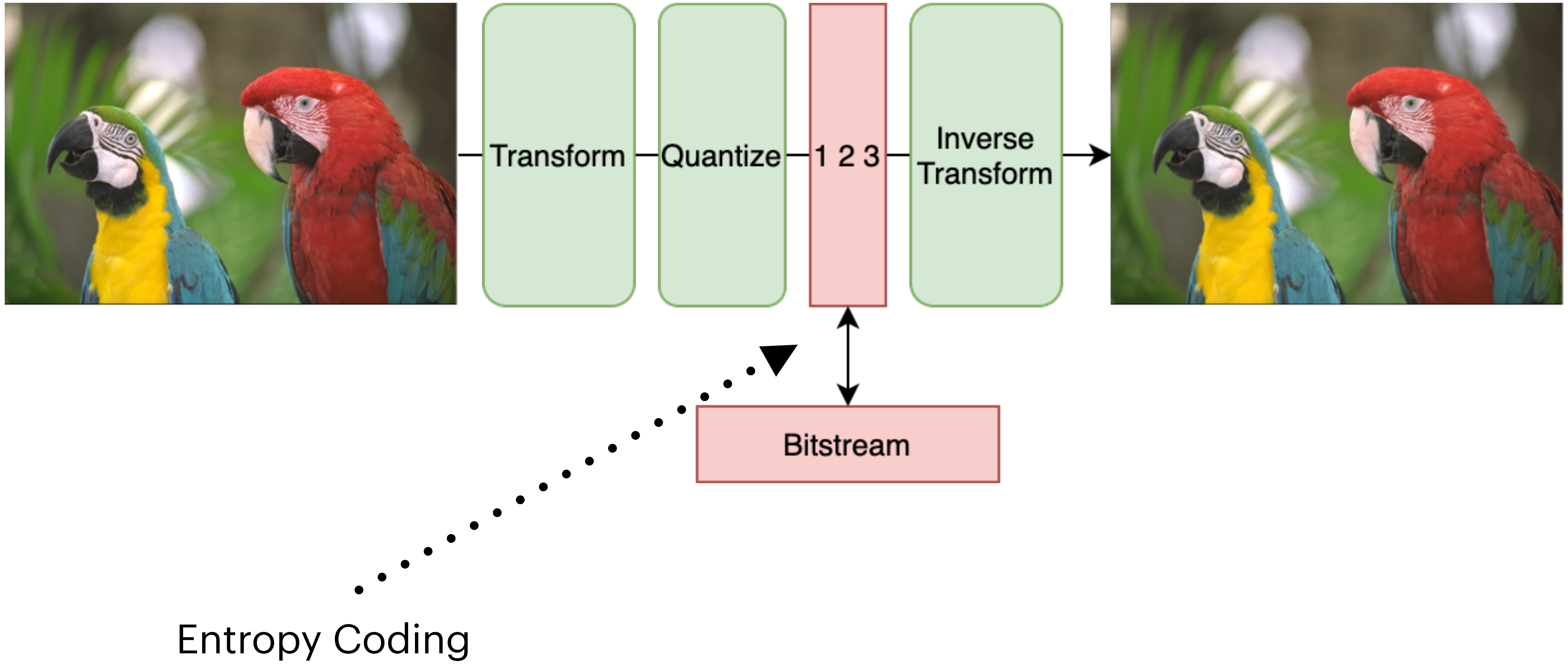
Hierarchical Autoregressive Modeling for Neural Video Compression

Ruihan Yang[1], Yibo Yang[1], Joe Marino[2], Stephan Mandt[1]

1:University of California, Irvine 2:California Institute of Technology

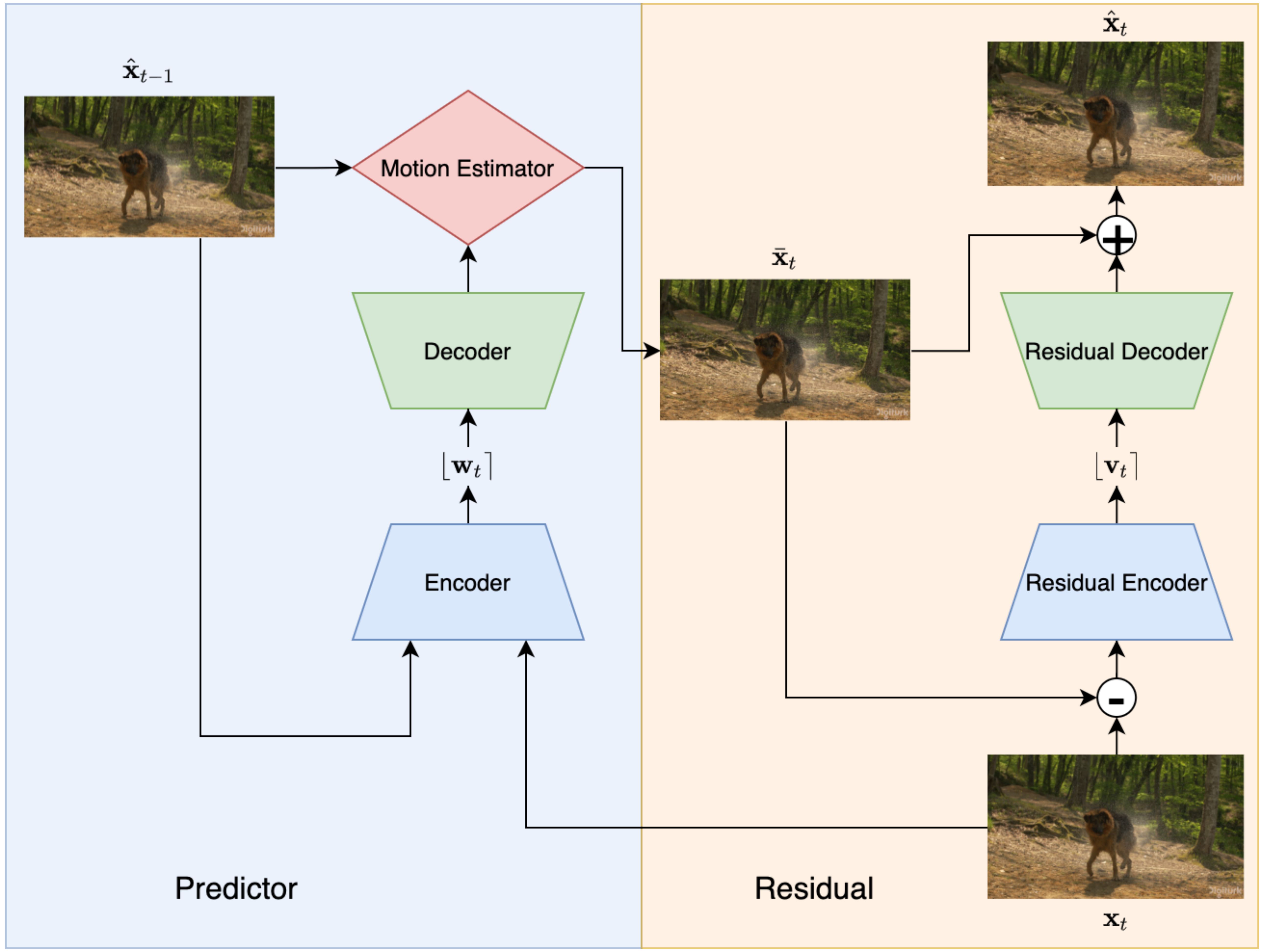
ICLR 2021

Lossy Compression



See “Variational image compression with a scale hyperprior (Balle et al. 2018)” for latent variable model for image compression.

Low-Latency Video Compression



Motivation & Contribution

- Draw the connection to hierarchical sequential latent variable model.
- Propose a general framework to improve the compression performance.
  - Additional module Inspired by normalizing flow.
  - Especially Improved residual compression.
- Our model outperforms the existing state of the art in rate-distortion performance.

Method

Agustsson et al. 2020

- Scale Space Flow (SSF)
- 3D optical flow with trilinear interpolation

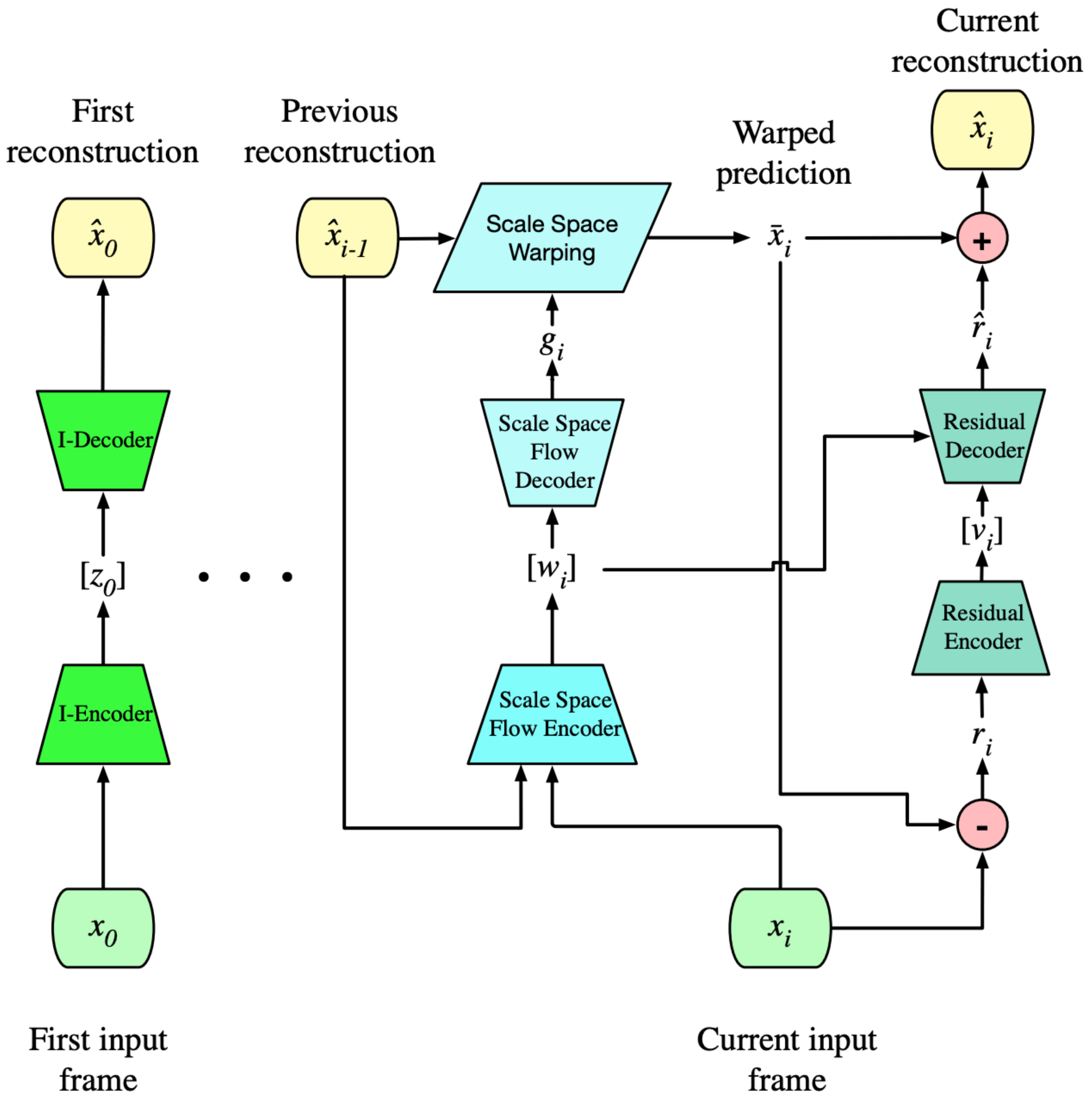


Figure adapted from Agustsson et al. 2020

Method

- Rethink the residual compression as a hierarchical model
  - Invertible Transform
  - Looks like VAE+Normalizing flow?
    - “Improving Sequential Latent Variable Models with Autoregressive Flows” (Marino et al. 2020)

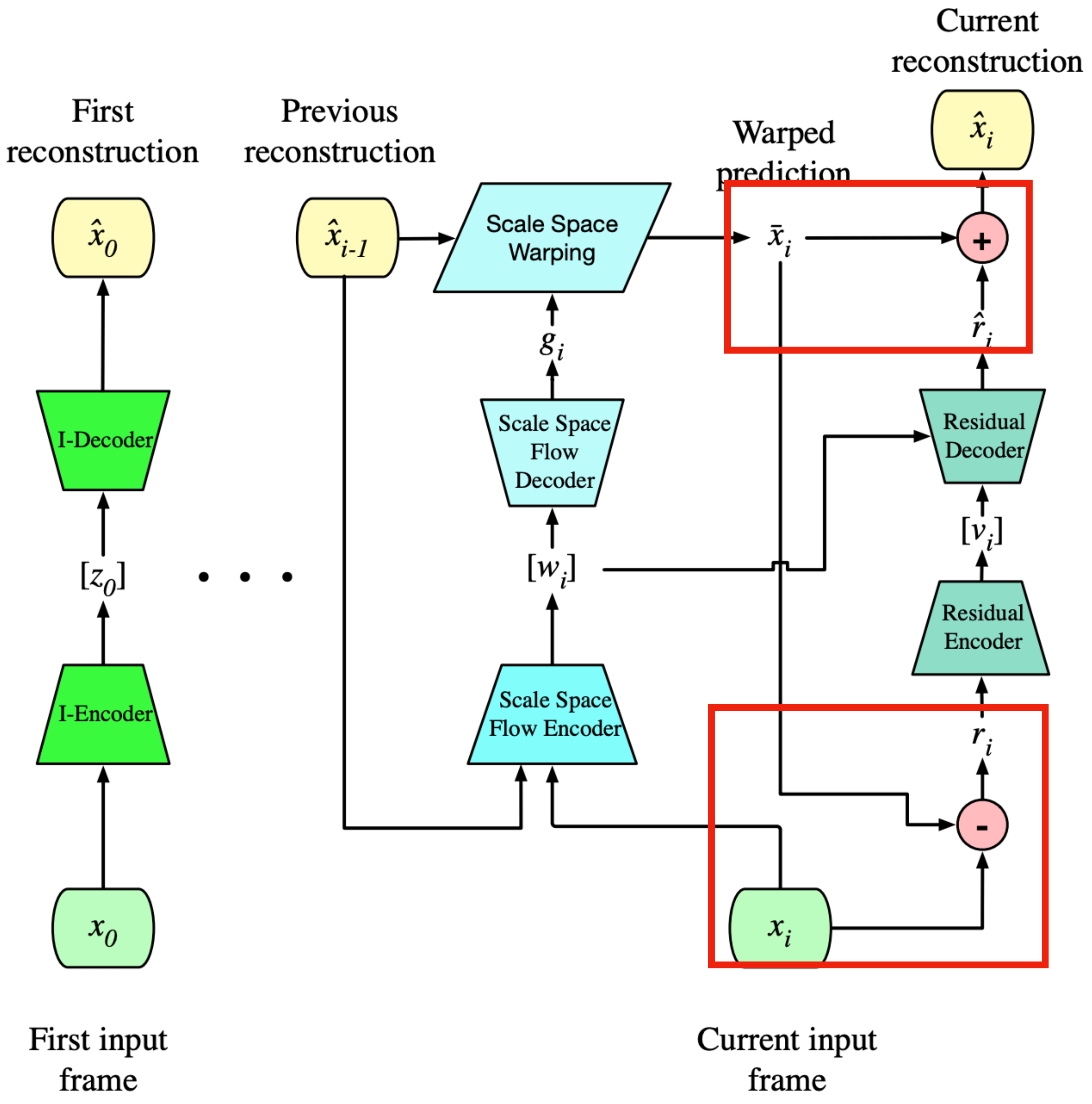


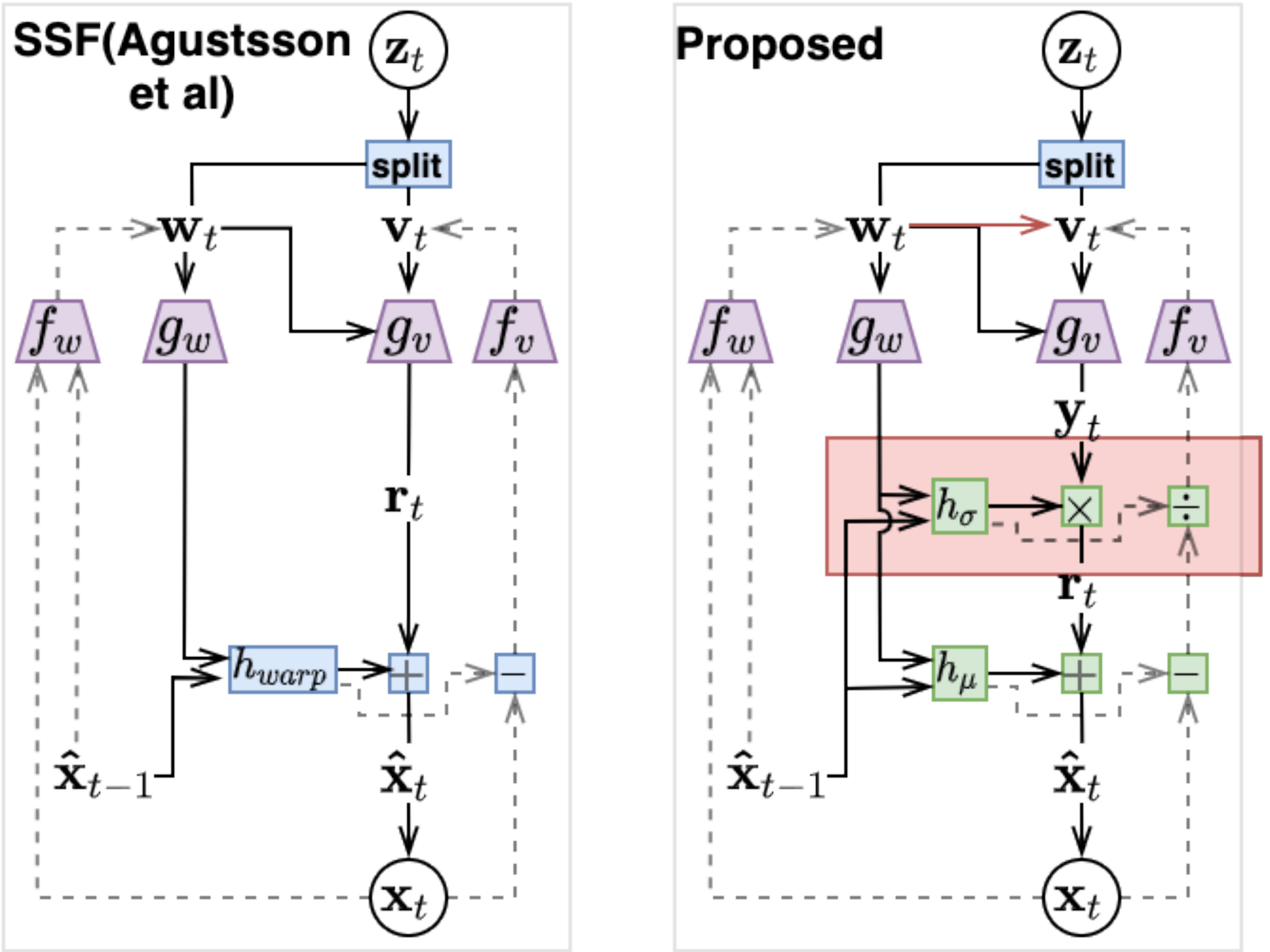
Figure adapted from Agustsson et al. 2020

Method

- Methods to model the invertible transform (coupling layer):
  - Additive coupling: NICE(Dinh et al. 2015): $x = y + m(z)$, $y = x - m(z)$
  - Affine coupling: RealNVP(Dinh et al. 2017): $x = s(z) \odot y + m(z)$, $y = (x - m(z))/s(z)$
  - Autoregressive: $x = x_t$, $z = x_{t-1}$, $y = y_t$

Method

- Stochastic Temporal Autoregressive Transform (STAT)
- Original decoding process:
  - $\hat{x}_t = h_{warp}(\hat{x}_{t-1}, g_w(w_t)) + g_v(v_t, w_t)$
- Proposed decoding process:
  - $\hat{x}_t = h_\mu(\hat{x}_{t-1}, g_w(w_t)) + h_\sigma(\hat{x}_{t-1}, g_w(w_t)) \odot g_v(v_t, w_t)$



Method


Previous reconstruction $\hat{\mathbf{x}}_{t-1}$ by STAT-SSF-SP
Magnitude of the proposed scale parameter $\hat{\boldsymbol{\sigma}}_t = h_\sigma(\hat{\mathbf{x}}_{t-1}, \lfloor \bar{\mathbf{w}}_t \rceil)$
0.4
0.2
0.0
Rate savings in residual encoding $\lfloor \bar{\mathbf{v}}_t \rceil$ relative to STAT-SSF (BPP=0.053)
Rate savings in residual encoding $\lfloor \bar{\mathbf{v}}_t \rceil$ relative to SSF (BPP=0.075)
100
50
0
−50
−100
SSF warping (mean) prediction $\hat{\boldsymbol{\mu}}_t = h_\mu(\hat{\mathbf{x}}_{t-1}, \lfloor \bar{\mathbf{w}}_t \rceil)$
Decoded noise $\hat{\mathbf{y}}_t = g_v(\lfloor \bar{\mathbf{v}}_t \rceil, \lfloor \bar{\mathbf{w}}_t \rceil)$
Decoded residual $\hat{\mathbf{r}}_t = \hat{\mathbf{y}}_t \odot \hat{\boldsymbol{\sigma}}_t$
Current reconstruction $\hat{\mathbf{x}}_t = \hat{\boldsymbol{\mu}}_t + \hat{\mathbf{r}}_t$ by STAT-SSF-SP (BPP=0.046)

Method

- Original:

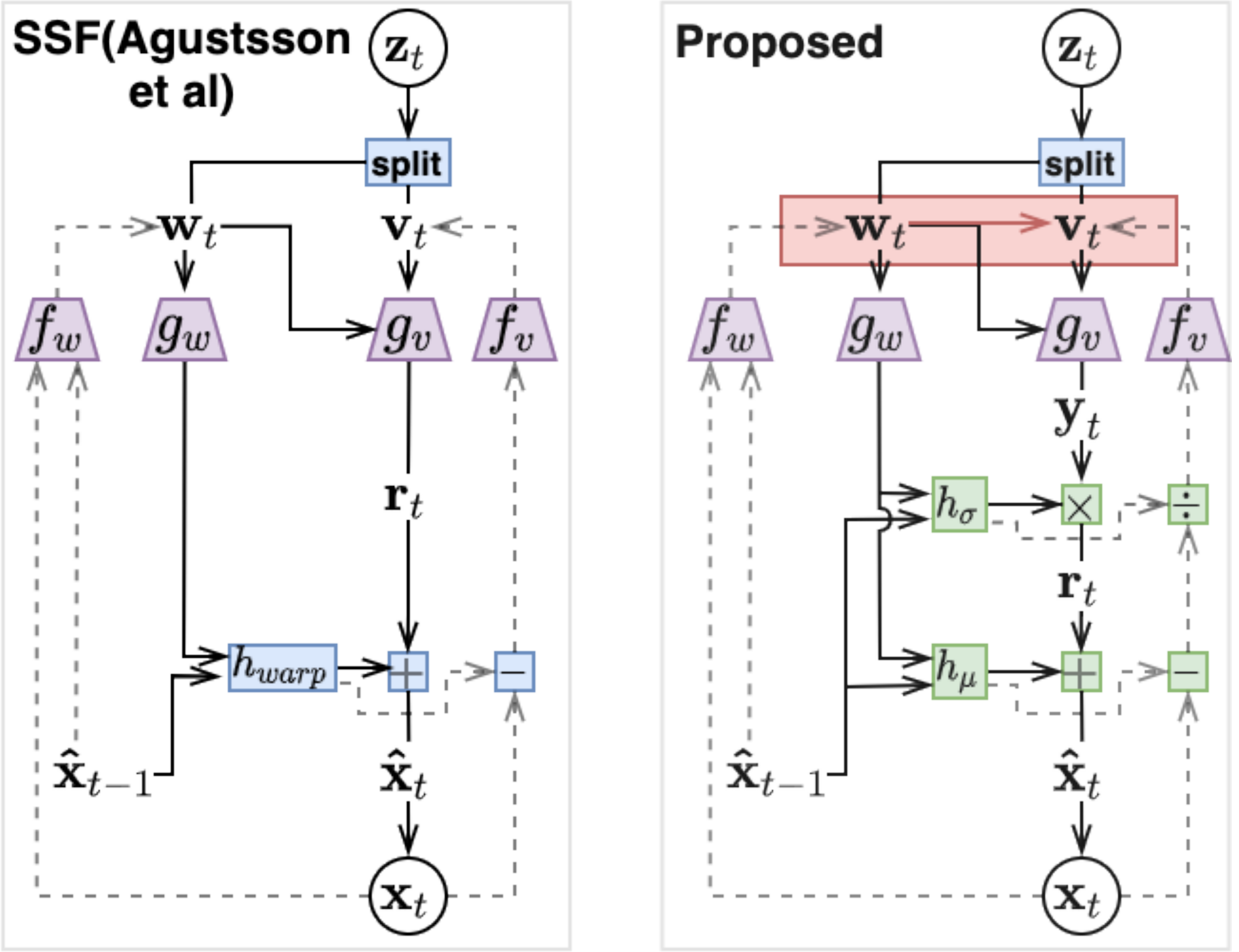
$$p(\mathrm{v}_t, \mathrm{w}_t) = p(\mathrm{w}_t)p(\mathrm{v}_t)$$

- **Structured Prior (Proposed)**:

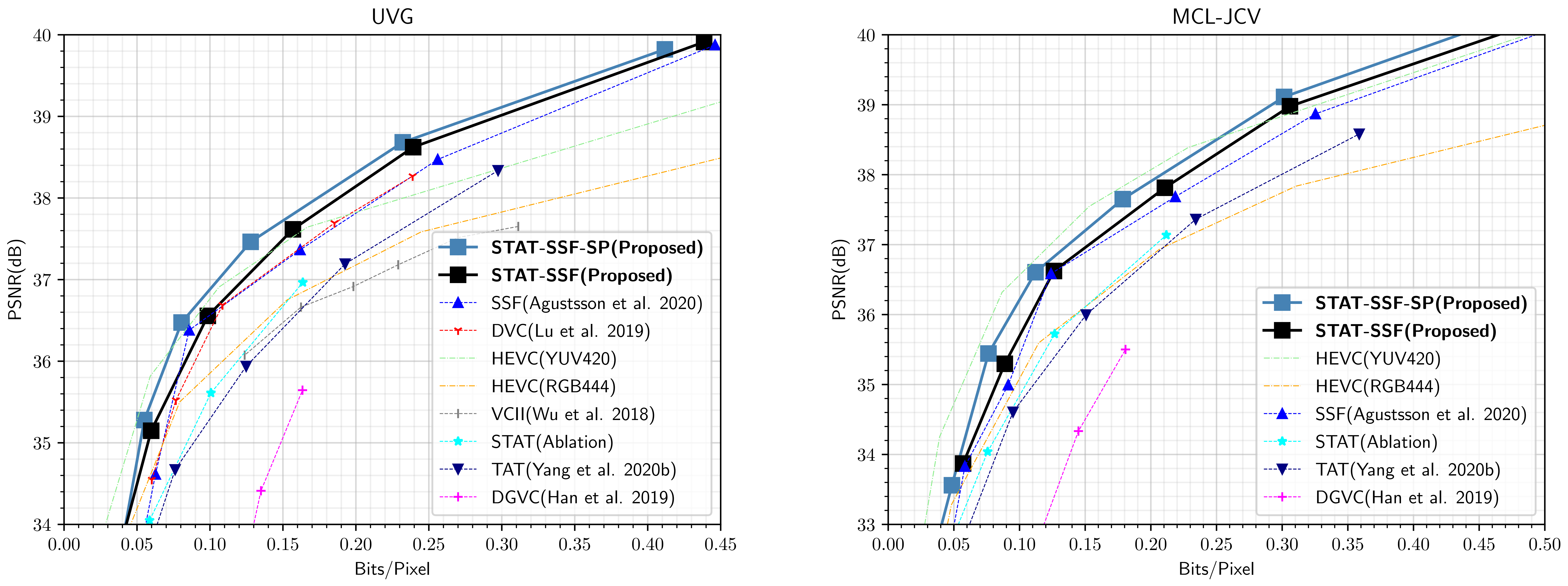
$$p(\mathrm{v}_t, \mathrm{w}_t) = p(\mathrm{w}_t)p(\mathrm{v}_t \,|\, \mathrm{w}_t)$$

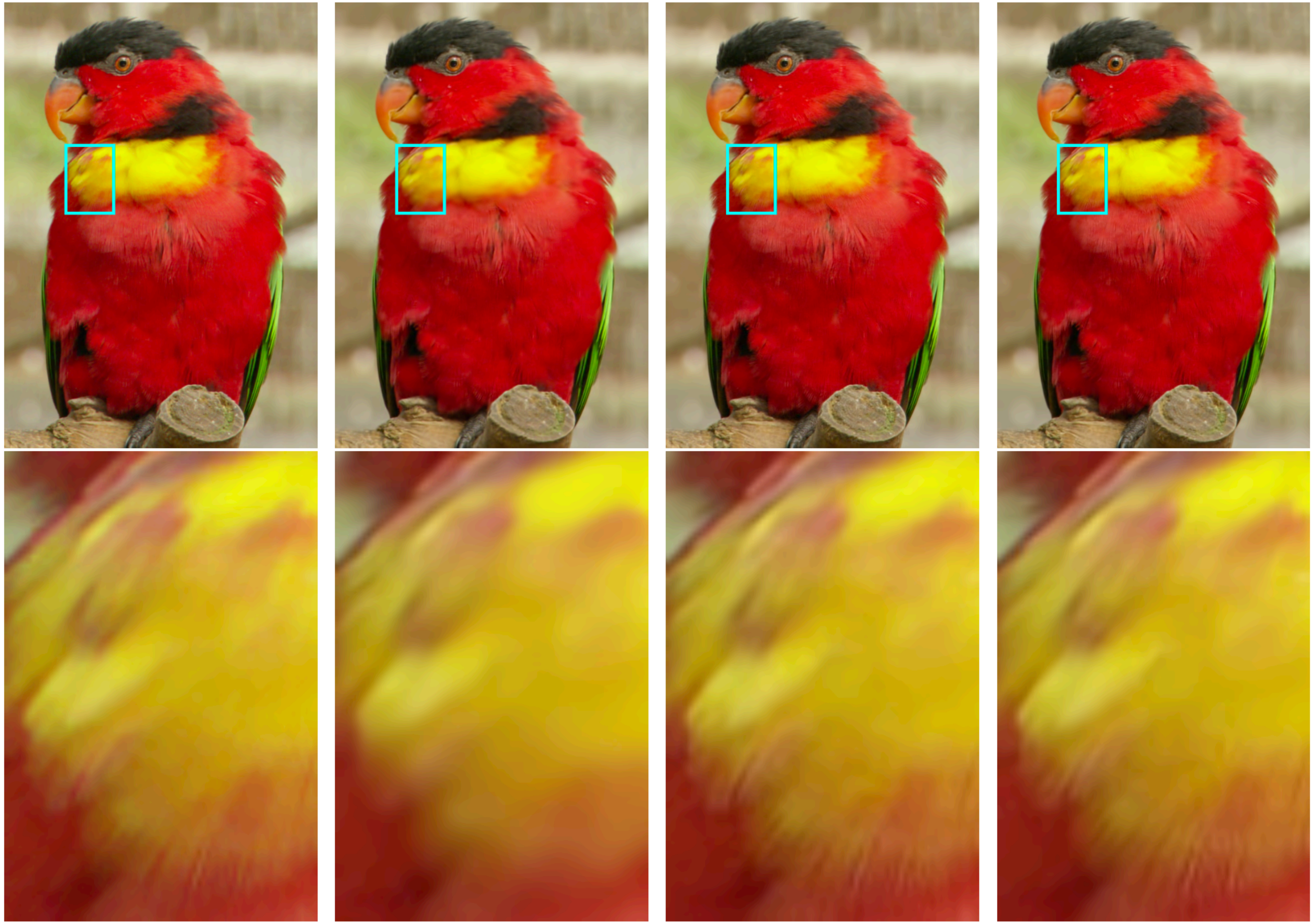
- Temporal Prior (Expensive):

$$p(\mathrm{v}_t, \mathrm{w}_t) = \prod p(\mathrm{w}_t \,|\, \mathrm{w}_{<t})p(\mathrm{v}_t \,|\, \mathrm{v}_{<t})$$



Experiment


UVG
MCL-JCV
PSNR(dB)
Bits/Pixel
STAT-SSF-SP(Proposed)
STAT-SSF(Proposed)
SSF(Agustsson et al. 2020)
DVC(Lu et al. 2019)
HEVC(YUV420)
HEVC(RGB444)
VCII(Wu et al. 2018)
STAT(Ablation)
TAT(Yang et al. 2020b)
DGVC(Han et al. 2019)



(a) HEVC;
BPP $= 0.087$,
PSNR $= 38.099$

(b) SSF;
BPP $= 0.082$,
PSNR $= 37.440$

(c) STAT-SSF (ours);
BPP = 0.0768,
PSNR $= 38.108$

(d) STAT-SSF-SP (ours);
BPP = 0.0551,
PSNR $= 38.097$

Thank you