Technical report · IF-TR-2026-06 · version 1.1

Masked pretraining for a from-scratch Vision Transformer on Indian road scenes: BluViT-S on IDD

Sherin Joseph Roy 18 November 2026 BluViT commit 59b8054 (no release tag) P1 · Perception

Funded entirely by BluRabbit Mobility Private Limited.

Abstract

BluViT is an image encoder for camera-based driving perception, trained from zero on Indian road imagery with no outside weights. This report tests the claim in its README that masked-autoencoder (MAE) pretraining is what makes a from-scratch Vision Transformer work with about ten thousand labelled images. BluViT-S, a ViT-S/16 with 2D rotary position embeddings and register tokens, was pretrained for 400 epochs on 42,320 unlabelled frames of the public India Driving Dataset (IDD) and fine-tuned for 26-class semantic segmentation on IDD's 14,027 training images. Scored on the official val split at the final epoch, it reached a mean IoU of 64.48 percent, against 37.31 for the same encoder trained from random initialisation, so the gate set before the runs passed. A control arm that removes a handicap of the fine-tune recipe reached 48.12, and a DeepLabV3 ResNet-50 trained from scratch reached 60.17. Each arm ran once, on one workstation GPU; nothing is measured on the lab's own footage, on a vehicle or on the target embedded device. Version 1.1 adds that a later ResNet-50, trained from scratch with a better recipe (IF-TR-2026-07), comes within one point of the pretrained ViT.

1 Problem

BluViT is meant to be the shared image encoder of a camera-only driving stack. Its README sets three conditions: it is trained from zero on Indian driving imagery, with no ImageNet weights and without the timm library; one backbone must serve wide and narrow cameras at different resolutions; and the first supervised task is semantic segmentation on the India Driving Dataset (IDD) [idd].

A Vision Transformer [vit] trained from scratch on a small labelled set is expected to be weak. The README's design rests on one claim: that masked-autoencoder pretraining [mae] on unlabelled frames is what makes a from-scratch ViT work on about ten thousand labelled images. It turns that claim into a gate, written before any run: an MAE-initialised model must beat a supervised-only baseline trained for the same 40 epochs, on mean IoU and on each of four classes reported separately in every run (person, rider, motorcycle and autorickshaw). If it does not, the README's stated remedy is more unlabelled frames, not more epochs.

This report describes the encoder, the study that tested the gate, and its results.

2 System design

Encoder

BluViT-S is a ViT-S/16: patch size 16, width 384, depth 12, six heads, at an input of 512x896, with `21.61 M` parameters in the encoder. The README gives a reason for each departure from the original ViT:

- **2D axial rotary position embeddings** [rope] [rope-vit] instead of an absolute position table, so one backbone serves cameras of different fields of view and resolutions, and inference size may differ from training size.
- **Four register tokens** [registers], because without them a small ViT's dense features carry high-norm artefact tokens in the background, which hurts a segmentation head.
- **QK-norm** [qknorm], **LayerScale** [layerscale] and **stochastic depth** [droppath], as inexpensive stabilisers for training from scratch in bf16.
- **A linear patch projection** instead of a convolutional stem, so MAE masking cannot leak information between patches. The README notes that a convolutional stem converges faster when training is supervised only, and chooses MAE compatibility.
- **A camera-id embedding** with six slots, unused in this study, for later multi-view training on the lab's rig without changing the backbone.

Pretraining and segmentation

The masked autoencoder masks 75 percent of the patches and reconstructs them with a 192-wide, four-block decoder and a normalised-pixel loss; the decoder is discarded after pretraining. The segmentation model taps four encoder blocks (4, 6, 8 and 12) into an all-MLP head in the style of SegFormer [segformer] with deep supervision, and is fine-tuned with layer-wise learning-rate decay [beit], EMA weights and class-weighted cross-entropy over IDD's 26 level-3 classes.

Labels

BluViT's own tool writes IDD's level-3 label masks from the polygon annotations. It was compared with the official AutoNUE `createLabels.py` [autonue] on all 16,063 Part I and II train and val annotations, with `0 px` differing. Two fixes were needed first: the fallback background maps to level-3 id 25, shared with sky, not to the ignore label; and polygons are rasterised from their raw floating-point vertices in file order, as the official script does. The training runs below use the official masks.

3 Method

Data

All data is public IDD. Supervised training uses the merged Part I and II segmentation train split (14,027 images); evaluation uses the official val split (2,036 images), held as a locked set whose files are re-hashed before every run. The test split, which has no public labels, is used for nothing.

The unlabelled MAE corpus has 42,320 frames from 479 sequences: frames of the IDD temporal release taken every fifth frame with near-duplicates removed, frames of the multimodal release's train and val drives, and the segmentation train images. Code asserts before pretraining that it contains no segmentation val or test image, no frame within a small perceptual-hash distance of one, no multimodal test frame and no lidar data; every temporal sequence that contains a val or test keyframe is excluded whole. The study's data report calls the corpus small for MAE.

Arms

Five arms were trained for 40 epochs each, with batch 8, 512x896 crops (scale 0.5 to 2.0, flip, colour jitter 0.3), bf16, EMA 0.999 and the same class weights and auxiliary loss. Every arm was scored on the val split after every epoch, on the full image at width 1280, with prediction and label resized to 1280x720 (nearest neighbour) as the IDD protocol specifies. The reported value is the **final epoch**: choosing the best epoch would select on the set being reported.

1. **Supervised only**: BluViT-S from random initialisation, with the fine-tune recipe (AdamW, lr 1e-4, cosine schedule, five warm-up epochs, layer-wise lr decay 0.75).
2. **MAE init, epoch 150**: the same command, with the encoder loaded from the MAE checkpoint at epoch 150.
3. **MAE init, epoch 400**: the same, from the end of pretraining. This and arm 1 are the gate's pair.
4. **Control**: arm 1 with layer-wise lr decay 1.0 (off), added at the author's decision after the gate's runs.
5. **CNN reference**: torchvision's DeepLabV3 ResNet-50 [deeplabv3] [resnet] with no pretrained weights, trained with SGD (lr 0.01, momentum 0.9, poly schedule) on the same data and scored the same way.

The three ViT arms run one command that differs only in the initialisation; the study's runner refuses to start if the training script, data loader or config differs by hash from the reference run. The seed rule, fixed in advance, was to add two more seeds to the gate's pair only if they ended within one mIoU point of each other.

1

> 1. Layer-wise lr decay 0.75 trains the patch embedding and the first block of a randomly initialised encoder at a few percent of the base learning rate. That suits fine-tuning a pretrained encoder and handicaps the supervised arm. The study kept the recipe identical across the gate's arms, recorded this as a known confound and ran arm 4 to measure it.

Pretraining

MAE ran for 400 epochs (batch 32, AdamW lr 1.5e-4, 20 warm-up epochs) on one NVIDIA GeForce RTX 5080, reading a 1280-wide resized cache of the corpus. Its loss on masked patches fell from `0.9053` after the first epoch to `0.3127` at epoch 50, `0.2988` at 150 and `0.2888` at 400, in `29.7 h` at a median of `179.6 img/s`.

4 Validation

Mean IoU

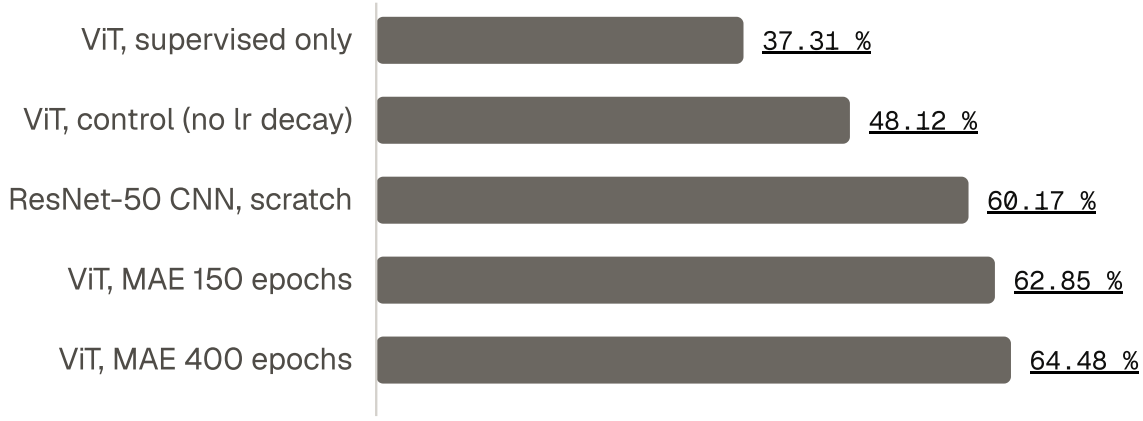


Figure 1. Mean IoU on the IDD val split (level 3, 26 classes), final epoch of 40, one seed per arm; the author's runs.

The supervised-only ViT reached `37.31 %`. Initialised from MAE, the same encoder and recipe reached `62.85 %` from the epoch-150 checkpoint and `64.48 %` from epoch 400, a gain of `27.17 points` for the gate's pair. The gap is far beyond the one-point seed rule, so one seed stands. The from-scratch CNN reached `60.17 %`.

The control removes most of the confound. Without layer-wise decay the supervised ViT reached `48.12 %`, and MAE initialisation is `16.36 points` above it. By the study's own reckoning, about two fifths of the gate's gap is the recipe, and the gain over the control is the figure to quote for what pretraining contributes.

Quantity	Value	Conditions	Basis
Mean IoU on IDD val, BluViT-S, supervised only (random init)	37.31 %	random initialisation, layer-wise lr decay 0.75 (the MAE arms' recipe); IDD level 3 (26 classes), official val split (2,036 images), final epoch of 40, EMA weights, scored at 1280x720 (IDD protocol); trained on 14,027 IDD train images, 512x896 crops, batch 8, bf16; one seed	Reproduced 2026-11-10
Mean IoU on IDD val, BluViT-S, supervised only, layer decay 1.0 (control)	48.12 %	control arm: random initialisation, layer-wise lr decay off (1.0); IDD level 3 (26 classes), official val split (2,036 images), final epoch of 40, EMA weights, scored at 1280x720 (IDD protocol); trained on 14,027 IDD train images, 512x896 crops, batch 8, bf16; one seed	Reproduced 2026-11-16
Mean IoU on IDD val, DeepLabV3 ResNet-50, from scratch (reference)	60.17 %	reference CNN: torchvision DeepLabV3 ResNet-50, no pretrained weights, SGD with a poly schedule; IDD level 3 (26 classes), official val split (2,036 images), final epoch of 40, EMA weights, scored at 1280x720 (IDD protocol); trained on 14,027 IDD train images, 512x896 crops, batch 8, bf16; one seed	Reproduced 2026-11-10
Mean IoU on IDD val, BluViT-S, MAE init (epoch 150)	62.85 %	encoder from the MAE checkpoint at epoch 150 of 400 (unlabelled IDD frames only); IDD level 3 (26 classes), official val split (2,036 images), final epoch of 40, EMA weights, scored at 1280x720 (IDD protocol); trained on 14,027 IDD train images, 512x896 crops, batch 8, bf16; one seed	Reproduced 2026-11-15
Mean IoU on IDD val, BluViT-S, MAE init (epoch 400)	64.48 %	encoder from the MAE checkpoint at epoch 400 (unlabelled IDD frames only); IDD level 3 (26 classes), official val split (2,036 images), final epoch of 40, EMA weights, scored at 1280x720 (IDD protocol); trained on 14,027 IDD train images, 512x896 crops, batch 8, bf16; one seed	Reproduced 2026-11-15
Gain in mean IoU from MAE initialisation (epoch 400) over supervised-only	27.17 points	gate pair: MAE init (epoch 400) minus supervised-only, both with layer-wise lr decay 0.75; verdict PASS (must win on mean IoU and on person, rider, motorcycle and autorickshaw); IDD level 3 (26 classes), official val split (2,036 images), final epoch of 40, EMA weights, scored at 1280x720 (IDD protocol); trained on 14,027 IDD train images, 512x896 crops, batch 8, bf16; one seed	Reproduced 2026-11-15
Gain in mean IoU from MAE initialisation (epoch 400) over the supervised control (layer decay 1.0)	16.36 points	MAE init (epoch 400) minus the control (supervised-only, layer-wise lr decay off); IDD level 3 (26 classes), official val split (2,036 images), final epoch of 40, EMA weights, scored at 1280x720 (IDD protocol); trained on 14,027 IDD train images, 512x896 crops, batch 8, bf16; one seed	Reproduced 2026-11-16

Table 1. Mean IoU per arm and the two differences the study states, IDD val, final epoch; run by the author. “Reproduced” rows were run by the author; “Reported” rows are quoted from the repository at the commit shown.

The four separately reported classes

The gate also required a gain on person, rider, motorcycle and autorickshaw. The MAE-initialised encoder beat the supervised arm on all four, and the control on all four as well. Against the from-scratch CNN it is ahead on each of the four, by smaller margins.

2

2. Added in version 1.1 (2026-11-18). A later run in the same study, IF-TR-2026-07, trained ResNet-50 from scratch with an FPN head and AdamW. It reached 63.57 %, within one point of the MAE-initialised ViT, and is ahead of it on person (66.99 %), rider (72.05 %) and motorcycle (75.59 %), behind it on autorickshaw (83.48 %). A from-scratch convolutional encoder, BluConv-T, reached 65.89 %. The ViT's lead over the CNN in this report holds for the DeepLabV3 recipe used here, not for convolutional networks in general.

Quantity	Value	Conditions	Basis
IoU, person, BluViT-S, MAE init (epoch 400)	64.15 %	encoder from the MAE checkpoint at epoch 400 (unlabelled IDD frames only); IDD level 3 (26 classes), official val split (2,036 images), final epoch of 40, EMA weights, scored at 1280x720 (IDD protocol); trained on 14,027 IDD train images, 512x896 crops, batch 8, bf16; one seed	Reproduced 2026-11-15
IoU, rider, BluViT-S, MAE init (epoch 400)	68.92 %	encoder from the MAE checkpoint at epoch 400 (unlabelled IDD frames only); IDD level 3 (26 classes), official val split (2,036 images), final epoch of 40, EMA weights, scored at 1280x720 (IDD protocol); trained on 14,027 IDD train images, 512x896 crops, batch 8, bf16; one seed	Reproduced 2026-11-15
IoU, motorcycle, BluViT-S, MAE init (epoch 400)	73.4 %	encoder from the MAE checkpoint at epoch 400 (unlabelled IDD frames only); IDD level 3 (26 classes), official val split (2,036 images), final epoch of 40, EMA weights, scored at 1280x720 (IDD protocol); trained on 14,027 IDD train images, 512x896 crops, batch 8, bf16; one seed	Reproduced 2026-11-15
IoU, autorickshaw, BluViT-S, MAE init (epoch 400)	84.92 %	encoder from the MAE checkpoint at epoch 400 (unlabelled IDD frames only); IDD level 3 (26 classes), official val split (2,036 images), final epoch of 40, EMA weights, scored at 1280x720 (IDD protocol); trained on 14,027 IDD train images, 512x896 crops, batch 8, bf16; one seed	Reproduced 2026-11-15
IoU, person, DeepLabV3 ResNet-50, from scratch (reference)	62.87 %	reference CNN: torchvision DeepLabV3 ResNet-50, no pretrained weights, SGD with a poly schedule; IDD level 3 (26 classes), official val split (2,036 images), final epoch of 40, EMA weights, scored at 1280x720 (IDD protocol); trained on 14,027 IDD train images, 512x896 crops, batch 8, bf16; one seed	Reproduced 2026-11-10
IoU, rider, DeepLabV3 ResNet-50, from scratch (reference)	66.57 %	reference CNN: torchvision DeepLabV3 ResNet-50, no pretrained weights, SGD with a poly schedule; IDD level 3 (26 classes), official val split (2,036 images), final epoch of 40, EMA weights, scored at 1280x720 (IDD protocol); trained on 14,027 IDD train images, 512x896 crops, batch 8, bf16; one seed	Reproduced 2026-11-10
IoU, motorcycle, DeepLabV3 ResNet-50, from scratch (reference)	71.27 %	reference CNN: torchvision DeepLabV3 ResNet-50, no pretrained weights, SGD with a poly schedule; IDD level 3 (26 classes), official val split (2,036 images), final epoch of 40, EMA weights, scored at 1280x720 (IDD protocol); trained on 14,027 IDD train images, 512x896 crops, batch 8, bf16; one seed	Reproduced 2026-11-10
IoU, autorickshaw, DeepLabV3 ResNet-50, from scratch (reference)	82.39 %	reference CNN: torchvision DeepLabV3 ResNet-50, no pretrained weights, SGD with a poly schedule; IDD level 3 (26 classes), official val split (2,036 images), final epoch of 40, EMA weights, scored at 1280x720 (IDD protocol); trained on 14,027 IDD train images, 512x896 crops, batch 8, bf16; one seed	Reproduced 2026-11-10
IoU, person, BluViT-S, supervised only, layer decay 1.0 (control)	43.6 %	control arm: random initialisation, layer-wise lr decay off (1.0); IDD level 3 (26 classes), official val split (2,036 images), final epoch of 40, EMA weights, scored at 1280x720 (IDD protocol);	Reproduced 2026-11-16

Quantity	Value	Conditions	Basis
		trained on 14,027 IDD train images, 512x896 crops, batch 8, bf16; one seed	
IoU, rider, BluViT-S, supervised only, layer decay 1.0 (control)	49.8 %	control arm: random initialisation, layer-wise lr decay off (1.0); IDD level 3 (26 classes), official val split (2,036 images), final epoch of 40, EMA weights, scored at 1280x720 (IDD protocol); trained on 14,027 IDD train images, 512x896 crops, batch 8, bf16; one seed	Reproduced 2026-11-16
IoU, motorcycle, BluViT-S, supervised only, layer decay 1.0 (control)	56.65 %	control arm: random initialisation, layer-wise lr decay off (1.0); IDD level 3 (26 classes), official val split (2,036 images), final epoch of 40, EMA weights, scored at 1280x720 (IDD protocol); trained on 14,027 IDD train images, 512x896 crops, batch 8, bf16; one seed	Reproduced 2026-11-16
IoU, autorickshaw, BluViT-S, supervised only, layer decay 1.0 (control)	70.18 %	control arm: random initialisation, layer-wise lr decay off (1.0); IDD level 3 (26 classes), official val split (2,036 images), final epoch of 40, EMA weights, scored at 1280x720 (IDD protocol); trained on 14,027 IDD train images, 512x896 crops, batch 8, bf16; one seed	Reproduced 2026-11-16
IoU, person, BluViT-S, supervised only (random init)	27.25 %	random initialisation, layer-wise lr decay 0.75 (the MAE arms' recipe); IDD level 3 (26 classes), official val split (2,036 images), final epoch of 40, EMA weights, scored at 1280x720 (IDD protocol); trained on 14,027 IDD train images, 512x896 crops, batch 8, bf16; one seed	Reproduced 2026-11-10
IoU, rider, BluViT-S, supervised only (random init)	35.77 %	random initialisation, layer-wise lr decay 0.75 (the MAE arms' recipe); IDD level 3 (26 classes), official val split (2,036 images), final epoch of 40, EMA weights, scored at 1280x720 (IDD protocol); trained on 14,027 IDD train images, 512x896 crops, batch 8, bf16; one seed	Reproduced 2026-11-10
IoU, motorcycle, BluViT-S, supervised only (random init)	43.65 %	random initialisation, layer-wise lr decay 0.75 (the MAE arms' recipe); IDD level 3 (26 classes), official val split (2,036 images), final epoch of 40, EMA weights, scored at 1280x720 (IDD protocol); trained on 14,027 IDD train images, 512x896 crops, batch 8, bf16; one seed	Reproduced 2026-11-10
IoU, autorickshaw, BluViT-S, supervised only (random init)	51.43 %	random initialisation, layer-wise lr decay 0.75 (the MAE arms' recipe); IDD level 3 (26 classes), official val split (2,036 images), final epoch of 40, EMA weights, scored at 1280x720 (IDD protocol); trained on 14,027 IDD train images, 512x896 crops, batch 8, bf16; one seed	Reproduced 2026-11-10

Table 2. Per-class IoU on the IDD val split, final epoch, one seed per arm: MAE init (epoch 400), the CNN reference, the control and the supervised-only arm; run by the author. “Reproduced” rows were run by the author; “Reported” rows are quoted from the repository at the commit shown.

Training without pretraining

Before pretraining was in the picture, the result was negative for the ViT: from random initialisation it reached `37.31 %` against the CNN’s `60.17 %`. The study had expected the CNN to be the weaker floor and rewrote its baseline statement once the

numbers were in: each from-scratch number ranks a recipe on 14,027 images in 40 epochs, not an architecture's ceiling. Neither recipe was tuned for this data.

Cost

Quantity	Value	Conditions	Basis
MAE pretraining wall clock, 400 epochs	29.7 h	masked autoencoder, 75 percent masking, normalised-pixel loss; 512x896, batch 32, bf16; 42,320 unlabelled IDD frames with nothing from val or test (asserted in code); summed over the run's resumed segments; loader reads a 1280-wide resized cache of the corpus	Reproduced 2026-11-15
MAE pretraining throughput, median	179.6 img/s	masked autoencoder, 75 percent masking, normalised-pixel loss; 512x896, batch 32, bf16; 42,320 unlabelled IDD frames with nothing from val or test (asserted in code); end to end including data loading	Reproduced 2026-11-15
Fine-tune wall clock, 40 epochs, BluViT-S, MAE init (epoch 400)	4.45 h	40 epochs including a val pass each epoch; one seed	Reproduced 2026-11-15
Fine-tune wall clock, 40 epochs, DeepLabV3 ResNet-50, from scratch	5.62 h	40 epochs including a val pass each epoch; one seed; the CNN shared the CPU with a cache build, so its time is slightly pessimistic (study report)	Reproduced 2026-11-10
Segmentation inference throughput on the val set	46.8 img/s	BluViT-S with the segmentation head, batch 1, full image at width 1280, including argmax and the resize to 1280x720; timed over 300 of the 2,036 val images; includes data loading	Reproduced 2026-11-09
BluViT-S encoder parameters	21.61 M	ViT-S/16: width 384, depth 12, 6 heads, 4 register tokens; the segmentation model with its head has 25.26 M	Reproduced 2026-11-15

Table 3. Training and inference cost on one NVIDIA GeForce RTX 5080 workstation GPU, and encoder size; measured by the author. "Reproduced" rows were run by the author; "Reported" rows are quoted from the repository at the commit shown.

A 40-epoch fine-tune took `4.45 h` for the ViT and `5.62 h` for the CNN, which shared the CPU with a cache build while it trained. Inference on the val set, at batch 1 and including data loading, ran at `46.8 img/s` on the same GPU. That is a workstation figure; it says nothing about an embedded device.

5 Limitations

- **One seed per arm.** The seed rule was applied as written, so the size of each gap is single-run. The step from MAE epoch 150 to 400 is small, and the study itself calls its size indicative only.
- **One dataset, one split.** Every number is on IDD's val split. Nothing is measured on the lab's own footage, on another dataset or on IDD's test split, and the encoder has not been tested on a camera of a different field of view, which its position embedding was chosen for.
- **The recipe was not tuned.** The ViT recipe was fixed before the runs and the CNN used a standard SGD recipe; both are floors. The CNN had no pretraining, so the MAE-initialised ViT being ahead of it does not rank the architectures; a better-trained ResNet-50 in IF-TR-2026-07 comes within one point of it without pretraining.

- **The gate's baseline was handicapped.** The gated supervised arm trained its early layers at a small fraction of the learning rate. The verdict stands as pre-registered, but the control arm, added afterwards, is the fairer comparison.
- **Small corpus.** The unlabelled corpus is about 2.6 hours of driving plus single images; the study notes that more frames would matter more than more epochs, and the effect of corpus size was not measured.
- **Interruptions.** Pretraining was resumed from checkpoints after memory exhaustion on the shared workstation, with its loader worker count reduced from 12 to 8 at epoch 70. Resuming re-seeds the sample order and changing the worker count changes which augmentations each sample draws; the augmentation itself is unchanged.
- **No device measurement.** The README targets an NVIDIA Jetson AGX Orin 64 GB, where neither training nor inference has been measured, and the ONNX and TensorRT export path was not exercised in this study.
- **The repository is not public.** The commands and commits below can be checked by the author but not by readers.

6 Reproducibility

The BluViT repository is local, with no remote and no release tag. Results come from three commits on its main branch: 1156461 (supervised-only arm), 4261989 (CNN reference) and 59b8054 (both MAE fine-tunes, the control and pretraining). Across all five arms the training script, data loader and config are identical by SHA-256, recorded per run.

- Pretraining: `python -m bluvit.train.pretrain_mae --config configs/mae_vit_s.yaml`, with the corpus manifest set in the config.
- Fine-tune (arms 1 to 4): `python -m bluvit.train.finetune_seg --config configs/seg_vit_s_idd.yaml train.epochs=40 seg.class_weights=configs/idd_class_weights.json data.workers=8 train.seed=1337` with `train.init_from=null`, the epoch-150 or epoch-400 MAE checkpoint, and for the control `train.layer_decay=1.0`.
- CNN reference: the study's `train_cnn_baseline.py`, with the recipe stated in the Method section.
- Masks: `python tools/idd_prepare.py masks` with `--verify-against` pointing at AutoNUE's `anue_labels.py`.

Hardware and software: one NVIDIA GeForce RTX 5080 (15.5 GB usable), an AMD Ryzen 9 9900X with about 30 GB RAM, Linux 7.0.0-42 (Ubuntu), PyTorch 2.14.1 with CUDA 13.0, bf16 throughout. The IDD releases are available from IIIT Hyderabad under their own terms; no IDD image is reproduced in this report.

References

1. G. Varma, A. Subramanian, A. Namboodiri, M. Chandraker, C. V. Jawahar. IDD: A dataset for exploring problems of autonomous navigation in unconstrained environments. WACV 2019. doi:10.1109/WACV.2019.00190
2. AutoNUE. public-code: dataset preparation and evaluation helpers for IDD, including createLabels.py and anue_labels.py. Source code. github.com/AutoNUE/public-code
3. A. Dosovitskiy et al. An image is worth 16x16 words: Transformers for image recognition at scale. ICLR 2021. arxiv.org/abs/2010.11929
4. K. He, X. Chen, S. Xie, Y. Li, P. Dollár, R. Girshick. Masked autoencoders are scalable vision learners. CVPR 2022. arxiv.org/abs/2111.06377
5. T. Darcet, M. Oquab, J. Mairal, P. Bojanowski. Vision Transformers need registers. ICLR 2024. arxiv.org/abs/2309.16588
6. J. Su, Y. Lu, S. Pan, A. Murtadha, B. Wen, Y. Liu. RoFormer: Enhanced Transformer with rotary position embedding. arXiv:2104.09864, 2021. arxiv.org/abs/2104.09864
7. B. Heo, S. Park, D. Han, S. Yun. Rotary position embedding for Vision Transformer. ECCV 2024. arxiv.org/abs/2403.13298
8. M. Dehghani et al. Scaling Vision Transformers to 22 billion parameters. ICML 2023. arxiv.org/abs/2302.05442
9. H. Touvron, M. Cord, A. Sablayrolles, G. Synnaeve, H. Jégou. Going deeper with image transformers. ICCV 2021. arxiv.org/abs/2103.17239

10. G. Huang, Y. Sun, Z. Liu, D. Sedra, K. Q. Weinberger. Deep networks with stochastic depth. ECCV 2016. arxiv.org/abs/1603.09382

11. E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, P. Luo. SegFormer: Simple and efficient design for semantic segmentation with Transformers. NeurIPS 2021. arxiv.org/abs/2105.15203

12. H. Bao, L. Dong, S. Piao, F. Wei. BEiT: BERT pre-training of image Transformers. ICLR 2022. arxiv.org/abs/2106.08254

13. L.-C. Chen, G. Papandreou, F. Schroff, H. Adam. Rethinking atrous convolution for semantic image segmentation. arXiv:1706.05587, 2017. arxiv.org/abs/1706.05587

14. K. He, X. Zhang, S. Ren, J. Sun. Deep residual learning for image recognition. CVPR 2016. doi:10.1109/CVPR.2016.90

Evidence

Quantity	Value	Conditions	Basis	Source and reproduction
Mean IoU on IDD val, BluViT-S, supervised only (random init)	37.31 %	random initialisation, layer-wise lr decay 0.75 (the MAE arms' recipe); IDD level 3 (26 classes), official val split (2,036 images), final epoch of 40, EMA weights, scored at 1280x720 (IDD protocol); trained on 14,027 IDD train images, 512x896 crops, batch 8, bf16; one seed	Reproduced 2026-11-10	BASELINE.md (study report), arm table; runs/seg_supervised_only/metrics.jsonl commit 1156461 (bluvit main; local repository, no release tag) · NVIDIA GeForce RTX 5080 (15.5 GB), AMD Ryzen 9 9900X, about 30 GB RAM, Linux 7.0.0-42 (Ubuntu), PyTorch 2.14.1 with CUDA 13.0, bf16 `Local repository, not public. python -m bluvit.train.finetune_seg --config configs/seg_vit_s_idd.yaml train.epochs=40 seg.class_weights=configs/idd_class_weights.json data.workers=8 train.init_from=null`
Mean IoU on IDD val, BluViT-S, supervised only, layer decay 1.0 (control)	48.12 %	control arm: random initialisation, layer-wise lr decay off (1.0); IDD level 3 (26 classes), official val split (2,036 images), final epoch of 40, EMA weights, scored at 1280x720 (IDD protocol); trained on 14,027 IDD train images, 512x896 crops, batch 8, bf16; one seed	Reproduced 2026-11-16	BASELINE.md (study report), arm table; runs/seg_supervised_only_ld1/metrics.jsonl commit 59b8054 (bluvit main; local repository, no release tag) · NVIDIA GeForce RTX 5080 (15.5 GB), AMD Ryzen 9 9900X, about 30 GB RAM, Linux 7.0.0-42 (Ubuntu), PyTorch 2.14.1 with CUDA 13.0, bf16 `Local repository, not public. python -m bluvit.train.finetune_seg --config configs/seg_vit_s_idd.yaml train.epochs=40 seg.class_weights=configs/idd_class_weights.json data.workers=8 train.init_from=null train.seed=1337 train.layer_decay=1.0`
Mean IoU on IDD val, BluViT-S, MAE init (epoch 150)	62.85 %	encoder from the MAE checkpoint at epoch 150 of 400 (unlabelled IDD frames only); IDD level 3 (26 classes), official val split (2,036 images), final epoch of 40, EMA weights, scored at 1280x720 (IDD protocol); trained on 14,027 IDD train images, 512x896 crops, batch 8, bf16; one seed	Reproduced 2026-11-15	BASELINE.md (study report), arm table; runs/seg_mae_init_ep150/metrics.jsonl commit 59b8054 (bluvit main; local repository, no release tag) · NVIDIA GeForce RTX 5080 (15.5 GB), AMD Ryzen 9 9900X, about 30 GB RAM, Linux 7.0.0-42 (Ubuntu), PyTorch 2.14.1 with CUDA 13.0, bf16 `Local repository, not public. python -m bluvit.train.finetune_seg --config configs/seg_vit_s_idd.yaml train.epochs=40 seg.class_weights=configs/idd_class_weights.json data.workers=8 train.init_from=runs/mae_vit_s/mae_ep150.pt train.seed=1337`

Quantity	Value	Conditions	Basis	Source and reproduction
Mean IoU on IDD val, BluViT-S, MAE init (epoch 400)	64.48 %	encoder from the MAE checkpoint at epoch 400 (unlabelled IDD frames only); IDD level 3 (26 classes), official val split (2,036 images), final epoch of 40, EMA weights, scored at 1280x720 (IDD protocol); trained on 14,027 IDD train images, 512x896 crops, batch 8, bf16; one seed	Reproduced 2026-11-15	BASELINE.md (study report), arm table; runs/seg_mae_init_ep400/metrics.jsonl commit 59b8054 (bluvit main; local repository, no release tag) · NVIDIA GeForce RTX 5080 (15.5 GB), AMD Ryzen 9 9900X, about 30 GB RAM, Linux 7.0.0-42 (Ubuntu), PyTorch 2.14.1 with CUDA 13.0, bf16 `Local repository, not public. python -m bluvit.train.finetune_seg --config configs/seg_vit_s_idd.yaml train.epochs=40 seg.class_weights=configs/idd_class_weights.json data.workers=8 train.init_from=runs/mae_vit_s/mae_last.pt train.seed=1337`
Mean IoU on IDD val, DeepLabV3 ResNet-50, from scratch (reference)	60.17 %	reference CNN: torchvision DeepLabV3 ResNet-50, no pretrained weights, SGD with a poly schedule; IDD level 3 (26 classes), official val split (2,036 images), final epoch of 40, EMA weights, scored at 1280x720 (IDD protocol); trained on 14,027 IDD train images, 512x896 crops, batch 8, bf16; one seed	Reproduced 2026-11-10	BASELINE.md (study report), arm table; runs/seg_cnn_deeplabv3_r50/metrics.jsonl commit 4261989 (bluvit main; local repository, no release tag) · NVIDIA GeForce RTX 5080 (15.5 GB), AMD Ryzen 9 9900X, about 30 GB RAM, Linux 7.0.0-42 (Ubuntu), PyTorch 2.14.1 with CUDA 13.0, bf16 `Local repository, not public. python scripts/train_cnn_baseline.py (study scripts, alongside bluvit)`
Gain in mean IoU from MAE initialisation (epoch 400) over supervised-only	27.17 points	gate pair: MAE init (epoch 400) minus supervised-only, both with layer-wise lr decay 0.75; verdict PASS (must win on mean IoU and on person, rider, motorcycle and autorickshaw); IDD level 3 (26 classes), official val split (2,036 images), final epoch of 40, EMA weights, scored at 1280x720 (IDD protocol); trained on	Reproduced 2026-11-15	BASELINE.md (study report), Gate section commit 59b8054 (bluvit main; local repository, no release tag) · NVIDIA GeForce RTX 5080 (15.5 GB), AMD Ryzen 9 9900X, about 30 GB RAM, Linux 7.0.0-42 (Ubuntu), PyTorch 2.14.1 with CUDA 13.0, bf16 `Local repository, not public. see the two arms' commands`

Quantity	Value	Conditions	Basis	Source and reproduction
		14,027 IDD train images, 512x896 crops, batch 8, bf16; one seed		
Gain in mean IoU from MAE initialisation (epoch 400) over the supervised control (layer decay 1.0)	`16.36 points`	MAE init (epoch 400) minus the control (supervised-only, layer-wise lr decay off); IDD level 3 (26 classes), official val split (2,036 images), final epoch of 40, EMA weights, scored at 1280x720 (IDD protocol); trained on 14,027 IDD train images, 512x896 crops, batch 8, bf16; one seed	Reproduced 2026-11-16	BASELINE.md (study report), Control section commit 59b8054 (bluvit main; local repository, no release tag) · NVIDIA GeForce RTX 5080 (15.5 GB), AMD Ryzen 9 9900X, about 30 GB RAM, Linux 7.0.0-42 (Ubuntu), PyTorch 2.14.1 with CUDA 13.0, bf16 `Local repository, not public. see the two arms' commands`
IoU, person, BluViT-S, supervised only (random init)	`27.25 %`	random initialisation, layer-wise lr decay 0.75 (the MAE arms' recipe); IDD level 3 (26 classes), official val split (2,036 images), final epoch of 40, EMA weights, scored at 1280x720 (IDD protocol); trained on 14,027 IDD train images, 512x896 crops, batch 8, bf16; one seed	Reproduced 2026-11-10	BASELINE.md (study report), per-class table; runs/seg_supervised_only/metrics.jsonl commit 1156461 (bluvit main; local repository, no release tag) · NVIDIA GeForce RTX 5080 (15.5 GB), AMD Ryzen 9 9900X, about 30 GB RAM, Linux 7.0.0-42 (Ubuntu), PyTorch 2.14.1 with CUDA 13.0, bf16 `Local repository, not public. python -m bluvit.train.finetune_seg --config configs/seg_vit_s_idd.yaml train.epochs=40 seg.class_weights=configs/idd_class_weights.json data.workers=8 train.init_from=null`

Quantity	Value	Conditions	Basis	Source and reproduction
IoU, rider, BluViT-S, supervised only (random init)	35.77 %	random initialisation, layer-wise lr decay 0.75 (the MAE arms' recipe); IDD level 3 (26 classes), official val split (2,036 images), final epoch of 40, EMA weights, scored at 1280x720 (IDD protocol); trained on 14,027 IDD train images, 512x896 crops, batch 8, bf16; one seed	Reproduced 2026-11-10	BASELINE.md (study report), per-class table; runs/seg_supervised_only/metrics.jsonl commit 1156461 (bluvit main; local repository, no release tag) · NVIDIA GeForce RTX 5080 (15.5 GB), AMD Ryzen 9 9900X, about 30 GB RAM, Linux 7.0.0-42 (Ubuntu), PyTorch 2.14.1 with CUDA 13.0, bf16 `Local repository, not public. python -m bluvit.train.finetune_seg --config configs/seg_vit_s_idd.yaml train.epochs=40 seg.class_weights=configs/idd_class_weights.json data.workers=8 train.init_from=null`
IoU, motorcycle, BluViT-S, supervised only (random init)	43.65 %	random initialisation, layer-wise lr decay 0.75 (the MAE arms' recipe); IDD level 3 (26 classes), official val split (2,036 images), final epoch of 40, EMA weights, scored at 1280x720 (IDD protocol); trained on 14,027 IDD train images, 512x896 crops, batch 8, bf16; one seed	Reproduced 2026-11-10	BASELINE.md (study report), per-class table; runs/seg_supervised_only/metrics.jsonl commit 1156461 (bluvit main; local repository, no release tag) · NVIDIA GeForce RTX 5080 (15.5 GB), AMD Ryzen 9 9900X, about 30 GB RAM, Linux 7.0.0-42 (Ubuntu), PyTorch 2.14.1 with CUDA 13.0, bf16 `Local repository, not public. python -m bluvit.train.finetune_seg --config configs/seg_vit_s_idd.yaml train.epochs=40 seg.class_weights=configs/idd_class_weights.json data.workers=8 train.init_from=null`
IoU, autorickshaw, BluViT-S, supervised only (random init)	51.43 %	random initialisation, layer-wise lr decay 0.75 (the MAE arms' recipe); IDD level 3 (26 classes), official val split (2,036 images), final epoch of 40, EMA weights, scored at 1280x720 (IDD protocol); trained on 14,027 IDD train images, 512x896 crops, batch 8, bf16; one seed	Reproduced 2026-11-10	BASELINE.md (study report), per-class table; runs/seg_supervised_only/metrics.jsonl commit 1156461 (bluvit main; local repository, no release tag) · NVIDIA GeForce RTX 5080 (15.5 GB), AMD Ryzen 9 9900X, about 30 GB RAM, Linux 7.0.0-42 (Ubuntu), PyTorch 2.14.1 with CUDA 13.0, bf16 `Local repository, not public. python -m bluvit.train.finetune_seg --config configs/seg_vit_s_idd.yaml train.epochs=40 seg.class_weights=configs/idd_class_weights.json data.workers=8 train.init_from=null`
IoU, person, BluViT-S, supervised only, layer decay 1.0 (control)	43.6 %	control arm: random initialisation, layer-wise lr decay off (1.0); IDD level 3 (26	Reproduced 2026-11-16	BASELINE.md (study report), per-class table; runs/seg_supervised_only_ld1/metrics.jsonl

Quantity	Value	Conditions	Basis	Source and reproduction
		classes), official val split (2,036 images), final epoch of 40, EMA weights, scored at 1280x720 (IDD protocol); trained on 14,027 IDD train images, 512x896 crops, batch 8, bf16; one seed		commit 59b8054 (bluvit main; local repository, no release tag) · NVIDIA GeForce RTX 5080 (15.5 GB), AMD Ryzen 9 9900X, about 30 GB RAM, Linux 7.0.0-42 (Ubuntu), PyTorch 2.14.1 with CUDA 13.0, bf16 `Local repository, not public. python -m bluvit.train.finetune_seg --config configs/seg_vit_s_idd.yaml train.epochs=40 seg.class_weights=configs/idd_class_weights.json data.workers=8 train.init_from=null train.seed=1337 train.layer_decay=1.0`
IoU, rider, BluViT-S, supervised only, layer decay 1.0 (control)	`49.8 %`	control arm: random initialisation, layer-wise lr decay off (1.0); IDD level 3 (26 classes), official val split (2,036 images), final epoch of 40, EMA weights, scored at 1280x720 (IDD protocol); trained on 14,027 IDD train images, 512x896 crops, batch 8, bf16; one seed	Reproduced 2026-11-16	BASELINE.md (study report), per-class table; runs/seg_supervised_only_ld1/metrics.jsonl commit 59b8054 (bluvit main; local repository, no release tag) · NVIDIA GeForce RTX 5080 (15.5 GB), AMD Ryzen 9 9900X, about 30 GB RAM, Linux 7.0.0-42 (Ubuntu), PyTorch 2.14.1 with CUDA 13.0, bf16 `Local repository, not public. python -m bluvit.train.finetune_seg --config configs/seg_vit_s_idd.yaml train.epochs=40 seg.class_weights=configs/idd_class_weights.json data.workers=8 train.init_from=null train.seed=1337 train.layer_decay=1.0`
IoU, motorcycle, BluViT-S, supervised only, layer decay 1.0 (control)	`56.65 %`	control arm: random initialisation, layer-wise lr decay off (1.0); IDD level 3 (26 classes), official val split (2,036 images), final epoch of 40, EMA weights, scored at 1280x720 (IDD protocol); trained on 14,027 IDD train images, 512x896 crops, batch 8, bf16; one seed	Reproduced 2026-11-16	BASELINE.md (study report), per-class table; runs/seg_supervised_only_ld1/metrics.jsonl commit 59b8054 (bluvit main; local repository, no release tag) · NVIDIA GeForce RTX 5080 (15.5 GB), AMD Ryzen 9 9900X, about 30 GB RAM, Linux 7.0.0-42 (Ubuntu), PyTorch 2.14.1 with CUDA 13.0, bf16 `Local repository, not public. python -m bluvit.train.finetune_seg --config configs/seg_vit_s_idd.yaml train.epochs=40 seg.class_weights=configs/idd_class_weights.json data.workers=8 train.init_from=null train.seed=1337 train.layer_decay=1.0`
IoU, autorickshaw, BluViT-S, supervised only, layer decay 1.0 (control)	`70.18 %`	control arm: random initialisation, layer-wise lr decay off (1.0); IDD level 3 (26 classes), official val split (2,036 images), final epoch of 40, EMA weights, scored	Reproduced 2026-11-16	BASELINE.md (study report), per-class table; runs/seg_supervised_only_ld1/metrics.jsonl commit 59b8054 (bluvit main; local repository, no release tag) · NVIDIA GeForce RTX 5080 (15.5 GB), AMD Ryzen 9 9900X,

Quantity	Value	Conditions	Basis	Source and reproduction
		at 1280x720 (IDD protocol); trained on 14,027 IDD train images, 512x896 crops, batch 8, bf16; one seed		about 30 GB RAM, Linux 7.0.0-42 (Ubuntu), PyTorch 2.14.1 with CUDA 13.0, bf16 `Local repository, not public. python -m bluvit.train.finetune_seg --config configs/seg_vit_s_idd.yaml train.epochs=40 seg.class_weights=configs/idd_class_weights.json data.workers=8 train.init_from=null train.seed=1337 train.layer_decay=1.0`
IoU, person, BluViT-S, MAE init (epoch 400)	64.15 %	encoder from the MAE checkpoint at epoch 400 (unlabelled IDD frames only); IDD level 3 (26 classes), official val split (2,036 images), final epoch of 40, EMA weights, scored at 1280x720 (IDD protocol); trained on 14,027 IDD train images, 512x896 crops, batch 8, bf16; one seed	Reproduced 2026-11-15	BASELINE.md (study report), per-class table; runs/seg_mae_init_ep400/metrics.jsonl commit 59b8054 (bluvit main; local repository, no release tag) · NVIDIA GeForce RTX 5080 (15.5 GB), AMD Ryzen 9 9900X, about 30 GB RAM, Linux 7.0.0-42 (Ubuntu), PyTorch 2.14.1 with CUDA 13.0, bf16 `Local repository, not public. python -m bluvit.train.finetune_seg --config configs/seg_vit_s_idd.yaml train.epochs=40 seg.class_weights=configs/idd_class_weights.json data.workers=8 train.init_from=runs/mae_vit_s/mae_last.pt train.seed=1337`
IoU, rider, BluViT-S, MAE init (epoch 400)	68.92 %	encoder from the MAE checkpoint at epoch 400 (unlabelled IDD frames only); IDD level 3 (26 classes), official val split (2,036 images), final epoch of 40, EMA weights, scored at 1280x720 (IDD protocol); trained on 14,027 IDD train images, 512x896 crops, batch 8, bf16; one seed	Reproduced 2026-11-15	BASELINE.md (study report), per-class table; runs/seg_mae_init_ep400/metrics.jsonl commit 59b8054 (bluvit main; local repository, no release tag) · NVIDIA GeForce RTX 5080 (15.5 GB), AMD Ryzen 9 9900X, about 30 GB RAM, Linux 7.0.0-42 (Ubuntu), PyTorch 2.14.1 with CUDA 13.0, bf16 `Local repository, not public. python -m bluvit.train.finetune_seg --config configs/seg_vit_s_idd.yaml train.epochs=40 seg.class_weights=configs/idd_class_weights.json data.workers=8 train.init_from=runs/mae_vit_s/mae_last.pt train.seed=1337`
IoU, motorcycle, BluViT-S, MAE init (epoch 400)	73.4 %	encoder from the MAE checkpoint at epoch 400 (unlabelled IDD frames only); IDD level 3 (26 classes), official val split (2,036 images), final epoch of 40, EMA	Reproduced 2026-11-15	BASELINE.md (study report), per-class table; runs/seg_mae_init_ep400/metrics.jsonl commit 59b8054 (bluvit main; local repository, no release tag) · NVIDIA GeForce RTX 5080 (15.5 GB), AMD Ryzen 9 9900X, about 30 GB RAM, Linux 7.0.0-42 (Ubuntu), PyTorch 2.14.1 with CUDA 13.0, bf16

Quantity	Value	Conditions	Basis	Source and reproduction
		weights, scored at 1280x720 (IDD protocol); trained on 14,027 IDD train images, 512x896 crops, batch 8, bf16; one seed		Local repository, not public. python -m bluvit.train.finetune_seg --config configs/seg_vit_s_idd.yaml train.epochs=40 seg.class_weights=configs/idd_class_weights.json data.workers=8 train.init_from=runs/mae_vit_s/mae_last.pt train.seed=1337
IoU, autorickshaw, BluViT-S, MAE init (epoch 400)	84.92 %	encoder from the MAE checkpoint at epoch 400 (unlabelled IDD frames only); IDD level 3 (26 classes), official val split (2,036 images), final epoch of 40, EMA weights, scored at 1280x720 (IDD protocol); trained on 14,027 IDD train images, 512x896 crops, batch 8, bf16; one seed	Reproduced 2026-11-15	BASELINE.md (study report), per-class table; runs/seg_mae_init_ep400/metrics.jsonl commit 59b8054 (bluvit main; local repository, no release tag) · NVIDIA GeForce RTX 5080 (15.5 GB), AMD Ryzen 9 9900X, about 30 GB RAM, Linux 7.0.0-42 (Ubuntu), PyTorch 2.14.1 with CUDA 13.0, bf16 Local repository, not public. python -m bluvit.train.finetune_seg --config configs/seg_vit_s_idd.yaml train.epochs=40 seg.class_weights=configs/idd_class_weights.json data.workers=8 train.init_from=runs/mae_vit_s/mae_last.pt train.seed=1337
IoU, person, DeepLabV3 ResNet-50, from scratch (reference)	62.87 %	reference CNN: torchvision DeepLabV3 ResNet-50, no pretrained weights, SGD with a poly schedule; IDD level 3 (26 classes), official val split (2,036 images), final epoch of 40, EMA weights, scored at 1280x720 (IDD protocol); trained on 14,027 IDD train images, 512x896 crops, batch 8, bf16; one seed	Reproduced 2026-11-10	BASELINE.md (study report), per-class table; runs/seg_cnn_deeplabv3_r50/metrics.jsonl commit 4261989 (bluvit main; local repository, no release tag) · NVIDIA GeForce RTX 5080 (15.5 GB), AMD Ryzen 9 9900X, about 30 GB RAM, Linux 7.0.0-42 (Ubuntu), PyTorch 2.14.1 with CUDA 13.0, bf16 Local repository, not public. python scripts/train_cnn_baseline.py (study scripts, alongside bluvit)
IoU, rider, DeepLabV3 ResNet-50, from scratch (reference)	66.57 %	reference CNN: torchvision DeepLabV3 ResNet-50, no pretrained weights, SGD with a poly schedule; IDD level 3 (26 classes), official val split (2,036 images), final epoch of 40, EMA	Reproduced 2026-11-10	BASELINE.md (study report), per-class table; runs/seg_cnn_deeplabv3_r50/metrics.jsonl commit 4261989 (bluvit main; local repository, no release tag) · NVIDIA GeForce RTX 5080 (15.5 GB), AMD Ryzen 9 9900X, about 30 GB RAM, Linux 7.0.0-42 (Ubuntu), PyTorch 2.14.1 with CUDA 13.0, bf16 Local repository, not public. python scripts/train_cnn_baseline.py (study

Quantity	Value	Conditions	Basis	Source and reproduction
		weights, scored at 1280x720 (IDD protocol); trained on 14,027 IDD train images, 512x896 crops, batch 8, bf16; one seed		scripts, alongside bluvit)
IoU, motorcycle, DeepLabV3 ResNet-50, from scratch (reference)	71.27 %	reference CNN: torchvision DeepLabV3 ResNet-50, no pretrained weights, SGD with a poly schedule; IDD level 3 (26 classes), official val split (2,036 images), final epoch of 40, EMA weights, scored at 1280x720 (IDD protocol); trained on 14,027 IDD train images, 512x896 crops, batch 8, bf16; one seed	Reproduced 2026-11-10	BASELINE.md (study report), per-class table; runs/seg_cnn_deeplabv3_r50/metrics.jsonl commit 4261989 (bluvit main; local repository, no release tag) · NVIDIA GeForce RTX 5080 (15.5 GB), AMD Ryzen 9 9900X, about 30 GB RAM, Linux 7.0.0-42 (Ubuntu), PyTorch 2.14.1 with CUDA 13.0, bf16 Local repository, not public. python scripts/train_cnn_baseline.py (study scripts, alongside bluvit)
IoU, autorickshaw, DeepLabV3 ResNet-50, from scratch (reference)	82.39 %	reference CNN: torchvision DeepLabV3 ResNet-50, no pretrained weights, SGD with a poly schedule; IDD level 3 (26 classes), official val split (2,036 images), final epoch of 40, EMA weights, scored at 1280x720 (IDD protocol); trained on 14,027 IDD train images, 512x896 crops, batch 8, bf16; one seed	Reproduced 2026-11-10	BASELINE.md (study report), per-class table; runs/seg_cnn_deeplabv3_r50/metrics.jsonl commit 4261989 (bluvit main; local repository, no release tag) · NVIDIA GeForce RTX 5080 (15.5 GB), AMD Ryzen 9 9900X, about 30 GB RAM, Linux 7.0.0-42 (Ubuntu), PyTorch 2.14.1 with CUDA 13.0, bf16 Local repository, not public. python scripts/train_cnn_baseline.py (study scripts, alongside bluvit)
MAE masked-patch loss, epoch 1 of 400	0.9053	masked autoencoder, 75 percent masking, normalised-pixel loss; 512x896, batch 32, bf16; 42,320 unlabelled IDD frames with nothing	Reproduced 2026-11-15	BASELINE.md (study report), MAE pretraining section; runs/mae_vit_s/metrics.jsonl commit 59b8054 (bluvit main; local repository, no release tag) · NVIDIA GeForce RTX 5080 (15.5 GB), AMD Ryzen 9 9900X, about 30 GB RAM, Linux 7.0.0-42 (Ubuntu), PyTorch 2.14.1 with CUDA 13.0, bf16

Quantity	Value	Conditions	Basis	Source and reproduction
		from val or test (asserted in code)		`Local repository, not public. python -m bluvit.train.pretrain_mae --config configs/mae_vit_s.yaml`
MAE masked-patch loss, epoch 50 of 400	`0.3127`	masked autoencoder, 75 percent masking, normalised-pixel loss; 512x896, batch 32, bf16; 42,320 unlabelled IDD frames with nothing from val or test (asserted in code)	Reproduced 2026-11-15	BASELINE.md (study report), MAE pretraining section; runs/mae_vit_s/metrics.jsonl commit 59b8054 (bluvit main; local repository, no release tag) · NVIDIA GeForce RTX 5080 (15.5 GB), AMD Ryzen 9 9900X, about 30 GB RAM, Linux 7.0.0-42 (Ubuntu), PyTorch 2.14.1 with CUDA 13.0, bf16 `Local repository, not public. python -m bluvit.train.pretrain_mae --config configs/mae_vit_s.yaml`
MAE masked-patch loss, epoch 150 of 400	`0.2988`	masked autoencoder, 75 percent masking, normalised-pixel loss; 512x896, batch 32, bf16; 42,320 unlabelled IDD frames with nothing from val or test (asserted in code)	Reproduced 2026-11-15	BASELINE.md (study report), MAE pretraining section; runs/mae_vit_s/metrics.jsonl commit 59b8054 (bluvit main; local repository, no release tag) · NVIDIA GeForce RTX 5080 (15.5 GB), AMD Ryzen 9 9900X, about 30 GB RAM, Linux 7.0.0-42 (Ubuntu), PyTorch 2.14.1 with CUDA 13.0, bf16 `Local repository, not public. python -m bluvit.train.pretrain_mae --config configs/mae_vit_s.yaml`
MAE masked-patch loss, epoch 400 of 400	`0.2888`	masked autoencoder, 75 percent masking, normalised-pixel loss; 512x896, batch 32, bf16; 42,320 unlabelled IDD frames with nothing from val or test (asserted in code)	Reproduced 2026-11-15	BASELINE.md (study report), MAE pretraining section; runs/mae_vit_s/metrics.jsonl commit 59b8054 (bluvit main; local repository, no release tag) · NVIDIA GeForce RTX 5080 (15.5 GB), AMD Ryzen 9 9900X, about 30 GB RAM, Linux 7.0.0-42 (Ubuntu), PyTorch 2.14.1 with CUDA 13.0, bf16 `Local repository, not public. python -m bluvit.train.pretrain_mae --config configs/mae_vit_s.yaml`
MAE pretraining wall clock, 400 epochs	`29.7 h`	masked autoencoder, 75 percent masking, normalised-pixel loss; 512x896, batch 32, bf16; 42,320 unlabelled IDD frames with nothing from val or test (asserted in code); summed over the run's resumed segments; loader reads a 1280-wide	Reproduced 2026-11-15	BASELINE.md (study report), MAE pretraining section commit 59b8054 (bluvit main; local repository, no release tag) · NVIDIA GeForce RTX 5080 (15.5 GB), AMD Ryzen 9 9900X, about 30 GB RAM, Linux 7.0.0-42 (Ubuntu), PyTorch 2.14.1 with CUDA 13.0, bf16 `Local repository, not public. python -m bluvit.train.pretrain_mae --config configs/mae_vit_s.yaml`

Quantity	Value	Conditions	Basis	Source and reproduction
		resized cache of the corpus		
MAE pretraining throughput, median	`179.6 img/s`	masked autoencoder, 75 percent masking, normalised-pixel loss; 512x896, batch 32, bf16; 42,320 unlabelled IDD frames with nothing from val or test (asserted in code); end to end including data loading	Reproduced 2026-11-15	BASELINE.md (study report), MAE pretraining section commit 59b8054 (bluvit main; local repository, no release tag) · NVIDIA GeForce RTX 5080 (15.5 GB), AMD Ryzen 9 9900X, about 30 GB RAM, Linux 7.0.0-42 (Ubuntu), PyTorch 2.14.1 with CUDA 13.0, bf16 `Local repository, not public. python -m bluvit.train.pretrain_mae --config configs/mae_vit_s.yaml`
Fine-tune wall clock, 40 epochs, BluViT-S, MAE init (epoch 400)	`4.45 h`	40 epochs including a val pass each epoch; one seed	Reproduced 2026-11-15	BASELINE.md (study report), arm table commit 59b8054 (bluvit main; local repository, no release tag) · NVIDIA GeForce RTX 5080 (15.5 GB), AMD Ryzen 9 9900X, about 30 GB RAM, Linux 7.0.0-42 (Ubuntu), PyTorch 2.14.1 with CUDA 13.0, bf16 `Local repository, not public. python -m bluvit.train.finetune_seg --config configs/seg_vit_s_idd.yaml train.epochs=40 seg.class_weights=configs/idd_class_weights.json data.workers=8 train.init_from=runs/mae_vit_s/mae_last.pt train.seed=1337`
Fine-tune wall clock, 40 epochs, DeepLabV3 ResNet-50, from scratch	`5.62 h`	40 epochs including a val pass each epoch; one seed; the CNN shared the CPU with a cache build, so its time is slightly pessimistic (study report)	Reproduced 2026-11-10	BASELINE.md (study report), arm table commit 4261989 (bluvit main; local repository, no release tag) · NVIDIA GeForce RTX 5080 (15.5 GB), AMD Ryzen 9 9900X, about 30 GB RAM, Linux 7.0.0-42 (Ubuntu), PyTorch 2.14.1 with CUDA 13.0, bf16 `Local repository, not public. python scripts/train_cnn_baseline.py (study scripts, alongside bluvit)`
Segmentation inference throughput on the val set	`46.8 img/s`	BluViT-S with the segmentation head, batch 1, full image at width 1280, including argmax and the resize to 1280x720; timed over 300 of the 2,036 val images; includes data loading	Reproduced 2026-11-09	bench.md (study benchmark, generated 2026-11-09) commit 8d8fa66 (bluvit main; local repository, no release tag) · NVIDIA GeForce RTX 5080 (15.5 GB), AMD Ryzen 9 9900X, about 30 GB RAM, Linux 7.0.0-42 (Ubuntu), PyTorch 2.14.1 with CUDA 13.0, bf16 `Local repository, not public. python scripts/step6_bench.py (study scripts)`

Quantity	Value	Conditions	Basis	Source and reproduction
BluViT-S encoder parameters	21.61 M	ViT-S/16: width 384, depth 12, 6 heads, 4 register tokens; the segmentation model with its head has 25.26 M	Reproduced 2026-11-15	runs/seg_mae_init_ep400/train.log (logged at start-up) commit 59b8054 (bluvit main; local repository, no release tag) · NVIDIA GeForce RTX 5080 (15.5 GB), AMD Ryzen 9 9900X, about 30 GB RAM, Linux 7.0.0-42 (Ubuntu), PyTorch 2.14.1 with CUDA 13.0, bf16 `Local repository, not public. logged by python -m bluvit.train.finetune_seg at start-up`
Differing pixels, BluViT's label masks vs the official IDD tool	0 px	all 16,063 Part I + II train and val masks; BluViT's tools/idd_prepare.py masks against masks written by the official AutoNUE createLabels.py (level3Id), after two fixes (fallback background id, polygon rasteriser)	Reproduced 2026-11-09	mask_parity.md (study report), run of 2026-11-09T13:29:17Z commit 08a4b6c (bluvit main; local repository, no release tag) · NVIDIA GeForce RTX 5080 (15.5 GB), AMD Ryzen 9 9900X, about 30 GB RAM, Linux 7.0.0-42 (Ubuntu), PyTorch 2.14.1 with CUDA 13.0, bf16 `Local repository, not public. python tools/idd_prepare.py masks --root <IDD_Segmentation> --verify-against <AutoNUE public-code>/helpers/anue_labels.py`
Mean IoU on IDD val, ResNet-50 + FPN head, AdamW, from scratch	63.57 %	AdamW lr 1e-3, poly; from scratch; IDD val (2,036 images, 26 classes, IDD protocol); final epoch of 40, EMA; one seed	Reproduced 2026-11-17	CONV_ENCODER.md (study report), accuracy table; runs/seg_r50_fpn_adamw/metrics.jsonl commit df0d01a (bluvit main; local repository, no release tag) · NVIDIA GeForce RTX 5080 (15.5 GB), AMD Ryzen 9 9900X, about 30 GB RAM, Linux 7.0.0-42 (Ubuntu), PyTorch 2.14.1 with CUDA 13.0, bf16 `Local repository, not public. python scripts/train_conv_seg.py --arch r50_fpn --opt adamw (study scripts, alongside bluvit)`
Mean IoU on IDD val, BluConv-T + FPN head, AdamW, from scratch	65.89 %	AdamW lr 1e-3, poly; from scratch; IDD val (2,036 images, 26 classes, IDD protocol); final epoch of 40, EMA; one seed	Reproduced 2026-11-16	CONV_ENCODER.md (study report), accuracy table; runs/seg_bluconv_t_adamw/metrics.jsonl commit df0d01a (bluvit main; local repository, no release tag) · NVIDIA GeForce RTX 5080 (15.5 GB), AMD Ryzen 9 9900X, about 30 GB RAM, Linux 7.0.0-42 (Ubuntu), PyTorch 2.14.1 with CUDA 13.0, bf16 `Local repository, not public. python scripts/train_conv_seg.py --arch bluconv_t --opt adamw (study scripts, alongside bluvit)`
IoU, person, IDD val, ResNet-50 + FPN	66.99 %	AdamW lr 1e-3, poly; from scratch; IDD val (2,036 images,	Reproduced 2026-11-17	CONV_ENCODER.md (study report), per-class table; runs/seg_r50_fpn_adamw/metrics.jsonl

Quantity	Value	Conditions	Basis	Source and reproduction
head, AdamW, from scratch		26 classes, IDD protocol); final epoch of 40, EMA; one seed		commit df0d01a (bluvit main; local repository, no release tag) · NVIDIA GeForce RTX 5080 (15.5 GB), AMD Ryzen 9 9900X, about 30 GB RAM, Linux 7.0.0-42 (Ubuntu), PyTorch 2.14.1 with CUDA 13.0, bf16 `Local repository, not public. python scripts/train_conv_seg.py --arch r50_fpn --opt adamw (study scripts, alongside bluvit)`
IoU, rider, IDD val, ResNet-50 + FPN head, AdamW, from scratch	72.05 %	AdamW lr 1e-3, poly; from scratch; IDD val (2,036 images, 26 classes, IDD protocol); final epoch of 40, EMA; one seed	Reproduced 2026-11-17	CONV_ENCODER.md (study report), per-class table; runs/seg_r50_fpn_adamw/metrics.jsonl commit df0d01a (bluvit main; local repository, no release tag) · NVIDIA GeForce RTX 5080 (15.5 GB), AMD Ryzen 9 9900X, about 30 GB RAM, Linux 7.0.0-42 (Ubuntu), PyTorch 2.14.1 with CUDA 13.0, bf16 `Local repository, not public. python scripts/train_conv_seg.py --arch r50_fpn --opt adamw (study scripts, alongside bluvit)`
IoU, motorcycle, IDD val, ResNet-50 + FPN head, AdamW, from scratch	75.59 %	AdamW lr 1e-3, poly; from scratch; IDD val (2,036 images, 26 classes, IDD protocol); final epoch of 40, EMA; one seed	Reproduced 2026-11-17	CONV_ENCODER.md (study report), per-class table; runs/seg_r50_fpn_adamw/metrics.jsonl commit df0d01a (bluvit main; local repository, no release tag) · NVIDIA GeForce RTX 5080 (15.5 GB), AMD Ryzen 9 9900X, about 30 GB RAM, Linux 7.0.0-42 (Ubuntu), PyTorch 2.14.1 with CUDA 13.0, bf16 `Local repository, not public. python scripts/train_conv_seg.py --arch r50_fpn --opt adamw (study scripts, alongside bluvit)`
IoU, autorickshaw, IDD val, ResNet-50 + FPN head, AdamW, from scratch	83.48 %	AdamW lr 1e-3, poly; from scratch; IDD val (2,036 images, 26 classes, IDD protocol); final epoch of 40, EMA; one seed	Reproduced 2026-11-17	CONV_ENCODER.md (study report), per-class table; runs/seg_r50_fpn_adamw/metrics.jsonl commit df0d01a (bluvit main; local repository, no release tag) · NVIDIA GeForce RTX 5080 (15.5 GB), AMD Ryzen 9 9900X, about 30 GB RAM, Linux 7.0.0-42 (Ubuntu), PyTorch 2.14.1 with CUDA 13.0, bf16 `Local repository, not public. python scripts/train_conv_seg.py --arch r50_fpn --opt adamw (study scripts, alongside bluvit)`

Table 4. Every measured quantity in this report, with its source, commit, hardware and the command that reproduces it. “Reproduced” rows were run by the author; “Reported” rows are quoted from the repository at the commit shown.

Cite this report

APA

Roy, S. J. (2026). Masked pretraining for a from-scratch Vision Transformer on Indian road scenes: BluViT-S on IDD (Technical Report IF-TR-2026-06, Version 1.1). IsoFrames. https://isoframes.io/reports/if-tr-2026-06/

BibTeX

```
@techreport{isoframes_if_tr_2026_06,
  author      = {Roy, Sherin Joseph},
  title       = {Masked pretraining for a from-scratch Vision Transformer on Indian road scenes:
BluViT-S on IDD},
  institution = {IsoFrames, BluRabbit Mobility},
  number      = {IF-TR-2026-06},
  type        = {Technical Report},
  year        = {2026},
  month       = nov,
  note        = {Version 1.1},
  url         = {https://isoframes.io/reports/if-tr-2026-06/}
}
```