2D-LFM: Lifting Foundation Model without 3D Supervision

Supplementary Material

1. Overview

In this supplementary document we provide additional details and results complementing the main paper:

- **Implementation details** covering the training details, sampling strategy, and learning hyperparameters used for large-scale training (Sec. 2).
- **Ablation studies** analyzing architectural and optimization choices (Sec. 3).
- **Extended experimental results** (Sec. 4), including scaling analysis on 200+ categories, and qualitative comparisons against state-of-the-art baselines.
- **Limitations and failure cases** of the proposed approach (Sec. 5).

Key Findings Our experiments reveal that scaling to hundreds of categories unlocks cross-category geometric transfer, where performance on underrepresented classes improves significantly by leveraging structural priors learned from data-rich categories. We find that the choice of positional encoding is critical for this scalability: while Graph-Laplacian encodings degrade as model depth and category diversity increase, our analytical RFF encodings enable stable transfer. This is best exemplified by the *Drosophila* category [3]; despite being a data-starved outlier, the model successfully exploits multi-view cues inherent in the training data to achieve near-perfect reconstruction, effectively amortizing the multi-view structure-from-motion constraints into a single forward pass.

2. Implementation and Training Details

2.1. Splits and Balanced Sampling.

We perform a randomized 80/20 train-validation split for each category. To mitigate the extreme class imbalance in our dataset - ranging from singleton classes (e.g., *Drosophila*) to massive motion-capture datasets, we employ a balanced sampling strategy. We define the epoch size based on the most populous category: let N_c be the sample count for category c, and $N_{\max} = \max_c N_c$. We oversample all smaller categories with replacement so that every category effectively contributes $N_{\max}$ samples per epoch. This ensures that gradient updates are equally distributed across diverse geometries, preventing the model from suffering "structural collapse" into the mean shape of the most common category.

2.2. Data Preprocessing and Normalization.

To ensure geometric invariance across heterogeneous domains ranging from microscopic insects to large automobiles, we normalize all 2D input landmarks and 3D ground truth structures to a canonical scale. Given a set of landmarks $\mathbf{X} \in \mathbb{R}^{N\times D}$ (where $D = 2$ or 3), we first center the shape at the origin and then scale it to unit Frobenius norm:

$$\mathbf{X}_{\text{norm}} = \frac{\mathbf{X} - \bar{\mathbf{X}}}{\|\mathbf{X} - \bar{\mathbf{X}}\|_F}$$

where $\bar{\mathbf{X}}$ is the centroid of the points. This normalization is applied consistently during training and inference, allowing the model to focus purely on structural connectivity and deformation rather than absolute metric scale.

2.3. Training Regimes and Composition.

We evaluate our approach in two distinct regimes to demonstrate scalability:

- **Standard Regime (48 Categories):** This configuration serves as our primary benchmark for ablation and comparison. It includes:
  - *Pascal3D+ Objects [8] (12 cats):* Aeroplane, bicycle, boat, bottle, bus, car, chair, dining table, motorbike, sofa, train, and TV monitor.
  - *Arctic Hand-Object [1] (12 cats):* Espresso machine, ketchup, microwave, mixer, scissors, waffle iron (plus left/right hand interactions for each).
  - *Animal [2, 9] & Insect [3] (24 cats):* A diverse collection including bear, cow, deer, dog, duck, fox, hippo, moose, puma, rabbit, raccoon, tiger, and notably *Drosophila* (fruit fly).
- **Foundation Regime (200+ Categories):** To test the limits of geometric transfer, and to substantiate the large-scale scaling trend projected in Figure 2 of the main paper, we scale the training set by incorporating the CUB-200-2011 Bird dataset [6] (193 fine-grained species) and massive human motion datasets (e.g., *AMASS* [5], *Human3.6M* [4]). In this regime, $N_{\max}$ is determined by the millions of frames available in the human datasets. Consequently, the effective epoch size expands by orders of magnitude compared to the standard regime. To maintain training feasibility, we define an "epoch" as a fixed number of iterations sampled from this balanced distribution, observing that the increased geometric diversity leads to faster convergence in wall-clock time.

2.4. Optimization and Hyperparameters

We observe distinct optimization dynamics depending on the dataset scale. Bulleted list 2.3 summarizes our data configuration.

2.5. Optimizer and Schedule.

We tailor the optimization strategy to the data regime. For the 48-category experiments, we found standard **Adam** outperformed AdamW (achieving lower validation MPJPE), likely because the smaller effective dataset size did not require aggressive weight-decay regularization. Conversely, for the Foundation Regime (232 categories), we transition to **AdamW** to prevent overfitting given the massive increase in training iterations.

2.6. Learning Rate Schedule.

Contrary to common practice in transformer training, we observed that aggressive warm-up and annealing schedules (e.g., OneCycleLR) led to early saturation in validation performance. We hypothesize that learning a universal geometric prior across multiple heterogeneous categories creates a high-variance loss landscape; rapid annealing forces the model to converge prematurely to local minima dominated by data-rich classes, ignoring the finer structure of underrepresented categories.

Consequently, we employ a **constant learning rate** schedule (1×10^{-5} for the foundation regime). We found that maintaining a sustained, stable step size was more effective than annealing, likely because the model requires continuous plasticity to balance the competing gradients of rigid (Pascal) and organic (Deformable) geometries throughout the entire training duration.

2.7. Compute Resources.

All models are trained with a per-GPU batch size of 64, over NVIDIA A100. The standard regime converges in less than 24 hours, while the foundation regime requires approximately 3-5 days to reach convergence (equivalent to $\sim$50 effective epochs).

3. Ablation Studies

We validate our architectural choices through controlled ablations on the Standard Regime (48 categories). Unless otherwise stated, all models use the default 16-layer transformer configuration.

3.1. Choice of Positional Encoding Function

We investigate the impact of the positional encoding (PE) function $\Phi(x)$. We compare our Random Fourier Features (RFF) against learnable embeddings (Graphical/Diffusion-style) and the standard sinusoidal encoding used in Vision Transformers (ViT).

Frequency Resolution. Standard ViT encodings use a broad frequency spectrum tailored for dense pixel grids: $\omega_k = 10000^{-2k/D}$. For sparse geometric landmarks, we hypothesized that a narrower bandwidth is required to distinguish nearby points. We test a "Modified ViT" encoding with increased frequency density:

$$\omega_k = 100^{-2k/D} \tag{1}$$

Table 1. **Impact of Positional Encoding Function.** MPJPE (mm) reported on Pascal3D+ validation set. RFF significantly outperforms graph laplacian and standard sinusoidal baselines.

Encoding Variant	Pascal3D	Model Config
ViT (Original)	92.3	16L, 256D
ViT (Modified Bandwidth)	54.0	16L, 256D
Graphical	40.0	16L, 256D
RFF (Ours)	**8.1**	16L, 256D

Table 1 confirms that standard ViT encodings fail to preserve precise geometry (92.3mm error). While the modified bandwidth improves performance (54.0mm), our analytical RFF encoding achieves significantly lower error (8.1mm). This suggests that preserving high-frequency spatial correspondence is critical for lifting, and RFFs provide the most robust correspondence conditions for this task.

3.2. Injection Strategy and Model Depth

We find that correspondence must be reinforced at *every* layer. We compare our "Every Layer" injection against standard input-only injection and intermediate strategies (e.g., alternating layers). Table 2 reveals a critical interaction between injection strategy and model depth:

Signal Dissipation: With standard "Input Only" injection, deeper models (16L) perform *worse* than shallow ones (4L) (100mm vs. 40mm). The correspondence signal "washes out" after multiple self-attention layers, causing the model to lose track of point identity.

Signal Preservation: By injecting PE at every layer, we reverse this trend. Deeper models (16L) now significantly outperform shallow ones (8.1mm vs. 17.4mm), effectively utilizing the increased capacity for geometric reasoning.

This result validates our architectural choice: to scale lifting transformers to foundation size (depth), continuous correspondence injection is a prerequisite.

3.3. Model Capacity and Optimization Sensitivity

Finally, we evaluate the impact of scaling model width and optimization stability within our best-performing "Every Layer" injection framework. Table 3 summarizes results on the 48-category benchmark. We find that increasing hidden dimension from 256D to 512D yields consistent gains (27.0mm $\rightarrow$ 25.0mm), and deeper models (20L) further reduce error to 21.9mm. Crucially, we observe that lower learning rates (1×10^{-5}) are essential for stability when scaling model size, preventing the early saturation observed at higher rates (1×10^{-4}).

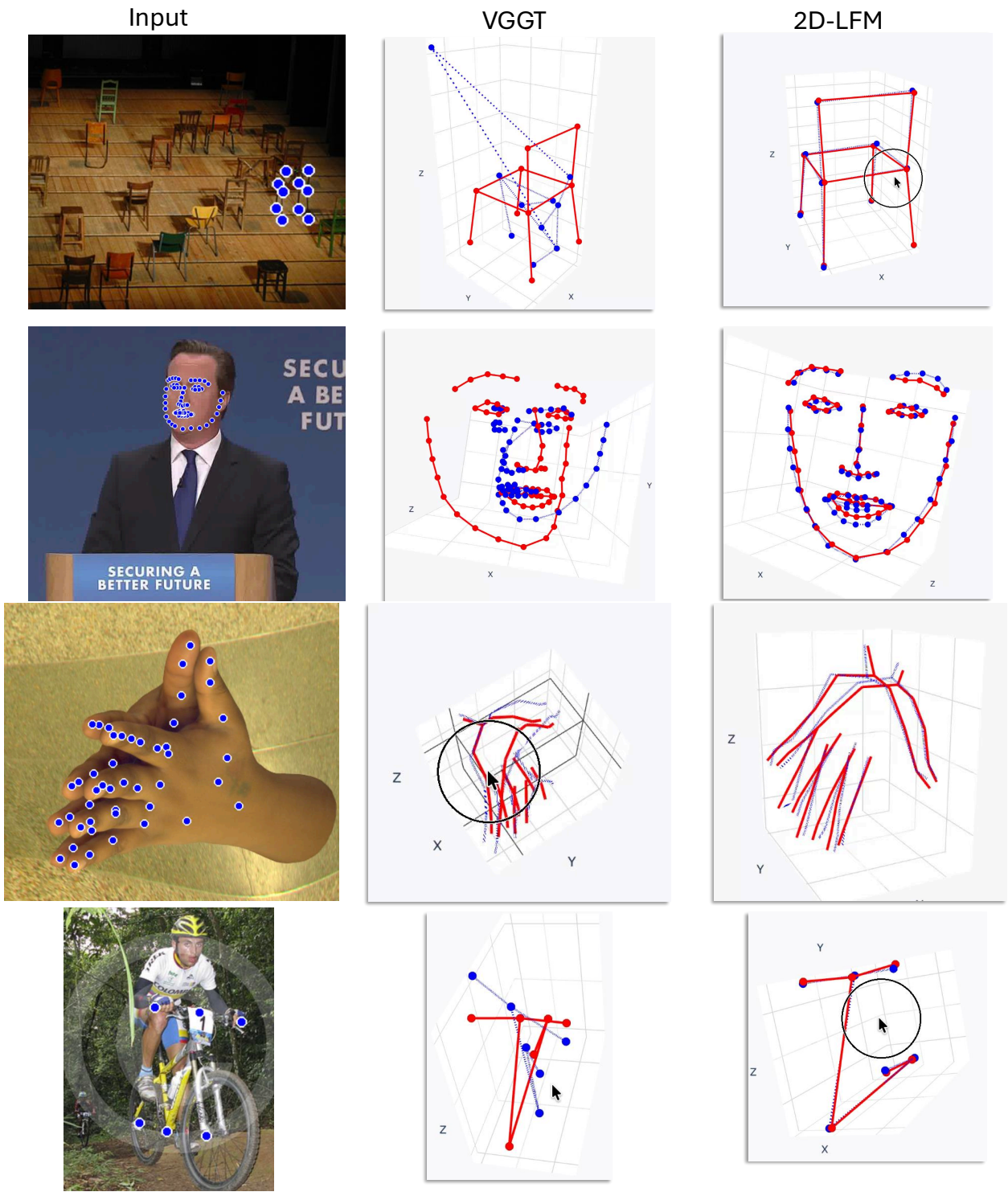

Figure 2. Additional qualitative comparisons between VGGT (middle) and our method (right) on categories where the scaled-orthographic assumption approximately holds. Our reconstructions better capture object-level landmark geometry in a wide range of categories.

4. Extended Experimental Results

We present a preliminary analysis of our large-scale training regime. The primary goal of this experiment is not to es-

Table 2. **Interaction between Injection Strategy and Model Architecture.** We compare "Deep & Narrow" (16L, 256D) vs. "Shallow & Wide" (4L, 1024D) models with approximately equal parameter counts. Without per-layer injection, deep models fail. With per-layer injection, deep models achieve the best performance.

Injection Strategy	Configuration	Pascal3D	Human3.6M
Input Only	Deep (16L, 256D)	100.0	43.2
	Wide (4L, 1024D)	40.0	36.7
First & Last	Deep (16L, 256D)	92.3	71.9
	Wide (4L, 1024D)	16.0	38.0
Even Layers	Deep (16L, 256D)	26.1	34.5
	Wide (4L, 1024D)	58.9	38.3
Every Layer (Ours)	**Deep (16L, 256D)**	**8.1**	**33.9**
	Wide (4L, 1024D)	17.4	38.2

Table 3. **Capacity and Optimization Sweep.** MPJPE (mm) on the 48-category regime. All models use "Every Layer" PE injection. Increasing capacity (depth/width) consistently improves performance, provided the learning rate is sufficiently low to handle the optimization variance.

Optimization	Configuration	MPJPE (mm)
Adam, LR 1×10^{-4}	16L, 256D	36.0
Adam, LR 1×10^{-5}	16L, 256D	27.0
Adam, LR 1×10^{-5}	16L, 512D	25.0
Adam, LR 1×10^{-5}	**20L, 512D**	**21.9**

tablish a converged benchmark, but to demonstrate that our **2D-LFM architecture is capable of scaling** to hundreds of heterogeneous categories without architectural modification or divergence.

4.1. Feasibility of Foundation-Scale Training

We scaled the training set from the standard 48 categories to the full foundation regime (232 categories). This introduces a massive increase in semantic and geometric diversity, including humans, hands, and hundreds of fine-grained bird species. We present results from our large-scale training regime, demonstrating the scalability of the 2D-LFM approach and its ability to generalize across heterogeneous geometric domains. We observe three key **observations**:

- **Architectural Robustness.** Training transformers on sparse geometric data often leads to instability (loss spikes) when data diversity is high. We observe that our per-layer PE injection maintains stable training dynamics even when the batch contains highly disparate geometries (e.g., *Cars* mixed with *Insects*). This validates that the architecture possesses the capacity to model a universal geometric prior.
- **Geometric Transfer to Singleton Classes.** Data-starved categories benefit disproportionately from scaling. We hypothesize that the model learns universal geometric constraints (e.g., rigidity, connectivity, and rotation rules) from data-rich categories like *Cars* and *Birds*, and transfers these priors to the insect domain.

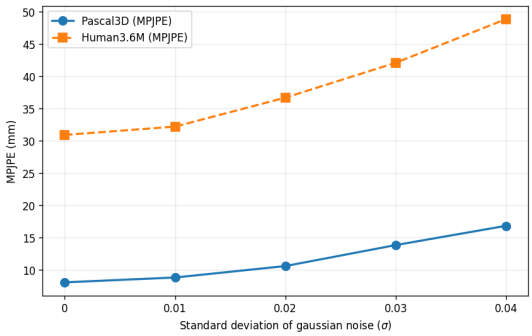

Figure 3. **Robustness to 2D keypoint noise.** MPJPE degrades gracefully as Gaussian noise σ increases, with no catastrophic failure.

4.2. Case Study: Drosophila and SfM Amortization

The improvement on the *Drosophila* category serves as a critical proof-of-concept for our method. The training data for this category consists of unannotated 2D landmarks from multiple viewpoints, but lacks 3D ground truth.

- **Single-Category Failure:** A model trained *only* on Drosophila achieves 23.4mm error, likely overfitting to 2D deformations rather than learning 3D structure.
- **Foundation Success:** The foundation model achieves 1.8mm error. By observing millions of rotating rigid bodies (cars) and articulated skeletons (humans) in the training set, the model effectively learns to solve the *Structure-from-Motion (SfM)* problem. At inference time, it recognizes the 2D landmark configuration as a specific 3D view of a rigid structure, effectively amortizing the SfM optimization into a single forward pass.

4.3. Robustness to Noisy (Real-World) Data

We evaluate robustness in two settings: (i) synthetic noise injection and (ii) off-the-shelf 2D detections.

Gaussian noise. We inject i.i.d. zero-mean Gaussian noise into normalized 2D keypoints at test time so that σ is comparable across categories. As shown in Fig. 3, MPJPE degrades smoothly and monotonically - small noise yields proportionally small error increases - indicating stable, non-catastrophic behavior. Performance remains competitive with the 2D-supervised baselines in Table 1 even under substantial noise; for reference, $\sigma = 0.04$ corresponds to $\sim$9 pixels jitter on a 224×224 crop, yet performance remains well below the naive transformer baseline ($>$90 mm).

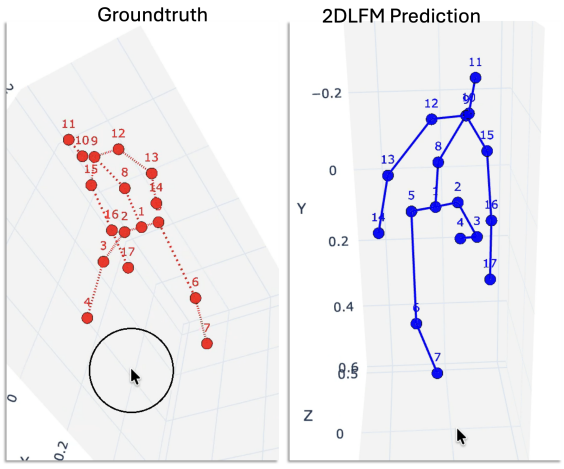

Figure 4. **Single-View Depth Ambiguity (Necker Reversal).** A typical failure mode in unsupervised lifting. The model predicts a skeleton (Blue) that is a depth-inverted version of the Ground Truth (Red). While both skeletons project to nearly identical 2D landmarks (bottom row), the 3D structure is flipped, illustrating the inherent ambiguity of monocular lifting without temporal cues.

Off-the-shelf ViTPose detections. To assess real-world applicability, we run ViTPose [10] on Human3.6M and feed the detected 2D keypoints directly into 2D-LFM. 2D-LFM achieves 34.4 mm MPJPE with ViTPose detections, compared to 30.9 mm with ground-truth keypoints - a modest increase consistent with the $\sigma \approx 0.01$ regime in Fig. 3, confirming that the model is robust to realistic upstream detection noise.

5. Limitations and Failure Cases

While our approach shows strong generalization, it inherits limitations inherent to unsupervised lifting and the specific training dynamics of foundation-scale models:

Single-View Depth Ambiguity. Relying on monocular 2D input creates an inherent ambiguity where a 3D structure and its depth-inverted counterpart project to identical 2D landmarks (the Necker reversal). Without temporal cues or 3D supervision to break this symmetry, the model may occasionally flip the depth of articulated parts: for example, reconstructing a forward-facing human as facing backward, or inverting the articulation of limbs relative to the camera. Figure 4 illustrates this failure mode: the model reconstructs a valid 3D skeleton (Blue) that is the depth-inverted reflection of the ground truth (Red), despite both having near-perfect 2D alignment.

Structural Outliers (The "Crocodile" Problem). Cross-category transfer relies on shared geometric priors (e.g., quadruped limb articulation). Categories that are both underrepresented and structurally unique can be effectively "washed out" by the dominant priors. For example, we observe high error on the *Crocodile* class (MPJPE $\approx$ 162mm); its long, straight tail is a distinct structural feature not shared by other animals. The model struggles to lift this unique topology from scratch while simultaneously satisfying the strong priors learned from other categories that discourage such linear structures.

References

[1] Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J Black, and Otmar Hilliges. Arctic: A dataset for dexterous bimanual hand-object manipulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12943–12954, 2023. 1

[2] Christopher Fusco, Shin-Fang Ch'ng, Mosam Dabhi, and Simon Lucey. Object agnostic 3d lifting in space and time. In *2025 International Conference on 3D Vision (3DV)*, pages 682–691. IEEE, 2025. 1

[3] Semih Günel, Helge Rhodin, Daniel Morales, João Campagnolo, Pavan Ramdya, and Pascal Fua. Deepfly3d, a deep learning-based approach for 3d limb and appendage tracking in tethered, adult drosophila. *Elife*, 8:e48571, 2019. 1

[4] 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

[5] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5442–5451, 2019. 1

[6] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 1

[7] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5294–5306, 2025. 1

[8] 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

[9] Jiacong Xu, Yi Zhang, Jiawei Peng, Wufei Ma, Artur Jesslen, Pengliang Ji, Qixin Hu, Jiehua Zhang, Qihao Liu, Jiahao Wang, et al. Animal3d: A comprehensive dataset of 3d animal pose and shape. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9099–9109, 2023. 1

[10] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose++: Vision transformer for generic body pose estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(2):1212–1230, 2023. 5