Label-Free Event-based Object Recognition via Joint Learning with Image Reconstruction from Events

–Supplementary materials–

Hoonhee Cho*, Hyeonseong Kim*, Yujeong Chae, and Kuk-Jin Yoon
Korea Advanced Institute of Science and Technology
{gnsgnsgml,brian617,yujeong,kjyoon}@kaist.ac.kr

This supplementary material provides how to employ web-crawled images, implementation and dataset details, and more qualitative results. In particular, the following contents are included in the supplementary material:

- Results of employing web-crawled images.
- Implementation details of the proposed method.
- Category split of datasets used in our experiments.
- Additional experiments and qualitative results.

1. Describing Notations and Abbreviations

We provide the descriptions of notations and abbreviations in Table 1 for better understanding.

Abbrev.	Description	Abbrev.	Description
$\mathcal{I}$	Intensity image	$\mathcal{L}_{att}$	Category-guided attraction loss
$\mathcal{E}$	Event stream	$\mathcal{L}_{rep}$	Category-agnostic repulsion loss
$\mathcal{G}$	Recon. network	PPI	Posterior probability indicator
f_i	i-th textual feature	TRCI	Temporally reversed consistency indicator
v_i	i-th visual feature	S_{PPI}	Selected sample indices from PPI
p_i	i-th category prob.	S_{TRCI}	Selected sample indices from TRCI
c	Predicted category	S_{RDS}	Intersection of S_{PPI} and S_{TRCI}
p_c	Posterior prob. for c	$\mathcal{L}_{con}$	Local-global recon. consistency loss
$\mathcal{E}^R$	Reversed event stream	$\mathcal{K}$	Num. of data samples
w_i	i-th prototype feature	$\mathcal{L}$	Num. of clusters

Table 1. Notions and abbreviations.

2. Employing Web-crawled Images

In the main paper, we split the dataset in half and used half of them as unpaired images of the other half, paying attention to the ease of experimentation or follow-up research. In this section, instead of using the original dataset, we collected unpaired images by performing web crawling from google. Examples of web-crawled images are shown in Fig. 1. As shown in Table 2, we report the results and confirm that using web-crawled images is comparable and even better than using unpaired images from the original dataset in all metrics. These findings affirm the practical efficacy of our approach in harnessing web crawling for improved performance.

*The first two authors contributed equally. In alphabetical order.

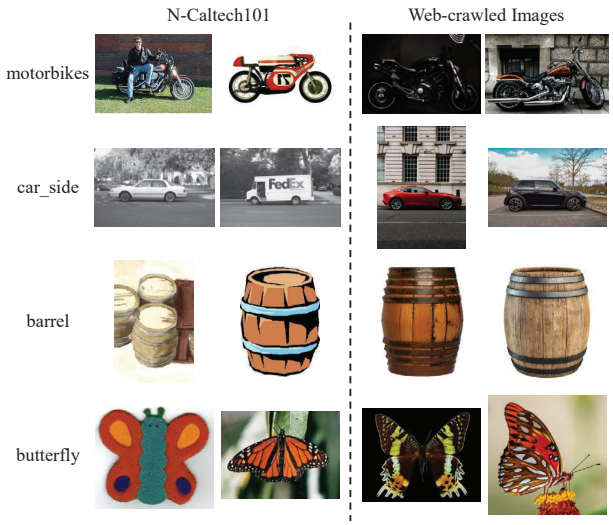

Figure 1. Examples from N-Caltech101 and web-crawled images of the corresponding categories.

Methods	N-Caltech101		
	Accuracy ↑	FID ↓	IS ↑
Ours (Text Prompt)	82.46	62.29	14.81
Ours (Visual Prototype: N-Caltech101)	<u>82.61</u>	<u>54.19</u>	<u>17.59</u>
Ours (Visual Prototype: Web Crawling)	**82.64**	**51.83**	**18.25**

Table 2. The results of our method with web crawling.

3. Implementation Details

We implement our framework on PyTorch [6] and adopt the Vision Transformer [1] with a patch size of 32, called ViT-B/32, as the visual encoder and the transformer [8] as the textual encoder. We design the image reconstruction network $\mathcal{G}$ with U-Net [7] consisting of two residual blocks [3]. For event representation, we utilize the Event Spike Tensor (EST) [2] using a bin size of 9, which is the same as the original paper. For the crop function $\mathcal{H}$ in Sec. 3.3 of the main paper, we use a crop size of 128×128 while we resize the EST to 224×224 for all datasets.

Dataset	N-Caltech101		N-ImageNet (Mini)	
Setting	standard	zero-shot	standard	zero-shot
Train	Faces_easy, Leopards, Motorbikes, accordion, airplanes, anchor, ant, barrel, bass, beaver, binocular, bonsai, brain, brontosaurus, buddha, butterfly, camera, cannon, car_side, ceiling_fan, cellphone, chair, chandelier, cougar_body, cougar_face, crab, crayfish, crocodile, crocodile_head, cup, dalmatian, dollar_bill, dolphin, dragonfly, electric_guitar, elephant, emu, euphonium, ewer, ferry, flamingo, flamingo_head, garfield, gerenuk, gramophone, grand_piano, hawksbill, headphone, hedgehog, helicopter, ibis, inline_skate, joshua_tree, kangaroo, ketch, lamp, laptop, llama, lobster, lotus, mandolin, mayfly, menorah, metronome, minaret, nautilus, octopus, okapi, pagoda, panda, pigeon, pizza, platypus, pyramid, revolver, rhino, rooster, saxophone, schooner, scissors, scorpion, sea_horse, snoopy, soccer_ball, stapler, starfish, stegosaurus, stop_sign, strawberry, sunflower, tick, trilobite, umbrella, watch, water_lilly, wheelchair, wild_cat, windsor_chair, wrench, yin_yang	Faces_easy, Motorbikes, accordion, airplanes, anchor, ant, barrel, beaver, binocular, bonsai, brain, brontosaurus, buddha, butterfly, camera, cannon, car_side, ceiling_fan, cellphone, chair, chandelier, cougar_body, cougar_face, crab, crayfish, crocodile_head, cup, dalmatian, dollar_bill, dolphin, electric_guitar, elephant, emu, ewer, ferry, flamingo_head, garfield, gerenuk, gramophone, grand_piano, hawksbill, helicopter, inline_skate, lobster, lotus, mandolin, mayfly, menorah, metronome, nautilus, octopus, okapi, pagoda, panda, pigeon, pizza, platypus, pyramid, revolver, rhino, rooster, schooner, scissors, scorpion, snoopy, soccer_ball, stapler, starfish, stegosaurus, stop_sign, strawberry, sunflower, tick, trilobite, umbrella, watch, water_lilly, wheelchair, wild_cat, windsor_chair	hamster, academic gown, airship, jackfruit, barbershop,cocktail shaker, Komodo dragon, sunglasses, grey fox,cello, comic book, goldfish, Bloodhound, porcupine,jaguar, kingsnake, altar, water buffalo, chiton, scarf, storage chest, tool kit, sea anemone, Border Terrier,menu, picket fence, forklift, yellow ladys slipper, chameleon, dragonfly, Pomeranian, European garden spider,Airedale Terrier, frilled-necked lizard, black stork, valley, radio telescope, leopard, crossword, Australian Terrier, Shih Tzu, husky, can opener, artichoke, assault rifle, fountain pen, harvestman, parallel bars, harmonica, half-track, snoek fish, pencil sharpener, submarine, muzzle, eastern diamondback rattlesnake, Miniature Schnauzer,missile, Komondor, grand piano, website, king penguin,canoe, red-breasted merganser, trolleybus, quail,poke bonnet, King Charles Spaniel, race car, Malinois, solar thermal collector, slug, bucket, dung beetle,Asian elephant, window screen, Flat-Coated Retriever, steel drum, snowplow, handkerchief, tailed frog, church,Chesapeake Bay Retriever, Christmas stocking, hatchet, hair clip, vulture, sidewinder rattlesnake, oscilloscope, worm snake, eel, wok, planetarium, Old English Sheepdog,platypus, Pembroke Welsh Corgi, alligator lizard, consomme, African rock python, hot tub, Tibetan Mastiff	hamster, academic gown, jackfruit, barbershop, Komodo dragon, sunglasses, grey fox, cello, comic book, goldfish, Bloodhound, porcupine, jaguar, altar, water buffalo, chiton, scarf, storage chest, tool kit, sea anemone, Border Terrier, menu, picket fence, forklift, yellow ladys slipper, chameleon, dragonfly, European garden spider, Airedale Terrier, frilled-necked lizard, black stork, valley, radio telescope, leopard, crossword, Australian Terrier, Shih Tzu, husky, can opener, assault rifle, fountain pen, harvestman, parallel bars, half-track, snoek fish, pencil sharpener, submarine, muzzle, eastern diamondback rattlesnake, Miniature Schnauzer, king penguin,canoe, red-breasted merganser, trolleybus, quail, poke bonnet, King Charles Spaniel, Malinois, solar thermal collector,bucket, dung beetle, Asian elephant, window screen,Flat-Coated Retriever, steel drum, snowplow, handkerchief, tailed frog, church, Chesapeake Bay Retriever, Christmas stocking, sidewinder rattlesnake, oscilloscope, worm snake, eel, wok, platypus, alligator lizard, consomme, African rock python
Test	Same as train	laptop, bass, joshua_tree, Leopards, ibis, hedgehog, minaret, crocodile, flamingo, headphone, ketch, saxophone, euphonium, dragonfly, wrench, llama, lamp, kangaroo, yin_yang, sea_horse	Same as train	hot tub, harmonica, grand piano, hatchet, Pomeranian, missile, slug, vulture, kingsnake, Pembroke Welsh Corgi, cocktail shaker, race car, airship, Tibetan Mastiff, Old English Sheepdog, planetarium, hair clip, Komondor, artichoke, website

Table 3. Category split for standard and zero-shot experimental settings in N-Caltech101 [5] and N-ImageNet (Mini) [4]. In the standard setting, categories are identical for training and testing. In the zero-shot setting, there is no common category for training and testing.

We use agglomerative clustering for the clustering algorithm in Sec. 3.4 of the main paper. For the number of clusters, we use $L = 3$ for N-Caltech101 and $L = 10$ for N-ImageNet (Mini). We use LAMB [9] optimizer with a weight decay of 1×10^{-4} and initialize the learning rate to 6×10^{-3}. We train the network with a batch size of 32. We set the weights λ_1, λ_2, λ_3 and parameter K in PPI as 1, 0.01, 1, and 6, respectively.

4. Dataset Details

In Table 3, we present the details about the category split of each dataset used in our experiments. In the main paper, the experiments involve two settings: standard and zero-shot. In the standard setting, the categories are identical for training and testing, evaluating the ability to recognize objects belonging to the same set of categories for training. In contrast, there is no common category for training and testing in the zero-shot setting. This evaluates the model's ability to generalize to novel categories that it has not been explicitly trained on.

5. Additional Experiments and Discussions

5.1. Using Ground Truth instead of Pseudo Label

Method	N-Caltech101	N-ImageNet (Mini)
Ours	82.46	30.16
Oracle	**84.28**	**32.98**

Table 4. The results of our method using ground truth.

Table 4 demonstrates that using ground truth in training indeed yields better performance than using pseudo labels. Nonetheless, through reliable sampling, our method achieves comparable performance, indicating the superiority of our approach.

5.2. Ablation Study on Repulsion Loss

Loss type	N-Caltech101	N-ImageNet (Mini)
Category-agnostic (Eq. 5)	**82.46**	**30.16**
Category-aware	81.48	28.54

Table 5. The results of using category-agnostic and category-aware repulsion losses.

In Table 5, we report the result using category-aware repulsion loss that separates only features belonging to the different categories. The result shows that category-agnostic repulsion is more effective.

5.3. Experiments with More Prompts

Prompts	N-Caltech101
"image of a [CLASS]."	**82.46**
"gray image of a [CLASS]."	81.30
"photo of a [CLASS]."	80.64
"reconstructed image of a [CLASS]."	81.62
"clean image of a [CLASS]."	81.04

Table 6. The results using various text prompts.

We conduct experiments with five prompt designs. As shown in Table 6, the simplest prompt shows the best performance.

6. Additional Qualitative Results

We additionally visualize the qualitative results on N-Caltech101 [5] and N-ImageNet (Mini) [4] in Fig. 2 and Fig. 3, respectively. Compared to other methods, the proposed method reconstructs images with more fine details.

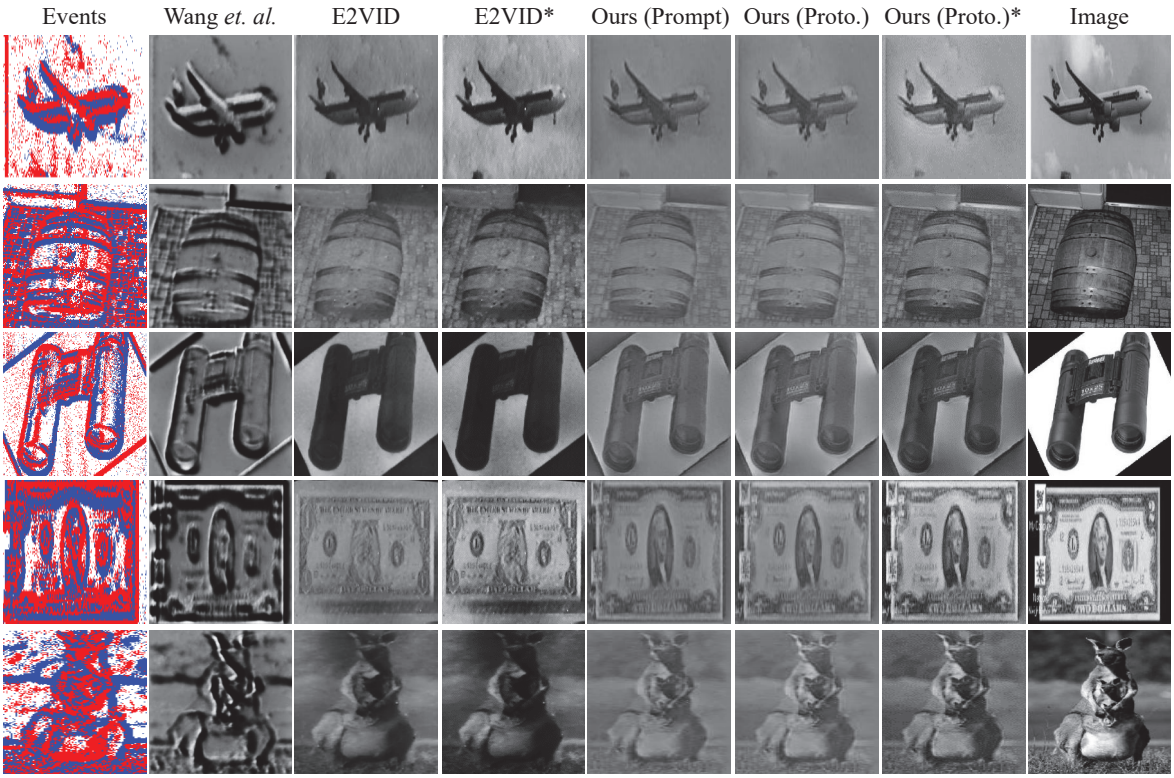

Figure 2. Additional qualitative results on N-Caltech101.

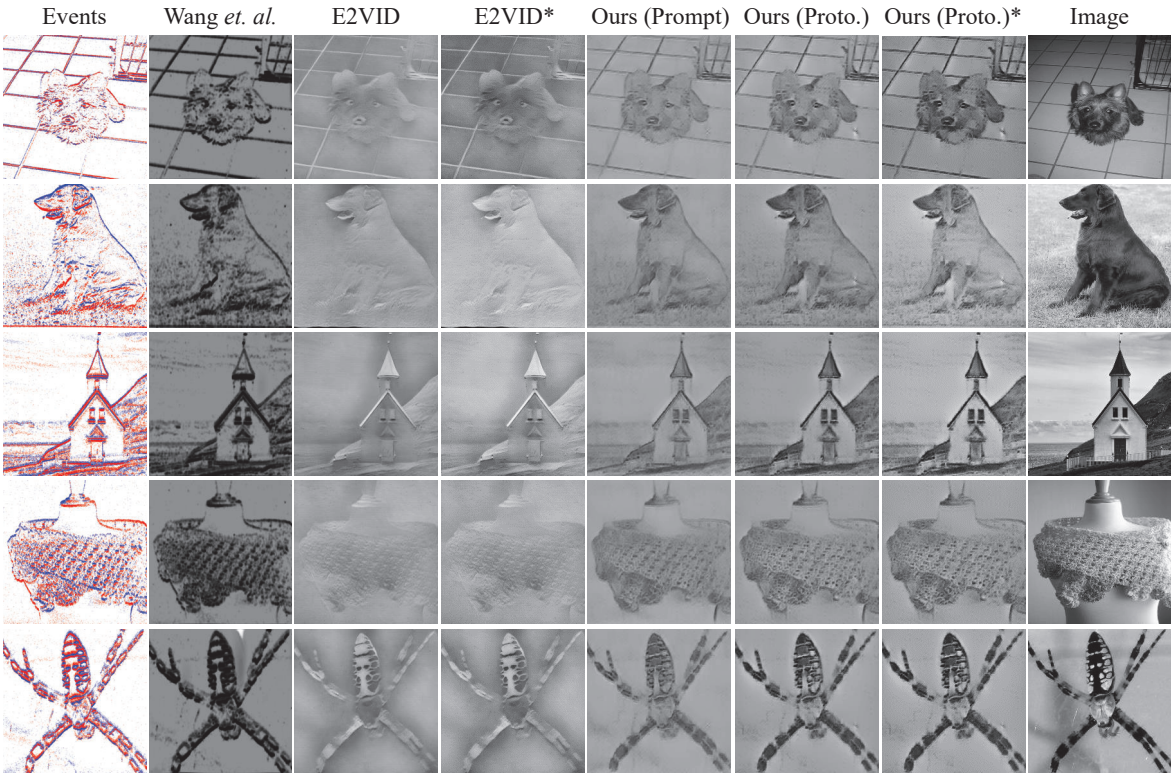

Figure 3. Additional qualitative results on N-ImageNet (Mini).

References

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 213–229. Springer, 2020. 1

[2] Daniel Gehrig, Antonio Loquercio, Konstantinos G Derpanis, and Davide Scaramuzza. End-to-end learning of representations for asynchronous event-based data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5633–5643, 2019. 1

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[4] Junho Kim, Jaehyeok Bae, Gangin Park, Dongsu Zhang, and Young Min Kim. N-imagenet: Towards robust, fine-grained object recognition with event cameras. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2146–2156, 2021. 2

[5] Garrick Orchard, Ajinkya Jayawant, Gregory K Cohen, and Nitish Thakor. Converting static image datasets to spiking neuromorphic datasets using saccades. *Frontiers in neuroscience*, 9:437, 2015. 2

[6] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 1

[7] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 1

[8] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 1

[9] Yang You, Jing Li, Sashank Reddi, Jonathan Hseu, Sanjiv Kumar, Srinadh Bhojanapalli, Xiaodan Song, James Demmel, Kurt Keutzer, and Cho-Jui Hsieh. Large batch optimization for deep learning: Training bert in 76 minutes. *arXiv preprint arXiv:1904.00962*, 2019. 2