MimiCAT: Mimic with Correspondence-Aware Cascade-Transformer for Category-Free 3D Pose Transfer

Zenghao Chai[1] Chen Tang[2] Yongkang Wong[1] Xulei Yang[3] Mohan Kankanhalli[1]

[1]School of Computing, National University of Singapore [2]MMLab, The Chinese University of Hong Kong [3]Institute for Infocomm Research, A*STAR

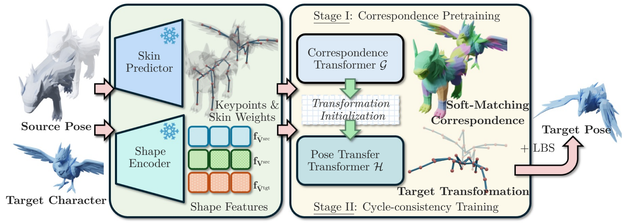
Project Page | Paper | Data

CVPR JUNE 3-7, 2026 DENVER COLORADO

Contributions

- **Task Goal:** Transfer a source pose onto a target character while preserving target geometry and pose semantics across categories. Extended 3D pose transfer to a more challenging **category-free** setting, with a cycle-consistency benchmark for evaluation.
- **Dataset Contribution:** Built **PokeAnimDB** with million-scale poses across diverse character categories, enabling large-scale and generalizable learning.
- **Technical Contribution:** Proposed **MimiCAT**, a cascade transformer that learns many-to-many soft correspondences across structurally different characters.
- **Results:** Achieved **state-of-the-art performance** on both humanoid-to-humanoid and cross-category settings, and demonstrated potential for downstream applications.

Overview of MimiCAT



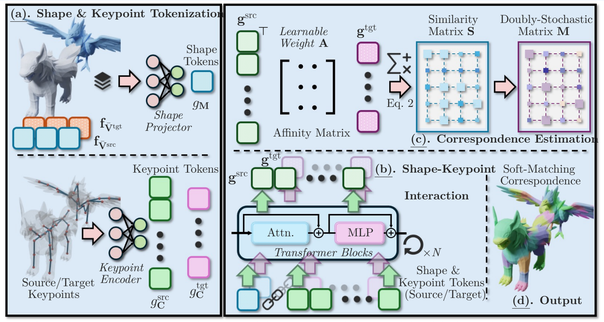
- **Pipeline:** Given the source pose and target character, MimiCAT first estimates soft correspondences, then predicts target transformations, and deforms the target via LBS.
- **Correspondence Transformer $\mathcal{G}$:** Learns a **many-to-many** soft correspondence matrix between variable-length keypoints of structurally diverse characters. We use **text-guided pseudo correspondences** from semantic keypoint labels, supervising the predicted affinity and correspondence with cosine similarity, Sinkhorn soft matching, and Hungarian hard assignment.
- **Pose Transfer Transformer $\mathcal{H}$:** The soft correspondence matrix $\mathbf{M}$ initializes target query positions and transformations by aggregating source motions, and $\mathcal{H}$ refines them with source deformation cues and target shape features; with $\mathcal{G}$ frozen, this stage is trained by **self-supervised cycle consistency** and pose-prior regularization.

PokeAnimDB Dataset

Dataset	#Char.	#Motion
Mixamo	118	2,000
AMASS	346	11,451
TruebonesZoo	70	1,000
PlanetZoo	251	251
DT4D-A	71	1,772
PokeAnimDB	**975**	**28,809**

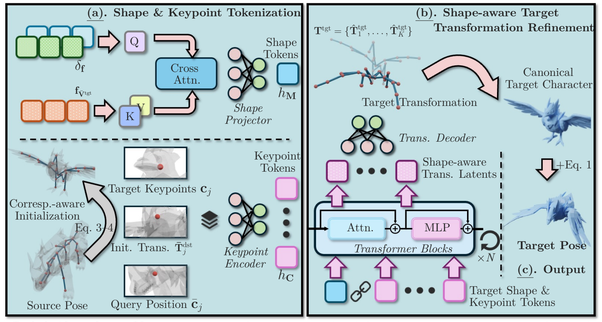
- **Scale & Diversity: 975** rigged characters, **28,809** artist-designed motions, and **4.47M** frames across humanoids, quadrupeds, birds, reptiles, fishes, and insects.
- **Semantic Labels:** Recorded bone names provide text-domain anchors across categories, e.g., shared "limb" semantics can associate human arms with bird wings.
- **Pose Prior:** A Transformer prior $\mathcal{F}$ models plausible keypoint rotations across diverse skeletons, improving realism and preventing degenerate transformations.

Correspondence Transformer $\mathcal{G}$



- **Tokenization:** Extract shape and keypoint tokens and build shared representations.
- **Interaction & Correspondence:** Transformer blocks fuse shape conditions with keypoint latents and estimate correspondences through learnable affinity weights.
- **Soft Matching:** Apply Sinkhorn algorithm to produce many-to-many matching.

Pose Transfer Transformer $\mathcal{H}$



- **Initialization:** Use a correspondence-aware initialization step to set target keypoint transformations and fuse deformation cues via cross-attention.
- **Refinement:** Feed shape and keypoint tokens into transformer blocks to obtain high-level representations, and decode them into refined target transformations.
- **Output:** Deform the canonical target mesh to produce the final posed geometry.

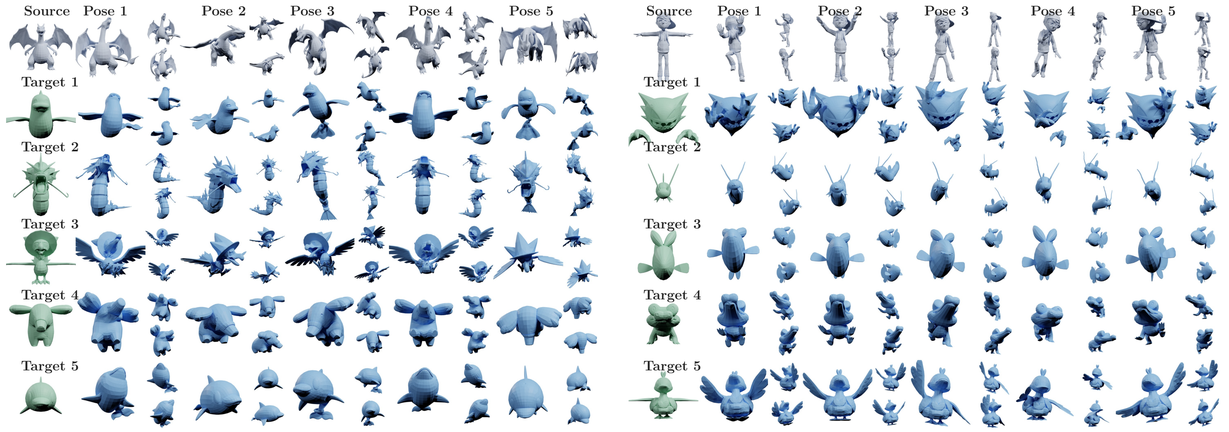
Quantitative Results

Methods	H2H		CCT	
	PMD↓	ELS↑	PMD↓	ELS↑
NPT	6.334	0.842	9.889	0.260
CGT	5.687	0.887	6.314	0.744
SFPT	3.616	0.888	4.312	0.913
TapMo	5.078	0.877	4.883	0.922
Ours	**3.570**	**0.923**	**4.264**	**0.927**

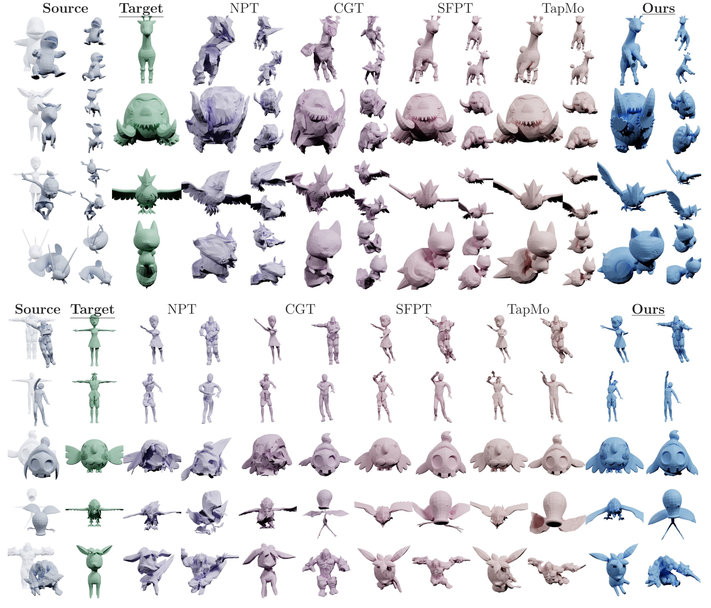
Rating	NPT	CGT	SFPT	TapMo	**Ours**
Pose↑	1.88	2.31	3.36	3.29	**4.08**
Geo.↑	1.56	2.52	3.53	3.61	**4.10**

- **Benchmarks:** Lowest PMD and highest ELS on both H2H and CCT settings.
- **User Study:** Highest ratings in pose similarity and geometry quality.

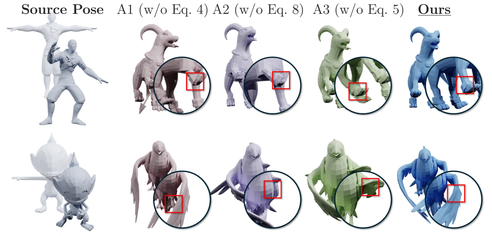
Qualitative Results



Comparison with Previous Methods

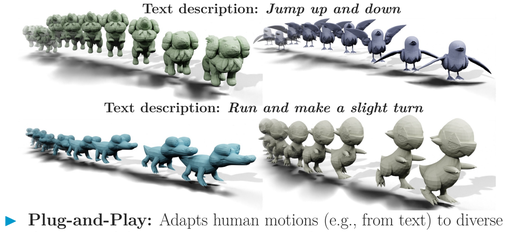


Ablation Study



- **Ablations:** Removing key components harms correspondence and degrades transfer quality. **Ours:** the full model delivers the most coherent pose transfer while better preserving target geometry.

Application



- **Plug-and-Play:** Adapts human motions (e.g., from text) to diverse characters while preserving geometry and semantics.