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

Efklidis Katsaros*, George Kalitsios*, Anastasia Kazakli*,
Panagiotis Radoglou-Grammatikis*†, Panagiotis Sarigiannidis†

***Abstract*—Multi-task learning constitutes the prevalent paradigm in numerous vision applications that cast an eye on runtime efficiency. At present however, deep multi-task networks are limited in single-image processing. While various motion descriptors have been proposed to estimate motion across frames in the video processing literature, the problem of incorporating motion compensation in multi-task learning is yet understudied. Moreover, the type of tasks typically integrated within multi-task architectures constitute only visual scene understanding tasks, i.e. tasks at the same level of hierarchy. In this work, we address multi-task video scene enhancement in combination with understanding for intra-oral scenes. Our work proposes a novel architecture derived from the multi-output, multi-scale, multi-task (MOST) family of models, that further incorporates optical flow into its design. We showcase that our work yields a) on-par performance with state-of-the-art convolutional networks across multiple tasks and architectures b) improved performance-vs-efficiency trade-off than combining single-task methods, i.e. up to $2\times$ faster runtimes. and c) low-latency and real-time processing at 25 FPS, when compiled with TensorRT at half precision, allowing for commercial outcomes.**

***Index Terms*—multi-task learning, video processing, optical flow, dental interventions**

I. Introduction

Recent years have seen tremendous progress in the application of machine learning models to the real world. As models grew more mature, the efforts to utilize them in real-life followed. In that context, Multi-task Learning (MTL) attracted significant research interest. MTL architectures are designed to exploit synergies among tasks, boosting overall performance while accelerating inference by reducing the need for separate forward passes through individual task-specific networks.

Numerous studies have explored how to exploit inter-task relationships and improve performance. For example, MTAN [18] constructed a shared feature space and utilized soft attention to dynamically extract task-relevant features for each decoder. PAD-Net [28] generated initial predictions for all tasks and improved them through attention-guided message passing for distillation. ATRC [2] facilitated interaction between tasks by learning various attention types, tailored to each task combination. Similarly, MTI-Net [25] supported feature propagation across tasks, enabling richer task-to-task information flow.

The aforementioned approaches are not suitable for quality enhancement because they are specifically tailored for scene understanding datasets and tasks like semantic segmentation, object detection, depth estimation, or surface normal estimation [11, 7, 5, 22]. These tasks operate within a similar level hierarchy, where information exchangeability is profound. In such cases, the output for each task can complement and improve the performance of the other. Furthermore, existing MTL solutions generally operate on single, static images, failing to exploit the potential benefits of information aggregation across successive video frames within the temporal domain.

In [14], the authors proposed the first video multi-task network for simultaneous visual scene enhancement and understanding. The work focused on a dental use case, showcasing a novel setup in which a microcamera is integrated into an adapter attached to a dental handpiece near the bur. This configuration enables real-time monitoring of the treatment area during drilling procedures. However, the compact nature of the camera leads to visual distortions, and the inherent hand movements of the dentist necessitate robust video stabilization. To address these challenges, the study proposed an algorithmic framework to compensate for degraded video quality, thereby facilitating the adoption of affordable microcameras in digital dentistry for improved intraoral visualization. While the model outperforms state-of-the-art single-taskers in the laboratory-acquired dataset [12], training the network on real intra-oral scenes [27] unveils further challenges. First, the data from patients exhibit substantially different artifact; motion blur is lesser while defocus blur is frequently present. Second, computing motion descriptors and aligning the frames becomes more intricate because the scenes in real-world interventions demonstrate variable depth.

To remedy those issues, we propose MOST-NET+, an architecture that extends MOST-NET to pixel-wise motion estimation via means of optical flow to enable more degrees of freedom for post-processing the video feeds. Our proposed network is the first to tackle video enhancement, semantic segmentation and optical flow estimation jointly. Thereafter, we validate our approach on the Vident-real dataset [27], a publicly available clinical dataset with real intra-oral surgeries. Our contributions are summarized as follows:

* K3Y Ltd, Sofia, Bulgaria. Emails: {ekatsaros, gkalitsios, nkazakli, pradoglou}@k3y.bg
† Department of Electrical and Computer Engineering, University of Western Macedonia, Kozani, Greece. Emails: {pradoglou, psarigiannidis}@uowm.gr

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 authors 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.

- Our network achieves on-par performance with state-of-the-art convolutional methods across multiple tasks and architectures.
- The proposed MOST-NET+ shows improved performance-vs-efficiency trade-off than combining single-task methods.
- When compiled with TensorRT at half precision, our network achieves low-latency and real-time processing at approximately 25 FPS, allowing for immediate commercialization.

II. Related Work

Multi-task learning has been widely explored for visual scene understanding, particularly in the context of autonomous driving. A seminal work is UberNet by Kokkinos et al. [16], a convolutional network that jointly addressed seven vision tasks in a single architecture, including but not limited to boundary detection, human parts segmentation, semantic segmentation, and object detection. A key innovation of this work was the ability to train a deep network using diverse datasets for each task group, under a limited memory budget. Liu et al. [18] introduced the Multi-Task Attention Network (MTAN), which combines a shared feature extractor with task-specific attention modules. These modules selectively aggregate task-relevant features from a global feature pool, improving task-specific predictions while leveraging shared representations. Zhang et al. [32] proposed the Pattern-Affinitive Propagation Network (PAPNet) to jointly predict depth, surface normals, and semantics. Their approach models task relationships using an affinity matrix, enabling refined predictions through a learned feature diffusion process. Similarly, Vandenhende et al. [24] argued that task affinities vary with scale, and designed a multi-scale architecture that uses spatial attention to enhance feature sharing across both tasks and resolutions. The authors argued that this design addresses the limitations of convolutional receptive fields at high resolutions by incorporating lower-scale context. More recently, Bruggemann et al. [2] introduced ATRC, a multi-task model that learns task-specific attention mechanisms to enable adaptive cross-task communication. Instead of manually designing attention types, their method uses a search-based strategy inspired by Neural Architecture Search (NAS) to discover effective interactions based on task demands.

While multi-task learning has been extensively studied for scene understanding, its application to visual scene enhancement remains relatively underexplored. Only a limited number of studies have investigated how enhancement tasks can benefit from joint modeling. For example, Cui et al. [8] proposed a dual-network approach to address low-resolution and motion blur in face images. Their system employed separate yet concurrent networks, enhanced with multi-scale fusion and attention mechanisms. Although designed to mitigate error propagation between tasks, the networks shared only gradients, which may lead to inefficiencies in both training and inference. Yu et al. [30] focused on medical image enhancement, proposing a multi-task GAN-based framework for simultaneous super-resolution and denoising of MRI scans. Their approach demonstrated potential in leveraging shared generation features, but remained limited in scope due to its domain specificity. Katsaros et al. [13] proposed to address video deblurring and denoising in dynamic scenes with a pure multi-task architecture. Therein, the authors learnt deformable offsets to align consecutive frames and restore the degraded frames with visually enhanced counterparts. In a separate direction, some works attempted to tackle multiple enhancement objectives using single-task architectures. Zhou et al. [34] addressed low-light enhancement and video deblurring, but treated the two tasks as a unified dense prediction problem without explicit task disentanglement or multi-task design. Similarly, Xu et al. [29] proposed Deep Parametric 3D Filters (DP3DF), a representation that integrates local spatiotemporal cues to simultaneously perform denoising, illumination enhancement, and super-resolution.

In contrast to aforementioned works, our method targets both visual scene enhancement and understanding tasks. Moreover, it targets the video domain instead of operating on the basis of single images. Leveraging the temporal domains offers informative visual cues for both enhancement and understanding tasks. Similarly, we argue that incorporating a model for the temporal domain within a multi-task architecture should be beneficial. In prior work, targeting dental interventions in lab conditions, Katsaros et al. [14] proposed the *MOST-NET* architecture and addressed video deblurring, denoising, color mapping, tooth segmentation and homography estimation. The study proposed a multi-task, decoder-focused model [24] for video processing, dubbed multi-output, multi-scale, multi-task, and applied it to video enhancement of dental scenes in laboratory settings. Specifically, the model framed color correction [31], denoising, and deblurring [26, 33, 13] as a unified dense prediction task. Additionally, it incorporated auxiliary tasks such as homography estimation [17] to stabilize the video stream [1], and tooth segmentation [4, 35], which was leveraged to reinitialize stabilization when needed. The architecture owes multiple scale-specific heads for each task, enabling a hierarchical processing scheme where predictions were propagated from coarser to finer resolution levels. It propagates the outputs bottom-up, from the lowest to the highest scale level. This formulation yields a two-fold benefit. First, it enables task synergy by loop-like modeling of task interactions in the encoder and decoder across scales. Second, it allows for refinement of predictions across scales and provides model insights by assessing which scales contribute to which task's performance improvement. Our work takes a step further and replaces homography with dense, pixel-wise optical flow estimation as a richer motion descriptor, refines the architecture and experiments with a clinical dataset of real life interventions.

III. Proposed Method

A. Problem Formulation

We revisit *MOST* (Multi-Output, Multi-Scale, Multi-Task), a general deep learning formulation for multi-task prediction.

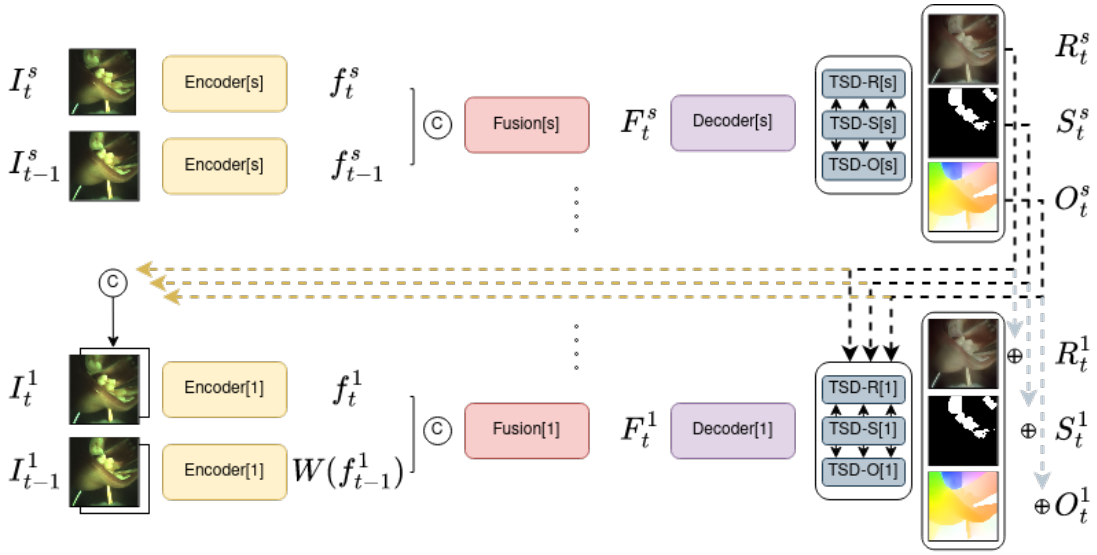


Fig. 1: The proposed architecture. MOST-NET+ extracts features from consecutive frames at different scales. Subsequently, it performs feature alignment and fusion at the fusion blocks to compensate the motion and filter out irrelevant context. Decoders follow at each scale to process the features and task-specific heads (TSD) predict outputs for all tasks at all scales. The outputs are propagated bottoms-up to enable information exchange and task refinement.

MOST facilitates the propagation of task outputs in two key ways: (i) across scale levels within the encoder and decoder to exploit cross-task interactions, and (ii) in a bottom-up manner to enable progressive refinement of predictions. We assume that the network produces T task-specific outputs at each scale, denoted as $\{\mathcal{O}_i^s\}_{i,s=1}^{T,S}$, where $s = 1$ corresponds to the original input resolution. These task outputs are propagated not only within the same scale (innerscale propagation), but also upsampled from coarser scales and passed forward to both the encoder layers and the task-specific branches in the decoder at finer scales. To guide this propagation, we define the following task-specific relationship:

$$u_i(\mathcal{O}_i^{s+1}) \approx \mathcal{O}_i^s, \tag{1}$$

where u_i denotes some operator, for instance, the upsampling operator for segmentation or the scaling operator for homography estimation.

In our setting, all tasks share training samples in $\mathcal{D}_{=}\{\{B\}_j, \{\mathcal{O}_i^s\}_j\}_{i,s,j=1}^{T,S,N}$, where $\{\mathcal{O}_i^s\}_j$ is a label related to task i at scale s for the j-th training sample $\{B\}_j$, while N denotes number of samples in training data. We omit the sample subscript for notational brevity. In the context of deep learning, the optimal set of parameters θ for some network $\mathcal{F}_\theta$ under the MOST formulation is derived by minimizing a penalization criterion:

$$\mathcal{L}(\theta) = \sum_i^T \sum_s^S \lambda_i L_i \left(\mathcal{O}_i^s, \hat{\mathcal{O}}_i^s(\theta) \right), \tag{2}$$

where λ_i is a scalar weighting value, $\hat{\mathcal{O}}_i^s(\theta)$ is an estimate of $\mathcal{O}_i^s$ for j-th sample in $\mathcal{D}$ and L_i is a distance measure.

B. Architecture

We present *MOST-NET+*, a novel instantiation of the *MOST* model introduced earlier. As shown in Fig. 1, 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.

C. Encoders

Feature Extraction: At each time step, *MOST-NET+* independently extracts features f_{t-1}^s and f_t^s from two input frames B_{t-1} and B_t across three scales. To implement the U-shaped downsampling architecture [21], deep features are obtained using 3×3 convolutions with strides of 1, 2, and 2 for scales $s = 1$, 2, and 3, respectively. Each convolution is followed by a ReLU activation. At the coarser scales ($s = 2, 3$), the downsampling process typically leads to some loss of spatial detail. To address this, we enhance the scale-specific representations in two ways. First, we concatenate the deep features with image-level features—that is, features derived from the downsampled version of the input image itself, following the approach of *MIMO-UNET* [6]. This concatenated representation is then passed through a stack of five residual blocks at each scale. Second, to facilitate

richer representations through cross-scale interaction, we apply Asymmetric Feature Fusion [6] to enhance the output of the residual blocks at scales $s = 1$ and 2. The resulting features f_t^s have output channel dimensions defined as 2^{s+4}. In contrast to the approach in [14], we omit Fourier transforms entirely, opting instead for standard residual blocks. Furthermore, at each time step t, the features from time step $t-1$ are not recomputed, but instead retrieved from a cached version, as in [20].

Fusion: At each scale, features f_t^s and $W_{\tilde{F}^s}(f_{t-1}^s)$ are concatenated and a channel attention mechanism [33] follows to selectively fuse them into f_t^s. *MOST-NET+* uses optical flow outputs from lower scales to warp encoder features from the previous time step as $W_{\tilde{F}}(f_{t-1}^s)$. Here, W denotes the warping operator while $\tilde{F}^s$ is an upscaled version of F^{s+1} for higher scales and the identity matrix for $s = 3$.

D. Decoders

Dense Outputs: The attended encoder features F_t^s are forwarded to the corresponding expansion blocks at each scale through skip connections. At the coarsest scale ($s = 3$), the attended features F_t^3 are directly processed by a stack of two residual blocks, each producing 128 output channels. Following this, the resolution is progressively restored using two transposed convolutions with a stride of 2. At finer scales ($s < 3$), the attended features F_t^s are first concatenated with the upsampled decoder features from the next lower scale. This combined representation is then passed through 3×3 convolutions to reduce the number of channels by half. The result is subsequently fed into two residual blocks with 64 and 32 output channels, respectively. These residual block outputs form scale-specific shared backbones, resulting in features denoted as g_t^s. Each scale is followed by lightweight task-specific branches to predict dense outputs. In this work, we expand these branches by employing three 3×3 convolutional layers, each separated by ReLU activations, to estimate M_t^s and R_t^s. *MOST-NET+* facilitates refinement of segmentations at higher scales by upsampling and reintegrating lower-scale predictions into the task-specific branches of subsequent finer scales.

Optical Flow Cascades: At each scale, optical flow estimation modules predict dense, two-channel flow fields representing the horizontal and vertical offsets between consecutive frames. These modules take as input a pair of concatenated feature maps and process them through a cascade of convolutional blocks to estimate a residual flow field, which is subsequently upsampled to a fixed resolution. The flow estimator design incorporates seven (3×3) convolsequenceutional layers at each scale, which progressively reduce the spatial dimensions of the feature maps and predict the optical flow at one-quarter ($\frac{1}{4}$) of the original resolution. Subsequently, the predicted flow fields are upsampled by a factor of four using bilinear interpolation. For example, at the third scale—where the input features are at $\frac{H}{4} \times \frac{W}{4}$ resolution—the feature maps are further downsampled to $\frac{H}{16} \times \frac{W}{16}$, and the flow is predicted at this reduced resolution before being upsampled back to $\frac{H}{4} \times \frac{W}{4}$. It is important to note that, for optical flow estimation, the architecture branches out only at the third (small) and second (medium) scales. We observed through experimentation that including the first (highest resolution) scale yields only marginal performance gains while significantly increasing computational cost, and thus we exclude it from the final model.

IV. Experiments

The experiments are performed on the *Vident-real* [27] clinical dataset suitable for multi-task video processing in intra-oral surgeries, encompassing restoration, teeth segmentation, and inter-frame homography estimation. The dataset comprises 100 real intra-oral surgical sequences, split into training (65 videos), validation (10 videos), and test sets (25 videos). The dataset is recorded at approximately 54 frames per second. Each frame captured during the interventions is paired with its high-quality counterpart, a teeth segmentation mask, and an inter-frame homography matrix. In this work, the homographies are replaced with optical flows distilled from the large version of *RAFT* [23]. Moreover, we restrict each video sequence for all train, validation and test splits to 100 frames to allow for faster experimentation cycles.

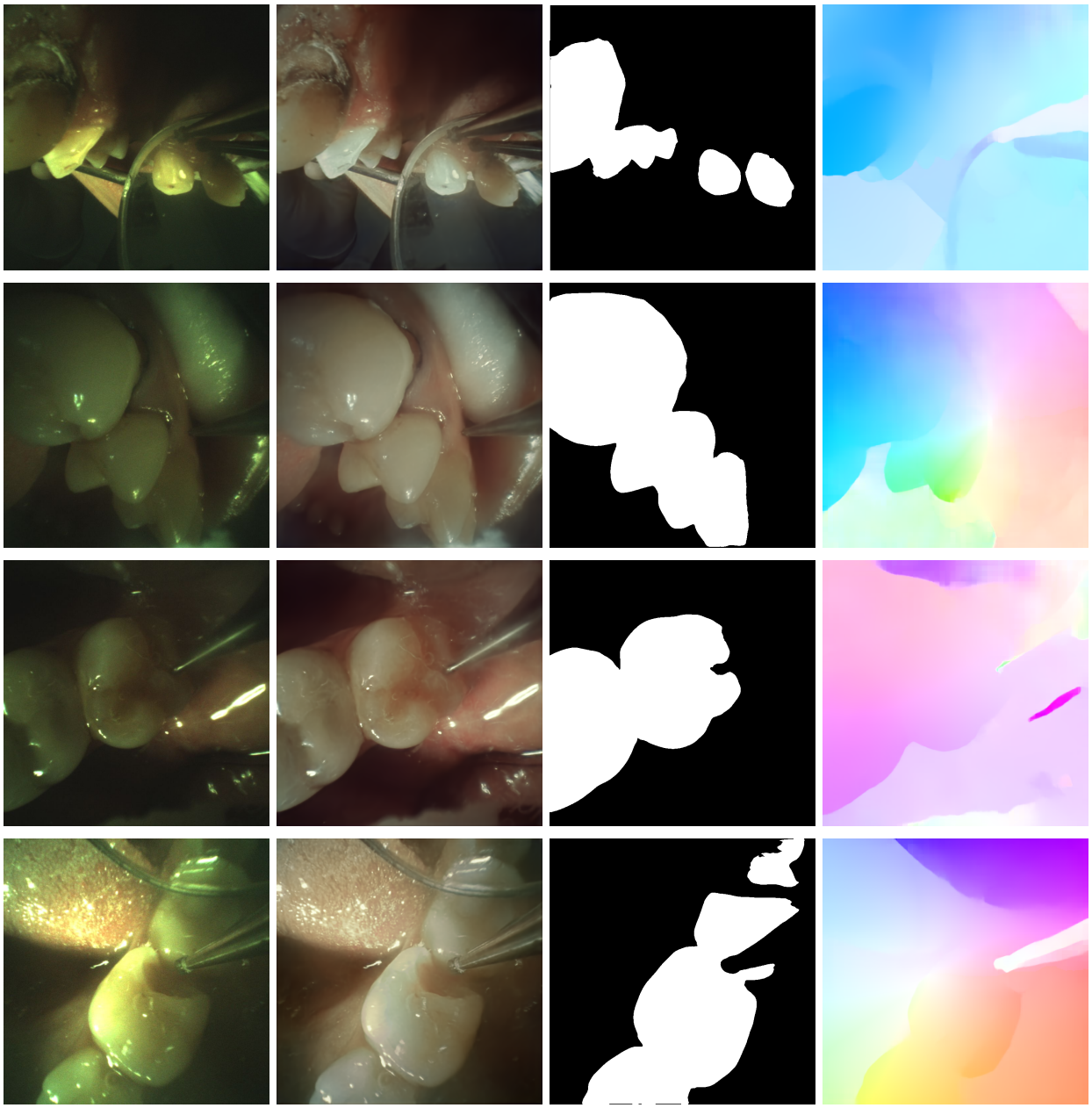
We use the Charbonnier loss [3] as L_1, the binary cross-entropy [19] as L_2 and the Endpoint Error (EPE) [9] as L_3 to train *MOST-NET+*. The task-specific, manually-derived loss weights, are set to 1×10^{-1}, 2×10^{-1} and 1×10^{-1} for λ_1, λ_2 and λ_3 respectively. For all experiments, we employ a batch size of 4, Adam [15] as the optimizer with a learning rate of 1×10^{-4} reduced to 1×10^{-6} with cosine annealing for θ. The training frames are augmented by horizontal and vertical flips with 0.5 probability, and color jittering.

To validate the effectiveness of the proposed approach, we conduct comparisons against representative baselines that balance accuracy and efficiency. For video restoration, we select *ESTRNN* [33] and *MIMO-UNET* [6] as lightweight yet performant architectures. For optical flow estimation, *RAFT* [23] and *FlowNet* [9], widely recognized for their strong performance across diverse datasets. In the segmentation task, we compare against *UNET++* [35] and *DeepLabv3+* [4] with a *ResNet50* [10] encoder, as established benchmarks in medical and general semantic segmentation. These models were selected due to their relevance, reported performance, and suitability for real-time or resource-constrained scenarios.

The performance of the proposed framework is evaluated using task-specific metrics: PSNR and SSIM are employed to assess the fidelity and structural consistency of restored frames, EPE quantifies the accuracy of optical flow estimation, and IoU measures the quality of semantic segmentation, focusing on intra-oral regions. To account for computational efficiency and real-time applicability, we additionally report the number of trainable parameters (#P) and inference speed in frames per second (FPS). For completeness, each metric is reported both on the test and validation sets, with validation values shown in parentheses.

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

Methods	PSNR	SSIM	EPE	IoU	#P(M)	FPS
BASELINE	17.87 (18.77)	0.829 (0.855)	9.24 (8.52)	0.270 (0.214)	-	-
FLOWNet [9]	-	-	2.63 (2.11)	-	38.7	52.7
RAFT [23]	-	-	1.81 (1.43)	-	5.3	5.1
MIMO-UNET [6]	25.83 (26.37)	0.967 (0.966)	-	-	6.8	4.6
ESTRNN [33]	28.65 (28.39)	0.977 (0.973)	-	-	2.3	10.6
UNET++ [35]	-	-	-	0.730 (0.788)	50.0	7.9
DLV3+ [4]	-	-	-	0.746 (0.765)	26.7	25.5
ESTRNN+RAFT+DLV3+	28.65 (28.39)	0.977 (0.973)	1.81 (1.43)	0.746 (0.765)	34.3	3.0
MOSTNET+(SW)	29.96 (29.27)	0.972 (0.965)	3.51 (2.81)	0.685 (0.723)	13.2	6.4
MOSTNET+(DW)	29.97 (28.99)	0.969 (0.963)	2.13 (1.70)	0.716 (0.739)	29.8	5.2



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

Fig. 2: Visualization outputs for all tasks MOST-NET+(DW) performs.

V. Results

Table I presents the results of our experiments. First, we establish lower-bound reference points: we compute baseline PSNR and SSIM between each input image and its ground-truth restoration to define the minimum reconstruction quality that any meaningful model should surpass. We then report baseline EPE (end-point error) under a zero-motion assumption using the optical-flow labels to gauge the average pixel displacement in the dataset. Finally, we include baseline IoU from a trivial "random" classifier that labels every pixel as "teeth" (and none as background) to indicate the segmentation performance achievable without any learned information.

Our proposed approach, *MOSTNET+*, is evaluated in two variants: *MOSTNET+*SW that outputs optical flow only on the small scale and *MOSTNET+*DW which predicts optical flow at the lowerst and the medium scale. Both variants consistently demonstrate competitive or superior performance across all tasks — optical flow estimation, image enhancement, and semantic segmentation — in a single, unified multi-task, multi-scale architecture. *MOSTNET+*DW achieves the best PSNR score of 29.97 and SSIM of 0.969, surpassing the image enhancement baselines *MIMO-UNET* and *ESTRNN*, despite their exclusive focus on that single task. Similarly, in optical flow estimation, *MOSTNET+*DW delivers an EPE of 2.13, better than specialized methods like *FlowNet* and approaching the highly-accurate *RAFT* with and EPE of 1.81, but with significantly better complexity and multi-task integration. In terms of semantic segmentation, our method achieves an IoU of 0.716, very close to the parameter-heavy *UNET++* at 0.730 and *DLV3+* at 0.746.

While delivering competitive performance, the *MOSTNET+* variants maintain a balanced model complexity. *MOSTNET+*SW is especially lightweight at 13.2M parameters, compared to much heavier models like *UNET++* (50.0M) while accommodating multiple tasks where *UNET++* performs only semantic segmentation. *MOSTNET+*DW, while slightly larger at 29.8M, remains more compact than the combined single-taskers in *ESTRNN+RAFT+DLV3+* with a total of 34.3M, while delivering strong performance across all tasks. In terms of computational cost, *MOSTNET+* also achieves a good balance, with runtime speeds suitable for practical applications, specifically, 6.4 FPS for SW, 5.2 FPS for DW, that is about $2\times$ faster than the forked *ESTRNN+RAFT+DLV3+* option. Moreover, we cast our *MOSTNET+*DW model on TensorRT to half precision inference and achieve runtimes of 25.4 FPS. Moreover, our model has low latency, since it does not utilize future frames, making it appropriate for industrial applications.

Unlike all other baselines that focus on isolated tasks, our *MOSTNET+* architecture performs joint prediction of optical flow, enhancement, and segmentation in a multi-scale fashion. This design not only reduces the need for multiple networks but also allows cross-task feature sharing, resulting in more efficient learning and better generalization. Despite the inherent challenge of multi-task learning, our method not only closes the performance gap with single-task models but often surpasses them, emphasizing the benefit of shared representations and spatial coherence across scales. Fig. 2 illustrates the qualitative performance of MOST-NET+(DW), showcasing the input frame, the restored output, the predicted segmentation mask, and the estimated optical flow, demonstrating consistent and coherent results across all tasks.

Despite these advantages, there are still failure cases for different tasks. Such are scenes where the input imagery is too dark and/or blurry, resulting in erroneous flow estimates. Another issue is that the optical flow is hindered under fast motion and performance degrades in such cases. We attribute the optical flow errors of such cases to the inherent difficulties of heavily degraded visual scenes as well as the optical flow pseudo-labels that we consider noisy to a large extent. Unfortunately, manually assessing the RAFT flow estimates to filter them out is not straight-forward for a human, and perhaps unsupervised flow learning schemes can assist.

Beyond its relevance in multi-task video analysis, our method is particularly well-suited for deployment in remote healthcare scenarios. Its low-latency and real-time performance enable on-device processing for smart intra-oral cameras or tele-dentistry systems, supporting diagnostic and interventional tasks without the need for cloud offloading. As such, it aligns with the goals of next-generation IoT-enabled e-health applications.

VI. Conclusion

In this work, we introduced *MOSTNET+*, a unified, multi-task, multi-scale architecture for simultaneous optical flow estimation, video enhancement, and semantic segmentation in intraoral video data. Motivated by the need for efficient, real-time solutions in clinical settings, we leveraged task synergies and hierarchical scale-specific modeling to design a robust framework. *MOSTNET+* delivers strong generalization and competitive performance in multiple tasks through joint modeling of enhancement and understanding across multiple scales. Experimental results demonstrate that our proposed variants, *MOSTNET+*SW and *MOSTNET+*DW, not only compete but often outperform state-of-the-art single-task networks, while maintaining lower computational complexity and runtime overhead. Moreover, our solution achieves approximately 25 FPS with low latency. These findings validate the eligibility of multi-task model deployment in real-time applications.

References

[1] Arwen Bradley, Jason Klivington, Joseph Triscari, and Rudolph van der Merwe. Cinematic-l1 video stabilization with a log-homography model. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1041–1049, 2021.

[2] David Brüggemann, Menelaos Kanakis, Anton Obukhov, Stamatios Georgoulis, and Luc Van Gool. Exploring relational context for multi-task dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15869–15878, 2021.

[3] Pierre Charbonnier, Laure Blanc-Feraud, Gilles Aubert, and Michel Barlaud. Two deterministic half-quadratic regularization algorithms for computed imaging. In *Proceedings of 1st International Conference on Image Processing*, volume 2, pages 168–172. IEEE, 1994.

[4] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018.

[5] Xianjie Chen, Roozbeh Mottaghi, Xiaobai Liu, Sanja Fidler, Raquel Urtasun, and Alan Yuille. Detect what you can: Detecting and representing objects using holistic models and body parts. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1971–1978, 2014.

[6] Sung-Jin Cho, Seo-Won Ji, Jun-Pyo Hong, Seung-Won Jung, and Sung-Jea Ko. Rethinking coarse-to-fine approach in single image deblurring. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4641–4650, 2021.

[7] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.

[8] Yun Cui, Chaoying Tang, and Qiaoyue Huang. Joint face super-resolution and deblurring using multi-task feature fusion network. In *7th International Conference on Vision, Image and Signal Processing (ICVISP 2023)*, volume 2023, pages 57–61. IET, 2023.

[9] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2758–2766, 2015.

[10] 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.

[11] Georgios Kalitsios, Vasileios Mygdalis, and Ioannis Pitas. Domain adaptation in power line segmentation: A new synthetic dataset. In *2023 IEEE International Conference on Image Processing (ICIP)*, pages 1820–1824, 2023.

[12] Efklidis Katsaros, Piotr Kopa Ostrowski, Anna Jezierska, Emilia Lewandowska, Jacek Rumiński, and Daniel Węsierski. Vident-lab: a dataset for multi-task video processing of phantom dental scenes, 2022.

[13] Efklidis Katsaros, Piotr K Ostrowski, Daniel Wesierski, and Anna Jezierska. Concurrent video denoising and deblurring for dynamic scenes. *IEEE Access*, 9:157437–157446, 2021.

[14] Efklidis Katsaros, Piotr K Ostrowski, Krzysztof Włódarczak, Emilia Lewandowska, Jacek Ruminski, Damian Siupka-Mróz, Łukasz Lassmann, Anna Jezierska, and Daniel Wesierski. Multi-task video enhancement for dental interventions. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part VII*, pages 177–187. Springer, 2022.

[15] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[16] Iasonas Kokkinos. Ubernet: Training a universal convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6129–6138, 2017.

[17] Hoang Le, Feng Liu, Shu Zhang, and Aseem Agarwala. Deep homography estimation for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7652–7661, 2020.

[18] Shikun Liu, Edward Johns, and Andrew J Davison. End-to-end multi-task learning with attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1871–1880, 2019.

[19] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

[20] Piotr Kopa Ostrowski, Efklidis Katsaros, Daniel Wesierski, and Anna Jezierska. Bp-evd: Forward block-output propagation for efficient video denoising. *IEEE Transactions on Image Processing*, 31:3809–3824, 2022.

[21] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[22] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. *ECCV (5)*, 7576:746–760, 2012.

[23] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*, pages 402–419. Springer, 2020.

[24] Simon Vandenhende, Stamatios Georgoulis, and Luc Van Gool. Mti-net: Multi-scale task interaction networks for multi-task learning. In *European Conference on Computer Vision*, pages 527–543. Springer, 2020.

[25] Simon Vandenhende, Stamatios Georgoulis, Wouter Van Gansbeke, Marc Proesmans, Dengxin Dai, and Luc Van Gool. Multi-task learning for dense prediction tasks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 2021.

[26] Xintao Wang, Kelvin CK Chan, Ke Yu, Chao Dong, and Chen Change Loy. Edvr: Video restoration with enhanced deformable convolutional networks. In *CVPR Workshops*, 2019.

[27] Daniel Węsierski, Anna Jezierska, Piotr Kopa Ostrowski, Emilia Lewandowska, Efklidis Katsaros, and Agata

Żółtowska. Vident-real: an intra-oral video dataset for multi-task learning, 2024.

[28] Dan Xu, Wanli Ouyang, Xiaogang Wang, and Nicu Sebe. Pad-net: Multi-tasks guided prediction-and-distillation network for simultaneous depth estimation and scene parsing. In *CVPR*, pages 675–684, 2018.

[29] Xiaogang Xu, Ruixing Wang, Chi-Wing Fu, and Jiaya Jia. Deep parametric 3d filters for joint video denoising and illumination enhancement in video super resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 3054–3062, 2023.

[30] Miao Yu, Miaomiao Guo, Shuai Zhang, Yuefu Zhan, Mingkang Zhao, Thomas Lukasiewicz, and Zhenghua Xu. Rirgan: An end-to-end lightweight multi-task learning method for brain mri super-resolution and denoising. *Computers in Biology and Medicine*, 167:107632, 2023.

[31] Mohan Zhang, Qiqi Gao, Jinglu Wang, Henrik Turbell, David Zhao, Jinhui Yu, and Yan Lu. RT-VENet: A convolutional network for real-time video enhancement. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 4088–4097, 2020.

[32] Zhenyu Zhang, Zhen Cui, Chunyan Xu, Yan Yan, Nicu Sebe, and Jian Yang. Pattern-affinitive propagation across depth, surface normal and semantic segmentation. In *CVPR*, pages 4106–4115, 2019.

[33] Zhihang Zhong, Ye Gao, Yinqiang Zheng, and Bo Zheng. Efficient spatio-temporal recurrent neural network for video deblurring. In *European Conference on Computer Vision*, pages 191–207. Springer, 2020.

[34] Shangchen Zhou, Chongyi Li, and Chen Change Loy. Lednet: Joint low-light enhancement and deblurring in the dark. In *ECCV*, 2022.

[35] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *Deep learning in medical image analysis and multimodal learning for clinical decision support*, pages 3–11. Springer, 2018.