Back to 3D: Few-Shot 3D Keypoint Detection Using Back-Projected 2D Features

Thomas Wimmer[1,2], Peter Wonka[3], Maks Ovsjanikov[1]

[1]LIX, École Polytechnique, IP Paris, France [2]Technical University of Munich, Germany [3]KAUST, Saudi Arabia

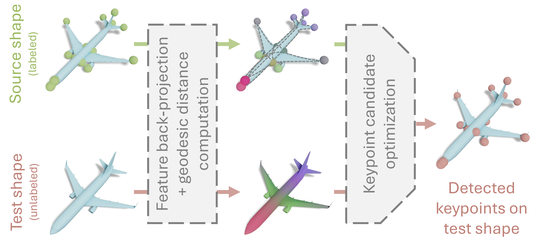
CVPR JUNE 17-21, 2024 SEATTLE, WA

Motivation

Keypoint detection requires localized understanding of geometric details, as well as of global semantics. Foundation models have shown strong generalization for various 2D vision tasks. Can we leverage these powerful priors for 3D keypoint detection?

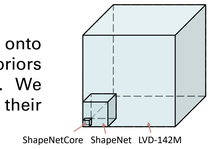
Problem setting: Given one or a few labeled source shapes, how can we detect the same set of keypoints on an unlabeled test shape?

Overall Pipeline



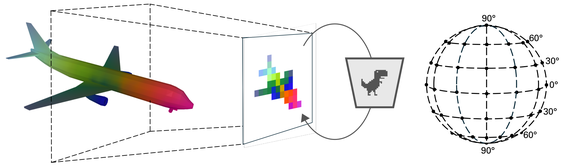
Previous approaches rely on axiomatic shape descriptors or 3D feature extractors pre-trained on (small) 3D datasets. We hypothesize that neither of these features can sufficiently capture the semantics needed for zero- or few-shot applications on unseen data distributions.

We propose to back-project 2D features onto 3D shapes to leverage the powerful priors learned in 2D for 3D shape analysis. We examine the resulting features for their geometric properties and stability.



To prevent "collapsed" solutions in the presence of symmetries, where multiple symmetric keypoints are detected at the same location on the test shape, we propose a **geodesic distance-based keypoint optimization**.

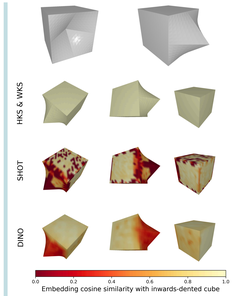
Back-Projection of 2D Features



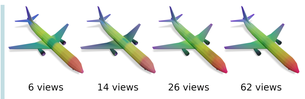
Gaussian Geodesic Re-Weighting of Features

Point visibility information is noisy for complex meshes. We propose a Gaussian smoothing of the back-projected features along the shape's surface to resolve this issue:

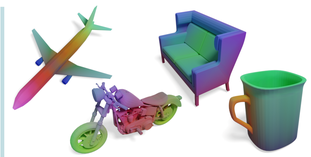
Feature Properties



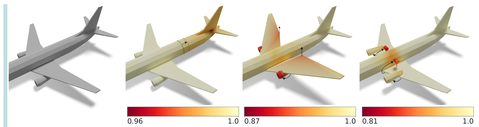
Compared to axiomatic shape descriptors, back-projected features provide a more intuitive understanding of shape changes.



Increase in feature quality and distinctiveness with increasing number of rendering viewpoints.



Extracted features contain rich semantic information and clearly differ for different semantic parts of the objects. Visualization using a PCA.



Extracted features contain rich semantic information and clearly differ for different semantic parts of the objects. Visualization using a PCA.

Keypoint Optimization

To capitalize on the rich semantic features, while preventing "collapsed" solutions due to symmetries, we propose an optimization of keypoint locations that aims to match the global distribution of keypoints on the test shape, alongside their features.

Optimize right-stochastic assignment matrix:

$S \in [0,1]^{n\times(k+1)}$ with $\hat{S} = S_{[1,\dots,n;1,\dots,k]}$

Back-projected features:

$F_{kp} \in \mathbb{R}^{k\times d_{emb}}$

$F_{cand} \in \mathbb{R}^{n\times d_{emb}}$

Pairwise geodesic distances:

$D_{kp} \in \mathbb{R}^{k\times k}$

$D_{cand} \in \mathbb{R}^{n\times n}$

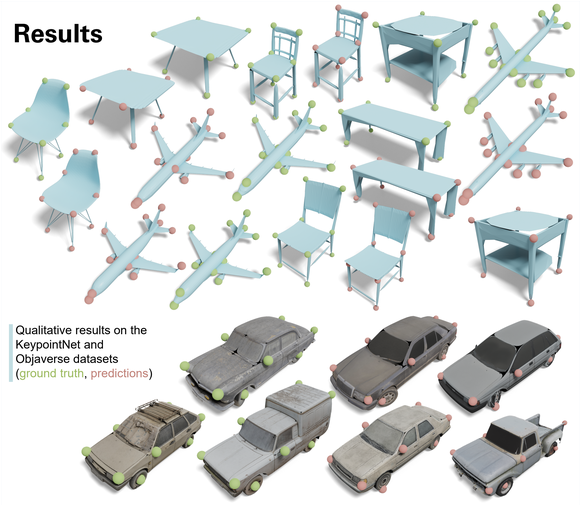
Optimization Objective:

$$L = L_{feature}$$

$$L_{feature} = \left\| \hat{S}^T F_{cand} - F_{kp} \right\|$$

$$L_{distance} = \left\| \hat{S}^T D_{cand} \hat{S} - D_{kp} \right\|$$

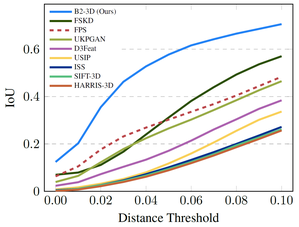
Results



Qualitative results on the KeypointNet and Objaverse datasets (ground truth, predictions)

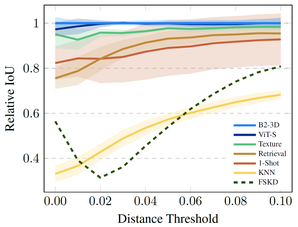
Results

Our method B2-3D outperforms previous methods by a large margin, almost doubling the performance of the previous state-of-the-art[1] on the KeypointNet dataset[2] for 3-shot keypoint detection.



Ablations

- B2-3D works well even in a 1-shot setting.
- Uniform coloring improves results compared to low-quality texture information.
- The keypoint optimization module avoids collapsed solutions (cf. [KNN]).
- DINO[3] is the best pre-trained 2D feature extractor.



Conclusions

- Large pre-trained 2D models can act as powerful priors for 3D shape analysis and enable zero- or few-shot applications.
- Back-projected features are learning-free and carry strong semantic information while being sufficiently geometry-aware.
- A simple geodesic distance-based optimization can successfully resolve symmetries.

References:

[1] Attaiki, Souhaib, and Maks Ovsjanikov. (NeurIPS, 2022)
"NCP: Neural correspondence prior for effective unsupervised shape matching."

[2] You, Yang, et al. (CVPR, 2020)
"Kpeointnet: A large-scale 3D keypoint dataset aggregated from numerous human annotations."

[3] Oquab, Maxime, Darcet, Timothée, Moutakanni, Théo, et al. (TMLR, 2024)
"DINOv2: Learning robust visual features without supervision."

Dutt, Niladri Shekhar, et al. (CVPR, 2024)
"Diffusion 3D Features (Diff3F): Decorating untextured shapes with distilled semantic features."

Banani, Mohamed El, et al. (arXiv, 2024)
"Probing the 3D awareness of visual foundation models."

Acknowledgements:
Thomas Wimmer is supported by the Konrad Zuse School of Excellence in Learning and Intelligent Systems (ELIZA) through the DAAD programme Konrad Zuse Schools of Excellence in Artificial Intelligence, sponsored by the German Federal Ministry of Education and Research. Parts of this work were supported by the ERC Starting Grant 758800 (EXPROTEA), ERC Consolidator Grant 101087347 (VEGA), ANR AI Chair AIGRETTE, and gifts from Adobe Inc. and Ansys Inc.