Multi-task Learning for Video Processing: Going with the Flow

George Kalitsios
K3Y LTD
gkalitsios@k3y.bg

Authors: Efklidis Katsaros, George Kalitsios, Panagiotis Radoglou-Grammatikis, Panagiotis Sarigiannidis

Acknowledgments

This project has received funding from the European Union’s Horizon Europe research and innovation programme under grant agreement No 101135930 (CoGNETs).

Disclaimer: Funded by the European Union.

Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Commission.

Neither the European Union nor the European Commission can be held responsible for them.

CoGNETs
Continuums of Game Nets

Motivation

Real-World Challenges affecting dental video quality during procedures:

- Microcameras attached to handpieces to obtain continuous, close-up views of the operative field, which is crucial for precision and safety.
- The small cameras introduce issues:
  - Handpiece vibration leads to visible frame shake.
  - Light changes, saliva, and water cause blur, noise, and distortion.
  - Depth and camera proximity leads to non-uniform motion.
- These issues compromise video clarity and increase surgeon discomfort.
- Existing solutions are either costly, inefficient, or not tailored to real-time use.
- Our work aims to provide an effective, real-time solution to enhance video quality during dental procedures.

Introduction

- Multi-Task Learning (MTL) improves efficiency by handling multiple tasks at once in a single network pass.
- Existing MTL models are limited to static image input without temporal modeling.
- Most works combine high-level tasks like semantic segmentation, object detection, depth estimation, which are at the same level of understanding.
- These setups often ignore low-level tasks like video enhancement or denoising, which are crucial for clear and stable video.
- Our approach integrates optical flow to capture motion and temporal information between frames.
- This allows our system to enhance and understand intra-oral surgical videos in real time, combining low-level and high-level tasks efficiently.

Contributions

- We introduce MOSTNET+, the first multi-task network for video enhancement, segmentation, and optical flow
- Our model is built with multi-scale and motion-aware components, allowing it to effectively learn both spatial and temporal dependencies.
- Achieves competitive accuracy across tasks compared to state-of-the-art single-task networks
- Offers a better performance-efficiency trade-off, running up to 2× faster than combining single-task models
- Reaches real-time inference at ~25 FPS with low latency using TensorRT in half-precision, making it a strong candidate for clinical use.

Related Work

MTL for Scene Understanding

UberNet [8] (Kokkinos et al.)

One of the first to tackle multiple tasks (segmentation, detection, etc.) with shared CNNs and diverse datasets.

MTAN [9] (Liu et al.)

Combines shared backbone with task-specific attention for better feature allocation.

PAPNet [10] (Zhang et al.)

Uses affinity matrices for joint prediction of depth, normals, and semantics.

ATRC [11] (Bruggeman et al.)

Uses Neural Architecture Search for learning optimal cross-task attention.

MTL for Scene Enhancement

MTFFNet [12] (Cui et al.)

Dual-stream network for deblurring and super-resolution of face images with limited interaction between tasks.

RIRGAN [13] (Yu et al.)

GAN-based MTL for medical denoising and super-resolution, tailored to specific domains and input types.

LEDNet [14] (Zhou et al.)

Tackled low-light enhancement and deblurring jointly, but without explicit task separation.

DP3DF [15] (Xu et al.)

Proposed DP3DF for joint denoising, enhancement, and super-resolution using local spatiotemporal cues.

Proposed MOST-NET+ Architecture

- MOST-NET+ (Multi-Output, Multi-Scale, Multi-Task), is a general deep learning framework for multi-task prediction.
- The encoder is composed of two main components: a feature extractor and a feature alignment module.
  - The Feature Extractor enriches representations at each scale by integrating both deep features and image-level features.
  - The Feature Alignment module is responsible for aligning features from the previous frame with those of the current frame and fusing their information using channel-wise attention.
- The Decoder produces dense outputs by branching out scale-wise, with each branch generating task specific predictions for its corresponding scale.
- These scale-wise decoders are also shared with the optical flow modules, which estimate and iteratively refine flow fields in a bottom-up cascading manner.

Proposed MOST-NET+ Architecture

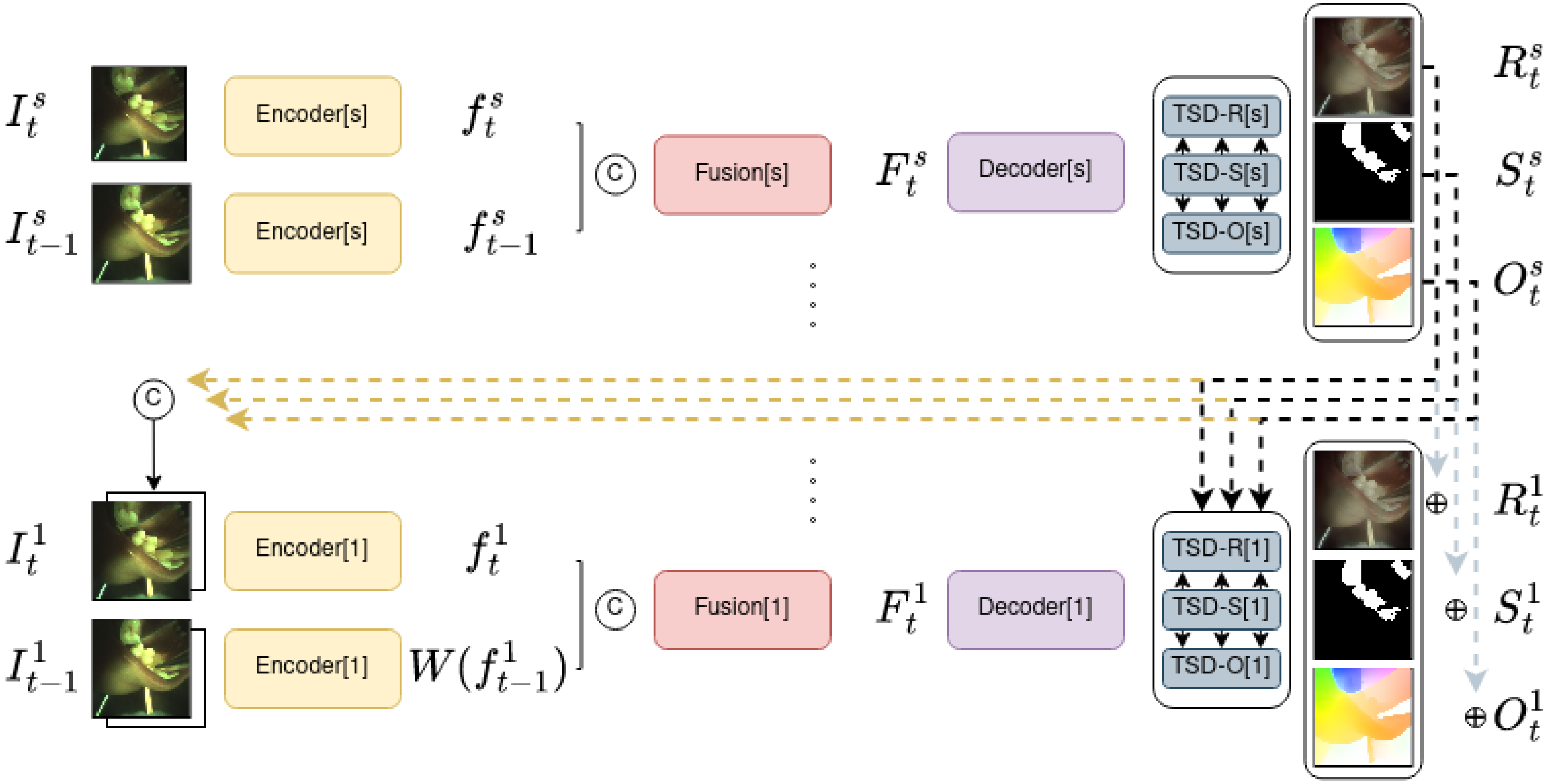


Fig. 1: The proposed architecture of MOST-NET+, which stands for Multi-Output, Multi-Scale, Multi-Task Network. It's a versatile deep learning framework we designed to handle multiple tasks at the same time within a single model.

Proposed MOST-NET+ Architecture

- *Enhancement Module*: Performs color correction, denoising, and deblurring at the pixel level to improve frame quality.
- *Optical Flow Module*: Tracks pixel motion between frames to stabilize the video.
- *Tooth Segmentation Module*: Provides reference points to reinitialize stabilization when tracking is lost.
- MOSTNET+ leverages positive interactions between these tasks:
  - Enhancement ↔ Optical Flow: Cleaner frames improve motion estimation accuracy.
  - Optical Flow ↔ Enhancement: Accurate motion cues enhance frame alignment and deblurring.
  - Segmentation ↔ Stabilization: Tooth segmentation anchors enable reliable reinitialization of stabilization.
- This synergy between tasks is what makes MOSTNET+ effective in stabilizing and enhancing video sequences.

Dataset

- We conducted our experiments on the **Vident-real Clinical Dataset** [7], which contains 100 real intra-oral surgical video sequences.
- The dataset is well-suited for multi-task learning and supports three tasks:
  - Video Restoration,
  - Teeth Segmentation,
  - and Optical Flow Estimation.
- Each frame in these video sequences is paired with a high-quality reference frame, a segmentation mask for the teeth, and optical flow labels extracted using the RAFT model.
- To ensure faster experimentation cycles, we limited each video sequence to 100 frames.
- We split the dataset into 65 training, 10 validation, and 25 test sequences.

Performance Evaluation

Optical Flow Estimation

Baseline Models Used for Comparison

- *RAFT [3], FlowNet [4]: strong performance across datasets*

Evaluation Metrics

- *EPE (End-Point Error): Measures flow accuracy — lower is better*

$$\mathrm{EPE} = \frac{1}{N}\sum_{i=1}^{N}\left\|\mathbf{f}_i^{\mathrm{pred}} - \mathbf{f}_i^{\mathrm{gt}}\right\|_2$$

Where:

N= number of pixels

$\mathbf{f}_i^{\mathrm{pred}}, \mathbf{f}_i^{\mathrm{gt}}$ = predicted and ground truth flow vectors

$\|\cdot\|_2$ = Euclidean norm

Performance Evaluation

Video Enhancement

Baseline Models Used for Comparison

- *ESTRNN [1], MIMO-Unet [2]*: lightweight, efficient architectures

Evaluation Metrics

- PSNR (Peak Signal-to-Noise Ratio): *Measures image* pixel-level *fidelity — higher is better*

$$\mathrm{PSNR} = 10 \cdot \log_{10}\left(\frac{\mathrm{MAX}^2}{\mathrm{MSE}}\right)$$

Where:

MAX= maximum pixel value(e.g., 255)

$$\mathrm{MSE} = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\left(I(i,j) - K(i,j)\right)^2$$

Performance Evaluation

Video Enhancement

Evaluation Metrics

- SSIM (Structural Similarity Index): *Assesses structural similarity — higher is better*
- It takes into account luminance, contrast, and texture, providing a more perceptual measure of quality.

$$\mathrm{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}$$

Where:

μ_x, μ_y = means of images

σ_x^2, σ_y^2 = variances

σ_{xy} = covariance

C_1, C_2 = stability constants

Performance Evaluation

Semantic Segmentation

Baseline Models Used for Comparison

- *UNet++ [5], DeepLabv3+ (ResNet-50 encoder) [6]: well-established in medical and general segmentation*

Evaluation Metrics

- IoU (Intersection over Union): *Segmentation quality — higher is better*

$$\mathrm{IoU} = \frac{|\mathrm{Prediction} \cap \mathrm{GroundTruth}|}{|\mathrm{Prediction} \cup \mathrm{GroundTruth}|}$$

Where:
$\cap$ = *pixel-wise intersection*
$\cup$ = *pixel-wise union of masks*

Performance Evaluation

- We evaluate our method across three tasks:
  - Video Enhancement,
  - Optical Flow Estimation,
  - Semantic Segmentation.
- We compare two versions of our model, MOSTNET+SW and MOSTNET+DW:
  - MOSTNET+SW: Optical flow predicted at a single (lowest) scale
  - MOSTNET+DW: Optical flow predicted at both lowest and medium scales
- Both variants consistently demonstrate competitive or superior performance across all tasks in a single, unified multi-task, multiscale architecture.

Evaluation results *PSNR, SSIM, EPE, IoU*
Vident-real clinical dataset

	Methods	PSNR $\uparrow$	SSIM $\uparrow$	EPE $\downarrow$	IoU $\uparrow$
	BASELINE	17.80/18.77	0.829/ 0.855	9.24/8.52	0.270/0.214
Optical Flow Estimation	**FLOWNet [4]**	-	-	2.63/2.11	-
	RAFT [3]	-	-	1.81/1.43	-
Video Enhancement	**MIMO-UNET [2]**	25.83/26.37	0.967/0.966	-	-
	ESTRNN [1]	28.65/28.39	0.977/0.973	-	-
Semantic Segmentation	**UNET++ [5]**	-	-	-	0.730/0.788
	DLV3+ [6]	-	-	-	0.746/0.765
Combined Single-taskers	**ESTRNN+RAFT+DLV3+**	28.65/28.39	0.977/0.973	1.81/1.43	0.746/0.765
Proposed Method	**MOSTNET+(SW)**	29.96/29.27	0.972/0.965	3.51/2.81	0.685/0.723
	MOSTNET+(DW)	29.97/28.99	0.969/0.963	2.13/1.70	0.716/0.739

TABLE I: Performance over PSNR, SSIM, EPE and IoU on the test/validation set

Evaluation results *P(M), FPS*

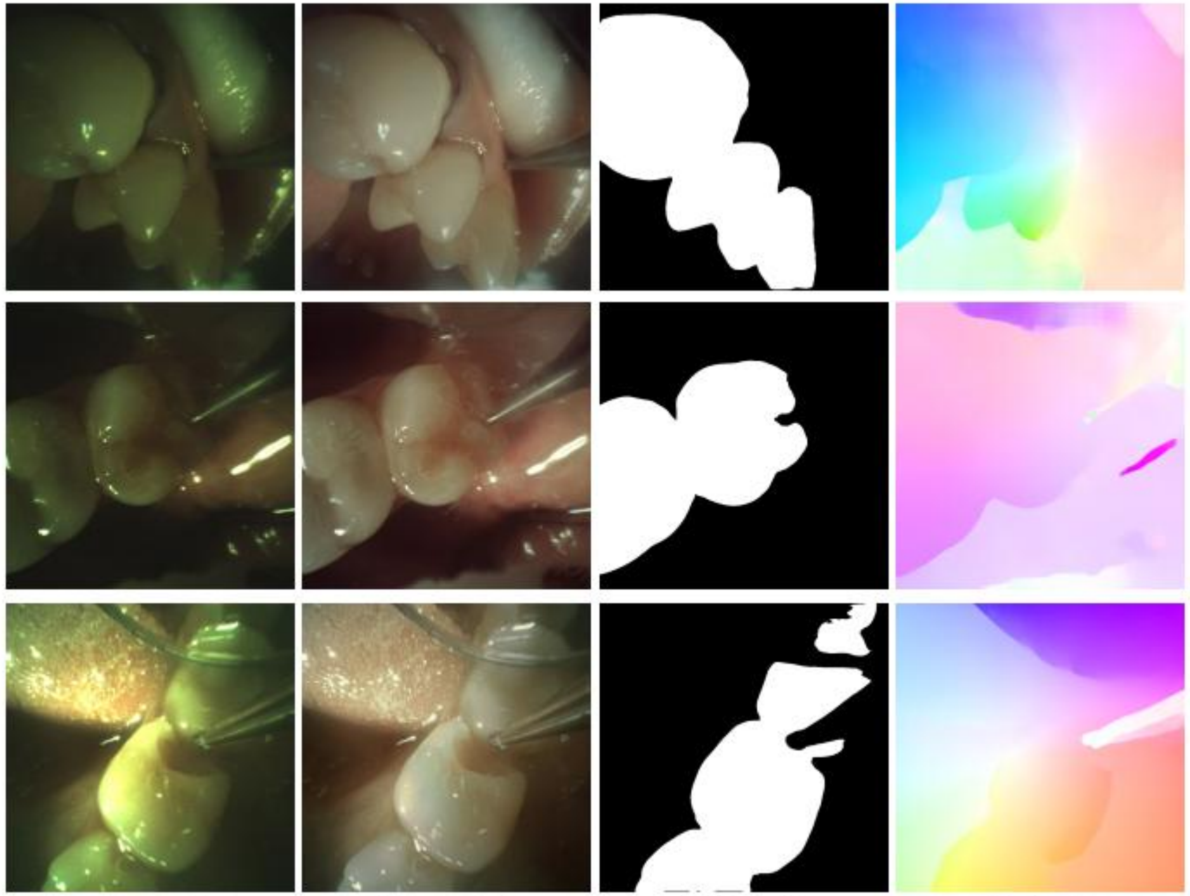
Vident-real clinical dataset

Group	Methods	#P(M)	FPS
	BASELINE	-	-
Optical Flow Estimation	FLOWNet [4]	38.7	52.7
	RAFT [3]	5.3	5.1
Video Enhancement	MIMO-UNET [2]	6.8	4.6
	ESTRNN [1]	2.3	10.6
Semantic Segmentation	UNET++ [5]	50.0	7.9
	DLV3+ [6]	26.7	25.5
Combined Single-taskers	ESTRNN+RAFT+DLV3+	34.3	3.0
Proposed Method	MOSTNET+(SW)	13.2	6.4
	MOSTNET+(DW)	29.8	5.2

TABLE I: Performance over P(M) and FPS.

- MOSTNET+SW (13.2M) is very lightweight, much smaller than large single-task models like UNet++ (50M).
- Even MOSTNET+DW (29.8M), the larger version, is more compact than running separate models for each task (34.3M).
- In terms of speed:
  - MOSTNET+SW runs at ~ 6.4 FPS
  - MOSTNET+DW runs at ~ 5.2 FPS — about twice as fast as separate models like ESTRNN, RAFT, and DLV3+ all together (3.0 FPS).
  - With TensorRT and half precision, MOSTNET+DW exceeds 25 FPS for real-time use
- Because of this efficiency and speed, MOSTNET+ is well-suited for next-generation IoT e-health systems.

Qualitative Performance
Vident-real clinical dataset



(a) Input B_t (b) Restored R_t (c) Mask M_t (d) Flow O_t

Conclusions

- Proposed MOSTNET+: a unified, multitask, multi-scale architecture designed for real-time intra-oral video processing.
- It jointly tackles video enhancement, optical flow estimation, and teeth segmentation—all within a single, efficient model.
- Designed for real-time clinical use with efficient, low-latency performance (~25 FPS)
- Utilizes task synergies and scale-specific modeling for improved robustness and generalization
- Variants MOSTNET+SW and MOSTNET+DW:
  - Outperform or match state-of-the-art single-task models
  - Maintain lower computational complexity and runtime overhead
- Demonstrates the potential of multi-task learning in real-time medical video applications

Thank You!

George Kalitsios

K3Y LTD

gkalitsios@k3y.bg

- **Multi-task Learning for Video Processing: Going with the Flow**

CoGNETs

Continuums of Game Nets

References

[1] ESTRNN – Zhong et al., "Efficient Spatio-Temporal Recurrent Neural Network for Video Deblurring," ECCV 2020.

[2] MIMO-UNet – Cho et al., "Rethinking Coarse-To-Fine Approach in Single Image Deblurring," ICCV 2021.

[3] RAFT – Teed and Deng, "RAFT: Recurrent All-Pairs Field Transforms for Optical Flow," ECCV 2020.

[4] FlowNet – Dosovitskiy et al., "FlowNet: Learning Optical Flow with Convolutional Networks," ICCV 2015.

[5] UNet++ – Zhou et al., "UNet++: A Nested U-Net Architecture for Medical Image Segmentation," MICCAI Workshops 2018.

[6] DeepLabv3+ – Chen et al., "Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation," ECCV 2018.

[7] Vident-Real – Węsierski et al., "Vident-Real: An Intra-Oral Video Dataset for Multi-Task Learning," 2024.

[8] UBERNet – Kokkinos, "UBERNet: Training a Universal Convolutional Neural Network for Low-, Mid-, and High-Level Vision Using Diverse Datasets and Limited Memory," CVPR 2017.

[9] MTAN– Liu et al., "End-to-End Multi-Task Learning with Attention," CVPR 2019.

[10] PAPNet – Zhang et al., "Pattern-Affinitive Propagation Across Depth, Surface Normal and Semantic Segmentation," CVPR 2019.

[11] ATRC– Brüggemann et al., "Exploring Relational Context for Multi-Task Dense Prediction," ICCV 2021.

[12] MTFFNet – Cui et al., "Joint Face Super-Resolution and Deblurring Using Multi-Task Feature Fusion Network," ICVISP 2023.

[13] RIRGAN – Yu et al., "RIRGAN: An End-to-End Lightweight Multi-Task Learning Method for Brain MRI Super-Resolution and Denoising," Computers in Biology and Medicine, 2023.

[14] LEDNet – Zhou et al., "LEDNet: Joint Low-Light Enhancement and Deblurring in the Dark," ECCV 2022.

[15] DP3DF – Xu et al., "Deep Parametric 3D Filters for Joint Video Denoising and Illumination Enhancement in Video Super Resolution," AAAI 2023.