DINeMo: Learning Neural Mesh Models with no 3D Annotations

Weijie Guo[1,2*]   **Guofeng Zhang**[1]   **Wufei Ma**[1]   **Alan Yuille**[1]

[1]Johns Hopkins University   [2] Peking University

Abstract

Category-level 3D/6D pose estimation is a crucial step towards comprehensive 3D scene understanding, which would enable a broad range of applications in robotics and embodied AI. Recent works explored neural mesh models that approach a range of 2D and 3D tasks from an analysis-by-synthesis perspective. Despite the largely enhanced robustness to partial occlusion and domain shifts, these methods depended heavily on 3D annotations for part-contrastive learning, which confines them to a narrow set of categories and hinders efficient scaling. In this work, we present DINeMo, a novel neural mesh model that is trained with no 3D annotations by leveraging pseudo-correspondence obtained from large visual foundation models. We adopt a bidirectional pseudo-correspondence generation method, which produce pseudo correspondence utilizing both local appearance features and global context information. Experimental results on car datasets demonstrate that our DINeMo outperforms previous zero- and few-shot 3D pose estimation by a wide margin, narrowing the gap with fully-supervised methods by 67.3%. Our DINeMo also scales effectively and efficiently when incorporating more unlabeled images, which demonstrate the advantages over supervised learning methods that rely on 3D annotations. Our project page is available here.

1. Introduction

Estimating the 3D location and 3D orientation of objects from a certain category is both a challenging and essential step towards comprehensive scene understanding [1, 2, 9, 13]. Models must learn to inference 3D formulations from 2D signals only, while also generalizing across different shapes and appearances within a category. However, generalization to out-of-distribution (OOD) scenarios, *e.g.*, partial occlusions, novel shapes and appearances, remain a fundamental challenge [6, 10, 38].

Inspired by cognitive studies on human vision [18, 34], recent works explored compositional 3D representations of objects, *e.g.*, neural mesh models, and approached a

range of 2D and 3D tasks with an analysis-by-synthesis process [9, 13, 30]. By learning a part-contrastive 3D feature representation of the objects, these methods can jointly optimize the 3D location, 3D orientation, and part visibility of objects with a feature construction loss. Despite the improved robustness to partial occlusion and domain shifts [38], these methods require various 3D annotations that enable part-contrastive learning, such as object orientations [15, 32] or human poses [7, 37]. These 3D annotations are often hard to obtain — they are time-consuming to annotate and require certain expertise from the annotators, which limits the applicability of these neural mesh models to a broader range of objects or to scaling up efficiently. To address the scarcity of 3D annotations, previous studies [14] explored generative models to produce synthetic images with 3D annotations. However, models trained on synthetic data must also be finetuned on real data with 3D annotations given the considerable domain gap between diffusion-generated and real images.

In this work, we present DINeMo, a novel neural mesh model that is trained without 3D annotations , enabled by leveraging large pretrained visual foundation models, such as DINOv2 [20]. Rather than relying on groundtruth keypoint locations as in previous works [9, 13], we train DINeMo with part-contrastive loss using pseudo-correspondence obtained by matching neural features from SD-DINO [35]. However, raw pseudo-correspondence from SD-DINO can be quite noisy, *e.g.*, keypoints are often mismatched between left and right. We argue that keypoint correspondence matching should consider both local information, *i.e.*, per patch feature similarities, and global context information, *i.e.*, 3D orientation of the object. Based on this motivation, we propose a novel bidirectional pseudo-correspondence generation, which consists of two steps: (i) matching a global pose label from raw keypoint correspondences, and (ii) refine local keypoint correspondences based on the predicted global pose label. Lastly we extend standard analysis-by-synthesis inference with Grounded-SAM[23] masks, achieving enhanced occlusion robustness.

Extensive results on car class demonstrate that our DINeMo outperforms previous zero- and few-shot 3D pose estimation methods by a wide margin on both in-

---

distribution and partial occlusion testing data, achieving a 67.3% reduction in the performance gap relative to fully supervised methods. Our approach also outperforms all previous methods on SPair71k [16] for semantic correspondence.

As DINeMo can be trained on object images without 3D annotations, we further study the scaling properties of our approach by involving abundant data from public image datasets, such as Stanford Cars [12]. Experimental results demonstrate that our DINeMo scales effectively and efficiently by involving more unlabeld images for training. This demonstrates the advantages of DINeMo over previous fully-supervised approaches [9, 13], which struggle to scale up due to the reliance on 3D annotations.

In summary, our main contributions are as follows: (1) We present DINeMo, a novel neural mesh model trained with pseudo-labels from our bidirectional pseudo-correspondence generation method. (2) Experimental results show that DINeMo outperforms previous zero- and few-shot methods by a wide margin, largely narrowing the gap with fully-supervised methods. (3) Our DINeMo scales effectively with more unlabeled images for training.

2. Related Work

Category-level pose estimation. Multiple strategies have been developed for category-level 3D pose estimation. Traditional methods formulate this task as a classification problem that predicts discrete pose labels [17, 28]. Another line of work follows a two-stage, keypoint-driven pipeline [39], where semantic keypoints are detected and used to estimate pose via Perspective-n-Point algorithm. More recent advancements shifted toward render-and-compare paradigms [9, 13, 30, 33], transitioning from generating pixel values to semantic feature representations. These methods frame pose estimation as an optimization problem that minimizes the discrepancy between features extracted from input image and ones rendered from a posed 3D mesh.

Keypoint correspondence. Several approaches have been developed to establish correspondences between 2D image keypoints and 3D mesh vertices—a fundamental component in tasks such as shape reconstruction and 3D pose estimation. At the category level, however, most existing methods focused on deformable object classes, particularly humans [4, 22] and animals [26]. This focus is largely attributed to the availability of large-scale datasets with dense image-to-template annotations, such as DensePose-COCO [4] for humans and DensePose-LVIS [19] for animals. Notably, recent work has demonstrated the feasibility of learning dense correspondences without explicit keypoint supervision [26], particularly in the animal domain. In this work, we extend this line of research to rigid object categories, which have received comparatively less attention.
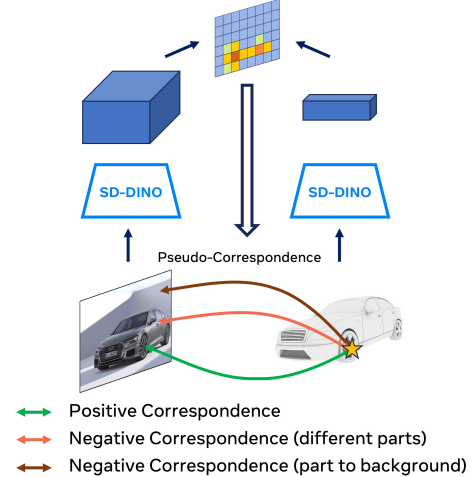

Figure 1. **Overview of DINeMo,** a novel neural mesh model trained on pseudo-correspondence obtained from large visual foundation models.

Render-and-compare approaches estimate 3D poses by minimizing the reconstruction error between feature representations projected from a 3D object and those extracted from the input image. This strategy can be interpreted as a form of approximate analysis-by-synthesis [3], which contrasts with purely discriminative methods and has demonstrated increased robustness to out-of-distribution scenarios. Particularly, such approaches have shown effectiveness in handling partial occlusions, proving beneficial in both object classification [11] and 3D pose estimation tasks [8, 30].

3. Methods

In this section we introduce our DINeMo, as illustrated in Fig. 1. We start by reviewing the neural mesh models in Sec. 3.1. Then we introduce our bidirectional pseudo-correspondence generation in Sec. 3.2. Lastly we present our inference method in Sec. 3.3, which leverages pre-trained segmentation model to optimize pose parameters.

3.1. Preliminaries: Neural Mesh Models

Neural mesh models [9, 13, 30] define a probabilistic generative model $p(F \mid \mathfrak{N})$ of feature activations F using a 3D neural mesh $\mathfrak{N} = \{\mathcal{V}, \mathcal{E}, \mathcal{C}\}$, where $\mathcal{V} = \{V_i \in \mathbb{R}^3\}_{i=1}^N$ is the set of mesh vertices, $\mathcal{E}$ is the edge set, and $\mathcal{C} = \{C_i \in \mathbb{R}^c\}_{i=1}^N$ is the learnable feature representation for each vertex. Given pose parameters m, we define the likelihood of a target feature map $F = f_\Phi(I)$ as

$$p(F \mid \mathfrak{N}, m, C_b) = \prod_{i \in \mathcal{FG}} p(f_i \mid \mathfrak{N}, m) \prod_{i' \in \mathcal{BG}} p(f_{i'} \mid C_b) \tag{1}$$

where f_Φ is the network backbone parameterized by Φ, $\mathcal{FG}$ and $\mathcal{BG}$ are set of foreground and background positions,

Figure 2. **Bidirectional pseudo-correspondence generation.** See Sec. 3.2.
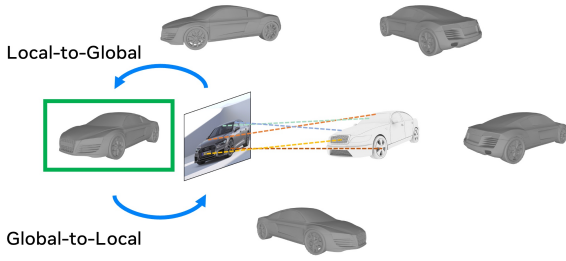

Figure 3. **Qualitative comparisons with and without our bidirectional pseudo-correspondence generation.** See Sec. 3.2.

and C_b is the background feature. The network parameters Φ and learnable feature $\{\mathcal{C}, C_b\}$ are optimized with a part-contrastive loss. During inference, we minimize the negative log-likelihood w.r.t. pose parameters m.

3.2. Bidirectional Pseudo-Correspondence

Previous works train neural mesh models on keypoint correspondences obtained from 3D annotations, such as object or human poses. In this work, we propose to leverage pseudo-correspondence from visual foundation models, eliminating the need for 3D annotations that are difficult to obtain.

As demonstrated in Fig. 1, we follow SD-DINO [35] and compute feature representations of the image and template mesh renderings from multiple views. Specifically, we extract and concatenate the DINOv2 [20] features and Stable Diffusion [24] features. We aggregate the per-vertex feature similarities from each view with the max operator and obtain the keypoint pseudo-correspondence from cosine similarities between normalized feature representations.

However, we found that raw pseudo-correspondences estimated from SD-DINO [35] feature similarities are noisy (see Fig. 3). In particular, object parts can be confused between left and right. This is because pseudo-correspondences from feature similarities only consider local apperances and lack of high-level consistency, *e.g.*, 3D object pose. We argue that keypoint correspondence matching should consider both local information, such as local appearances, as well as global context information, *i.e.*, 3D formulation of the object.

We propose **bidirectional pseudo-correspondence generation**, which generates keypoint corresopndence considering both low-level and high-level semantics. Our method consists of two steps: (i) Local-to-global: in the first stage we obtain raw pseudo-correspondences from SD-DINO [35] and then determine the 3D object orientation by majority voting. (ii) Global-to-local: in the second stage we refine the raw pseudo-correspondences by downweighting matching scores with vertices that are not visible from the estimated 3D orientation by a fixed constant. Our approach effectively integrates both low-level feature similarities and
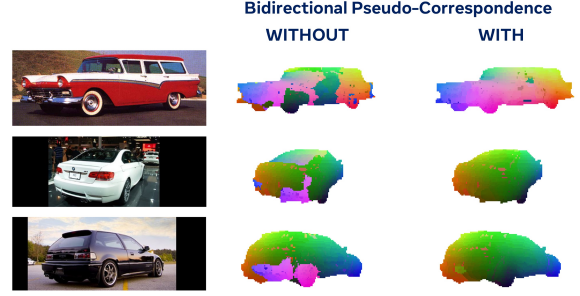
high-level context information and generates more consistent keypoint pseudo-correspondence as shown in Fig. 3.

3.3. Inference

During inference, we minimize the negative log-likelihood w.r.t. pose parameters m with gradient descent

$$
\begin{aligned}
&\mathcal{L}_{\mathrm{NLL}}(F, \mathfrak{N}, m, C_b) \\
&= -\ln p(F \mid \mathfrak{N}, m, C_b) \\
&= -\sum_{i \in \mathcal{FG}} \ln p(f_i \mid \mathcal{N}, m) - \sum_{i' \in \mathcal{BG}} \ln p(f_{i'} \mid C_b) \quad (2)
\end{aligned}
$$

To find foreground and background regions, previous works [9, 13, 30] introduce an one-hot map $\mathcal{Z}$ indicating if each 2D location is visible or not based on feature activations. Empirically we find that the one-hot map often involves a significant portion of the background, leading to degraded performance on partial occlusion data. Given the recent advancements of segmentation methods, we first predict $\mathcal{Z}$ with a SAM2 [21] model and then optimizes the negative log-likelihood loss while fixing $\mathcal{Z}$.

4. Experiments

In this section we present our experiments results on 3D pose estimation and semantic correspondence in Sec. 4.1. We study the scaling properties of our DINeMo in Sec. 4.2. For details about our experimental setup, please refer to Sec. 8 in supplementary materials.

4.1. Main Results

3D Object Pose Estimation We evaluate our model on the car split of the PASCAL3D+ dataset [32] for in-distribution 3D pose estimation and occluded PASCAL3D+ dataset [29] for partial occlusion generalization. As we can see from the results in Tab. 1, as a zero-shot method, our DINeMo outperforms previous zero- and few-shot methods by a wide margin, *i.e.*, by 27.3% over NVS [31] (50-shot) and by 10.5% over 3D-DST [14]. Moreover, our DINeMo significantly narrows the gap between zero- and few-shot

Methods	L0		L1		L2		L3	
	Acc@$\frac{\pi}{6}$	Acc@$\frac{\pi}{18}$	Acc@$\frac{\pi}{6}$	Acc@$\frac{\pi}{18}$	Acc@$\frac{\pi}{6}$	Acc@$\frac{\pi}{18}$	Acc@$\frac{\pi}{6}$	Acc@$\frac{\pi}{18}$
Fully-Supervised								
Resnet50 [5]	95.5	63.5	80.0	40.7	57.0	21.4	36.9	7.6
NOVUM [9]	97.9	94.9	91.9	78.0	77.1	52.3	49.8	23.8
Zero- and Few-Shot								
NVS [31] (7-shot)	63.8	36.4	-	-	-	-	-	-
NVS [31] (50-shot)	65.5	39.8	-	-	-	-	-	-
3D-DST [14] (0-shot)	82.3	65.4	-	-	-	-	-	-
DINeMo (ours) (0-shot)	**92.8**	**78.6**	87.9	68.1	73.7	51.5	43.9	23.1

Table 1. **3D object pose estimation on the car split of Pascal3D+ [32] and occluded PASCAL3D+ [29].** Our DINeMo outperforms previous zero- and few-shot 3D pose estimation methods by a wide margin, narrowing the gap with fully-supervised methods by 67.3%.

Methods	PCK@0.1
DINOv2-ViT-S/14[20]	48.4
DINOv2-ViT-B/14[20]	52.8
DIFT [27]	48.3
SD-DINO[35]	53.8
Telling Left from Right [36]†	60.8
DINeMo (ours)	**59.1**

Table 2. **Semantic correspondence evaluation on car split of SPair71k [16].** The metric is *per point* PCK, following previous works [35, 36]. Our DINeMo outperforms all previous methods by a wide margin and achieves comparable performance with [36] that use index to flip source keypoints at test time.

Figure 4. **Scaling properties of DINeMo.** See Sec. 4.2.

methods and fully-supervised methods by 67.3% (from previous 15.6% to 5.1%). Lastly our DINeMo demonstrate enhanced robustness to partial occlusion, largely outperforming fully-supervised ResNet50 baseline and falling behind fully-supervised NOVUM only by a small gap.

Semantic Correspondence We evaluate our model on the car split of the SPair71k dataset [16] using the PCK@0.1 metric. Results show that our model achieve a significant improvement over previous works, *e.g.* by 10.7% compared to the DINOv2-ViT-S/14 backbone that our method builds on and by 5.3% compared to SD-DINO [35]. Our method also achieves comparable performance with Telling Left from Right [36], which utilize extra information to flip source keypoints at test time.

4.2. Scaling Properties

As our DINeMo is trained on object images without 3D annotations, we study the scaling properties w.r.t. different training data sizes. Specifically, we trained a variety of DINeMo models on different numbers of unlabeled images, ranging from 2048 images (comparable to the training set size in PASCAL3D+ [32]) up to 15,000 images (as in Stanford Cars dataset [12]). As shown from the results in Fig. 4,
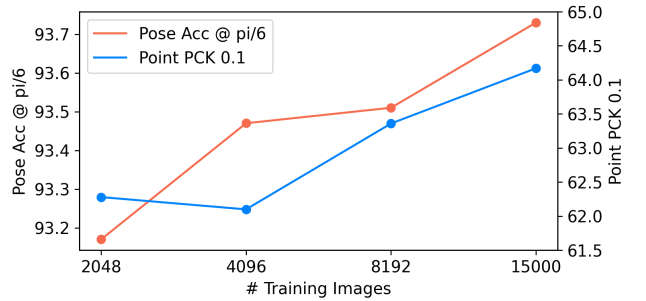
DINeMo scales well with more unlabeled images used during training, *i.e.*, the pose accuracy at $\pi/6$ increases from 93.1% to 93.7% and the per point PCK increases from 62.3 to 64.2. This highlights the advantages over previous supervised learning methods — DINeMo does not require 3D annotations that are hard to obtain, and scales effectively and efficiently by involving more unlabeled images for training, which are abundant on the Internet [25].

5. Conclusions

In this work we present DINeMo, a novel neural mesh model that is trained with no 3D annotations-only from pseudo-correspondence obtained from large visual foundation models. We propose a novel bidirectional pseudo-correspondence generation method that can effectively utilize both local appearance features and global context information to produce more 3D-consistent pseudo-correspondence. Experimental results on car datasets demonstrate that our DINeMo outperforms previous zero- and few-shot 3D pose estimation methods by a wide margin, narrowing the gap with fully-supervised methods by 67.3%. By incorporating more unlabeled images during training, our DINeMo also scales effectively and efficiently, demonstrating the advantages over fully supervised methods that rely on scarce 3D annotations.

6. Acknowledgements

References

[1] Garrick Brazil, Abhinav Kumar, Julian Straub, Nikhila Ravi, Justin Johnson, and Georgia Gkioxari. Omni3D: A large benchmark and model for 3D object detection in the wild. In *CVPR*, Vancouver, Canada, 2023. IEEE. 1

[2] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020. 1

[3] Ulf Grenander. A unified approach to pattern analysis. In *Advances in computers*, pages 175–216. Elsevier, 1970. 2

[4] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild, 2018. 2

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4, 1

[6] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019. 1

[7] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1325–1339, 2013. 1

[8] Shun Iwase, Xingyu Liu, Rawal Khirodkar, Rio Yokota, and Kris M. Kitani. Repose: Fast 6d object pose refinement via deep texture rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3303–3312, 2021. 2

[9] Artur Jesslen, Guofeng Zhang, Angtian Wang, Wufei Ma, Alan Yuille, and Adam Kortylewski. Novum: Neural object volumes for robust object classification. In *European Conference on Computer Vision*, pages 264–281. Springer, 2024. 1, 2, 3, 4

[10] Adam Kortylewski, Ju He, Qing Liu, and Alan L Yuille. Compositional convolutional neural networks: A deep architecture with innate robustness to partial occlusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8940–8949, 2020. 1

[11] Adam Kortylewski, Qing Liu, Huiyu Wang, Zhishuai Zhang, and Alan Yuille. Combining compositional models and deep networks for robust object classification under occlusion. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1333–1341, 2020. 2

[12] Jonathan Krause, Hailin Jin, Jianchao Yang, and Li Fei-Fei. Fine-grained recognition without part annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5546–5555, 2015. 2, 4

[13] Wufei Ma, Angtian Wang, Alan Yuille, and Adam Kortylewski. Robust category-level 6d pose estimation with coarse-to-fine rendering of neural features. In *European Conference on Computer Vision*, pages 492–508. Springer, 2022. 1, 2, 3

[14] Wufei Ma, Qihao Liu, Jiahao Wang, Angtian Wang, Xiaoding Yuan, Yi Zhang, Zihao Xiao, Guofeng Zhang, Beijia Lu, Ruxiao Duan, et al. Generating images with 3d annotations using diffusion models. *arXiv preprint arXiv:2306.08103*, 2023. 1, 3, 4

[15] Wufei Ma, Guofeng Zhang, Qihao Liu, Guanning Zeng, Adam Kortylewski, Yaoyao Liu, and Alan L Yuille. Imagenet3d: Towards general-purpose object-level 3d understanding. *Advances in Neural Information Processing Systems*, 37:96127–96149, 2024. 1

[16] Juhong Min, Jongmin Lee, Jean Ponce, and Minsu Cho. Spair-71k: A large-scale benchmark for semantic correspondence. *arXiv preprint arXiv:1908.10543*, 2019. 2, 4, 1

[17] Arsalan Mousavian, Dragomir Anguelov, John Flynn, and Jana Kosecka. 3d bounding box estimation using deep learning and geometry. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 7074–7082, 2017. 2

[18] Ulric Neisser et al. *Cognitive Psychology*. Appleton-Century-Crofts, 1967. 1

[19] Natalia Neverova, Artsiom Sanakoyeu, Patrick Labatut, David Novotny, and Andrea Vedaldi. Discovering relationships between object categories via universal canonical maps, 2021. 2

[20] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 1, 3, 4

[21] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 3

[22] Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J. Guibas. Humor: 3d human motion model for robust pose estimation, 2021. 2

[23] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, Zhaoyang Zeng, Hao Zhang, Feng Li, Jie Yang, Hongyang Li, Qing Jiang, and Lei Zhang. Grounded sam: Assembling open-world models for diverse visual tasks, 2024. 1

[24] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 3

[25] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022. 4

[26] Aleksandar Shtedritski, Christian Rupprecht, and Andrea Vedaldi. Shic: Shape-image correspondences with no keypoint supervision. In *ECCV*, 2024. 2

[27] Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 4, 1

[28] Shubham Tulsiani and Jitendra Malik. Viewpoints and keypoints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1510–1519, 2015. 2

[29] Angtian Wang, Yihong Sun, Adam Kortylewski, and Alan L Yuille. Robust object detection under occlusion with context-aware compositionalnets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12645–12654, 2020. 3, 4, 1

[30] Angtian Wang, Adam Kortylewski, and Alan Yuille. Nemo: Neural mesh models of contrastive features for robust 3d pose estimation. *arXiv preprint arXiv:2101.12378*, 2021. 1, 2, 3

[31] Angtian Wang, Shenxiao Mei, Alan L Yuille, and Adam Kortylewski. Neural view synthesis and matching for semi-supervised few-shot learning of 3d pose. *Advances in Neural Information Processing Systems*, 34:7207–7219, 2021. 3, 4, 1

[32] Yu Xiang, Roozbeh Mottaghi, and Silvio Savarese. Beyond pascal: A benchmark for 3d object detection in the wild. In *IEEE winter conference on applications of computer vision*, pages 75–82. IEEE, 2014. 1, 3, 4

[33] Jiahao Yang, Wufei Ma, Angtian Wang, Xiaoding Yuan, Alan Yuille, and Adam Kortylewski. Robust category-level 3d pose estimation from synthetic data. *arXiv preprint arXiv:2305.16124*, 2023. 2

[34] Alan Yuille and Daniel Kersten. Vision as bayesian inference: analysis by synthesis? *Trends in cognitive sciences*, 10(7):301–308, 2006. 1

[35] Junyi Zhang, Charles Herrmann, Junhwa Hur, Luisa Polania Cabrera, Varun Jampani, Deqing Sun, and Ming-Hsuan Yang. A tale of two features: Stable diffusion complements dino for zero-shot semantic correspondence. *Advances in Neural Information Processing Systems*, 2023. 1, 3, 4

[36] Junyi Zhang, Charles Herrmann, Junhwa Hur, Eric Chen, Varun Jampani, Deqing Sun, and Ming-Hsuan Yang. Telling left from right: Identifying geometry-aware semantic correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3076–3085, 2024. 4, 1

[37] Yi Zhang, Pengliang Ji, Angtian Wang, Jieru Mei, Adam Kortylewski, and Alan L Yuille. 3d-aware neural body fitting for occlusion robust 3d human pose estimation. *ICCV*, 2023. 1

[38] Bingchen Zhao, Shaozuo Yu, Wufei Ma, Mingxin Yu, Shenxiao Mei, Angtian Wang, Ju He, Alan Yuille, and Adam Kortylewski. Ood-cv: A benchmark for robustness to out-of-distribution shifts of individual nuisances in natural images. In *European conference on computer vision*, pages 163–180. Springer, 2022. 1

[39] Xingyi Zhou, Arjun Karpur, Linjie Luo, and Qixing Huang. Starmap for category-agnostic keypoint and viewpoint estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 318–334, 2018. 2

DINeMo: Learning Neural Mesh Models with no 3D Annotations

Supplementary Material

7. Author Contribution Statement

WG conducted the experiments. WM and GZ led the study and wrote the paper. AY provided high-level conceptual guidance and feedback.

8. Experimental Setup

Benchmarks For 3D pose estimation, we follow previous works [9, 13] and evaluate our DINeMo on the car split of PASCAL3D+ [32] for in-distribution testing and on the car split of occluded PASCAL3D+ [29] for partial occlusion generalization. For semantic correspondence, we evaluate our DINeMo on the car split of the SPair71k dataset[16].

Baselines For 3D pose estimation, we consider two types of baselines: (i) *fully-supervised models*, *e.g.*, ResNet50 [5] and NOVUM [9], which serve as a reference to the state-of-the-art performance by utilizing all 3D pose annotations in PASCAL3D+ [32]; (ii) *zero- and few-shot models*, *i.e.*, NVS [31] and 3D-DST [14]. Specifically, NVS advanced label efficient training by synthesizing feature maps from novel views and obtain correspondences for unlabeled images. 3D-DST proposed to synthesize images with 3D annotations using diffusion models, improving zero-shot performance and enhancing model robustness.For semantic correspondence, we consider DINOv2[20] and Diffusion features[27], as well as the unsupervised version of SD-DINO[35] and Telling Left from Right[36].
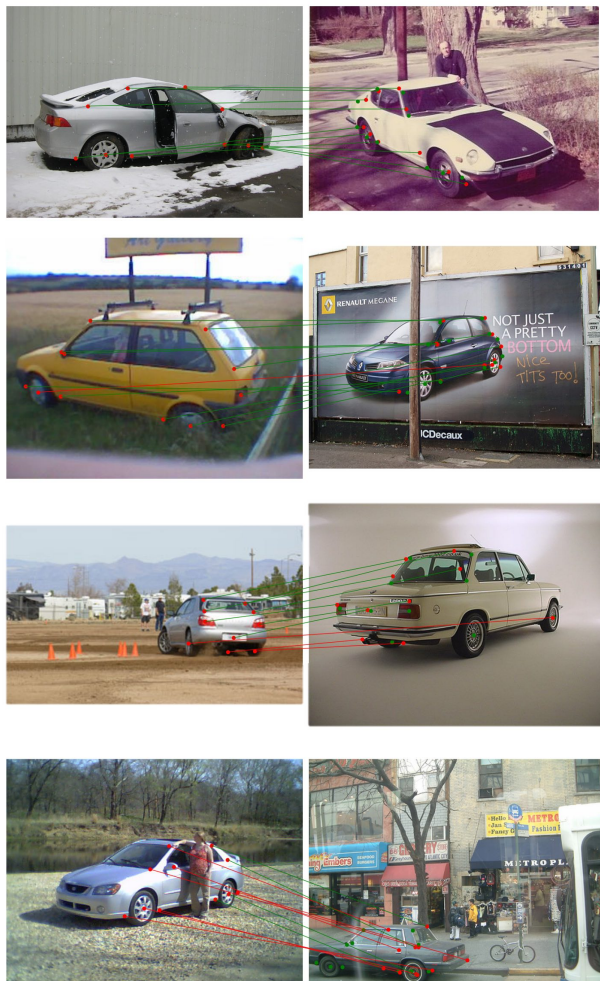
9. Qualitative Examples

We present some qualitative comparisons between DINOv2 and our DINeMo in Fig. 5. We also present some qualitative examples of 3D pose estimation of our DINeMo in Fig. 6.

10. Public Release

Code. All code of our DINeMo will be made available upon acceptance of the paper.

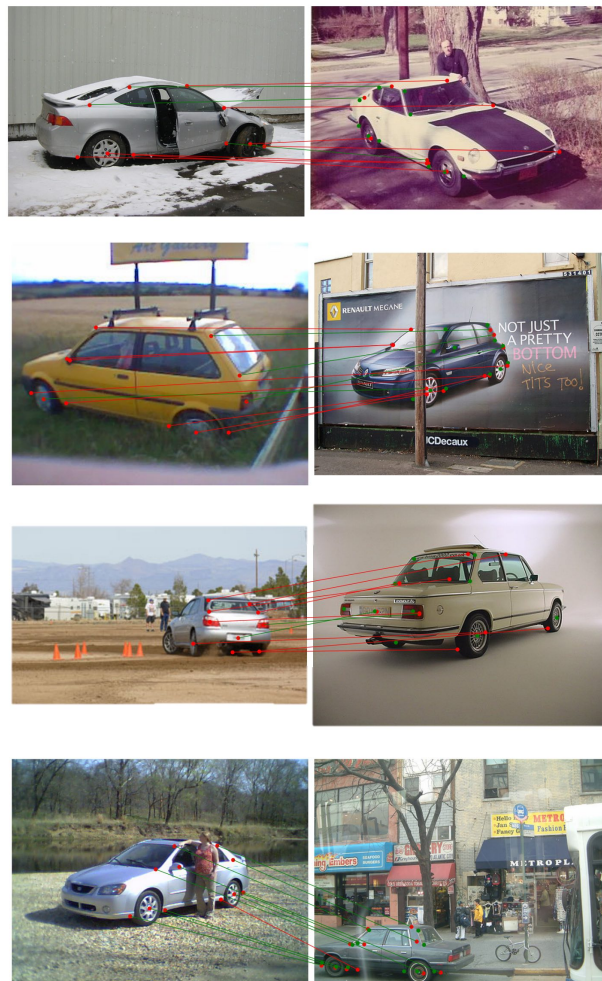DINeMo (Ours)   DINOv2

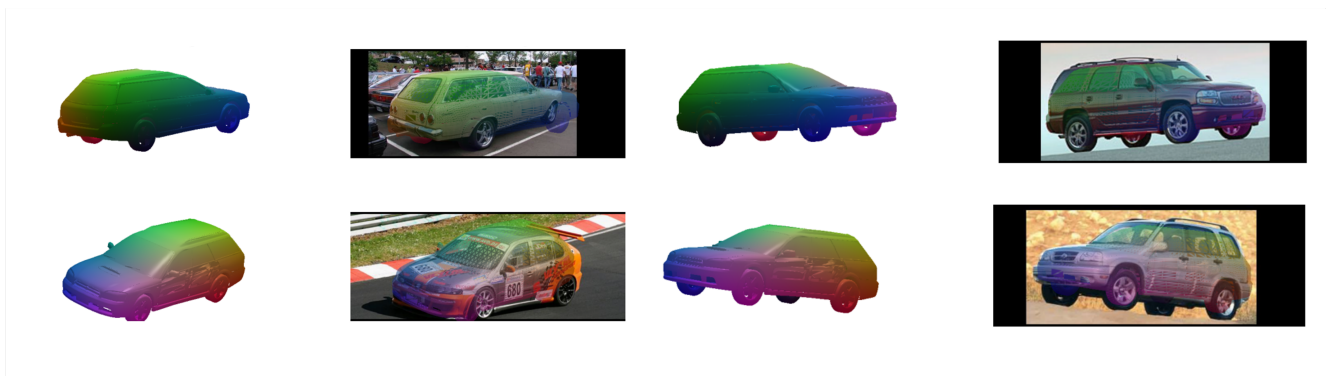Figure 5. **Qualitative comparisons between DINOv2 (left) and our DINeMo (right) on the SPair71k[16] dataset.**

Figure 6. **Qualitative pose estimation results on the Pascal3D+[32] dataset.**