# 2024/12/10 multimedia

- shindo

**Bridging the Gap Between Image Coding for Machines and Humans**

(ICIP 2022)

画像認識のための画像圧縮手法を、人間視聴用に拡張するための手法の提案

ICM手法は機械のための画像を復号する一方で、人間には適さない画像を出力する

→　復号画像にブロックノイズのようなノイズが加わる(特に low bitrate)

ICM手法のLICの一部分をファインチューニングすることで、機械・人間ともに使用できる画像復号ができるようになる

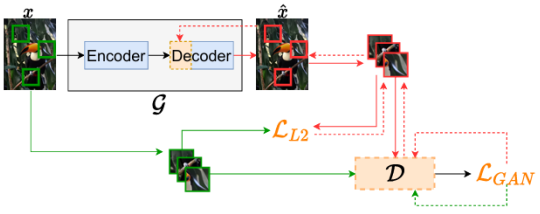


**Fig. 1**. The overview of the finetuning scheme using PatchGAN discriminator. The dashed lines denote gradient backpropagation flows and dotted boxes denote the parameters getting updated by the optimizer. The green lines indicate the data from the input of the ICM codec and the red lines indicate the data from the output of the ICM codec.

LICのデコーダの頭２層をGAN-lossを用いて学習させる

エンコーダのweightはフリーズしているので、レートは変わらない

若干認識精度が落ちる

PSNR、SSIMが上がる

結果は以下の通り

Base codec　　Finetuned codec　　Limitedly finetuned codec (*LI*)

**Fig. 2**. The codec finetuned with PatchGAN *(middle)* effectively removes the checkerboard artifacts commonly found in the decoded images of the NN-based convolutional codec such as the base model *(left)*, while the codec finetuned with limited adversarial impact *(right)* only mildly suppresses the artifacts. More examples available at: `https://flysofast.github.io/human-finetuned-icm/`.

全く計算量/モデルサイズを増やすことないところが、いいところ。

- minghao

**How Do Neural Spoofing Countermeasures Detect Partially Spoofed Audio?**

(arXiv 2024)

フェイク音声の検出タスク　(特に、部分的にフェイクを合成された音声)

部分的にフェイクな音声を検出するためには、元の音声とフェイクの音声の変換点の理解が重要だと仮定できる

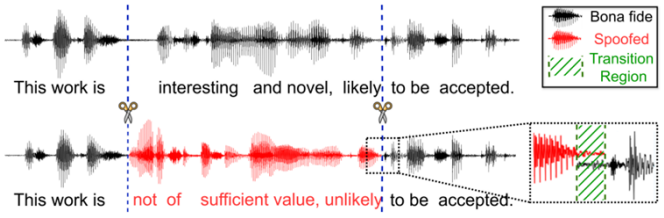


Figure 1: *Illustration of partially spoofed speech changing the meaning of a sentence.*

**Grad-CAM**を用いて、フェイク検出器が、音声のどのパーツに注目しているかを数値化する

→　検出器が正しく判別できているとき：音声のフェイクと元のパーツの**境界に注目している**

→　検出器が正しく判別できないとき：音声のフェイクと元のパーツの**境界に注目していない**

Table 2: *RCQs (%) of SSL-Res1D models when predicting spoof and bona fide classes' scores for partially spoofed samples, across five different segment types. Models are trained on ASVspoof 2019 LA [38] or PartialSpoof [10]. The grey color represents the relative size relationship among the values of the five types of segments within each trial, with deeper shades indicating larger values.*

| Training Set | Grad-CAM Target Class (k) | PartialSpoof Development Set Utt. EER | Bona fide Speech | Spoofed Speech | Transition Region | Bona fide Non-speech | Spoofed Non-speech | PartialSpoof Evaluation Set Utt. EER | Bona fide Speech | Spoofed Speech | Transition Region | Bona fide Non-speech | Spoofed Non-speech |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ASVspoof19 | Spoof | 4.67 | -3.60 | 31.38 | 4.07 | -29.77 | -13.42 | 3.60 | -3.62 | 25.81 | 21.69 | -33.43 | -7.87 |
| | Bona fide | | -0.77 | -35.07 | 39.14 | 30.24 | 12.75 | | -7.11 | -27.24 | 37.03 | 38.72 | 15.26 |
| PartialSpoof | Spoof | 0.35 | -0.06 | 1.87 | 15.03 | -6.95 | -1.81 | 0.73 | -0.62 | -0.90 | 25.36 | -4.72 | 0.99 |
| | Bona fide | | -3.65 | -7.90 | 49.85 | 1.27 | 0.89 | | -8.63 | -4.57 | 58.01 | 10.22 | -1.09 |

検出器が誤っているときの注目領域↓　transition region の値が小さい

Table 3: *RCQs (%) of five different types of speech segments for misclassified partially spoofed samples with SSL-Res1D model.*

| Test Set | Bona fide Speech | Spoofed Speech | Transition Region | Bona fide Non-speech | Spoofed Non-speech |
|---|---|---|---|---|---|
| Dev. | 12.22 | 17.75 | -6.76 | -19.89 | -20.18 |
| Eval. | 26.76 | 27.66 | -10.42 | -49.19 | -28.16 |

これらの結果から、Grad-CAM を用いた注目領域の評価が、タスクの精度評価につかえるかもしれない。

- takabe

**Language-driven Semantic Segmentation**

(ICLR 2022)

open-vocabulary segmentation task

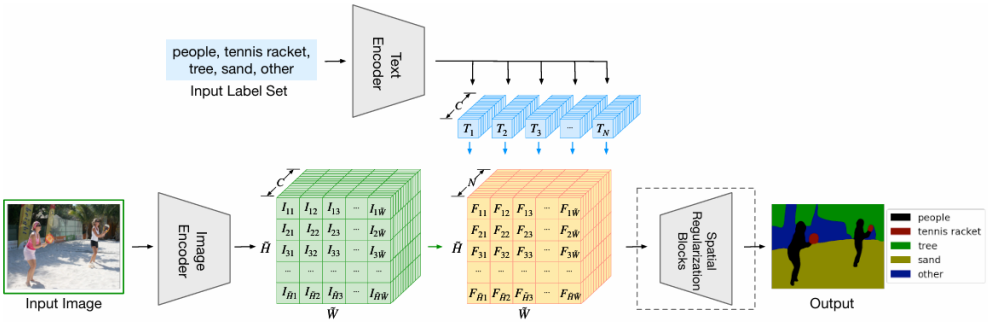


Figure 2: Overview. A text encoder embeds labels into a vector space. An image encoder extracts per-pixel embeddings from the image and correlates the feature of each pixel to all label embeddings. The image encoder is trained to maximize the correlation between the text embedding and the image pixel embedding of the ground-truth class of the pixel. A final spatial regularization block spatially regularizes and cleans up the predictions.

CLIP text-encoderとimage-encoderを用いて、テキスト(検出対象物体)と画像から、特徴量抽出。

CLIPにより、同一特徴空間に落とすことで、open-vocabularyな認識タスクが可能になる

text-enc.はフリーズ、image-enc., spatial reg.を学習させる

spatial-reg. は、空間的に小さくなった特徴量を、元の画像サイズに戻し、セマンティックセグメンテーションを完遂させるために採用されている。

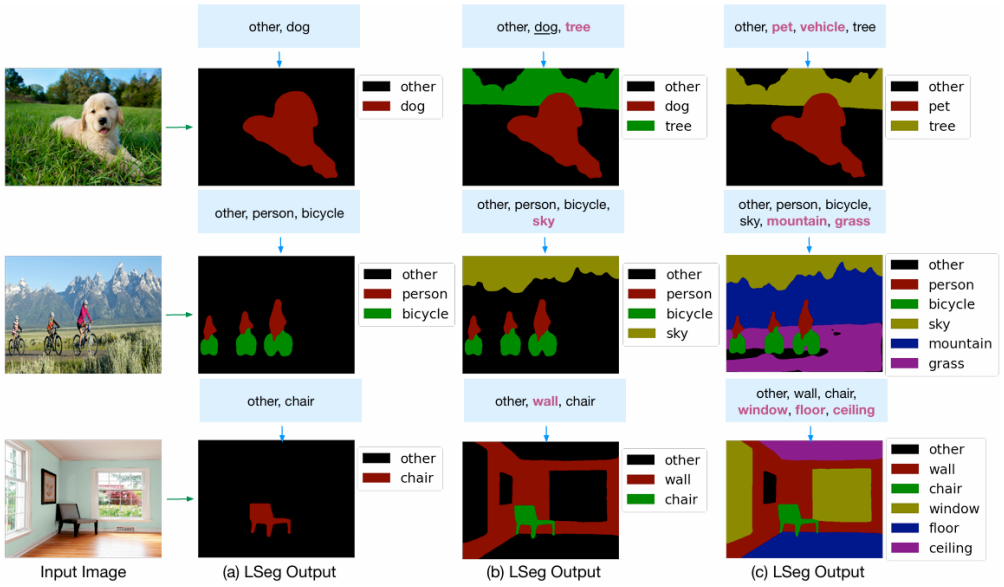


Figure 1: Example results. LSeg is able to handle unseen labels as well as label sets of arbitrary length and order. This enables flexible synthesis of zero-shot semantic segmentation models on the fly. From left to right, labels that are removed between runs are underlined, whereas labels that are added are marked in **bold red**.

結果は上図の通りであり、zero-shotのセグメンテーションが可能となっている。

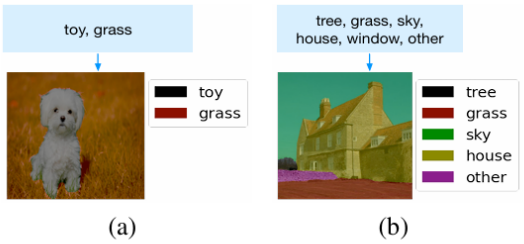


Figure 6: Failure cases.

問題点は、何かしらのクラスを必ず割り当てる必要があるという点。”other”というクラスの配置が常に必要である。ない場合は、存在するクラスから何か一つ割り当て。