CLAIMS

1. A computer-implemented method comprising:

obtaining estimated motion information of a target frame in a video and reference feature information for a reference frame and a reference reconstructed frame of the reference frame for the target frame, wherein the reference feature information comprises a plurality of reference feature maps ;

determining, using a context extraction model, context information for the target frame based on the estimated motion information, the reference reconstructed frame, and the reference feature information, wherein the context information comprises weights and the motion offsets for the target frame;

generating, in a conversion between the target frame and a bitstream of the video, a target reconstructed frame of the target frame using a frame coding model and based on at least the context information;

wherein determining the context information comprises:

partitioning the plurality of reference feature maps into a first number (G) of feature map groups;

for each feature map group of the first number (G) of feature map groups,

determining a second number (N) of motion offsets from the reference frame to the target frame based on the estimated motion information, the reference reconstructed frame, and the feature map group, the second number being greater than one, and

performing respective motion alignments on the feature map group based on the second number of motion offsets (N), to obtain the second number of aligned feature map groups, wherein the motion alignments are to align the reference feature information to a space of the target frame; and

determining the context information based on the aligned feature map groups determined for the first number of feature map groups.

2. The method of claim 1, wherein determining the context information based on the aligned feature map groups determined for the first number of feature map groups comprises:

reordering the aligned feature map groups determined for the first number of feature map groups;

merging the reordered aligned feature map groups, to obtain the first number of merged feature map groups, wherein each merged feature map group is merged from the second number of aligned feature maps selected from the reordered aligned feature map groups; and

determining the context information based on the first number of merged feature map groups.

3. The method of claim 1, wherein partitioning the plurality of reference feature maps into the first number of feature map groups comprises:

partitioning the plurality of reference feature maps into the first number of feature map

groups comprises along a channel dimension.

4. The method of claim 1, wherein generating the target reconstructed frame comprises:

determining a quantized code representation for the target frame, the quantized code representation comprising elements organized along a channel dimension and spatial dimensions;

partitioning the elements of the quantized code representation into a plurality of channel element groups along the channel dimension;

partitioning each channel element group of the plurality of channel element groups into a plurality of spatial element groups corresponding to a plurality of spatial positions; and

performing a plurality of entropy coding operations, wherein in each entropy coding operation, a plurality of spatial element groups corresponding to the plurality of spatial positions in the plurality of channel element groups are entropy-coded, and combinations of the plurality of spatial element groups that are entropy-coded in the plurality of entropy coding operations are different.

5. The method of claim 4, wherein performing the plurality of entropy coding operations comprises: for a given entropy coding operation of the plurality of entropy coding operations,

performing the given entropy coding operation based on at least an entropy coding result of a plurality of spatial element groups from a preceding entropy coding operation of the given entropy coding operation.

6. The method of claim 1, wherein the context extraction model and the frame coding model are trained using a plurality of sample frames in a sample video, the plurality of sample frames being assigned with a plurality of weights, and at least one weight being higher than other weights among the plurality of weights, and

wherein a training objective for training the context extraction model and the frame coding model is determined as follows:

determining a plurality of errors between the plurality of sample frames and a plurality of reconstructed sample frames output using the context extraction model and the frame coding model; and

weighting the plurality of errors with the plurality of assigned weights.

7. The method of claim 6, wherein values of the plurality of weights follow a hierarchical structure.

8. The method of claim 1, wherein the conversion comprises:
encoding the target frame as at least a part of the bitstream, or
decoding the target frame from the bitstream.

9. The method of claim 1, further comprising:

generating, in the conversion between the target frame and the bitstream of the video, target feature information of the target frame using the frame coding model and based on at least the context information.

10. An electronic device comprising:

a processor; and

a memory coupled to the processor and comprising instructions stored thereon which, when executed by the processor, cause the device to perform acts comprising:

obtaining estimated motion information of a target frame in a video and reference feature information for a reference frame and a reference reconstructed frame of the reference frame for the target frame, wherein the reference feature information comprises a plurality of reference feature maps;

determining, using a context extraction model, context information for the target frame based on the estimated motion information, the reference reconstructed frame, and the reference feature information, wherein the context information comprises weights and the motion offsets for the target frame;

generating, in a conversion between the target frame and a bitstream of the video, a target reconstructed frame of the target frame using a frame coding model and based on at least the context information ;

wherein determining the context information comprises:

partitioning the plurality of reference feature maps into a first number of feature map groups;

for each feature map group of the first number of feature map groups,

determining a second number of motion offsets from the reference frame to the target frame based on the estimated motion information, the reference reconstructed frame, and the feature map group, the second number being greater than one, and

performing respective motion alignments on the feature map group based on the second number of motion offsets, to obtain the second number of aligned feature map groups, wherein the motion alignments are to align the reference feature information to a space of the target frame; and

determining the context information based on the aligned feature map groups determined for the first number of feature map groups.

11. The electronic device of claim 10, wherein the frame coding model includes an encoder, a concatenator, a decoder, and a frame generator.

12. The electronic device of claim 10, wherein the context extraction model includes a concatenator, a motion alignment unit, an offset predictor, an adder, and a cross-group aggregator.

13. The device of claim 10, wherein determining the context information based on the aligned feature map groups determined for the first number of feature map groups comprises:

reordering the aligned feature map groups determined for the first number of feature map groups;

merging the reordered aligned feature map groups, to obtain the first number of merged feature map groups, wherein each merged feature map group is merged from the second number of aligned feature maps selected from the reordered aligned feature map groups; and

determining the context information based on the first number of merged feature map groups.

14.The device of claim 11, wherein generating the target reconstructed frame comprises:

determining a quantized code representation for the target frame, the quantized code representation comprising elements organized along a channel dimension and spatial dimensions;

partitioning the elements of the quantized code representation into a plurality of channel element groups along the channel dimension;

partitioning each channel element group of the plurality of channel element groups into a plurality of spatial element groups corresponding to a plurality of spatial positions; and

performing a plurality of entropy coding operations, wherein in each entropy coding operation, a plurality of spatial element groups corresponding to the plurality of spatial positions in the plurality of channel element groups are entropy-coded, and combinations of the plurality of spatial element groups that are entropy-coded in the plurality of entropy coding operations are different.

15. A computer program product being tangibly stored in a computer storage medium and comprising computer executable instructions that, when executed by a device, cause the device to perform acts comprising the method of any one of claim 10 to 14.