints, accounting for occlusion or missing annotations. Next, to ensure a strict zero-shot setup for unseen-category evaluation, we remove all classes overlapping with the SPair-71k training categories, namely: *cat body, sheep body, horse body, dog body, cow body, goldenretriever face, germanshepherddog face, bighornsheep face, przewalskihorse face, car, bus*. For categories that only partially overlap with SPair-71k but contribute a substantial number of novel keypoints, we retain them for the unseen-keypoints evaluation. In particular, the *apparel items* super-category partially overlaps with the SPair-71k *person* class, but provides many fine-grained garment-specific keypoint definitions that are entirely absent from SPair-71k. The same rationale applies to the *human face* super-category, which offers a dense and diverse set of facial landmarks. After filtering, we group the cate-

Table 13. **Statistics on MP-100 benchmark.**

SPair-71k training categories		**Split type**	**MP-100 categories used in our benchmark**
ID	Category	**Unseen keypoints**	**Human face:**
1	aeroplane	Human face	Human face.
2	bicycle	Categories: 1	**Apparel item:**
3	bird	Avg. keypoints/pair: 68	short sleeved outwear, short sleeved shirt, skirt, short sleeved dress, vest dress,
4	boat	#Keypoint definitions: 68	long sleeved dress, long sleeved outwear, long sleeved shirt,
5	bottle	Apparel item	sling, sling dress, trousers, vest.
6	bus	Categories: 12	**Animal body:**
7	car	Avg. keypoints/pair: 27	macaque body, locust body, fly body, antelope body, cheetah body, fox body,
8	cat	#Keypoint definitions: 282	leopard body, panther body, rat body, squirrel body, beaver body, deer body,
9	chair	**Unseen categories**	giraffe body, lion body, pig body, rhino body, weasel body, bison body,
10	cow	Animal body	elephant body, gorilla body, otter body, polar bear body, skunk body, wolf body,
11	dog	Categories: 32	hippo body, bobcat body, raccoon body, hamster body, panda body,
12	horse	Avg. keypoints/pair: 15	rabbit body, spider monkey body, zebra body.
13	motorbike	#Keypoint definitions: 101	**Animal face:**
14	person	Animal face	alpaca face, californian sea lion face, chipmunk face, ferret face, gibbons face,
15	plant	Categories: 26	guanaco face, proboscis monkey face, arctic wolf face, camel face,
16	sheep	Avg. keypoints/pair: 9	common warthog face, gentoo penguin face, grey seal face, klipspringer face,
17	train	#Keypoint definitions: 9	fennec fox face, blackbuck face, cape buffalo face, dassie face, gerbil face,
18	tv/monitor	Home furniture	grizzly bear face, olive baboon face, quokka face, bonobo face, capybara face,
		Categories: 4	fallow deer face, onager face, pademelon face.
		Avg. keypoints/pair: 12	**Home furniture:**
		#Keypoint definitions: 45	couch, table, bed, swivel chair.

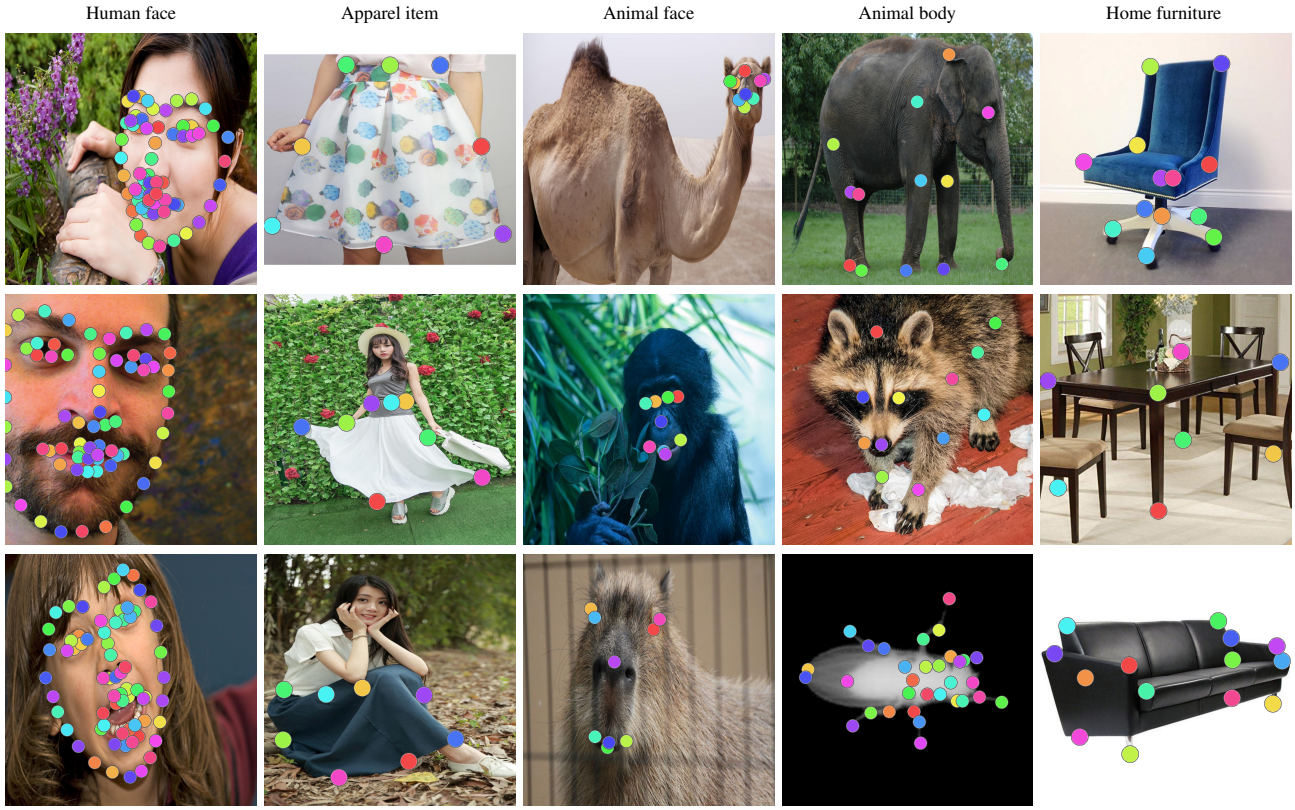

Figure 6. **MP-100 samples used in our benchmark.** Each column shows representative instances from the five macro-domains.

gories into five domains, *i.e.*, *human face, apparel items, furniture, animal face, animal body*. We sample 2k image pairs within each domain using stratified sampling to avoid class imbalance.

Table 13 summarizes the resulting benchmark splits. For acquisition details and annotation conventions, we refer readers to the original MP-100 paper [48]. To give an intuitive understanding of the visual diversity present across the domains used in our benchmark, we include representative samples in Fig. 6. The samples highlight the substantial variation in appearance, pose, texture, and structure across human faces, apparel items, animal species, and furniture categories, as well as the richness and heterogeneity of their keypoint definitions, which make this a challenging testbed for semantic correspondence.

Unseen keypoints. This setting evaluates a model's ability to generalize to *new keypoint definitions* for object categories seen during training. For instance, *human face* appears in SPair-71k as part of the broader *person* class, but only with seven coarse keypoints (eyes, ears, nose, mouth, chin). In contrast, MP-100 provides a 68-point landmark definition capturing fine-grained facial geometry. A similar situation arises for apparel items: although clothing is implicitly present within the *person* category in SPair-71k, MP-100 introduces rich keypoint annotations on garments (*e.g.*, sleeve corners, skirt borders), with 282 novel keypoint definitions over 12 categories. These cases allow us to test fine-grained keypoint transfer under class-level familiarity.

Unseen categories. Here, we evaluate generalization to object types for which *no* keypoint annotations are available during training. We remove any category present in SPair-71k to enforce a strict zero-shot setting. The *animal body* (32 categories) and *animal face* (26 categories) domains include species entirely absent from SPair-71k. The *home furniture* domain includes *couch*, *table*, *bed*, and *swivel chair*. Although *swivel chair* may be loosely related to *chair*, it does not appear in SPair-71k, and its keypoint definition (*e.g.*, rotating base) differs meaningfully. We therefore choose to include it in the *unseen-category* split.

Split protocol. We organize the benchmark into five evaluation splits: two for the *unseen-keypoint* scenario (human face and apparel items), and three for the *unseen-category* scenario (animal body, animal face, and home furniture). This structure allows us to separately examine *(i)* generalization to fine-grained keypoint definitions within familiar categories, and *(ii)* generalization across entirely novel object types with no lexical or semantic overlap with the SPair-71k training set.

G.1. Additional results on MP-100

Table 14 extends the main paper evaluation by reporting the accuracy across PCK thresholds on all MP-100 splits. Overall, the trends observed at PCK@0.10 remain stable at both finer (PCK@0.05) and coarser (PCK@0.15) thresholds. On *unseen categories*, MARCO remains the strongest method across all domains and all thresholds. Averaged across domains, it improves over the strongest prior method by +4.3 % at PCK@0.05, +4.5 % at PCK@0.10, and +5.4 % at PCK@0.15. The gains are especially pronounced on *Home furniture*, but remain consistent also on *Animal body* and *Animal face*, indicating robust transfer to categories never seen during training. On *unseen keypoints*, the comparison is more nuanced. For *Human face*, the strongest competitor is the zero-shot DIFT baseline: MARCO is slightly weaker at PCK@0.05, but becomes the best method at coarser thresholds, improving over DIFT by +0.2 % at PCK@0.10 and +1.4 % at PCK@0.15. For *Apparel items*, MARCO consistently outperforms SD+DINO, with gains of +1.5 %, +5.7 %, and +9.0 % at PCK@0.05, PCK@0.10, and PCK@0.15, respectively. Interestingly, several zero-shot methods remain highly competitive, and in some cases outperform supervised baselines. This supports our main observation: training with sparse keypoint annotations limits its transfer to new landmark vocabularies. In contrast, MARCO preserves the transferable structure of the pre-trained backbone, yielding more stable results across both unseen keypoints and unseen categories. Overall, these results reinforce MP-100 as a challenging and informative benchmark for assessing correspondence generalization.

G.2. Per-category results on MP-100

For completeness, Tabs. 15 and 16 report detailed per-category results on MP-100 using PCK@0.10. We compare MARCO with representative prior approaches, namely DINOv2 [37], GeoSC [54], and Jamais Vu [32]. Table 15 focuses on *unseen categories*, while Tab. 16 reports results on *unseen keypoints* within known categories. A clear pattern emerges on *unseen categories*: MARCO achieves the best accuracy on the large majority of classes, with especially strong gains on highly variable categories such as *macaque* (+7.6 over Jamais Vu), *rhino* (+6.0), *cape buffalo* (+8.2), *olive baboon* (+11.3), *bed* (+14.7), and *couch* (+9.0). These results indicate that the proposed dense self-distillation improves transfer not only across new animal species, but also to object families with very different geometry, such as home furniture. At the same time, the table also highlights genuinely difficult categories where all methods remain relatively close, or where a strong frozen foundation model remains competitive, such as *table*, *locust*, *fly*, and *polar bear*. This suggests that some categories are limited less by semantic transfer and more by intrinsic ambiguity, symmetry, or annotation difficulty. On *unseen keypoints*, MARCO is consistently best across all apparel categories and on *human face*, often with large margins. In particular, the gains over Jamais Vu reach +10.0 on *sling*, +14.6 on *sling dress*, +11.1 on *long sleeved shirt*, and +9.2 on *short sleeved dress*, while remaining positive on all other categories. These improvements are notable because this setting introduces new landmark vocabularies within categories already seen during training, directly testing whether the model has learned dense semantic structure rather than memorizing the supervised keypoints. Interestingly, the unsupervised DINOv2 baseline is already fairly strong in some categories, especially *human face*, confirming that foundation features contain substantial transferable structure; however, MARCO consistently improves on top of this prior structure and yields the strongest overall generalization.

H. Qualitative Examples

To complement the quantitative analyses, Fig. 7 provides a series of qualitative visualizations, illustrating the behavior of MARCO across a wide range of settings. We include examples from **SPair-71k**, highlighting the model's ability to produce accurate and spatially coherent correspondences under significant appearance

Table 14. **Generalization on MP-100** [48]. This table extends Tab. 2 of the main paper, reporting results across PCK thresholds on our proposed MP-100 benchmark. We evaluate methods trained on SPair-71k, as well as zero shot baselines, indicated with *. Supervised methods often fall short of zero-shot approaches, highlighting a generalization gap in existing approaches. In contrast, MARCO maintains robust performances across domains, outside of the training distribution. Per-image PCK (in %, ↑), best **bold**, second best underlined.

	Unseen keypoints						**Unseen categories**								
	Human face			Apparel items			Animal body			Home furniture			Animal face		
	0.05	0.10	0.15	0.05	0.10	0.15	0.05	0.10	0.15	0.05	0.10	0.15	0.05	0.10	0.15
DINOv2 * [37]	41.8	66.2	76.9	24.9	44.7	58.3	22.6	36.1	46.2	30.5	44.2	53.0	13.5	33.3	46.7
DIFT * [43]	68.9	87.3	92.8	29.3	48.2	59.0	19.1	31.0	39.9	34.9	46.9	54.1	10.0	26.5	38.8
SD + DINO * [53]	68.3	85.3	90.8	29.4	50.2	62.1	23.8	36.1	45.2	36.5	49.2	56.1	17.4	39.6	53.1
GECO [14]	40.8	82.9	86.3	23.4	41.7	56.6	23.2	31.9	43.7	37.9	48.1	58.2	17.2	38.2	50.0
Geo-SC [54]	63.5	85.2	91.1	23.8	42.9	54.9	27.1	38.9	47.6	40.4	49.6	55.0	24.0	49.2	61.3
Jamais Vu [32]	64.0	85.5	91.7	25.0	45.7	58.8	27.0	39.3	48.3	42.5	52.7	58.7	22.9	47.7	59.8
MARCO *(ours)*	**64.3**	**87.5**	**94.2**	**30.9**	**55.9**	**71.1**	**29.9**	**42.3**	**51.5**	**48.8**	**60.4**	**66.9**	**26.6**	**52.6**	**64.7**

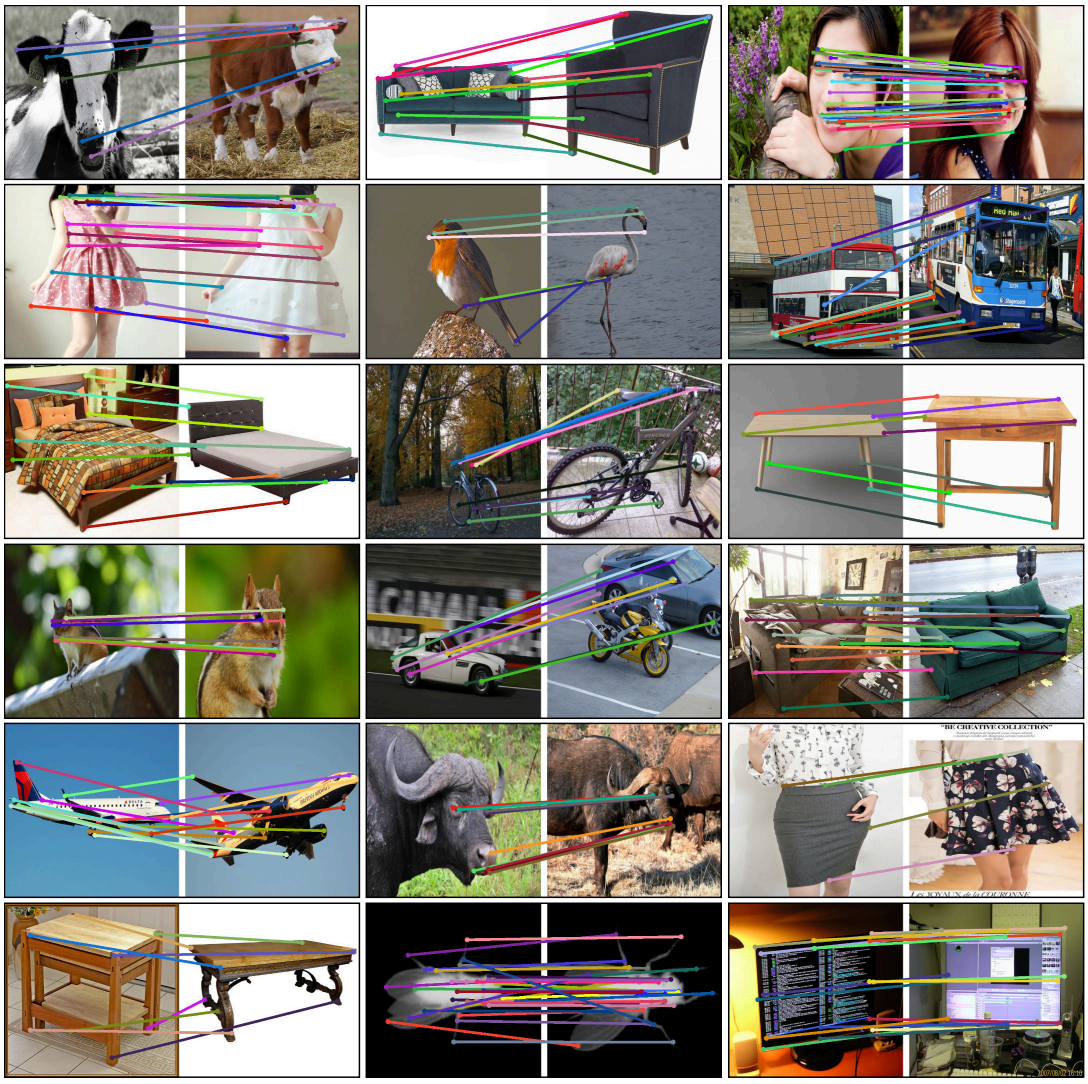

Figure 7. **Qualitative results with MARCO.** Examples from Spair-71k [34] and MP-100 [48].

Table 15. **Per-category evaluation on MP-100 [48]**: unseen categories. Per-image PCK@0.10 (in %, ↑).

Unseen categories

Domain	Category	DINOv2	Geo-SC	Jamais Vu	MARCO
Animal body	macaque	46.5	53.6	54.3	**61.9**
	locust	64.7	70.4	**71.1**	65.8
	fly	63.3	70.8	**73.4**	66.8
	antelope	37.9	42.8	41.7	**48.1**
	cheetah	36.8	40.3	41.2	**44.0**
	fox	46.0	50.5	51.3	**55.7**
	leopard	32.4	35.5	37.8	**39.4**
	panther	40.6	41.1	40.0	**46.8**
	rat	26.5	26.5	27.3	**29.1**
	squirrel	35.2	41.4	41.4	**44.9**
	beaver	29.8	21.6	19.4	**28.9**
	deer	41.2	48.0	50.1	**53.7**
	giraffe	31.1	**35.3**	34.6	35.2
	lion	37.8	42.6	43.4	**47.9**
	pig	19.8	22.4	22.3	**26.2**
	rhino	44.3	51.7	52.8	**58.8**
	weasel	42.6	44.3	45.4	**50.7**
	bison	31.0	36.0	37.2	**40.0**
	elephant	23.4	25.5	25.6	**31.4**
	gorilla	33.4	31.8	32.0	**36.1**
	otter	33.5	34.0	32.9	**35.9**
	polar bear	31.3	35.8	35.8	**33.1**
	skunk	34.3	31.0	34.5	**38.6**
	wolf	35.4	41.1	40.0	**42.1**
	hippo	26.3	28.1	28.3	**30.5**
	bobcat	32.5	35.9	36.8	**40.6**
	raccoon	34.3	35.0	**38.2**	37.5
	hamster	38.2	40.5	41.2	**45.6**
	panda	24.1	25.0	25.9	**27.6**
	rabbit	39.2	41.1	39.3	**42.4**
	spider monkey	31.6	30.0	28.8	**33.8**
	zebra	33.2	36.4	37.4	**37.9**
Animal face	alpaca	22.1	28.9	29.9	**34.4**
	californian sea lion	23.7	30.3	30.8	**40.7**
	chipmunk	39.3	**61.7**	56.9	60.0
	ferret	38.7	**69.8**	69.3	68.0
	gibbons	23.7	44.1	45.9	**51.5**
	guanaco	21.0	26.8	27.5	**36.8**
	proboscis monkey	25.3	40.2	39.8	**48.1**
	arctic wolf	43.2	63.3	63.4	**65.8**
	camel	24.2	32.1	31.6	**39.0**
	common warthog	44.1	**61.1**	55.4	58.3
	gentoo penguin	28.2	33.3	34.0	**38.3**
	grey seal	32.1	49.3	47.6	**51.8**
	klipspringer	38.7	49.1	47.9	**53.8**
	fennec fox	37.1	**63.7**	61.3	62.8
	blackbuck	17.5	29.9	29.1	**34.8**
	cape buffalo	34.3	53.2	53.2	**61.4**
	dassie	36.1	**58.7**	55.2	57.2
	gerbil	31.4	**51.5**	46.6	51.1
	grizzly bear	44.0	57.8	56.4	**60.4**
	olive baboon	39.4	48.3	49.8	**61.1**
	quokka	34.6	51.6	48.2	**56.5**
	bonobo	38.1	**63.0**	61.5	62.7
	capybara	36.3	45.0	43.5	**49.1**
	fallow deer	29.4	42.6	38.6	**44.4**
	onager	44.7	53.0	50.2	**53.5**
	pademelon	41.5	**70.1**	68.1	67.5
Home furniture	couch	61.1	61.7	69.1	**78.1**
	table	24.0	**25.9**	**25.9**	23.9
	bed	37.7	44.9	47.0	**61.7**
	swivel chair	52.9	65.2	67.3	**75.8**

changes, occlusions, and viewpoint variations. We further showcase results on our **MP-100 benchmark**, focusing on both the *unseen-keypoint* and *unseen-category* regimes. These examples

Table 16. **Per-category evaluation on MP-100 [48]**: unseen keypoints. Per-image PCK@0.10 (in %, ↑).

Unseen keypoints

Domain	Category	DINOv2	Geo-SC	Jamais Vu	MARCO
Apparel item	short sleeved outwear	46.5	45.3	48.2	**58.9**
	short sleeved shirt	48.8	49.6	52.7	**61.3**
	skirt	40.0	32.6	34.7	**43.9**
	short sleeved dress	48.2	46.8	50.6	**60.0**
	vest dress	54.4	56.2	60.7	**69.9**
	long sleeved dress	43.7	42.2	45.8	**55.0**
	long sleeved outwear	42.5	42.4	46.1	**52.6**
	long sleeved shirt	42.5	41.9	43.9	**55.0**
	sling	42.6	33.7	33.8	**51.5**
	sling dress	48.6	45.9	48.6	**63.2**
	trousers	35.3	37.4	39.4	**44.7**
	vest	42.5	40.4	43.4	**52.6**
Human face	human	66.2	85.2	85.5	**87.5**

demonstrate how MARCO adapts to novel landmark definitions it has never been trained on, and how its predictions remain stable even for novel object types with distinct shapes and geometries. Overall, the qualitative results visually confirm the strong generalization capabilities encouraged by our dense self-distillation framework.

References

[55] K. Sun, B. Xiao, D. Liu, and J. Wang. Deep high-resolution representation learning for Human Pose Estimation. In *CVPR*, pages 5693–5703, 2019. iv

[56] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly. Parameter-efficient transfer learning for NLP. In *ICML*, pages 2790–2799, 2019. ii