2024/10/15 multimedia

- shindo

Semantically Structured Image Compression via Irregular Group-Based Decoupling

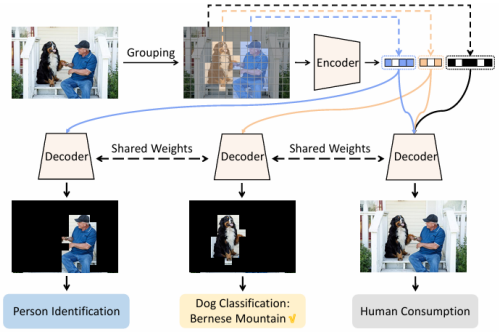
(ICCV 2023)

様々なアプリケーション(画像認識/視聴)に対応するための画像圧縮手法の１検討

ICMに関しては、ROI-based のアプローチをとっており、ROIを作成するための前処理は必要

ROIを用いて画像を分割して圧縮、この際、分割された画像同士は完全に独立なものを目指す(独立でないと、bitに重なりが生じ圧縮の観点からよろしくない)

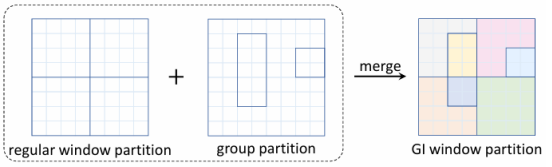
提案手法は、一度の圧縮で、分割された画像群を、互いに独立な形で扱う



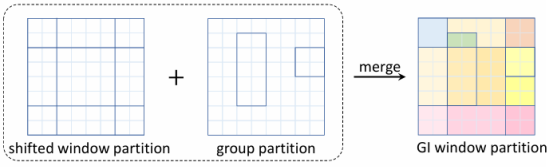
→　patch-wise で上の目的と相性の良い　SwinTを用いる

SwinTでは、不要ピクセルをmaskし、attention計算に使用しないことが容易である

そこで、下の図のように、SwinTを拡張し、GI SwinT (group independent)を実装



(a) GI window partition under regular window partition.



(b) GI window partition under shifted window partition.

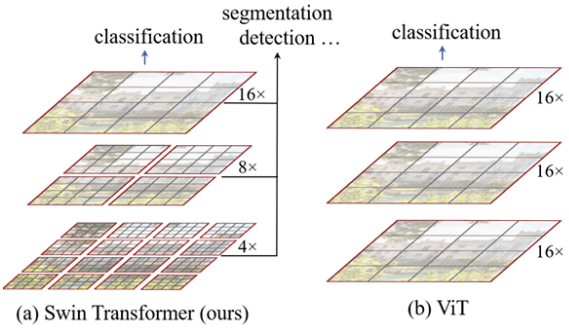
Swin Transformer: Hierarchical Vision Transformer using Shifted Windows

(ICCV 2021)

ViT の拡張

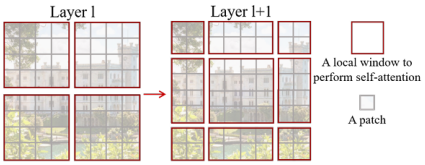
(1) 複数種類のpatch-sizeを導入し、それに対応するwindowを用意　→　階層構造を設計

　様々なスケールで画像から特徴量を抽出できるため、多くのタスクのbackboneとして有効



(2) shift windowを導入し、(一つ前の)window間の関係の検討を可能にした

　windowを半サイズ分ずらすことで、一つ前のwindowでは検討できなかった、patch内の関連性を取得



- taiju

Distribution Extrapolation Diffusion Model for Video Prediction

(CVPR 2024)

将来予測に関する論文

条件づけるフレームから、motion生成。

条件フレームとmotionから

diffusionを用いて将来のmotionを生成し、それをもとに将来画像を生成

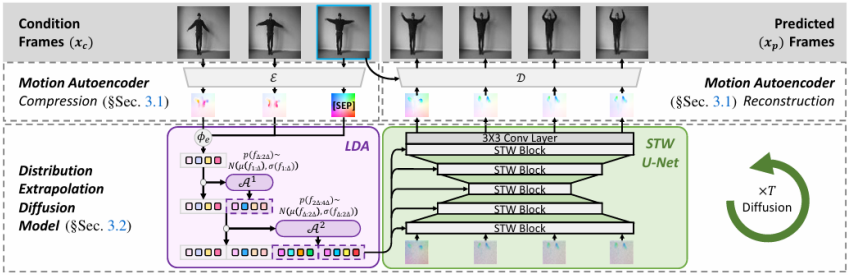


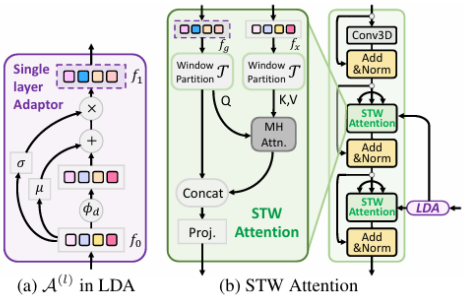
Figure 3. **Pipeline of ExtDM.** ExtDM consists of three main components: Motion autoencoder constructs a bijection transform between the pixel space and motion space via compression and reconstruction. The layered distribution adaptor extrapolates the features of future frames as a shifted distribution derived from condition frames. Furthermore, the built STW U-Net takes the extrapolated feature as guidance and conducts sparse and stride attention among spatiotemporal dimensions for encouraging feature interactions.

motion生成だから、画像生成よりも軽量で速い

LDA：現在のmotion(特徴量)から、将来のmotionの特徴量を作成

現在のmotionの特徴量から、平均分散を計算し、それをシフトすることで、将来のmotion特徴量の推定に使う

STW U-Net： 将来のmotionの特徴量からmotion画像を作成するdiffusion model



(a) $\mathcal{A}^{(l)}$ in LDA (b) STW Attention

Figure 4. **Illustration of the detailed structure** of (a) single layer adaptor and (b) spatiotemporal window block. For more details please refer to Alg. 1.

LDAの出力をattention 構造に入力することで、diffusion(STW U-Net)を条件づける

- takabe

GaussianDreamer: Fast Generation from Text to 3D Gaussians by Bridging 2D and 3D Diffusion Models

(CVPR 2024)

3D 生成へのdiffusionの可能性　テキスト→３D

3d diffusion はデータセット不足から難易度高い 細かいテクスチャ厳しめ

2d diffusionの利用　そもそも2d最適化　３Dとしての一貫性の欠如

提案

3d diffusionで大まかな３D空間を作成、2d diffusion で細かいテクスチャの生成

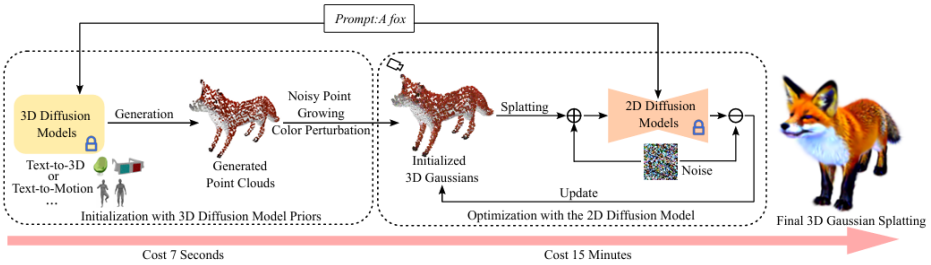


Figure 2. Overall framework of GaussianDreamer. Firstly, we utilize a 3D diffusion model to generate the initialized point clouds. After executing noisy point growing and color perturbation on the point clouds, we use them to initialize the 3D Gaussians. The initialized 3D Gaussians are further optimized using the SDS method [55] with a 2D diffusion model. Finally, we render the image using the 3D Gaussians by employing 3D Gaussian Splatting [26]. We can use one of various 3D diffusion models to generate the initialized point clouds. In this case, we take text-to-3D and text-to-motion diffusion models as examples.

手法

テキスト→３D点群→点群増やす→**初期の３Dガウシアン→画像変換→diffusionでより良い画像に変換→３Dガウシアン→**画像変換

この処理を繰り返す

2Ddiffusion modelを使うことにより、3D diffusion modelよりかは、生成時間が短い

noisy point growing は以下のように点群を補間するものである

この処理が無いと、初期ガウシアン不足により、まともな空間が得られる、生成画像がぼやける

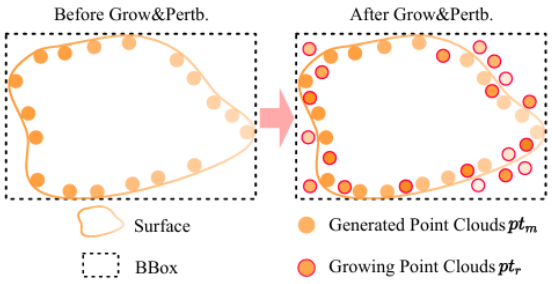


Figure 3. The process of noisy point growing and color perturbation. "Grow&Pertb." denotes noisy point growing and color perturbation.

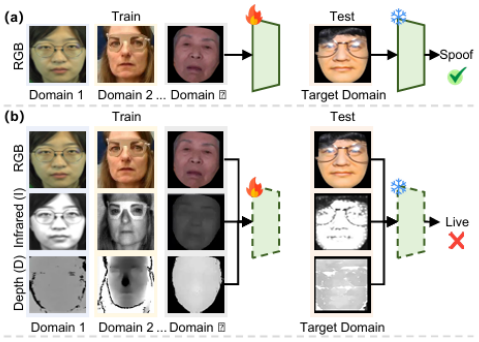
- fujinami

Suppress and Rebalance: Towards Generalized Multi-Modal Face Anti-Spoofing

(CVPR 2024)

Face Anti-Spoofing に関する論文。深度画像や赤外線画像など、入力として複数データを利用したmulti-modalを提案

モーダルを増やすことによるロバスト性の獲得・それらのモーダルのバランスをとる手法の提案



→　ReGrad と U-adapter の提案

ReGrad：各条件下におけるモーダルのバランス調整

U-adapter：信頼性の担保

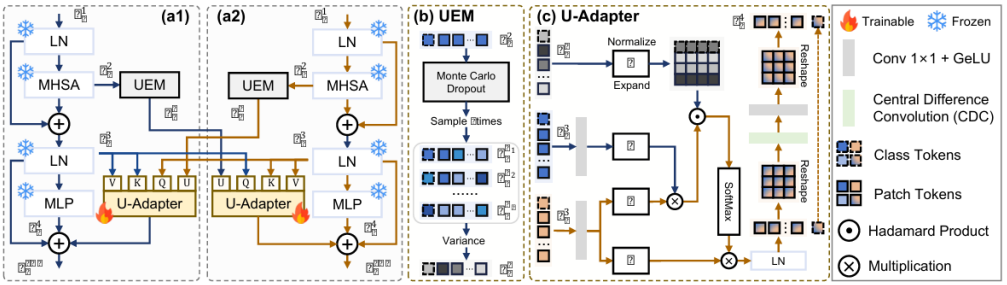


Figure 3. (a1)-(a2) Illustration of fine-tuning ViT with proposed U-Adapters, showcasing the interaction between the RGB (R) and Depth (D) modalities. Note that only parameters of U-Adapters are trainable. (b) Uncertainty Estimation Module (UEM) used for recognizing unreliable tokens. (c) Detailed structure of U-Adapter, which adopts cross-modal fusion and suppresses the interference of unreliable tokens on other modalities. After fusion, discriminative central difference information is integrated for fine-grained spoof representation.

U-adapter は、不確かな特徴量を切り捨てるために、attention moduleに取り込まれる

二つのモーダルをattention moduleでfusionする際に、U(uncertainty)を要素として取り入れる。

UKQV入力のfusionで、お互いに有効な特徴量を取り入れるものが U-adapterである

この仕組みは、各モーダルの出力における信頼性を確保するのに役立つ