Distillo, ergo sum

Heterogeneous knowledge distillation in medical computer vision

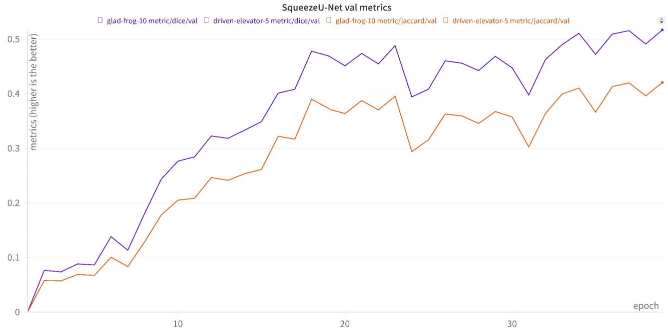
Giacomo Bellini, Ludovica Mazza, Matteo Rampolla, Daniele Solombrino

Our goals

- **Heterogeneous** Knowledge Distillation → Transformer to CNN
  - TransU-Net → SqueezeU-Net
- Formulate HKD-related **hypotheses**
- Build **code** to verify hypotheses
- Look at **results** → verify hypotheses

Step 1: train from scratch

- Reasons:
  - Need for complete access to entire training process
  - Low quality checkpoints currently available
- Setup:
  - 60/20/20 train/validation/test **split**
  - 2200/900/900 train/val/test **slices**
  - 40 epochs
- TransU-Net
  - **Transformer**, 100M trainable parameters
  - **IoU** 0.68, **Dice** 0.78 (val)
- SqueezeU-Net
  - **CNN**, 0.4M trainable parameters
  - **IoU** 0.42, **Dice** 0.51 (val)

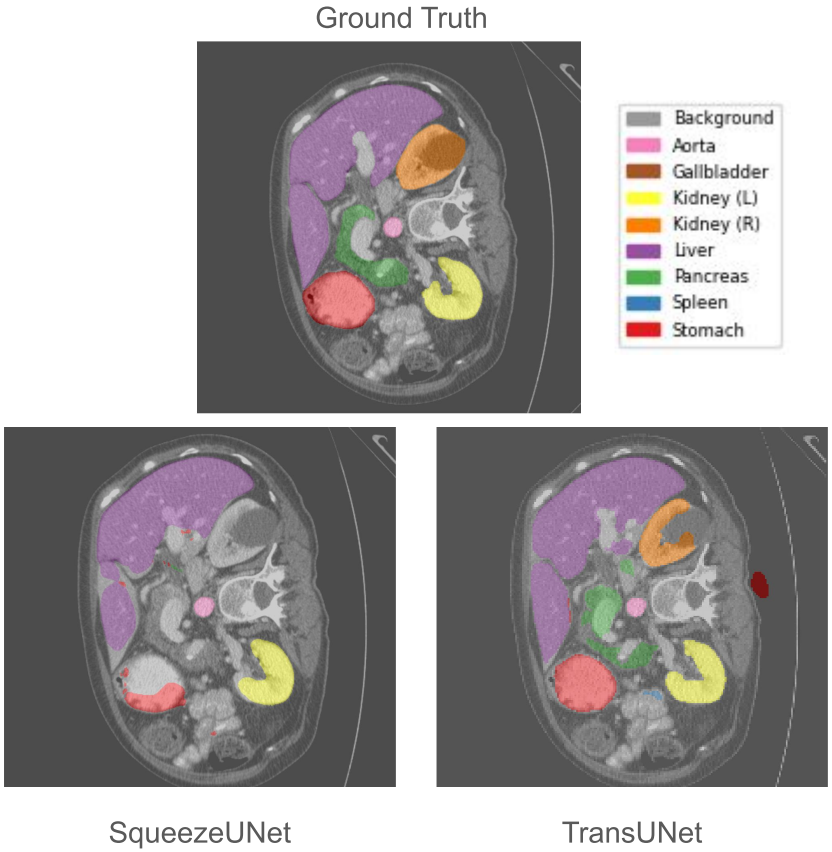


Step 2: models qualitative evaluation

- Prediction vs. Ground truth
- Saliency
- GradCAM

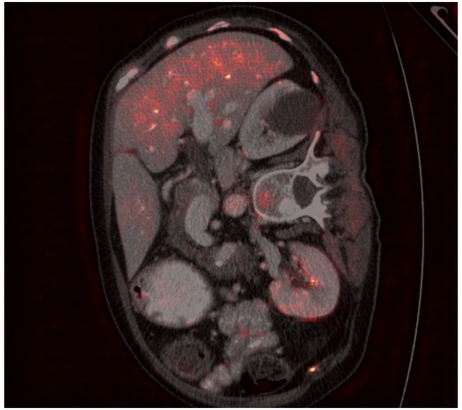
Step 2: models qualitative evaluation

- **Prediction vs. Ground truth ✅**
- Saliency
- GradCAM

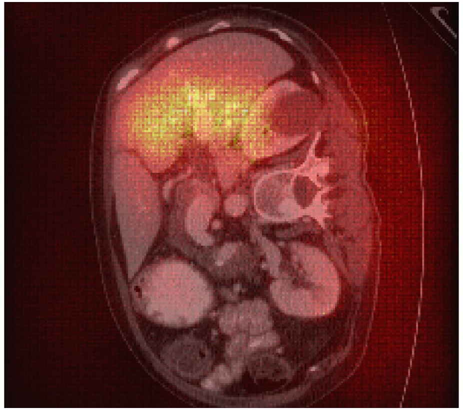


Step 2: models qualitative evaluation

- Prediction vs. Ground truth
- **Saliency** ✅
- GradCAM



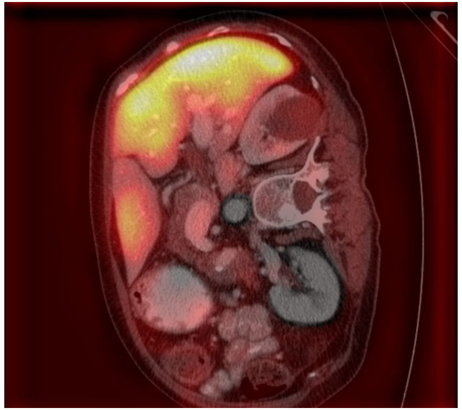
SqueezeUNet



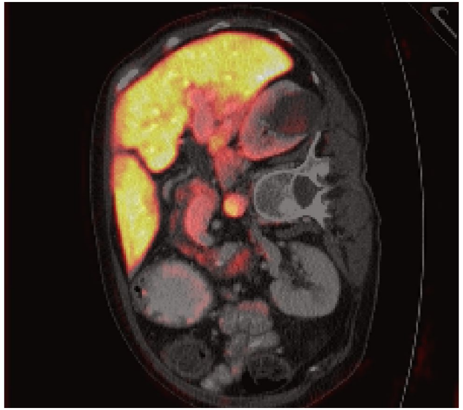
TransUNet

Step 2: models qualitative evaluation

- Prediction vs. Ground truth
- Saliency
- **GradCAM ✅**



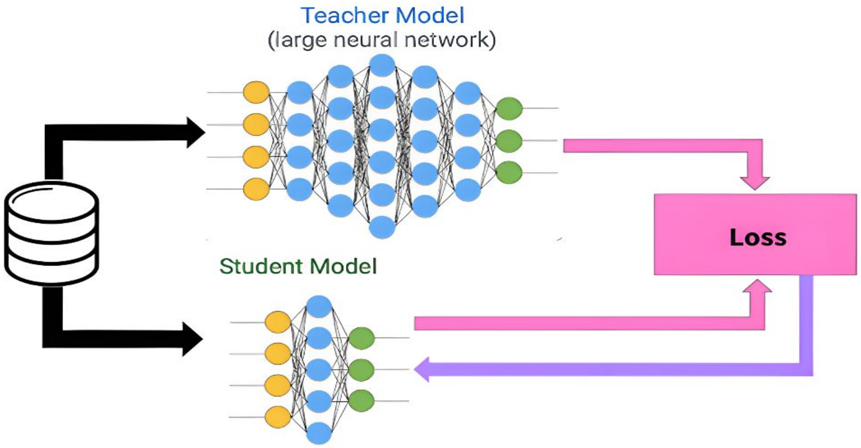
SqueezeUNet



TransUNet

Step 3: hypothesis 1 setup

- **Logits** matching
- Dice term + Cross-Entropy term + KL-divergence term → hypothesis 1 loss
  - Dice and CE → **segmentation** loss (TransU-Net paper)
  - KL-div → **HKD** loss (KD paper)



Step 3: hypothesis 1 claim

Distilled CNN → more global context captured w.r.t. train from scratch 💡

- Higher validation **metrics**?
- Better **qualitative** results?
  - Prediction vs. Ground truth
  - GradCAM

Step 3: hypothesis 1 results

Distilled CNN → more global context captured

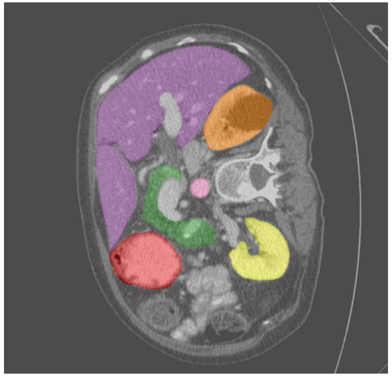
- Higher validation metrics
  - IoU 0.49 (was 0.42), Dice 0.61 (vs. 0.51) ✅
- Better qualitative results
  - Prediction vs. Ground truth
  - GradCAM

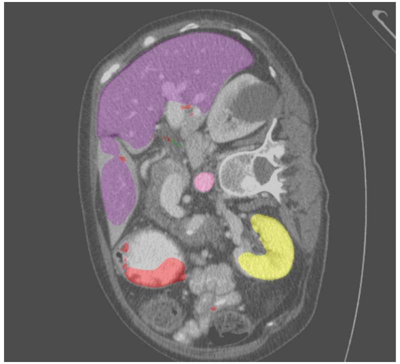
Step 3: hypothesis 1 results

Distilled CNN → more global context captured

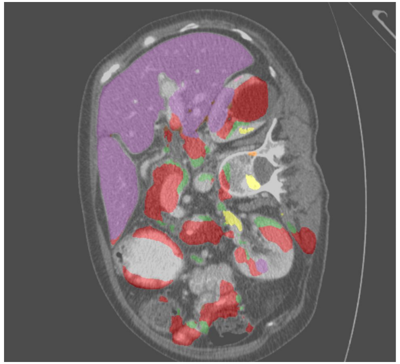
- Higher validation metrics
  - IoU 0.49 (was 0.42), Dice 0.61 (vs. 0.51)
- Better qualitative results ⚠️
  - **Prediction vs. Ground truth** ✅
  - GradCAM

Ground Truth





SqueezeUNet



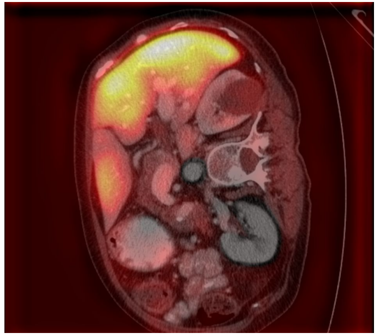
KD SqueezeUNet

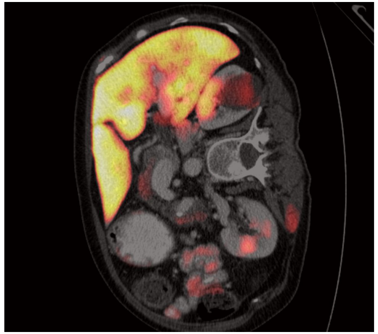
Step 3: hypothesis 1 results

Distilled CNN → more global context captured

- Higher validation metrics
  - IoU 0.49 (was 0.42), Dice 0.61 (vs. 0.51)
- Better qualitative results ⚠️
  - Prediction vs. Ground truth
  - **GradCAM** ✅

SqueezeUNet





KD SqueezeUNet

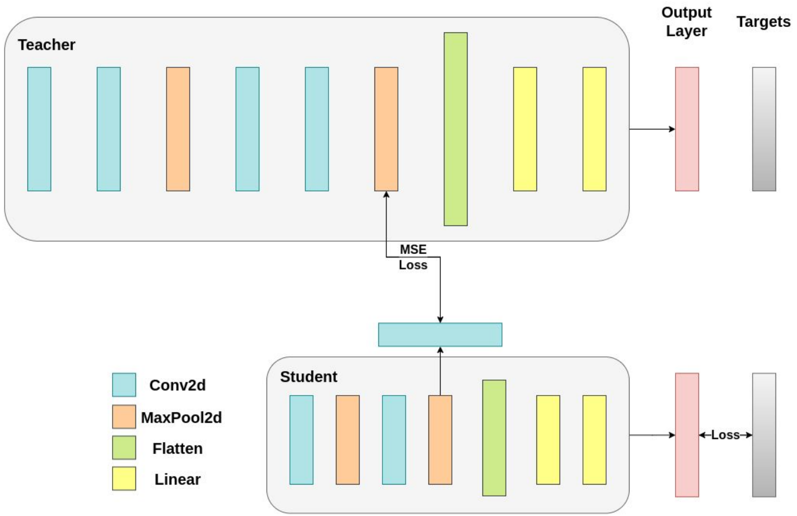
Step 3: hypothesis 1 comment

Distilled CNN → more global context captured ✅

- Higher validation metrics
  - IoU 0.49 (was 0.42), Dice 0.61 (vs. 0.51)
- Better qualitative results
  - Prediction vs. Ground truth
  - GradCAM
- Student forced to **mimic** teacher "**way of thinking**"

Step 4: hypothesis 2 setup

- (Latent) **embedding** matching
- Dice term + Cross-Entropy term + L1 term → hypothesis 2 loss
    - Dice and CE → **segmentation** loss (TransU-Net paper)
    - L1 → HKD loss (KD paper)



Step 4: hypothesis 2 claim

Easier to **organize data** in the embedding space than matching probabilistic distribution 💡

[Relative representations enable zero-shot latent space communication @ ICLR 2023]

- Higher validation **metrics**?
- Better **qualitative** results?
  - Prediction vs. Ground truth
  - Saliency
  - GradCAM

Step 4: hypothesis 2 results

Easier to **organize data** in the embedding space than matching probabilistic distribution 💡

Relative representations enable zero-shot latent space communication @ ICLR 2023

- Higher validation **metrics**? ❌
    - Distillation loss **diverges** (even after hyperparameter fine-tuning!)

Step 4: hypothesis 2 comment

Easier to **organize data** in the embedding space than matching probabilistic distribution 💡

[Relative representations enable zero-shot latent space communication @ ICLR 2023]

- 0.4M parameters not enough→ **4M, 6.7M, 16M** → ✅
- Increasing student parameters **helps** the latent space structures converge

Future works

- **Latent space**
  - Visualizations
  - Advance alignments
- More **advanced** heterogeneous KD techniques
  - Coaching a Teachable Student @ CVPR 2023
  - Generative HKD with Masked Image Modeling
  - HKD using Information Flow Modeling @ CVPR 2020
- **Homogeneous** KD
  - TinyViT: Fast Pretraining Distillation for Small ViTs @ ECCV 2022

Recap

- **HKD** → Transformer to CNN → 100M params to 0.4M
- Synapse dataset → Medical **image segmentation**
- Hypothesis 1: HKD makes CNN more aware of **global contexts** ✅
  - Logits matching
  - Student forced to **mimic** teacher "**way of thinking**"
- Hypothesis 2: latent matching **easier** than logits matching ❌
  - Increasing student params → latent space structure starts to converge → ✅
- **Challenges**
  - Deep Learning interpretability
  - (Joint) hyperparameter tuning