Vision AI

2021 arXiv Trends

2021-05

hutom

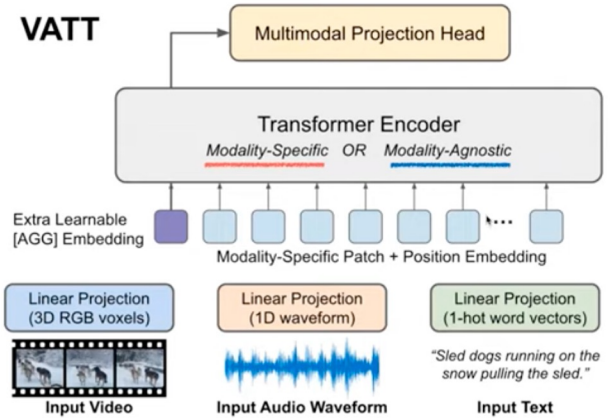
no.	Paper Title	Correspondence	h-index
1	**VATT: Transformers for Multimodal Self-Supervised Learning from Raw Video, Audio and Text**	Hassan Akbari	6
2	**Taming Transformers for High-Resolution Image Synthesis**	Patrick Esser* Robin Rombach*	6 3
3	**DatasetGAN: Efficient Labeled Data Factory with Minimal Human Effort**	Yuxuan Zhang	2
4	**Auto-Tuned Sim-to-Real Transfer**	Yuqing Du* Olivia Watkins*	2 2
5	**Self-supervised Object detection from audio-visual correspondence**	Triantafyllos Afouras* Yuki M. Asano*	10 6

VATT: Transformers for Multimodal Self-Supervised Learning from Raw Video, Audio and Text

https://arxiv.org/pdf/2104.11178.pdf

VATT: Transformers for Multimodal Self-Supervised Learning from Raw Video, Audio and Text

Hassan Akbari[*1,2], Linagzhe Yuan[1], Rui Qian[*1,3], Wei-Hong Chuang[1], Shih-Fu Chang[2], Yin Cui[1], Boqing Gong[1]
[1]Google [2]Columbia University [3]Cornell University
{lzyuan,whchuang,yincui,bgong}@google.com {ha2436,sc250}@columbia.edu {rq49}@cornell.edu



- keyword : Transformer, Multi-modal Representation Learning
- Multi-modal Representation learning from Unlabeled data (self-supervised learning)
- Multi-modality : Video, Audio and Text
- Transformer architecture + Constrastive learning
- DropToken : Simple technique for computational efficiency
- Pretraining : Howto100M dataset (136.6 million clips)
- Downstream tasks
  - action recognition
  - audio event classification
  - zero-shot video retrieval (text to relevant video)
  - image classification (직접적으로 Image를 이용해 학습하지 않았지만 image net 1k dataset 에서대해서 평가진행)

VATT: Transformers for Multimodal Self-Supervised Learning from Raw Video, Audio and Text

Architecture

- **Tokenization layer**
- modality-specific tokenization layer takes as input the raw signals, and returns a sequence of vectors to be fed to the transformers

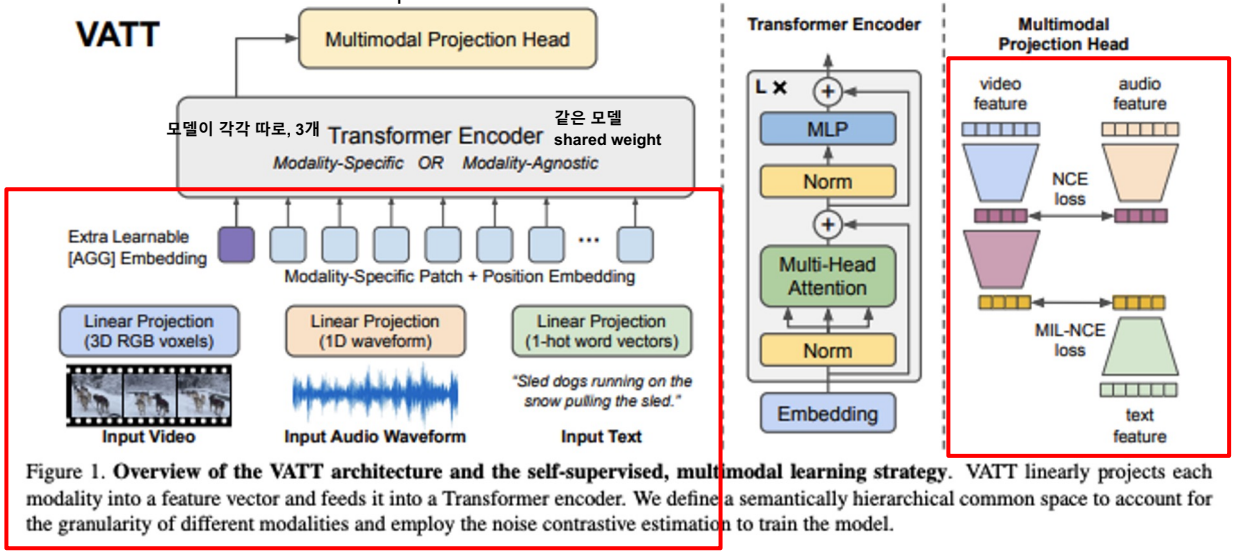
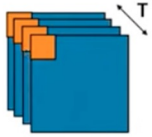


Figure 1. **Overview of the VATT architecture and the self-supervised, multimodal learning strategy.** VATT linearly projects each modality into a feature vector and feeds it into a Transformer encoder. We define a semantically hierarchical common space to account for the granularity of different modalities and employ the noise contrastive estimation to train the model.

Video: we partition an entire video clip of size $T \times H \times W$ to a sequence of $\lceil T/t \rceil \cdot \lceil H/h \rceil \cdot \lceil W/w \rceil$ patches, where each patch contains $t \times h \times w \times 3$ voxels. We apply a linear projection on the entire voxels in each patch to get a d-dimensional vector representation. This projection is performed by a learnable weight $\boldsymbol{W}_{vp} \in \mathbb{R}^{t \cdot h \cdot w \cdot 3 \times d}$.



Positional encoding

$$\boldsymbol{E}_{\text{Temporal}} \in \mathbb{R}^{\lceil T/t \rceil \times d}$$
$$\boldsymbol{E}_{\text{Horizontal}} \in \mathbb{R}^{\lceil H/h \rceil \times d}$$
$$\boldsymbol{E}_{\text{Vertical}} \in \mathbb{R}^{\lceil W/w \rceil \times d}$$
$$\boldsymbol{e}_{i,j,k} = \boldsymbol{e}_{\text{Temporal}_i} + \boldsymbol{e}_{\text{Horizontal}_j} + \boldsymbol{e}_{\text{Vertical}_k}$$

video : 각 패치마다 d-dimension vector 들이 transformer input

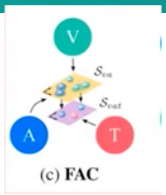
audio : t-segment, frequency domain 으로 변환하지 않고 raw audio file domain에서 사용

Audio: the raw audio waveform is a 1D input with length T, and we partition it to $\lceil T'/t' \rceil$ segments each containing t' waveform amplitudes. Similar to video, we apply a linear projection with a learnable weight $\boldsymbol{W}_{ap} \in \mathbb{R}^{t' \times d}$ to all elements in a patch to get a d-dimensional vector representation. We use $\lceil T'/t' \rceil$ learnable embeddings to encode the position of each waveform segment.

use $\lceil T'/t' \rceil$ learnable embeddings

Common space Projection

- Multimodal, modality가 다르더라도 같은 feature space 보내주어야 함, FAC 사용.
- first, vision & audio
- second, vision+audio & text



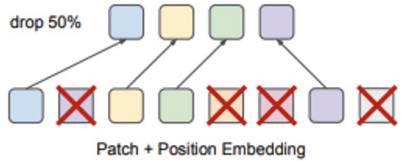
NCE Loss : vision & audio

MIL-NCE Loss : Vision & text

$$\mathcal{L} = \text{NCE}(\boldsymbol{z}_{v,va}, \boldsymbol{z}_{a,va}) + \lambda \text{MIL-NCE}(\boldsymbol{z}_{v,vt}, \{\boldsymbol{z}_{t,vt}\})$$

$$\text{NCE}(\boldsymbol{z}_{v,va}, \boldsymbol{z}_{a,va}) = -\log\left(\frac{\exp(\boldsymbol{z}_{v,va}^{\top}\boldsymbol{z}_{a,va}/\tau)}{\sum_{i=1}^{B}\exp(\boldsymbol{z}^{i\top}_{v,va}\boldsymbol{z}^{i}_{a,va}/\tau)}\right), \quad (4)$$

$$\text{MIL-NCE}(\boldsymbol{z}_{v,vt}, \{\boldsymbol{z}_{t,vt}\}) = -\log\left(\frac{\sum_{\boldsymbol{z}_{t,vt}\in\mathcal{P}(\boldsymbol{z}_{v,vt})}\exp(\boldsymbol{z}_{v,vt}^{\top}\boldsymbol{z}_{t,vt}/\tau)}{\sum_{\boldsymbol{z}_{t,vt}\in\mathcal{P}(\boldsymbol{z}_{v,vt})\cup\mathcal{N}(\boldsymbol{z}_{v,vt})}\exp(\boldsymbol{z}_{v,vt}^{\top}\boldsymbol{z}_{t,vt}/\tau)}\right),$$



DropToken

- Vanila Transformer, input 길이가 길어지면 계산비용 길이의 제곱만큼 늘어남, 길이를 줄이는 것이 좋은 선택.
- DropToken is applied to the video, audio inputs.
- Randomly sample a portion of the tokens and then feed the sampled sequence.
- text 에 비해 audio, video가 시각축에 대해 연속적인 데이터가 등장(중복)하므로 중간 부분 날리더라도 sequence에 대한 정보 충분히 기록(?)

VATT: Transformers for Multimodal Self-Supervised Learning from Raw Video, Audio and Text

Experiments

- **No clear difference between the Modality agnostic feature and Modality-specific**
- modality에 상관없이 하나의 모델로 공유하게 되어도 modality specific과 비슷한 능력이 생기는 듯하다?

Two Multimodal video datasets are used for pretraining of VATT

- HowTo100M : 1.2 million unique videos (multiple clips with audio & narration scripts) // 136 million video-audio-text triplets.
- AudioSet : 10-sec sample clips (video + audio) 2 million video from youtube // text = zero vector sequence input

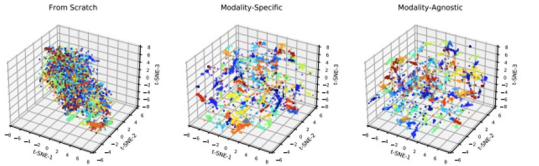


Figure 3. t-SNE visualization of the feature representations extracted by the vision Transformer trained from scratch on Kinetics-400 validation set, the modality-specific VATT's vision Transformer after fine-tuning, and the modality-agnostic Transformer after fine-tuning. For better visualization, we show 100 random classes from Kinetics-400.

DownStream Tasks

- Video action recognition : UFC101, HMDB51, Kinetics-400, 600
- Audio event classification : Audio Set
- Zero-shot video Retrieval : YouCook2, MSR-VTT
- Image Classification : ImageNet 1k

Various sizes of transformer are used for the experiments.

Model	Layers	Hidden Size	MLP Size	Heads	Params
Small	6	512	2048	8	20.9 M
Base	12	768	3072	12	87.9 M
Medium	12	1024	4096	16	155.0 M
Large	24	1024	4096	16	306.1 M

Table 1. Details of the Transformer architectures in VATT.

Action Recognition Results

METHOD	TOP-1	TOP-5	TFLOPS
ARTNet [98]	69.2	88.3	6.0
I3D [16]	71.1	89.3	-
R(2+1)D [30]	72.0	90.0	17.5
MFNet [60]	72.8	90.4	-
Inception-ResNet [2]	73.0	90.9	-
bLVNet [32]	73.5	91.2	0.84
A^2-Net [22]	74.6	91.5	-
TSM [61]	74.7	-	-
S3D-G [102]	74.7	93.4	-
Oct-I3D+NL [21]	75.7	-	0.84
D3D [88]	75.9	-	-
GloRe [23]	76.1	-	-
I3D+NL [98]	77.7	93.3	10.8
ip-CSN-152 [92]	77.8	92.8	-
MoViNet-A5 [51]	78.2	-	0.29
CorrNet [17]	79.2	-	6.7
LGD-3D-101 [75]	79.4	94.4	-
SlowFast [34]	79.8	93.9	7.0
X3D-XXL [33]	80.4	94.6	5.8
TimeSFormer-L [10]	80.7	94.7	7.14
VATT-Base	79.6	94.9	9.09
VATT-Medium	81.1	**95.6**	15.02
VATT-Large	**82.1**	95.5	29.80
VATT-MA-Medium	79.9	94.9	15.02

Table 2. Results for video action recognition on Kinetics-400.

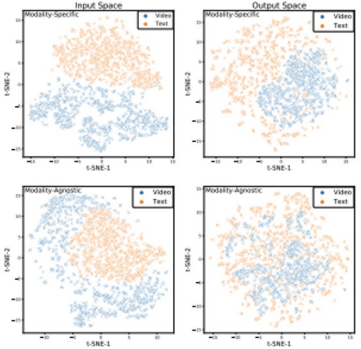


Figure 4. t-SNE visualization of the input space vs. output space for modality-specific and modality-agnostic backbones when different modalities are fed.

Image Classification

Supervised->

METHOD	PRE-TRAINING DATA	TOP-1	TOP-5
iGPT [18]	ImageNet	66.5	-
ViT-Base [29]	JFT	**79.9**	-
VATT-Base	-	64.7	83.9
VATT-Base	HowTo100M	78.7	93.9

Table 6. Finetuning results for ImageNet classification.

Taming Transformers for High-Resolution Image Synthesis

https://arxiv.org/pdf/2012.09841.pdf

Taming Transformers for High-Resolution Image Synthesis

Patrick Esser* Robin Rombach* Björn Ommer
Heidelberg Collaboratory for Image Processing, IWR, Heidelberg University, Germany
*Both authors contributed equally to this work

Figure 1. Our approach enables transformers to synthesize high-resolution images like this one, which contains 1280x460 pixels.

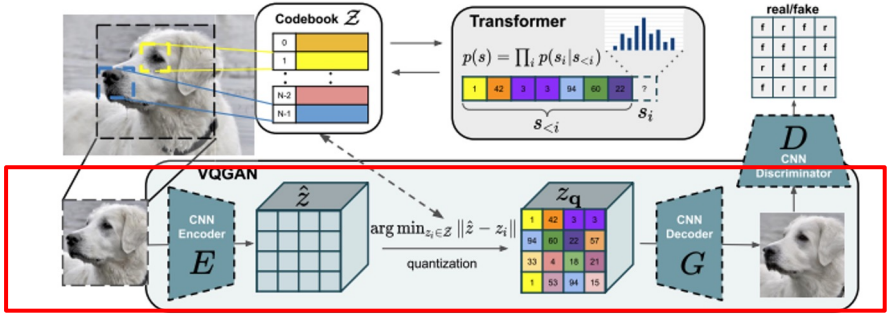
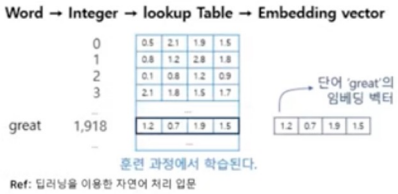


Figure 2. Our approach uses a convolutional *VQGAN* to learn a codebook of context-rich visual parts, whose composition is subsequently modeled with an autoregressive transformer architecture. A discrete codebook provides the interface between these architectures and a patch-based discriminator enables strong compression while retaining high perceptual quality. This method introduces the efficiency of convolutional approaches to transformer based high resolution image synthesis.

- **Language Model In Transformers**
  - **word vector의 embedding 방법**
  - **Lookup Table 활용**
    - **input 마다 lookup table존재**
    - **input vector = learnable vector**



- **VQGAN**
- Discrete Codebook (Lookup Table) 그대로 차용
- **Encoding - Vector Quantization - Reconstruction**
- Encoding : 16x16 input patch -> [CNN Encoder] -> 1 patch (input img의 1/16크기의 tensor)
- Vector Quantization : N개의 패치를 encoding 할 수 있는 Codebook의 value와 유클리디안 distance (argmin)
- Reconstruction : Quantization 된 Zq가 input, 즉 처음보는 data구성분포가 아닌 CodeBook에 존재하는 n개의 data로 구성분포
- 이후 Discriminator 를 이용해 adversarial loss 계산

Taming Transformers for High-Resolution Image Synthesis

- **Encoder & Decoder**
  - **Reconstruction Loss**

$$z_{\mathbf{q}} = \mathbf{q}(\hat{z}) = \left(\underset{z_k \in \mathcal{Z}}{\arg\min} \|\hat{z}_{ij} - z_k\| \right) \in \mathbb{R}^{h \times w \times n_z}. \quad (2)$$

유클리디안 distance

The reconstruction $\hat{x} \approx x$ is then given by

$$\hat{x} = G(z_{\mathbf{q}}) = G\left(\mathbf{q}(E(x))\right). \quad (3)$$

Reconstruction 값

Here, $\mathcal{L}_{\text{rec}} = \|x - \hat{x}\|^2$ is a reconstruction loss, s

Decoder가 Quantization된 Zq를 보고 이미지를 잘 Reconstruction 해낼 수 있도록 만들어주는 [G]에 대한 Loss

  - **Reconstruction Loss + Quantization**

Quantization 부분, 미분가능 ?
Lookup table latent vector = Learnable vector
학습을 통해 계속 업데이트

Encoder를 통해 나온 값 = Quantization 된 값이 닮아가도록 학습

$$\mathcal{L}_{\text{VQ}}(E, G, \mathcal{Z}) = \|x - \hat{x}\|^2 + \|\text{sg}[E(x)] - z_{\mathbf{q}}\|_2^2 + \beta \|\text{sg}[z_{\mathbf{q}}] - E(x)\|_2^2. \quad (4)$$

Here, $\mathcal{L}_{\text{rec}} = \|x - \hat{x}\|^2$ is a reconstruction loss, sg$[\cdot]$ denotes the stop-gradient operation, and $\|\text{sg}[z_{\mathbf{q}}] - E(x)\|_2^2$ is the so-called "commitment loss" with weighting factor β [62].

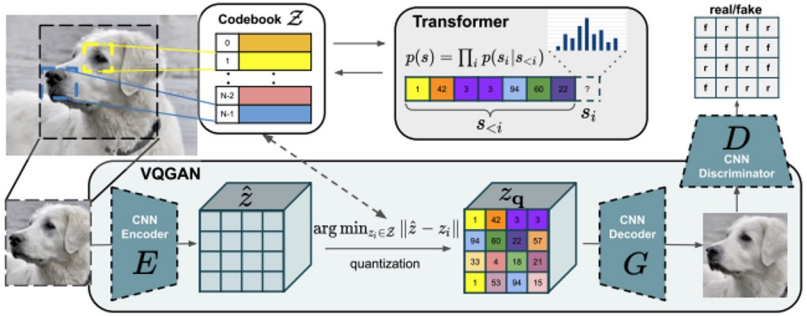


Figure 2. Our approach uses a convolutional *VQGAN* to learn a codebook of context-rich visual parts, whose composition is subsequently modeled with an autoregressive transformer architecture. A discrete codebook provides the interface between these architectures and a patch-based discriminator enables strong compression while retaining high perceptual quality. This method introduces the efficiency of convolutional approaches to transformer based high resolution image synthesis.

  - **GAN Loss**

{E,G,Z} = Generator, Discriminator Vanila GAN Loss

$$\mathcal{L}_{\text{GAN}}(\{E, G, \mathcal{Z}\}, D) = [\log D(x) + \log(1 - D(\hat{x}))] \quad (5)$$

The complete objective for finding the optimal compression model $\mathcal{Q}^* = \{E^*, G^*, \mathcal{Z}^*\}$ then reads

$$\mathcal{Q}^* = \underset{E,G,\mathcal{Z}}{\arg\min} \max_D \mathbb{E}_{x \sim p(x)} \Big[\mathcal{L}_{\text{VQ}}(E, G, \mathcal{Z}) + \lambda \mathcal{L}_{\text{GAN}}(\{E, G, \mathcal{Z}\}, D) \Big], \quad (6)$$

where we compute the adaptive weight λ according to

$$\lambda = \frac{\nabla_{G_L}[\mathcal{L}_{\text{rec}}]}{\nabla_{G_L}[\mathcal{L}_{\text{GAN}}] + \delta} \quad (7)$$

hyper param이 아닌
Adaptive Parameter.
Reconstruction Loss Gradient 와
GAN Loss Gradient의 비율을 보고
Adaptive하게 맞춰줄수 있도록 사용(?)

Taming Transformers for High-Resolution Image Synthesis

- After Encoder & Decoder Training
  - Transformer - Unconditional Generation (Seq to Seq, in NLP, 단음단어를 예
  - VQGAN을 통해 학습한 모델을 Transformer가 학습

1. Code book에서 Randint를 통해 첫번째 Code Vector 고르기
2. s = key, Zq = value(latent vector)

$$s_{ij} = k \text{ such that } (z_{\mathbf{q}})_{ij} = z_k.$$

$$\mathcal{L}_{\text{Transformer}} = \mathbb{E}_{x \sim p(x)} \left[-\log p(s) \right].$$

Log likelihood Maximize = NLL loss + Softmax

지금까지 나왔던 s를 보고 다음 번호가 무엇인가를 예측(=Sequence 예측).
즉, Code Book 중에서 다음에 나올 Sample 이 무엇인가를 예측하게 된다.

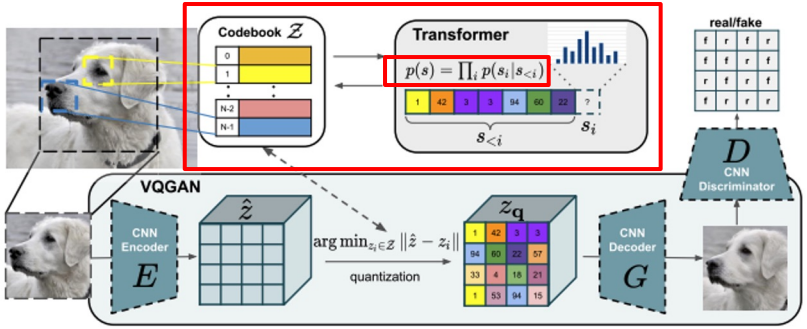


Figure 2. Our approach uses a convolutional *VQGAN* to learn a codebook of context-rich visual parts, whose composition is subsequently modeled with an autoregressive transformer architecture. A discrete codebook provides the interface between these architectures and a patch-based discriminator enables strong compression while retaining high perceptual quality. This method introduces the efficiency of convolutional approaches to transformer based high resolution image synthesis.

Inference step
- Generative 모델 = 분포에 대한 학습, 학습한 분포로 부터 Sampling 하는 것
- Softmax 분포에 따라 Random Sampling

High Resolution ?
- 이미지가 커지면 커질수록 생성되는 Patch들이 많아지므로
  마지막 Patch 만들때는 이전에 만든 Patch 들에 대한 Attention을 구하기 위해 연산량이

- Sliding Window를 통해 인접한 patch들과의 Attention만 연산진행

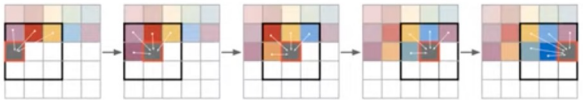


Figure 3. Sliding attention window.

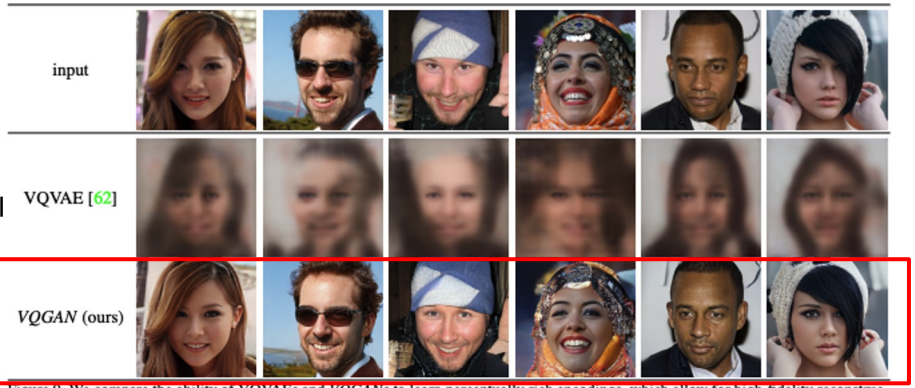


Figure 9. We compare the ability of VQVAEs and *VQGAN*s to learn perceptually rich encodings, which allow for high-fidelity reconstructions with large factors f. Here, using the same architecture and $f = 16$, VQVAE reconstructions are blurry and contain little information about the image, whereas *VQGAN* recovers images faithfully. See also Sec. B.

Taming Transformers for High-Resolution Image Synthesis

신기합니다 .. * 0 *

Super Resolution

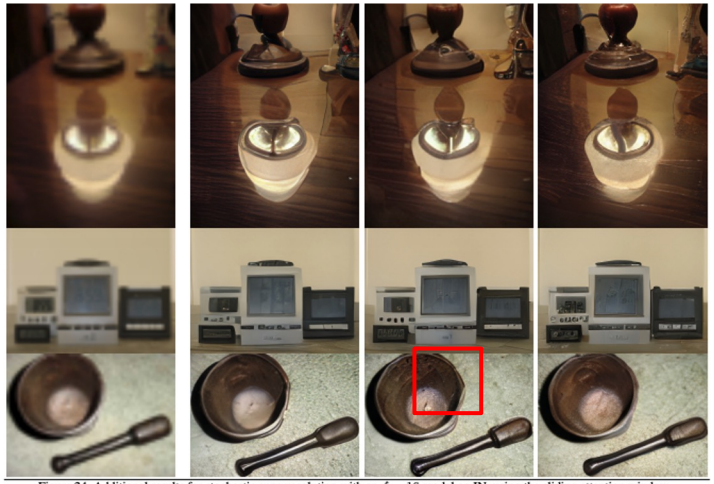


Figure 24. Additional results for stochastic superresolution with an $f = 16$ model on IN, using the sliding attention window.

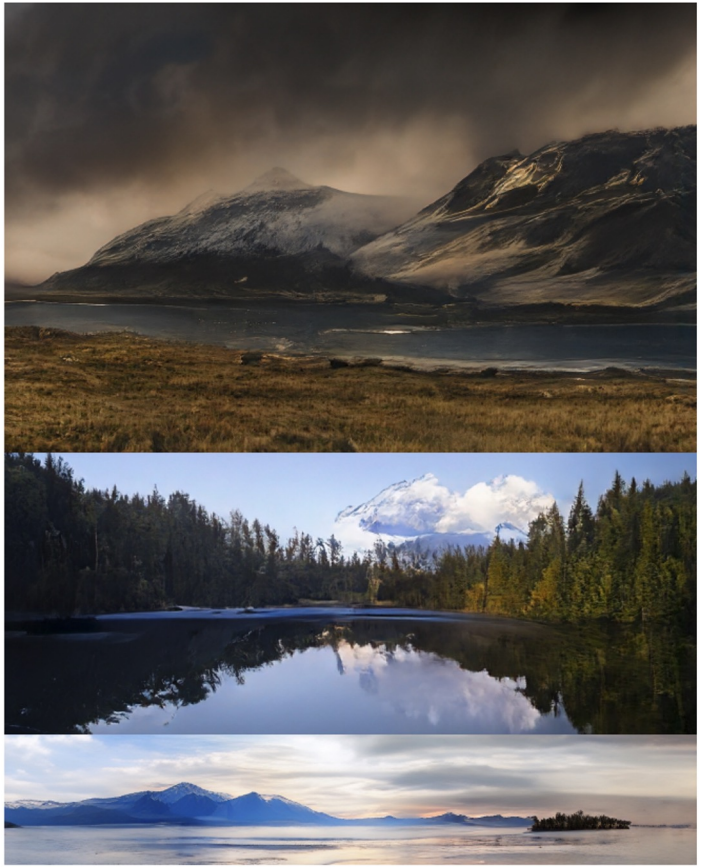


Figure 17. Samples generated from semantic layouts on S-FLCKR. Sizes from top-to-bottom: 1280 × 832, 1024 × 416 and 1280 × 240 pixels.

Softmax, Random Sampling = 약간 pixel 구성이 다름

https://arxiv.org/pdf/2104.06490.pdf

DatasetGAN: Efficient Labeled Data Factory with Minimal Human Effort

Yuxuan Zhang[1,5*] Huan Ling[1,2,3,*] Jun Gao[1,2,3] Kangxue Yin [1]
Jean-Francois Lafleche[1] Adela Barriuso[4] Antonio Torralba[4] Sanja Fidler[1,2,3]
NVIDIA[1] University of Toronto[2] Vector Institute[3] MIT[4] University of Waterloo[5]
y2536zha@uwaterloo.ca, {huling, jung, kangxuey, jlafleche}@nvidia.com
adela.barriuso@gmail.com, torralba@mit.edu, sfidler@nvidia.com

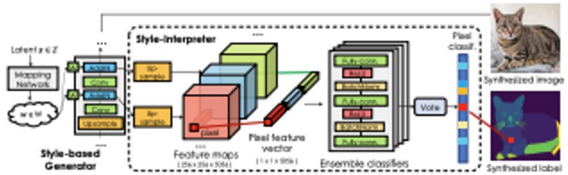


Figure 2: Overall architecture of our DATASETGAN. We upsample the feature maps from StyleGAN to the highest resolution for constructing pixel-wise feature vectors for all pixels on the synthesized image. An ensemble of MLP classifiers is then trained for interpreting the semantic knowledge in the feature vector of a pixel into its part label.

DatasetGAN

- 인간의 최소한의 노력으로 high-quality semantically segmented images 의 대규모 데이터 세트를 생성하는 자동 절차

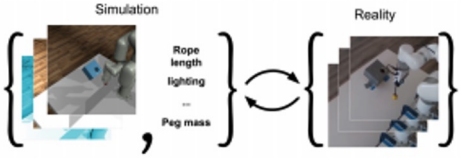
- StyleGAN 을 사용하여 realistic images + label을 생성함

1) 생성된 데이터 세트는 실제 데이터 세트와 마찬가지로 모든 컴퓨터 비전 아키텍처를 훈련하는데 사용될 수 있음.
2) DatasetGAN 을 보여주기 위해, 34개의 인간 얼굴 부분과 32개의 자동차 부품에 대한 데이터셋을 생성함.
3) 제안 접근 방식은 semi-supervised baselines 을 크게 능가하며, supervised methods 와 동일한 효과를 보임.

https://arxiv.org/pdf/2104.07662.pdf

Auto-Tuned Sim-to-Real Transfer

Yuqing Du*[1], Olivia Watkins*[1], Trevor Darrell[1], Pieter Abbeel[1], Deepak Pathak[2]
[1] UC Berkeley, [2] Carnegie Mellon University



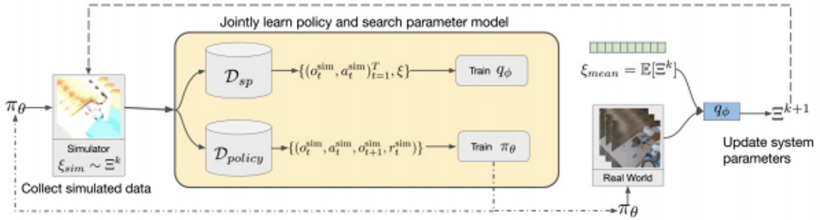


Fig. 2: Overall System: Using any off-the-shelf RL algorithm, we use simulated data to train both our policy and a Search Parameter Model (SPM) q_ϕ which predicts whether a candidate set of system parameters ξ is higher or lower than those which produced an observed trajectory. We iteratively update our simulation by running our policy in the real world and using our SPM to predict which direction to update our simulator to make it closer to the real world.

- **4월 중순부터 꾸준히 이슈가 되고 있는 논문(Top Recent),** *회사연구에 도움이 되는 논문인지는 아닌..*
  - Real world에서의 robot learning = costly and time consuming
  - 하지만 시뮬레이션에서의 학습은 real world 로 transfer 될 때, 'reality gap' 으로 인해 실패하는 경우가 많음.

- **Paper Approach**
- rewards (보상) 를 정의하거나, state를 추정할 필요 없이
  real world 의 RGB 영상 이미지만을 사용해 시뮬레이터 시스템 파라미터를 real world 에 자동으로 matching 되도록 tuning 하는 방법을 제안함.
- **Real World 시스템 파라미터에 접근하기 위해, 시뮬레이션 시스템 파라미터를 반복적으로 변환하여 파라미터를 Auto-tuning**
- Search Param Model (SPM) 제안
  - SPM: 주어진 파라미터가 실제 파라미터보다 높은지 낮은지를 예측하는 모델
- 로봇 제어 작업에 대한 제안 모델 방법을 Sim-to-Sim, Sim-to-Real 전송 모두에서 평가함.
- 단순한 도메인 randomization 에 대한 개선.

Self-supervised Object detection from audio-visual correspondence

https://arxiv.org/pdf/2104.06401.pdf

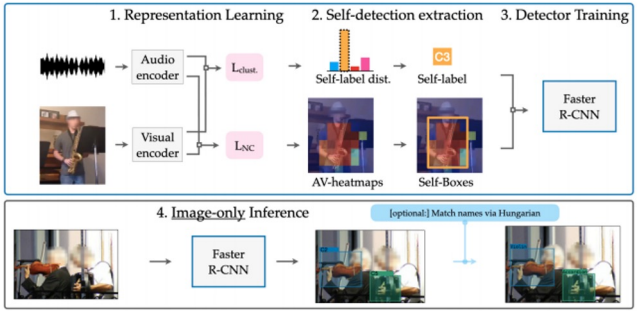


Figure 2: Self-supervised object detection from audio-visual correspondence: We combine noise-constrastive and clustering-based self-supervised learning to generate self-detections (boxes and labels) and use those as targets to train a detector. The trained detector can be used to detect objects from many categories on images without requiring audio.

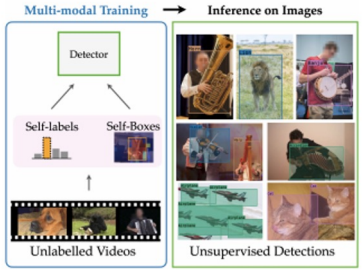


Figure 1: **We train an object detector simply by watching videos.** Without using any manual annotations, we learn to detect different objects in images, by first self-labelling boxes and object categories and then using those as targets to teach a detector. The detection results shown are outputs from our trained model; for visualisation purposes we show Hungarian-matched labels.

Supervision 없이 object detectors 를 학습하는 문제 해결

- Differently from Weakly-Supervised Object Detection (We do not assume Image-level class Labels)
- Instead, Extracting Signal from Audio-visual data

-> object detector를 'teach' 하기 위해 audio-visual data 에서 audio component 를 추출함.

- 해당 문제는 sound source localisation 과 관련이 있지만 다음의 이유로 쉽게 해결하기 어려움.
  - Detector must classify the objects by type
  - Enumerate each instance of the object
  - Object is silent

Method	No labels?	VGGSound mAP$_{30}$	VGGSound mAP$_{50}$	VGGSound mAP$_{[50:95:5]}$	Audioset mAP$_{30}$	Audioset mAP$_{50}$	Audioset mAP$_{[50:95:5]}$	OpenImages mAP$_{30}$	OpenImages mAP$_{50}$	OpenImages mAP$_{[50:95:5]}$
PCL (WSOD) [90]	✗	54.9	27.7	7.6	39.0	17.5	4.4	37.9	14.5	3.5
Ours - weak sup.	✗	67.6	42.9	14.2	50.6	30.9	10.3	48.9	33.7	9.5
Center Box*	✓	29.6	5.6	1.5	15.1	3.5	0.7	20.7	4.2	0.8
Selective Search* [94]	✓	5.2	1.1	0.4	2.8	0.4	0.1	7.4	2.1	0.7
COCO-trained RPN*	✗	33.4	7.5	1.6	19.0	4.1	0.8	24.4	11.1	2.6
Ours - self-boxes*	✓	48.1	29.6	10.0	27.8	14.1	4.8	NA	NA	NA
Ours - full	✓	**52.3**	**39.4**	**14.7**	**44.3**	**28.0**	**9.6**	**39.9**	**28.5**	**7.6**

Table 1: **Self-supervised object detection.** We report object detection metrics across three test datasets and find our method is far superior to other unsupervised approaches and outperforms even the weakly supervised baseline in most metrics. For methods denoted by *, we report class-agnostic evaluation numbers. *Center Box* denotes a simple baseline predicting random sized boxes in the middle of the frame. *Ours-weak sup.* is a variant of our model trained with the video-level category annotations in combination with our self-extracted boxes. The class-agnostic performance of the self-boxes that are used to train the detector reveals that the latter greatly outperforms them, which highlights the benefit of our approach.

First Designing, a Self-supervised framework with a contrastive objective that jointly learns to classify and localise objects.

- 제안하는 object detector 방식으로 기존의 unsupervised and weakly-supervised detectors 의 object detection과 sound source localization task 를 능가함.
  - Musical Instruments, airplanes, cats

End of the Document