Video Inpainting

- To fill in the damaged or missing portions of a video

- The main research purpose :
  - Object Removal / Video Restoration [1-3,5,8,9-12]
    - With known mask
  - Blind Video De-captioning [4]
    - With unknown mask like caption
  - Free-form Video Inpainting [6,9]
    - The mask is of arbitrary shape

- The difference with the image inpainting
  - A temporal consistency should be considered
    - **Spatial + Temporal** in video inpainting

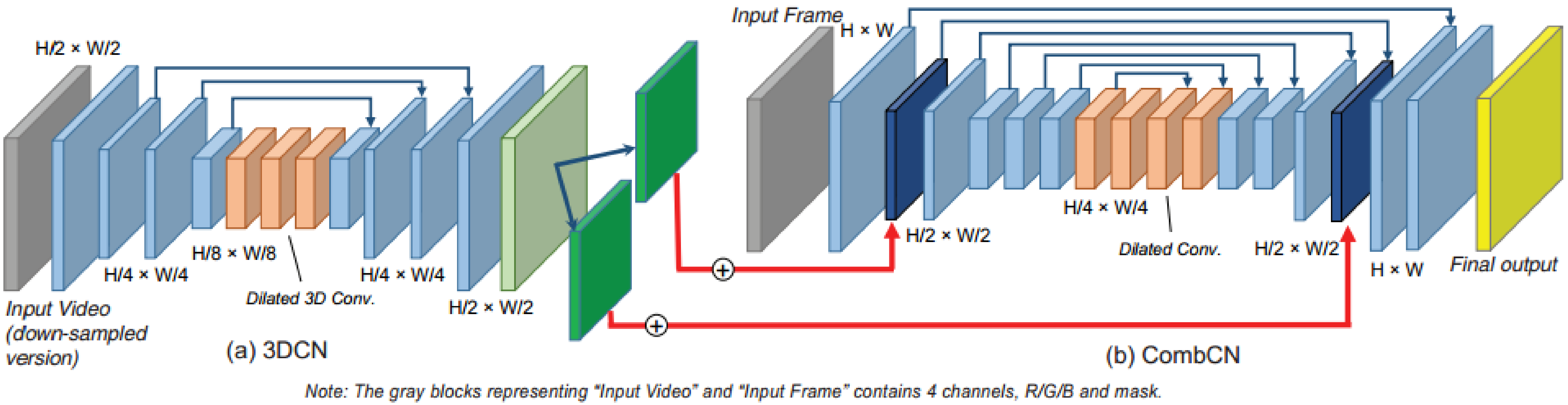
Object Removal

Blind Video De-captioning

Free-form Video Inpainting

Three typical architectures in references

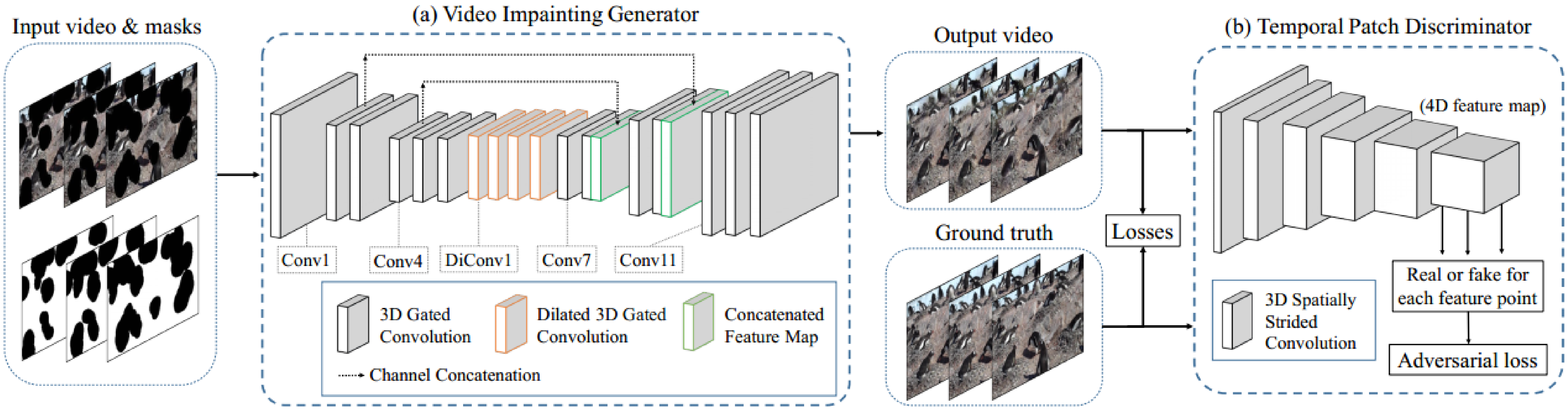
1. Baseline [1], also similar to [4]



- 3D-CNN to learn the temporal information
- The outputs of 3D-CNN are added to CombCN (3D-2D combined completion network)

Three typical architectures in references

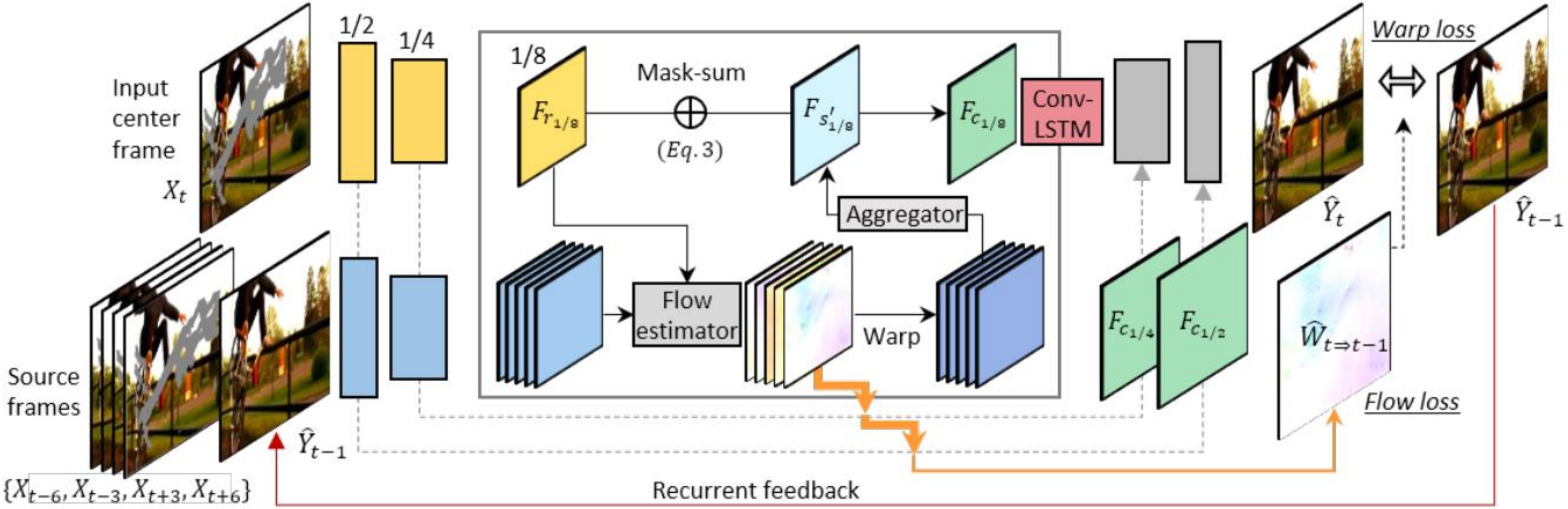
2. Generator + discriminator [9] also similar to [6]



- U-net-like generator network
- Discriminator to enhance the temporal consistency and video quality

Three typical architectures in references

3. Optical Flow-based [3], also similar to [2, 5, 8, 12]



- Learn the flow features in each frame
- Align the reference flow features to the source feature domain(current frame) to produce aligned feature map
- Recurrence and Memory using ConvLSTM

Three typical architectures in references

4. Others [6, 7, 10, 11]

- Other way to find feature map between current frame and reference frames
  - Optical flow based alignment is not suitable for slow-moving videos [11]
    - LGTSM (Learnable Gated Temporal Shift Module) [6]
    - Homography estimator [7]
    - Asymmetric Attention Block [10]
    - Shared alignment encoders + alignment regressors [11]

Summary of networks in references

Index	Model	Network
[1]		3DCN + CombCN
[2]	DFC-Net	Three DFC-S and each inputs gradually enlarged as 1/2, 2/3, 1
[3]	VINET	Encoder-Decoder Network + Flow composition + ConvLSTM
[4]	BVDNet	Parallel 3DCN and 2DCN encoder + 2DCN decoder
[5]	VORNet	Warping Network + Inpainting Network(Image inpainting)+refinement network
[6] [6]	LGTSM	U-net like generator and a TSMGAN discriminator
[7]		Homography-guided warping + Align-and-Attend Video Inpainter +FlowNet
[8]		ImageCN + FlowNet + Flow blending network + ConvLTSM
[9]		3D Gated CNN (U-net) + Temporal PatchGAN (TPatchGAN) discriminator
[10]	OPN	Encoder to parallel produce key and value features + Asymmetric Attention Block
[11]		Optical flow based Alignment network + Copy-and-Paste network (context matching module)
[12]	DIP-based	DIP-based generative network to generate inpainted video and flow

Summary of Loss and Dataset

Index	Loss	Dataset
[1]	L1 loss in 3DCN and CombCN separately	FaceForensics, 300VM, Caltech
[2]	L1 loss with hard flow example mining	DAVIS, YouTube-VOS
[3]	Reconstruction loss(L1+ssim), temporal loss(flow+warp)	YouTube-VOS, Other mask | DAVIS
[4]	Reconstruction loss(L1+ssim+gradient), temporal loss	ECCV Challenge datasets
[5]	Reconstruction loss, perceptual loss, PatchGAN + TempoGAN loss	SVOR from YouTube-VOS
[6]	L1 loss, perceptual loss and style loss, TSMGAN loss	FaceForensics, FVI
[7]	Align loss in Homography, Reconstruction loss(hole, valid) + temporal loss(flow, warp)+ imGAN and vidGAN loss in inpainting	Places2 image + irregular mask in Homography and Youtube-VOS in inpainting
[8]	Spatial loss, Short-term temporal loss, Long-term temporal loss	FaceForensics, DAVIS+VIDEVO
[9]	L1 loss(w/o+w mask), perceptual and style loss, T-PatchGAN loss	FaceForensics, FVI
[10]	Reconstruction loss(peel, valid), perceptual and style loss, total variation regularization term	YouTube-VOS++
[11]	Align loss, Reconstruction loss(hole(visible,invisible), no-hole), perceptual and style loss, total variation regularization term	Places + Crawled Youtube videod
[12]	Image generation and flow generation loss, consistency and perceptual loss	DAVIS + 13 videos

Summary of Common tools

Tools	Papers	Function
U-Net	[1, 3, 6*, 8, 9]	Skip-connections
FlowNet	[2, 3, 4, 5, 7, 8]	Flow extraction
PWCNet	[3, 12]	Coarse-to-fine structure
ConvLSTM	[3, 5, 8]	Improve the temporal stability recurrently
VGG	[5, 6, 9, 10, 11, 12]	Compute the perceptual distance
PatchGAN	[5, 9*]	To motivate our model to generate realistic images
No skip connections	[6]	There are many masked areas in the down-sampling layers

- [4] uses temporal-pooling skip connections

Conclusion

- All of the papers are based on the encoder-decoder network.

- The main challenge is to integrate the temporal and spatial information well
  - Most of the papers are aligning the reference features to the current frame
    - They use both previous and later frames
  - Few papers use image inpainting to fill up the invisible hole

- Some papers use different GAN discriminator to improve the temporal and spatial consistency

Reference

[1] Wang C, Huang H, Han X, et al. **Video inpainting by jointly learning temporal structure and spatial details**[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2019, 33: 5232-5239.
[2] Xu R, Li X, Zhou B, et al. **Deep Flow-Guided Video Inpainting**[J]. arXiv preprint arXiv:1905.02884, 2019.(CVPR)
[3] Kim D, Woo S, Lee J Y, et al. **Deep video inpainting**[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 5792-5801.
[4] Kim D, Woo S, Lee J Y, et al. **Deep blind video decaptioning by temporal aggregation and recurrence**[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 4263-4272.
[5] Chang Y L, Yu Liu Z, Hsu W. **VORNet: Spatio-temporally Consistent Video Inpainting for Object Removal**[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2019: 0-0.
[6] Chang Y L, Liu Z Y, Lee K Y, et al. **Learnable gated temporal shift module for deep video inpainting**[J]. arXiv preprint arXiv:1907.01131, 2019.(BMVC)
[7] Woo S, Kim D, Park K Y, et al. **Align-and-Attend Network for Globally and Locally Coherent Video Inpainting**[J]. arXiv preprint arXiv:1905.13066, 2019.
[8] Ding Y, Wang C, Huang H, et al. **Frame-Recurrent Video Inpainting by Robust Optical Flow Inference**[J]. arXiv preprint arXiv:1905.02882, 2019.
[9] Chang Y L, Liu Z Y, Lee K Y, et al. **Free-form video inpainting with 3d gated convolution and temporal patchgan**[C]//Proceedings of the IEEE International Conference on Computer Vision. 2019: 9066-9075.
[10] Oh S W, Lee S, Lee J Y, et al. **Onion-Peel Networks for Deep Video Completion**[C]//Proceedings of the IEEE International Conference on Computer Vision. 2019: 4403-4412.
[11] Lee S, Oh S W, Won D Y, et al. **Copy-and-Paste Networks for Deep Video Inpainting**[C]//Proceedings of the IEEE International Conference on Computer Vision. 2019: 4413-4421.
[12] Zhang H, Mai L, Xu N, et al. **An Internal Learning Approach to Video Inpainting**[C]//Proceedings of the IEEE International Conference on Computer Vision. 2019: 2720-2729.