Vision Study

HyperCon: Image-to-Video Model Transfer for Video-To-Video Translation Tasks

WACV 2021

Yeojeong Park

https://arxiv.org/pdf/1912.04950.pdf

Contents

Introduction

HyperCon

Experiments : Video Style Transfer

Experiments : Video Inpainting

Conclusion

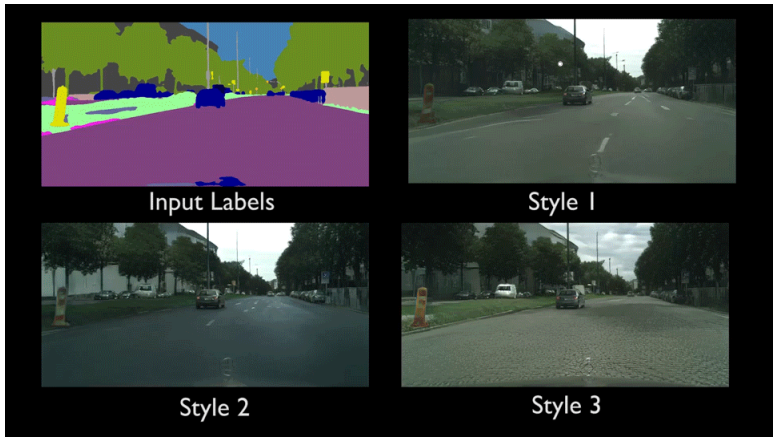
Video-to-Video translation

Video inpainting

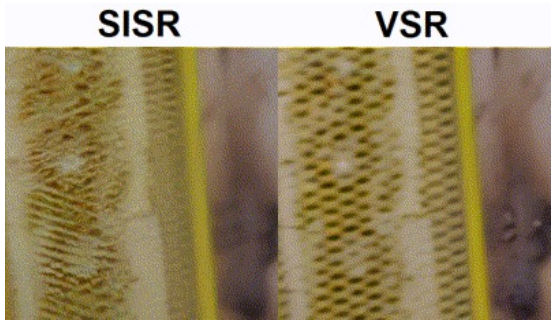
Video Style Transfer

Video tasks' goal:
1) Satisfy the intended translation
2) No flickering artifacts

→ Temporal-consistent!



Vid2Vid synthesis



Video Super Resolution

Related works

Techniques to address temporal consistency

1) Optical flow-based loss / Network layer(3D, RNN)

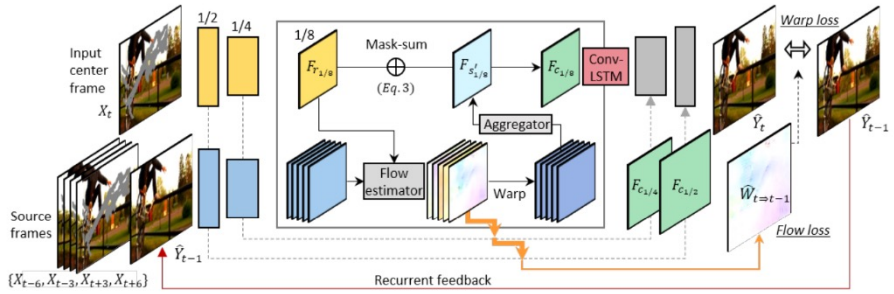


Figure 2. **The overview of VINet.** Our network takes in multiple frames ($X_{t-6}, X_{t-3}, X_t, X_{t+3}, X_{t+6}$) and the previously generated frame ($\hat{Y}_{t-1}$), and generates the inpainted frame ($\hat{Y}_t$) as well as the flow map ($\hat{W}_{t\Rightarrow t-1}$). We employ both flow sub-networks and mask sub-networks at 4 scales (1/8, 1/4, 1/2, and 1) to aggregate and synthesize feature points progressively. For temporal consistency, we use a recurrent feedback and a temporal memory layer (ConvLSTM) along with two losses: flow loss and warp loss. The orange arrows denote the $\times 2$ upsampling for residual flow learning as in [25] for 5 streams, while the thinner orange arrow is for only the stream from $\hat{Y}_{t-1}$. The mask sub-networks are omitted in the figure for the simplicity.

Kim et al., Deep Video Inpainting

- Relevant losses (e.g. recon loss, style loss)
- Require models and losses that are defined exclusively on videos/tasks

2) A Blind video consistency model

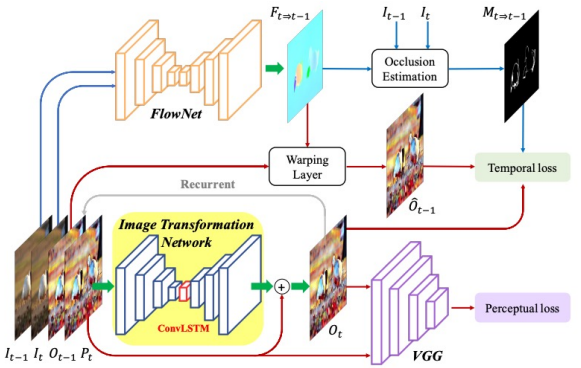


Fig. 2: **Overview of the proposed method.** We train an image transformation network that takes I_{t-1}, I_t, O_{t-1} and processed frame P_t as inputs and generates the output frame O_t which is temporally consistent with the output frame at the previous time step O_{t-1}. The output O_t at the current time step then becomes the input at the next time step. We train the image transformation network with the VGG perceptual loss and the short-term and long-term temporal losses.

Lai et al., Learning Blind Video Temporal Consistency

- Task-independent approach
- Require dense correspondences btw unpro cessed frames
  → inappropriate to inpainting

Motivation

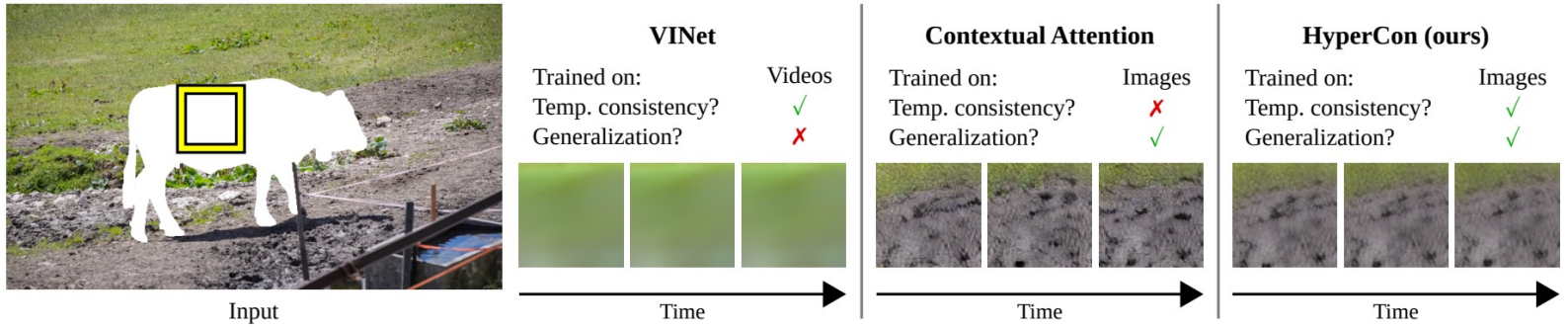


Figure 1: Video-to-video translation models designed and trained from scratch, *e.g.*, VINet [18] are temporally consistent, but exhibit poor generalization performance due to the limited size of high-fidelity video datasets (note the lack of defined texture). Image-to-image models, *e.g.*, Contextual Attention [37], generalize well thanks to large image datasets, but lack temporal consistency (note the changing texture). HyperCon leverages the generalization performance conferred by image datasets while enforcing the temporal consistency properties of video-to-video models.

- SOTA Video inpainting model are temporally consistent but fails to hallucinate a sensible text ure
- SOTA Image inpainting model produce realistic textures but exhibits temporal inconsistency
- Training with Large scale of Image dataset generalize better to new data.
- Goal: transform img-to-img model into a strong vid-to-vid model without finetuning
- Automatically induce temporal consistency while achieving the same visual effect as the img-to-img model

Contributions

- we motivate image-to-video model transfer as a way to leverage the superior generalization performance of image models for video-to-video translation without sacrificing temporal consistency.
- we propose **HyperCon**, which supports a wider span of tasks than prior video consistency work thanks to its support for both masked and unmasked inputs.
- we show that HyperCon performs favorably compared to state-of-the-art video consistency and inpainting methods without the need to be finetuned on these tasks.

HyperCon：overview

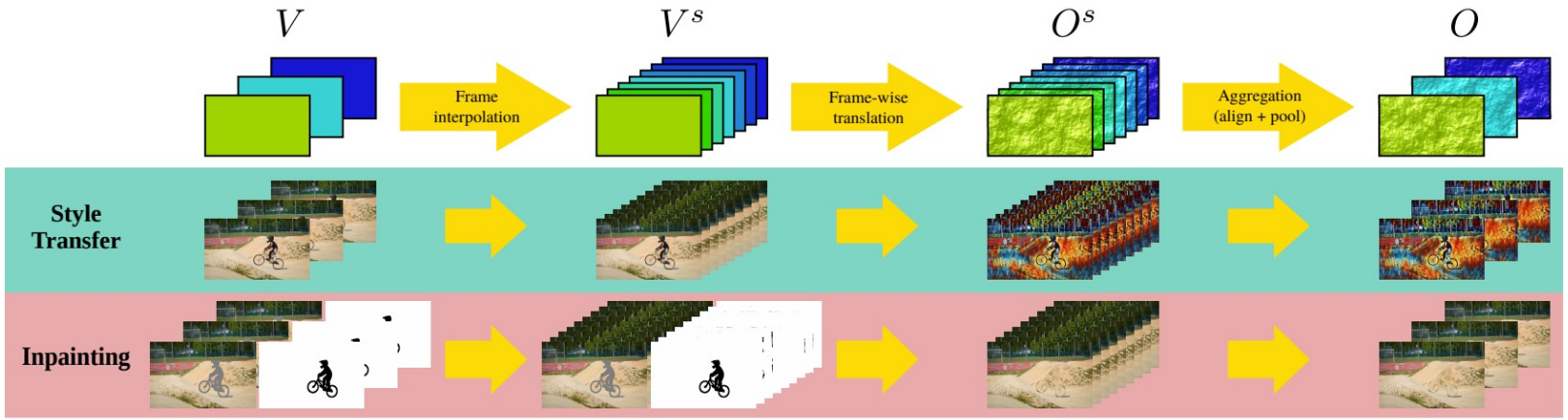


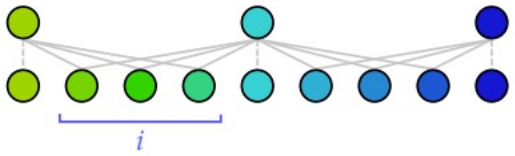
Figure 2: Visual overview of our Hyperconsistency (HyperCon) method. We begin by artificially inserting frames into the input video V with a frame interpolation network to produce an interpolated video V^s. Then, we independently translate each frame in the interpolated video with an image-to-image translation model. Finally, we aggregate frames (*i.e.*, align with optical flow and pool pixel-wise) within a local sliding window to produce the final temporally consistent output video O. This can be applied to tasks with or without masked inputs (*e.g.*, inpainting and style transfer, respectively).

1. Interpolating video：Insert i frames between each pair of frames
2. Translating：independently translate frames
3. Temporal Aggregation：aggregating frames within overlapping windows

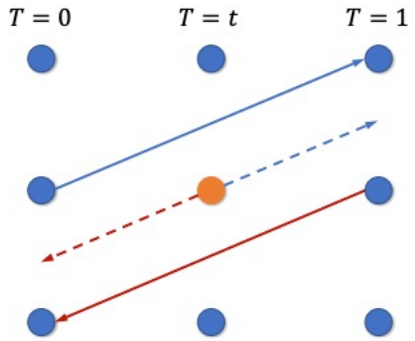
Generating the Interpolated Video

$consecutive\ frames\ v_a, v_{a+1}$
$intermediate\ frame\ index\ b\ \in \{1, \dots, i\}$
1. $predict\ two\ warping\ grids\ \left(F^s_{a+b' \to a}, F^s_{a+b' \to a+1}\right)$
2. $predict\ weight\ w_{a+b'}$

$$b' = \frac{b+1}{i+1}$$

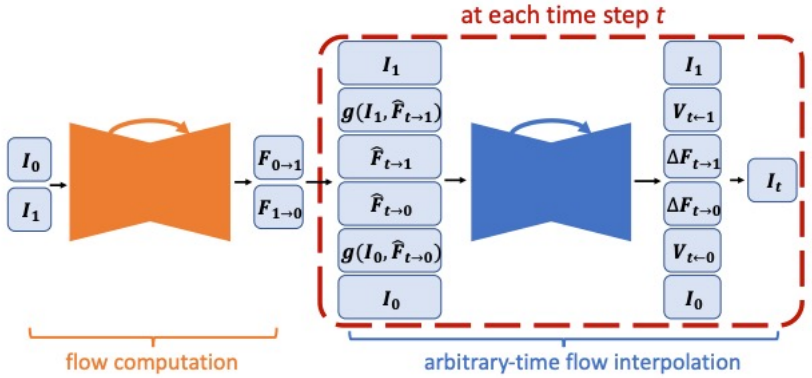


Interpolation (Section 2.1)



$$(F^s_{a+b' \to a}, F^s_{a+b' \to a+1}, w_{a+b'}) = \text{wrpgrd}(v_a, v_{a+1}, b'), \quad (1)$$

$$v^s_j = (1 - w_{a+b'}) \odot \text{warp}(v_a, F^s_{a+b' \to a}) + w_{a+b'} \odot \text{warp}(v_{a+1}, F^s_{a+b' \to a+1}). \quad (2)$$

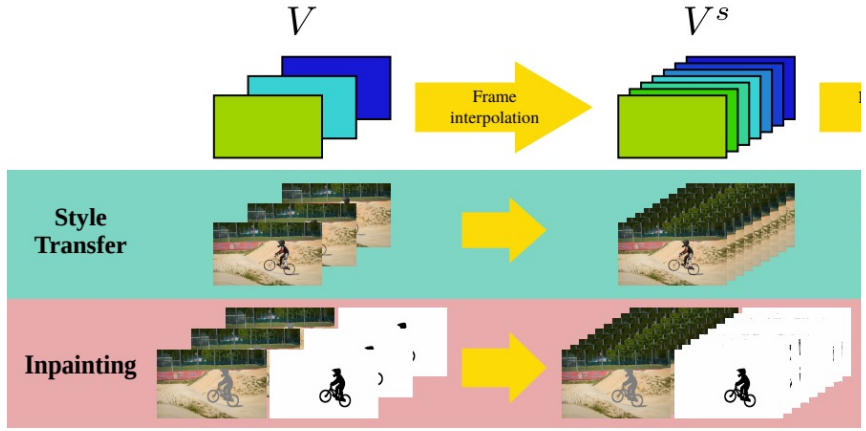


Iiang et al., Super SloMo : High Quality Estimation of Multiple Intermediate Frames for Video Interpolation(https://arxiv.org/pdf/1712.00080.pdf)

Generating the Interpolated Video

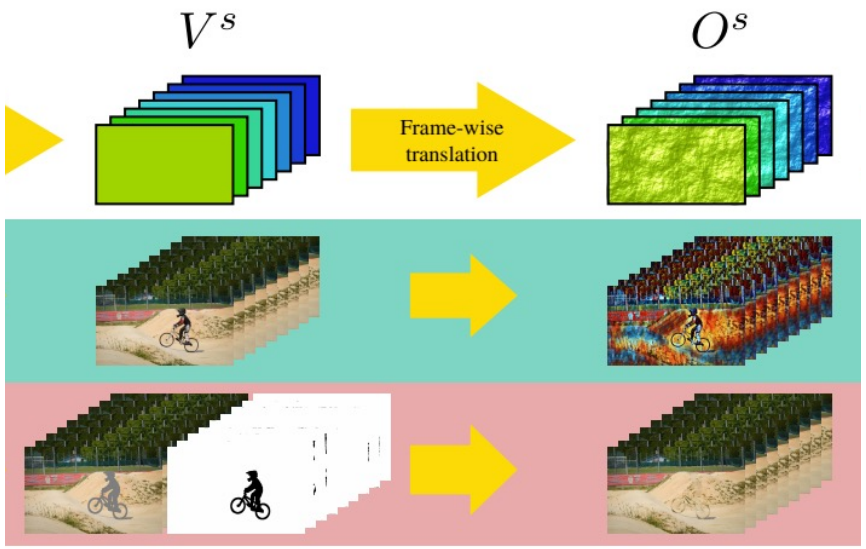
Input

After SloMo



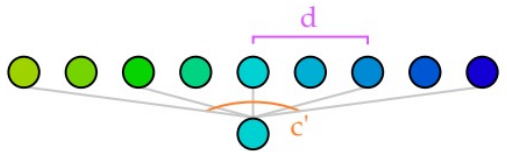
Iiang et al., Super SloMo : High Quality Estimation of Multiple Intermediate Frames for Video Interpolation(https://arxiv.org/pdf/1712.00080.pdf)

Translating the Interpolated Video



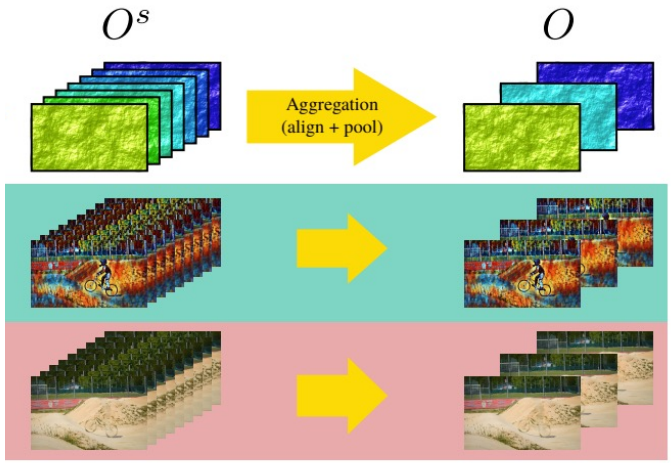
- Simply translate each frame in V^s independently
  - $o_j^s = g(v_j^s),\ \ j \in \{1, \ldots, N'\}$
- O^s is not temporally consistent yet.
- Most spatial regions will exhibit consensus within small temporal windows.
- We can remove artifacts by **mapping neighboring frames to one frame!**

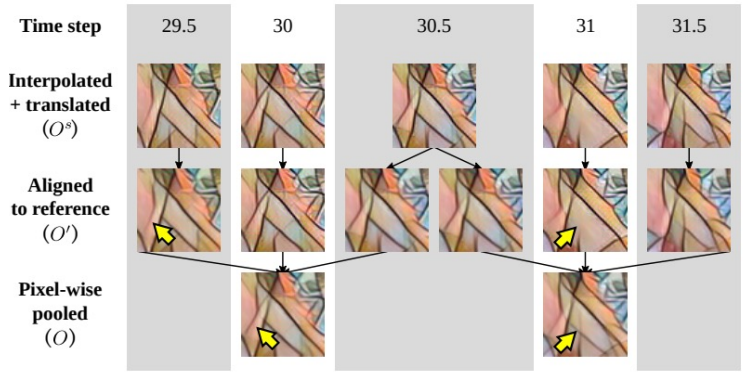
Temporal Aggregation



Temporal Aggregation (Section 2.3)

γ : # of frames in the sliding window
d : temporal dilation factor
$c' = 2\gamma + 1$



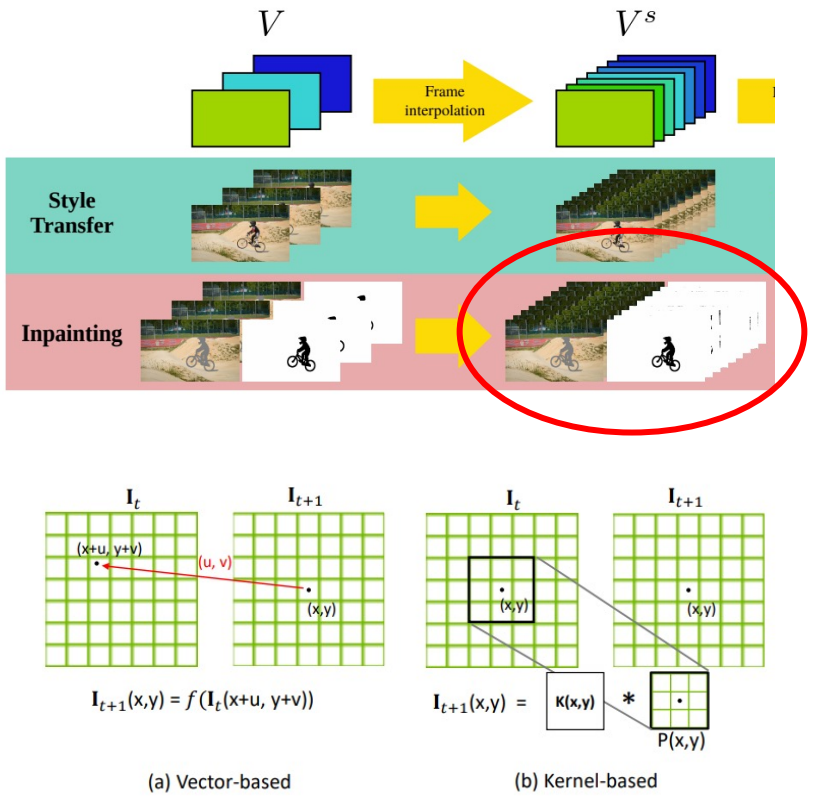


$$o'_{j,k} = \begin{cases} o^s_j & k = 0 \\ \text{warp}(o^s_{j+dk}, F^a_{j \to j+dk}) & k \neq 0 \end{cases}, k \in \{-\gamma, \ldots, \gamma\}, \quad (4)$$

$$o_j = \text{pool}(o'_{j,k} \mid k \in \{-\gamma, \ldots, \gamma\}). \quad (5)$$

- Among similar interpolated-frames, flickering artifacts
- Select pixel by a majority vote
  → Automatically incorporate stable components

HyperCon for Masked Videos



- Motion of the interpolated mask video must match motion of RGB interpolated video
- Vector-based interpolation
  - Often appear degraded by noise
  - Not increasing parameters
- Kernel-based interpolation
  - Good results for small displacement
  - Require large kernels to capture large motions
- They choose **vector-based sampling**
  - Same warping grid can be applied to both RGB and mask
- Thresholding turns partially-masked pixels into fully-masked pixels

$$\dot{m}_j^s = (1 - w_{a+b'}) \odot \text{warp}(m_a, F^s_{a+b' \to a}) + w_{a+b'} \odot \text{warp}(m_{a+1}, F^s_{a+b' \to a+1}),$$
$$m_j^s = \text{thresh}(\dot{m}_j^s, 1).$$

Evaluation Metrics

$$PSNR = 10 \log_{10}\left(\frac{R^2}{MSE}\right)$$

$$SSIM(x, y) = [l(x, y)]^{\alpha} \cdot [c(x, y)]^{\beta} \cdot [s(x, y)]^{\gamma}$$

$$e_{\text{warp}}(o_a, o_{a+1}) = \frac{1}{\sum_p M_a^f(p)} \sum_p M_a^f(p) \|D_a(p)\|_2^2 .$$

$$D_a = o_a - \text{warp}(o_{a+1}, F_{a \to a+1})$$

- Evaluate temporal consistency
  - Patch-based consistency measure PSNR, SSIM
  - Warping error
- Evaluate Quality of style transfer
  - FID

- Evaluate reconstruction quality of inpainting
  - LPIPS Distance
- Evaluate spatio-temporal similarity of video inpainting
  - VLPIPS
- Evaluate realism of video inpainting
  - VFID

Hyperparameter Analysis

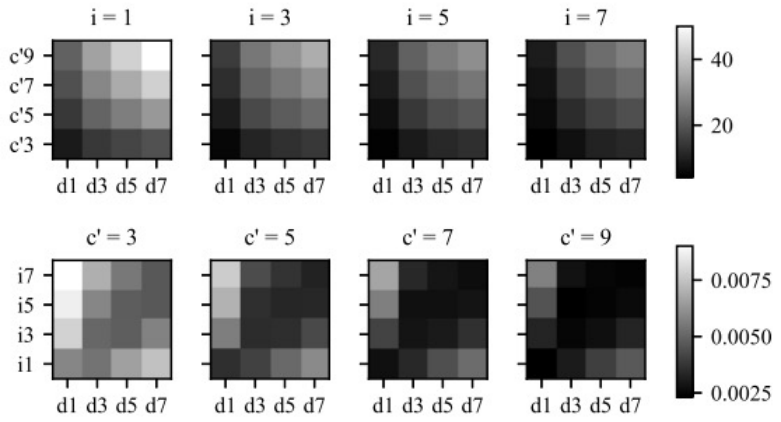
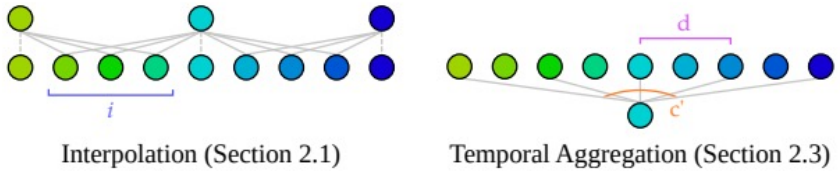


Figure 5: Hyperparameter grid search for *rain-princess* on the DAVIS train/val set. Row 1: FID (style adherence). Row 2: E_{warp} (temporal consistency). Lower is better.

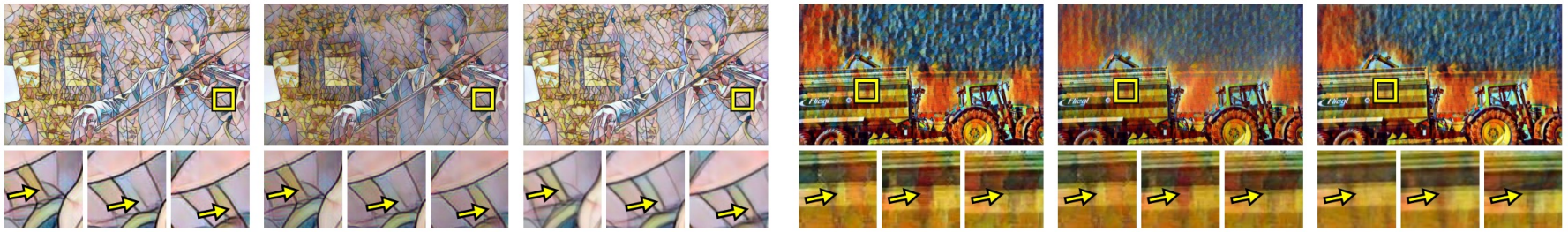


Interpolation (Section 2.1) Temporal Aggregation (Section 2.3)

- Style adherence and temporal consistency are competing objectives
- Style adherence is maximized when aggregates over few frames
- Temporal consistency is maximized when many frames are aggregated
- Here, they consider FID first, then choose the one that min. E_{warp}.

Results

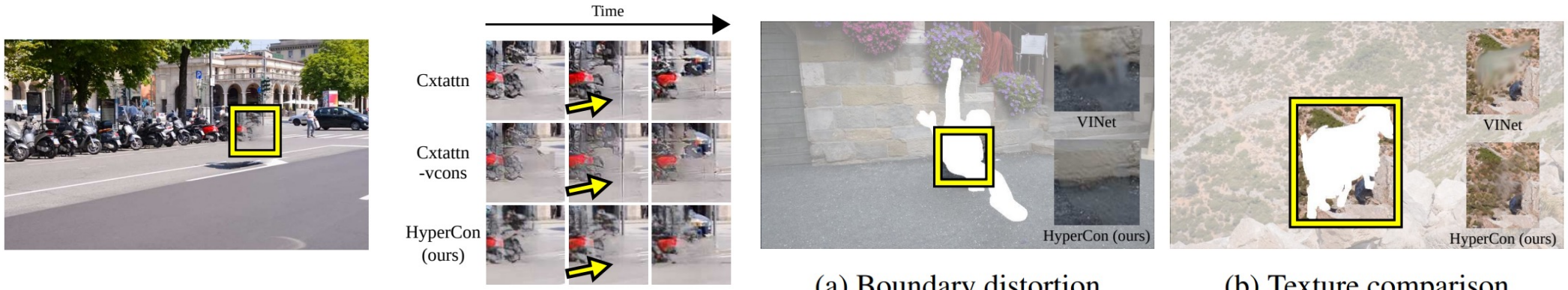
		mosaic				rain-princess			
Dataset	Method	FID↓	E_{warp}↓	C_{PSNR}↑	C_{SSIM}↑	FID↓	E_{warp}↓	C_{PSNR}↑	C_{SSIM}↑
DAVIS 2017	FST [17]	-	0.024829 ± 0.001174	16.72 ± 1.43	0.5166 ± 0.0152	-	0.013934 ± 0.000903	19.75 ± 1.39	0.6030 ± 0.0156
	FST-vcons [19]	24.50	0.010194 ± 0.000500	20.51 ± 1.37	0.5830 ± 0.0140	10.87	0.007071 ± 0.000477	22.73 ± 1.35	0.6717 ± 0.0136
	HyperCon (ours)	**18.04**	**0.008810 ± 0.000493**	**21.04 ± 1.37**	**0.6662 ± 0.0157**	**10.53**	**0.006110 ± 0.000487**	**23.38 ± 1.35**	**0.7480 ± 0.0143**
DAVIS 2019	FST [17]	-	0.029379 ± 0.001684	14.88 ± 0.29	0.4737 ± 0.0184	-	0.017614 ± 0.001309	17.61 ± 0.38	0.5550 ± 0.0189
	FST-vcons [19]	24.89	0.011999 ± 0.000672	18.79 ± 0.29	0.5451 ± 0.0169	14.77	0.008938 ± 0.000644	20.86 ± 0.38	0.6365 ± 0.0159
	HyperCon (ours)	**23.11**	**0.010677 ± 0.000710**	**19.20 ± 0.32**	**0.6105 ± 0.0192**	**13.59**	**0.008117 ± 0.000730**	**21.13 ± 0.43**	**0.6960 ± 0.0175**
ActivityNet	FST [17]	-	0.017895 ± 0.000847	19.29 ± 0.81	0.6478 ± 0.0134	-	0.008428 ± 0.000516	23.45 ± 0.81	0.7469 ± 0.0116
	FST-vcons [19]	26.37	0.006727 ± 0.000311	22.58 ± 0.38	0.7075 ± 0.0115	9.07	0.003923 ± 0.000240	25.97 ± 0.42	0.8008 ± 0.0093
	HyperCon (ours)	**10.09**	**0.005946 ± 0.000298**	**23.61 ± 0.77**	**0.7848 ± 0.0102**	**5.50**	**0.003379 ± 0.000233**	**26.95 ± 0.76**	**0.8544 ± 0.0083**



(a) FST [17] (b) FST-vcons [19] (c) HyperCon (ours) (d) FST [17] (e) FST-vcons [19] (f) HyperCon (ours)

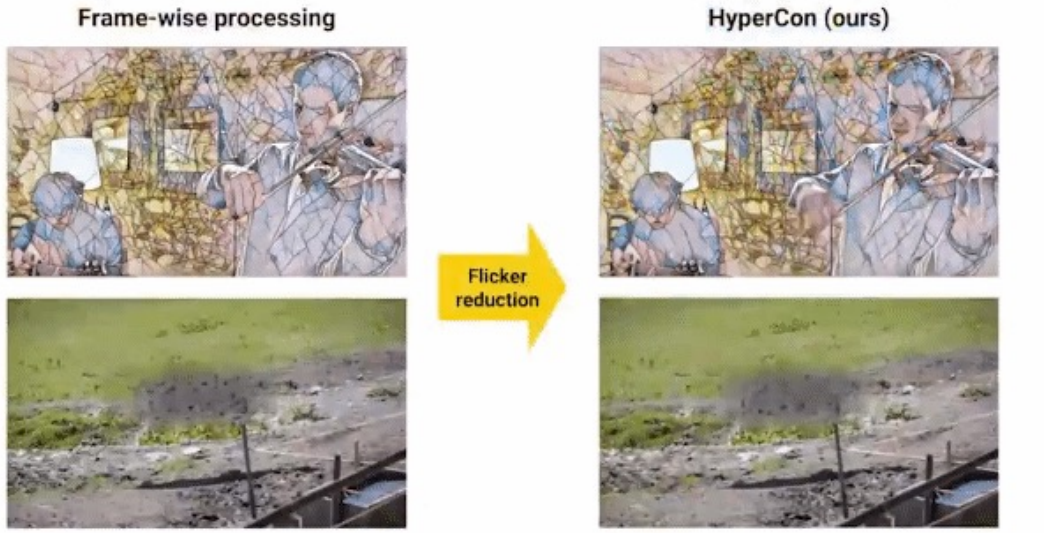
Results

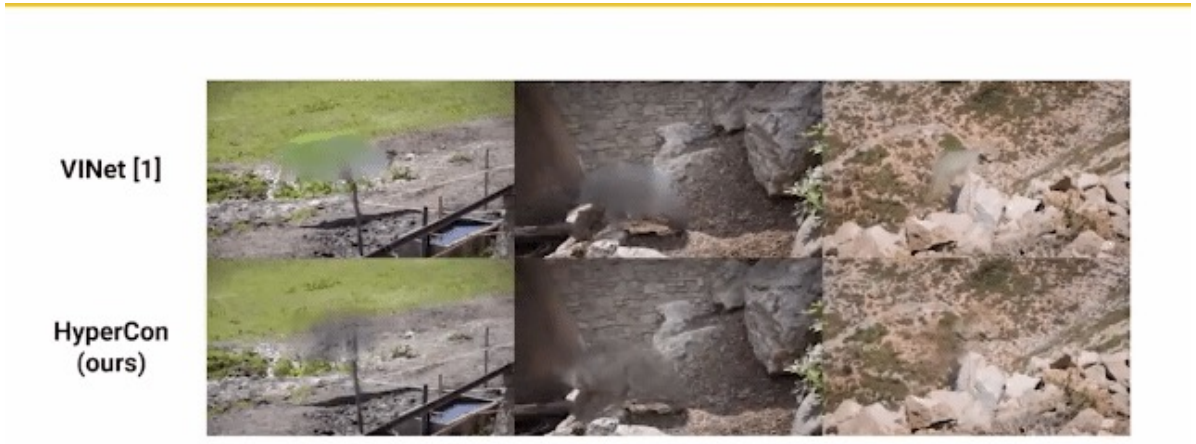
Method	DAVIS 2017					DAVIS 2019					ActivityNet				
	$D_{\text{LPIPS}}\downarrow$	$D_{\text{VLPIPS}}\downarrow$	FID$\downarrow$	VFID$\downarrow$	$E_{\text{warp}}\downarrow$	$D_{\text{LPIPS}}\downarrow$	$D_{\text{VLPIPS}}\downarrow$	FID$\downarrow$	VFID$\downarrow$	$E_{\text{warp}}\downarrow$	$D_{\text{LPIPS}}\downarrow$	$D_{\text{VLPIPS}}\downarrow$	FID$\downarrow$	VFID$\downarrow$	$E_{\text{warp}}\downarrow$
Cxtattn [37]	0.0457	0.5838	20.94	1.435	0.002186	0.0442	0.5575	**15.55**	1.361	0.002539	**0.0432**	0.5981	**21.5173**	1.4417	0.000894
Cxtattn-vcons [19]	0.0480	0.6076	23.23	1.502	0.001780	0.0478	0.5964	18.52	1.490	0.002166	0.0448	0.6067	23.1642	1.4383	0.000689
VINet [18]	0.0616	0.6062	29.24	1.465	0.001882	0.0539	0.5455	18.22	**1.195**	0.002292	0.0608	0.6139	29.3806	1.3783	**0.000678**
HyperCon-mean (ours)	**0.0450**	**0.5272**	**18.49**	**1.073**	**0.001540**	**0.0437**	**0.5179**	16.07	1.274	**0.001847**	0.0454	**0.5728**	22.5111	**1.2251**	**0.000640**
HyperCon-median (ours)	**0.0424**	**0.5217**	**17.75**	**1.074**	**0.001614**	**0.0419**	**0.5089**	**15.24**	**1.254**	**0.001950**	**0.0441**	**0.5812**	**22.2705**	**1.2601**	0.000683



(a) Boundary distortion (b) Texture comparison

Qualitative results





Conclusion

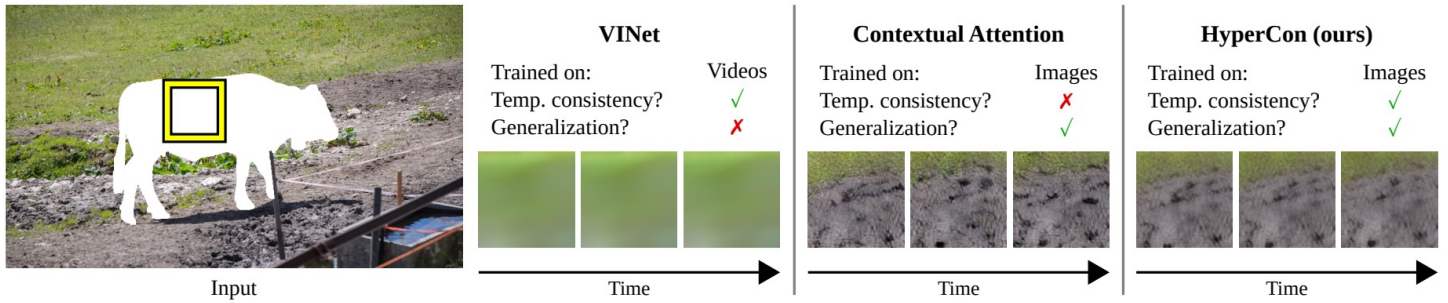


Figure 1: Video-to-video translation models designed and trained from scratch, *e.g.*, VINet [18] are temporally consistent, but exhibit poor generalization performance due to the limited size of high-fidelity video datasets (note the lack of defined texture). Image-to-image models, *e.g.*, Contextual Attention [37], generalize well thanks to large image datasets, but lack temporal consistency (note the changing texture). HyperCon leverages the generalization performance conferred by image datasets while enforcing the temporal consistency properties of video-to-video models.

Limitations

- Not enforce long-range temporal dependencies
- Error propagation issue

Future works

- Integrate three steps into one trainable parameter-efficient model
  → share information between steps

…

Thank you :–)