CoDeF: Content Deformation Fields for Temporally Consistent Video Processing

Hao Ouyang[1,2*] Qiuyu Wang[2*] Yuxi Xiao[2,3*] Qingyan Bai[1,2] Juntao Zhang[1]
Kecheng Zheng[2] Xiaowei Zhou[3] Qifeng Chen[1†] Yujun Shen[2†]
[1]HKUST [2]Ant Group [3]CAD&CG, ZJU

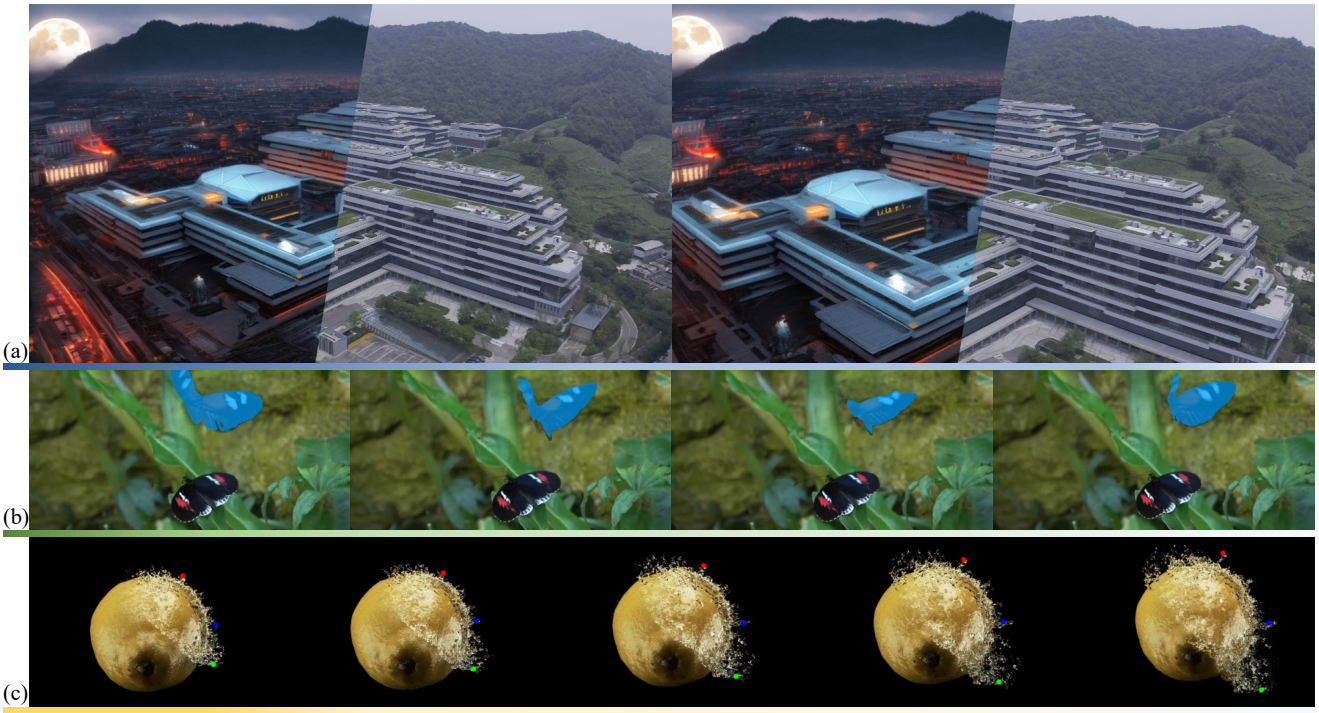

Figure 1. **Versatile applications** of CoDeF including (a) text-guided video-to-video translation (left half: translated frames, right half: input frames), (b) video object tracking, and (c) video keypoint tracking. It is noteworthy that, with the proposed type of video representation, we manage to directly *lift* image algorithms for video processing *without any tuning* on videos.

Abstract

We present the content deformation field (CoDeF) as a new type of video representation, which consists of a ***canonical content field*** *aggregating the static contents in the entire video and a* ***temporal deformation field*** *recording the transformations from the canonical image (i.e., rendered from the canonical content field) to each individual frame along the time axis. Given a target video, these two fields are jointly optimized to reconstruct it through a carefully tailored rendering pipeline. We advisedly introduce some regularizations into the optimization process, urging the canonical content field to inherit semantics (e.g., the object shape) from the video. With such a design, CoDeF naturally supports lifting image algorithms for video processing, in the sense that one can apply an image algorithm to the canonical image and effortlessly propagate the outcomes to the entire video with the aid of the temporal deformation field. We experimentally show that CoDeF is able to lift image-to-image translation to video-to-video translation and lift keypoint detection to keypoint tracking without any training. More importantly, thanks to our lifting strategy that deploys the algorithms on only one image, we achieve superior cross-frame consistency in processed videos compared to existing video-to-video translation approaches, and even manage to track non-rigid objects like water and smog. Code is made available at https://qiuyu96.github.io/CoDeF/.*

1. Introduction

The field of image processing has witnessed remarkable advancements, largely attributable to the power of generative models trained on extensive datasets, yielding exceptional quality and precision. However, the processing of video content has not achieved comparable progress. One challenge lies in maintaining high temporal consistency, a task complicated by the inherent randomness of neural networks. Another challenge arises from the nature of video datasets themselves, which often include textures of inferior quality compared to their image counterparts and necessitate greater computational resources. Consequently, the quality of video-based algorithms significantly lags behind those focused on images. This contrast prompts a question: *is it feasible to represent video in the form of an image to seamlessly apply established image algorithms to video content with high temporal consistency*?

In pursuit of this objective, researchers have suggested the generation of video mosaics from dynamic videos [40, 47] in the era preceding deep learning, and the utilization of a neural layered image atlas [15, 22, 65] subsequent to the proposal of implicit neural representations. Nonetheless, these methods exhibit two principal deficiencies. First, the capacity of these representations, particularly in faithfully reconstructing intricate details within a video, is restricted. Often, the reconstructed video overlooks subtle motion details, such as blinking eyes or slight smiles. The second limitation pertains to the typically distorted nature of the estimated atlas, which consequently suffers from impaired semantic information. Existing image processing algorithms, therefore, do not perform optimally as the estimated atlas lacks sufficient naturalness.

We propose a novel approach to video representation that utilizes a 2D hash-based image field coupled with a 3D hash-based temporal deformation field. The incorporation of multi-resolution hash encoding [28] for the representation of temporal deformation significantly enhances the ability to reconstruct general videos. This formulation facilitates tracking the deformation of complex entities such as water and smog. However, the heightened capability of the deformation field presents a challenge in estimating a natural canonical image. An unnatural canonical image can also estimate the corresponding deformation field with a faithful reconstruction. To navigate this challenge, we suggest employing annealed hash during training. Initially, a smooth deformation grid is utilized to identify a coarse solution applicable to all rigid motions, with high-frequency details added gradually. Through this coarse-to-fine training, the representation achieves a balance between the naturalness of the canonical and the faithfulness of the reconstruction. We observe a noteworthy enhancement in reconstruction quality compared to preceding methods. This improvement is quantified as an approximately 2.3 increase in PSNR, along with an observable increase in the naturalness of the canonical image. Our optimization process requires a mere approximate 300 seconds to estimate the canonical image with the deformation field while the previous implicit layered representations[15] takes more than 10 hours.

Building upon our proposed content deformation field, we illustrate lifting image processing tasks such as prompt-guided image translation, super-resolution, and segmentation, to the realm of videos. Our approach to prompt-guided video-to-video translation employs ControlNet [67] on the canonical image, propagating the translated content via the learned deformation. The translation process is conducted on a single canonical image and obviates the need for time-intensive inference models (*e.g.*, Diffusion models) across all frames. Our translation outputs exhibit marked improvements in temporal consistency and texture quality over the state-of-the-art zero-shot video translations with generative models [35, 63]. When contrasted with Text2Live, which relies on a neural layered atlas, our model is proficient in handling more complex motion, producing more natural canonical images, and thereby achieving superior translation results. Additionally, we extend the application of image algorithms such as super-resolution, semantic segmentation, and keypoints detection to the canonical image, leading to their practical applications in video contexts. This includes video super-resolution, video object segmentation, video keypoints tracking, among others. Our proposed representation consistently delivers superior temporal consistency and high-fidelity synthesized frames, demonstrating its potential as a novel tool in video processing.

2. Related Work

Implicit Neural Representations. Implicit representations in conjunction with coordinate-based Multilayer Perceptrons (MLPs) have demonstrated its powerful capability in accurately representing images [4, 49, 51], videos [15, 20, 49, 65], and 3D/4D representations [25, 26, 30–33, 56]. These techniques have been employed in a range of applications, including novel view synthesis [26], image super-resolution [4], and 3D/4D Reconstruction [56, 61]. Furthermore, for the purpose of speeding up the training, a various of acceleration [28, 46] techniques have been explored to replace the original Fourier positional encoding with some discrete representation like multi-resolution feature grid or hash table. Moreover, the adoption of an implicit deformation field [19, 31, 32, 34] has displayed a remarkable capability to overfit dynamic scenes. Inspired by these works, our primary objective is to reconstruct videos by utilizing a canonical image which inherit semantics for video processing purposes.

Consistent Video Editing. Our research is closely aligned

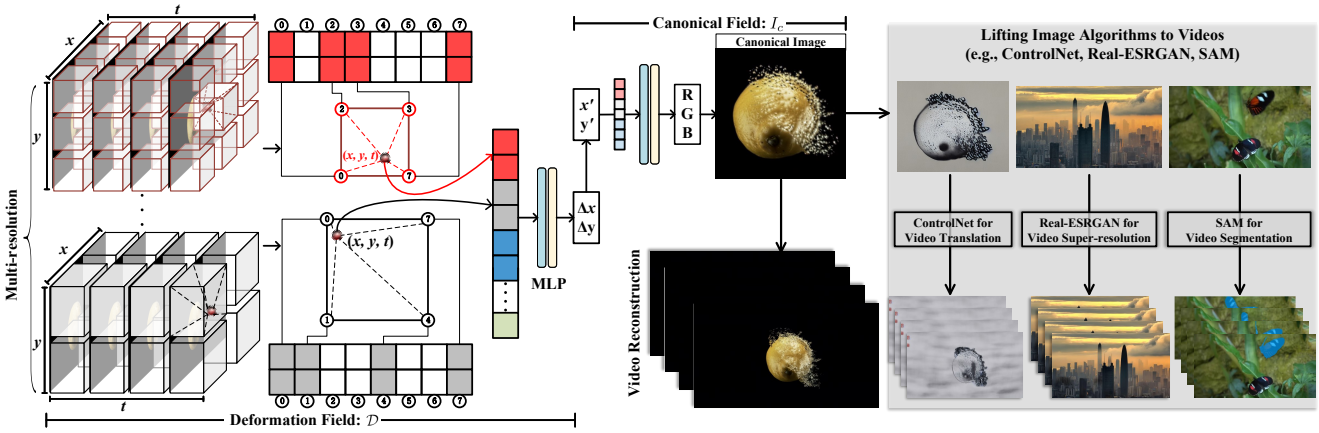

Figure 2. **Illustration of the proposed video representation**, CoDeF, which factorizes an arbitrary video into a 2D content canonical field and a 3D temporal deformation field. Each field is implemented with a multi-resolution 2D or 3D hash table using an efficient MLP. Such a new type of representation naturally supports *lifting image algorithms for video processing*, in the way of directly applying the established algorithm on the canonical image (*i.e.*, rendered from the canonical content field) and then propagating the results along the time axis through the temporal deformation field.

with the domain of consistent video editing [14, 15, 18, 22], which predominantly features two primary approaches: propagation-based methods and layered representation-based techniques. Propagation-based methods [12–14, 43, 53, 59] center on editing an initial frame and subsequently disseminating those edits throughout the video sequence. While this approach offers advantages in terms of computational efficiency and simplicity, it may be prone to inaccuracies and inconsistencies during the propagation of edits, particularly in situations characterized by complex motion or occlusion. Conversely, layered representation-based techniques [15, 22, 23, 40, 47] entail decomposing a video into distinct layers, thereby facilitating greater control and flexibility during the editing process. Text2Live [1] introduces the application of CLIP [37] models for video editing by modifying an optimized atlas [15] using text inputs, thereby yielding temporally consistent video editing results. Our work bears similarities to Text2Live in the context of employing an optimized representation for videos. However, our methodology diverges in several aspects: we optimize a more semantically-aware canonical representation incorporating a hash-based deformable design and attain higher-fidelity video processing.

Video Processing via Generative Models. The advancement of diffusion models has markedly enhanced the synthesis quality of text-to-image generation [6, 10, 50], surpassing the performance of prior methodologies [24, 41, 64, 66]. State-of-the-art diffusion models, such as GLIDE [29], Dall-E 2 [38, 39], Stable Diffusion [42], and Imagen [45], have been trained on millions of images, resulting in exceptional generative capabilities. While existing text-to-image (T2I) models enable free-text generation, incorporating additional conditioning factors [2, 3, 8, 27, 44, 54, 57, 67] such as edge, depth map, and normal map is essential for achieving precise control. In an effort to enhance controllability, researchers have proposed several approaches. InstructPix2Pix [2], on the other hand, fine-tunes T2I models using synthesized image condition pairs. ControlNet [67] introduces additional control conditions for Stable Diffusion through an auxiliary branch, thereby generating images that faithfully adhere to input condition maps. A recent research direction concentrates on the processing of videos utilizing text-to-image (T2I) models exclusively. Approaches like Tune-A-Video [63], Text2Video-Zero [16], FateZero [35], Vid2Vid-Zero [58], and Video-P2P [21] explore the latent space of DDIM [50] and incorporate cross-frame attention maps to facilitate consistent generation. Nevertheless, these methods may experience compromised temporal consistency due to the inherent randomness of generation, and the control condition may not be achieved with precision.

Text-to-video generation has emerged as a prominent research area in recent years, with prevalent approaches encompassing the training of diffusion models or autoregressive transformers on extensive datasets. Although text-to-video architectures such as NUWA [62], CogVideo [11], Phenaki [55], Make-A-Video [48], Imagen Video [9], and Gen-1 [7] are capable of generating video frames that semantically correspond to the input text, they may exhibit limitations in terms of precise control over video conditions or low resolution due to substantial computational demands.

3. Method

Problem Formulation. Given a video V comprised of frames $\{I_1, I_2, ..., I_N\}$, one can naively apply the image processing algorithm $\mathcal{X}$ to each frame individually for corresponding video tasks, yet may observe undesirable

inconsistencies across frames. An alternative strategy involves enhancing algorithm $\mathcal{X}$ with a temporal module, which requires additional training on video data. However, simply introducing a temporal module is hard to guarantee theoretical consistency and may result in performance degradation due to insufficient training data.

Motivated by these challenges, we propose representing a video $\mathcal{V}$ using a flattened canonical image I_c and a deformation field $\mathcal{D}$. By applying the image algorithm $\mathcal{X}$ on I_c, we can effectively propagate the effect to the whole video with the learned deformation field. This novel video representation serves as a crucial bridge between image algorithms and video tasks, allowing directly lifting of state-of-the-art image methodologies to video applications.

The proposed representations ought to exhibit the following essential characteristics:

- **Fitting Capability for Faithful Video Reconstruction.** The representation should possess the ability to accurately fit large rigid or non-rigid deformations in videos.
- **Semantic Correctness of the Canonical Image.** A distorted or semantically incorrect canonical image can lead to decreased image processing performance, especially considering that most of these processes are trained on natural image data.
- **Smoothness of the Deformation Field.** The ensurance of the smoothness in the deformation field is an essential feature that guarantees temporal consistency and correct propagation.

3.1. Content Deformation Fields

Inspired by the dynamic NeRFs [31, 32], we propose to represent the video in two distinct components: the canonical field and the deformation field. These two components are realized through the employment of a 2D and a 3D hash table, respectively. To enhance the capacity of these hash tables, two minuscule MLPs are integrated. We present our proposed representation for reconstructing and processing videos, as illustrated in Fig. 2. Given a video $\mathcal{V}$ comprising frames $\{I_1, I_2, ..., I_N\}$, we train an implicit deformable model tailored to fit these frames. The model is composed of two coordinate-based MLPs: the deformation field $\mathcal{D}$ and the canonical field $\mathcal{C}$.

The canonical field $\mathcal{C}$ serves as a continuous representation encompassing all flattened textures present in the video $\mathcal{V}$. It is defined by a function $\mathbf{F} : \mathbf{x} \rightarrow \mathbf{c}$, which maps a 2D position $\mathbf{x} : (x, y)$ to a color $\mathbf{c} : (r, g, b)$. In order to speed up the training and enable the network to capture the high-frequency details, we adopt the multi-resolution hash encoding $\gamma_{2\text{D}} : \mathbb{R}^2 \rightarrow \mathbb{R}^{2+F\times L}$ to map the coordinate $\mathbf{x}$ into a feature vector, where L is the number of levels for multi-resolution and F is the number of feature dimensions for per layer. The function $\gamma_{2\text{D}}(\mathbf{x}) = (\mathbf{x}, \mathbf{F}_1(\mathbf{x}), ..., \mathbf{F}_L(\mathbf{x}))$ facilitates the model's ability to capture high-frequency details, where $\mathbf{F}_i(\mathbf{x})$ is the features linearly interpolated by $\mathbf{x}$ at i^{th} resolution. The deformation field $\mathcal{D}$ captures the observation-to-canonical deformation for every frame within a video. For a specific frame I_i, $\mathcal{D}$ establishes the correspondence between the observed and canonical positions. Dynamic NeRFs [31, 32] implement the deformation field in 3D space by using the Fourier positional encoding and an extra learnable time code. This implementation ensures the smoothness of the deformation field. Nevertheless, this straightforward implementation can not be seamlessly transferred into video representation for two reasons (*i.e.* low training efficiency and inadequate representative capability). Therefore, we propose to represent the deformation field as a 3D hash table with a tiny MLP following. Specifically, an arbitrary position $\mathbf{x}$ in the t^{th} frame is first encoded by a 3D hash encoding function $\gamma_{3\text{D}}(\mathbf{x}, t)$ to get high-dimension features. Then a tiny MLP $\mathcal{D} : (\gamma_{3\text{D}}(\mathbf{x}, t)) \rightarrow \mathbf{x}'$ maps the embedded features its corresponding position $\mathbf{x}'$ in canonical field. We elaborate our 3D hash encoding based deformation field in detail as follows.

3D Hash Encoding for Deformation Field. Specifically, an arbitrary point in the video can be conceptualized as a position $\mathbf{x}_{3\text{D}} : (x, y, t)$ within an orthogonal 3D space.We represent our video space using the 3D hash encoding technique, as depicted on the left side of Fig. 2. This technique encapsulates the 3D space as a multi-resolution feature grid. The term *multi-resolution* refers to a composition of grids with varying degrees of resolution, and *feature grid* denotes a grid populated with learnable features at each vertex. In our framework, the multi-resolution feature grid is organized into L distinct levels. The dimensionality of the learnable features is represented as F. Furthermore, the resolution of the l^{th} layer, denoted as N_l, exhibits a geometric progression between the coarsest and finest resolutions, denoted collectively as $[N_{\text{min}}, N_{\text{max}}]$, using

$$N_l = \lfloor N_{\text{min}} \cdot b^l \rfloor, b = \exp\left(\frac{\ln N_{\text{max}} - \ln N_{\text{min}}}{L-1}\right). \quad (1)$$

Considering the queried points $\mathbf{x}_{3\text{D}}$ at l^{th} layer, the input coordinate is scaled by that level's grid resolution. And the queried features of $\mathbf{x}_{3\text{D}}$ are tri-linear interpolated from its 8-neighboring corner points(seen in Fig. 2). For attaining the corner points of $\mathbf{x}_{3\text{D}}$, rounding down and up are first operated as

$$\lfloor \mathbf{x}^l_{3\text{D}} \rfloor = \lfloor \mathbf{x}_{3\text{D}} \cdot N_l \rfloor, \lceil \mathbf{x}^l_{3\text{D}} \rceil = \lceil \mathbf{x}_{3\text{D}} \cdot N_l \rceil, \quad (2)$$

and we map its each corner to an entry in the level's respective feature vector array, which has fixed size of at most T. For the coarse level, the parameters of low resolution grid are fewer than T, where the mapping is $1:1$. Thus, the features can be directly looked up by its index. On

the contrary. For the finer resolution, the point is mapped by the hash function,

$$h(\mathbf{x}^l_{\text{3D}}) = \left(\oplus_{i=1}^{d} x_i \pi_i\right) \mod T, \tag{3}$$

where $\oplus$ denotes the bit-wise XOR operation and $\{\pi_i\}$ are unique large prime numbers following [28].

The output color value at coordinate $\mathbf{x}$ for frame t can be computed as

$$\mathbf{c} = \mathcal{C}(\mathcal{D}(\gamma_{\text{3D}}(\mathbf{x}, t))). \tag{4}$$

This output can be supervised using the ground truth color present in the input frame.

3.2. Model Design

The proposed representation can effectively model and reconstruct both the canonical content and the temporal deformation for an arbitrary video. However, it faces challenges in meeting the requirements for robust video processing. In particular, while 3D hash deformation possesses powerful fitting capability, it compromises the smoothness of temporal deformation. This trade-off makes it notably difficult to maintain the inherent semantics of the canonical image, creating a significant barrier to the adaptation of established image algorithms for video use. To achieve precise video reconstruction while preserving the inherent semantics of the canonical image, we propose the use of annealed multi-resolution hash encoding. To further enhance the smoothness of deformation, we introduce flow-based consistency. In challenging cases, such as those involving large occlusions or complex multi-object scenarios, we suggest utilizing additional semantic information. This can be achieved by using semantic masks in conjunction with the grouped deformation fields.

Annealed 3D Hash Encoding for Deformation. For the finer resolution, the hash encoding enhance the complex deformation fitting performance but introducing the discontinuity and distortion in canonical field (Seen in Fig. 7). Inspired by the annealed strategy utilized in dynamic NeRFs [31], we employ the annealed hash encoding technique for progressive frequency filter for deformation. More specifically, we use a progressive controlling weights for those features interpolated in different resolution. The weight for the j^{th} layer in training step k is computed as

$$w_j(k) = \frac{1 - \cos(\pi \cdot \text{clamp}(m(k - N_{\text{beg}})/N_{\text{step}} - j, 0, 1))}{2}, \tag{5}$$

where N_{beg} is a predefined step for beginning annealing and m represents a hyper parameters for controlling the annealing speed, and N_{step} is the number for annealing step.

Flow-guided Consistency Loss. Corresponding points identified by flows with high confidence should be the same points in the canonical field. We compute the flow-guided consistency loss according to this observation. For two consecutive frames I_i and I_{i+1}, we employ RAFT[52] to detect the forward flows $\mathcal{F}_{i\to i+1}$ and backward flows $\mathcal{F}_{i+1\to i}$. The confident region of a frame I_i can be defined as

$$M_{\text{flow}} = |\text{Warp}(\text{Warp}(I_i, \mathcal{F}_{i\to i+1}), \mathcal{F}_{i+1\to i}) - I_i| < \epsilon, \tag{6}$$

where ϵ represents a hyperparameter for the error threshold.

This loss can be formulated as

$$\mathcal{L}_{\text{flow}} = \sum \|\mathcal{D}(\gamma_{3D}(\mathbf{x}, t)) - \mathcal{D}(\gamma_{3D}(\mathbf{x} + \mathcal{F}^{\mathbf{x}}_{t\to t+1}, t+1)) - \mathcal{F}^{\mathbf{x}}_{t\to t+1}\| * M^{\mathbf{x}}_{\text{flow}}, \tag{7}$$

where $\mathcal{F}^{\mathbf{x}}_{t\to t+1}$ and $M^{\mathbf{x}}_{\text{flow}}$ are the optical flow and the flow confidence at $\mathbf{x}$. The flow loss efficiently regularize the smoothness of the deformation field especially for the smooth region.

Grouped Content Deformation Fields. Although the representation can learn to reconstruct a video using a single content deformation field, complex motions arising from overlapped multi-objects may lead to conflicts within one canonical. Consequently, the boundary region might suffer from inaccurate reconstruction. For challenging instances featuring large occlusions, we propose an option to introduce the layers corresponding to multiple content deformation fields. These layers would be defined based on semantic segmentation, thereby improving the accuracy and robustness of video reconstruction in these demanding scenarios. We leverage the Segment-Anything-track (SAM-track) [5] to attain the segmentation of each video frame

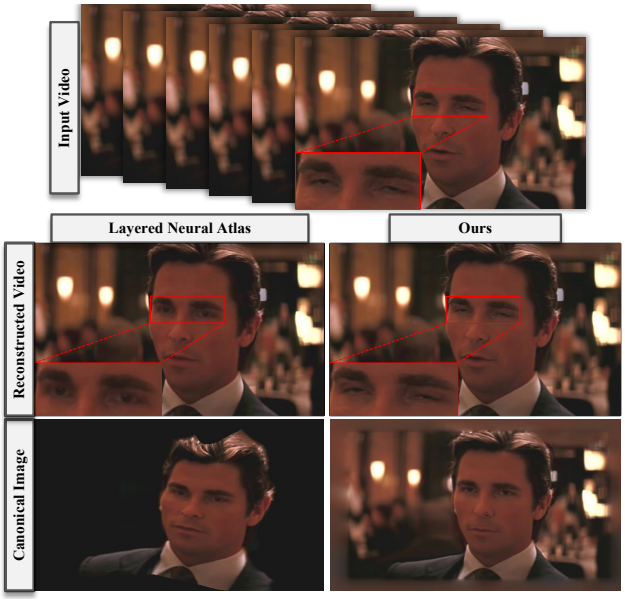

Figure 3. **Qualitative comparison** between layered neural atlas [15] and our CoDeF regarding *video reconstruction*, which reflects the capacity of the video representation and also plays a fundamental role in faithful video processing. Details are best appreciated when zoomed in.

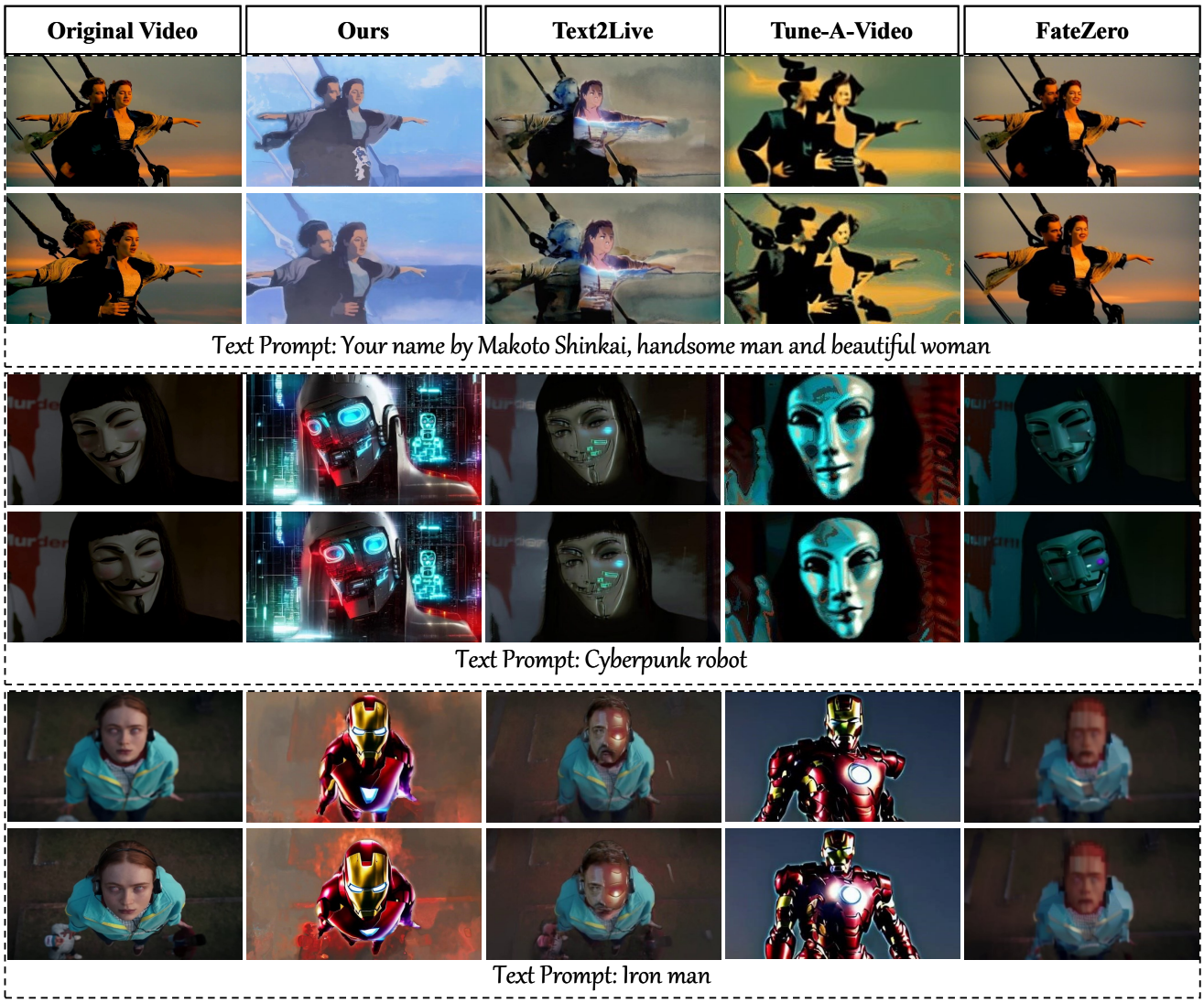

Figure 4. **Qualitative comparison** on the task of *text-guided video-to-video translation* across different methods, including Text2Live [1], Tune-A-Video [63], FateZero [35], and directly lifting ControlNet [67] through our `CoDeF`. We strongly encourage the readers to see the supplementary videos for a detailed evaluation of temporal consistency and synthesis quality.

I_i into K semantic layers with mask $M_0^i, ..., M_{K-1}^i$. And for each layer, we use a group of canonical fields and deformation fields to represent those separate motion of different objects. These models are subsequently formulated as groups of implicit fields: $\mathcal{D} : \{\mathcal{D}_1, ..., \mathcal{D}_K\}, \mathcal{C} : \{\mathcal{C}_1, ..., \mathcal{C}_K\}$. In theory, for semantic layer k in frame i, it is sufficient to sample pixels in the region M_k^i for efficient reconstruction. However, hash encoding can result in random and unstructured patterns in unsupervised regions, which decreases the performance of image-based models trained on natural images. To tackle this issue, we sample a number of points outside of the region M_k^i and train them using L_2 loss with the ground truth color. In this way, we effectively regularize $\bar{M}_k^i$ with the background loss $\mathcal{L}_{\text{bg}}$. Consequently, the canonical image attains a more natural appearance, leading to enhanced processing results.

Training Objectives. The representation is trained by minimizing the objective function $\mathcal{L}_{\text{rec}}$. This function corresponds to the L_2 loss between the ground truth color and the predicted color $\mathbf{c}$ for a given coordinate $\mathbf{x}$. To regularize and stabilize the training process, we introduce additional regularization terms as previously discussed. The total loss is calculated using the following equation

$$\mathcal{L} = \mathcal{L}_{\text{rec}} + \lambda_1 * \mathcal{L}_{\text{flow}}, \tag{8}$$

where λ_1 represents the hyper-parameters for loss weights. It's important to note that when training the grouped deformation field, we include an additional regularizer, denoted as $\lambda_2 * \mathcal{L}_{\text{bg}}$.

Figure 5. **Video object tracking** results achieved by *lifting* an image segmentation algorithm [17] through our CoDeF.

3.3. Application to Consistent Video Processing

Upon the optimization of the content deformation field, the canonical image I_c is retrieved by setting the deformation of all points to zero. It is important to note that the size of the canonical image can be flexibly adjusted to be larger than the original image size depending on the scene movement observed in the video, thereby allowing more content to be included. The canonical image I_c is then utilized in executing various downstream algorithms for consistent video processing. We evaluated the following state-of-the-art (SOTA) algorithms: *(1) ControlNet* [67]: Used for prompt-guided video-to-video translation. *(2) Segment-anything (SAM)* [17]: Applied for video object tracking. *(3) R-ESRGAN* [60]: Employed for video super-resolution. Additionally, the canonical image allows users to conveniently edit the video by directly modifying the image. We further illustrate this capability through multiple manual video editing examples.

4. Experiments

4.1. Experimental Setup

We conduct experiments to underscore the robustness and versatility of our proposed method. Our representation is robust with a variety of deformations, encompassing rigid and non-rigid objects, as well as complex scenarios such as smog. The default parameters for our experiments are set with the anneal begin and end steps at 4000 and 8000, respectively. The total iteration step is capped at 10000. On a single NVIDIA A6000 GPU, the average training duration is approximately 5 minutes when utilizing 100 video frames. It should be noted that the training time varies with several factors such as the length of the video, the type of motion, and the number of layers. By adjusting the training parameters accordingly, the optimization duration can be varied from 1 to 10 minutes.

4.2. Evaluation

The evaluation of our representation is concentrated on two main aspects: the quality of the reconstructed video with the estimated canonical image, and the quality of downstream video processing. Owing to the lack of accurate evaluation metrics, conducting a precise quantitative analysis remains challenging. Nevertheless, we include a selection of quantitative results for further examination.

Table 1. **Quantitative comparison of video reconstruction.**

Video	LNA [15]	**CoDeF** (K=1)	**CoDeF** (K=2)	*w/o* flow
Blackswan	29.92	31.51	**31.97**	31.47
Boat	31.51	34.13	**34.73**	34.28
Kite-surf	28.37	34.26	**34.35**	34.14

Reconstruction Quality. In a comparative analysis with the Neural Image Atlas, our model, as demonstrated in Fig. 3, exhibits robustness to non-rigid motion, effectively reconstructing subtle movements with heightened precision (*e.g.* eyes blinking, face textures). Quantitatively, the video reconstruction PSNR of our algorithm on DAVIS video datasets is 2.3 dB higher and we report the metrics of some selected sequences in Tab. 1. In comparison between the atlas and our canonical image, our results provide a more natural representation, and thus, facilitate the easier application of established image algorithms. Besides, our method makes a significant progress in training efficiency, *i.e.*, 5 minutes (ours) *vs.* 10 hours (atlas).

Downstream Video Processing. We provide an expanded range of potential applications associated with the proposed representations, including video-to-video translation, video keypoint tracking, video object tracking, video super-resolution. (More results are attached in the supplement.)

Video-to-video Translation. By applying image translation to the canonical image, we can perform video-to-video translation. A qualitative comparison is presented encompassing several baseline methods that fall into three distinct categories: (1) per-frame inference with image translation models, such as ControlNet [67]; (2) layered video editing, exemplified by Text2LIVE [1]; and (3) diffusion-based video translation, including Tune-A-Video [63] and FateZero [35]. As depicted in Fig. 4, the per-frame image translation models yield high-fidelity content, accompanied by significant flickering. The alternative baselines exhibit compromised generation quality or comparatively low temporal consistency. A thorough comparison is better appreciated by viewing the supplementary videos.

We further provide a quantitative comparison in Tab. 2, using the same settings for the CLIP score [37] as in Tune-

Table 2. **Quantitative comparison with baselines.** *Note that Text2LIVE [1] is optimized with CLIP loss.

Method	Frame Consisitency		Textual Alignment	
	CLIP Score	User Score	CLIP Score	User Score
Text2LIVE* [1]	**99.17**	4.21	**30.72**	2.32
FateZero [35]	96.75	3.75	23.21	3.34
Tune-A-Video [63]	94.40	2.12	26.02	3.89
CoDeF	98.54	**4.76**	27.43	**4.13**

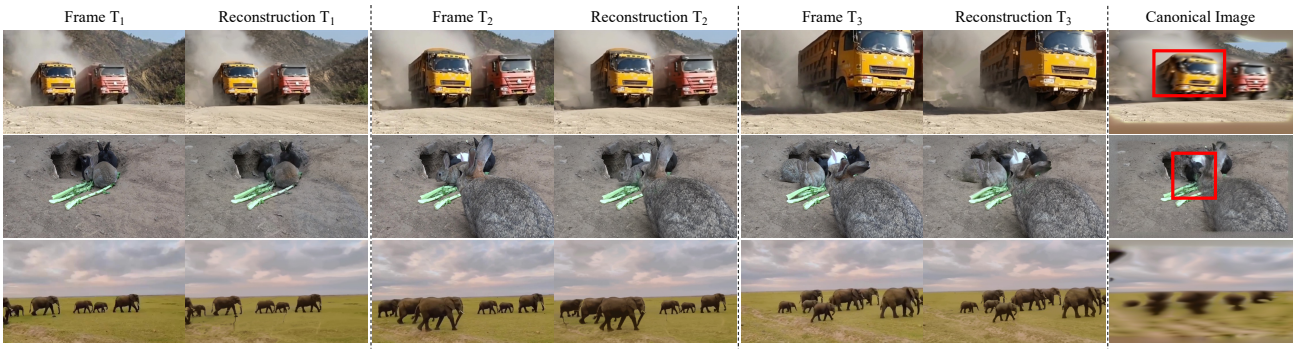

Figure 6. **Limitations** on more complicated video seqeunces.

Figure 7. **Ablation study** on the effectiveness of annealed hash. The unnaturalness in the canonical image will harm the performance of downstream tasks.

A-Video [63]. As Text2LIVE [1] is optimized using CLIP loss, it is inevitable that the CLIP-based measurement is higher. For the user study, we present participants with both the original and translated videos alongside the text prompt. Participants are asked to rate the translation results based on temporal consistency and textual alignment on a scale from 1 to 5, with 5 being the best possible score. From the 1120 responses gathered from 56 users, our method outperforms all other baselines.

Video Object Tracking. Using the segmentation algorithms on the canonical image, we are able to facilitate the propagation of masks throughout all video frames. As illustrated in Fig. 5, our pipeline proficiently yields masks that maintain consistency across all frames.

4.3. Ablation Study

To validate the effect of the proposed modules, we conduct an ablation study. On substituting the 3D hash encoding with positional encoding, there is a notable decrease in the reconstruction PSNR of the video by 3.1 dB. In the absence of the annealed hash, the canonical image loses its natural appearance, as evidenced by the presence of multiple hands in Fig. 7. Furthermore, without incorporating the flow loss, smooth areas are noticeably affected by pronounced flickering. We also investigate how varying the group number K affects outcomes in Tab. 1, which shows that larger K results in better reconstruction results.

4.4. Limitation

We address the limitations of our method with examples from the OVIS [36] dataset. Fig. 6 clearly showcases the method's shortcomings in creating a coherent canonical image of complex scenes. We show three challenging examples: a significant scale change from distant to close-up views (as with "Trucks"), the sudden emergence of new objects within the scene ("Rabbits"), and the presence of multiple, rapidly-moving small entities ("Elephants"). In these tests, our approach can occasionally yield a blurred canonical representation, omit finer details, or fail to maintain the semantic integrity of the objects' shapes.

5. Conclusion and Discussion

In this paper, we investigate representing videos as content deformation fields, focusing on achieving temporally consistent video processing. Our approach demonstrates promising results in terms of both fidelity and temporal consistency. However, there remain several challenges to be addressed in future work. One of the primary issues pertains to the per-scene optimization required in our methodology. We expect that improvements in feed-forward implicit field methods could be used in this area. Another challenge arises in scenarios involving extreme changes in viewing points. To tackle this issue, the incorporation of 3D prior knowledge may prove beneficial, as it can provide additional information and constraints. Lastly, the handling of large non-rigid deformations remains a concern. To address this, one potential solution involves employing multiple canonical images. Despite the challenges, our work provide great potential for future improvement.

References

[1] Omer Bar-Tal, Dolev Ofri-Amar, Rafail Fridman, Yoni Kasten, and Tali Dekel. Text2live: Text-driven layered image and video editing. In *Eur. Conf. Comput. Vis.*, 2022. 3, 6, 7, 8

[2] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023. 3

[3] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. *arXiv preprint arXiv:2304.08465*, 2023. 3

[4] Yinbo Chen, Sifei Liu, and Xiaolong Wang. Learning continuous image representation with local implicit image function. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021. 2

[5] Yangming Cheng, Liulei Li, Yuanyou Xu, Xiaodi Li, Zongxin Yang, Wenguan Wang, and Yi Yang. Segment and track anything. *arXiv preprint arXiv:2305.06558*, 2023. 5

[6] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *Adv. Neural Inform. Process. Syst.*, 2021. 3

[7] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. *arXiv preprint arXiv:2302.03011*, 2023. 3

[8] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 3

[9] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 3

[10] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Adv. Neural Inform. Process. Syst.*, 2020. 3

[11] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022. 3

[12] Allan Jabri, Andrew Owens, and Alexei Efros. Space-time correspondence as a contrastive random walk. In *Adv. Neural Inform. Process. Syst.*, 2020. 3

[13] Varun Jampani, Raghudeep Gadde, and Peter V Gehler. Video propagation networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017.

[14] Ondřej Jamriška, Šárka Sochorová, Ondřej Texler, Michal Lukáč, Jakub Fišer, Jingwan Lu, Eli Shechtman, and Daniel Sỳkora. Stylizing video by example. *ACM Trans. Graph.*, 38(4):1–11, 2019. 3

[15] Yoni Kasten, Dolev Ofri, Oliver Wang, and Tali Dekel. Layered neural atlases for consistent video editing. *ACM Trans. Graph.*, 40(6):1–12, 2021. 2, 3, 5, 7

[16] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. *arXiv preprint arXiv:2303.13439*, 2023. 3

[17] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 7

[18] Chenyang Lei, Yazhou Xing, Hao Ouyang, and Qifeng Chen. Deep video prior for video consistency and propagation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(1):356–371, 2022. 3

[19] Tianye Li, Mira Slavcheva, Michael Zollhoefer, Simon Green, Christoph Lassner, Changil Kim, Tanner Schmidt, Steven Lovegrove, Michael Goesele, Richard Newcombe, et al. Neural 3d video synthesis from multi-view video. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022. 2

[20] Zhengqi Li, Simon Niklaus, Noah Snavely, and Oliver Wang. Neural scene flow fields for space-time view synthesis of dynamic scenes. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021. 2

[21] Shaoteng Liu, Yuechen Zhang, Wenbo Li, Zhe Lin, and Jiaya Jia. Video-p2p: Video editing with cross-attention control. *arXiv preprint arXiv:2303.04761*, 2023. 3

[22] Erika Lu, Forrester Cole, Tali Dekel, Weidi Xie, Andrew Zisserman, David Salesin, William T Freeman, and Michael Rubinstein. Layered neural rendering for retiming people in video. *arXiv preprint arXiv:2009.07833*, 2020. 2, 3

[23] Erika Lu, Forrester Cole, Weidi Xie, Tali Dekel, Bill Freeman, Andrew Zisserman, and Michael Rubinstein. Associating objects and their effects in video through coordination games. In *Adv. Neural Inform. Process. Syst.*, 2022. 3

[24] Elman Mansimov, Emilio Parisotto, Jimmy Lei Ba, and Ruslan Salakhutdinov. Generating images from captions with attention. *arXiv preprint arXiv:1511.02793*, 2015. 3

[25] Mateusz Michalkiewicz, Jhony K Pontes, Dominic Jack, Mahsa Baktashmotlagh, and Anders Eriksson. Implicit surface representations as layers in neural networks. In *Int. Conf. Comput. Vis.*, 2019. 2

[26] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *Eur. Conf. Comput. Vis.*, 2020. 2

[27] Chong Mou, Xintao Wang, Liangbin Xie, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*, 2023. 3

[28] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph.*, 41(4):102:1–102:15, 2022. 2, 5

[29] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 3

[30] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019. 2

[31] Keunhong Park, Utkarsh Sinha, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Steven M Seitz, and Ricardo

Martin-Brualla. Nerfies: Deformable neural radiance fields. In *Int. Conf. Comput. Vis.*, 2021. 2, 4, 5

[32] Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M. Seitz. Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. *ACM Trans. Graph.*, 40(6), 2021. 2, 4

[33] Songyou Peng, Michael Niemeyer, Lars Mescheder, Marc Pollefeys, and Andreas Geiger. Convolutional occupancy networks. In *Eur. Conf. Comput. Vis.*, 2020. 2

[34] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021. 2

[35] Chenyang Qi, Xiaodong Cun, Yong Zhang, Chenyang Lei, Xintao Wang, Ying Shan, and Qifeng Chen. Fatezero: Fusing attentions for zero-shot text-based video editing. *arXiv preprint arXiv:2303.09535*, 2023. 2, 3, 6, 7

[36] Jiyang Qi, Yan Gao, Yao Hu, Xinggang Wang, Xiaoyu Liu, Xiang Bai, Serge Belongie, Alan Yuille, Philip HS Torr, and Song Bai. Occluded video instance segmentation: A benchmark. *International Journal of Computer Vision*, 130(8):2022–2039, 2022. 8

[37] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Int. Conf. Mach. Learn.*, 2021. 3, 7

[38] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 3

[39] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *Int. Conf. Mach. Learn.*, 2021. 3

[40] Alex Rav-Acha, Pushmeet Kohli, Carsten Rother, and Andrew Fitzgibbon. Unwrap mosaics: A new representation for video editing. In *SIGGRAPH*, pages 1–11, 2008. 2, 3

[41] Scott E Reed, Zeynep Akata, Santosh Mohan, Samuel Tenka, Bernt Schiele, and Honglak Lee. Learning what and where to draw. In *Adv. Neural Inform. Process. Syst.*, 2016. 3

[42] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022. 3

[43] Manuel Ruder, Alexey Dosovitskiy, and Thomas Brox. Artistic style transfer for videos. In *Pattern Recognition: 38th German Conference, GCPR 2016, Hannover, Germany, September 12-15, 2016, Proceedings 38*, pages 26–36. Springer, 2016. 3

[44] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023. 3

[45] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. In *Adv. Neural Inform. Process. Syst.*, 2022. 3

[46] Sara Fridovich-Keil and Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022. 2

[47] Jonathan Shade, Steven Gortler, Li-wei He, and Richard Szeliski. Layered depth images. In *Proceedings of the 25th annual conference on Computer graphics and interactive techniques*, 1998. 2, 3

[48] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022. 3

[49] Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. In *Adv. Neural Inform. Process. Syst.*, 2020. 2

[50] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 3

[51] Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. In *Adv. Neural Inform. Process. Syst.*, 2020. 2

[52] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Eur. Conf. Comput. Vis.*, 2020. 5

[53] Ondřej Texler, David Futschik, Michal Kučera, Ondřej Jamriška, Šárka Sochorová, Menclei Chai, Sergey Tulyakov, and Daniel Sỳkora. Interactive video stylization using few-shot patch-based training. *ACM Trans. Graph.*, 39(4):73–1, 2020. 3

[54] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023. 3

[55] Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, and Dumitru Erhan. Phenaki: Variable length video generation from open domain textual description. *arXiv preprint arXiv:2210.02399*, 2022. 3

[56] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689*, 2021. 2

[57] Tengfei Wang, Ting Zhang, Bo Zhang, Hao Ouyang, Dong Chen, Qifeng Chen, and Fang Wen. Pretraining is all you need for image-to-image translation. *arXiv preprint arXiv:2205.12952*, 2022. 3

[58] Wen Wang, Kangyang Xie, Zide Liu, Hao Chen, Yue Cao, Xinlong Wang, and Chunhua Shen. Zero-shot video editing using off-the-shelf image diffusion models. *arXiv preprint arXiv:2303.17599*, 2023. 3

[59] Xiaolong Wang, Allan Jabri, and Alexei A Efros. Learning correspondence from the cycle-consistency of time. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019. 3

[60] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *Int. Conf. Comput. Vis.*, 2021. 7

[61] Yiming Wang, Qin Han, Marc Habermann, Kostas Daniilidis, Christian Theobalt, and Lingjie Liu. Neus2: Fast learning of neural implicit surfaces for multi-view reconstruction. *arXiv preprint arXiv:2212.05231*, 2022. 2

[62] Chenfei Wu, Jian Liang, Lei Ji, Fan Yang, Yuejian Fang, Daxin Jiang, and Nan Duan. Nüwa: Visual synthesis pre-training for neural visual world creation. In *Eur. Conf. Comput. Vis.*, 2022. 3

[63] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Weixian Lei, Yuchao Gu, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. *arXiv preprint arXiv:2212.11565*, 2022. 2, 3, 6, 7, 8

[64] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018. 3

[65] Vickie Ye, Zhengqi Li, Richard Tucker, Angjoo Kanazawa, and Noah Snavely. Deformable sprites for unsupervised video decomposition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022. 2

[66] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Int. Conf. Comput. Vis.*, 2017. 3

[67] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023. 2, 3, 6, 7