CityGuessr: City-Level Video Geo-Localization on a Global Scale

Parth Parag Kulkarni[1], Dr. Gaurav Kumar Nayak[2], Dr. Mubarak Shah[1]

[1]Center for Research in Computer Vision, University of Central Florida

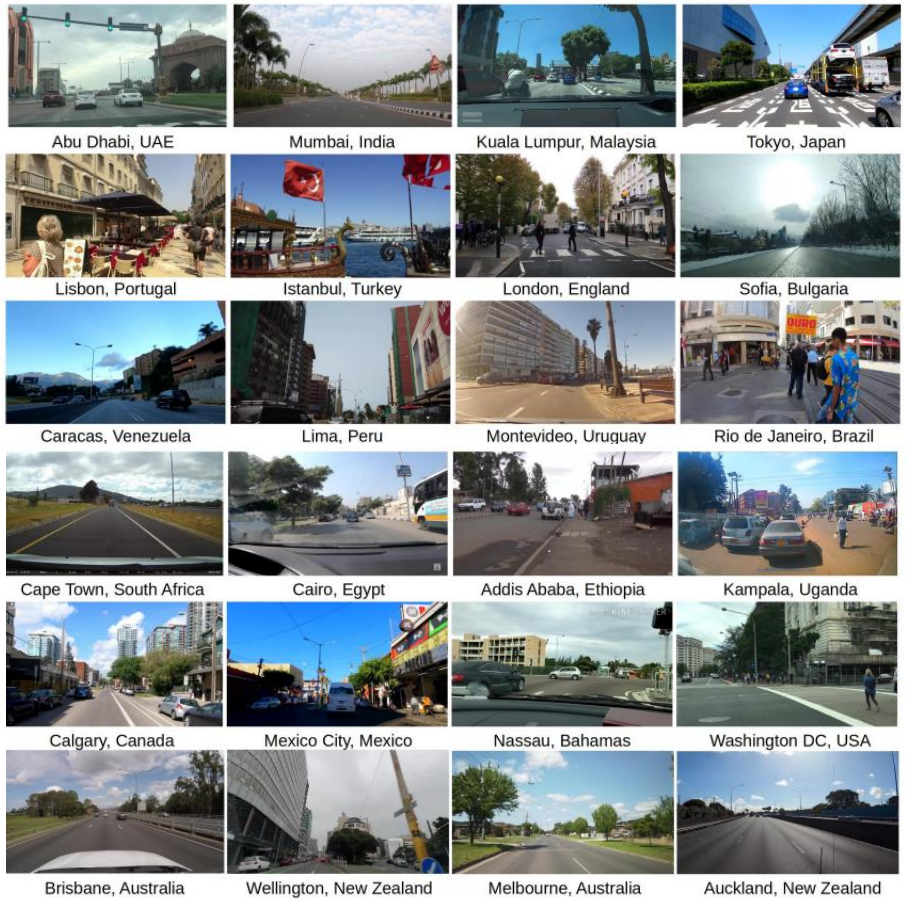
[2]Mehta Family School of DS & AI, Indian Institute of Technology Roorkee, India

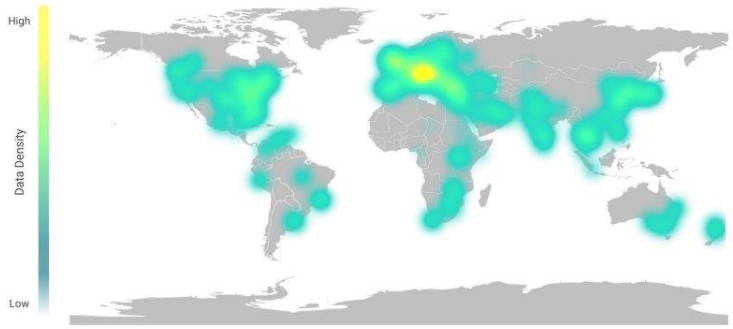
Preview

- We formulate a novel problem of worldwide video geolocalization
- first global-scale video dataset named ‘CityGuessr68k’(68,269 videos, 166 cities)
- transformer-based architecture with two primary components
  - Self-Cross Attention module for incorporating scenes
  - TextLabel Alignment strategy for distilling knowledge from textlabels in feature space
- performance results on CityGuessr68k as well as Mapillary(MSLS)[1] datasets.

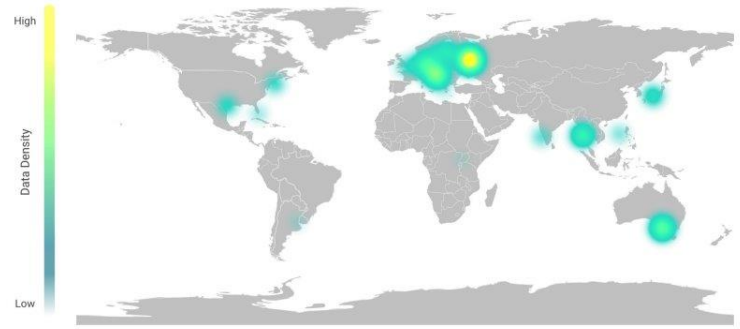
CityGuessr68k Dataset

- ~ 68000 first-person driving and walking videos from 166 cities around the world
- annotated with hierarchical location labels
- primary benchmarking dataset for this task





(a) Distribution of videos in CityGuessr68k



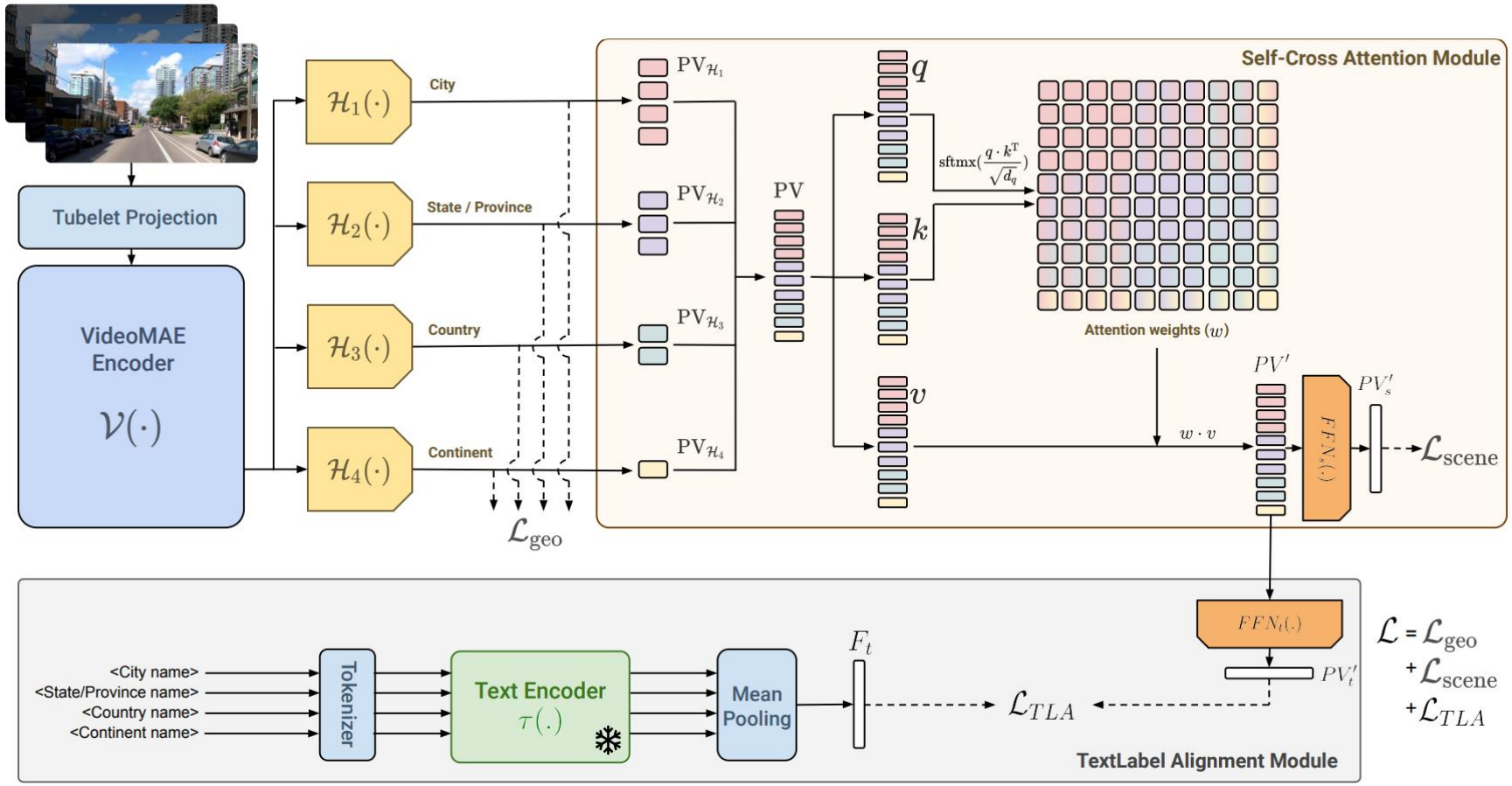
(b) Distribution of videos in Mapillary(MSLS)

Problem Statement

“Given an input video, determining which city in the world, this video was recorded in”

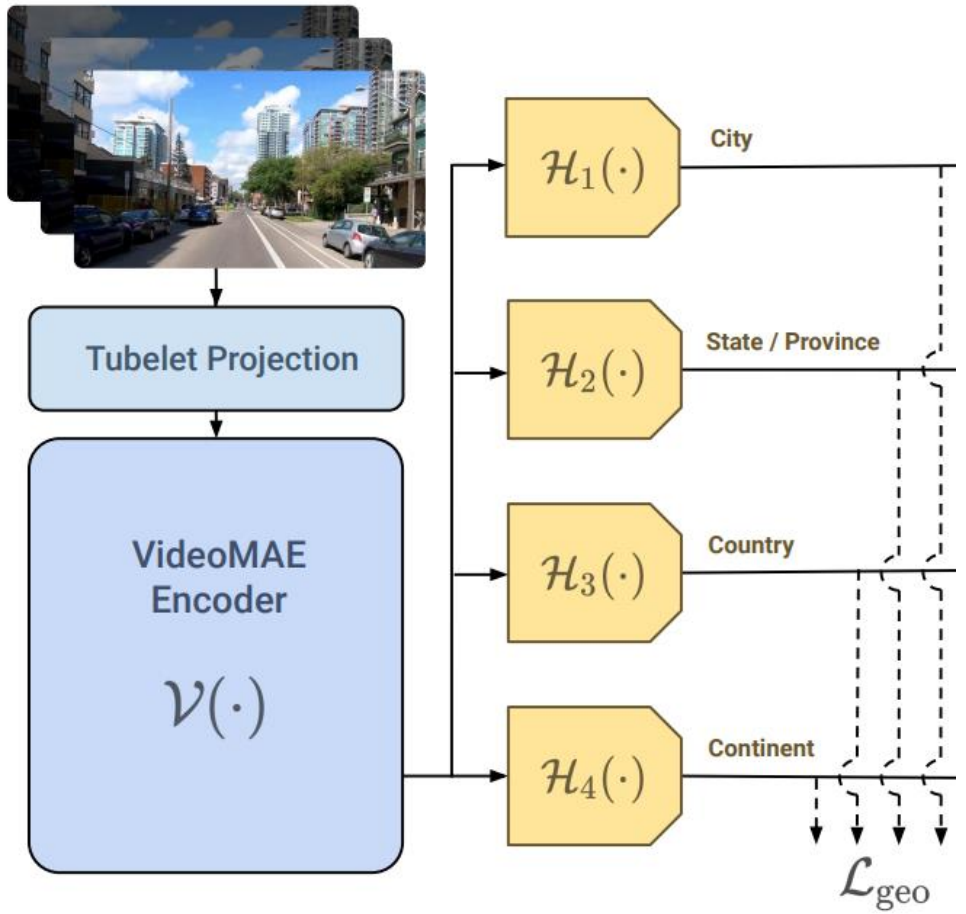
- Consequently, determining state/province, country and continent

Model

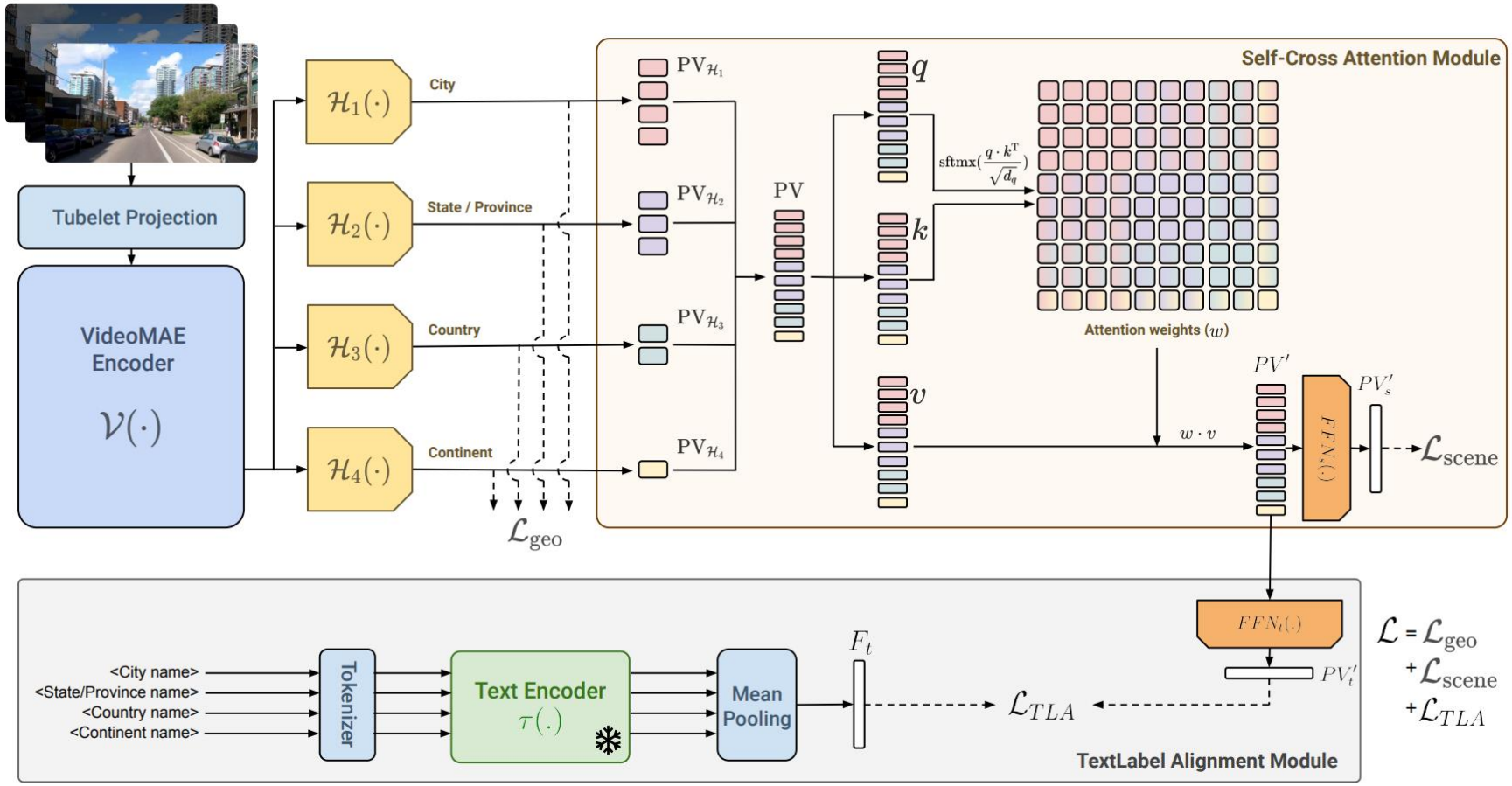

Tubelet Projection
VideoMAE Encoder
$\mathcal{V}(\cdot)$
$\mathcal{H}_1(\cdot)$
$\mathcal{H}_2(\cdot)$
$\mathcal{H}_3(\cdot)$
$\mathcal{H}_4(\cdot)$
City
State / Province
Country
Continent
$\mathcal{L}_{geo}$
$PV_{\mathcal{H}_1}$
$PV_{\mathcal{H}_2}$
$PV_{\mathcal{H}_3}$
$PV_{\mathcal{H}_4}$
PV
Self-Cross Attention Module
q
k
v
$\text{sftmx}(\frac{q \cdot k^T}{\sqrt{d_q}})$
Attention weights (w)
$w \cdot v$
PV'
$FFN_s(.)$
PV'_s
$\mathcal{L}_{scene}$
<City name>
<State/Province name>
<Country name>
<Continent name>
Tokenizer
Text Encoder
$\tau(.)$
Mean Pooling
F_t
$\mathcal{L}_{TLA}$
$FFN_t(.)$
PV'_t
TextLabel Alignment Module
$\mathcal{L} = \mathcal{L}_{geo} + \mathcal{L}_{scene} + \mathcal{L}_{TLA}$

Model – Encoder Backbone and Classifiers

- VideoMAE [2] backbone
- Pretrained on Kinetics-400 [3]
- Outputs a 384-dimensional feature vector
- Passed into each classifier
- All outputs used for computing geolocalization loss

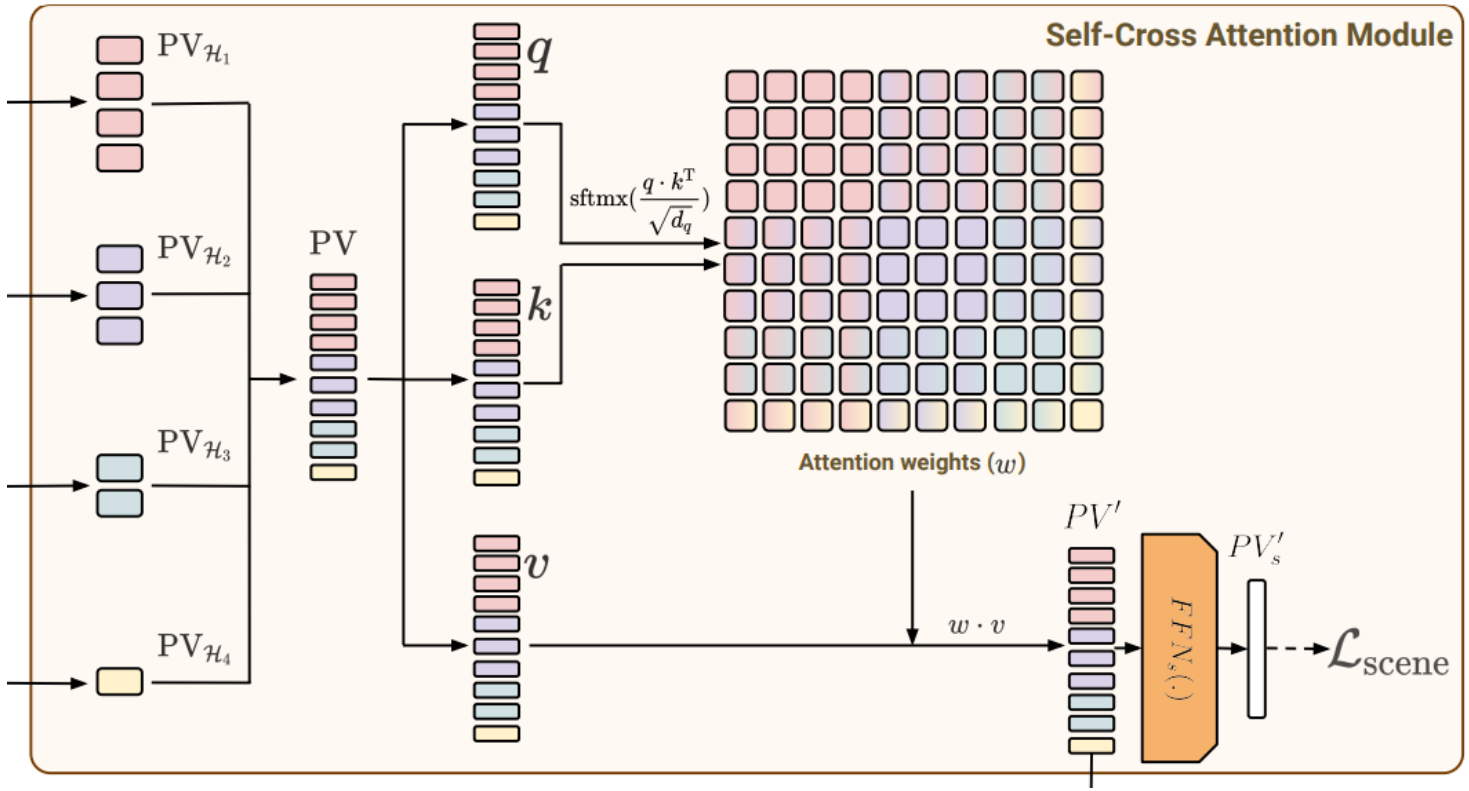


Model


Tubelet Projection
VideoMAE Encoder
$\mathcal{V}(\cdot)$
$\mathcal{H}_1(\cdot)$
$\mathcal{H}_2(\cdot)$
$\mathcal{H}_3(\cdot)$
$\mathcal{H}_4(\cdot)$
City
State / Province
Country
Continent
$\mathcal{L}_{geo}$
$PV_{\mathcal{H}_1}$
$PV_{\mathcal{H}_2}$
$PV_{\mathcal{H}_3}$
$PV_{\mathcal{H}_4}$
PV
Self-Cross Attention Module
q
k
v
$\text{sftmx}(\frac{q \cdot k^T}{\sqrt{d_q}})$
Attention weights (w)
$w \cdot v$
PV'
$FFN_s(.)$
PV'_s
$\mathcal{L}_{scene}$
<City name>
<State/Province name>
<Country name>
<Continent name>
Tokenizer
Text Encoder
$\tau(.)$
Mean Pooling
F_t
$\mathcal{L}_{TLA}$
$FFN_t(.)$
PV'_t
TextLabel Alignment Module
$\mathcal{L} = \mathcal{L}_{geo} + \mathcal{L}_{scene} + \mathcal{L}_{TLA}$

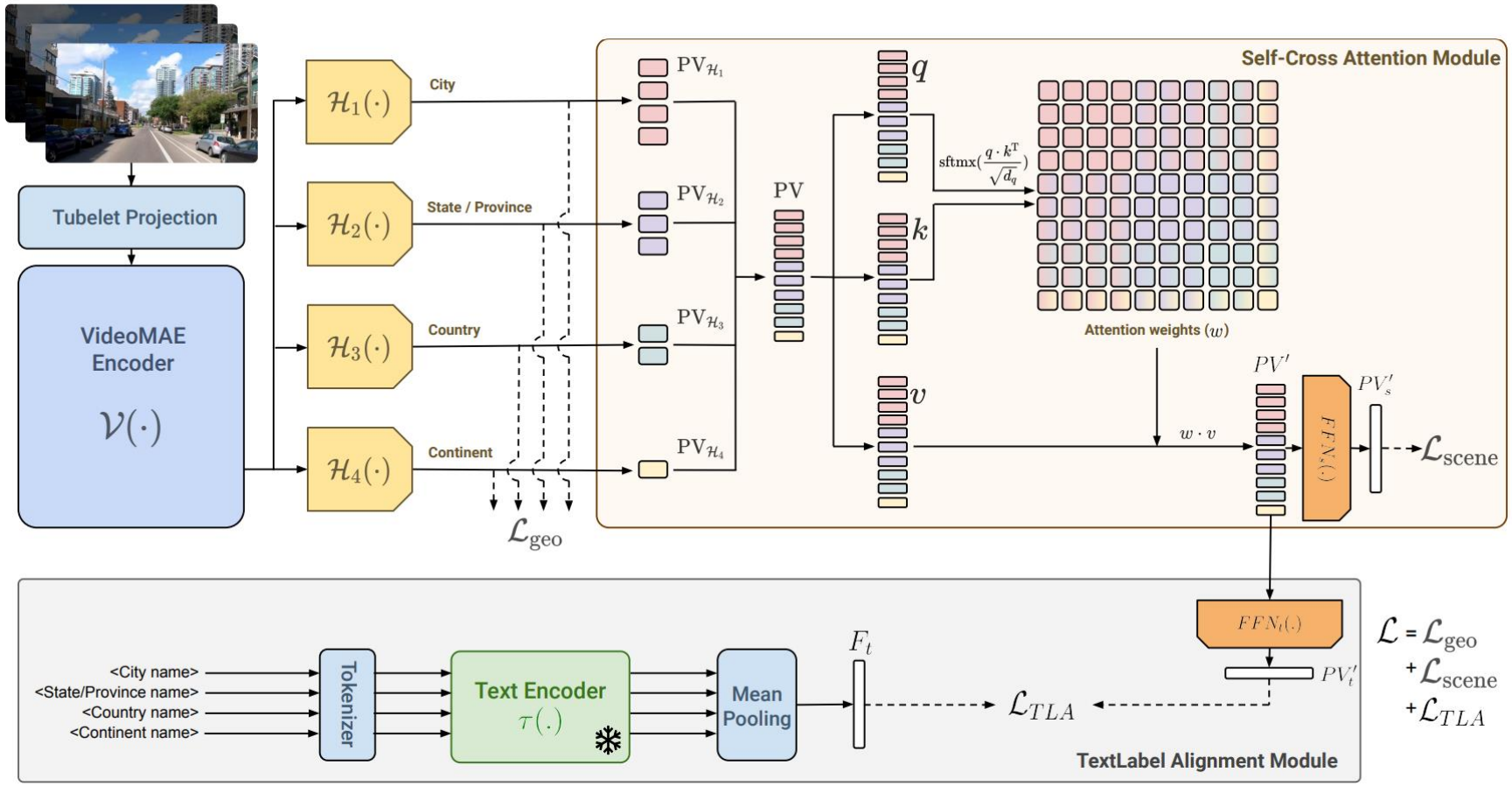
Model – Scene Recognition

- Auxiliary task
- For scene identification
  - fuse the knowledge from all 4 hierarchies
- Tokens of all hierarchies :
  - Self attention with itself
  - Cross attention with others

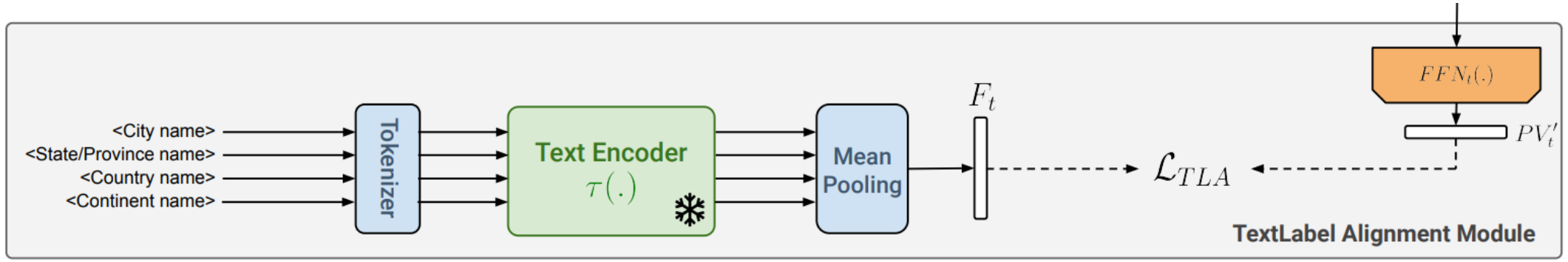


- Output used to compute scene loss
- We use soft scene labels to capture scene knowledge from all frames of a video

Model


Tubelet Projection
VideoMAE Encoder
$\mathcal{V}(\cdot)$
$\mathcal{H}_1(\cdot)$
$\mathcal{H}_2(\cdot)$
$\mathcal{H}_3(\cdot)$
$\mathcal{H}_4(\cdot)$
City
State / Province
Country
Continent
$\mathcal{L}_{geo}$
$PV_{\mathcal{H}_1}$
$PV_{\mathcal{H}_2}$
$PV_{\mathcal{H}_3}$
$PV_{\mathcal{H}_4}$
PV
Self-Cross Attention Module
q
k
v
$\text{sftmx}(\frac{q \cdot k^T}{\sqrt{d_q}})$
Attention weights (w)
$w \cdot v$
PV'
$FFN_s(.)$
PV'_s
$\mathcal{L}_{scene}$
<City name>
<State/Province name>
<Country name>
<Continent name>
Tokenizer
Text Encoder
$\tau(.)$
Mean Pooling
F_t
$\mathcal{L}_{TLA}$
$FFN_t(.)$
PV'_t
TextLabel Alignment Module
$\mathcal{L} = \mathcal{L}_{geo} + \mathcal{L}_{scene} + \mathcal{L}_{TLA}$

Model – TextLabel Alignment Strategy



- Associating a name with a picture/video of a location is helpful for humans
- distill knowledge from the textlabels of all hierarchies to model features

Inference – Hierarchical Evaluation

Country $$P'(C_k^{H_3}|V) = P(C_k^{H_3}|V) * P(C_l^{H_4}|V)$$

State/Province $$P'(C_j^{H_2}|V) = P(C_j^{H_2}|V) * P'(C_k^{H_3}|V) * P(C_l^{H_4}|V)$$

City $$P'(C_i^{H_1}|V) = P(C_i^{H_1}|V) * P'(C_j^{H_2}|V) * P'(C_k^{H_3}|V) * P(C_l^{H_4}|V)$$

- predictions for fine-grained hierarchies could be improved with the assistance of the coarser hierarchies

- Independent :
  - Predict every hierarchy independently

- Codependent :
  - Predict the finest hierarchy only
  - Then back trace the coarser hierarchy predictions

UCF

Experiments, Results and Discussion

Image v/s Video

Backbone	Setting	City	State	Country	Continent
MAE	First frame	52.1	52.6	55.3	70.4
	Mid frame	48.9	49.3	54.6	69.8
	Last frame	48.1	48.4	53.4	69.3
	Random frame	55.8	56.3	60.8	74.1
VideoMAE	video	64.5	64.5	65.9	74.4

Impact of Proposed models

Scenes	TLA	City	State	Country	Continent
-	-	64.5	64.5	65.9	74.4
Majority	-	66.9	67.3	72.1	81.1
Soft	-	67.9	68.4	72.4	81.6
Soft	city only	69.1	69.5	73.7	83.1
Soft	all hierarchies	**69.6**	**70.2**	**74.8**	**83.8**

Independent v/s Codependent Hierarchical Evaluation

Model	City	State	Country	Continent
w/o hierarchical eval.	69.1	69.6	72.5	79.2
Independent	69.6	69.8	72.5	79.2
Codependent	**69.6**	**70.2**	**74.8**	**83.8**

Comparison with State-of-the-art

Model	City	State	Country	Continent
PlaNet [37]	55.8	56.3	60.8	74.1
ISNs [20]	59.5	59.9	64.1	75.9
GeoDecoder [3]	64.2	64.5	69.5	79.9
Timesformer [2]	60.9	61.4	66.1	78.4
VideoMAE [32]	64.5	64.5	65.9	74.4
Ours	**69.6**	**70.2**	**74.8**	**83.8**

UCF

Performance on Mapillary(MSLS)

Model	City	State	Country	Continent
VideoMAE	67.6	67.6	68.2	81.9
Ours	**72.8**	**72.8**	**73.2**	**88.1**

Examples of Localizations – city correct

Pred. City : Hiroshima

Pred. City : Montevideo

Pred. City : Johannesburg

Pred. City : Las Vegas

Pred. City : Harare

Pred. City : Busan

Pred. City : Agra

Pred. City : Amsterdam

Conclusion

- We formulated a novel problem of worldwide video geolocalization
- We introduced a new global level video dataset, CityGuessr68k, containing 68,269 videos from 166 cities.
- We also proposed a baseline approach which consists of
  - Self-Cross Attention module for incorporating an auxiliary task of scene recognition
  - TextLabel Alignment strategy to distill knowledge from location labels in feature space.
- We demonstrated the efficacy of our method on our dataset as well as on Mapillary(MSLS) dataset
- Future direction
  - to explore the generalizability of the combination of Self-Cross Attention module and TextLabel Alignment to other hierarchical video classification tasks.

UCF

References

[1] Warburg, F., Hauberg, S., Lopez-Antequera, M., Gargallo, P., Kuang, Y., Civera, J.: Mapillary street-level sequences: A dataset for lifelong place recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2626–2635 (2020) 1, 3, 4, 5, 6, 13

[2] Tong, Z., Song, Y., Wang, J., Wang, L.: Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. Advances in neural information processing systems 35, 10078–10093 (2022) 7, 13

[3] Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al.: The kinetics human action video dataset. arXiv preprint arXiv:1705.06950 (2017) 7, 11

UCF