

SPECSIA: Stylization Dataset for Novel-View Enhancement in Drawing-based 3D Animation

Kyuwon Kim¹, Sunjae Yoon², and Chang D. Yoo^{*1}

¹ School of Electrical Engineering, KAIST, Daejeon, Republic of Korea

² Department of AI, Chung-Ang University, Seoul, Republic of Korea

¹ {rbdnjs7830, cd_yoo}@kaist.ac.kr ² sunjaeyoon@cau.ac.kr

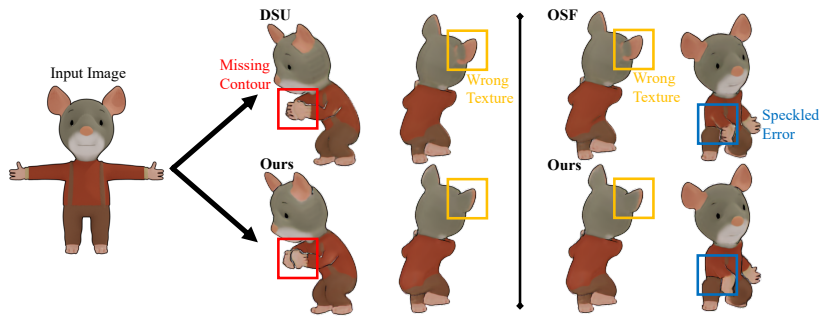


Fig. 1: Comparative animation results using DraViE (ours), DrawingSpinUp (DSU) [37], and the Occlusion-robust Stylization Framework (OSF) [34]. DraViE reduces novel-view artifacts while preserving the input drawing style.

Abstract. Generating animation from a single 2D drawing is challenging because the output must preserve character appearance while remaining plausible and temporally coherent under motion. Existing drawing-based 3D animation pipelines often use sample-wise 2D refinement to align animated renderings with the input image, but such optimization tends to overfit to the observed view and fails to correct projection-induced artifacts in novel views. To address this limitation, we introduce SPECSIA-15K, a paired stylization dataset containing 14,980 artifact-corrupted projection/refinement-target pairs from 1,498 3DBiCar characters. We further present DraViE (Drawing-based View Enhancement), a lightweight plug-and-play module trained with data-level priors to remove novel-view artifacts while preserving style and motion plausibility. Experiments show consistent gains in novel-view fidelity and temporal coherence with lower per-character adaptation cost than sample-wise fine-tuning. Project page: <https://rbdnjs7830.github.io/SPECSIA/>

Keywords: Drawing-based 3D animation · Novel-view enhancement · Stylization dataset

* Corresponding author.

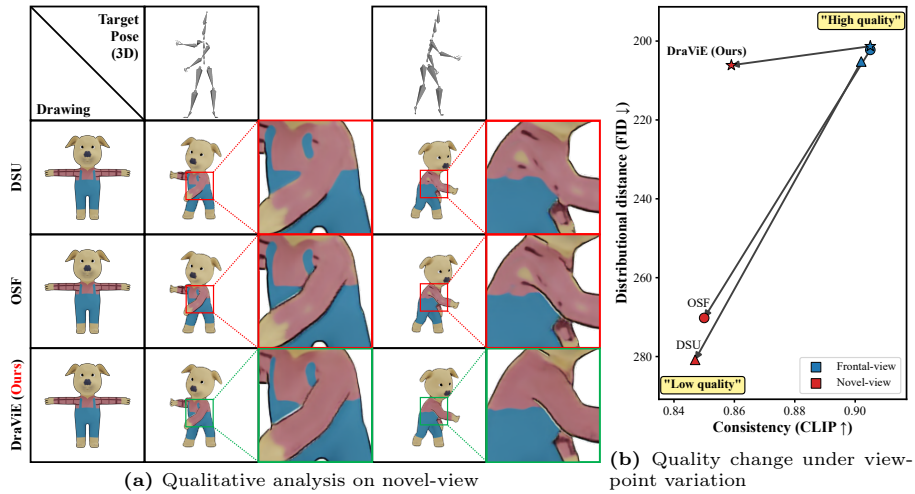


Fig. 2: Quality degradation in novel views. (a) Existing pipelines produce artifacts in novel-view renderings, such as missing contours and speckle-like regions (zoomed in red boxes). DSU: DrawingSpinUp [37], OSF: Occlusion-robust Stylization Framework [34]. (b) Joint plot of CLIP (semantic consistency) [20] and FID (distributional fidelity) [9] for frontal vs. novel views. Triangles/circles indicate DSU/OSF, and blue/red points indicate frontal/novel views. Lines connect the two for each method, highlighting the degradation in novel views. The plot follows the same evaluation setting as Table 1.

1 Introduction

Single-image animation [11, 24, 34, 37] has become increasingly important for applications such as character animation, content creation, and virtual avatars. However, generating time-varying motion from a single input image remains challenging: the output should preserve the input character’s visual identity (e.g., color, texture, and silhouette) while maintaining physically plausible motion and temporal coherence [8, 32, 34, 37]. This challenge is further amplified because the input is typically limited to a single view (typically a frontal view), making it difficult to maintain stable appearance in unseen or novel views. Figs. 1 and 2 show representative results and the motivating novel-view degradation.

Many recent approaches adopt 3D-assisted pipelines that reconstruct a 3D structure (e.g., a mesh) from the input, apply target motion in 3D, and render the animated result back to 2D [17, 28]. However, correcting errors in 3D is expensive and tightly coupled to the reconstruction pipeline. As a practical alternative, drawing-based animation systems rely on sample-wise 2D refinement by aligning the input image with 2D projections of the rendered animation [34, 37]. While effective at improving appearance in the observed view, such sample-wise optimization often overfits to the input view and degrades quality in unseen views, because alignment objectives enforce pixel-/feature-level consistency and cannot reliably correct artifacts originating from 3D rendering and projection.

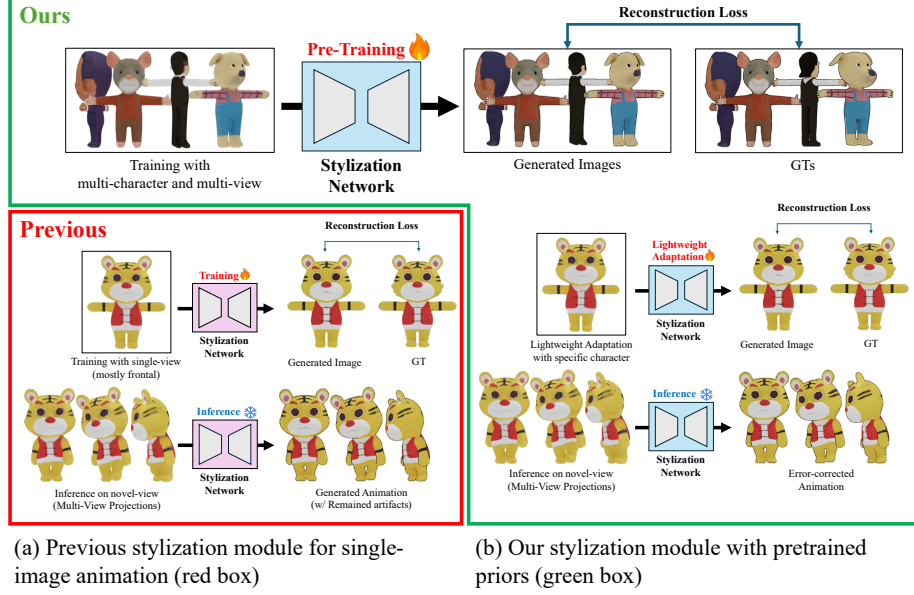


Fig. 3: Comparison between the previous stylization pipeline and our approach. (a) Illustration of the previous stylization network with sample-wise alignment. Because sample-wise alignment is tied to the input pose, the module may fail to correct errors that emerge in 3D rendering or 2D projection. (b) Our approach: stylization network with a data-level pre-trained prior.

A key bottleneck, therefore, is the lack of scalable supervision for improving unseen-view quality under large viewpoint changes and view-dependent projection artifacts. This scarcity makes it difficult to learn correction priors that generalize across characters, motions, and viewpoints. To address this gap, we introduce **SPECSIA-15K**, a stylization dataset designed for novel-view enhancement in drawing-based 3D animation. SPECSIA-15K is constructed from multi-view renderings of 3DBiCar [18]: we sample 10 views per object and form 14,980 paired examples of artifact-corrupted 2D projections and corresponding targets for refinement. This paired supervision enables data-level learning of correction priors that generalize across diverse characters, motions, and viewpoints.

Leveraging SPECSIA-15K, we present **DraViE** (Drawing-based View Enhancement), a lightweight pre-trained module that performs plug-and-play post-correction on rendered outputs. Trained at the data level rather than via sample-wise optimization, DraViE learns priors that correct novel-view artifacts while preserving drawing style and motion plausibility. Finally, while data-level priors improve generalization, test-time inputs may still exhibit unique drawing styles (e.g., line thickness or color palette). To handle this in a practical and low-cost manner, we optionally perform a lightweight adaptation on the single reference drawing using a small number of updates. This improves alignment to the specific

input style while preserving the learned priors and avoiding strong frontal-view overfitting. As a result, DraViE enables view-consistent, high-quality animation with lower per-character adaptation cost in our setting, while remaining easy to integrate into diverse motion generation pipelines as a simple post-processing component.

Overall, this paper reveals the limitations of sample-wise alignment in existing drawing-based 3D animation pipelines, which often overfit to the observed view. To enable scalable supervision for novel-view enhancement, we introduce SPECSIA-15K, a scalable stylization dataset that supports learning and evaluation. Building on this dataset, we propose DraViE, a data-level pre-trained, plug-and-play module for novel-view error correction, yielding improvements in novel-view fidelity and temporal coherence across motions and viewpoints.

2 Related Works

2.1 Image-to-3D Animation & Hand-drawn Animation

Prior work on single-image animation combines deformation modeling with image synthesis to transfer motion while preserving appearance [1, 2]. Recent works based on editing- and mesh-guided methods further improve controllability for non-rigid editing and structure-aware manipulation [14, 15, 27]. Nevertheless, under single-view inputs, depth ambiguity and occlusion make it difficult to maintain identity and temporal coherence, especially under viewpoint changes [24, 32, 35]. These limitations have motivated 3D-assisted pipelines that reconstruct and animate a 3D representation, often aided by multi-view generation to improve reconstruction quality [16, 17, 28]. However, for hand-drawn characters, rendering and projection can distort lines and textures, and existing pipelines commonly rely on sample-wise optimization to match the input style [34, 37]. In this context, the scarcity of scalable supervision for improving unseen-view quality remains a key challenge for learning view-consistent refinement. This motivates data-level resources and learning objectives that explicitly provide scaled supervision for improving novel-view quality in drawing-based 3D animation.

2.2 Image Stylization

Image stylization aims to transform "style" factors (e.g., texture and color) while preserving the content and identity of the input. Early work formulated stylization via content-style decomposition [7], and later studies adopted GAN-based image-to-image translation for broader domain transfer [12, 38]. More recently, diffusion models have extended stylization beyond pure style change toward quality enhancement and error correction by leveraging generative priors under diverse degradations [5, 10, 19, 23, 26]. In video and animation, applying stylization independently to each frame often breaks temporal coherence, motivating approaches that enforce temporal consistency for coherent stylization, generation, and diffusion video editing [6, 8, 22, 33]. This issue is particularly pronounced for

hand-drawn character animation, where style is strongly defined by lines and contours. Recent pipelines take 3D-animated renderings and projected 2D images as input and apply stylization networks to refine appearance and match the intended drawing style [34,37]. However, these methods focus on local refinement and may have limited capacity to correct view-dependent rendering/projection errors that emerge in novel views. Therefore, a stylization formulation that incorporates priors for error correction supported by scaled data across viewpoints is crucial for robust novel-view enhancement in drawing-based animation.

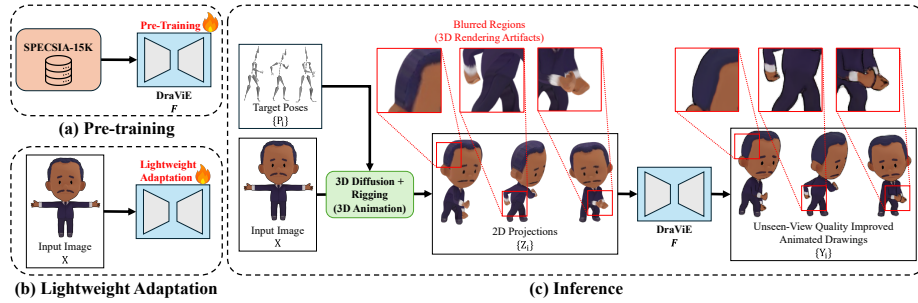


Fig. 4: DraViE: Drawing-based View Enhancement. (a) We pretrain DraViE on SPECSIA-15K. (b) We then optionally perform lightweight adaptation to the input character. (c) Given an input drawing and a target motion, a 3D animation system (e.g., 3D diffusion [17] and rigging [3]) produces novel-view projections that often contain view-dependent projection-induced artifacts. DraViE applies prior-guided 2D refinement as a plug-and-play module, reducing view-dependent degradations while preserving the original drawing style. Red boxes highlight artifacts corrected by DraViE.

3 Method

We address novel-view degradation and artifacts in drawing-based 3D animation pipelines by introducing (i) **SPECSIA-15K**, a paired stylization dataset for projection refinement, and (ii) **DraViE** (Drawing-based View Enhancement), a lightweight pre-trained post-correction module. Our goal is not to replace upstream 3D animation pipelines, but to provide scalable supervision and a modular 2D refinement prior for reducing view-dependent projection-level artifacts. Figs. 4 and 3 summarize the workflow and comparison to sample-wise stylization.

A typical drawing-based animation pipeline first reconstructs an animatable 3D character from a single input drawing X using an image-to-3D model (e.g., diffusion-based reconstruction [17]), applies rigging to enable pose-driven articulation [3], and renders animated frames under a motion sequence $\{P_i\}$. Here, $i \in \{1, \dots, T\}$ indexes the animation frame (time step), and P_i denotes the target pose at frame i . Given a target pose P_i , the system produces a rendered 2D projection Z_i from the user’s viewpoint. In practice, Z_i often exhibits line/contour

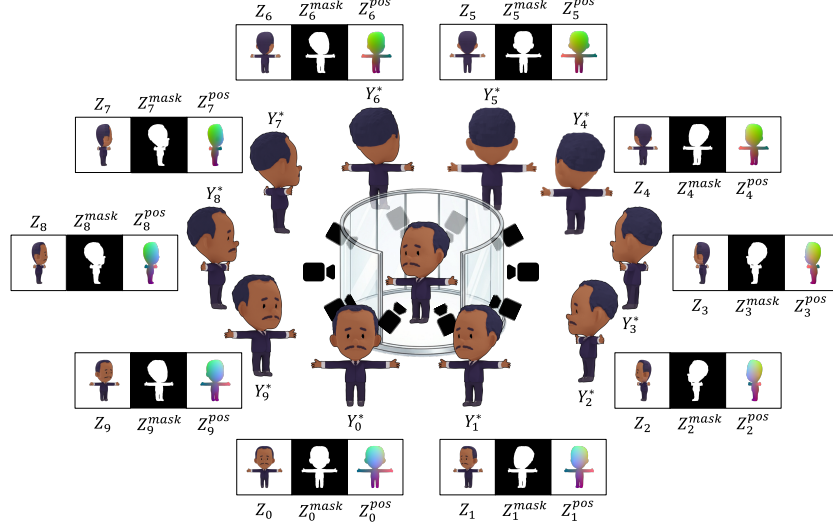


Fig. 5: SPECSIA-15K Dataset Construction. We construct paired viewpoint-aligned training data from 1,498 3DBiCar [18] characters by rendering 10 viewpoints as artifact-free ground truth $\{Y_v^*\}_{v=0}^9$. For each character, we reconstruct a 3D model from a single frontal view and project it to the same viewpoints to obtain artifact-prone projections $\{Z_v\}_{v=0}^9$. Each training sample consists of a poor input triplet $(Z_v, Z_v^{mask}, Z_v^{pos})$ (RGB projection, foreground mask, and positional hint) paired with the corresponding GT rendering Y_v^* , yielding 14,980 poor-GT pairs for 2D projection refinement.

distortions, texture degradation, and view-dependent artifacts, particularly in unseen viewpoints. While correcting these errors directly in 3D is expensive and tightly coupled to the reconstruction pipeline, many systems employ 2D refinement modules on projected images for efficiency and modularity. However, existing approaches are optimized in a sample-wise manner (often per character), which can overfit to the observed view and fail to generalize to novel views.

DraViE performs 2D post-correction on projected frames using an encoder-decoder network F (e.g., U-Net [21]). Following prior works that use spatial conditioning to resolve local ambiguities [13, 37], we incorporate a positional hint Z^{pos} as an input to the refinement network. Given the projection Z_i and auxiliary signals (mask and positional hint), DraViE outputs a refined frame:

$$Y_i = F(Z_i, Z_i^{mask}, Z_i^{pos}), \quad (1)$$

where Y_i corrects artifacts in Z_i while preserving the drawing style and motion plausibility. This enables refinement without modifying the upstream pipeline.

3.1 SPECSIA-15K Dataset

To learn priors for novel-view error correction, we require scalable supervision that connects (i) artifact-prone projected results and (ii) corresponding artifact-

free targets (Fig. 5). Such scalable data is scarce for hand-drawn character animation, which has motivated character-specific, sample-wise optimization in prior work. We fill this gap by introducing **SPECSIA-15K**, a paired dataset specialized for 2D projection refinement across viewpoints in drawing-based 3D animation.

Data Source. We use drawing-style character assets from 3DBiCar [18]. For each character, we generate multi-view artifact-free targets by directly rendering the original asset, and multi-view artifact-prone projections by reconstructing a 3D character from a single reference view and re-projecting it to novel views.

View Sampling and Pairing. For each character, we uniformly sample 10 viewpoints along yaw in $0^\circ \sim 360^\circ$ and render artifact-free targets $\{Y_v^*\}_{v=0}^9$. We then select the frontal view Y_0^* as the single-view reference input to an image-to-3D reconstruction pipeline (e.g., diffusion-based reconstruction and rigging [3, 17]). The reconstructed 3D model is projected into the same 10 viewpoints, yielding artifact-prone projections $\{Z_v\}_{v=0}^9$. In practice, Z_v contains line/contour distortions, texture degradation, and view-dependent artifacts that are amplified in unseen viewpoints. This paired setup aligns inputs and targets by viewpoint. In addition to the RGB projection Z_v , we provide two auxiliary signals:

- **Foreground mask** Z_v^{mask} : binary mask indicating valid object regions.
- **Positional hint** Z_v^{pos} : aligned map encoding canonical surface coordinates.

Each sample forms a triplet input $(Z_v, Z_v^{mask}, Z_v^{pos})$ paired with target Y_v^* .

Rendering Protocol and Auxiliary Signals.

Artifact-free targets Y_v^* . We render 3DBiCar characters in Blender [4] using an orthographic camera and transparent background, producing RGBA PNG frames at 512×512 resolution. To match diverse drawing-like appearances, we overlay a contour outline using Blender Freestyle as a style augmentation. Specifically, we sample the outline thickness uniformly from $\{1, 2, 3, 4\}$ pixels and sample the line color in grayscale from $[0, 255]$ (black to white), then apply it within each rendered frame. We use the same camera setup across views and uniformly rotate the camera rig around the character to obtain 10 yaw-spaced views.

Artifact-prone projections Z_v . To generate Z_v , we reconstruct a 3D character from the single reference view Y_0^* using an image-to-3D model and apply rigging [3, 17]. We then render the reconstructed character under the same set of viewpoints. Because the reconstructed model can be imperfect, these projections naturally exhibit systematic projection errors such as missing textures, warped contours, and view-inconsistent patterns, particularly in novel views.

Foreground mask Z_v^{mask} . We generate Z_v^{mask} by rendering a silhouette-only pass (emission material) for each view and extracting the alpha mask. To reduce view-dependent misalignment between the projected mesh and the target silhouette, we optionally adjust camera/object alignment by searching small translations and pitch/roll offsets that maximize mask IoU with the target silhouette. This alignment step improves the correspondence between Z_v and Y_v^* during pair construction without altering the underlying reconstruction pipeline.

Positional hint Z_v^{pos} . Inspired by the positional conditioning strategy used in DrawingSpinUp [37] and Stylizing Video by Example [13], we generate Z_v^{pos} by encoding canonical mesh coordinates as colors. Specifically, we min-max normalize mesh vertex coordinates and assign $(\tilde{x}, \tilde{y}, \tilde{z}) \in [0, 1]^3$ to vertex colors, then render it as an RGB map aligned with each projected frame. This conditioning provides a spatial cue about where a patch originates on the character surface.

Dataset Scale and Usage. SPECSIA-15K includes 1,498 characters and 10 viewpoints per character, resulting in 14,980 samples. SPECSIA-15K provides poor-GT pairs tailored to learning robust priors for projection refinement. By training on diverse characters and viewpoints, DraViE learns corrections that generalize beyond the reference view and improve stability in unseen viewpoints.

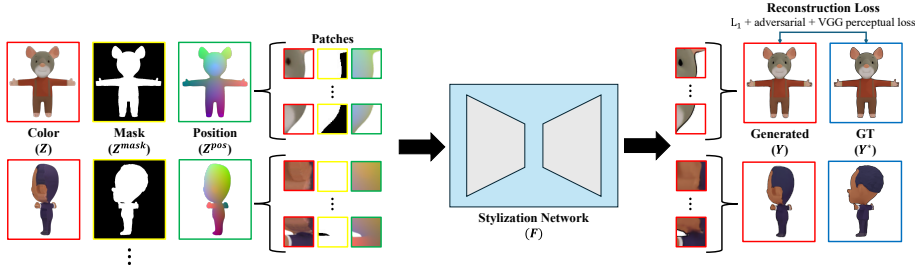


Fig. 6: Pretraining DraViE model. Each character has 10 different views, and each RGB projection is paired with its mask and positional hint. These triplets are split into patches, and a stylization network is trained to reconstruct these patches. Before inference, the pre-trained DraViE model is optionally adapted to the input drawing.

3.2 Pre-training DraViE with Data-level Priors

We pre-train F on SPECSIA-15K to learn general refinement priors for recurring artifact patterns in 3D renderings and 2D projections. We follow the overall training scheme of DrawingSpinUp with minor adaptations to our setting [37].

Training Objective. The model predicts Y from (Z, Z^{mask}, Z^{pos}) and is supervised by Y^* , the corresponding artifact-free target rendering:

$$F(Z, Z^{mask}, Z^{pos}) \approx Y^*. \quad (2)$$

Pre-training differs from sample-wise optimization in that it learns correction priors over a dataset distribution, improving generalization to unseen viewpoints.

Patch-wise Training. Hand-drawn style is governed by local cues (e.g., stroke thickness and texture granularity), which are difficult to capture from full-resolution images. Moreover, training on full frames can bias the model toward dominant pose patterns (e.g., the canonical T-pose), leading to pose-specific overfitting; patch-wise training mitigates this issue by emphasizing local style cues. As Fig. 6 shows, we sample a patch center guided by Z^{mask} and crop patches of size 32×32 from (Z, Z^{mask}, Z^{pos}) and Y^* . The model predicts the refined patch Y_{patch} , and we supervise it with a combination of L_1 loss, adversarial loss, and VGG perceptual loss between Y_{patch} and Y_{patch}^* . Each training step, we choose a patch from a random character and a random viewpoint. At inference time, since F is fully convolutional, it can process full-resolution inputs (not limited to 32×32), enabling practical refinement of full-frame projected renderings.

3.3 Lightweight Adaptation (Optional)

While the pre-trained model generalizes across diverse characters, test-time inputs may exhibit unique drawing styles (e.g., line thickness or color palette). We optionally perform a lightweight adaptation to calibrate DraViE to the input character, improving character-specific appearance while preserving data-level correction priors. Unlike full sample-wise optimization, this adaptation is computationally inexpensive and mitigates strong overfitting to the reference view. Concretely, we initialize from the SPECSIA-15K-pretrained generator and fine-tune it on the input character for a single epoch using the same patch-wise objective as pre-training. We keep the architecture and loss unchanged and only reduce the generator learning rate and the number of epochs. We sample 32×32 patches from the character’s input pose renderings (with the same auxiliary inputs) and optimize a combination of L_1 loss, VGG perceptual loss, and adversarial loss. This brief update improves input-style alignment (colors/strokes/textures) for the given character while retaining the learned artifact-correction prior.

3.4 Inference

Given an input drawing X and motion sequence $\{P_i\}$, the 3D animation system produces projected frames $\{Z_i\}$ along with auxiliary signals $\{Z_i^{mask}, Z_i^{pos}\}$ for all animation timesteps in the sequence. DraViE refines each projected frame as:

$$Y_i = F(Z_i, Z_i^{mask}, Z_i^{pos}). \quad (3)$$

By leveraging learned priors, DraViE reduces view-dependent artifacts, preserves drawing style, and improves temporal coherence across novel views.

4 Experiments

4.1 Experimental Settings

Implementation Details. For image-to-3D reconstruction, we use Wonder3D [17]. For rigging, we use Adobe Mixamo¹ with a 65-joint skeleton. Unless otherwise specified, we render sequences with the same camera and rendering settings across methods to ensure a fair and controlled comparison setting.

Data and Baselines. SPECSIA-15K contains multi-view projections of 1,498 3DBiCar characters with 10 views per character. We use a character-disjoint split of 1,298/100/100 characters for train/validation/test, respectively, so that no character identity is shared between pre-training and evaluation. For evaluation, we render a near-frontal view for each test character as the single reference input. This choice reflects practical user inputs that are close to frontal but not strictly frontal. To further assess generalization to hand-drawn inputs, we additionally select 20 textured characters from Amateur Drawings [25], excluding characters with nearly monochromatic body textures. Each test character is driven by 20 motion sequences (10 frontal-dominant and 10 novel-view motions²), resulting in $(100 + 20) \times 20 = 2,400$ animation samples in total. We compare against two drawing-based 3D animation pipelines, DrawingSpinUp (DSU) [37] and the Occlusion-robust Stylization Framework (OSF) [34]. For all methods, we use identical 3D reconstruction and rigging inputs to isolate the effect of the 2D post-correction module. Additional clean-target, dataset-split, view-sector, efficiency, protocol, and qualitative analyses are provided in the supplementary material.

4.2 Evaluation Metrics

We evaluate the output video quality in terms of temporal coherence, fidelity, and overall realism. For temporal coherence, we compute (i) the average cosine similarity of adjacent-frame CLIP image embeddings [20], capturing semantic/appearance consistency over time, and (ii) the average SSIM [30] between adjacent frames, measuring structural stability. For fidelity, we report LPIPS [36] between each generated frame and the reference appearance image (the character texture), assessing perceptual identity preservation. For overall realism, we compute FID [9] between the set of reference images and the set of generated frames (sampled from all videos), measuring distribution-level visual quality. All metrics are averaged over 10 runs with different random seeds.

4.3 Experimental Results

Qualitative Comparisons. Fig. 7 compares outputs across different 3D animation systems. We evaluate the methods under dynamic motions (e.g., turning

¹ We use the online platform (<https://www.mixamo.com>, accessed: 2026-03-09).

² Detailed Mixamo motions are listed in Supplementary Material Sec. A.

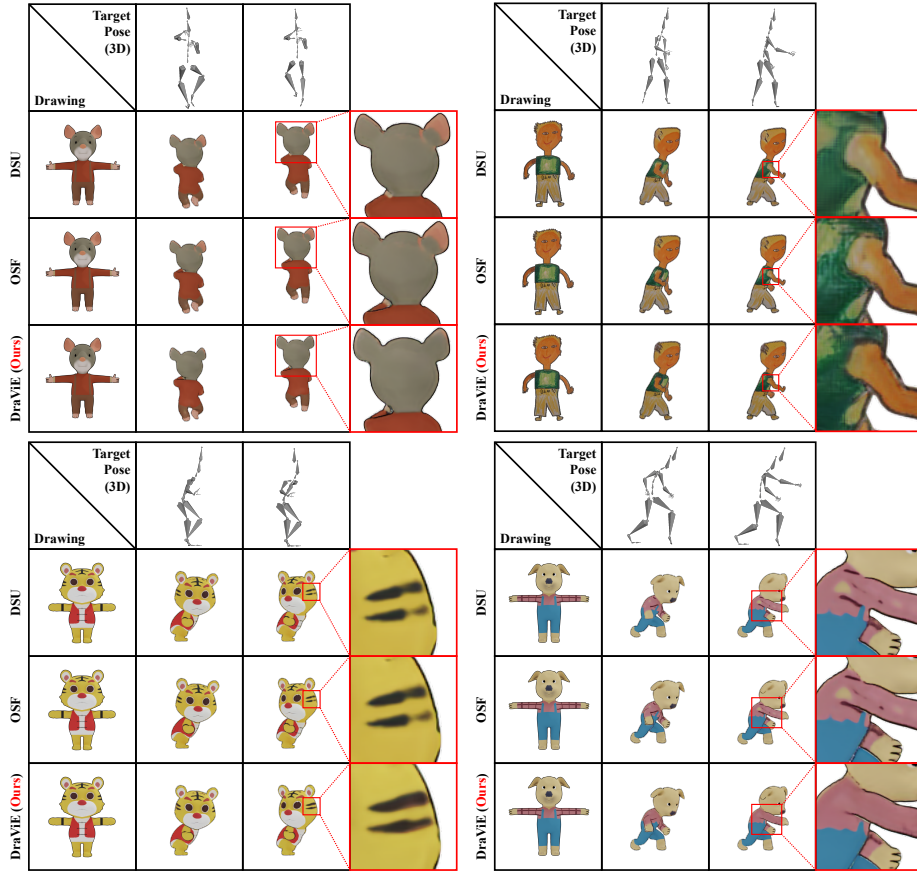


Fig. 7: Qualitative comparisons across different drawing-based 3D animation models. The red boxes highlight zoomed views of the stylization applied to animated drawing objects. DSU: DrawingSpinUp [37], OSF: Occlusion-robust Stylization Framework [34].

and punching) on diverse characters. All methods preserve the input drawing style. However, the artifact patterns differ. DrawingSpinUp (DSU) [37] tends to lose contour details under large motions, resulting in blurred textures. Both DSU and the Occlusion-robust Stylization Framework (OSF) [34] show speckle-like artifacts amplified in novel views due to rendering and projection errors. In contrast, DraViE reduces both blurring and speckling, effectively correcting projection-induced artifacts that prior methods often fail to address.³

Quantitative Results. Table 1 reports quantitative comparisons on drawing-based 3D animation systems. We evaluated the animation quality on motions dominated by frontal views and novel views. DraViE achieves performance com-

³ More qualitative results are provided in supplemental material section B.

Table 1: Quantitative evaluations on drawing-based 3D animation models. DSU: DrawingSpinUp [37], OSF: Occlusion-robust Stylization Framework [34].

Method	Frontal-view				Novel-view			
	CLIP↑	SSIM↑	LPIPS↓	FID↓	CLIP↑	SSIM↑	LPIPS↓	FID↓
DSU [37]	0.902	0.840	0.212	205.19	0.847	0.833	0.250	280.83
OSF [34]	0.905	0.841	0.211	202.24	0.850	0.835	0.249	270.21
DraViE (Ours)	0.905	0.846	0.208	201.31	0.859	0.840	0.245	206.13

Table 2: Cross-pipeline generalization and blinded user preference. Left: each triplet reports DSU/OSF/DraViE using the same SPECSIA-pretrained checkpoint and the same adaptation protocol, without backend-specific retraining. Right: preference values are percentages from a 30-person blinded A/B study.

Upstream	CLIP↑	LPIPS↓	FID↓	Method	Art.	Style	Overall
Wonder3D	.874/.877/. .882	.231/.230/. .225	243/236/ 203	Prior	26.7	40.0	13.3
InstantMesh	.831/.835/. .840	.277/.275/. .272	242/233/ 205	DraViE	73.3	60.0	86.7
CRM	.807/.810/. .848	.272/.271/. .267	271/266/ 239				

parable to prior methods on frontal-view animations, indicating that our correction does not degrade quality in the observed view. More importantly, DraViE yields consistent gains in novel views across all metrics, with higher CLIP [20] and SSIM [30] and lower LPIPS [36] and FID [9]. The large reduction in novel-view FID suggests that DraViE produces frames that are closer to the reference distribution and contain fewer projection-induced artifacts. These quantitative improvements align with the qualitative comparisons in Fig. 7.

Cross-pipeline Generalization and User Preference. To examine whether DraViE learns a transferable correction prior rather than a backend-specific denoiser, we apply the same SPECSIA-pretrained checkpoint to projections generated by Wonder3D [17], InstantMesh [31], and CRM [29], using the same lightweight adaptation protocol without backend-specific retraining. As shown in Table 2, DraViE consistently improves over DSU and OSF across all upstream backends, suggesting that the learned prior transfers across different reconstruction artifacts. The upstream models produce different artifact patterns; some projections contain texture bleeding near front/back surface boundaries, while others show clay-like or over-smoothed surface appearance. Despite these differences, DraViE improves the output by targeting projection-level errors such as speckling, boundary corruption, and texture inconsistency, rather than relying on a specific reconstruction backend. We also conduct a 30-person blinded A/B preference study with hidden method names and randomized order. Each participant compares DraViE against a randomly selected prior method under artifact reduction, style preservation, and overall visual quality criteria. Participants prefer DraViE across all three criteria, which supports the quantitative gains and indicates that the improvements are perceptually visible. Cross-pipeline qualitative

Table 3: Quantitative ablations on (i) lightweight adaptation and (ii) positional hint.

Setting	Frontal-view				Novel-view			
	CLIP $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	FID $\downarrow$	CLIP $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	FID $\downarrow$
w/o LA	0.900	0.840	0.212	207.15	0.838	0.834	0.250	221.48
w/ LA	0.905	0.846	0.208	201.31	0.859	0.840	0.245	206.13
w/o Z^{pos}	0.900	0.845	0.212	205.81	0.844	0.839	0.250	318.39
w/ Z^{pos}	0.905	0.846	0.208	201.31	0.859	0.840	0.245	206.13

examples, descriptor-space backend discrepancy, view-sector analysis, efficiency results, and additional analyses are provided in the supplementary material.

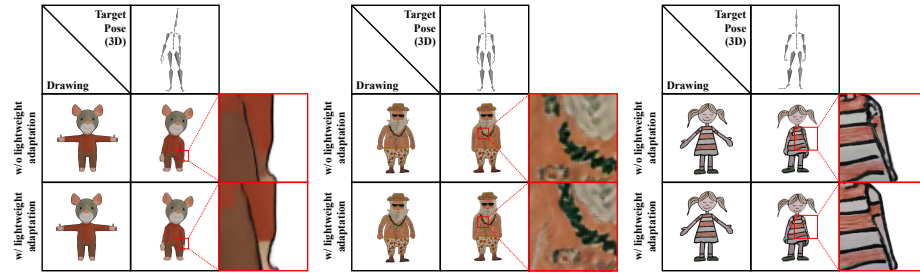


Fig. 8: Qualitative comparison of DraViE with and without the lightweight adaptation. Without adaptation, the pre-trained prior removes projection artifacts but often deviates from the input drawing appearance. It leads to mismatched colors and over-smoothed or altered strokes (e.g., hand color, leaf pattern, and stripe line thickness).

4.4 Ablation Study

Lightweight Adaptation. We introduce lightweight adaptation to better align DraViE with the appearance of the input drawing (e.g., stroke thickness, local textures, and color tones) while retaining the pre-trained correction prior. Table 3 shows that even without adaptation, DraViE already provides strong correction performance, especially on the frontal view. It indicates that the pre-trained model already captures common projection error patterns and can suppress view-dependent artifacts. However, the outputs can deviate from the input drawing appearance in novel views, which is consistent with lower CLIP [20] and higher LPIPS [36] and FID [9]. With lightweight adaptation, DraViE improves perceptual alignment to the input, and the gains are more pronounced in novel views (e.g., CLIP: 0.838 $\rightarrow$ 0.859, FID: 221.48 $\rightarrow$ 206.13). Fig. 8 qualitatively supports this trend. Without adaptation, the model tends to over-smooth textures and strokes, whereas adaptation better preserves detailed textures (e.g., hand colors, leaf patterns, and stripe strokes) while still correcting projection artifacts.

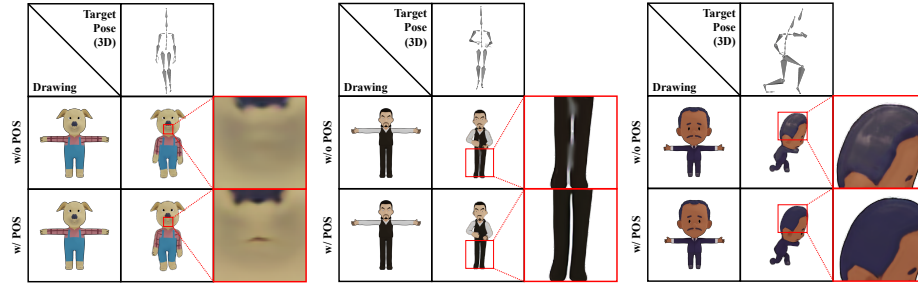


Fig. 9: Qualitative comparison of DraViE trained with and without the positional hint. Without positional hint, the model often fails to recover fine structures (e.g., mouth) and leaves blurring artifacts at specific regions (e.g., hair, legs).

Input Signal Ablation. DraViE uses two auxiliary inputs, mask and positional hint. The mask is essential for patch-wise training, while the positional hint provides the spatial location of each input patch. In this ablation, we show the effect of the positional hint by training DraViE with and without it. Fig. 9 shows that removing the positional hint significantly degrades output quality. Without positional information, the model observes only local appearances and cannot distinguish where a patch comes from on the character. As a result, the model tends to apply conservative smoothing over corrupted regions and fails to reconstruct geometric structures (e.g., mouth) and often leaves speckling artifacts induced from the projections (e.g., around hair and legs). In contrast, conditioning on the positional hint enables DraViE to learn location-aware correction priors, leading to sharper structure recovery and cleaner textures under viewpoint changes. Table 3 also shows that removing the positional hint degrades the quality of animation, confirming its role in stable refinement.

5 Limitations

DraViE is designed as a modular 2D post-correction component rather than a replacement for upstream 3D reconstruction, rigging, or temporal modeling. This makes the method easy to attach to existing drawing-based animation pipelines, but it also bounds the correction quality by the rendered proxy and limits the algorithmic scope of the model. In particular, DraViE cannot fully recover from severe upstream failures such as incorrect topology, missing large regions, or erroneous rigging.

For example, when the reconstructed mesh merges the legs into a single region,

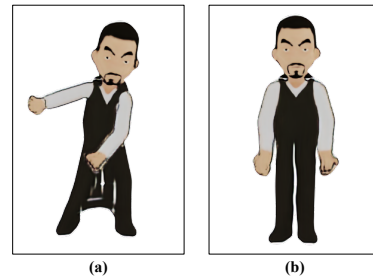


Fig. 10: Failure case. (a) Fused legs. (b) Hallucinated gap.

DraViE cannot reliably reconstruct separation. In some cases, the learned prior may hallucinate a leg gap and refine the projection to look separated, despite the input exhibiting a fused-leg structure (Fig. 10). This failure highlights a fundamental limitation of 2D post-correction: it may favor a plausible silhouette prior over the true input geometry when upstream topology is incorrect.

SPECSIA-15K is constructed with an offline rendering, reconstruction, and rigging pipeline. Although the supplementary material reports cross-pipeline evaluations, clean-target, and dataset-split, the dataset may encode systematic biases from the source assets and upstream reconstruction backend. Finally, DraViE processes frames independently and does not use explicit temporal attention or optical flow. The shared 3D proxy and positional hints improve stability in practice, but rapid rotations, large occlusions, or high-frequency details that are missing in the projection may still lead to flickering or over-smoothed details.

6 Conclusion

We introduced **SPECSIA-15K**, a paired multi-view stylization dataset for projection refinement in drawing-based 3D animation, and **DraViE**, a lightweight plug-and-play refinement module trained with data-level correction priors.

SPECSIA-15K provides scalable supervision for correcting projection artifacts under dynamic motion and varying viewpoints, moving beyond sample-wise refinement that can overfit to the observed view. By explicitly pairing artifact-prone projections with clean multi-view targets, SPECSIA-15K enables learning correction priors that are difficult to obtain from single-character optimization alone. Building on this dataset, DraViE improves novel-view fidelity and temporal coherence while preserving the input drawing style, and integrates into existing animation pipelines without modifying their core components.

These results suggest that dataset-level supervision is a practical direction for improving robustness in drawing-based 3D animation systems. At the same time, projection-level refinement cannot fully repair severe reconstruction or rigging failures, motivating future work on less biased data construction, stronger 3D reconstruction, and explicit temporal modeling.

We will release the training/evaluation code, split lists, rendering parameters, metadata, and redistributable SPECSIA noisy/clean pairs when permitted by the corresponding third-party licenses.

Acknowledgements

This work was supported by Institute for Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.RS-2021-II211381, Development of Causal AI through Video Understanding and Reinforcement Learning, and Its Applications to Real Environments) and (No.RS-2022-II220184, Development and Study of AI Technologies to Inexpensively Conform to Evolving Policy on Ethics)

References

1. Aberman, K., Wu, R., Lischinski, D., Chen, B., Cohen-Or, D.: Learning character-agnostic motion for motion retargeting in 2d. *ACM Transactions on Graphics (TOG)* **38**(4), 75 (2019)
2. AlBahar, B., Lu, J., Yang, J., Shu, Z., Shechtman, E., Huang, J.B.: Pose with Style: Detail-preserving pose-guided image synthesis with conditional stylegan. *ACM Transactions on Graphics* **40**(6) (2021)
3. Baran, I., Popović, J.: Automatic rigging and animation of 3d characters. *ACM Trans. Graph.* **26**(3), 72-es (Jul 2007)
4. Blender Foundation: Blender. <https://www.blender.org/> (2024), version 3.6.14, accessed: 2026-03-09
5. Choi, J., Kim, S., Jeong, Y., Gwon, Y., Yoon, S.: Ilvr: Conditioning method for denoising diffusion probabilistic models. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. pp. 14367–14376 (October 2021)
6. Gao, C., Gu, D., Zhang, F., Yu, Y.: Reconet: Real-time coherent video style transfer network. In: Jawahar, C.V., Li, H., Mori, G., Schindler, K. (eds.) *ACCV* (6). *Lecture Notes in Computer Science*, vol. 11366, pp. 637–653. Springer (2018)
7. Gatys, L.A., Ecker, A.S., Bethge, M.: Image style transfer using convolutional neural networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 2414–2423 (2016)
8. Guo, Y., Yang, C., Rao, A., Liang, Z., Wang, Y., Qiao, Y., Agrawala, M., Lin, D., Dai, B.: Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. In: *The Twelfth International Conference on Learning Representations* (2024)
9. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: GANs trained by a two time-scale update rule converge to a local nash equilibrium. In: *Advances in Neural Information Processing Systems*. vol. 30, pp. 6626–6637 (2017)
10. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: *Advances in Neural Information Processing Systems*. vol. 33, pp. 6840–6851 (2020)
11. Hu, L., Gao, X., Zhang, P., Sun, K., Zhang, B., Bo, L.: Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 8153–8163 (2024)
12. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 1125–1134 (2017)
13. Jamriška, O., Sochorová, v., Texler, O., Lukáč, M., Fišer, J., Lu, J., Shechtman, E., Sýkora, D.: Stylizing video by example. *ACM Trans. Graph.* **38**(4) (Jul 2019)
14. Koo, G., Yoon, S., Hong, J.W., Yoo, C.D.: Flexiedit: Frequency-aware latent refinement for enhanced non-rigid editing. In: *Computer Vision – ECCV 2024*. pp. 363–379. Springer (2024)
15. Koo, G., Yoon, S., Lee, Y., Hong, J.W., Yoo, C.D.: Flowdrag: 3d-aware drag-based image editing with mesh-guided deformation vector flow fields. In: *Proceedings of the 42nd International Conference on Machine Learning. Proceedings of Machine Learning Research*, vol. 267, pp. 31467–31485. PMLR (2025)
16. Liu, Y., Lin, C., Zeng, Z., Long, X., Liu, L., Komura, T., Wang, W.: Syncdreamer: Generating multiview-consistent images from a single-view image. In: *The Twelfth International Conference on Learning Representations* (2024)

17. Long, X., Guo, Y.C., Lin, C., Liu, Y., Dou, Z., Liu, L., Ma, Y., Zhang, S.H., Habermann, M., Theobalt, C., Wang, W.: Wonder3d: Single image to 3d using cross-domain diffusion. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 9970–9980 (2024)
18. Luo, Z., Cai, S., Dong, J., Ming, R., Qiu, L., Zhan, X., Han, X.: Rabbit: Parametric modeling of 3d biped cartoon characters with a topological-consistent dataset. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 12825–12835 (2023)
19. Nichol, A.Q., Dhariwal, P.: Improved denoising diffusion probabilistic models. In: Meila, M., Zhang, T. (eds.) *Proceedings of the 38th International Conference on Machine Learning*. *Proceedings of Machine Learning Research*, vol. 139, pp. 8162–8171. PMLR (18–24 Jul 2021)
20. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Meila, M., Zhang, T. (eds.) *Proceedings of the 38th International Conference on Machine Learning*. *Proceedings of Machine Learning Research*, vol. 139, pp. 8748–8763. PMLR (18–24 Jul 2021)
21. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. pp. 234–241. Springer International Publishing, Cham (2015)
22. Ruder, M., Dosovitskiy, A., Brox, T.: Artistic Style Transfer for Videos, pp. 26–36. Springer International Publishing (2016)
23. Saharia, C., Ho, J., Chan, W., Salimans, T., Fleet, D.J., Norouzi, M.: Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **45**(4), 4713–4726 (2023)
24. Siarohin, A., Lathuilière, S., Tulyakov, S., Ricci, E., Sebe, N.: First order motion model for image animation. In: *Advances in Neural Information Processing Systems*. vol. 32 (2019)
25. Smith, H.J., Zheng, Q., Li, Y., Jain, S., Hodgins, J.K.: A method for animating children’s drawings of the human figure. *ACM Trans. Graph.* **42**(3) (jun 2023)
26. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021. OpenReview.net (2021)
27. Sorkine, O., Cohen-Or, D., Lipman, Y., Alexa, M., Rössl, C., Seidel, H.P.: Laplacian surface editing. In: *Proceedings of the 2004 Eurographics/ACM SIGGRAPH Symposium on Geometry Processing*. pp. 175–184. SGP ’04, Association for Computing Machinery, New York, NY, USA (2004)
28. Tang, J., Ren, J., Zhou, H., Liu, Z., Zeng, G.: Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. In: *The Twelfth International Conference on Learning Representations* (2024)
29. Wang, Z., Wang, Y., Chen, Y., Xiang, C., Chen, S., Yu, D., Li, C., Su, H., Zhu, J.: CRM: Single image to 3d textured mesh with convolutional reconstruction model. In: *Computer Vision – ECCV 2024*. pp. 57–74. Springer (2024)
30. Wang, Z., Bovik, A., Sheikh, H., Simoncelli, E.: Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing* **13**(4), 600–612 (2004)
31. Xu, J., Cheng, W., Gao, Y., Wang, X., Gao, S., Shan, Y.: Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models. arXiv preprint arXiv:2404.07191 (2024)

32. Xu, Z., Zhang, J., Liew, J.H., Yan, H., Liu, J.W., Zhang, C., Feng, J., Shou, M.Z.: Magicanimate: Temporally consistent human image animation using diffusion model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1481–1490 (2024)
33. Yoon, S., Koo, G., Kim, G., Yoo, C.D.: Frag: Frequency adapting group for diffusion video editing. In: Proceedings of the 41st International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 235, pp. 57315–57330. PMLR (2024)
34. Yoon, S., Koo, G., Lee, Y., Hong, J.W., Yoo, C.D.: Occlusion-robust stylization for drawing-based 3d animation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 12263–12273 (2025)
35. Yoon, S., Koo, G., Lee, Y., Yoo, C.D.: Tpc: Test-time procrustes calibration for diffusion-based human image animation. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 37, pp. 118654–118677 (2024)
36. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 586–595 (2018)
37. Zhou, J., Xiao, C., Lam, M.L., Fu, H.: Drawingspinup: 3d animation from single character drawings. In: SIGGRAPH Asia 2024 Conference Papers. Association for Computing Machinery, New York, NY, USA (2024)
38. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 2223–2232 (2017)